Library Hi Tech

An analysis of file format control in institutional repositories

Miquel Termens Mireia Ribera Anita Locher

Received 5 October 2014
Revised 7 March 2015
Accepted 18 March 2015

An analysis of file format control in institutional repositories

Miquel Termens, Mireia Ribera and Anita Locher
Library & Information Science Department,
Universitat de Barcelona, Barcelona, Spain

Abstract

Purpose – The purpose of this paper is to analyze the file formats of the digital objects stored in two of the largest open-access repositories in Spain, DDUB and TDX, and determines the implications of these formats for long-term preservation, focussing in particular on the different versions of PDF.
Design/methodology/approach – To be able to study the two repositories, the authors harvested all the files corresponding to every digital object and some of their associated metadata using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and Open Archives Initiative Object Reuse and Exchange (OAI-ORE) protocols. The file formats were analyzed with DROID software and some additional tools.
Findings – The results show that there is no alignment between the preservation policies declared by institutions, the technical tools available, and the actual stored files.
Originality/value – The results show that file controls currently applied to institutional repositories do not suffice to grant their stated mission of long-term preservation of scientific literature.

Keywords Digital preservation, Institutional repositories, File format, PDF
Paper type Research paper

1. Introduction

The risks involved in long-term preservation of digital objects are complex to categorize, and there is no consensus on the best solutions for each specific case (Vermaaten *et al.*, 2012; Graf and Gordea, 2013). Although some experts state that since internet adoption – and particularly since mainstream use of the Web began – no format has been deprecated severely enough to prevent its use (Rusbridge, 2006; Rosenthal, 2010; Rosenthal, 2013), format obsolescence is the most commonly cited technical problem challenging content preservation (Lawrence *et al.*, 2000; Pearson and Webb, 2008). This complexity justifies the focus of the paper, oriented toward analyzing current management practices of two technical characteristics of the files uploaded to repositories, their format and their encryption. The paper will not enter into details of their implications on long-term preservation policies.

All repositories store digital objects with a dual aim: first, to promote their dissemination; and second, to guarantee their preservation (Ware, 2004; van Westrienen and Lynch, 2005; Kennan and Wilson, 2006). The first aim is the most evident and was often the initial reason for creating the repositories. The second aim is often not explicitly mentioned and repository holders do not guarantee its fulfillment through either established policies or resources. It is a common practice to focus technical and economic efforts on attracting and disseminating new content, and to

This study received a grant from the project *El acceso abierto* (*open access*) *a la ciencia en España*. 2012-2014. Plan Nacional I+D+i, código CSO2011-29503-C02-01. The authors thank Yvonne Friese of the Deutsche Zentralbibliothek für Wirtschaftswissenschaften for the use of her PDF scripts. The authors also thank the CBUC and the UB's CRAI for their help with the data interpretation.

Library Hi Tech
Vol. 33 No. 2, 2015
pp. 162-174

0737-8831
DOI 10.1108/LHT-10-2014-0098

leave preservation-related tasks to later stages, when the repository has been in use for some years. Furthermore, many repository managers strongly believe that in the short run preservation problems are not serious and can be solved by traditional computer security measures. This attitude is currently widespread, but we question its validity and wonder how long repository managers can ignore the need for proper control of archived files and metadata.

Previous studies have dealt with features of formats stored in large preservation repositories. Jackson (2012) led the most comprehensive research in this area, working with about 2.5 billion files collected between 1996 and 2010 from the JISC UK Web Domain Data set. His results show that in this period contents in HTML format, with a major presence on the Web, evolved from version 2.0 (the oldest one) to XHTML 1.0 (the newest one at the time of writing), and that older HTML versions were progressively replaced by newer ones.

The PDF format is widely used in repositories and is favored by repository holders because of its good preservation properties. A recent study with managers of the 118 repositories of the Association of Research Libraries in the USA found that they considered PDF as the third preferred choice for perdurability, after TIFF and WAV. (Rimkus *et al.*, 2014). It is also particularly noteworthy that there is a PDF version oriented specifically to preservation, the PDF/Archiving version (PDF/A-1 ISO 19005-1:2005 and PDF/A-2 ISO 19005-2:2011). However, a recent study of Swedish PhD dissertations deposited during the period 2003-2012 (Fischer and Lundell, 2013) found that many of the PDF/A files failed to meet the standards for this format.

A special feature of PDF is the fact that document owners can protect files against user actions such as modification, printing, copying, and even reading. These restrictions affect preservation because they make it difficult to migrate the content to a format that may be considered more suitable in the future.

This study considers whether the currently used file formats are the best ones to guarantee – or at least permit – long-term preservation of stored contents? We present a case study of the digital objects stored in two Spanish open-access repositories, paying special attention to the PDF format. We analyze the formats, the versions and the level of encryption of the objects and confront them with the long-term preservation mission of the repositories holding them.

