Survey of Digital Preservation Systems

DLF Spring Forum 2007:
Pasadena, CA USA

April 24th, 2007

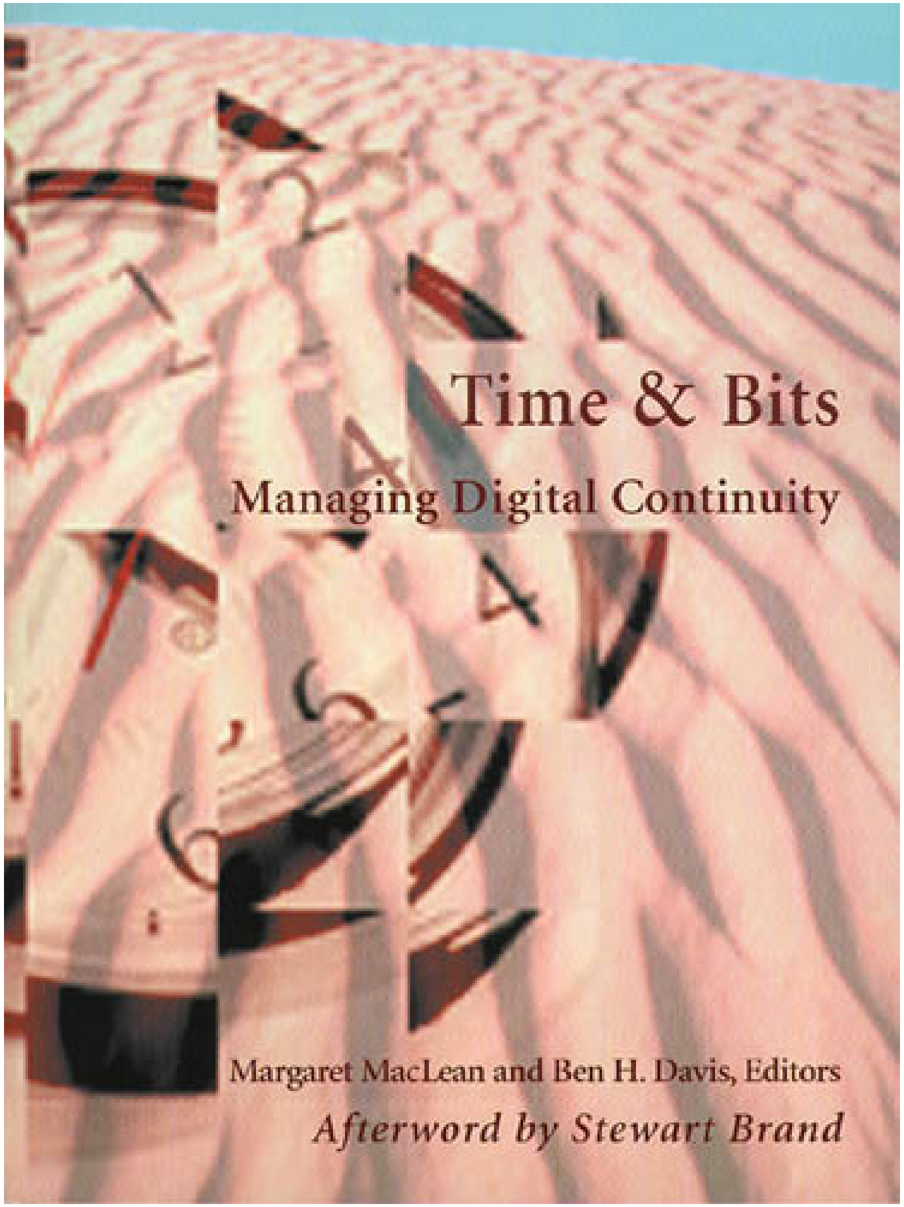
Commentator: Daniel Davis

Karim Boughida & Sally Hubbard

(Getty Research Institute)

Time & Bits (Getty 1998)

The Getty Information Institute and the Long Now Foundation. Largely visionary, and proceedings now available only in paper form as a Getty publication.



Time & Bits

Managing Digital Continuity

Margaret MacLean and Ben H. Davis, Editors

Afterword by Stewart Brand

Civilization now happens digitally. And it has no memory. This is no way to run a civilization.

(Steward Brand,
Time & Bits 1998)

Context

- Time and Bits (Getty 1998)
- Getty Research Institute TDR Working Group (2006)
- TDR vs DPS
- What are the characteristics of a DPS: OAIS and TDR-RLG (TRAC) compliance?
- Who is in production?

Getty Research Institute TDR working group (2006)

- Review of current preservation capability (minimal and uncoordinated)
- Lack of understanding of problem, e.g. why DAMs doesn’t provide solution
- Need to prepare for/deal with born digital as well as surrogate objects
- Need to upgrade/re-think infrastructure (e.g multi-tier storage a conceptual shift)

Getty Research Institute TDR working group (2006)

- Unable to recommend a technology at this time
- Developing policy
- Planning pilot
- Housekeeping: surveys of existing media in collections, etc.
- Propaganda/education

Survey Data

- First survey of its kind focusing on systems (RLG 1998, Cornell 2005)
- 316 responses
- Some preliminary email follow-up, No supplementary interviews

Survey

- Scope: Libraries, archives, museums, publishers, scientific data, historians (biased towards libraries, and English speaking world)
- Number of questions: 32
- Representative sample population
- Distributed via emails, listservs, blogs, phone
- No pre-testing for survey design
- No mandatory questions

