

*Learned Publishing*, 24:35–49  
doi:10.1087/20110107

## Introduction: preservation as a permanent problem

Ever since the founding of the ancient library of Alexandria, estimated to have happened around 300 B.C., it has been a repeated ambition throughout history to ensure longevity if not perpetuity of publications, to keep at least one copy safe somewhere, and to establish a well-secured and ever-cumulating corpus of knowledge that can withstand the ravages of time. One of the stories of Alexandria<sup>1</sup> holds that it started off with the private collection of Aristotle and another that by decree of the Egyptian ruler Ptolemy III every visitor was required to surrender all books and scrolls in their possession. Scribes would copy these works swiftly and keep the original in the library while the original proprietor received the copied version – in that way a vast collection was built up quickly. As many as 400,000–700,000 books and parchment rolls are estimated to have been in the library during its peak years. And whether it was accidental fire or a vengeful Caesar or brute barbarians that brought the library to its end a few centuries later (the historical variations on cause and course of events are plenty), it has ever since symbolized the importance of preservation of publications.

The publishing world is well aware of the dangers of decay. In the print world, there are the perils of embrittling and yellowing paper, and destructive chemical reactions caused by the wrong ink, the wrong glue, the wrong paper, or the wrong treatment. And just as in Alexandrian times when visiting travellers were supposedly expected to bring their books to the library, publishers have always been very collaborative in depositing their works in national archives and national libraries, which usually require one copy of every work published, a legal regulation that has been in existence for several centuries in

# Avoiding a Digital Dark Age for data: why publishers should care about digital preservation

Eefke SMIT,<sup>a</sup> Jeffrey VAN DER HOEVEN,<sup>b</sup> and David GIARETTA<sup>c</sup>

<sup>a</sup>International Association of STM Publishers,

<sup>b</sup>Koninklijke Bibliotheek (KB), The Netherlands, and <sup>c</sup>European Alliance for Permanent Access

**ABSTRACT:** *This paper provides an overview of the needs and threats for digital preservation and summarizes the findings from project PARSE.Insight. This project, co-funded by the EU, contains one of the first large worldwide surveys about digital preservation including most players of the STM information chain: researchers, libraries, data managers, publishers, and research funders. One of the conclusions is that in the present data deluge, it is extremely important that all players in the information chain work together on proper digital preservation of all research output, to ensure its future usability, understandability, and authenticity. This poses a new role for publishers who can ensure better discoverability and citability via good linking and integration of data and publications.*

© Eefke Smit, Jeffrey van der Hoeven, and David Giaretta 2011



Eefke Smit



Jeffrey van der Hoeven



David Giaretta

*anything digital  
is fragile and  
susceptible to  
decay*

most of Europe and around the world. That was the world of paper, parchment, and papyrus.

Now we have entered the digital era. Digital media have become the dominant way in which we create, shape, and exchange information. Government, businesses, research organizations, libraries, and archiving institutions, as well as individuals, have become completely dependent on digital information. The rate at which the research world has become digital – in fact much earlier and faster than many other sectors – poses a threat to the longevity and retrievability of scientific information. Science depends entirely on the knowledge gained in the past to further advance.

Strangely enough, some people seem to think that in a digital world, these perils of decay are less prominent or complicated because ‘it is now so much easier to store it all somewhere digitally’ and because of the misperception that ‘as long as it is digital, it is safe’. But anyone who has tried to open an old WordPerfect file, or play a video-recording made a few years ago, no matter how well and safely stored, will have found that this is not so easy. Moreover, often the transition or migration to newer formats is to no avail either, if something as nasty as so-called bit-rot has taken hold of the digital object. Digital obsolescence is a nasty trap in today’s life, making a lot of material no longer readable, whether it is because of outdated file formats, unavailability of the right readers or software, or simply (and often inexplicably) due to bit-rot. Or more subtly, one might be able to see or print a table of numbers, e.g. of scientific data, but have no idea what they mean or how to use them.

In short, anything digital is fragile and susceptible to decay. As the amounts of digital information and data are exploding in volume, digital preservation is becoming an ever-pressing issue. The prospect of losing the digital records of science is therefore highly alarming: a Digital Dark Age may be looming.

The term Digital Dark Age<sup>2</sup> was coined by Terry Kuny more than a decade ago at an IFLA conference and applies to a possible future situation where it will be difficult or

impossible to read historical documents, because of inadequate digital preservation. This could cause the period around the turn of the 21st century, when viewed from the future, to be comparable to the Dark Ages in the sense that there will be a relative lack of written record. A famous real example<sup>3</sup> is with NASA whose early space records suffered in this way: for over a decade, magnetic tapes from the 1976 Viking landings on Mars were unprocessed. When later analysed, the data were unreadable as they were in an unknown format and the original programmers had either died or left the organization. The images were eventually extracted through many months of puzzling through the data and examining how the recording machines functioned.

The key question is: are we doing any better now? This paper aims to present an overview of the present status of digital preservation in the scholarly arena, drawing from the results of EU-project PARSE.Insight<sup>4</sup> which has carried out the largest digital preservation survey to date. Overall, it showed that among publishers there are well-established preservation strategies and practices for journal articles. Well over 90% of the journals covered in the PARSE surveys have a digital preservation policy, mostly as a legacy to the print world, thereby safeguarding the future state of the research publications that serve as the official versions of record. These digital journal collections are usually the subject of close collaboration between trusted electronic depots of national archives (e.g. the Koninklijke Bibliotheek (KB), the National Library of the Netherlands) and sector-wide initiatives (e.g. Portico, CLOCKSS in the US) to which many of the preservation tasks are outsourced.

The situation is not so good for scholarly and research output other than these official ‘version of record’ publications in scholarly journals, such as (raw) research data or datasets and multimedia formats or communications via social networks. These will pose a new challenge during the years to come for digital preservation. In the light of the emerging data deluge, their volume is likely to increase even more, but equally so is

the variation and multitude of formats in which they will appear.

This paper advocates close collaboration among all stakeholders in the information chain of scientific, technical and medical publications to solve this issue in the right way, including researchers, their institutes, their policymakers, funders, libraries, data-centres, archivists, national libraries, and of course publishers.

