PARSE insight

Permanent Access to the Records of Science in Europe

Community insight

Approach & overview of results

International workshop on Preservation, Access and Re-use of Scientific Data

21 September 2009

Jeffrey van der Hoeven (Koninklijke Bibliotheek)

KB

Koninklijke Bibliotheek

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME

Overview

- Community Insight
  - Objectives
  - Approach
  - Stakeholders
- Highlights of survey results
- Implications for the roadmap

Objectives

"Gain insight into what is happening regarding digital preservation in research in Europe."

About 1.33 million researchers in Europe!

Source: Eurostat (figure based on 2006)

Approach

- Top-down:
  - Desk research
  - Targeted surveys to stakeholders in science
  - Interviews
  - Workshops and conferences
- Bottom-up: case studies in 3 communities:
  - Case 1: High Energy Physics (HEP)
  - Case 2: Earth Observation (EO)
  - Case 3: Social Sciences & Humanities (SSH)

Insight: stakeholders

Funding/policy

- National Funding organisations
- European funding
- Corporate funding

Research

- Research institutes (non-profit)
- Universities
- Academic libraries

Publishing

- General (cross-community) publishers
- Specific (community) publishers

Data management (preservation)

- Data centres (profit / non-profit)
- Libraries
- Archives

Surveys to stakeholders

Funding/policy

ESF, Alliance for Permanent Access, national funding agencies

Research

Elsevier mailinglist (35,000 people), ESF, MCFA, Eurodoc, ALLEA, YEAR, Digital Humanities Observatory, etc.

Data management (preservation)

LIBER, DPE, DPC, NCDD, DCC, D-lib Magazine, PADI, JISC mailing lists, CASPAR, Planets, etc.

Publishing

International Association of STM publishers, Directory of Open Access Journals (DOAJ)

Surveys to stakeholders

Funding/policy

< responses

Research

1397 responses

Data management (preservation)

273 responses

Publishing

186 responses

In general

- **Preservation of digital data =**

  the process of storing digital information in such a way that it remains accessible, understandable and usable over the long term.

- **Digital research data =**

  whole spectrum of research output in digital format. This includes raw data and publications.

About researchers

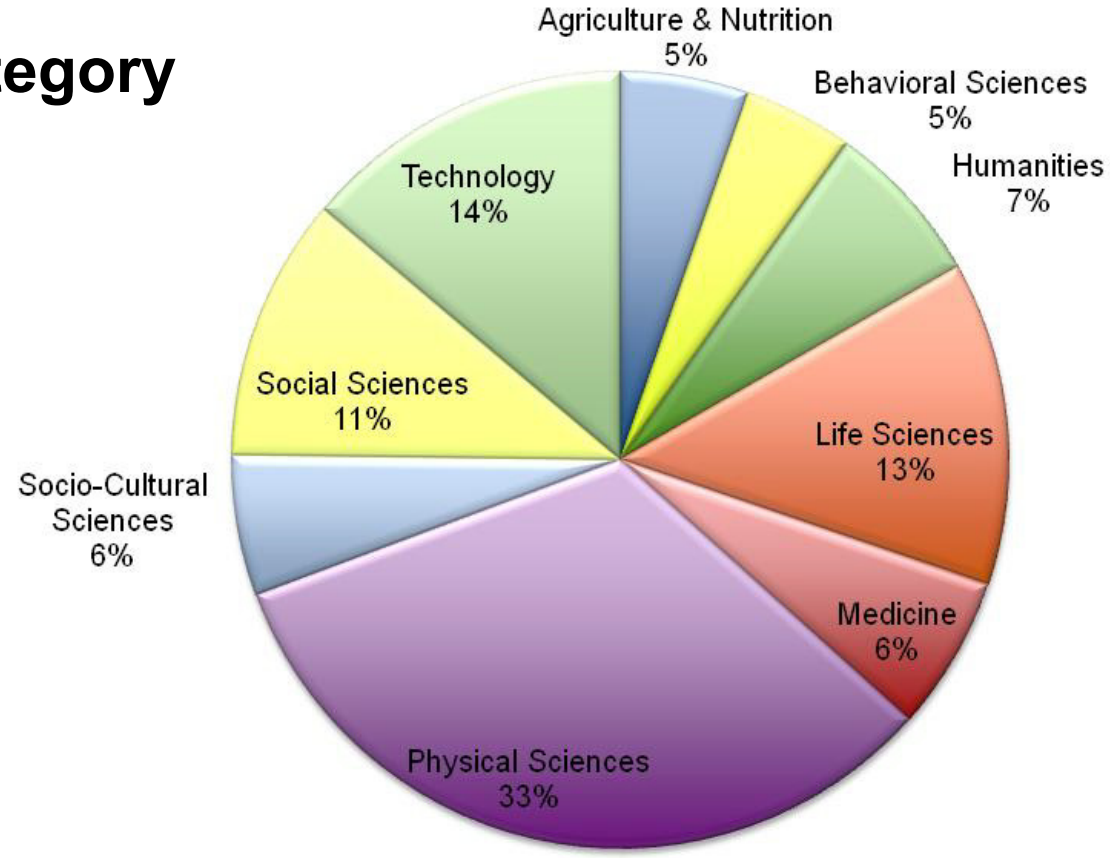
Communities aggregated to:

- Agriculture & Nutrition
- Behavioural Sciences
- Humanities
- Life Sciences
- Medicine
- Social Sciences
- Physical Sciences
- Socio-Cultural Sciences
- Technology

Based on KNAW classification (Royal Netherlands Academy of Arts and Sciences)

About researchers

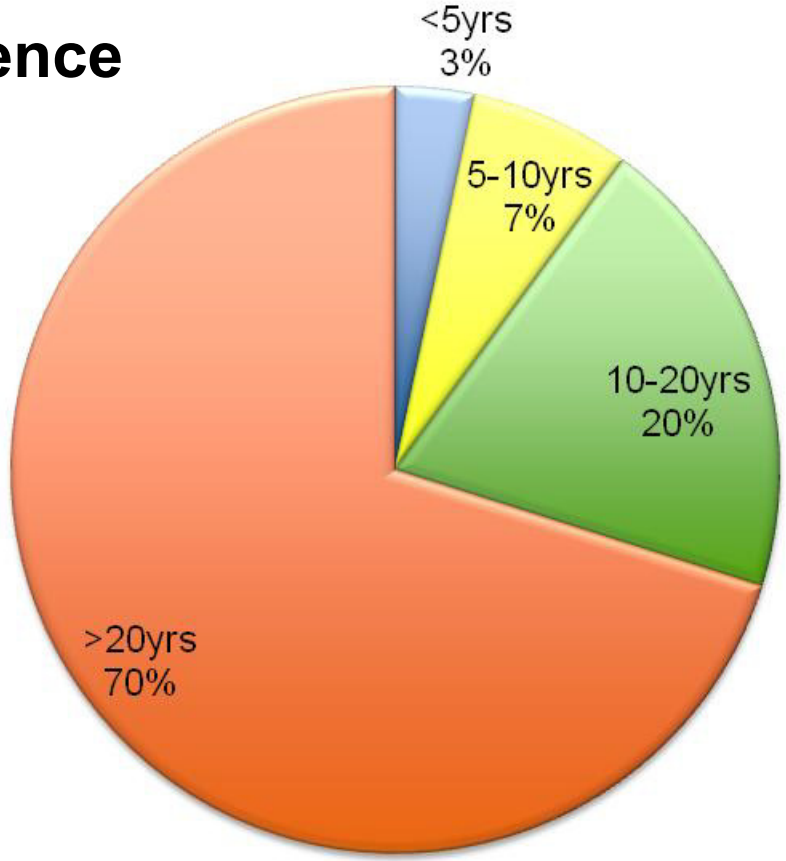
Per category



EU 44%, USA 33%, Other 23%

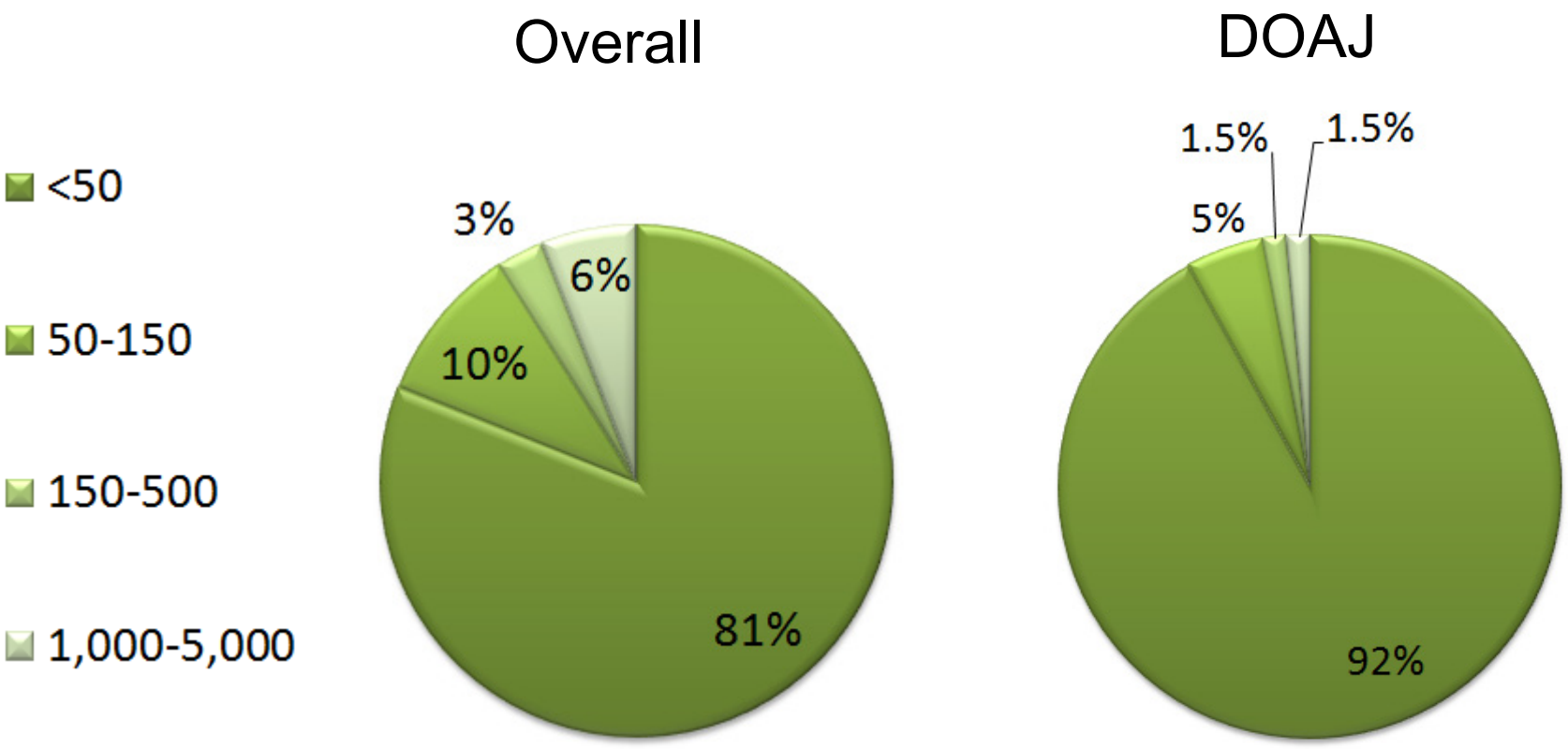
About researchers

Level of experience



About publishers

Number of journals



Koninklijke Bibliotheek

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME

Survey topics & structure

- Perceptions of preservation
- Preservation in research – the state of affairs
- Preservation in research – the outlook
- Cross-disciplinary use of research data
- Roles and responsibilities

(R) = Research

(DM) = Data Management

(P) = Publishing

Perceptions of preservation

Reasons for preservation

1. It is unique.
2. It potentially has economic value.
3. It may stimulate inter-disciplinary collaborations.
4. It allows for re-analysis of existing data.
5. It may serve validation purposes in the future.
6. It will stimulate the advancement of science (new research can build on existing knowledge).
7. If research is publicly funded, the results should become public property and therefore properly preserved.

Reasons for preservation (R)

It is unique

It potentially has economic value

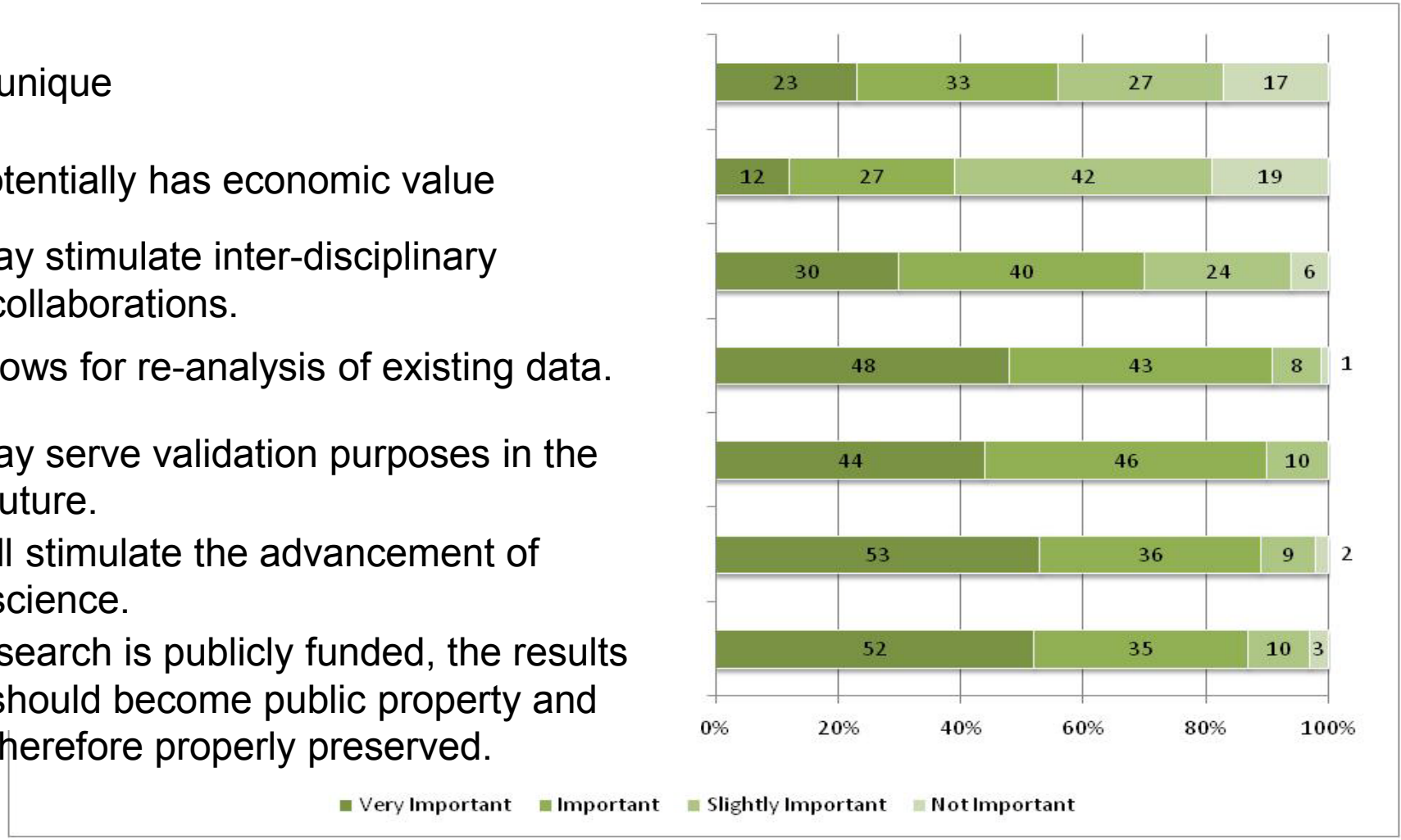
It may stimulate inter-disciplinary collaborations.