Definitions

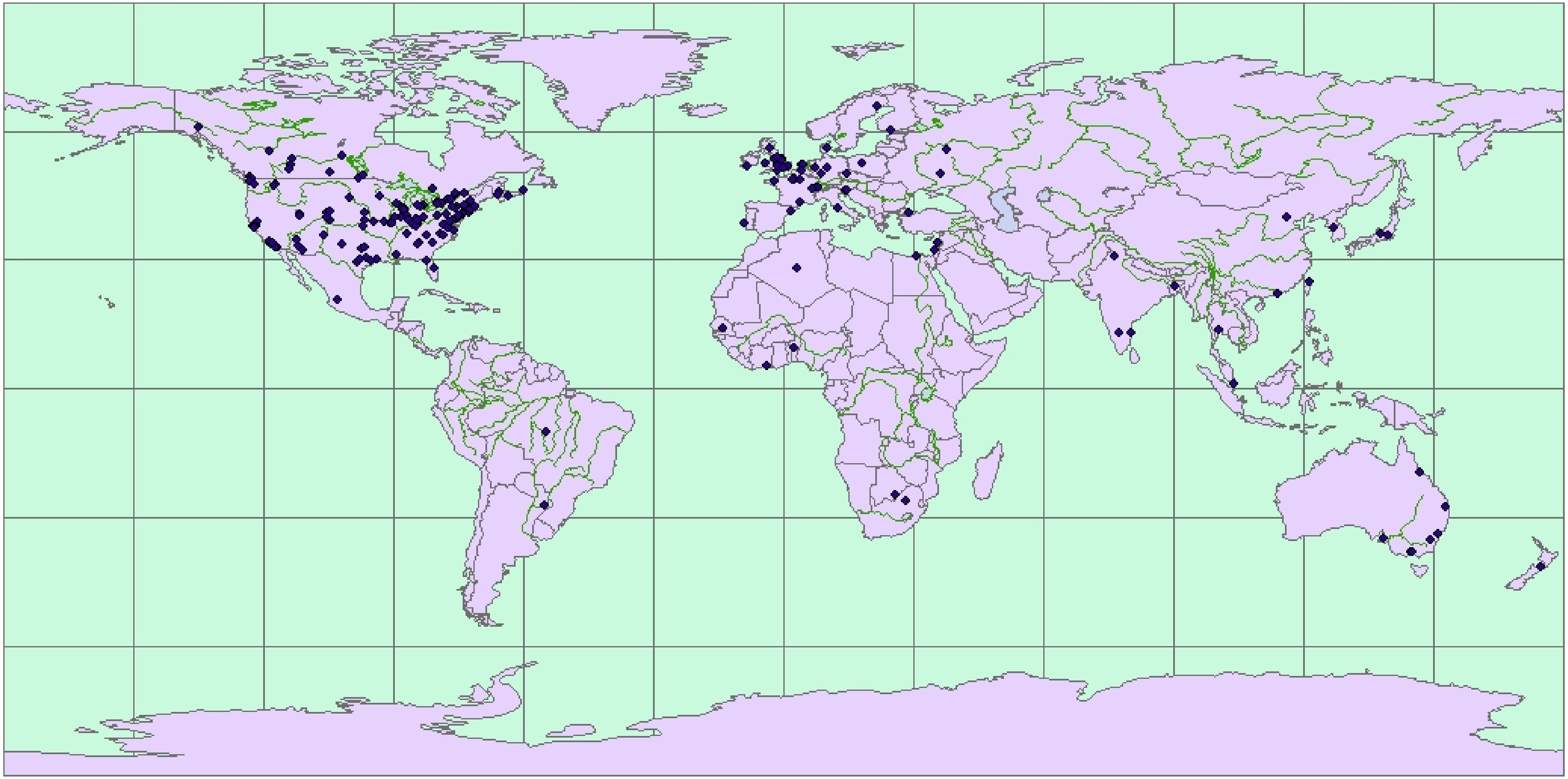
The survey gave only a broad definition of a DPS, and respondents used various meanings of the term:

- Any collections of digital material
- IR
- DAM
- OAISoid (resembling OAIS)
- OAIS

Definitions

- Definitions varied across fields as well as with depth of expertise or experience
- Points on a continuum, e.g. a DAM as a beginning?
- Reflective of a fundamental misunderstanding?
- Indicative of a new, coalescing, field?
- Probably all of the above

All Responses (316)



All Responses: North and Central America

[removed temporarily]

All Responses: EU

[removed temporarily]

All Responses: Africa

[removed temporarily]

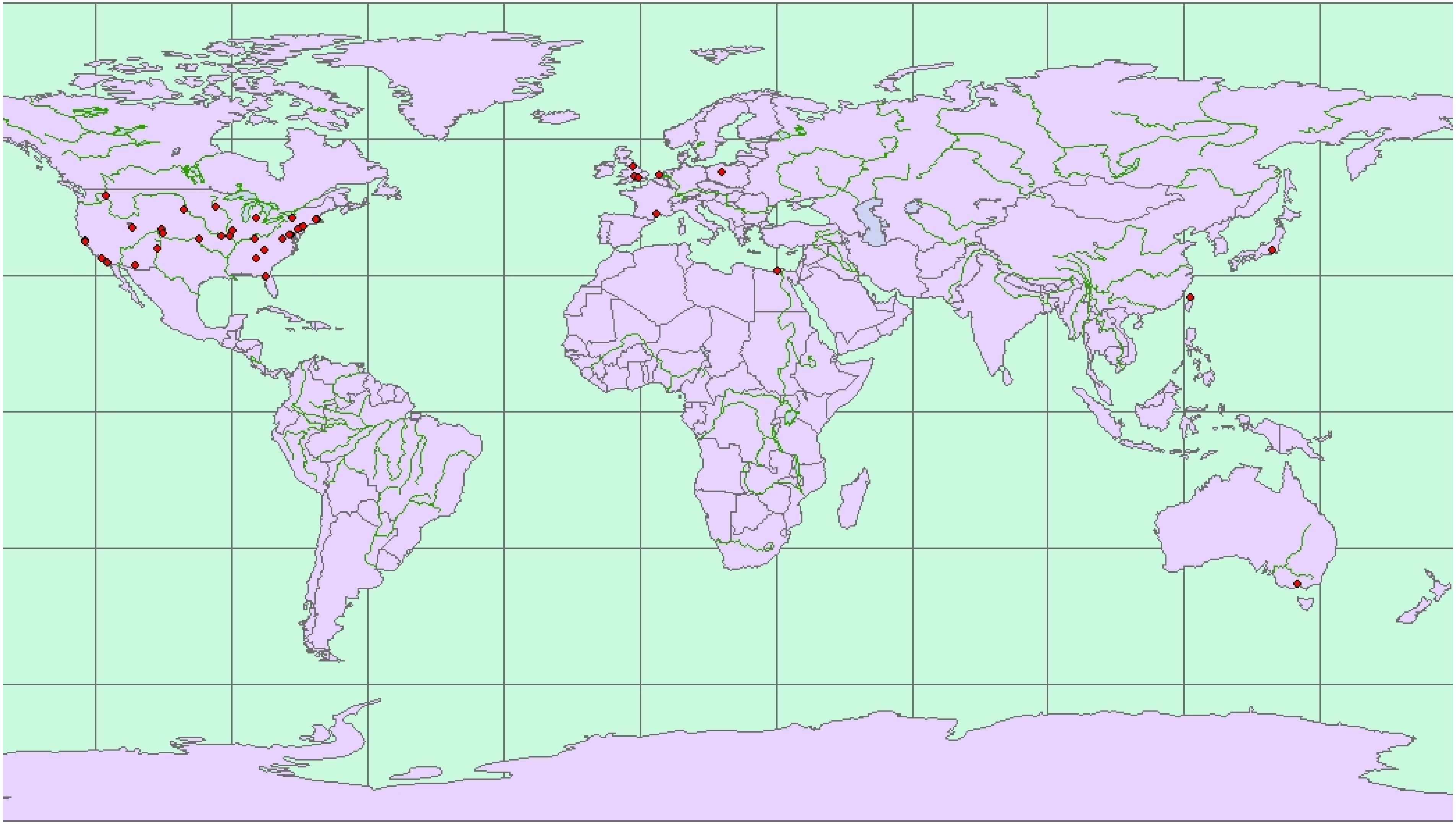
and S.America

[removed temporarily]

All Responses: Asia

[removed temporarily]

All Responses: Implementation (46)