Or, to quote the motto of the Alliance for Permanent Access:

Digital preservation is too big an issue for individual institutions or even sectors to address on their own. The required effort is simply not feasible and international exchange and collaboration are therefore essential.<sup>5</sup>

#### Digital preservation: what is it ?

Even among the well-known custodians of preservation, there is quite a bit of confusion as to what the concept of digital preservation actually entails. Often it is incorrectly mixed up with related concepts, or people focus only on smaller parts of digital preservation and mistake these for the whole creature. Some examples:

- 1 *Digitization projects.* Even if in some way related, digital preservation is *not* the same as digitization projects. Certain digitization projects, as started in recent years, have the objective to curate, preserve, and give access to large collections of print, incunabula, and manuscripts, often documents that are centuries old and need protection in their present form against ongoing wear and tear. The confusion comes from the mistaken belief that anything digital will last forever. Well it does not, or to repeat Jeff Rothenberg of ACM: 'Digital Information lasts forever, or for the next five years, whichever comes first.'<sup>6</sup> And exactly for this reason, anything digitally born and created needs preservation measures just as much as anything digitized at a later age.
2. *Archiving projects.* Archives defined as simply collections of historical documents are an important source for which (digital) preservation measures should be in place.

However, archiving in this sense is not the same thing as preservation – one could say that archiving is the first step in terms of selecting what is worthwhile preserving; but there is a lot needed next. In a 'webworld' archiving projects are faced with new challenges of so-called dynamic archiving: keeping archives of ever-changing information, as in the web-archiving now being implemented by several national libraries with the aim of providing snapshots of the Web at any given time in the past. There is a strong overlap with digital preservation, but it is certainly not the same thing. See the British Library example,<sup>7</sup> but also New Zealand, Norway, etc.

3. *Long-term storage of digital data.* According to the US-based Blue Ribbon Task Force,<sup>8</sup> most institutes currently see digital preservation as a 'storage' issue. But while proper storage may be an essential element to good digital preservation, again it should not be made synonymous to it. Long-term storage (usually via multiple back-ups) ensures that data will be retained, but is in itself no guarantee that the data can still be read and understood in the future, because systems are often replaced and formats change or tacit knowledge or appropriate software and documentation is lost.
4. *Open access.* For some, digital preservation, and even more so when the jargon tends towards a term like 'permanent access', is an important step to open access, reasoning that over time copyright will run out and good digital preservation will build a collection that everyone in due course can use. Some (mis)use the digital preservation podium to turn what they call 'dark archives' into 'open archives'. Again, this reflects a confusion of terms and definitions, mixing up the availability over time with access rights. Open access is an emerging business model, and has to do with the absence of access restrictions for users while the costs for obtaining information have shifted elsewhere, either to the author or to his/her sponsors. Permanent access, on the other hand, has to do with the availability of the material over time, independent of

*there is quite a bit of confusion as to what the concept of digital preservation actually entails*

access rights, business models, and shifted costs. It would be a pity if the good cause of digital preservation were to get mixed up with the open access discussion, simply because digital preservation is too important, independent of whatever business model is put in place to ensure users get access to the information. The only relation between the two is that open access material is equally dependent on good digital preservation as is any other digital information with or without access entitlements.<sup>9</sup>

So, if digital preservation is not the same as digitization, archiving (used in the collection sense described above), long-term storage, or open access, then what is it?

According to the definition as used in the Open Archival Information System (OAIS)<sup>10</sup> and by Giaretta,<sup>11</sup> two key elements are central in digital preservation:

- *usability and understandability* over time of a digital object;
- *authenticity* over time of a digital object.

These two criteria (being able to reuse and understand, and authenticity) add exactly what any of the above aspects were missing: digital preservation is all of the above and more.

In other words: preservation makes no sense if the preserved objects cannot be made retrievable, discoverable, accessible, and reusable, and if they cannot at the same time provide assurance that they are what they claim to be. While digital objects best exist for preservation purposes in standardized formats (as is often done in large digitization projects), and need to be selected or archived in well-documented collections so that nothing worthwhile gets lost, while they need to be stored long term in a safe way that withstands the ravages of time, and while there must be ways to access these data and information, everything in these collections must also remain usable and understandable to make preservation really meet its goals. For authenticity, one needs to make sure that evidence is collected about what has befallen the digital objects and who has been responsible for them over time.

### Seven threats to digital preservation

The preservation of digital objects can be threatened in several different ways, in addition to the obvious dramatic ones such as floods, earthquakes, and political upheaval. The report of EU-project PARSE.Insight lists seven threats in a useful and transparent way:<sup>12</sup>

1. Users may be unable to understand or use the data, e.g. because of the semantics, format, processes, or algorithms involved.
2. Non-maintainability of essential hardware, software, or support environment may make the information inaccessible.
3. The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity.
4. Access and use restrictions may not be respected in the future, jeopardizing proper reuse.
5. Loss of ability to identify the location of data.
6. The current custodian of the data, whether an organization or project, may cease to exist at some point in the future.
7. The ones we trust to look after the digital holdings may let us down.

In the large-scale surveys as carried out under the EU project PARSE-Inisight (see [www.parse-insight.eu](http://www.parse-insight.eu)) the majority of respondents from research communities (50–70%) regard these threats as ‘important’ or even ‘very important’ – see Figure 1. Lack of sustainable hardware, software, or support that may make the information inaccessible raised the highest concerns (80% ‘important’ or ‘very important’), while access restrictions cause the fewest worries (19% very important, 37% important). Uncertainty about the future existence of custodians of the data are another high-scoring concern, together with uncertainty about the origin and authenticity of data.<sup>12</sup>

### Strategies for digital preservation

Several strategies are being entertained for digital preservation. While the notion increasingly takes hold that a proper preservation environment encompasses both systems *and* people, i.e. not just technical solutions but also an environment of organi-

*two key elements are central: usability and understandability; authenticity over time*