It allows for re-analysis of existing data.

It may serve validation purposes in the future.

It will stimulate the advancement of science.

If research is publicly funded, the results should become public property and therefore properly preserved.



Reasons for preservation (DM)

It is unique

It potentially has economic value

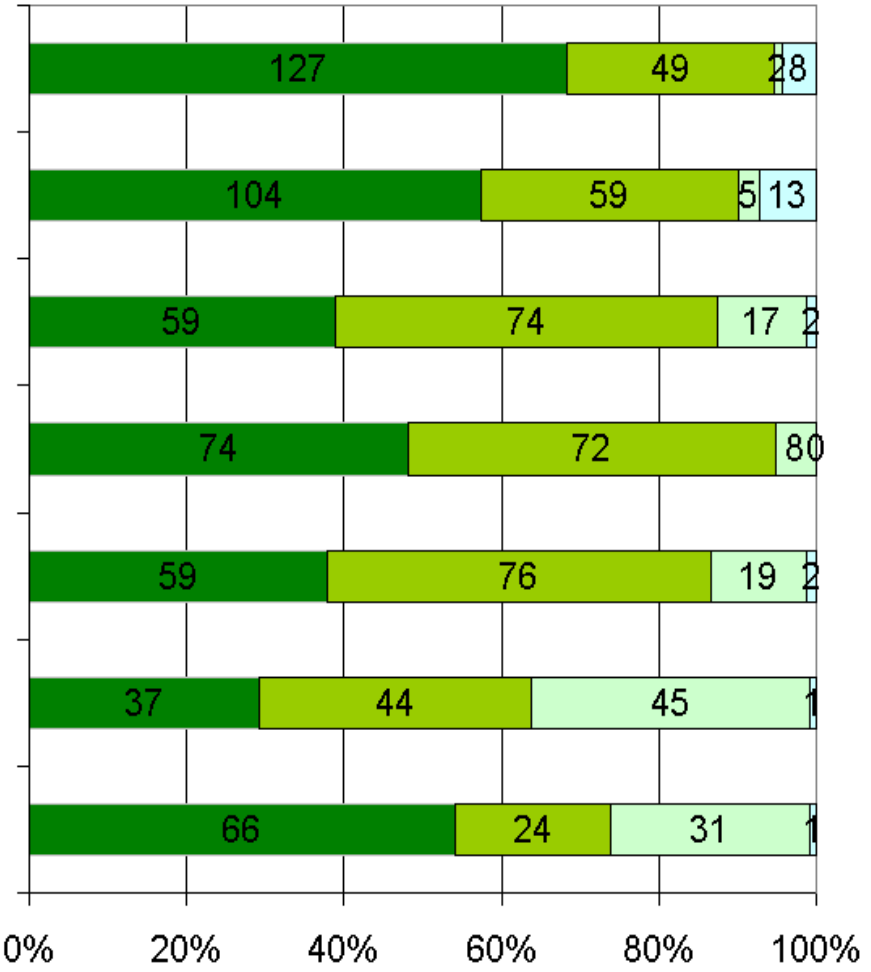
It may stimulate inter-disciplinary collaborations.

It allows for re-analysis of existing data.

It may serve validation purposes in the future.

It will stimulate the advancement of science.

If research is publicly funded, the results should become public property and therefore properly preserved.



Koninklijke Bibliotheek

Reasons for preservation (P)

It is unique

It potentially has economic value

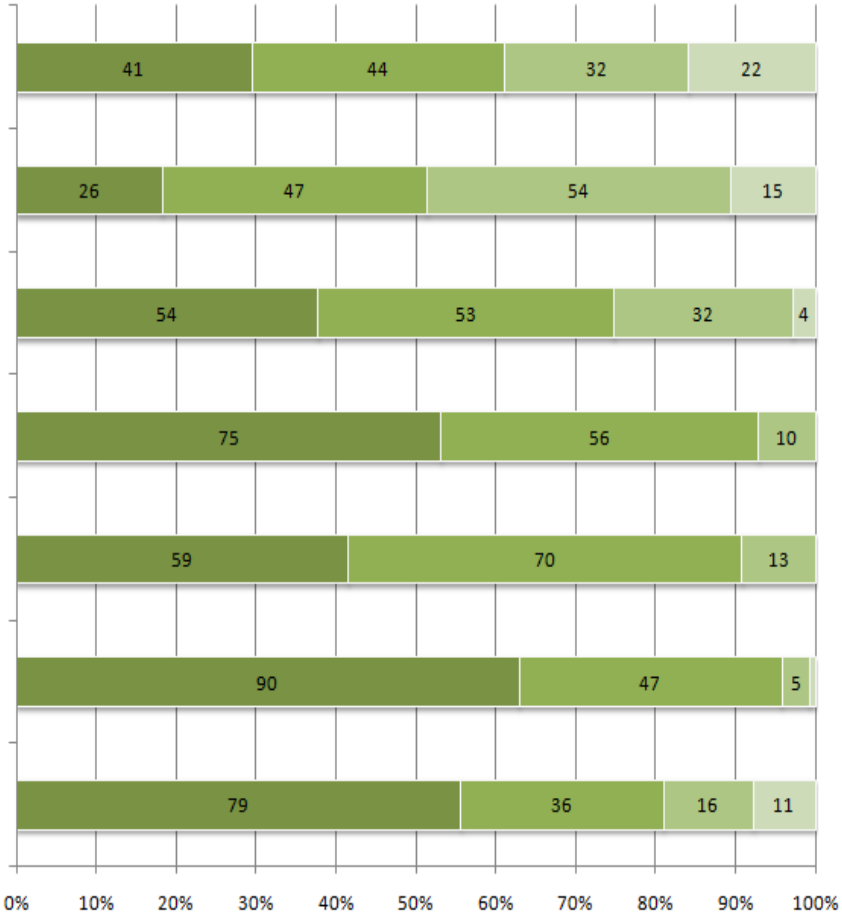
It may stimulate inter-disciplinary collaborations.

It allows for re-analysis of existing data.

It may serve validation purposes in the future.

It will stimulate the advancement of science.

If research is publicly funded, the results should become public property and therefore properly preserved.



Threats to preservation

1. The ones we trust to look after the digital holdings may let us down.
2. The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future.
3. Loss of ability to identify the location of data.
4. Access and use restrictions (e.g. Digital Rights Management) may not be respected in the future.
5. Evidence may be lost because the origin and authenticity of the data may be uncertain.
6. Lack of sustainable hardware, software or support of computer environment may make the information inaccessible.
7. Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved.

Threats to preservation (R)

The ones we trust to look after the digital holdings may let us down

The current custodian of the data may cease to exist

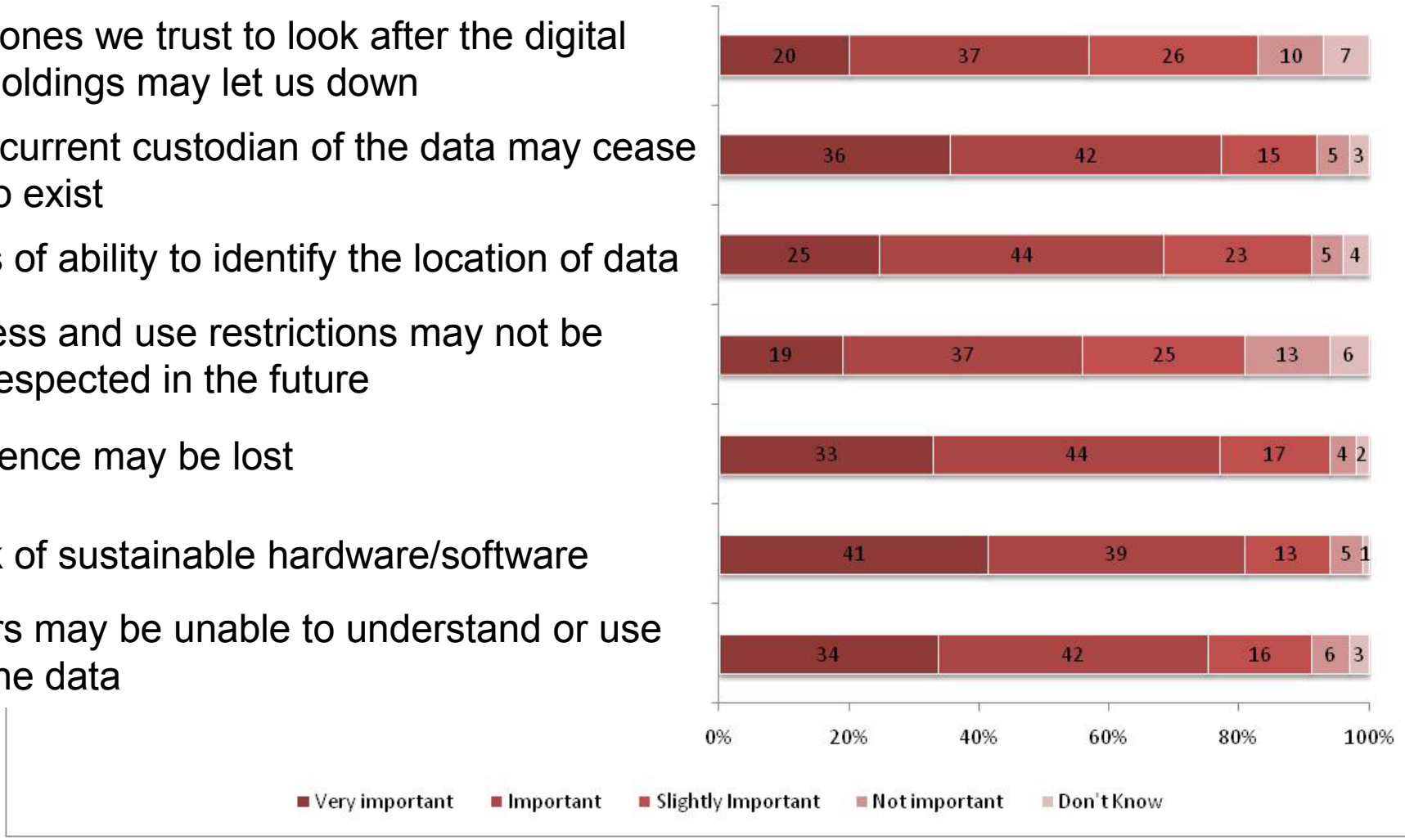
Loss of ability to identify the location of data

Access and use restrictions may not be respected in the future

Evidence may be lost

Lack of sustainable hardware/software

Users may be unable to understand or use the data



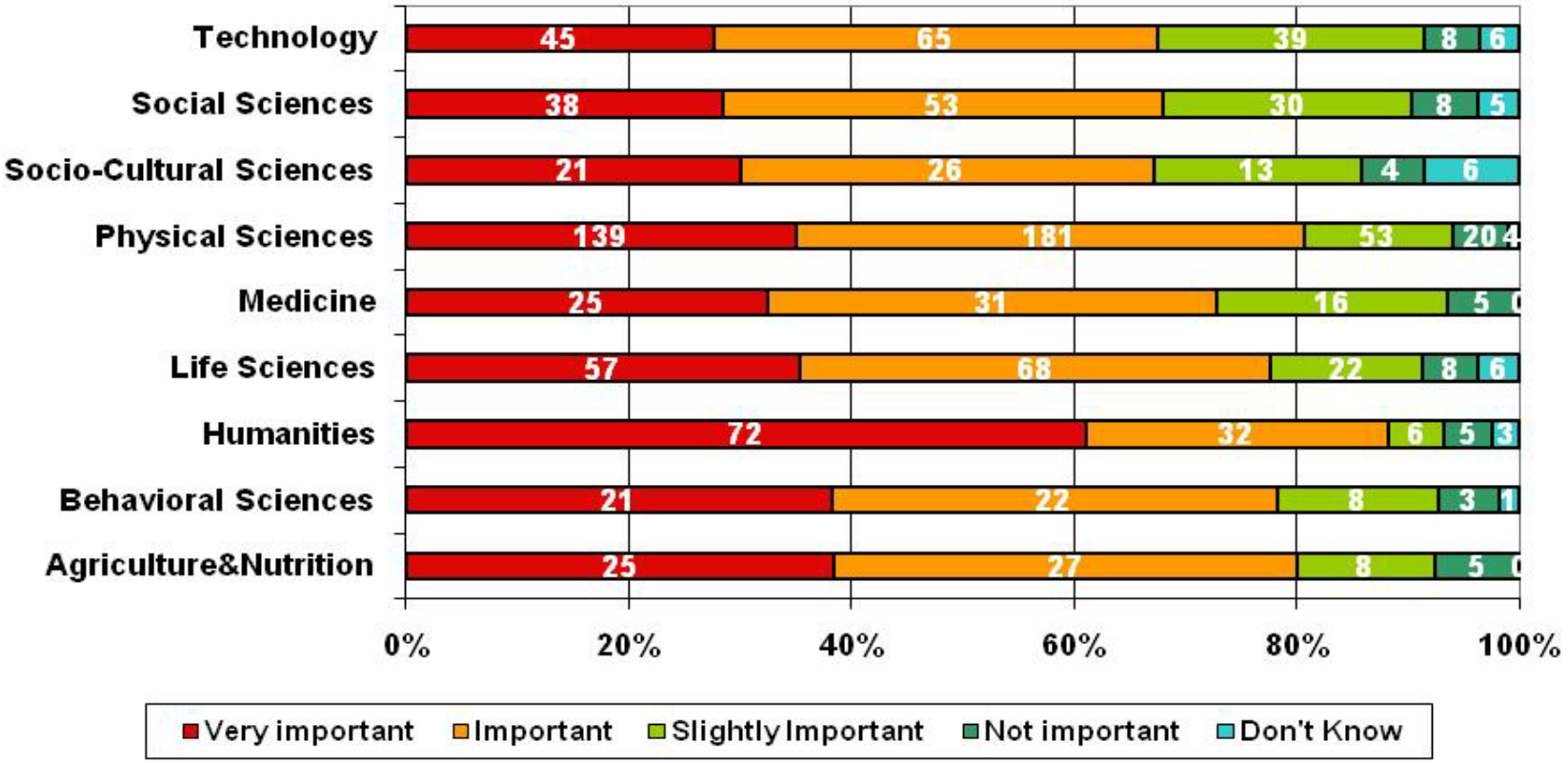
Koninklijke Bibliotheek

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME

Threats to preservation (R)

Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved.