Implementation: Americas

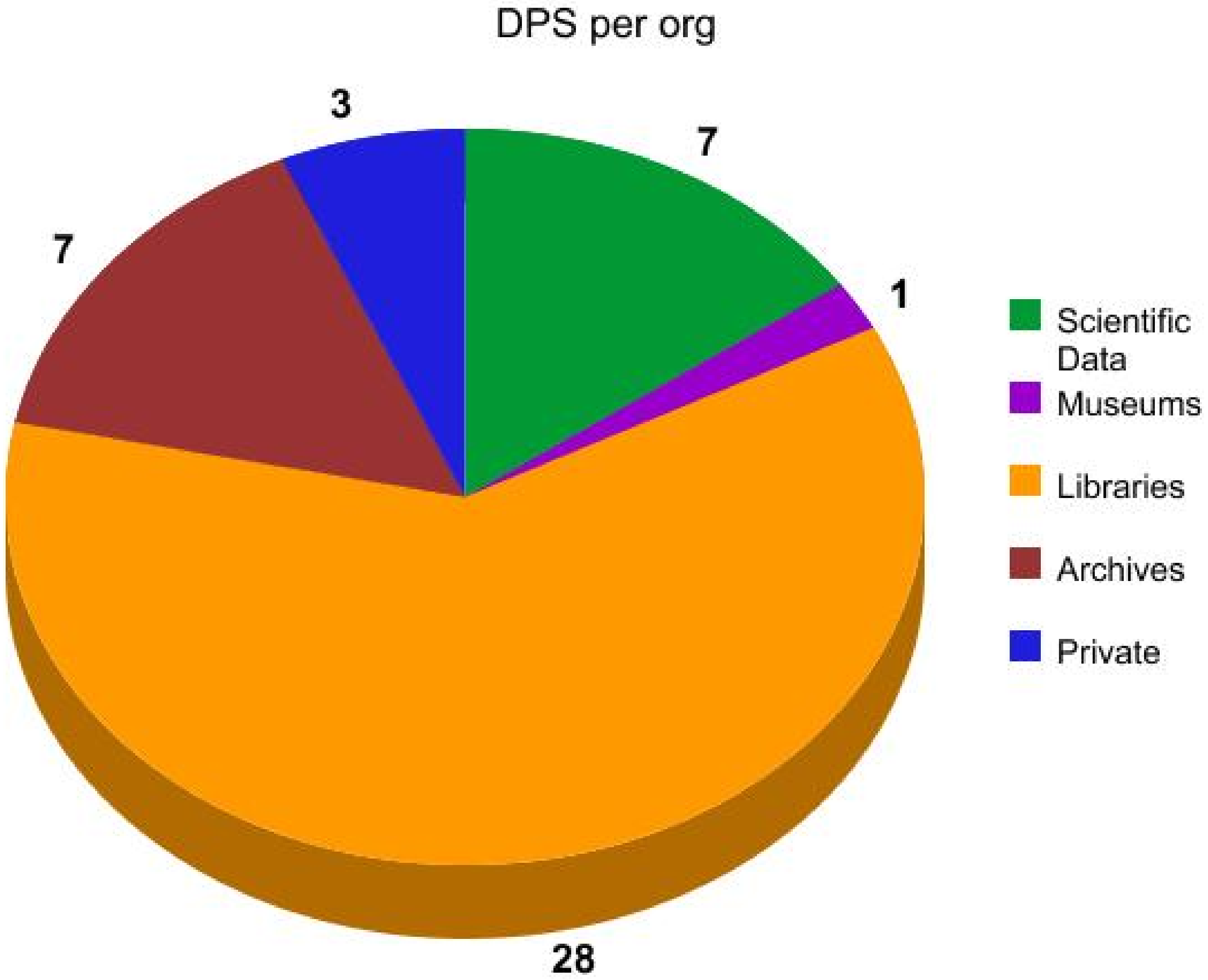
[removed temporarily]

Implementation: EU, Africa

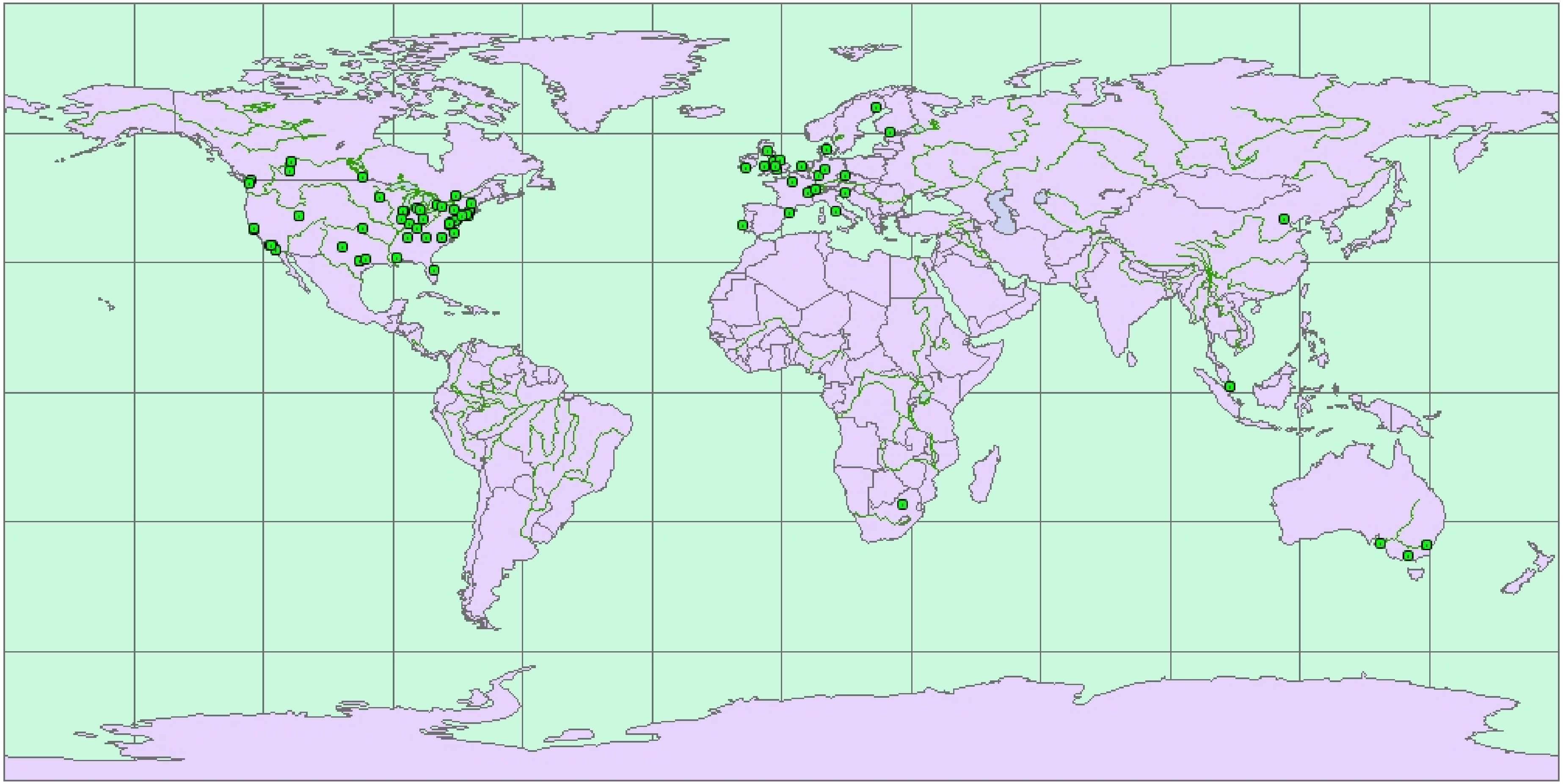
[removed temporarily]

Implementation: Asia, Australia

[removed temporarily]


DPS per org
3
7
7
1
28
Scientific Data
Museums
Libraries
Archives
Private

All Responses: Planning (96)