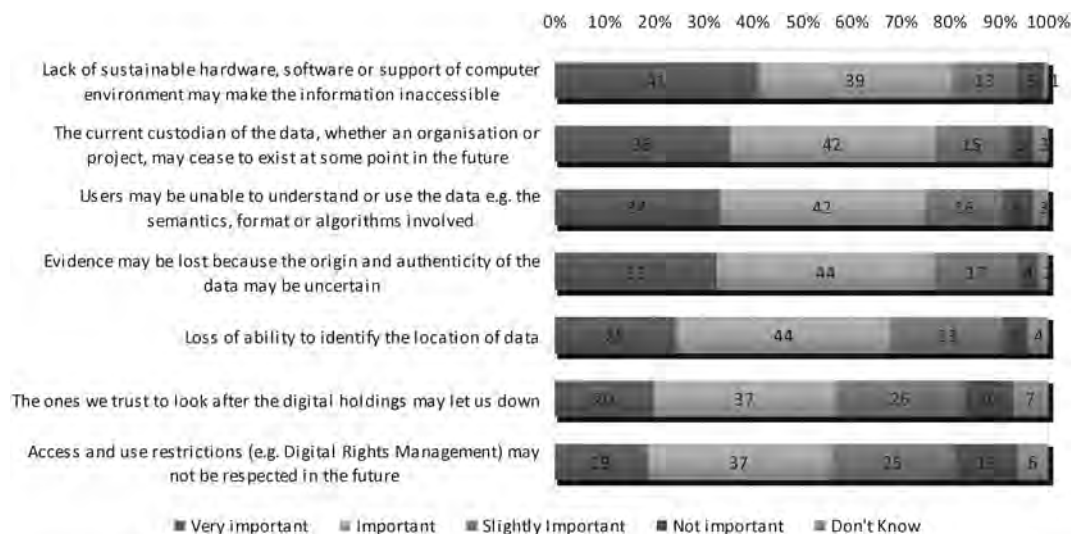


Figure 1. Threats to digital preservation,  $n = 1,209$

Source: PARSE.insight survey report.

*most of the strategies take a strong IT perspective*

zation and infrastructure, most of the strategies take a strong IT perspective and deal mainly with the technical processes and activities necessary for digital preservation. See for example a very useful summary on Wikipedia of Digital Preservation.<sup>13</sup>

From this technical and IT perspective, the main digital preservation strategies can be summarized as follows:

1. *Normalization/standardization*: systematically ensuring that the file structure and format of the digital objects are similar, following the same agreed norm and standard (usually the simplest and commonly used, e.g. most text documents are preserved as PDFs);
2. *Refreshing*: because of physical decay of storage media, the digital data is transferred frequently onto a copy of the same storage medium to avoid data alterations, or copied to other media or otherwise repackaged (replication or repackaging).
3. *Transformation* (often referred to as *migration*): digital data is transformed, e.g. to more current formats that can be accessed with modern system environments.
4. *Emulation*: in the case of emulation the problem of system obsolescence is solved from the other angle – the data is not *per se* transferred or transformed but any

obsolete hardware and software is replaced by a new software program, called an emulator, to make the old data format accessible and readable again.

5. *Semantic metadating*: keeping the descriptions of the digital object and ensuring that they can be understood by the target user community in the future – this is often the most important mechanism for scientific data since it includes the tacit knowledge about the object. The semantics, i.e. the meaning, is usually independent of the format.
6. *Combination*: in most cases there is a combination of all of these strategies, with normalization and refreshing usually as standard elements. Most document centres and archives also apply regular migration and see emulation more as a rescue remedy of last resort. This may change, however, once emulation practices develop further and data archives become bigger and more diverse, posing new challenges to migration strategies. Emulation is of less use for scientific data as researchers usually want to be able to use the latest analysis systems. In this case transformation will sometimes be used, but the detailed descriptions and especially the semantics, which is usually independent of format, must be available.

Proper metadating has become a more prominent element more recently as a vital step for future understandability and authenticity.

From an organizational viewpoint and concerning a workable infrastructure that relies on sound common practices, digital preservation strategies are increasingly concerned with policy, economic and social aspects, such as:

1. *Common standards* for metadata such as descriptive information and representation information to ensure that information can be curated, found and understood over time.
2. *Central registries* of what is preserved where, information on provenance and authenticity, hardware and software needed, to ensure retrievability and access.
3. A system of (interoperable) *persistent identifiers* that help to locate well-preserved sources, interoperable with existing identifiers to ensure discoverability.
4. *Certification and auditing* procedures and processes that ensure trustworthy digital repositories for long-term preservation.
5. *Economic sustainability* over time for digital preservation actions – a useful step was made by the Blue Ribbon task Group as quoted earlier, but an important part of the strategy is to plan for the worst and ensure that digital holdings can be handed over to the next member in the preservation chain without loss of information
6. A *holistic approach* involving all stakeholders in the information chain, ensuring that generators of digital information and objects include a data-management plan in their research proposals including good storage, archiving, and preservation, and that any later link in the chain reinforces this, from data managers to publishers, librarians, research funders, policymakers, etc.
7. *Future citability and rights management* to ensure that researchers receive credits for their work.
8. *Proper training* of (young) researchers on preservation issues for their research output.

There is also a growing notion on the impor-

tance of *convergence*. Digital preservation is not helped if everybody is doing something differently somewhere else. Paradoxically, digital preservation is also under great risk if everybody everywhere is doing exactly the same thing. In preparing ourselves for an unknown future, with unforeseen developments, we should render many different options without getting fragmented, and avoid putting all eggs in one basket.

Hence, a sensible digital preservation strategy:

- Combines different technical strategies and is open to any new ones.
- Ensures a network of several trustworthy deposit places and archives internationally.
- Involves all stakeholders in the information chain, from authors, to data managers to publishers to libraries and archives.
- Ensures an infrastructure with interoperable identifiers and other services that support preservation, such as consistent metadating.

#### **The present state of digital preservation: EU project PARSE.Insight**

EU project PARSE.Insight<sup>4,9,12</sup> was one of the first international projects that aimed to present an overview of the present state of affairs in digital preservation among the main stakeholders: research institutes, libraries and data managers, funders, and publishers. Together with eight other participants from the research and library world, the International Association of STM Publishers was one of the project partners.