Threats to preservation (DM)

The ones we trust to look after the digital holdings may let us down

The current custodian of the data may cease to exist

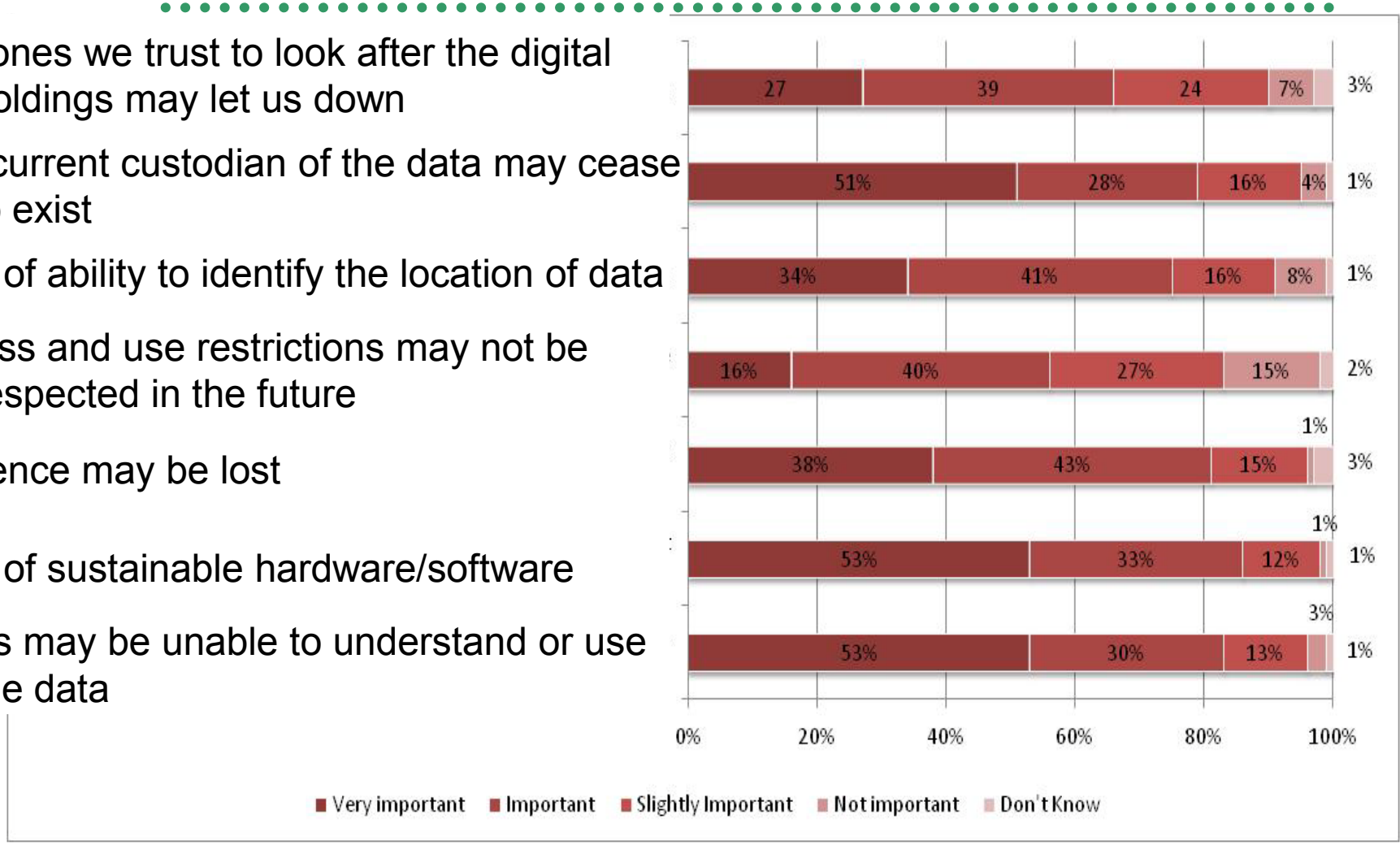
Loss of ability to identify the location of data

Access and use restrictions may not be respected in the future

Evidence may be lost

Lack of sustainable hardware/software

Users may be unable to understand or use the data



Koninklijke Bibliotheek

Threats to preservation (P)

The ones we trust to look after the digital holdings may let us down

The current custodian of the data may cease to exist

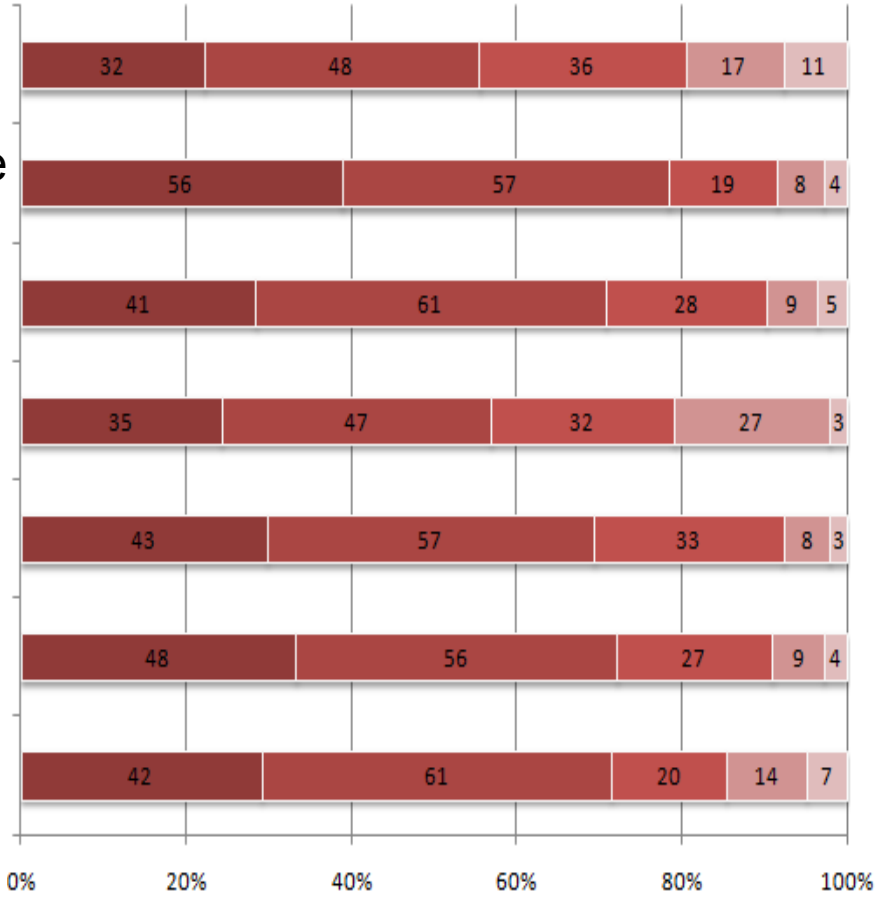
Loss of ability to identify the location of data

Access and use restrictions may not be respected in the future

Evidence may be lost

Lack of sustainable hardware/software

Users may be unable to understand or use the data



Koninklijke Bibliotheek

PARSE insight

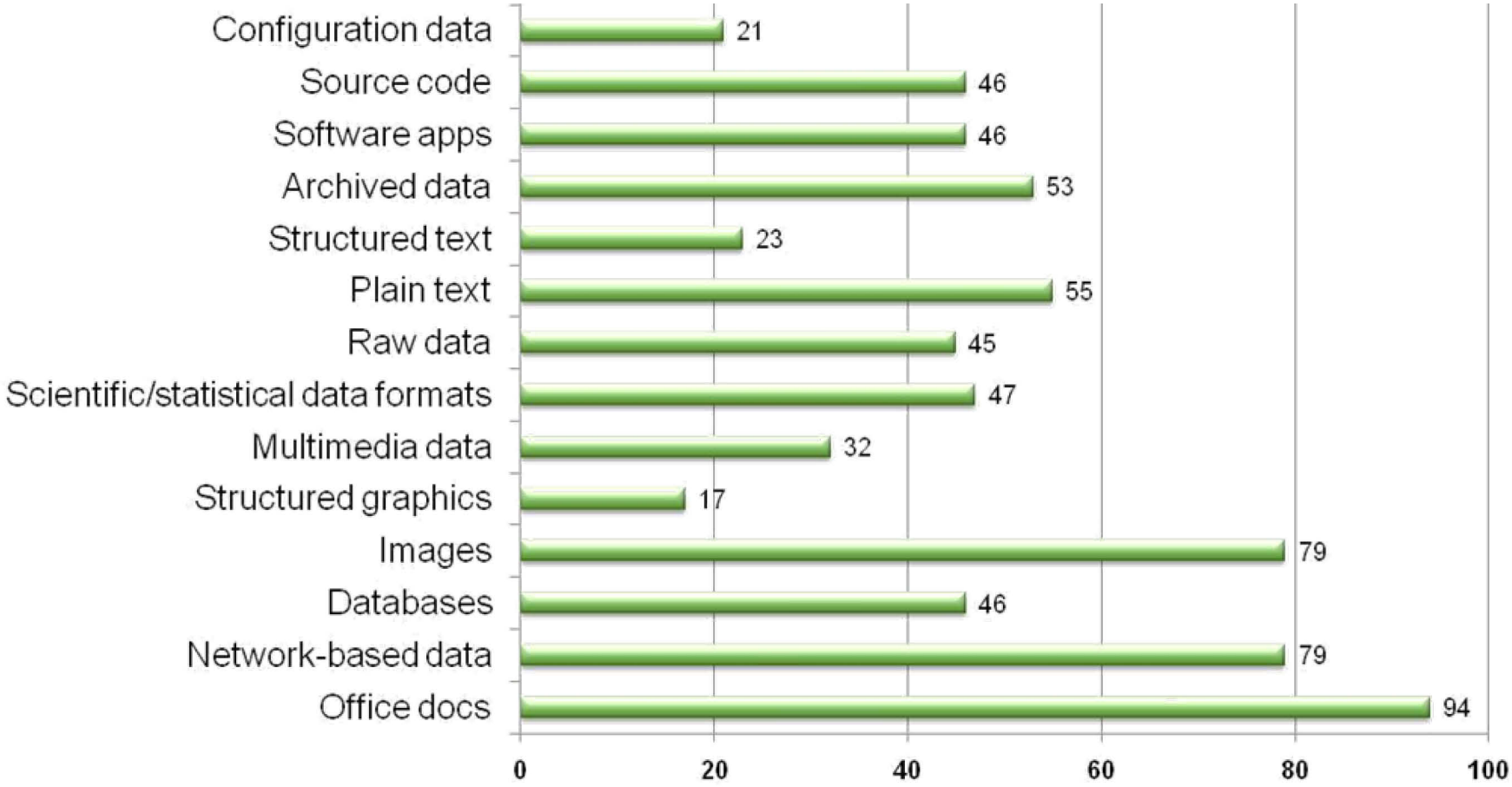
Preservation in research
–
current state of affairs

KB
Koninklijke Bibliotheek

e-infrastructure

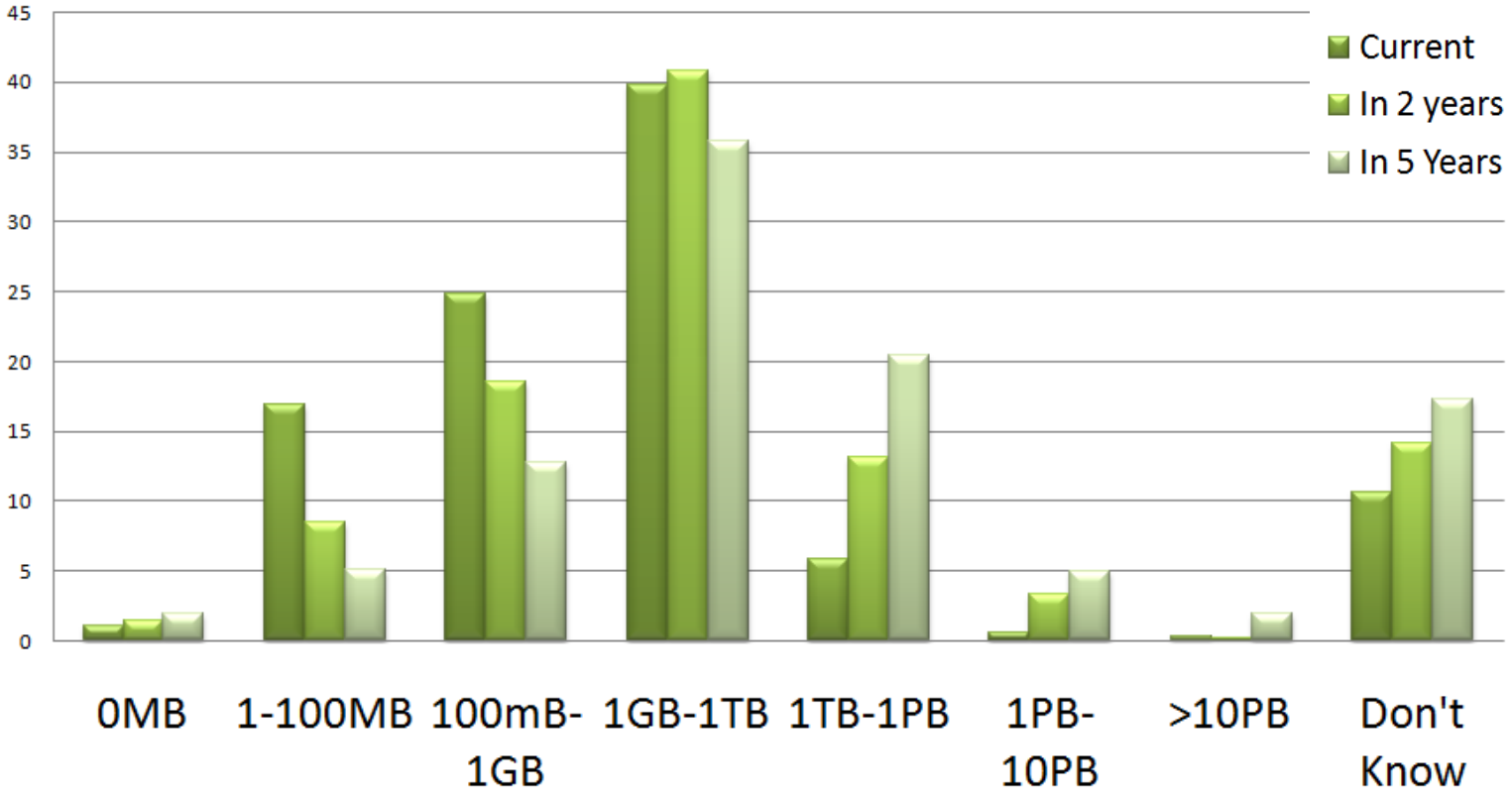
SEVENTH FRAMEWORK PROGRAMME

Data spectrum (R)



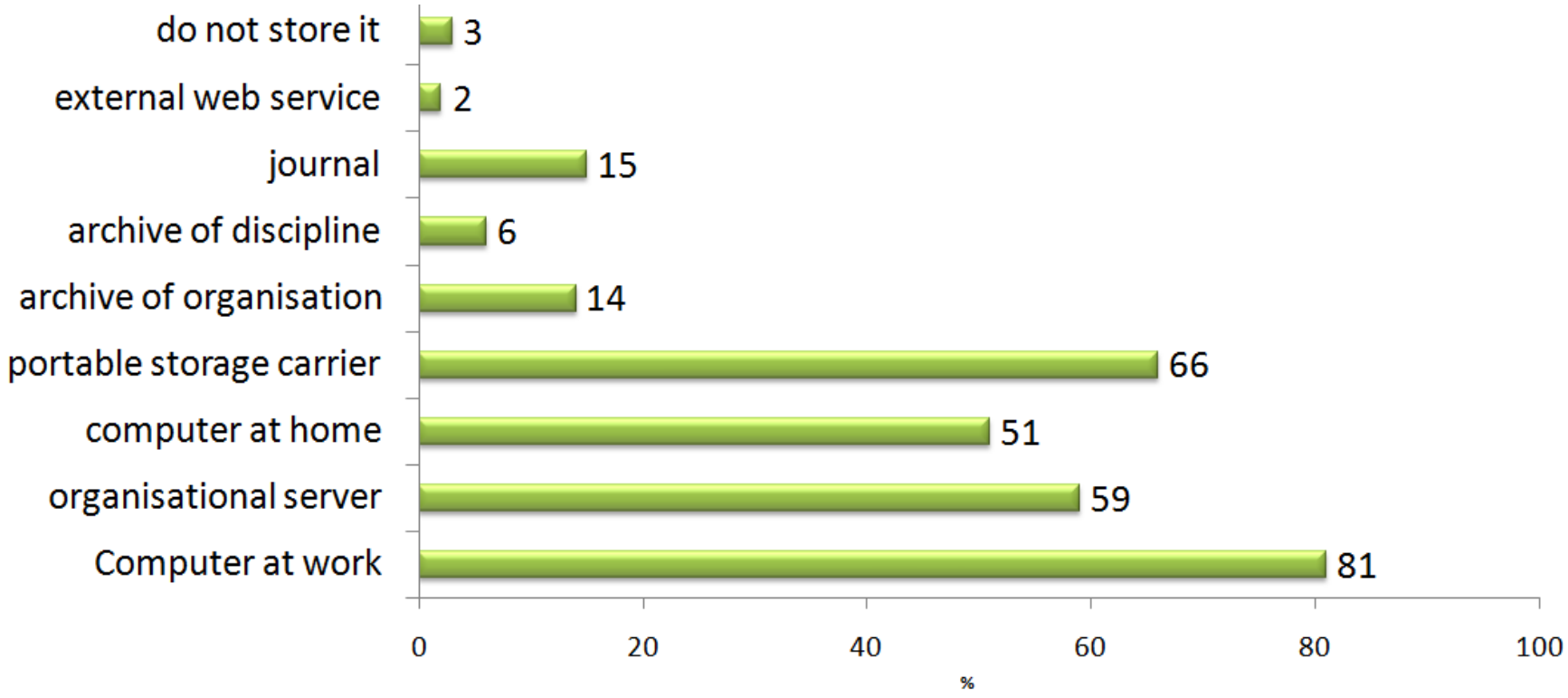
Data storage (R)