Planning: Americas

[removed temporarily]

Planning: EU

[removed temporarily]

Planning: Asia, Australia

[removed temporarily]

Planning: Africa

[removed temporarily]

DLF members and allies

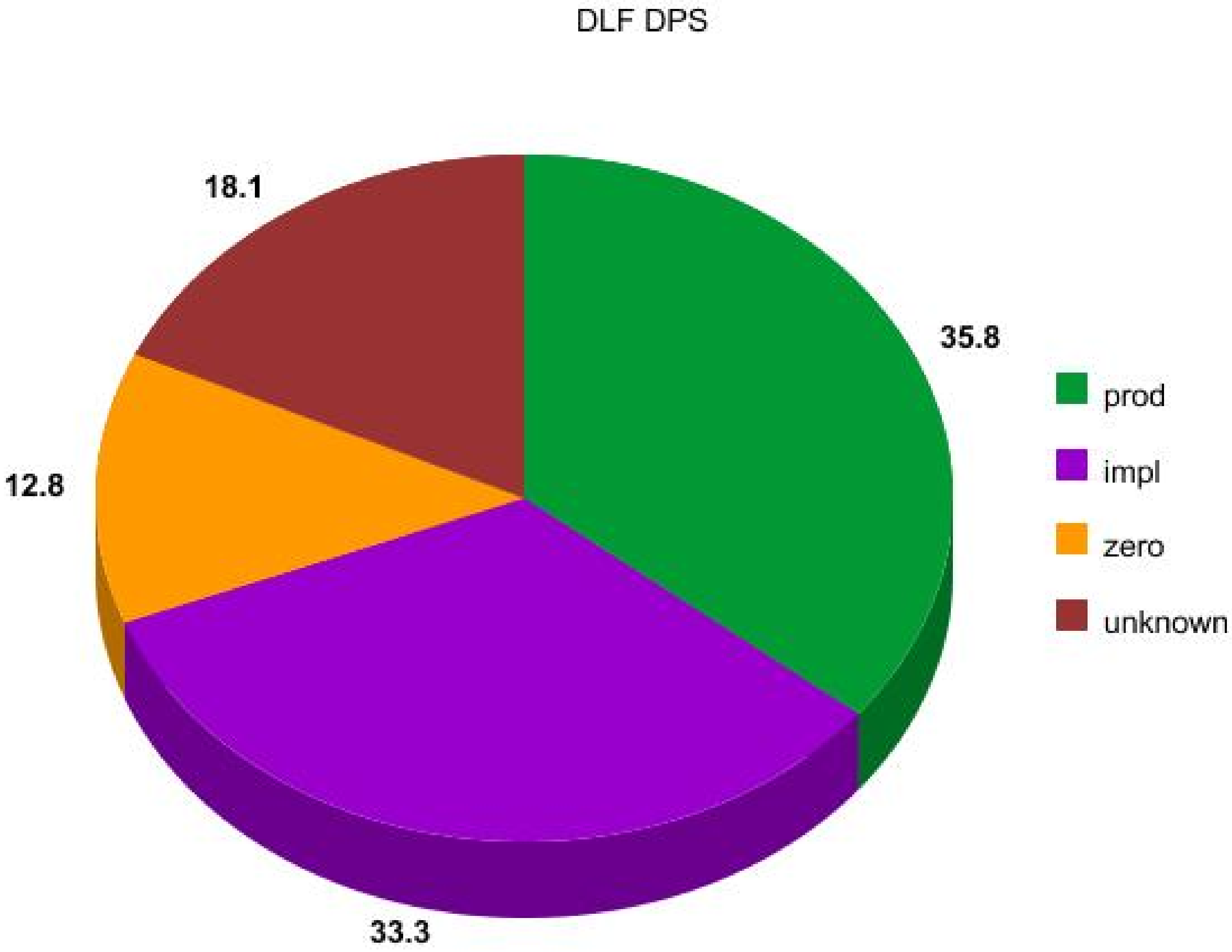
- 39 members and allies
- 84.7 % responded (no response from 5)
- 5 or 12.8 %: No planning or implementation projects (assumption)
- Prod and impl-plan: next slide

DLF: DPS in production

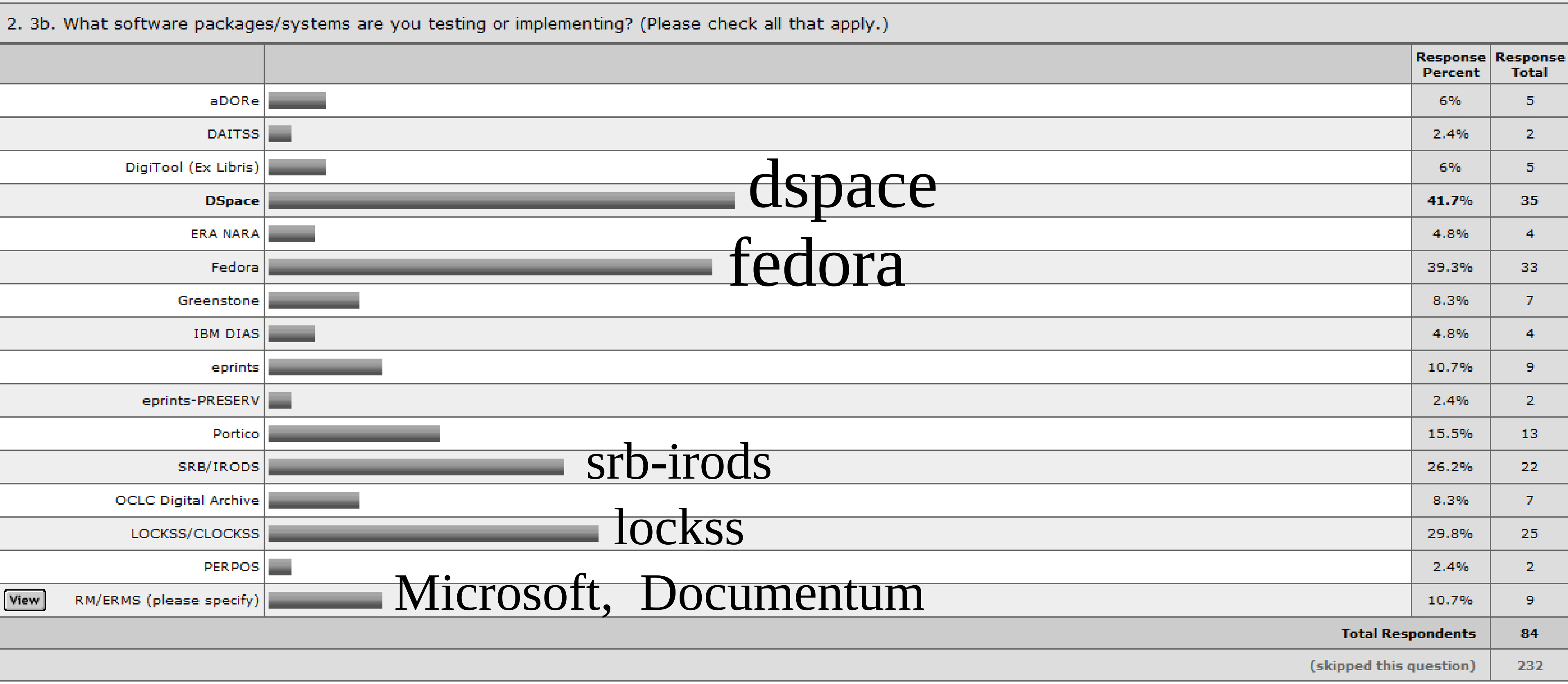
- 14 in production: Government and university libraries
- 35.8 % are in production

DLF: DPS in impl.-plan.