2. Methodology

The analysis was carried out on the content of two open-access repositories:

(1) DDUB: Dipòsit digital de la Universidad de Barcelona (UB) (http://diposit.ub.edu/). This is the institutional repository of the UB (www.ub.edu), a public university that is second in Spain in number of students. According to its web site, this repository, created in 2006, contains the "open-access digital versions of publications related to the teaching, research, and institutional activities of the UB's teaching staff and other members of the university community." DDUB is managed by the Centro de Recursos para el Aprendizaje y la Investigación (CRAI), the library of the UB. Document ingestion is decentralized and is carried out by the people in charge of the repository collections. Collection owners may be administrative or technical staff or faculty. In this repository there are no restrictions on the format of the file to be ingested. There are also low requirements concerning metadata in order to avoid creating barriers to the submission of documents, as metadata are introduced by the deposit author. However, to ensure quality the data were validated by CRAI

staff, following the Dublin Core schema. The University of Barcelona is institutionally committed to the open-access movement (University, 2003), as evidenced by an institutional mandate favoring open-access and the existence of the Knowledge Dissemination Office within the CRAI, which promotes the dissemination of the UB's scientific production in open access.

(2) TDX: Tesis Doctorales en Red (www.tdx.cat/ and www.tesisenred.net). This is a subject-oriented cooperative repository created in 2001. It contains PhD dissertations of 18 public and private Spanish universities and one Andorran university, and is the largest PhD repository in Spain. TDX is managed by the Consorci de Biblioteques Universitàries de Catalunya (www.csuc.cat/ca/biblioteques-cbuc), a university library consortium located in Barcelona city, but the submission of PhD dissertations is managed by each participant university; this distribution of responsibilities facilitates a high level of control of the files and metadata ingested. TDX also belongs to the International Networked Digital Library of Theses and Dissertations (www.ndltd.org/) and to the North American MetaArchive consortium (www.metaarchive.org/), which is specialized in digital preservation and establishes protocols and a duplication system to ensure the long-term preservation of contents held by its members. The vast majority of dissertations are stored in PDF format, and most of the files are created by the authors. However, for the massive digitalization of old dissertations, the files were created by a third-party company (Anglada *et al.*, 2002a, b). In both TDX and DDUB the metadata follow the Dublin Core schema.

We chose these repositories because they hold a sufficient volume of digital objects to reach significant results and because they represent two of the most common types of repository: one is linked to the activity of an institution, and the other is linked to a subject or a document type. Both repositories use Dspace software (www.dspace.org/) created by the Massachusetts Institute of Technology. Within DDUB (the institutional repository), we can check the coherence between the theory of an institutional open-access policy and the reality of the objects stored in the repository. The DDUB also explicitly mentions long-term preservation in its objectives (http://diposit.ub.edu/dspace/quees_es.jsp9). Within TDX (the subject-oriented repository), we can check whether greater homogeneity in submission workflows really leads to higher homogeneity in the technical format of contents. TDX does not include long-term preservation explicitly in its objectives (www.tdx.cat/pmf#objectius) but does mention the measures taken for this purpose. None of the repository policies and documents analyzed makes any reference to data encryption, so we expected to find some PDF files with some kind of protection.

To be able to study the two repositories, we harvested all the final files (bitstreams) corresponding to every digital object and some of their associated metadata in local storage. Harvesting was done using the OAI-PMH and OAI-ORE protocols. OAI-PMH (www.openarchives.org/pmh/) is a protocol for collecting digital object metadata through http calls to the repository server; in our case study we collected the following metadata: identifier, title, collection, name and extension of the component files (bitstreams), cataloguing data, and publishing data. We used the OAI-ORE (www.openarchives.org/ore/) protocol to collect the final files of each object.

After collection, the files were analyzed using DROID software, version 6.1.3 (www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm), created by the National Archives of the UK. DROID identifies uniquely the digital format and version of a file, thanks to the PRONOM (www.nationalarchives.gov.uk/PRONOM/)

(Brown, 2005) database information. This functionality of the program fostered its adoption in digital repositories with a twofold purpose: as an entry control to avoid the submission of formats that the institution has not selected as priority; and as a way to identify the stored file format in real time (Brody *et al.*, 2007; Hitchcock *et al.*, 2007) and to include this technical information in the digital object in order to establish preservation policies suitable to the specific problems of stored files (Tarrant *et al.*, 2011).

In our study, DROID was used to identify the file format and version of the collected digital objects. For each file we registered the following data: PUID designation in the PRONOM database, format name, file extension, format version, and file size in kilobytes. The total number of files analyzed by DROID was 89,947 from DDUB and 41,925 from TDX; this number is higher than the actual number of collected files because for compressed files DROID analyzes each component file separately.