How much digital data do you have? Now and future



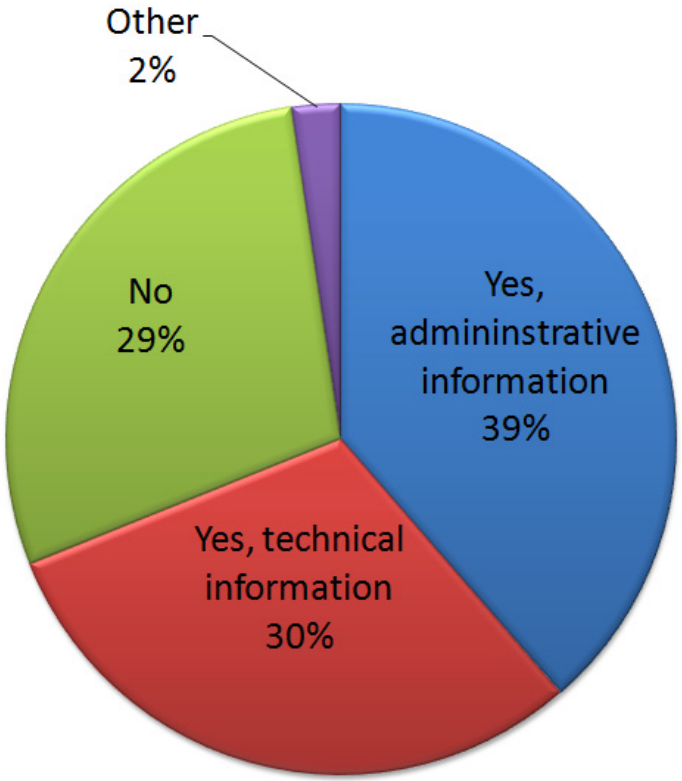
Data storage (R)

Where do you keep your digital data?



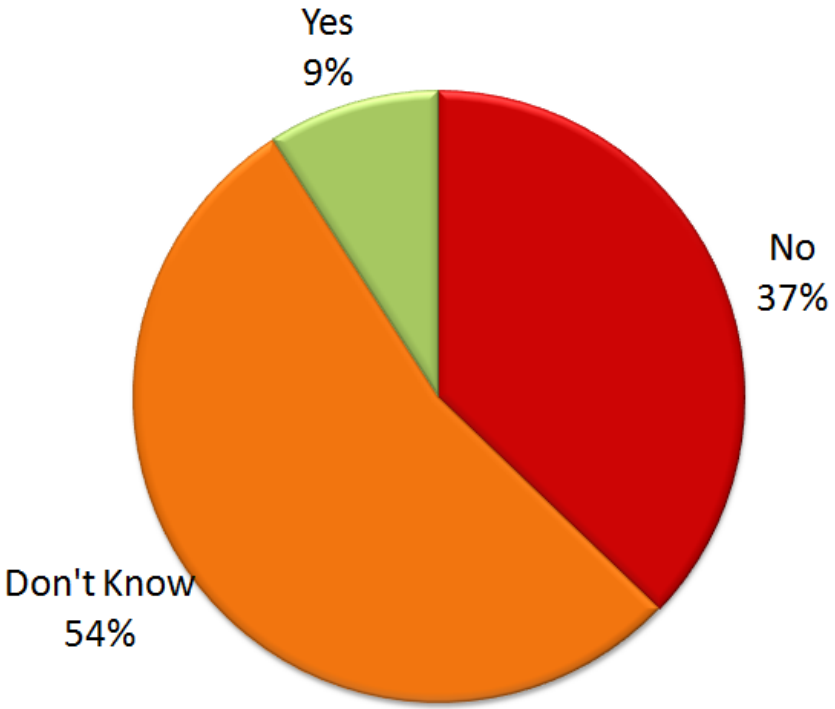
Data storage (R)

Do you as researcher assign any additional information to your digital research data?



Data storage (R)

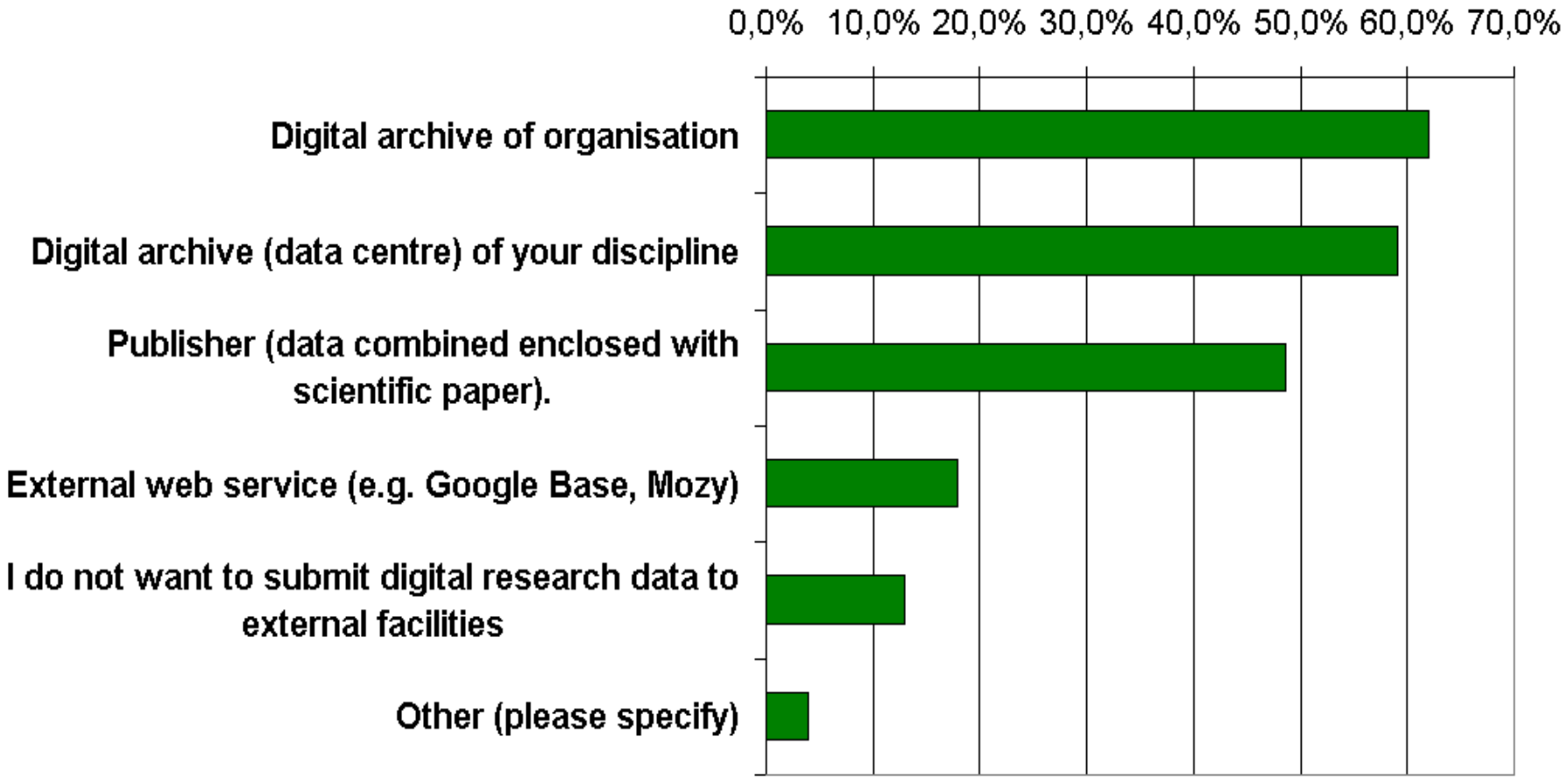
Is there a preservation facility for preserving digital research data which can be used by all projects within your discipline?



Koninklijke Bibliotheek

Data preservation (R)

Where would you like to store your data?

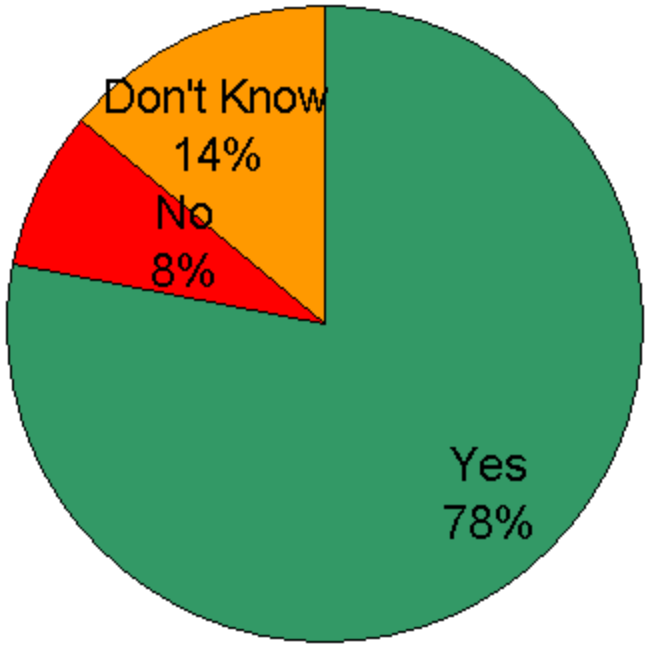


Koninklijke Bibliotheek

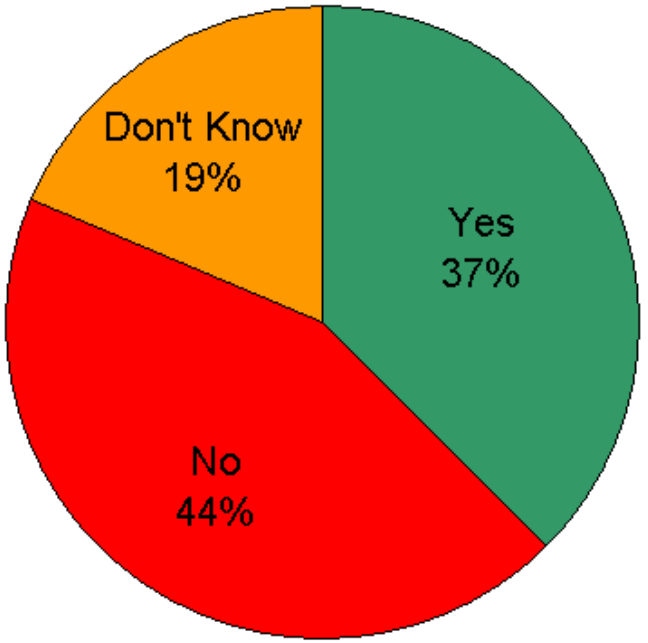
Data preservation (P)

Does your organisation have a disaster recovery policy for its digital content?

STM

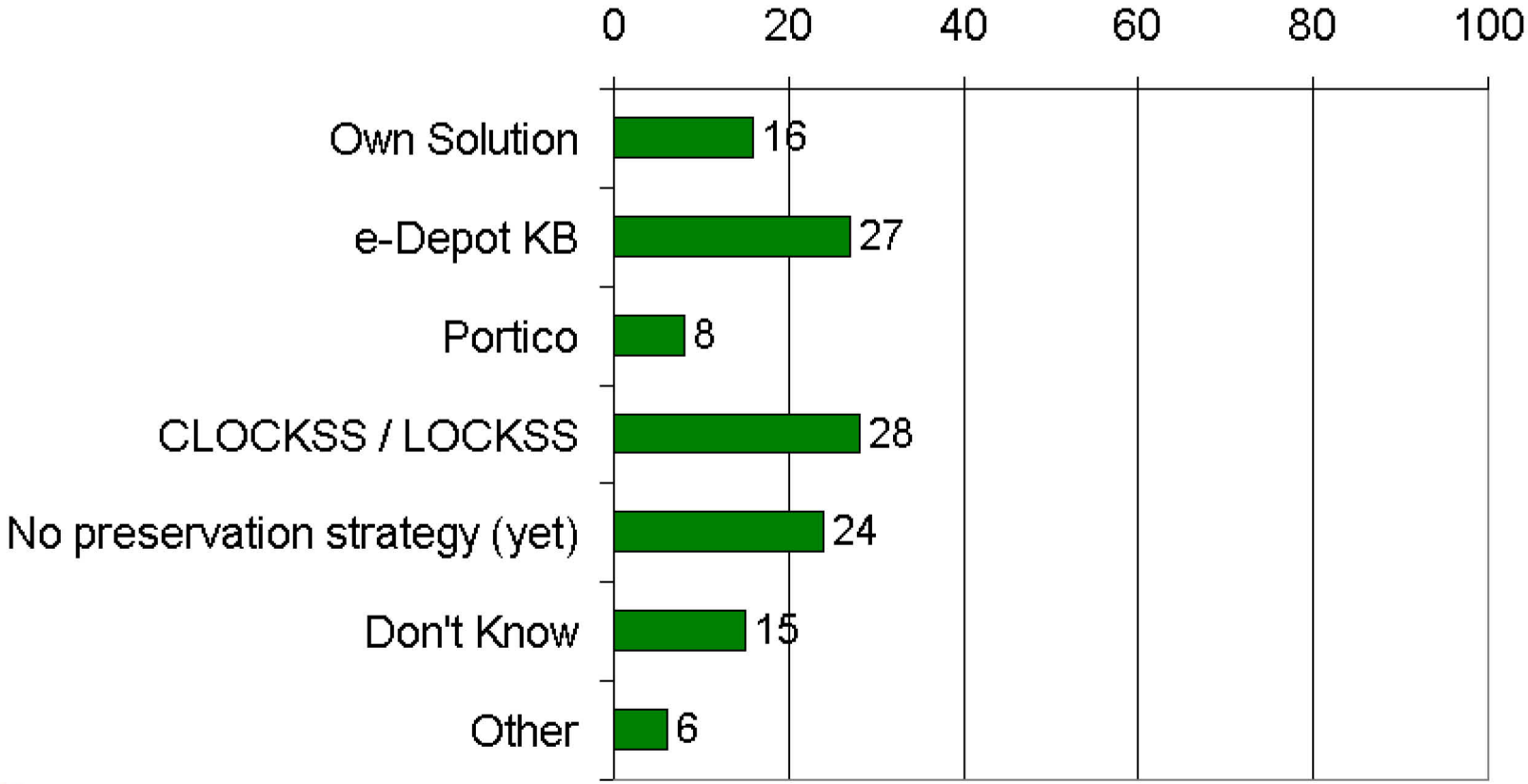


DOAJ



Data preservation (P)

Have you systematically organised the preservation of your digital publications?