- 13 implementing-planning: all university libraries
- 33.3 % impl-plan
- RECAP: 35.8 prod + 33.3 impl + 12.8 no= 81.9 %
- 18.1 % (unknown) [see pie next]


DLF DPS
18.1
35.8
12.8
33.3
prod
impl
zero
unknown

Software packages/systems



2. 3b. What software packages/systems are you testing or implementing? (Please check all that apply.)

		Response Percent	Response Total
aDORe		6%	5
DAITSS		2.4%	2
DigiTool (Ex Libris)		6%	5
DSpace		**41.7%**	**35**
ERA NARA		4.8%	4
Fedora		39.3%	33
Greenstone		8.3%	7
IBM DIAS		4.8%	4
eprints		10.7%	9
eprints-PRESERV		2.4%	2
Portico		15.5%	13
SRB/IRODS		26.2%	22
OCLC Digital Archive		8.3%	7
LOCKSS/CLOCKSS		29.8%	25
PERPOS		2.4%	2
View RM/ERMS (please specify)		10.7%	9
		Total Respondents	**84**
		(skipped this question)	232

dspace

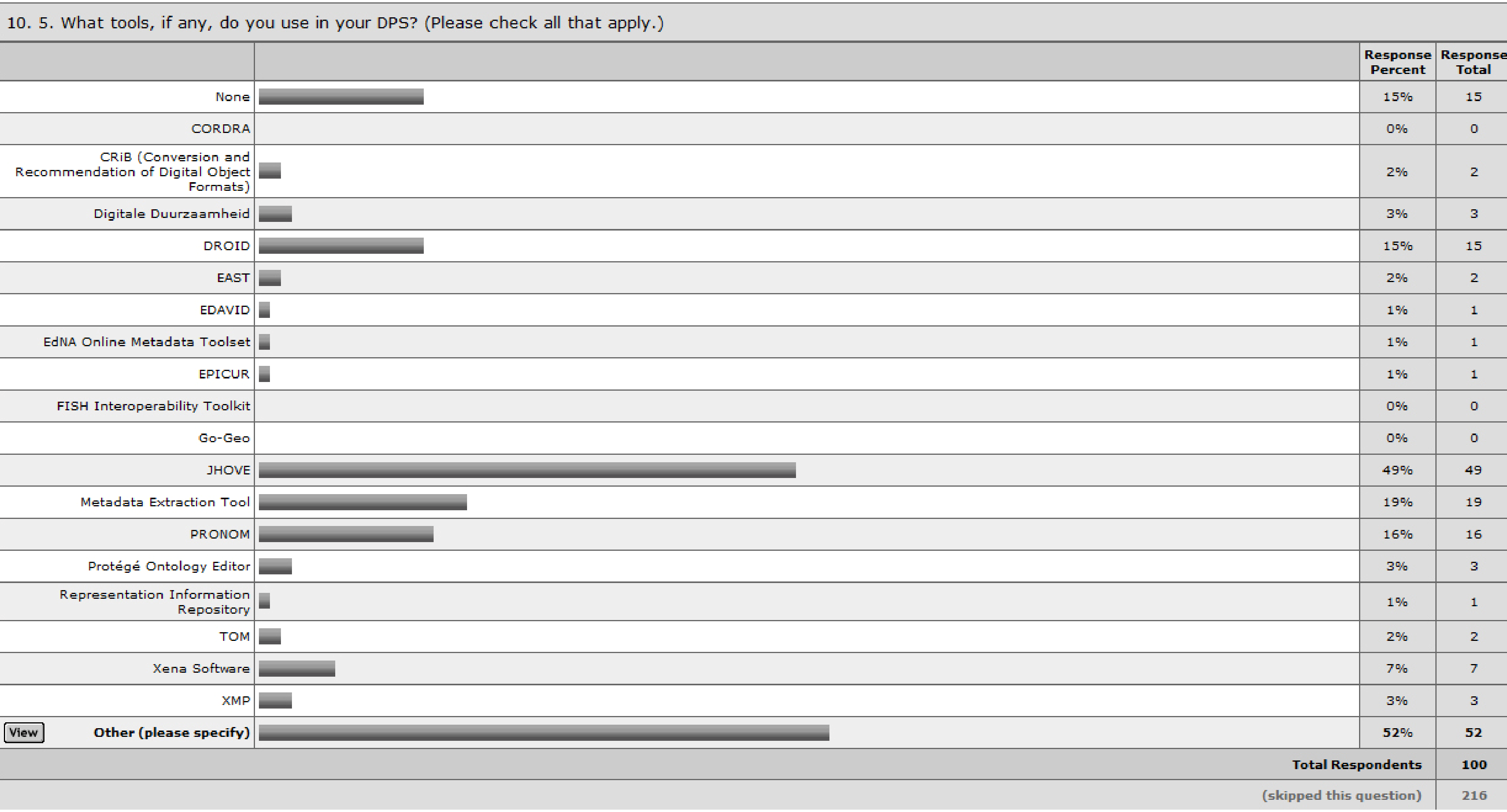
fedora

srb-irods

lockss

Microsoft, Documentum

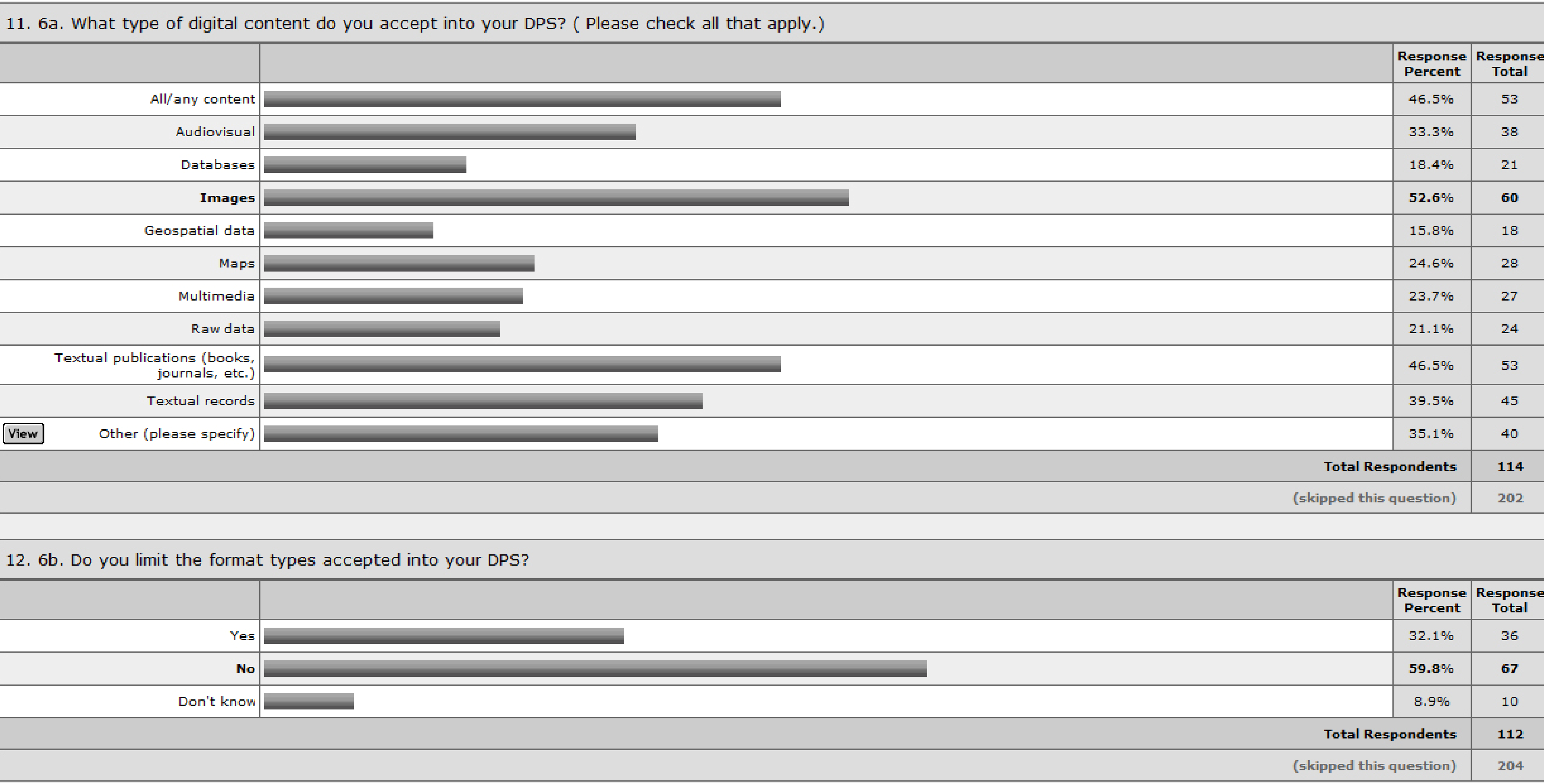
Tools



10. 5. What tools, if any, do you use in your DPS? (Please check all that apply.)

	Response Percent	Response Total
None	15%	15
CORDRA	0%	0