DROID is capable of recognizing only a limited set of file formats: the ones included in the PRONOM database. It was unable to recognize the format of some files from DDUB, including some formats that are common within some disciplines included in DDUB. After a superficial analysis of these files we found that many of them contained programming code, such as C language code or Python code, and a minority of them were corrupted files. Other authors have warned of DROID's limitations for correctly recognizing PDF versions, namely PDF/A (Jackson, 2012), but this restriction did not affect the results of our experiment.

The collection process was automated with the development of a harvester in Java language. The data of DDUB were collected on September 27, 2013 and those of TDX on October 4, 2013. Our harvester program also performed data integrity validations, called the DROID program to analyze the files, and presented the results in a CSV file.

Finally, we checked encryption and protection against modification, copy or printing on a random sample of collected PDF files. This check was done with the help of a script created with iText open software libraries (http://sourceforge.net/projects/itext/) provided by Yvonne Friese from the Deutsche Zentralbibliothek für Wirtschaftswissenschaften in Kiel.

3. Results

A total of 41,925 files were retrieved in TDX, of which 41,798 were in 29 formats correctly identified by DROID, and the others in 125 formats not identified by DROID. The distribution of the most popular formats is displayed in Table I.

As was expected for PhD dissertations, in TDX the vast majority of files were formatted in PDF. Other formats were identified very rarely in the attachments to the body of the document.

Although we had to discard a total of 273 files with incorrect submission data, the number of files was large enough to observe how different versions of PDF were used

Format	Extension	Files	
Acrobat PDF	pdf	38,364	91.51%
JPEG File Interchange Format	jpg	2,949	7.03%
Plain Text File	txt	215	0.51%
Graphics Interchange Format	gif	71	0.17%
Others	others	326	0.78%
Total		41,925	100.00%

Table I. Distribution of ingested formats within TDX

over time. This distribution is displayed in Table II and Figure 1. As was expected, newer versions were used over time. The most popular PDF versions were 1.4 and 1.6, which were used in 35.67 and 42.88 percent of the files, respectively. New versions were adopted with some delay and old versions were still used even when updated versions were available. Indeed, use of PDF version 1.7, the last existing one, was very low and PDF/A, one of the most suitable file formats for preservation, was hardly used at all.

The situation of DDUB regarding formats is detailed in Table III. This repository displays a great dispersion of formats, as a logical consequence of the great variety of documents stored, including articles, reports, final theses, official reports, images, and

Year	PDF 1.1	PDF 1.2	PDF 1.3	PDF 1.4	PDF 1.5	PDF 1.6	PDF 1.7	PDF/A	PDF/X	Total annual
2000				1						1
2001	20	203	561	650		4				1,438
2002	5	143	1,353	1,016	9	8				2,534
2003		37	789	2,041	6	13				2,886
2004	1	96	541	1,935	19	39				2,631
2005		230	49	2,343	585	44				3,251
2006		136	259	1,882	773	204				3,254
2007		116	286	1,745	253	988	1			3,389
2008		20	115	1,109	554	3,536	1			5,335
2009		14	94	302	308	3,570	6	1	1	4,296
2010		4	57	354	146	3,507	7	3		4,078
2011		4	48	84	101	2,797	38			3,072
2012			19	124	127	1,548	33			1,851
2013					1	74				75
Total	26	1,003	4,171	13,586	2,882	16,332	86	4	1	38,091
%	0.07	2.63	10.95	35.67	7.57	42.88	0.23	0.01	0.00	100.00

Table II.
Evolution of the version of PDF files ingested at TDX (2000-2013)

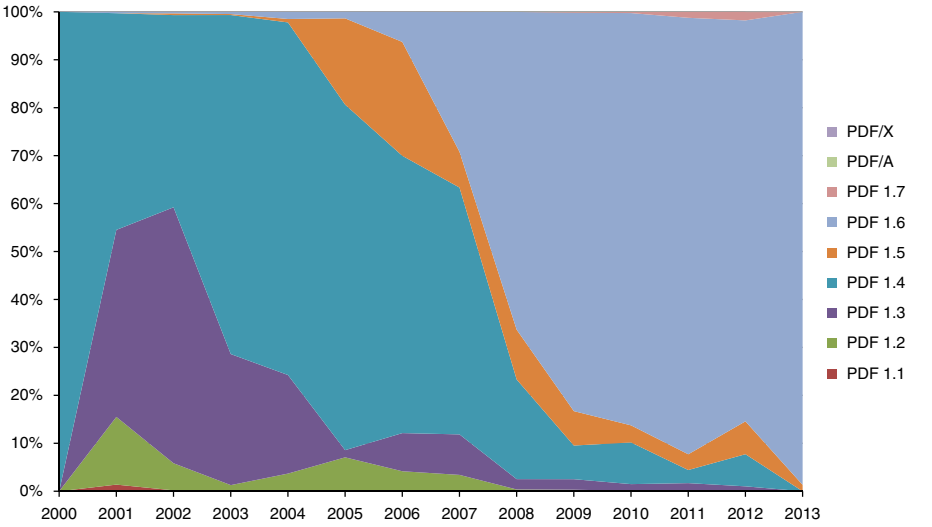

Figure 1.
Evolution of the version of PDF files ingested at TDX (2000-2013)