Full reports of the project can be found on [www.parse-insight.eu](http://www.parse-insight.eu). Below is a summary of the outcomes relevant to a publisher's perspective on the subject of this paper. Within the PARSE.Insight study, the European research landscape was subdivided into four main stakeholders: researchers, data managers, publishers, and funders. Several methods were deployed to gather information from these stakeholders on their practices, ideas, and needs to guarantee long-term access to research output. These methods comprise desk research, in-depth interviews, case studies into three specific research communities, and large-scale sur-

*Paradoxically, digital preservation is also under great risk if everybody everywhere is doing exactly the same thing*

*In practical terms, raw data, processed data, publications, and post-publication material are all covered by the same term*

veys. The aim of the PARSE.Insight study was to provide good insight into the current state of affairs and needs regarding digital preservation of research output in Europe and give recommendations about future investments for building a sustainable digital ecology. The study was conducted from March 2008 to June 2010, co-financed by the EU Seventh Framework Programme (FP7). All graphs depicted here below originate from the PARSE.Insight survey report.<sup>14</sup>

#### Main outcomes

All four stakeholder groups in this study agree that preservation of research output is important. Reasons such as *it may stimulate the advancement of science* and *it allows for reanalysis* were acknowledged. But the fact that preservation is by no means simple was commonly understood as well. Threats to digital preservation such as *lack of sustainable hardware, software, and evidence may be lost because the origin and authenticity of the data may be uncertain* were acknowledged by all stakeholders although they vary sometimes. For example, in the disciplines of high energy physics and earth observations, experiments simply cannot be redone easily (if at all). For a complete list of reasons and threats, see below in this summary report.

#### Digital research data

In the context of PARSE.Insight the term 'data' is used for all research output. In practical terms, raw data, processed data, publications, and post-publication material are all covered by the same term. A distinction between these sorts of research data is only made when necessary (e.g. when policies for publications are compared with other data).

To researchers, the reasons for preservation of research data are clear – see Figure 2. Researchers mention as the main threats to preservation the lack of technical support, human errors, and the lack of structural funding for preservation measures and infrastructure, as can be understood from Figure 3. Evidently, threats which are more current to their day-to-day work score higher than those that are not so likely to occur in the short term; though it does not make these less real.

#### Publishers' preservation policies

Most publishers have their house well in order regarding digital preservation. This is one of the main conclusions of project PARSE.Insight. 84% of the larger publishers (>50 journals in their catalogue) have a preservation policy in place for their publica-

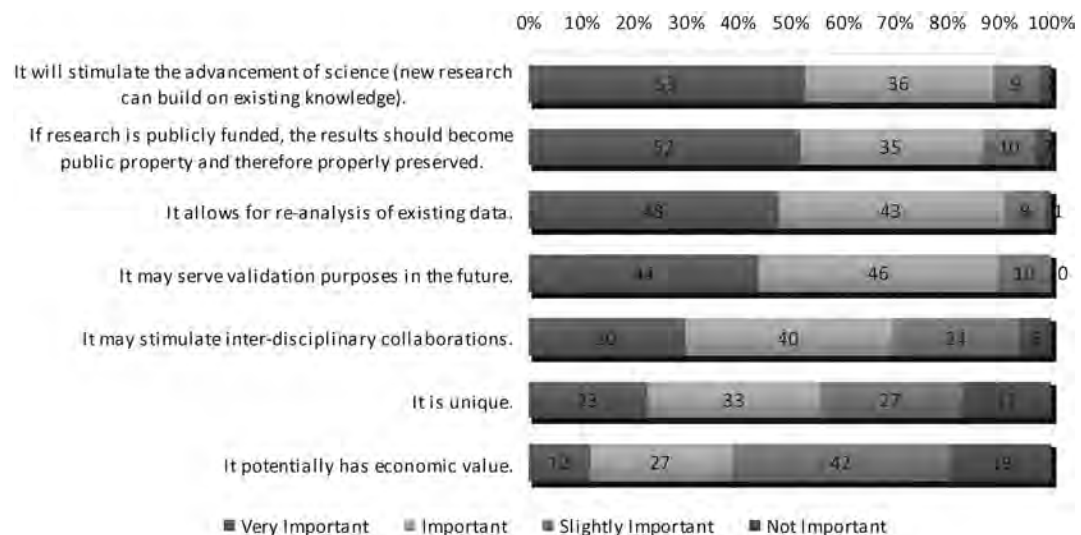


Figure 2. Reasons for preservation of research data, n = 1,213

Source: PARSE.insight survey report.

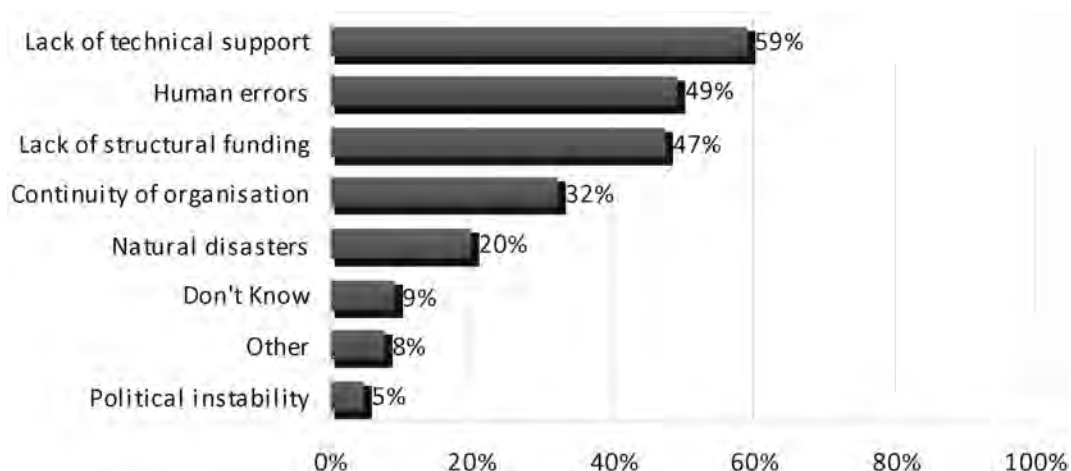


Figure 3. General threats to currently preserved digital data, n = 1,190

Source: PARSE.insight survey report.