CRiB (Conversion and Recommendation of Digital Object Formats)	2%	2
Digitale Duurzaamheid	3%	3
DROID	15%	15
EAST	2%	2
EDAVID	1%	1
EdNA Online Metadata Toolset	1%	1
EPICUR	1%	1
FISH Interoperability Toolkit	0%	0
Go-Geo	0%	0
JHOVE	49%	49
Metadata Extraction Tool	19%	19
PRONOM	16%	16
Protégé Ontology Editor	3%	3
Representation Information Repository	1%	1
TOM	2%	2
Xena Software	7%	7
XMP	3%	3
View **Other (please specify)**	**52%**	**52**
	Total Respondents	**100**
	(skipped this question)	216

Digital Content/Format



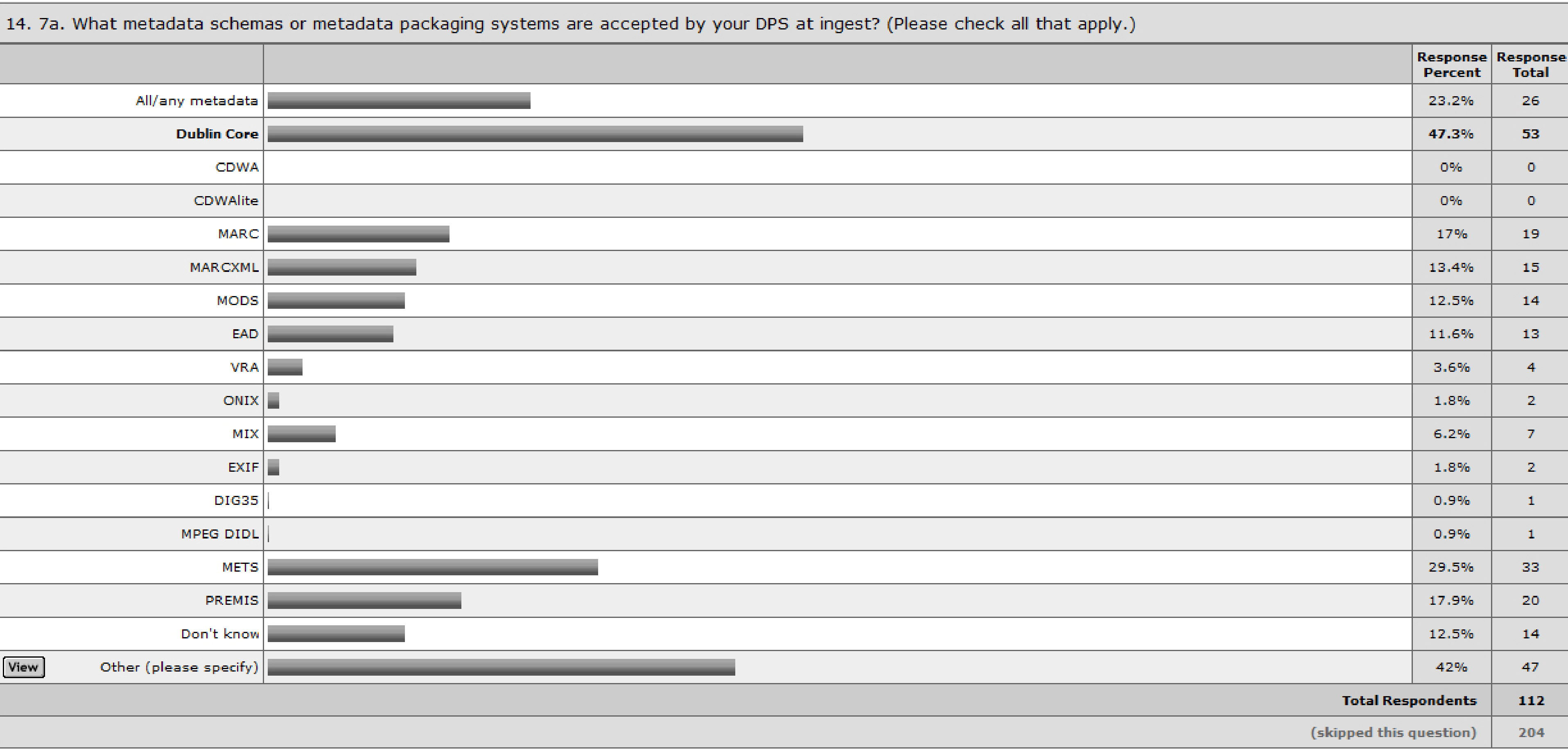
11. 6a. What type of digital content do you accept into your DPS? (Please check all that apply.)

	Response Percent	Response Total
All/any content	46.5%	53
Audiovisual	33.3%	38
Databases	18.4%	21
Images	**52.6%**	**60**
Geospatial data	15.8%	18
Maps	24.6%	28
Multimedia	23.7%	27
Raw data	21.1%	24
Textual publications (books, journals, etc.)	46.5%	53
Textual records	39.5%	45
View Other (please specify)	35.1%	40
	Total Respondents	**114**
	(skipped this question)	202

12. 6b. Do you limit the format types accepted into your DPS?