Format	Extension	Files	
Acrobat PDF	pdf	24,471	28.27%
Portable Network Graphics	png	12,346	14.26%
Graphics Interchange Format	gif	9,082	10.49%
Java Archive Format	jar, java, jsp, js	6,569	7.59%
Hypertext Markup Language	htm, html, xhtml	5,568	6.43%
GZIP Format	rda	4,958	5.73%
Windows Metafile Image	wmf	4,888	5.65%
Python Script	py	4,188	4.84%
Extensible Markup Language	xml, xsd	2,654	3.07%
JPEG File Interchange Format	jpg, jpeg	2,459	2.84%
Plain Text File	txt	1,098	1.27%
Microsoft Powerpoint Design Template	ppt, pptx, pot	1,080	1.25%
Others	others	7,201	8.32%
Total		86,562	100.00%

Table III. Distribution of ingested formats at DDUB

computer programs. This variety can be appreciated better if the formats are grouped into categories, as in Table IV. Among them we detected a lot of files with a wrong extension or temporary files that should not have been published.

Even with this varied format, PDF documents were still very popular, accounting for 28.27 percent of the files. In Table V we show the evolution of PDF versions ingested over time in DDUB; two files were not considered because their submission date was wrong. Again, we can observe a high use of versions 1.4 and 1.6, survival of old versions over time and the almost null adoption of version 1.7 and PDF/A. There is therefore a contradiction between professional consensus recommending the use of PDF/A as the preferred preservation format and its actual use by authors (Zhao, 2011; Oettler, 2013).

If we display these data graphically, as in Figure 2, we can see that the evolution of versions is not as regular as it might seem from the numbers. There was an abnormally high use of version 1.6 in 2008 and of version 1.5 in 2010. These abnormalities were caused, as CRAI staff confirmed to us, by the submission of files resulting from a massive digitalization of documents that was outsourced by the library. Taking out the atypical values of 2008 and 2010, as shown in Figure 2, the evolution of versions is more regular, with new versions superseding old ones (Figure 3).

Category	Occurrences	
Graphic	28,977	33.48%
Text	26,165	30.23%
Software code	14,154	16.35%
Web page	5,568	6.43%
Compressed file	5,092	5.88%
Slides	1,086	1.25%
Audio	187	0.22%
Video	115	0.13%
Miscellaneous	1,001	1.16%
Not identified	4,217	4.87%
Total	86,562	100.00%

Table IV. Distribution by category of ingested formats at DDUB

The use of encryption strategies in PDF files is shown in Table VI for DDUB and in Table VII for TDX.

In DDUB 1,651 files (41.28 percent) were encrypted so as to allow no modifications, and 27 of them even had additional restrictions, such as not allowing screen readers (a technical aid used by blind persons) to access the text.

In TDX 4,307 files (71.95 percent) were encrypted so as to allow no modifications, and 633 of them even had additional restrictions, such as not allowing screen readers to access the text.

In order to check whether such restrictions could affect the later migration to PDF/A format, we conducted a series of tests with a small sample of these files using the PDF/A Manager software by PDFTron (www.pdftron.com/pdfamanager/index.html) and Adobe Acrobat Pro X. Although there were minor differences in errors and migration rates between the two applications, it was found that a large percentage of protected PDF documents could be migrated.

Table V.
Evolution of the version of PDF files ingested at DDUB (2006-2013)

Year	PDF 1.0	PDF 1.1	PDF 1.2	PDF 1.3	PDF 1.4	PDF 1.5	PDF 1.6	PDF 1.7	PDF/A	Annual total
2006			48	46	105	19	2			220
2007			4	36	45	15	11			111
2008			10	100	84	20	409			623
2009		88	94	350	325	33	170			1,060
2010		108	209	371	544	1,704	423			3,359
2011	1	62	134	238	660	166	484	3		1,748
2012	1	55	148	594	1,251	583	1,180	37	2	3,851
2013	30	6	195	525	3,936	1,285	7,439	80	1	13,497
Total	32	319	842	2,260	6,950	3,825	10,118	120	3	24,469
%	0.13	1.30	3.44	9.24	28.40	15.63	41.35	0.49	0.01	100.00