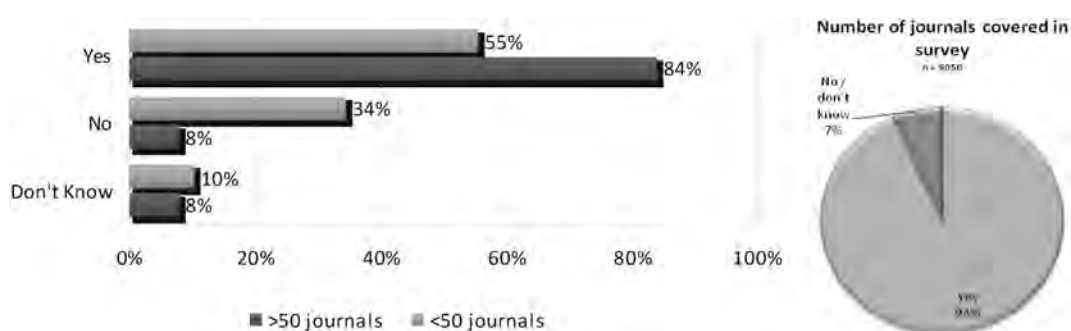


Figure 4. Prevalence of preservation policy, n = 153 publishers

Source: PARSE.insight survey report.

tions. Among the smaller (often single-title open access) publishers this is only 55% (see Figure 4). Due to the large impact of the publishers with many journals in their catalogue (Springer, Elsevier, Wiley, etc.), >93% of the journals covered in this survey are covered by a preservation policy. The survey received responses from publishers who together publish >9,000 journals which is close to 40% of all active, peer-reviewed journals (see the 2009 STM Report on [http://www.stm-assoc.org/2009\\_10\\_13\\_MWC\\_STM\\_Report.pdf](http://www.stm-assoc.org/2009_10_13_MWC_STM_Report.pdf))

Most larger publishers rely for the preservation of their publications on outsourcing to trusted parties such as Portico and the KB.

While this means that the preservation of journal publications is well managed, the

situation is less bright for supplementary material submitted by authors to publishers in conjunction with their manuscript. While >70% of larger publishers (and <60% of smaller) accept such supplementary material from authors, covering >90% of all journals, only 10% take special preservation measures for these data and other material. Another 20% treat them similar to the normal publications (which is most probably not sufficient to ensure future reuse) and close to 70% do not take any preservation measures for supplementary material (Figure 5).

Apart from the supplemental data to publications, the publication itself is expected to change. 63% of the respondents of publishers in the survey believe that *publications will become interactive and multimedia* (e.g.

*the situation is less bright for supplementary material submitted by authors to publishers*

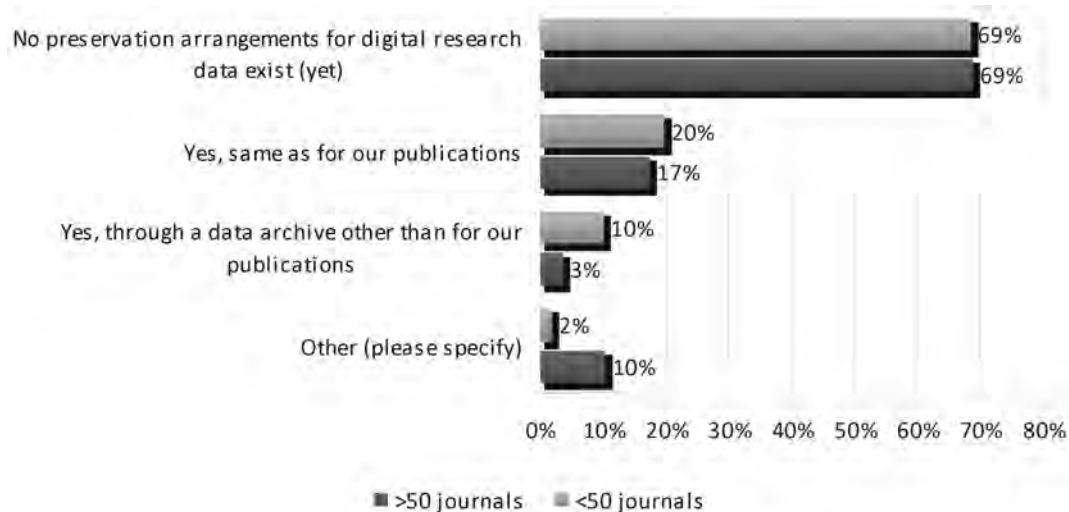


Figure 5. Prevalence of preservation arrangements for underlying digital research data,  $n = 150$  publishers

Source: PARSE.insight survey report.

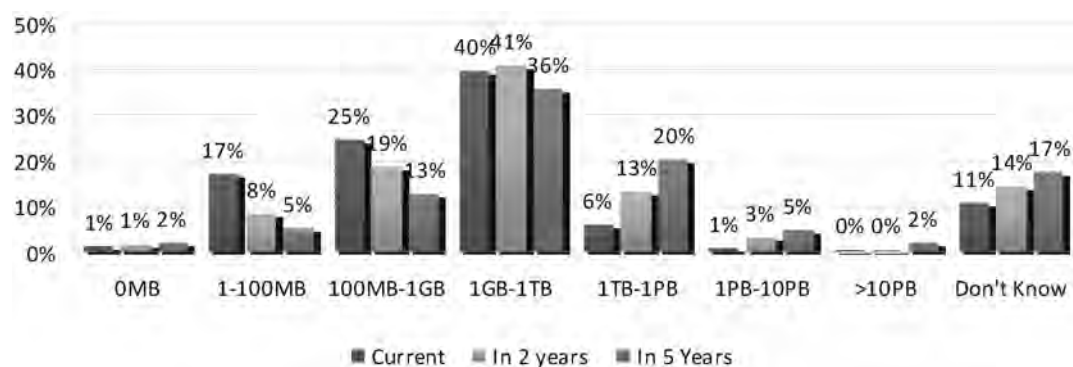


Figure 6. Estimated data stored per project,  $n = 1,296$

Source: PARSE.insight survey report.

adding animations, sound, related Web content, research data, discussion forum).

#### How to handle underlying data

The fact that official publications are much better protected in terms of digital preservation than the underlying research data and other supplementary material is perhaps not surprising, but is a worrying conclusion. Even more so if we consider that we are currently facing a data-explosion on the Web. Researchers in this survey expect that the present gigabyte average for data volumes in

research projects will develop into a tera- and petabyte era within the next 3–5 years (Figure 6).

Therefore, publishers have to be prepared. Funders, grant-providers, journal editors, and the science community in general will increasingly wish to be able to consult the underlying data of research projects in combination with publications. Does this mean that publishers should take on a new role for storing and making available research data? In some cases perhaps yes, but most probably so in a good collaboration with other players in this data-sphere. While

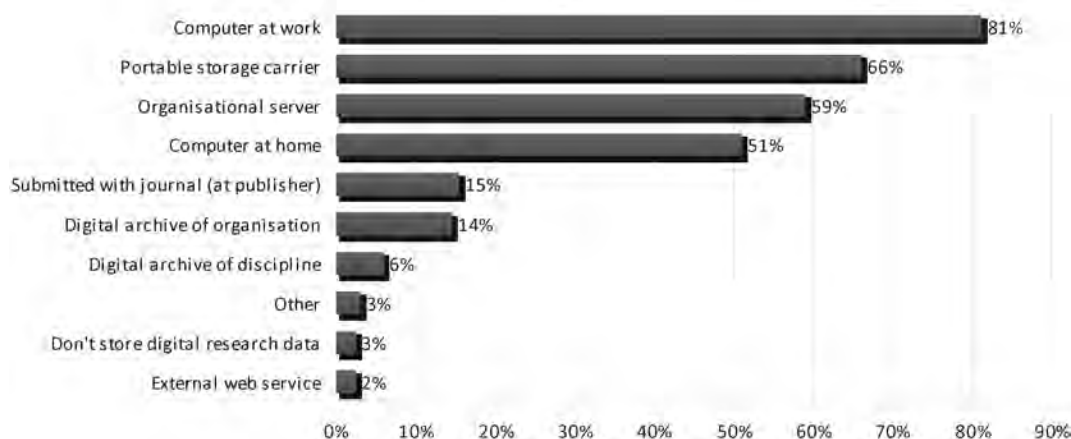


Figure 7. Where do researchers store data?,  $n = 1,202$

Source: PARSE.insight survey report.

nearly half of the authors surveyed say that they would like to submit their data to publishers, together with the article, right now, only 15% of researchers submit their data to publishers – most of the data are on their computer, at work (81%) or at home (51%) and on portable disks (66%), or at best on the server of their organization (59%) (Figure 7).

Proper digital archives, at their organization or for their discipline, are not yet very popular (14% and 6%).

#### Who should be data custodians?

Who can best manage and handle the growing volume of research data as compared to the official version-of-record publications? We have seen the multitude of practices that presently exist among researchers with huge variation between different disciplines. Publishers have mixed feelings and most see this as primarily a role for authors, their institutes, the national libraries, and the research community. This is probably a wise position, certainly in view of the diversity in data types and the challenging requirements and related costs involved in accurate preservation. A likely development is that new, specialized datacentres will emerge who take an expert role in this. However, there is also a significant 40% of publishers who see a role for themselves, probably to ensure that data

and publications can be consulted easily in combination (Figure 8).

These outcomes only emphasize how important it is to have proper links between research data and the related publications. 84% of researchers find it important that these links exist. Also, over 60% of researchers locate existing data via publications. In this context it is relevant to point out new initiatives such as Datacite, a consortium of over 10 large (national) libraries around the world (see [www.datacite.org](http://www.datacite.org)). The consortium offers registration services for datasets from researchers when they have deposited these in trustworthy datacentres. Via Datacite, the dataset gets a DOI assigned. Publishers can use this DOI for the dataset to ensure persistent links from the submitted publication to that dataset, and the Datacite registry ensures the reverse links from the dataset to all related publications.

#### PARSE.Insight conclusions

As the survey outcomes of PARSE.Insight show, the current state of affairs on preservation of research output in Europe is diverse and fragmented. Preservation of publications is covered pretty well by publishers and data managers/libraries, but looking at the broader spectrum of research data (e.g. data sets, software) the outcome is less bright. Some data repositories do exist, but few disciplines are covered and organization

*right now, only 15% of researchers submit their data to publishers*

*how important it is to have proper links between research data and the related publications*

differs from country to country. Only a few guidelines exist and the ones under development are created mostly in isolation. Funders can play a significant role but currently are more focused on access to data in the short term than looking at data preservation for future generations. Publishers should extend their preservation responsibilities to more than just their own publications, with a more prominent role in ensuring the right connections that will be persistent over time, between official publications and related research output.

To cope with current threats of preservation, all stakeholders studied by PARSE.Insight agree that a science data infrastructure is required. But a simple description of the roles specifying who should do what in this matter is far from clear. Roles should be defined more explicitly and business models should be developed. For this, strong co-ordination is required. In addition, awareness of what exactly digital preservation is and what should be done needs to be increased.

#### **Digital preservation: why should STM publishers care?**

Back to what publishers in STM can and should do. STM Publishing is about providing high-quality services to authors around the world; it is about serving the dissemination of knowledge and offering discoverability through the growing body of literature; and it is about long-term stewardship for the body of knowledge that builds up cumulatively in all publications. Authors publish in order not to perish, they hope to live in perpetuity via their ideas and breakthroughs. They publish for validation, acknowledgement, and attributable authenticity via certification. As longitudinal studies show,<sup>15</sup> establishing 'precedence' and 'recognition' is for authors an increasing motivation to publish in research journals ([http://www.stmassoc.org/2006\\_09\\_01\\_Scientific\\_Publishing\\_in\\_Transition\\_White\\_Paper.pdf](http://www.stmassoc.org/2006_09_01_Scientific_Publishing_in_Transition_White_Paper.pdf)).

These provide enough reasons why publishers should care about preservation in general. Moreover, the US-based Task Force on Archiving Digital Information identified

as long ago as 1996<sup>16</sup> one of the main principles for digital preservation: information creators/providers/owners have initial responsibility for archiving their digital information objects and thereby ensuring the long-term preservation of those objects.

Ever since the rise and fall of the library of Alexandria, most publishers have taken this responsibility seriously. Apart from inherent reasons related to the nature of digital decay, the digital age provides an extra set of reasons why publishers should care. Here are a few:

1. The official version of record that a scientific publication represents is increasingly being accompanied with a multitude of additional, supplementary, related, auxiliary, and underlying versions or material and links to related material on the Web. These resources often come in many different formats, non-text (tables, spreadsheets, and a multitude of scientific data format) or even multimedia (videos, podcasts, model animations, and simulations). Moreover, the representation of a publication itself has started to change with such things as outbound links to references, embedded sound and animation.
2. The data deluge on the Internet is very apparent, certainly in the area of STM where nearly every discipline is undergoing a strong funding trend in computational and often data-centric research. Research funders have started to require from grant proposals that there is a plan for data management: how the data generated via the project will be stored, maintained and made available for reuse. This means there will be an increasing number of datasets added to articles, but also more pointers from articles to data-resources and vice versa.
3. With the advent of an open access movement and self-archiving requirements for any and all kinds of research output into institutional repositories (independent of their level of success in serving researchers and scientists), publications may become increasingly multiversions and multi-sourced, providing a hub that links to many different sources such as gene databases, compound information, and raw

research data in subject specific repositories.

For all these new and evolving developments around research articles, digital preservation is not yet well established, if at all, as the findings of the PARSE.Insight study have shown. Those same surveys show that for nearly all official journal publications, digital preservation has been catered for – can publishers allow themselves to say: so why bother with the rest? Or in other words: why can't publishers simply say: we have taken care of the preservation of our own digital publications – it is up to others to do the rest. Well . . . yes they can, and no, they cannot.

If we view the publishing world as a closely interlinked ecosystem, then this system requires the interaction, collaboration, and contribution of every stakeholder to maintain it. Or to take the analogy of a chain consisting of many links, then this chain will be as strong as its weakest link. If publications increasingly become a part of a wider network of information, original data, shared databases, and other sources containing multimedia expressions and even social media, then digital preservation should encompass all of this. And the formal publication, the version of record, should firmly sit in the middle of all of this. Its links to evidentiary, auxiliary, and supplementary material should be persistent and sustainable, its metadata should provide, permanently, all the necessary information that facilitates reusability, understandability, and supports its authenticity. In fact, the official article may grow to be one of the main sources of high-quality metadata of all underlying 'stuff'. To do this well over time, digital preservation should be at the heart of this.

To establish such a situation, it is important that publishers engage with all other stakeholders in the scientific information chain about issues, policies, and infrastructures to increase sustainability of research output. This will help create an ecosystem in which publishers contribute and fulfil their role of establishing a reliable body of knowledge via trusted publications that rests on the following foundations:

- Persistent identifiers that help link official publications with all other underlying research output in perpetuity.
- A network of safe and trusted repositories for research output, certified along agreed auditing standards, to which publishers can reliably link in perpetuity for auxiliary and underlying material of research papers and to which they can refer their authors to deposit such material if a publishers decides not to include it in the journal or on the journal site.
- New common practices about citability of research output other than official publications, or parts of it, ensuring reliable information on authenticity and provenance of research output.
- Agreed codes of conduct for the availability of auxiliary or underlying material to research publications (e.g. if conclusions are heavily based on data analysis, the original data must be made available for the reviewers or even the readers).
- Peer-review standards for the assessment of datasets or even raw research data and other auxiliary material, adhering to general quality standards for data.
- Agreed access management rules for such material.
- And so on.

#### Next steps

The International Association of STM Publishers was happy to participate in project PARSE.Insight for exactly these reasons. With representatives from research institutes, from national libraries, from research funding organizations and (via STM) from publishers, the project brought together all essential key players in the scientific information chain. This led to useful discussions and good project outcomes (see above). The discussions on a roadmap that secures a proper infrastructure for digital preservation will continue under two newly approved EU projects: ODE and APARSEN. ODE stands for Opportunities for Data Exchange and will run from the end of 2010 until the end of 2012. APARSEN is a so-called Network of Excellence (APARSEN = Alliance for Permanent Access to the Records of Science in Europe – Network) and will run from

*If we view the publishing world as a closely interlinked ecosystem, then this system requires the interaction, collaboration, and contribution of every stakeholder to maintain it*

*the official article may grow to be one of the main sources of high-quality metadata of all underlying 'stuff'*

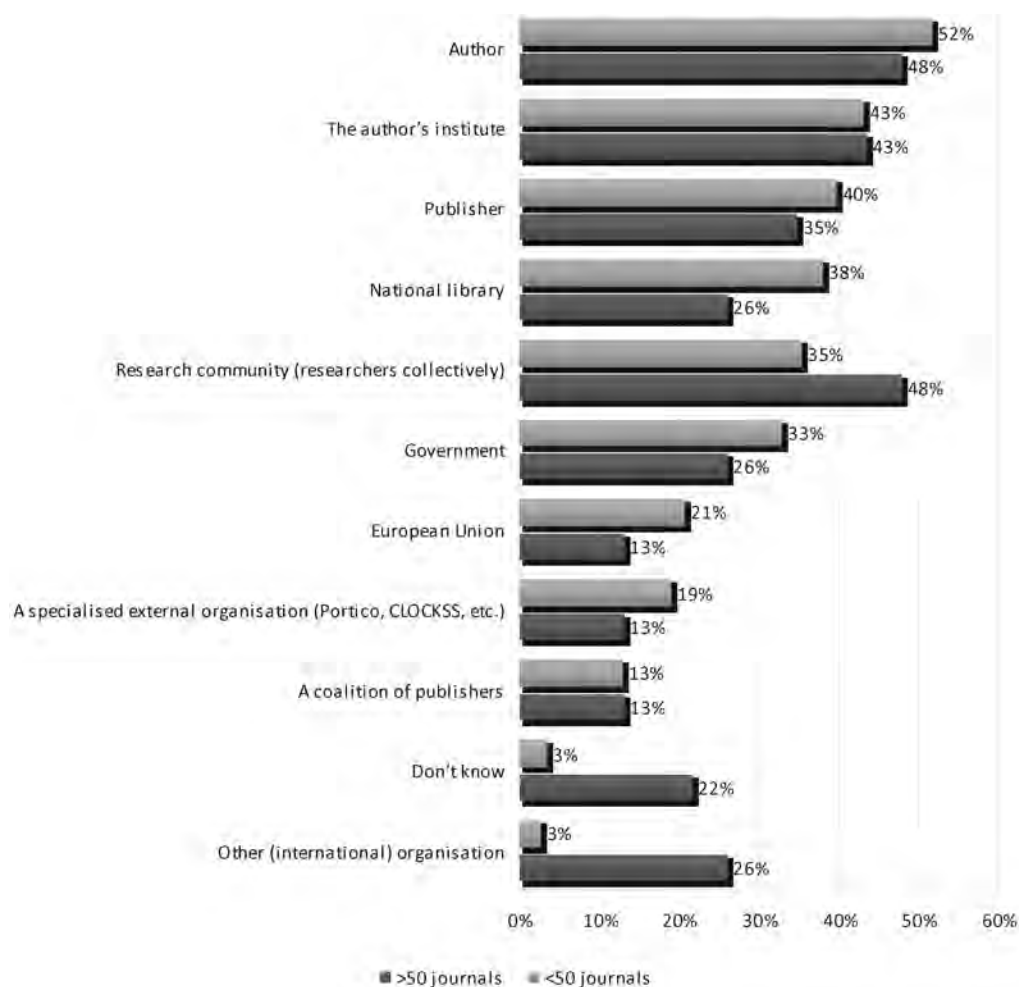


Figure 8. Responsibility for preservation of digital data,  $n = 139$  publishers

Source: PARSE.insight survey report.

the end of 2010 until the end of 2014. The original group of nine participants in PARSE.Insight has grown to 30 participants in APARSEN, with a much wider international and organizational span. On the agenda are all aspects as described here above that will serve as a foundation to a proper infrastructure, of systems, organization and people, for digital preservation. Keywords on which the networks will focus are trust, sustainability, usability and access across all stakeholders in the digital preservation of research output. The concerted actions serve one joint aim: to avoid a Digital Dark Age for research data.