Regarding data loss…(R)

We asked researchers about experience of significant loss of digital data in their work field.

"Bad tapes, dead drives"

"It happened often due to crashes of HDD"

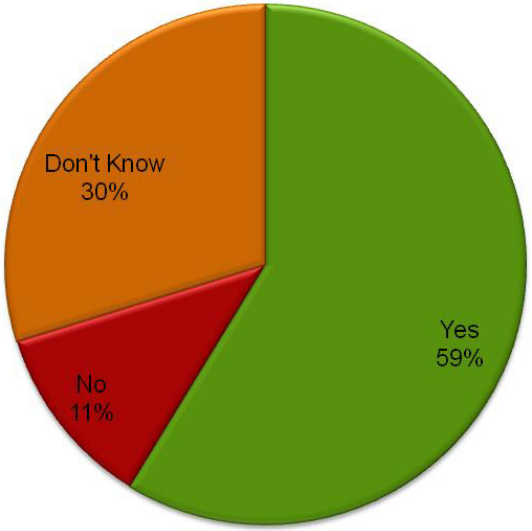
"Stolen desktop machine, not properly backed up"

"Data collected in 1970s-1980s was on paper tape (no longer used). Data collected in 1990s was on personal computers (HP) junked with no retrievable backup. I am currently analyzing old calibration data with a foreign colleague and we need to recreate the data from printouts.."

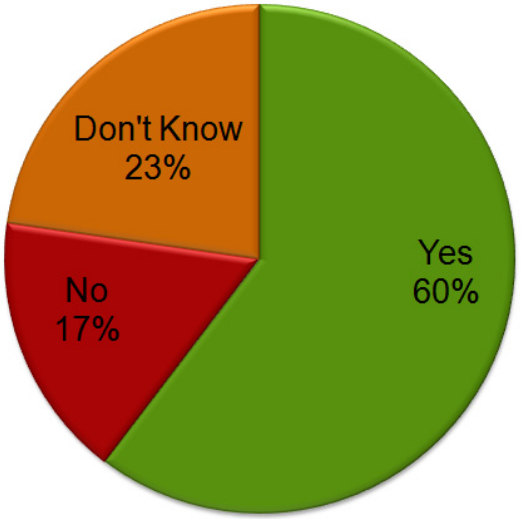
Preservation in research
–
the outlook

Do we need an infrastructure?

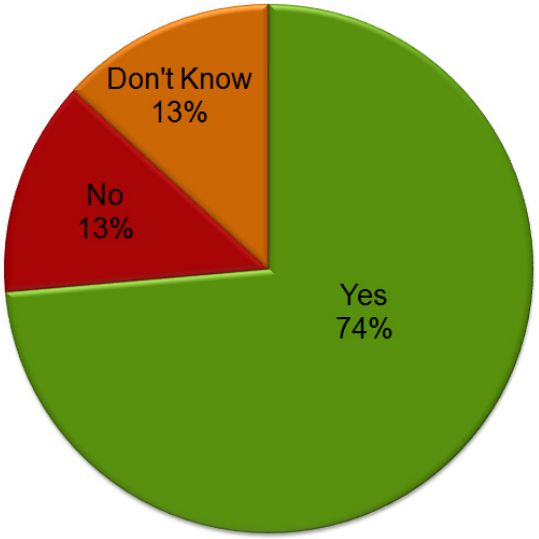
Research



Data Management



Publishing



Koninklijke Bibliotheek

What is an infrastructure?

Koninklijke Bibliotheek

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME

Ideas for an infrastructure..

"Distributed but common used storage and preservation system. Services for migration. More research in emulation".

Ideas for an infrastructure..

"National centres should be founded. In close international cooperation they should provide common principles, strategy, means and tools for converting and preserving the data in long-term period."

Ideas for an infrastructure..

"Swiss banks accepted by all European Union institutions”

Ideas for an infrastructure..

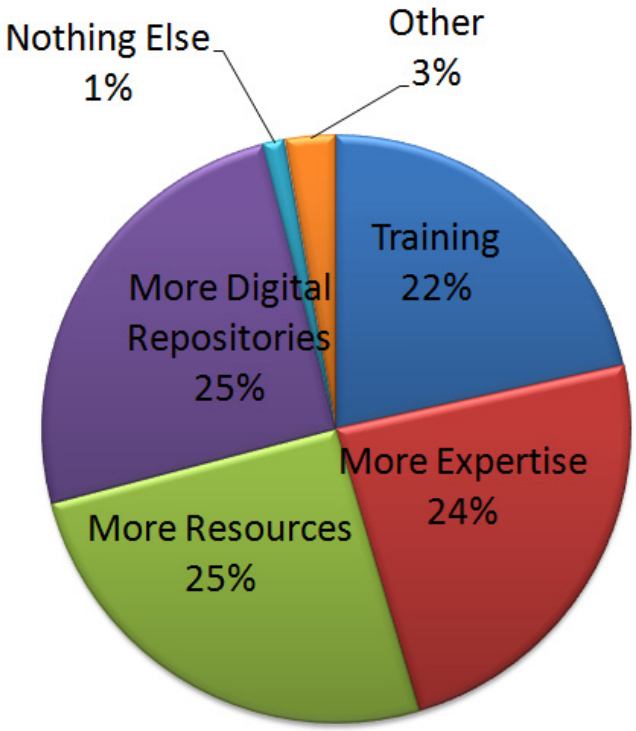
"international standards"

Ideas for an infrastructure..

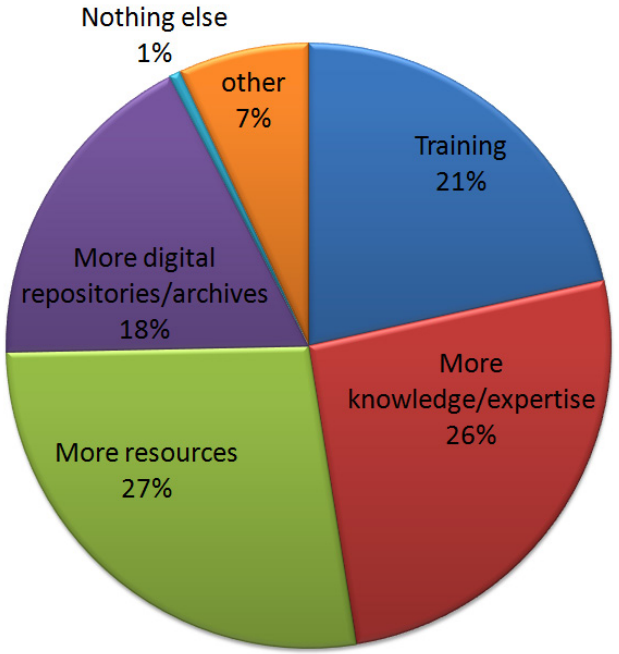
"no idea at the moment"

Other needs

Research

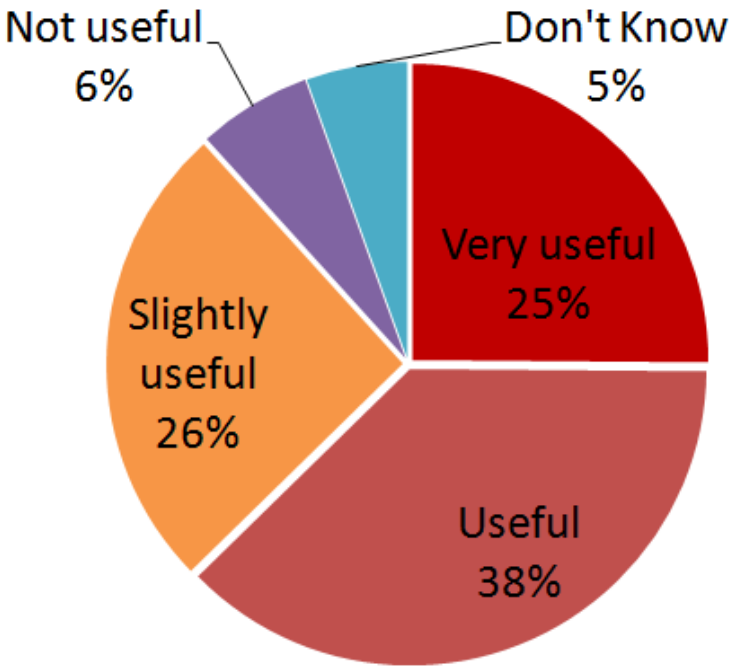


Data Management



Other needs (R)

Do we need an international knowledge platform/forum on digital preservation?



The publisher's business model

The publisher will become a provider of information management services to free...

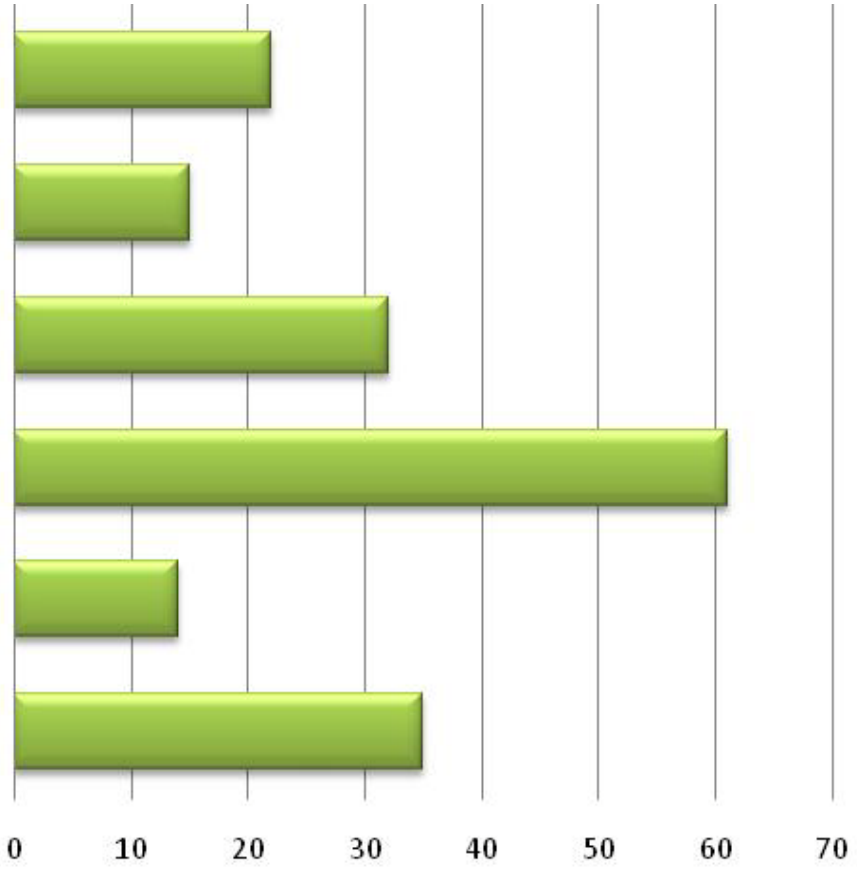
Publishers will become news aggregators, selecting and combining latest research...

Most research results will be Open Access and available for free via institutional...

A hybrid model, combining subscription-based journals and open access journals,...

Open access journals will become mainstream via the author-pays model.

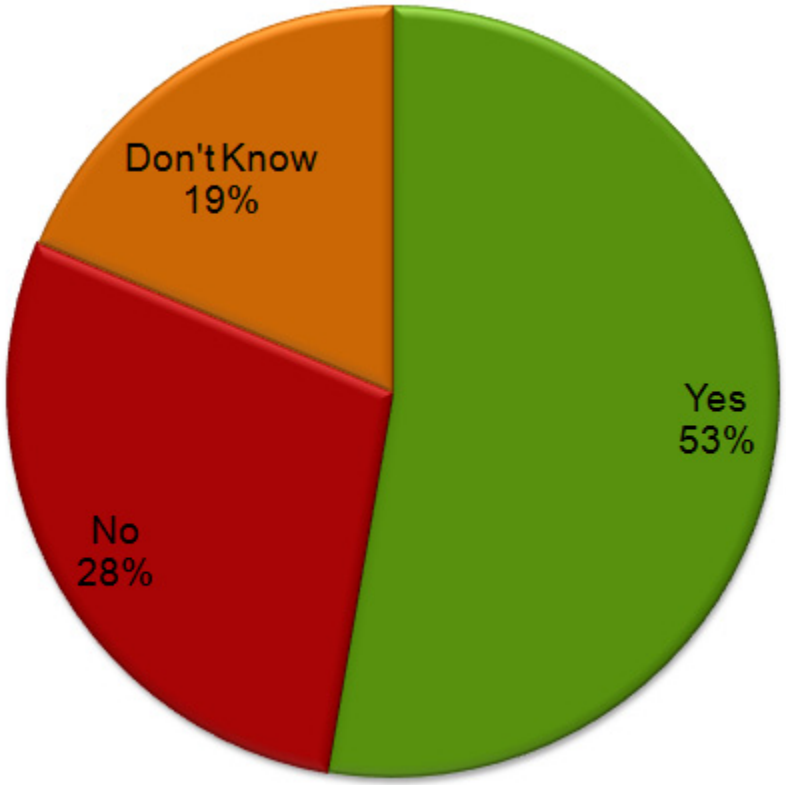
The publication process will not change much.



Cross-disciplinary use of research data

Sharing of data (R)

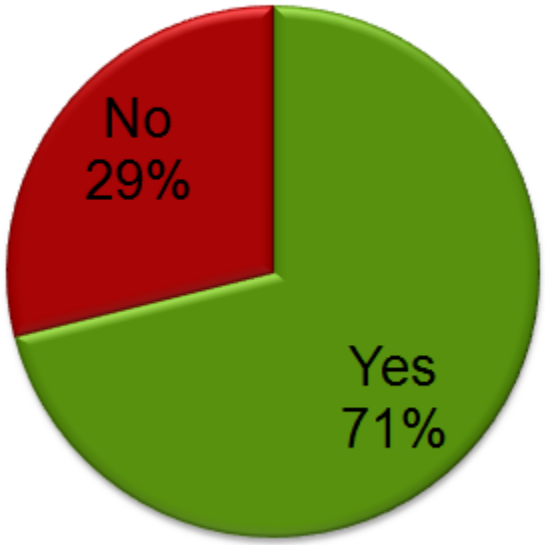
Did you ever need digital research data gathered by other researchers that was not available?



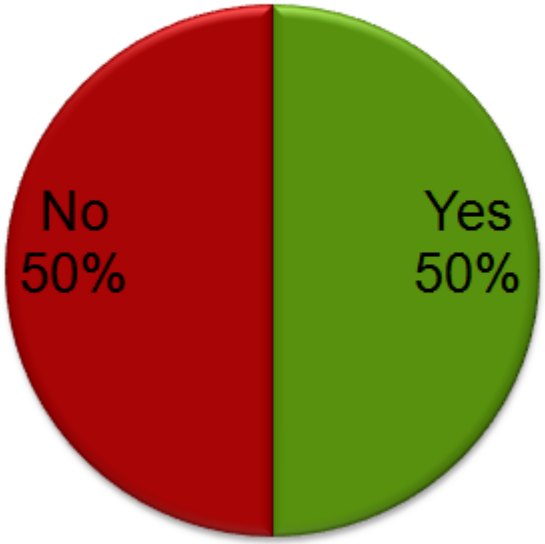
Sharing of data (R)

Do you presently make use of research data gathered by other researchers?