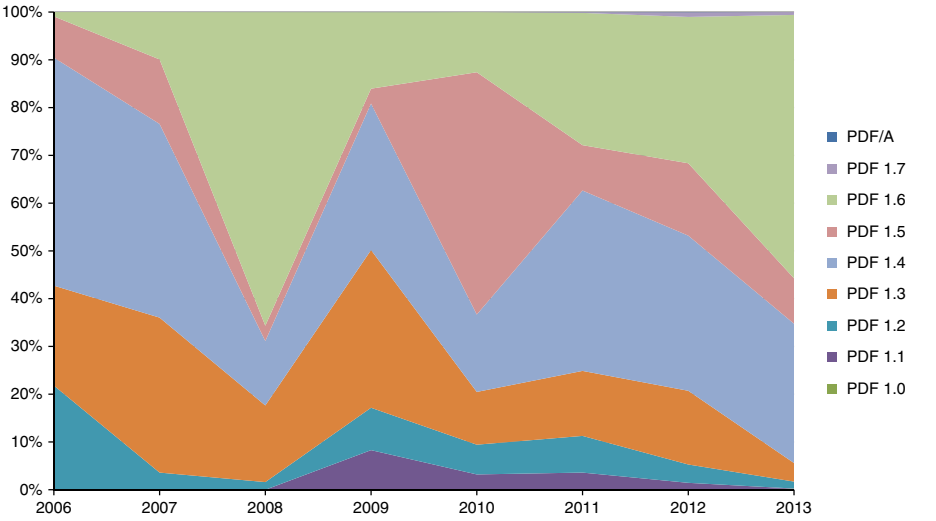

Figure 2.
Evolution of the version of PDF files ingested at DDUB (2006-2013)

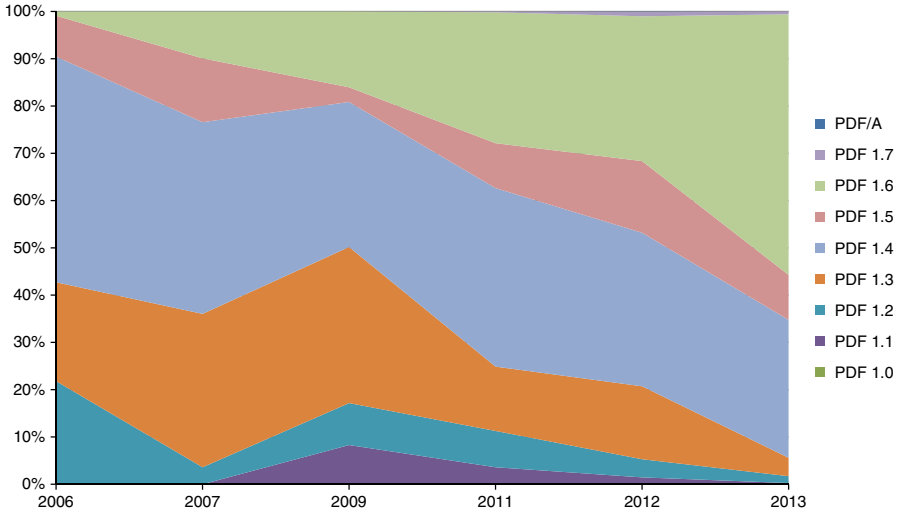

Figure 3. Evolution of the version of PDF files ingested at DDUB (2006-2013), without the abnormal years 2008 and 2010

Encryption level	Quantity		%
Unencrypted	2,349		58.72
Encrypted	1,651		41.28
Printing and changes not allowed		1,624	
Other types of protection/encryption, not identified		27	
Total sample	4,000		100.00

Table VI. Encryption level in a sample of PDF files ingested at DDUB (2006-2013)

Encryption level	Quantity		%
Unencrypted	1,679		28.05
Encrypted	4,307		71.95
Printing and changes not allowed		3,674	
Other types of protection/encryption, not identified		633	
Total sample	5,986		100.00

Table VII. Encryption level in a sample of PDF files ingested at TDX (2000-2013)

4. Discussion

The results show an evolution in the format versions, with more recent versions superseding old ones, though in DDUB massive submissions of files resulting from retrospective conversion of documents distorted the evolution of versions.

One implicit question in this trend is whether the format evolution observed in the analyzed files follows the pace of market availability. It is easy to know the moment when the company distributed new versions (Table VIII), but not so easy to know when new versions were effectively available in the market and integrated in authoring tools. Nevertheless, a gap of about four years is observed between the date of format distribution and the date of format adoption in the analyzed files. Furthermore, the pace

of adoption has been slowing down over time, creating an even larger gap between the available versions and those that are used.