	Response Percent	Response Total
Yes	32.1%	36
No	**59.8%**	**67**
Don't know	8.9%	10
	Total Respondents	**112**
	(skipped this question)	204

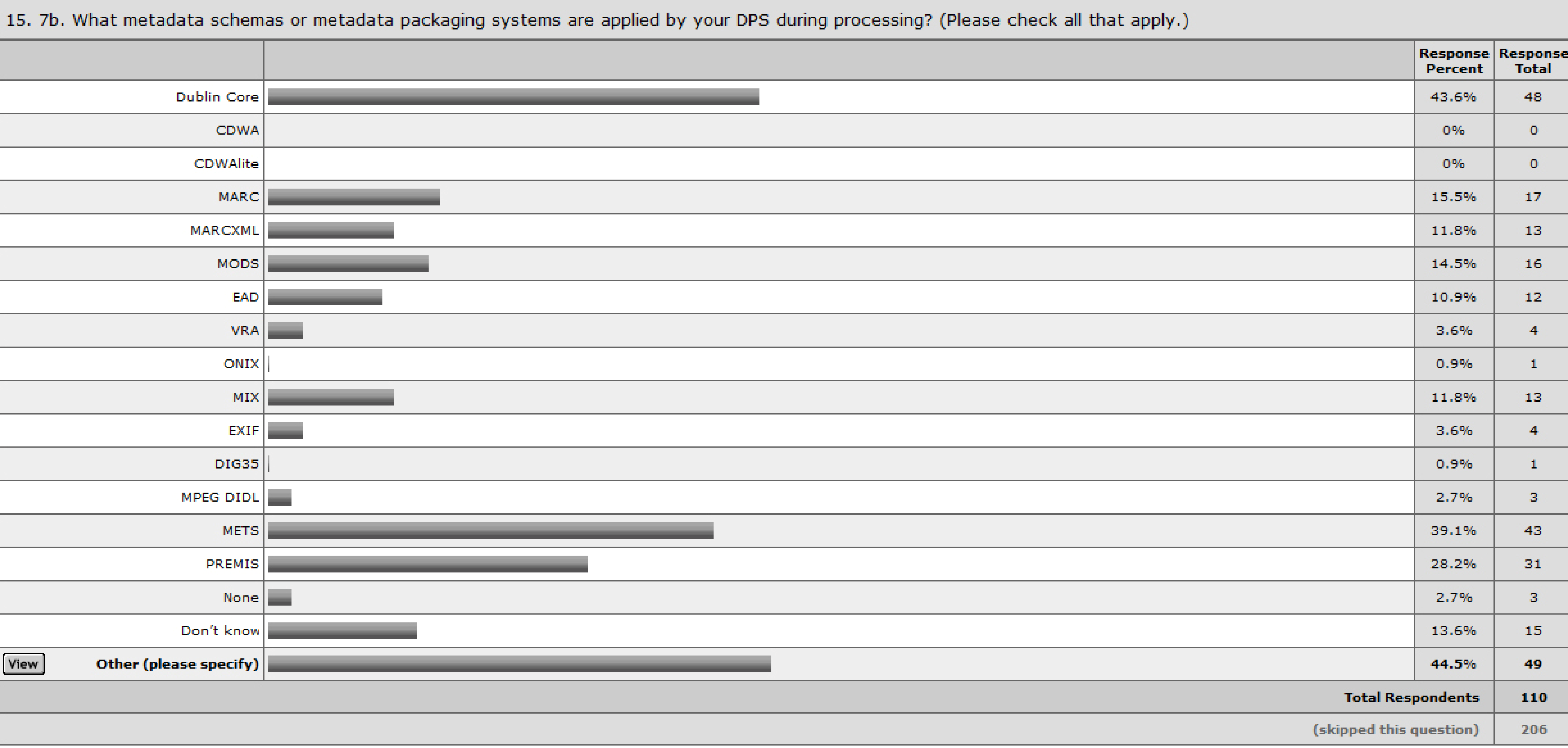
Metadata at Ingest



14. 7a. What metadata schemas or metadata packaging systems are accepted by your DPS at ingest? (Please check all that apply.)

	Response Percent	Response Total
All/any metadata	23.2%	26
Dublin Core	**47.3%**	**53**
CDWA	0%	0
CDWAlite	0%	0
MARC	17%	19
MARCXML	13.4%	15
MODS	12.5%	14
EAD	11.6%	13
VRA	3.6%	4
ONIX	1.8%	2
MIX	6.2%	7
EXIF	1.8%	2
DIG35	0.9%	1
MPEG DIDL	0.9%	1
METS	29.5%	33
PREMIS	17.9%	20
Don't know	12.5%	14
View Other (please specify)	42%	47
	Total Respondents	**112**
	(skipped this question)	204

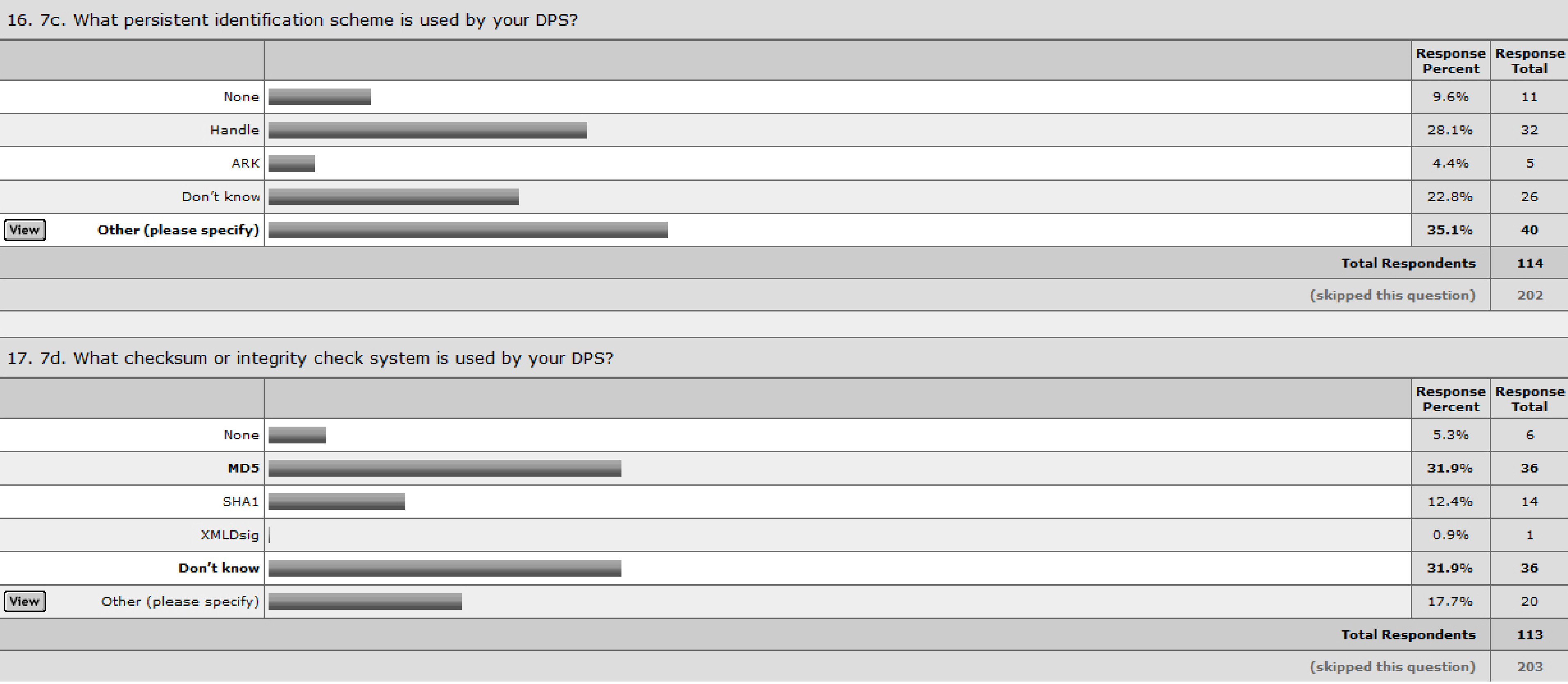
Metadata Applied



15. 7b. What metadata schemas or metadata packaging systems are applied by your DPS during processing? (Please check all that apply.)

		Response Percent	Response Total
Dublin Core		43.6%	48
CDWA		0%	0
CDWAlite		0%	0
MARC		15.5%	17
MARCXML		11.8%	13
MODS		14.5%	16
EAD		10.9%	12
VRA		3.6%	4
ONIX		0.9%	1
MIX		11.8%	13
EXIF		3.6%	4
DIG35		0.9%	1
MPEG DIDL		2.7%	3
METS		39.1%	43
PREMIS		28.2%	31
None		2.7%	3
Don't know		13.6%	15
View **Other (please specify)**		**44.5%**	**49**
		Total Respondents	**110**
		(skipped this question)	206

Identification/Integrity



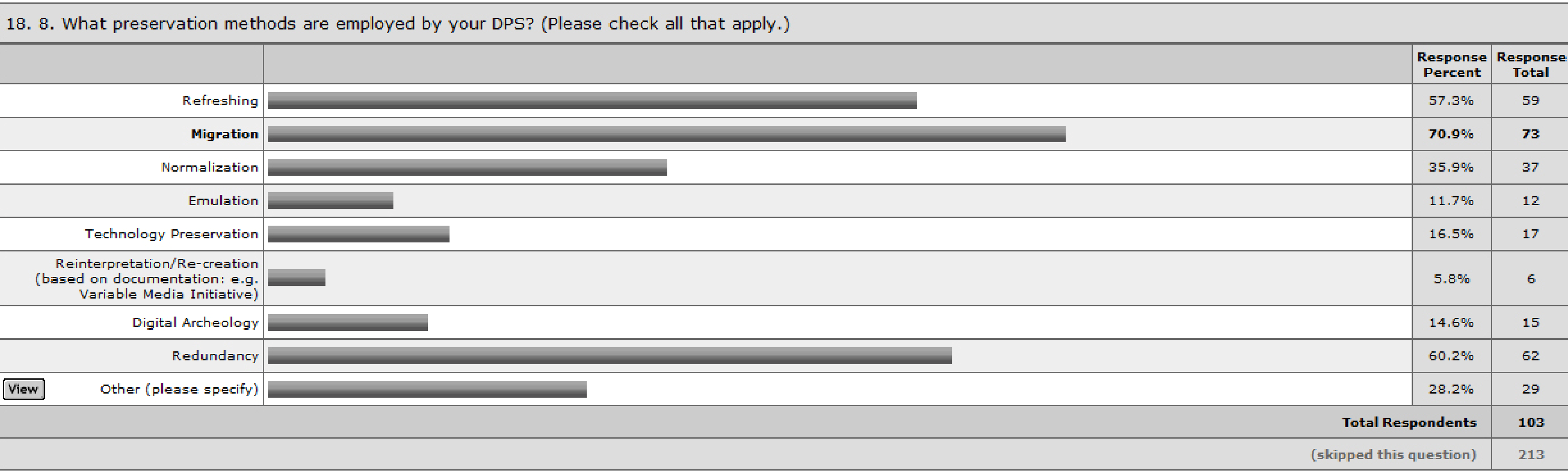
16. 7c. What persistent identification scheme is used by your DPS?

		Response Percent	Response Total
	None	9.6%	11
	Handle	28.1%	32
	ARK	4.4%	5
	Don't know	22.8%	26
View	**Other (please specify)**	**35.1%**	**40**
		Total Respondents	**114**
		(skipped this question)	202

17. 7d. What checksum or integrity check system is used by your DPS?

		Response Percent	Response Total
	None	5.3%	6
	MD5	**31.9%**	**36**
	SHA1	12.4%	14
	XMLDsig	0.9%	1
	Don't know	**31.9%**	**36**
View	Other (please specify)	17.7%	20
		Total Respondents	**113**
		(skipped this question)	203

Preservation Methods



18. 8. What preservation methods are employed by your DPS? (Please check all that apply.)

		Response Percent	Response Total
Refreshing		57.3%	59
Migration		**70.9%**	**73**
Normalization		35.9%	37
Emulation		11.7%	12
Technology Preservation		16.5%	17
Reinterpretation/Re-creation (based on documentation: e.g. Variable Media Initiative)		5.8%	6
Digital Archeology		14.6%	15
Redundancy		60.2%	62
Other (please specify)		28.2%	29
Total Respondents			**103**
(skipped this question)			213

How many TBs?

- Biggest: 25 PB (scientific data community), 1<x<9 PB (national library),4 PB (university library)
- 66 % have less then 10 TB

Storage costs per TB

- Variable: 3 Main Classes
  - Low Range: $400
  - Mid Range: $2000-$4000
  - High Range: $10,000-$15,000 (usually backup incl.)
- “Much not yet analyzed”

How many digital objects or records were ingested in the last year [2006]?

- Biggest: 360 Million, 80 Million, 30 Million, 20 Million, 10 Million
- No clear pattern, but roughly one third have fewer then 10,000 objects

What is your estimated annual growth per year? (in TB)

- Biggest: 4 PB, 1 PB (scientific data)
- Medium: 200 TB, 90 TB (national libraries)
- One third have less than 10 TB

What is the average size of digital objects or records? (in MB)

- Broad range: 1 KB – 6 GB
- The institution that ingested 20 million per year, has an average object size of 1 MB
- The institution that ingested 80 million has “too many variations to determine meaningful size”

Hardware/Storage

- IBM
- CAS (content addressed storage)
  - SUN
  - EMC (Microsoft ally)

What are the start up costs for your DPS?

- Subscription Model: $13,500
- “Big” (governmental/national libraries): $40 Million, $21 Million, $NZ 24 Million
- “Medium” (scientific data and national libraries): $3.5 Million
- “Small” (university and state libraries): $180,000, $150,000

What are the annual costs for your DPS?

- Subscription Model: $13,500
- “Big” (governmental/national libraries): $10 Million, $5.3 Million
- “Medium” (scientific data and national libraries): None given
- “Small” (university and state libraries): None given

Next Steps

- Distribution of preliminary results to respondents (in spreadsheet form)
- More detailed follow up of survey
- Possible 10th anniversary revisiting of Time and Bits
- A Getty digital preservation system …

{kboughida; shubbard}@getty.edu

www.getty.edu/research/institute

Tel: 310-440-7335

The Getty Research Institute
1200 Getty Center Drive, Suite 1100
Los Angeles, CA 90049-1688 USA