Within discipline



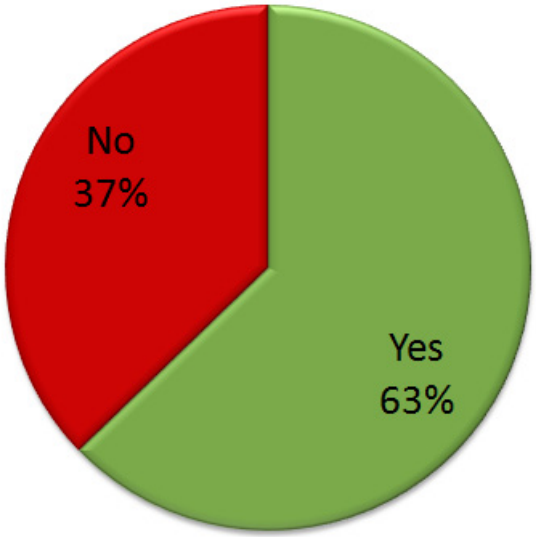
Outside discipline



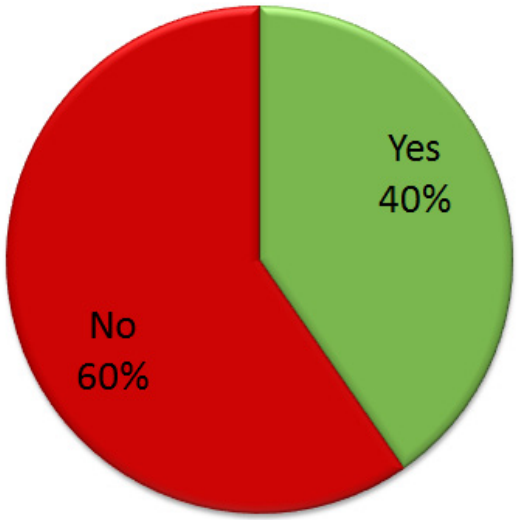
Sharing of data (R)

Would you like to make use of research data gathered by other researchers?

Within discipline

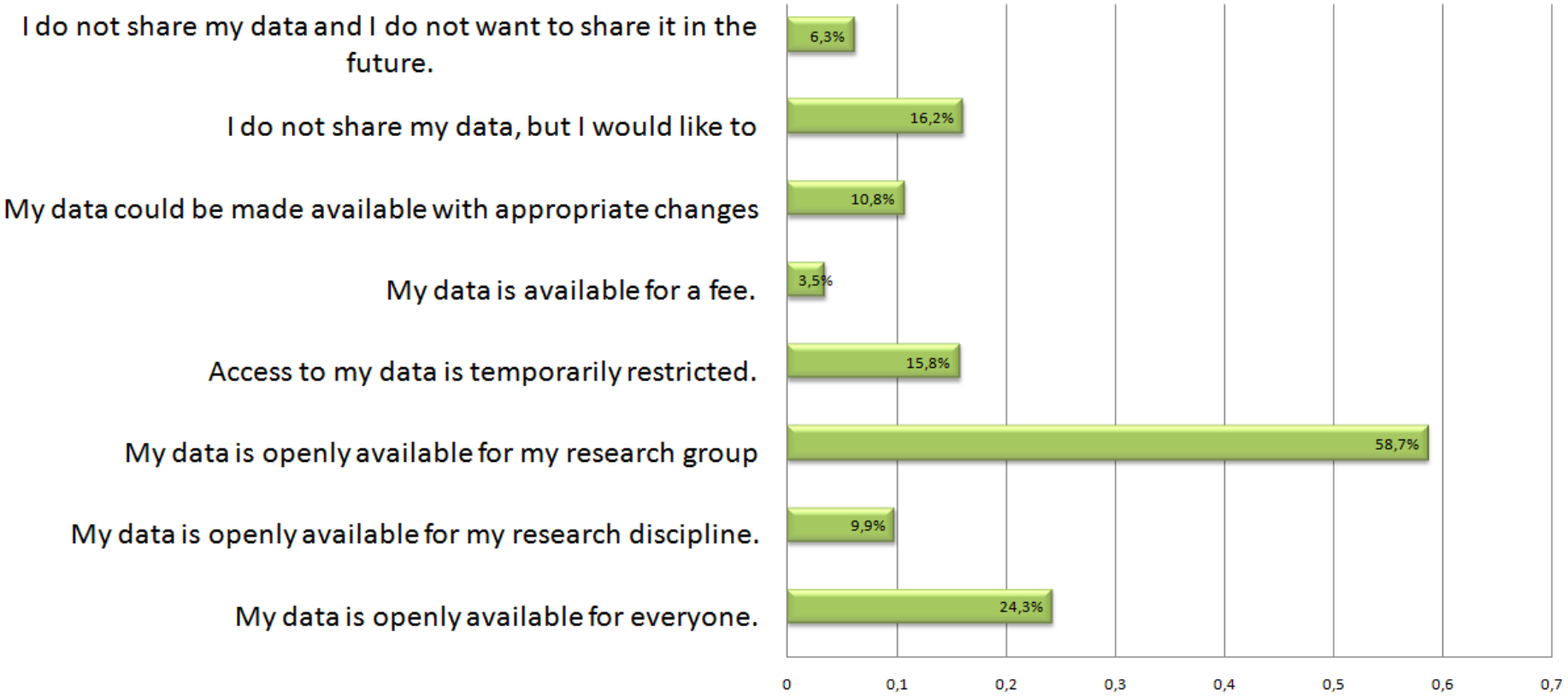


Outside discipline



Sharing of data (R)

How open is your data?



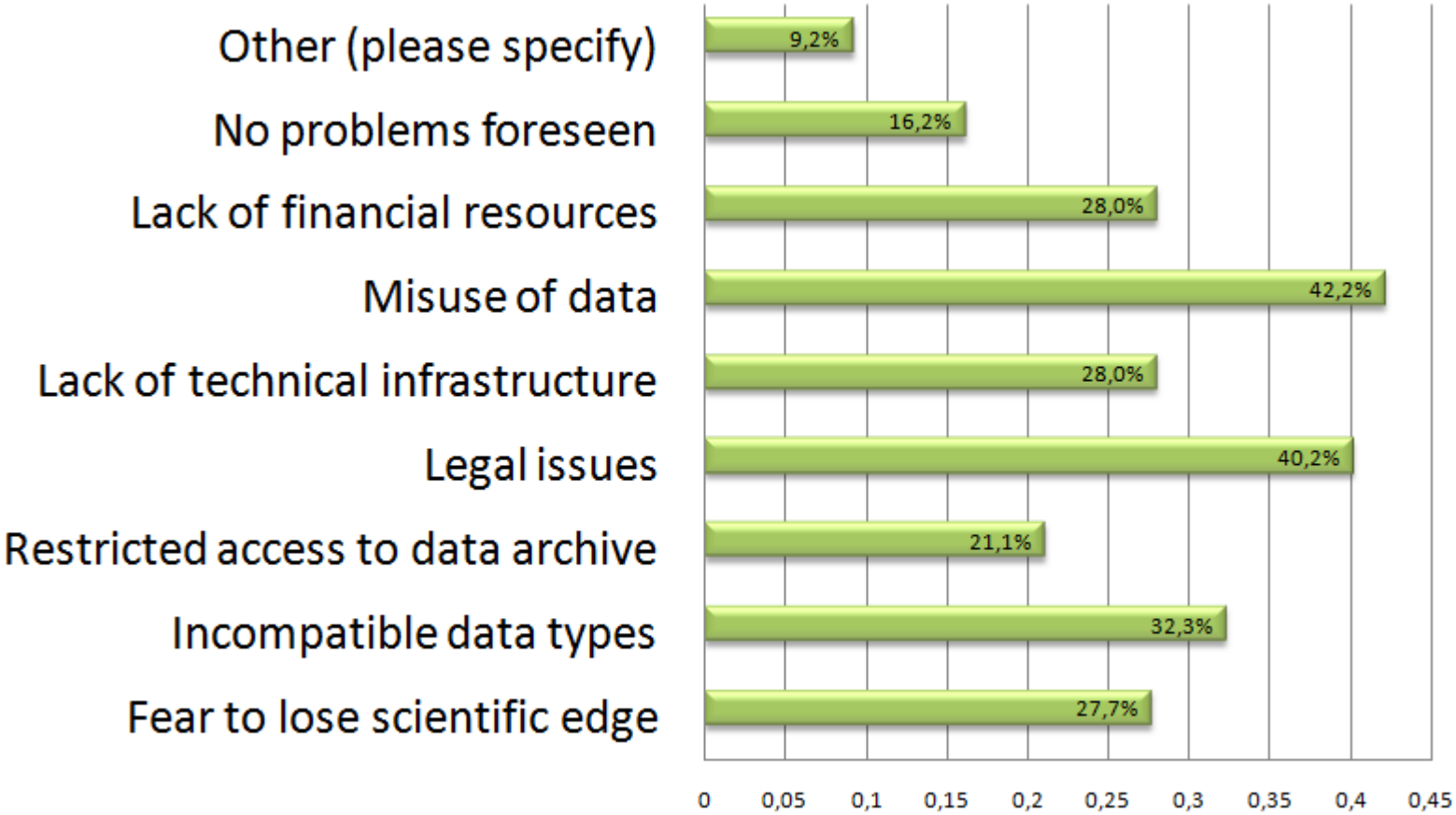
Koninklijke Bibliotheek

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME

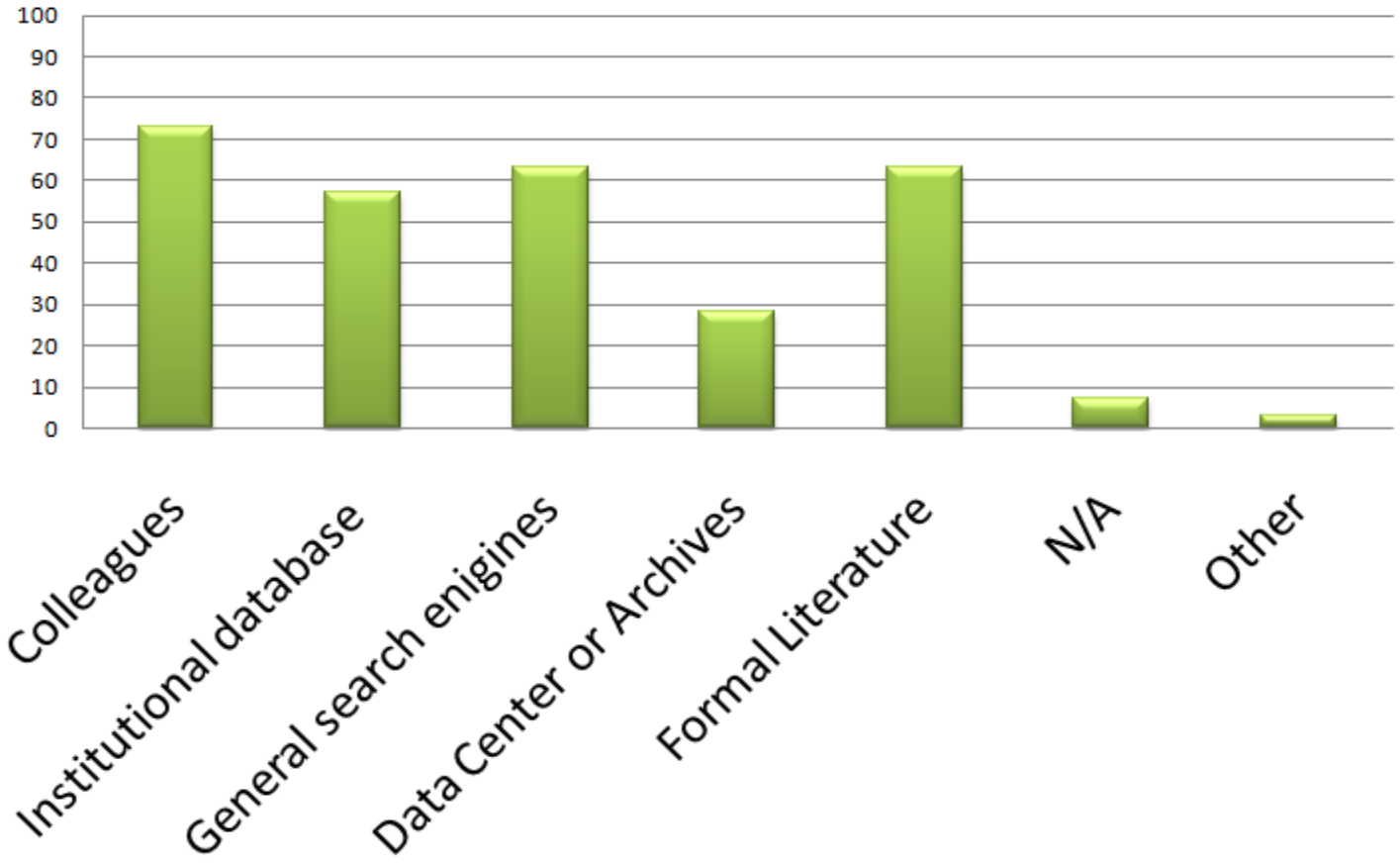
Sharing of data (R)

Which constrains do you see in making data open?



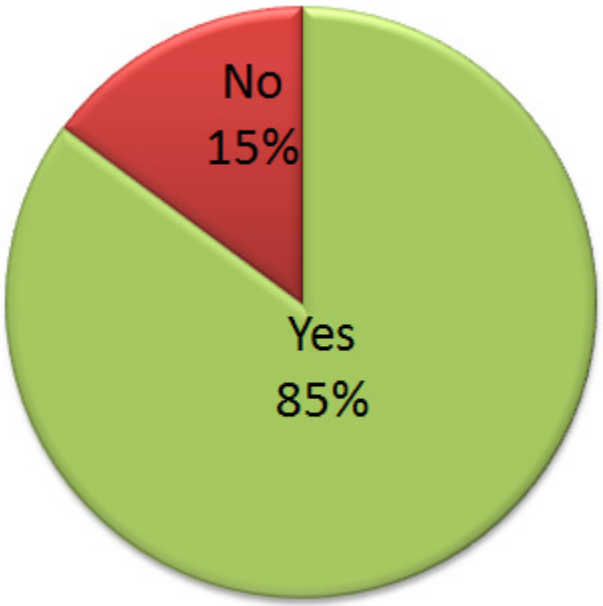
Sharing of data (R)

How do you locate and access digital research data?