There may be several reasons for this mismatch between the availability of new versions and their actual use. In regular submissions, the failure to use updated versions of the format could be explained by the kind of tools used by submitting authors. These tools generated old version, or even non-standard PDF documents (De Vorsey and McKinney, 2010). Universities are especially slow in updating authoring tools because they distribute software to their employees and often make no changes until the equipment is renewed, every four or five years. Moreover, as some authors have already warned (McLellan, 2010), there are pitifully few free software tools that create PDF/A files and even fewer that create correct PDF/A files. A final explanation for the use of old software could be the lack of incentives for authors to use more recent versions as they get no benefit from updating the tools, or at least do not perceive one. This line of reasoning could also explain the minimal use of the PDF/A format.

In DDUB it is surprising that the latest version of PDF was not used in the massive upload of files generated by a third-party company. It seems that the contractor did not establish the obligation of using the latest version of the format to generate files in the technical requirements, owing to negligence or ignorance. For its part, the outsourced company used a software that created files in an old version of the format, also owing to negligence or just to avoid new investments.

While the research on formats by the JISC UK Web Domain Data set (Jackson, 2012) revealed a perfect substitution of old HTML versions for the newest ones, in DDUB and TDX the changes are less perfect. Comparing the above study with ours, one cannot help noticing the different profiles of the authors and the tools available to them. In the JISC UK Web Domain Data set a great number of HTML pages belong to web sites professionally managed by web masters with technical knowledge, who are committed to exploiting the inherent benefits of new versions of standards and, out of necessity, regularly update their tools. By contrast, in DDUB and TDX the authors tend to have little technical knowledge and no incentives to improve their tools, so they often use outdated or inappropriate ones. These characteristics are not conducive to a good preservation of repository contents.

On the other hand, our results on PDF format are consistent with those obtained for PhD dissertations in Sweden by Fischer and Lundell (2013), who found a large percentage of PDF/A files failed to meet the standards for this format. This percentage had grown over the last few years because the authors now submit the files directly,

Table VIII.
Company publication of the different versions of PDF format

Version	Year
PDF 1.0	1993
PDF 1.1	1996
PDF 1.2	1996
PDF 1.3	2000
PDF 1.4	2001
PDF 1.5	2003
PDF 1.6	2004
PDF/A	2005
PDF 1.7	2006

Source: Font: http://en.wikipedia.org/wiki/Portable_Document_Format

whereas they had previously been generated and submitted by library staff. Fischer and Lunden conclude that "the adoption of standards for electronic documents is marginal" and that "there is a significant risk that, over time, a large proportion of Swedish doctoral dissertations will become inaccessible." Morrissey (2012) also warned about some technical inconsistencies in the PDF family of standards, and especially in PDF/A, which could affect the correct display of contents in the future.

In reference to encryption, the results are extremely high: 41.3 percent of the sample in DDUB and 72 percent in TDX had some kind of protection. Theoretically, this protection could prevent a future use or migration of the files to another format, if the holding institution would consider it appropriate. In fact, there are many programs that allow these protections to be hacked, but this would be a violation of author use conditions and a break of copyright laws in most countries, and would add technical difficulty. In the case of a repository, it is assumed that authors have explicitly authorized holders to migrate contents in order to grant their preservation. Though our small conversion test seems to demonstrate that encryption has no real impact on migration, other thorough and statistically valid tests should be done to validate this hypothesis, as there is a great deficit of reliable tests on preservation systems (Seadle, 2011; Koo and Chou, 2013). In any case, the uncertain feasibility of future migrations and their unknown effects on document rendering, raise serious concerns on the viability of migration as a suitable technique for long-term preservation.

5. Conclusions

We noted at the beginning of this paper that different repositories display very different contents and organization structures, but also share objectives and workflows and even have similar software architectures. We therefore believe that our research results on two specific repositories could be useful to the whole service community, allowing each repository owner to apply the lessons learned to some specific aspect of their interest.

The vast majority of repositories pursue two aims: to promote the dissemination of their contents and to guarantee their preservation. Though these aims are closely linked and are indissoluble from a programmatic point of view, at a practical level their coexistence generates technical and management conflicts. To foster dissemination, it is recommended to favor the submission of a great deal of documents by a great deal of authors, without overwhelming them with technical requirements. To foster good management of a repository and avoid peaks of workload, it is also recommended that documents be uploaded directly by authors rather than in massive submissions by technical staff (Carr and Brody, 2007). However, to foster the preservation of these objects, great control of files and metadata is needed, involving a major workload for technical staff in control tasks or stricter author submission requirements.

The literature on the subject considers the submission system to be one of the main barriers to greater repository adoption by researchers because it is too complex and time-consuming and seems to require some technical knowledge (Kim, 2010, 2011; Covey, 2011). Indulgence in object ingestion is clearly promoting the ingestion of more documents in repositories, but it is leading to the storage of incorrect files and ones that will never be used. These files also need storage space and maintaining them over time involves an increase in operating costs. However, establishing greater control of file formats at the submission stage does not seem to be a recommended policy, and other types of solution need to be explored.