#### References

1. Library of Alexandria: <http://www.crystalinks.com/libraryofalexandria.html>
2. Kuny, T. A Digital Dark Age? Challenges in the preservation of electronic information . 63RD IFLA (International Federation of Library Associations and Institutions) Council and General Conference, September 1997.
3. Blakeslee, S. Lost on earth – wealth of data found in space. *New York Times*, March 1990. <http://query.nytimes.com/gst/fullpage.html?res=9C0CE2DE1638F933A15750C0A966958260&sec=&spn=&pagewanted=all>
4. Project PARSE.Insight see [www.parse-insight.eu](http://www.parse-insight.eu). PARSE is an acronym for: Permanent Access to the Records of Science in Europe.
5. Alliance for Permanent Access: <http://www.alliancepermanentaccess.org>

6. Rothenberg, J. 2001: <http://www.amibusiness.com/dps/rothenberg-arma.pdf>
7. British Library Web Archiving: <http://www.bl.uk/aboutus/stratpolprog/digi/webarch/index.html>
8. Blue Ribbon Task Force: <http://brtf.sdsc.edu/>
9. See also: [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D3-6\\_InsightReport.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf)
10. OAIS, described in ISO 14721:2003, see [www.iso.ch](http://www.iso.ch), and available without charge from [www.ccsds.org](http://www.ccsds.org)
11. Giaretta, D. *Advanced Digital Preservation*. Berlin, Springer Verlag, 2011. ISBN 978-3-642-16808-6.
12. See [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D2-2\\_Roadmap.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf)
13. [http://en.wikipedia.org/wiki/Digital\\_preservation](http://en.wikipedia.org/wiki/Digital_preservation)
14. [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D3-4\\_SurveyReport\\_final\\_hq.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf)
15. Mabe, M. and Ware M. The STM Report: Overview of Scientific, Technical and Medical Publishing, 2009: see [http://www.stm-assoc.org/2009\\_10\\_13\\_MWC\\_STM\\_Report.pdf](http://www.stm-assoc.org/2009_10_13_MWC_STM_Report.pdf); and Ware M. Scientific Publishing in Transition, 2006, White Paper: [http://www.stm-assoc.org/2006\\_09\\_01\\_Scientific\\_Publishing\\_in\\_Transition\\_White\\_Paper.pdf](http://www.stm-assoc.org/2006_09_01_Scientific_Publishing_in_Transition_White_Paper.pdf)
16. Task Force on Archiving Digital Information, 1996: <http://www.oclc.org/research/activities/past/rlg/digpresstudy/final-report.pdf>

#### Recommended reading

- Riding the Wave, How Europe Can Gain from the Rising Tide of Scientific Data, High Level Expert Group to the European Commission, October 2010. <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>
- Patel, M. and Ball, A. 2008 Challenges and issues relating to the user of representation information for the digital curation of crystallography and engineering data. *International Journal of Digital Curation*, 3(1): 76-88. Available from [www.ijdc.net/index.php/ijdc/article/download/64/43](http://www.ijdc.net/index.php/ijdc/article/download/64/43)
- Conway, E. 2009 Curating Atmospheric Data for Long Term Use: Infrastructure and Preservation Issues for the Atmospheric Sciences community. Digital Curation Centre Scarp project case study.
- Conway, E. and Dunckley, M. Preservation network models: creating stable networks of information to ensure the long term use of scientific data. In: Proc. Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data (PV 2009), Villafranca del Castillo, Spain, 2009.

- Conway, E., Giaretta, D., and Dunckley, M. Curating scientific research data for the long term: a preservation analysis method in context. Proc. 6th International Conference on Preservation of Digital Objects (iPres 2009), San Francisco.
- Matthews, B., Shaon, A., Bicarregui, J., Jones, C., Woodcock, J., and Conway, E. Towards a methodology for software preservation. Proc. 6th International Conference on Preservation of Digital Objects (iPres 2009), San Francisco.
- It is interesting to note that the Vatican Library has adopted the FITS image data format for digitization of its 80,000 manuscripts. Report of Vatican use of FITS format – see [http://www.dariah.eu/index.php?option=com\\_wordpress&p=27&Itemid=198](http://www.dariah.eu/index.php?option=com_wordpress&p=27&Itemid=198)
- Tzitzikas, Y., Marketakis, Y., and Antoniou, G. Task-based dependency management for the preservation of digital objects using rules. In: Proc. 6th Hellenic Conference on Artificial Intelligence (SETN'10), Athens, 2010
- Marketakis, Y., Tzanakis, M., and Tzitzikas, Y. PreScan: towards automating the preservation of digital objects. Proc. International ACM Conference on Management of Emergent EcoSystems (MEDES'09), Lyon, 2009, pp. 404-411.
- Tzitzikas, Y. and Marketakis, Y. 2010. Automating the ingestion and transformation of embedded metadata. *ERCIM News*, 2010(80).
- National Science Foundation Cyberinfrastructure Council. 2007. Cyberinfrastructure Vision for 21st Century Discovery. Retrieved from <http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf>

#### Eefke SMIT

*International Association of STM Publishers*  
 Director Standards and Technology  
 Email: [smit@stm-assoc.org](mailto:smit@stm-assoc.org)

#### Jeffrey VAN DER HOEVEN

Researcher  
 Koninklijke Bibliotheek (KB)  
 The Netherlands

#### David GIARETTA

Director  
 European Alliance for Permanent Access