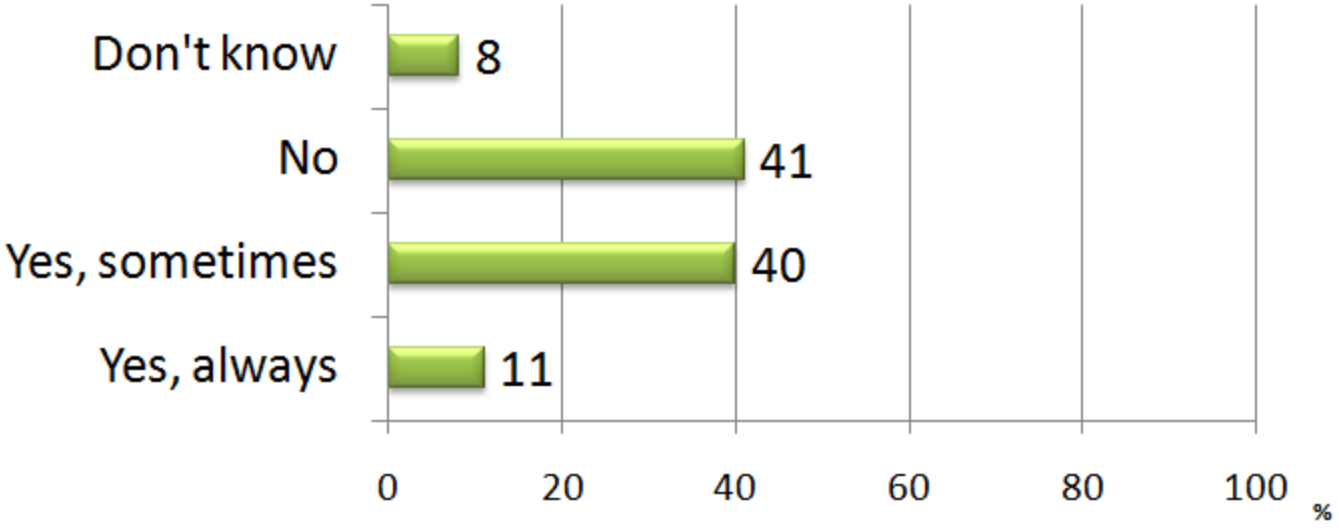
Linking of data (R)

As researcher, do you think it is useful to link underlying research to formal literature?



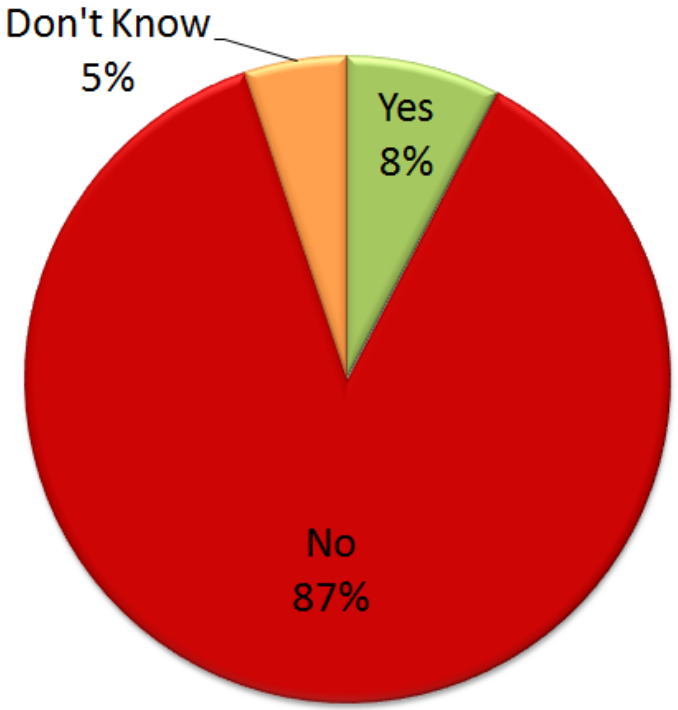
Linking of data (P)

Do you link references in your journals to underlying digital research data?



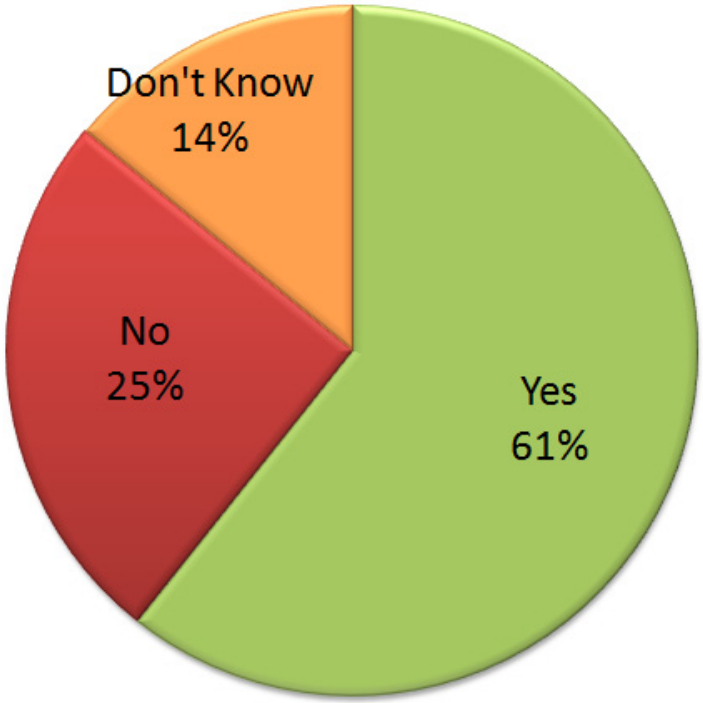
Linking of data (P)

Do you as publisher charge separate fees when users want to access data associated with publications?



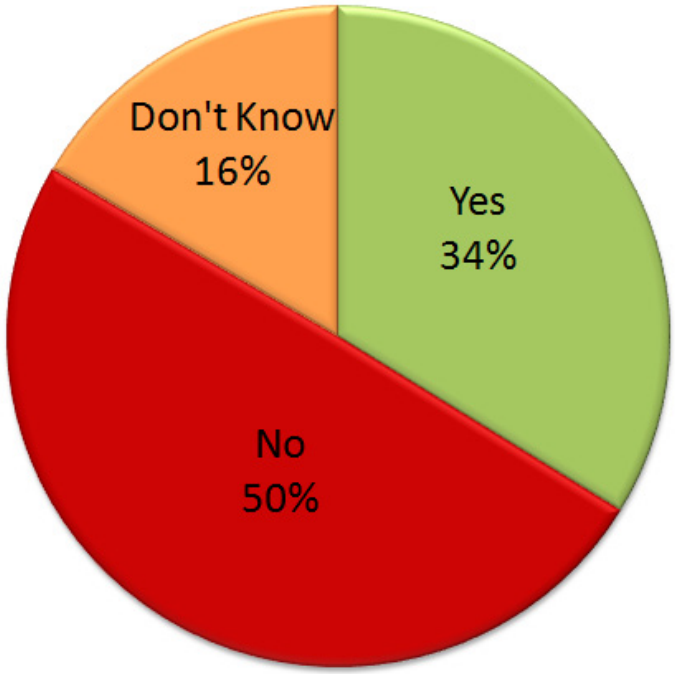
Linking of data (P)

Can authors submit their underlying digital research data with their publication to the publisher?



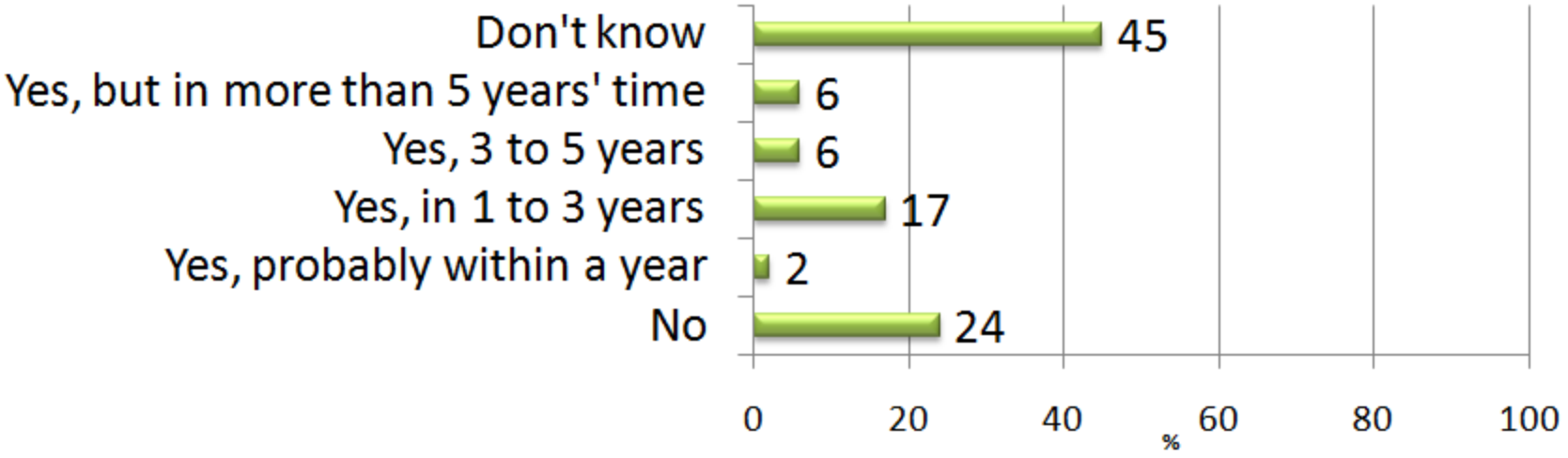
Linking of data (P)

Have procedures been established for the peer review of the content and formats of the digital research data?



Linking of data (P)

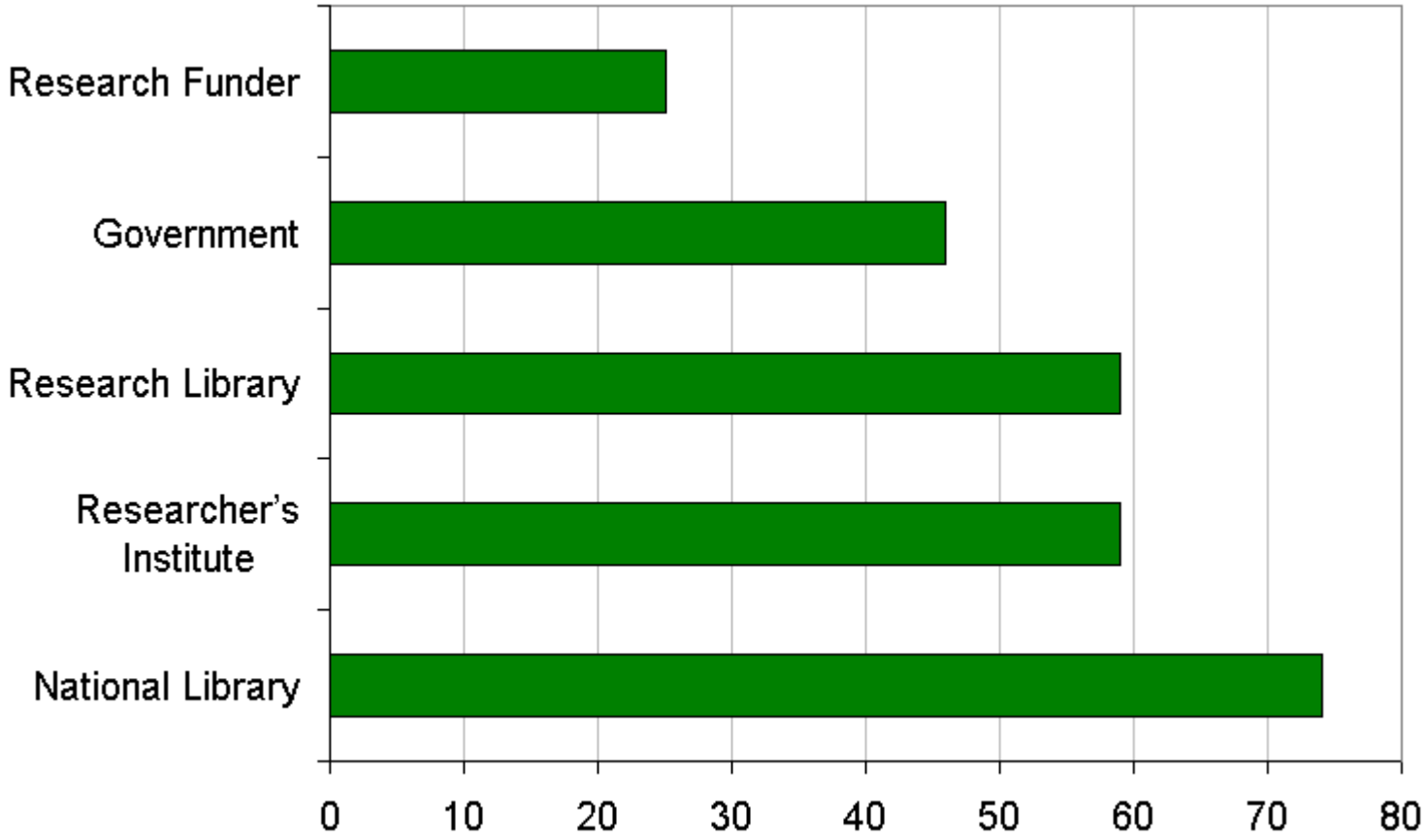
Are you planning to accept digital research data in the near future?



Roles & Responsibilities

Who is responsible? (DM)

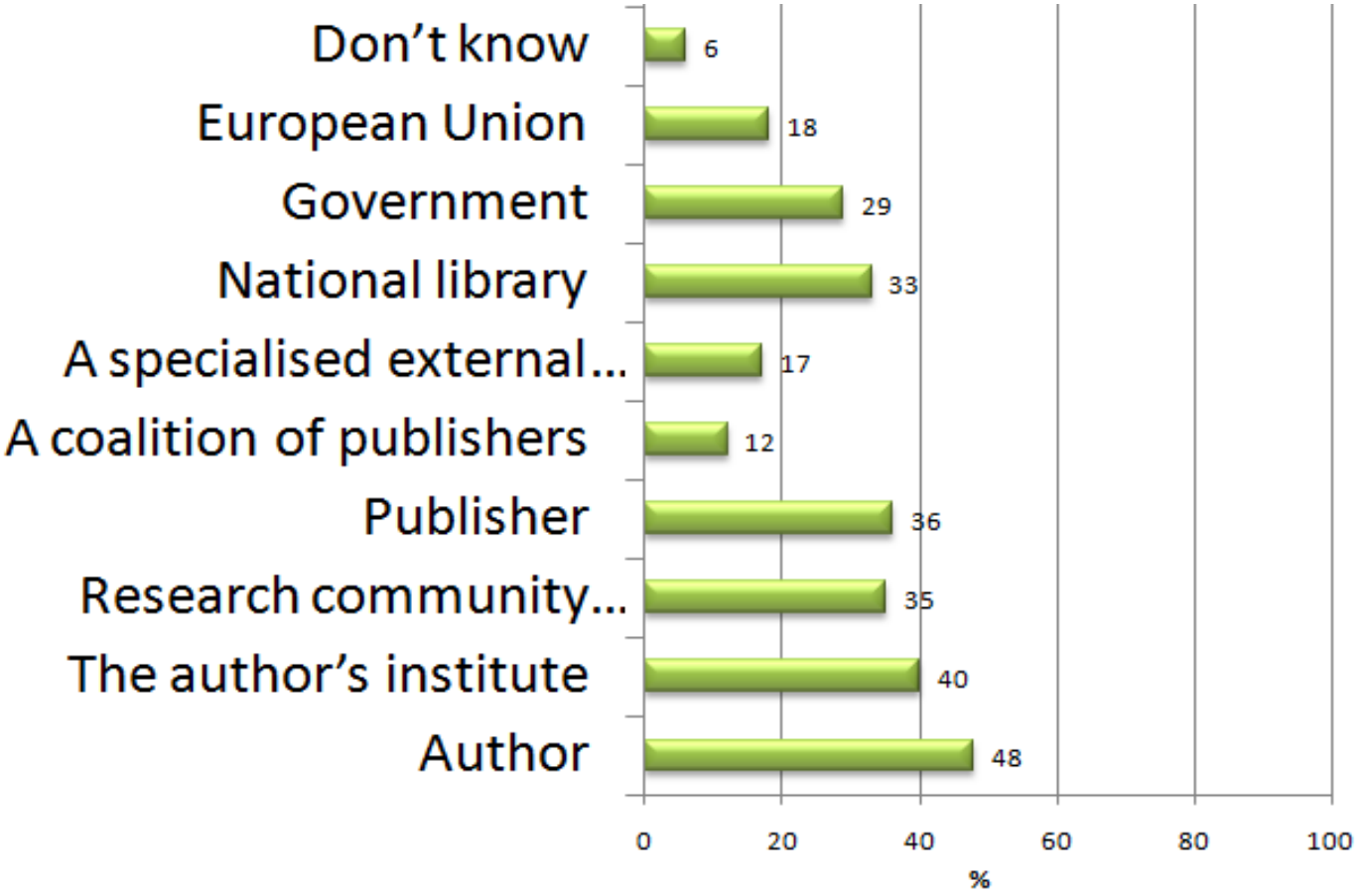
For preservation of other research output