In institutional repositories such problems are even greater because their mission requires them to accept almost any type of document. The technical quality of their

contents will largely depend on whether a policy of digital documents is in place in the hosting institution. We found the contents of DDUB to be extremely heterogeneous, with a low use of recent versions of PDF, an even lower user of PDF/A, the presence of PDF files with reading restrictions, and a wide variety of formats of minority use.

The results seem to suggest that repository managers (and perhaps the training given to staff) focus on metadata more than on the technical features of files. Dublin Core metadata are fully introduced by library staff in TDX and validated by them in DDUB, but bitstreams are accepted and ingested without any subsequent control. Therefore, efficient management of repositories should include revision and correction of file formats in order to guarantee their preservation.

One remaining question is whether institutions will be able to allow users to render such a large number of formats in the future. Some institutions could argue that the repository's mission ends with giving access to contents, and that their rendering is a user problem, but it is difficult to justify the organizational and economic effort invested in preserving files that are likely to involve problems of use in the future. Normalizing and unifying file formats will result in better user support and could become a management priority in large repositories. On another hand it seems evident that once the files are ingested in the repository, it is costly and even dangerous to apply a systematic policy of unifying formats due to the difficulties caused by encryption issues and the potential errors introduced by transformations. Finally, we shall not forget that a format change could involve a loss of authenticity if it is not applied with the proper legal and technical support.

References

Anglada, L., Comellas, N. and Ros, R. (2002b), "Sharing solutions: gathering catalan academic and scholarly publications online", *8th International Conference of the European University Information Systems (EUNIS 2002)*, *Porto*, available at: http://www.csuc.cat/sites/default/files/docs/eunis114.pdf (accessed 21 May 2015).

Anglada, L., Bárcena, I., Cambras, J., Comellas, N., Huguet, M. and Ros, R. (2002a), "Acceso electrónico a las tesis doctorales de Cataluña", *El Profesional de la información*, Vol. 11 No. 1, pp. 28-34.

Brody, T., Carr, L., Hey, J.M.N., Brown, A. and Hitchcock, S. (2007), "PRONOM-ROAR: adding format profiles to a repository registry to inform preservation services", *The International Journal of Digital Curation*, Vol. 2 No. 2, available at: http://ijdc.net/index.php/ijdc/article/view/53/25 (accessed 21 May 2015).

Brown, A. (2005), "Automating preservation: new developments in the PRONOM service", *RLG DigiNews*, Vol. 9 No. 2, available at: http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file959.html#article1

Carr, L. and Brody, T. (2007), "Size isn't everything. Sustainable repositories as evidenced by sustainable deposit profiles", *D-Lib Magazine*, Vol. 13 Nos 7-8, available at: www.dlib.org/dlib/july07/carr/07carr.html (accessed 21 May 2015).

Covey, D.T. (2011), "Recruiting content for the institutional repository: the barriers exceed the benefits", *Journal of Digital Information*, Vol. 12 No. 3, pp. 1-18.

De Vorsey, K. and McKinney, P. (2010), "Digital preservation in capable hands: taking control of risk assessment at the national library of New Zealand", *Information Standards Quarterly*, Vol. 22 No. 2, pp. 41-44.

Fischer, T. and Lundell, B. (2013), "Swedish dissertations: archived for the future?", *Proceedings of the 17th International Academic MindTrek Conference: Making Sense of Converging Media*, *ACM*, *New York, NY*, pp. 176-179.

Graf, R. and Gordea, S. (2013), "A risk analysis of file formats for preservation planning", *iPres 2013* Lisboa, available at: http://purl.pt/24107/1/iPres2013_PDF/A%20Risk%20Analysis%20of%20File%20Formats%20for%20Preservation%20Planning.pdf (accessed 21 May 2015).

Hitchcock, S., Brody, T., Hey, J.M.N. and Carr, L. (2007), "Digital preservation service provider models for institutional repositories: towards distributed services", *D-Lib Magazine*, Vol. 13 Nos 5-6, available at: http://dlib.org/dlib/may07/hitchcock/05hitchcock.html (accessed 21 May 2015).

Jackson, A. (2012), "Formats over time: exploring UK web history", *iPRES 2012, Proceedings of the 9th International Conference on Preservation of Digital Objects*, *Toronto*, pp. 103-106.

Kennan, M.A. and Wilson, C. (2006), "Institutional repositories: review and an information systems perspective", *Library Management*, Vol. 27 Nos 4/5, pp. 236-248.

Kim, J. (2010), "Faculty self-archiving: motivations and barriers", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 9, pp. 1909-1922.

Kim, J. (2011), "Motivations of faculty self-archiving in institutional repositories", *Journal of Academic Librarianship*, Vol. 37 No. 3, pp. 246-254.

Koo, J. and Chou, C.C.H. (2013), "PDF to PDF/A: evaluation of converter software for implementation in digital repository workflow", *New Review of Information Networking*, Vol. 18 No. 1, pp. 1-15.

Lawrence, G.W., Kehoe, W.R., Rieger, O.Y., Walters, W.H. and Kenney, A.R. (2000), *Risk Management of Digital Information: a File Format Investigation*, Council on Library and Information Resources, Washington, DC, available at: http://www.clir.org/pubs/reports/pub93/pub93.pdf (accessed 21 May 2015).

McLellan, E. (2010), "Selecting preservation file formats", *Information Standards Quarterly*, Vol. 22 No. 2, pp. 30-33.

Morrissey, S.M. (2012), "The network is the format: PDF and the long-term use of digital content", *Archiving 2012 – Preservation Strategies and Imaging Technologies for Cultural Heritage Institutions and Memory Organizations, Final Program and Proceedings*, Society for Imaging Science and Technology, Springfield, VA, pp. 200-203.

Oettler, A. (2013), *PDF/A in a Nutshell 2.0. PDF for Long-Term Archiving*, Association for Digital Document Standards, Berlin, available at: www.pdfa.org/wp-content/uploads/2013/05/PDFA_in_a_Nutshell_211.pdf (accessed 21 May 2015).

Pearson, D. and Webb, C. (2008), "Defining file format obsolescence: a risky journey", *International Journal of Digital Curation*, Vol. 3 No. 1, available at: http://ijdc.net/index.php/ijdc/article/view/76/44 (accessed 21 May 2015).

Rimkus, K., Padilla, T., Popp, T. and Martin, G. (2014), "Digital preservation file format policies of ARL member libraries: an analysis", *D-Lib Magazine*, Vol. 20 Nos 3/4, available at: www.dlib.org/dlib/march14/rimkus/03rimkus.html (accessed 21 May 2015).

Rosenthal, D.S.H. (2010), "Format obsolescence: assessing the threat and the defenses", *Library Hi Tech*, Vol. 28 No. 2, pp. 195-210.

Rosenthal, D.S.H. (2013), "In-browser emulation", *DSHR's Blog*, November 26, available at: http://blog.dshr.org/2013/11/in-browser-emulation.html (accessed 21 May 2015).

Rusbridge, C. (2006), "Excuse me […] some digital preservation fallacies?", *Ariadne*, No. 46, available at: www.ariadne.ac.uk/issue46/rusbridge/ (accessed 21 May 2015).

Seadle, M. (2011), "Archiving in the networked world: metrics for testing", *Library Hi Tech*, Vol. 29 No. 3, pp. 557-564.

Tarrant, D., Hitchcock, S. and Carr, L. (2011), "Where the semantic web and web 2.0 meet format risk management: P2 registry", *International Journal of Digital Curation*, Vol. 6 No. 1, pp. 165-182, available at: www.ijdc.net/index.php/ijdc/article/view/171/239 (accessed 21 May 2015).

van Westrienen, G. and Lynch, C.A. (2005), "Academic institutional repositories. Deployment status in 13 nations as of mid 2005", *D-Lib Magazine*, Vol. 11 No. 9, available at: www.dlib.org/dlib/september05/westrienen/09westrienen.html (accessed 21 May 2015).

Vermaaten, S., Lavoie, B. and Caplan, P. (2012), "Identifying threats to successful digital preservation: the SPOT model for risk assessment", *D-Lib Magazine*, Vol. 18 Nos 9/10, available at: www.dlib.org/dlib/september12/vermaaten/09vermaaten.html (accessed 21 May 2015).

Ware, M. (2004), "Institutional repositories and scholarly publishing", *Learned Publishing*, Vol. 17 No. 2, pp. 115-124.

Zhao, F. (2011), "On choosing the digital document's file format for long-term preservation", *IEEE 3rd International Conference on Communication Software and Networks, Xi'an*, IEEE, pp. 370-372.

Further reading

University of Barcelona (2011), "The University of Barcelona's open acces policy", Approved by The Governing Council on 7th June, available at: http://diposit.ub.edu/dspace/bitstream/2445/27711/1/2011_06_UB_OA_Policy.pdf (accessed 5 October 2014).

About the authors

Dr Miquel Termens is a Professor of the Department of Library and Information Science, at the University of Barcelona (Spain). He is an expert in digital library, digitization project management and digital preservation. He is currently the Scientific Director of the Centre for Digitization and Digital Preservation System at the University of Barcelona. Dr Miquel Termens is the corresponding author and can be contacted at: termens@ub.edu

Dr Mireia Ribera is a Professor of the Department of Library and Information Science, at the University of Barcelona (Spain). She is an expert in digital accessibility for disabled people, especially on access to scientific literature.

Anita Locher holds a Master's Degree in Digital Content Management from the University of Barcelona (Spain). She is currently preparing her PhD thesis about digital preservation issues in geographical data or geodata.