Koninklijke Bibliotheek

Who is responsible? (P)

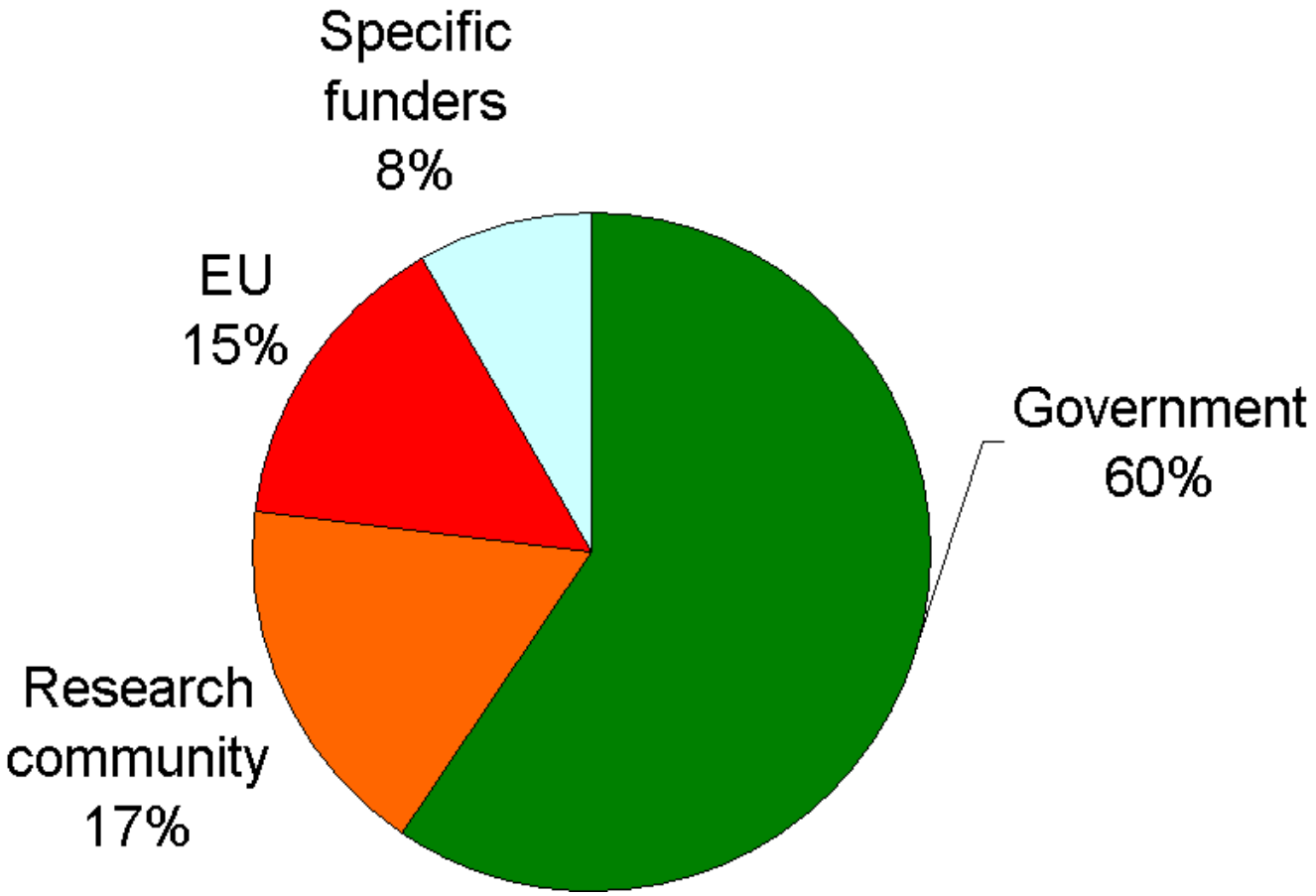
For preservation of other research output



Koninklijke Bibliotheek

About funding (DM)

Who pays for the preservation of data?



About funding (DM)

- 75% to 85% of data management believe funding to be an issue; now, but also in 5 or 10 years.

About funding

Who should pay for data preservation?

Researchers say : Government (national funding)

Data managers say : Government (national funding)

Publishers say : Government (national funding)

Who should pay for preservation of publications?

Researchers say : Government (national funding)

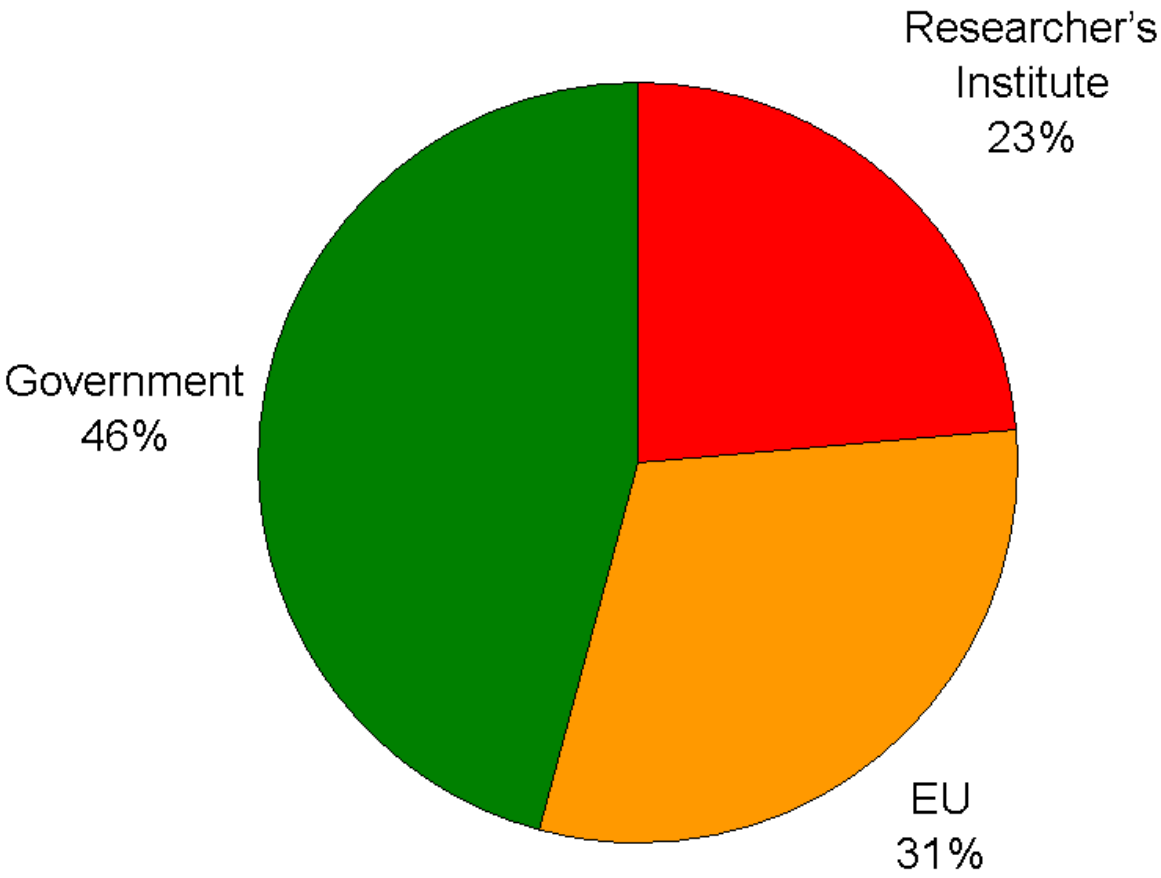
Data managers say : Government (national funding)

Publishers say : Government (national funding)

Who should pay? (DM)

For preservation of other research output

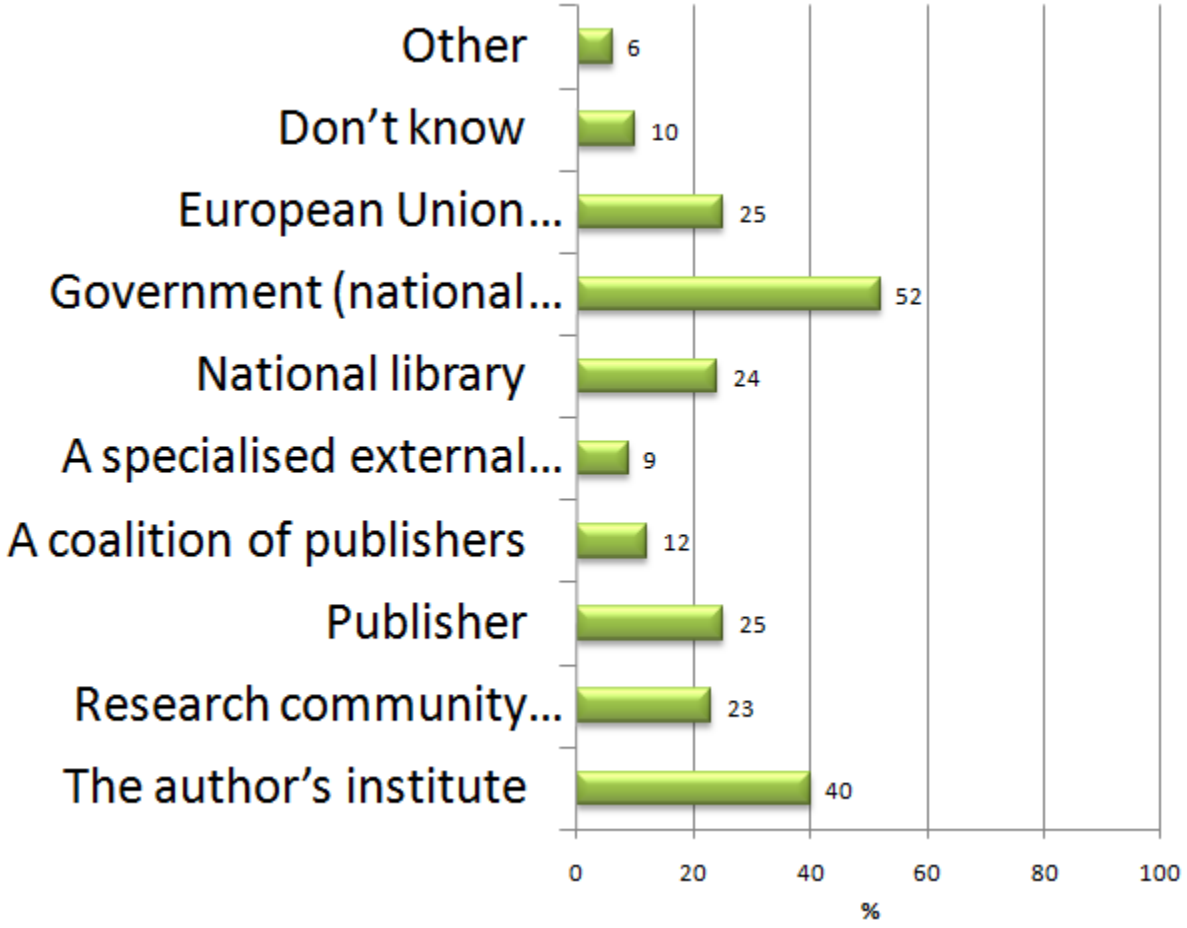
Top 3



Koninklijke Bibliotheek

Who should pay? (P)

For preservation of other research output



Koninklijke Bibliotheek

Implications for the roadmap

Need for an e-Science infrastructure in Europe

But we need to identify:

New roles and responsibilities

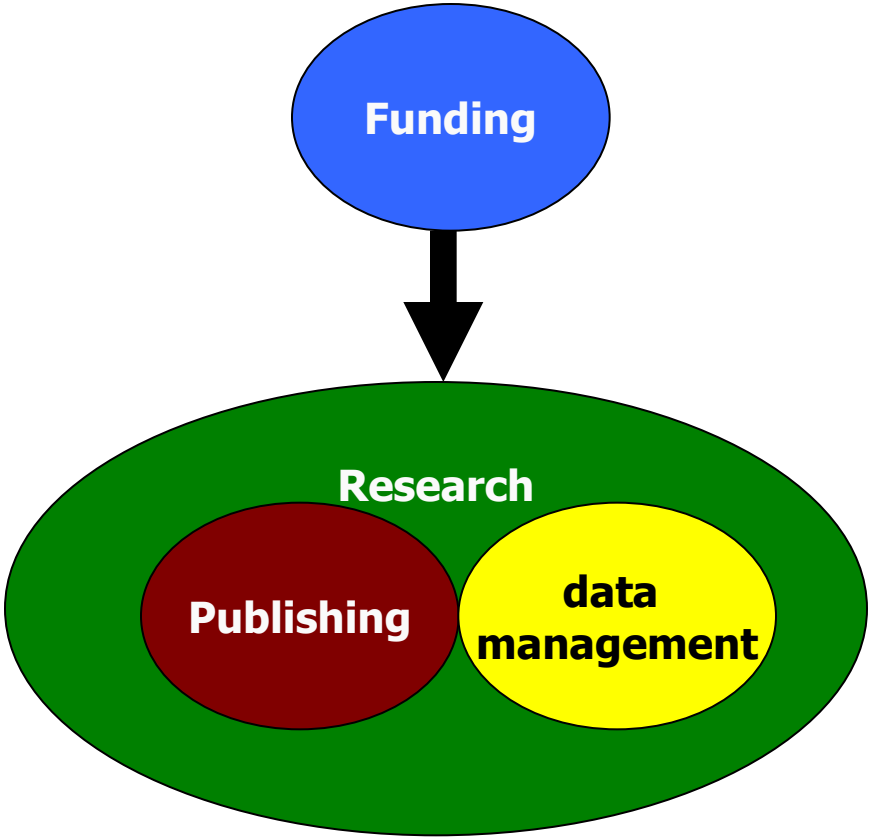
New business models

Regarding roles…

- Focus is on (open) access
- Emerging institutional repositories

result in:

- 'Chaos' of repositories*
- A lot of redundant work
- Less time & money for core research activities
- Standardisation?
- Sustainability?



*** Patricia Manson (EC), Conference "Alliance for Permanent Access", Budapest, Nov 2008**

Implications

- Are there reasons missing?
- Are there other technical/non-technical threats?

Regarding distrust in external organisations dealing with preservation:

- Is this distrust justified?
- How can it be countered?

Regarding the gap in desire and practice of sharing data:

- How can openness of data be stimulated?
- What would the role of funders be?

Regarding publishing:

- Do you agree with the opinion that other parties are better suited for taking care of digital preservation?
- Do you believe that the future business model of publishers will become a hybrid of open access / subscription-based?
- That underlying research data should be linked with the publication?

Thank you!

Contact:

Jeffrey.vanderhoeven@kb.nl

Preserve your publications at the KB?

Go to www.kb.nl

Koninklijke Bibliotheek

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME