THE NATIONAL ARCHIVES OF FINLAND

Mass Digitization Project

Leverage from the EU 2014–2020

European Union
European Regional Development Fund

Survey of Digitization in Archives

Lauri Hirvonen

The National Archives of Finland

19 December 2017

National Archives of Finland Rauhankatu 17 PB 258, FI-00171 HELSINKI Finland Tel. + 358 29 533 7000 Fax + 358 9 176 302 kirjaamo@arkisto.fi http://www.arkisto.fi

Table of Contents

Table of Contents 1
1 Introduction 2
1.1 Purpose of the Survey 2
1.2 Finnish Mass Digitization Project 2
2 Data 4
2.1 Data collection 4
2.2 Description of Data 4
3 Results of the Survey 7
3.1 Scale of Digitization 7
3.2 Purpose of Digitization 10
3.3 Technical Specifications for Digitization 13
3.3.1 File formats and Scanning Specifications 13
3.3.2 OCR 16
3.3.3 Labour Costs of Digitization 18
4 Summary 23
5 Appendices 25
5.1 Appendix 1 – Questionnaire and answers 25
5.2 Appendix 2- Graphs 30
5.3 Appendix 3 – Tables 38

1 Introduction

1.1 Purpose of the Survey

The aim of this survey is to provide an overview of current status of digitization mainly in archival sector. The study seeks to assess the amount of archival material digitized by different archives, scope of yearly digitization and preferred scanning specifications and file formats. The survey was conduct as a part of mass digitization planning project of the National Archives of Finland. Thus it sought to find whether any actor in the archival sector was engaged or planning a very large scale digitization of materials with an aim to dispose original records after digitization. It became apparent quite quickly that there were no direct comparisons to the planned Finnish mass digitization project. Private sector was left out of the scope of the survey, even though a number of businesses engage in large scale digitization of bills and so forth. The type of materials digitized in the private sector often differ from those held at the archival institutions and demand different kind of digitization process compared to public records. Nevertheless the National Archives of Finland is delighted that a number businesses answered the survey even though it was not specifically advertised to them.

1.2 Finnish Mass Digitization Project

One of the driving reasons for conducting the survey was the need for international comparison in digitization at the National Archives of Finland. Finnish government made a resolution to digitize records currently at the hands of various government agencies by 2030 in July 2017. A project tasked

to plan the digitization of government records was founded in the summer of 2017 with funding from European Regional Development fund. During the summer and autumn the project surveyed the number of analogue records currently at the archives of individual agencies. They have in total circa 380 linear kilometres of records, of which around 170 linear kilometres will be stored permanently and transferred to custody of the National Archives of Finland. This amounts to roughly to the amount of records currently at the National Archives. Rest of the material are shorter term operative records, which will not be transferred to custody of the National Archives of Finland. Due to the sheer amount of records, it was decided that international comparisons would be needed for planning of the coming digitization. Towards this end the National Archives of Finland decided to conduct an online survey aimed at archives to get comparison data.

2 Data

2.1 Data collection

The survey was conducted online during October 2017. Responses were gathered through direct invitations to digitization personnel at archives, and via advertisements to ICARUS (International Centre for Archival Research) mailing list and on various social media. The original intention of the National Archives of Finland was to conduct the survey in two phases. The first phase was the general survey of digitization included in this report, and the second was intended to be a more in-detail questionnaire on practicalities of mass digitization to few institutions selected from the respondents of the first phase. The second phase of the questionnaire was deemed to be too complex and time consuming to respond to and was cancelled.

2.2 Description of Data

The survey got forty responses in total and it was viewed 883 times. The twenty of the responses were made by national archives. Eighteen of them were from countries in the European Union, but US National Archives and Records Administration and the National Archives of Australia also responded to the survey. Ten of the participants were from other archives ranging from state archives to city archives and specialised archives. In addition, four university libraries answered the questionnaire. Five answers were made by companies or private foundations. The answers were mainly European from organizations, but there were also answers from Australia, India, Israel and USA.

The survey contained in thirty one questions, some of which had sub-questions.[1] The questionnaire of the survey was intended to scope out the scale of digitization of different organizations, scanning specifications and whether anyone had engaged on in a digitization approaching the intended scale of the Finnish mass digitization planning project. As the number of responses is rather small and they are from varied organizations, the results should be seen as generalizations of current status of digitization in the archival sector mainly in European context.

The questionnaire had a problem with its wording, which was not in all cases clear in to respondents. This was partly caused by thigh time schedule of the mass digitization project had at the National Archives of Finland. The terminology for mass scale digitization would have needed to be more specific as it quite ambivalent term. It was intended to cover high volume, almost factory like, digitization, which was planned in the National Archives of Finland's mass digitization project. As there were no other projects on the same level, the term could be understood in variety of different ways.

The estimate of digitized items in linear meters was included in the survey as a comparison for physical scale of digitization. It is not the best way to express the number of digitized items, and not all of the responses could provide estimates in linear meters.[2] Depending on the nature of analogue records number of images scanned from one linear meter can vary significantly.[3] The National

[1] See Appendix for the questionnaire.
[2] Also not all archives count the extent of their holdings in relation to linear meters or feet. The US National Archives and Records Administration counts physical storage space in cubic feet, which is not directly transferrable to linear meters.
[3] On average one linear meter of records is roughly equivalent of 8000 images, when scanned from Finnish archival records.

Archives of Finland and Norway have both digitized roughly 5000 linear meters of original material, but Finland has digitized circa 60 million images and Norway circa 48 million images from the same amount of physical material. Because of the difficulty of estimating the scale of digitization in linear meters, it was used in the survey to estimate whether reported number of digitized images contained any typos.

One of the answers reported a significantly higher amount of digitized files compared to rest of the responses. The Portuguese Direção Geral do Livro dos Arquivos e das Bibliotecas stated that they have 1.75 billion digitized images, which represented roughly 5000 linear meters of original material. As the institution is a general directorate for both archives and libraries, it seems likely that the number images contain a large amount digitized images from library collections. This makes it somewhat difficult to compare the Portuguese numbers to those of other nations, which are represented only by their national archives. In a later question Direção Geral do Livro dos Arquivos e das Bibliotecas stated that they are engaged or going to engage in mass digitization circa 5000 linear meters of material or forty million images. This later number is actually quite close to number of images reported for the same amount linear meters by Finnish and Norwegian national archives. The different numbers of digital images and linear meters highlight the fact that the type of digitized materials can significantly affect the amount of digital images produced.

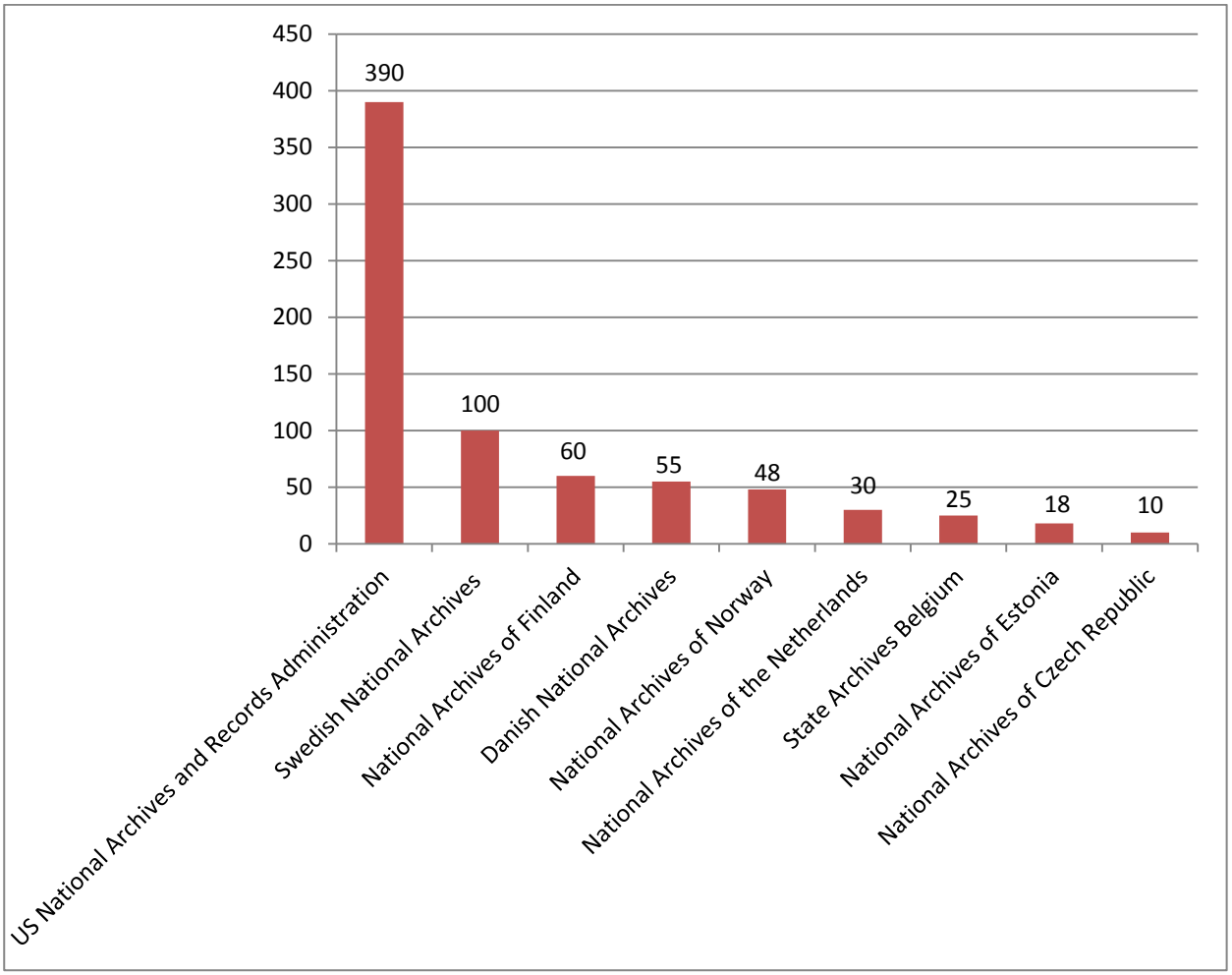
3 Results of the Survey

3.1 Scale of Digitization

Based on the number of digitized items were the largest institutions were: US National Archives and Records Administration, the Swedish National Archives,[4] the National Archives of Finland, the National Archives of Denmark, the National Archives of Norway, the National Archives of the Netherlands, State Archives Belgium, the National Archives of Estonia and Portuguese Direção Geral do Livro dos Arquivos e das Bibliotecas. All of them had over ten million digitized images in their holdings.[5] Appendix 3 contains a table of with all responses.

The numbers for yearly digitization follow the same pattern as the institutions with largest digital holdings. It should be noted that yearly fluctuations can affect the placing of archives if there are, for example, projects that produce large quantities of digitized files. The National Archives of Australia was unable to provide an estimate of the number of their digitized files, but could provide numbers for their yearly digitization, which is the reason they do not appear in the totals of digitization. These same archives are planning large scale production during the coming year.

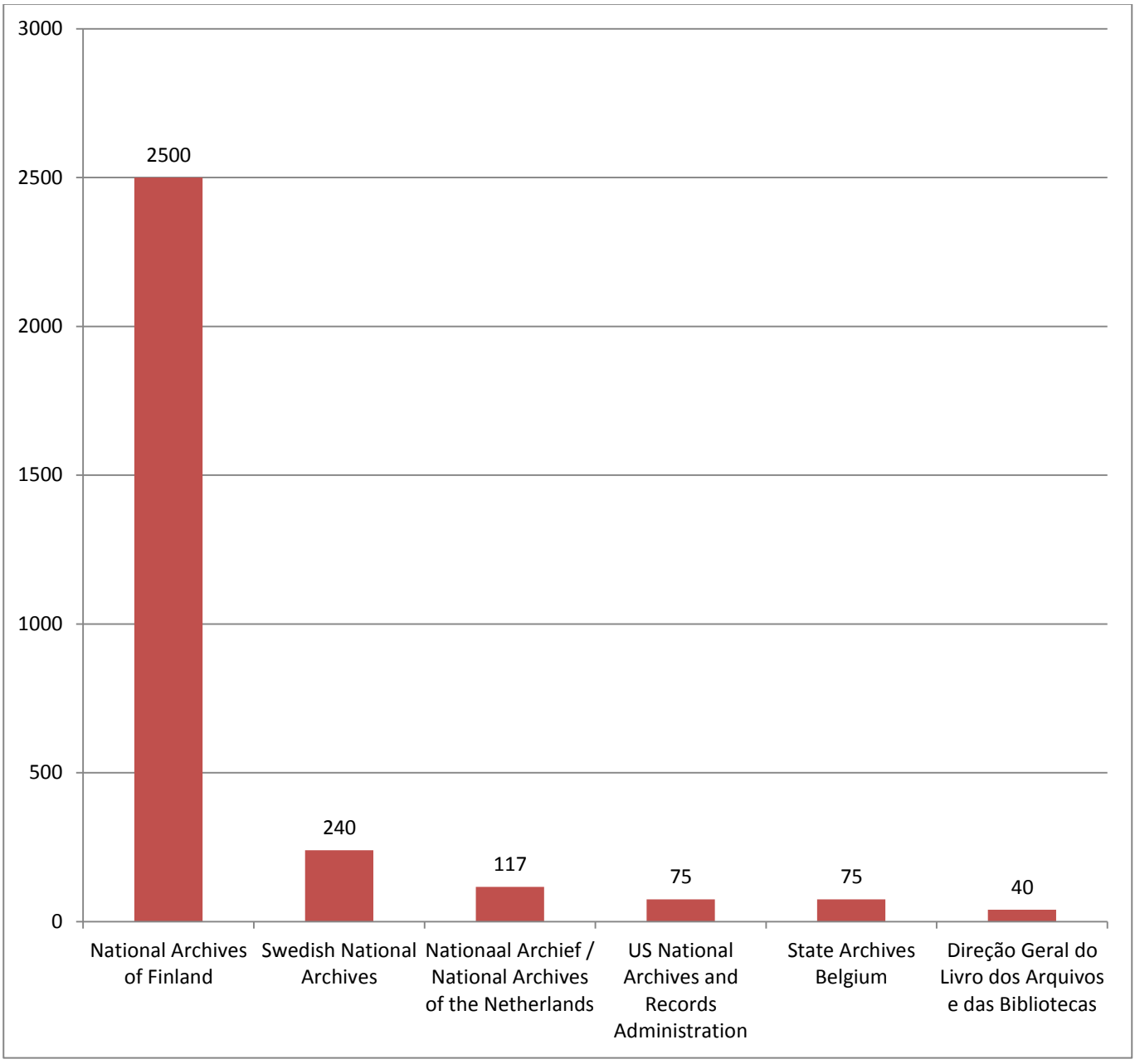
[4] The digitization centre of the Swedish National Archives, Mediakonverteringscentrum (MKC) also digitizes materials for the Swedish Royal Library and other organizations. This digitization is not included in the survey.
[5] Direção Geral do Livro dos Arquivos e das Bibliotecas of Portugal reported the highest number of digitized images at 1.7 billion images or 5000 linear meters. It is not clear what proportion of the number is made of items digitized for libraries such as books and newspapers and how the digitized files are spread to across different institutions. When linear meters are compared the Direção Geral do Livro dos Arquivos e das Bibliotecas has digitized roughly the same amount of physical material as the National Archives of Finland and Norway.



Graph 1 Number of digitized images (million images)

Finland aims to digitize over 200 linear kilometres of modern government records during the next decade. If the digitization project is executed, it will be by far the largest digitization in the archival sector as the number of images to be produced is circa 2.5 billion. There are other significant digitization projects in the future. The Swedish National Archives are going to digitize circa 10 percent of their holdings, while the National Archives of the Netherlands, the National Archives of Norway and State Archives Belgium are each going to digitize large amounts of their most used holdings. Norway did not have an estimate on the amount of images they would digitize, but they estimate that the analogue material to be digitized is 130 linear kilometres in size, which is roughly

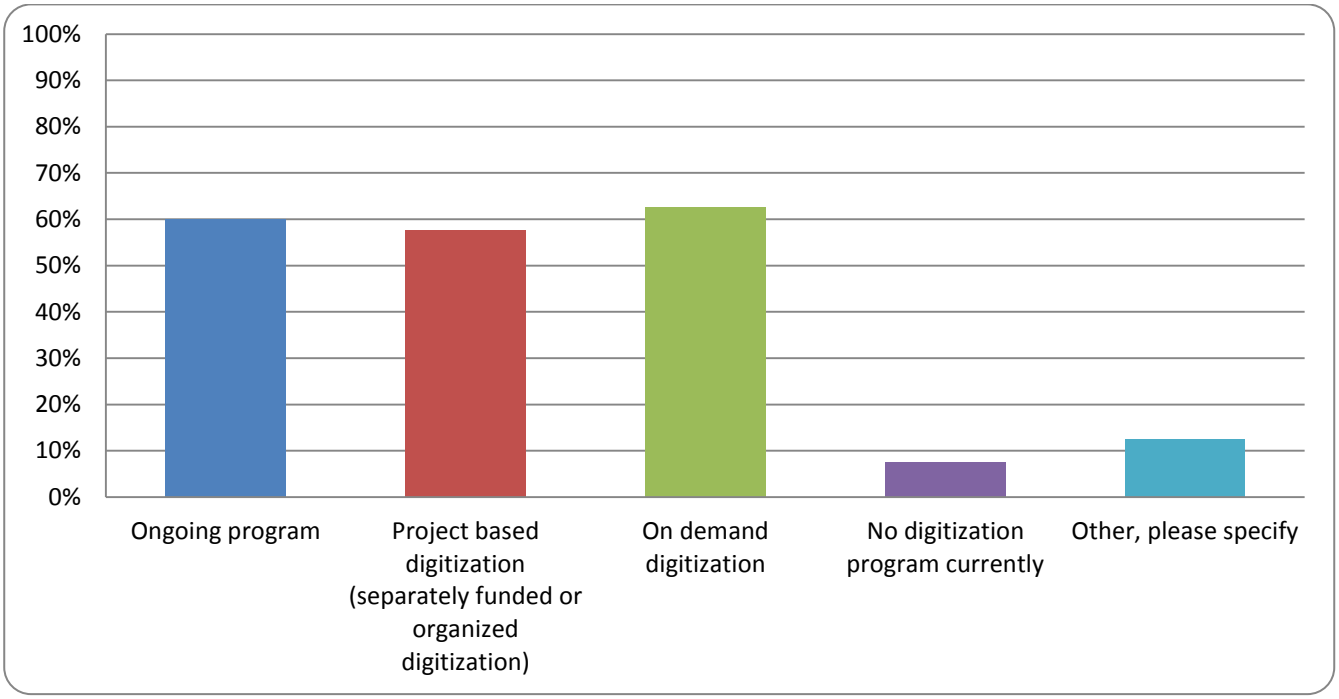
equivalent to figures estimated by the National Archives of the Netherlands and State Archives Belgium.



Graph 2 Planned digitization projects (million images)

3.2 Purpose of Digitization

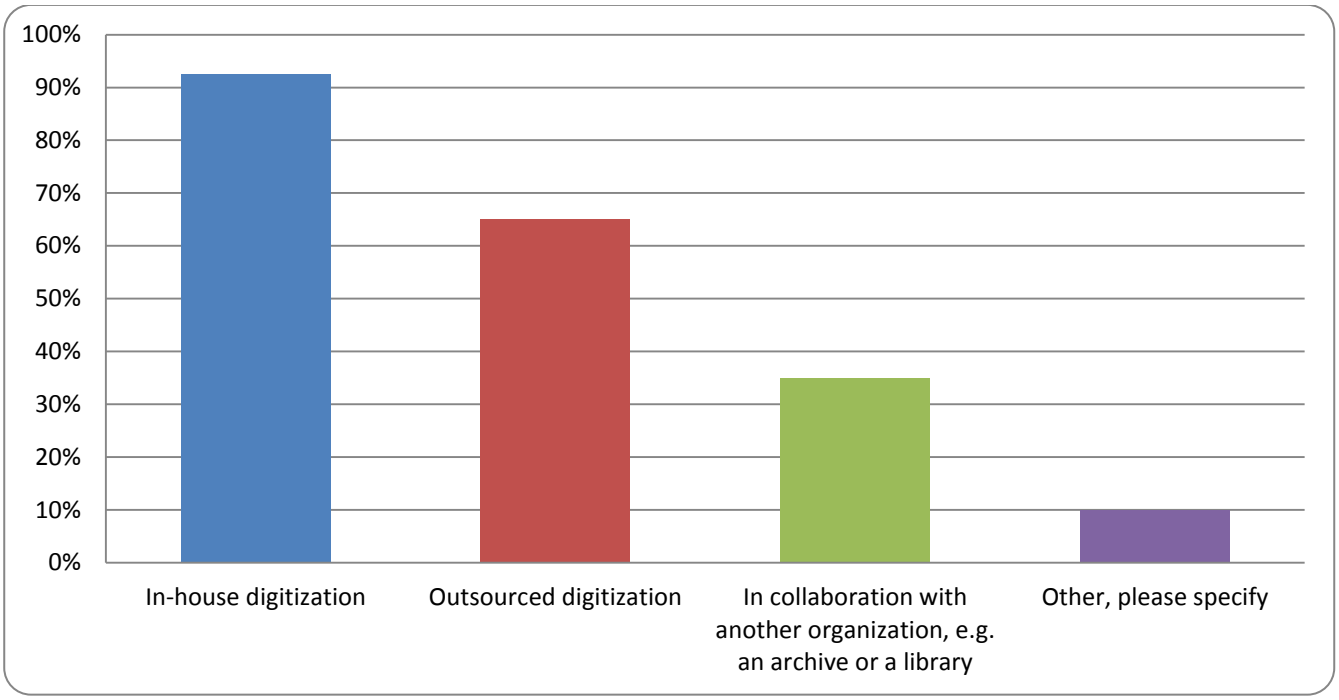
Digitization is organized amongst the participants either as an ongoing digitization program or on project basis. The distribution is exactly the same (60%). One third of the respondents digitized their material through a digitization program and as focused projects. The type of the organization did not affect how they organized digitization. If respondents digitized more than million images yearly, they had organized their digitization on program and project base. Portuguese Direção Geral do Livro dos Arquivos e das Bibliotecas of Portugal is an exemption of this as they did not report project based digitization. Two thirds of respondents (65%) also provided on demand digitization.



Graph 3 Organization of digitization

Nearly all respondents (93%) digitized material in-house, but majority (65%) also had used outsourced digitization. In addition, two fifths (38%) of the respondents had collaborated in digitization with other memory institutions. Unsurprisingly more material a respondent had digitized

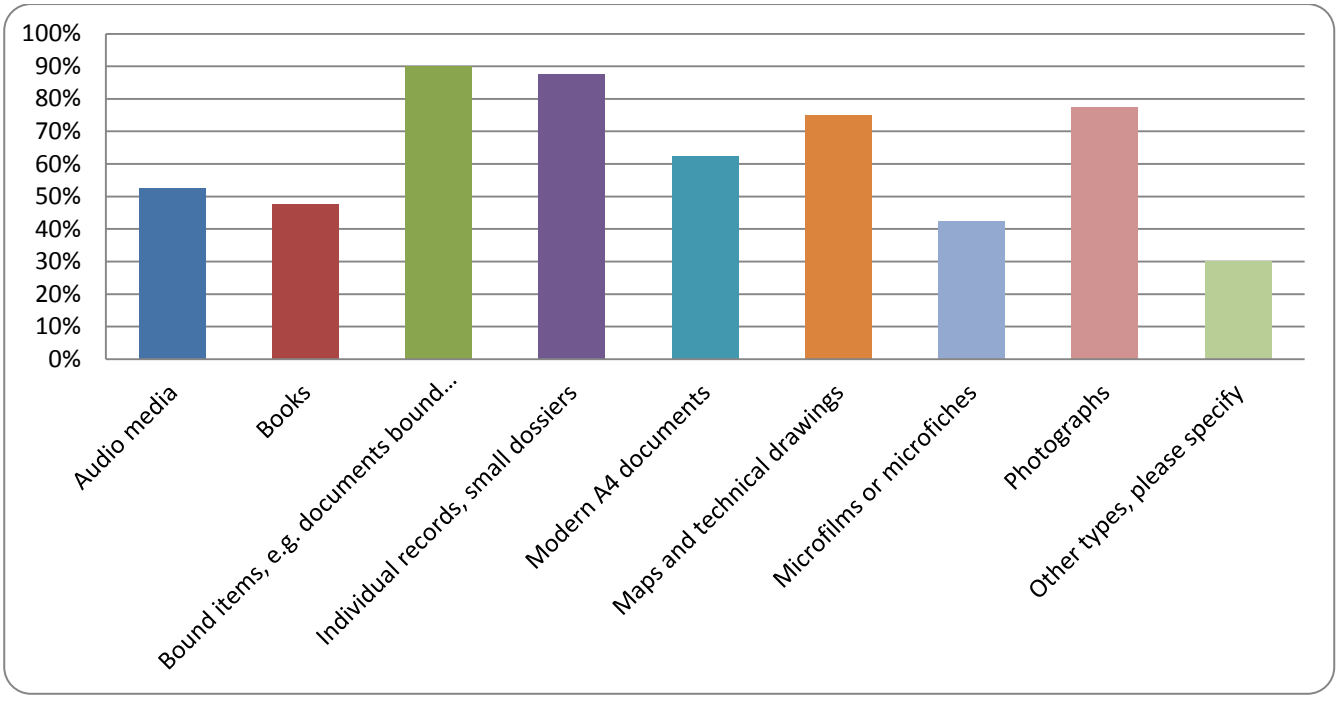
more likely it had used outsourced service providers or collaborated with others. As no further questions were asked of particularities of digitization activities, the exact nature of digitization programs or particularities of on demand digitization is not apparent from the survey.



Graph 4 In-house and outsourced digitization

The most digitized materials are what can be labelled as typical archival material ranging from loose documents to bound records and maps. Most of the respondents have digitized photographs, and nearly half have digitized audio media (53%), books (50%) and microfilms or microfiches (45%). In addition, a few had digitized audio-visual material. When an organization had conducted a large scale digitization project, bound items were the most common type of material digitized. Generally digitization has concentrated on the most used material in archives or older and more fragile items. Much of this material used to be bound, which explains that bound items are most common mass digitized materials. One third of the respondents (34%) had at some point unbound digitized items

in order to help scanning.[6] One respondent stated that they have unbound items to help scanning at times, and stated that the unbound loose original records will be kept afterwards in appropriate archival enclosure afterwards as originals are not needed in reading room after digitization. The question of unbinding stems from possibility of using paper guillotines at the National Archives of Finland to speed up unbinding of relatively modern documents with ample margins without damaging the information of the records. The variety of digitize materials is not surprising as most of the answers were from general archives that do not specialise in on type of material exclusively.

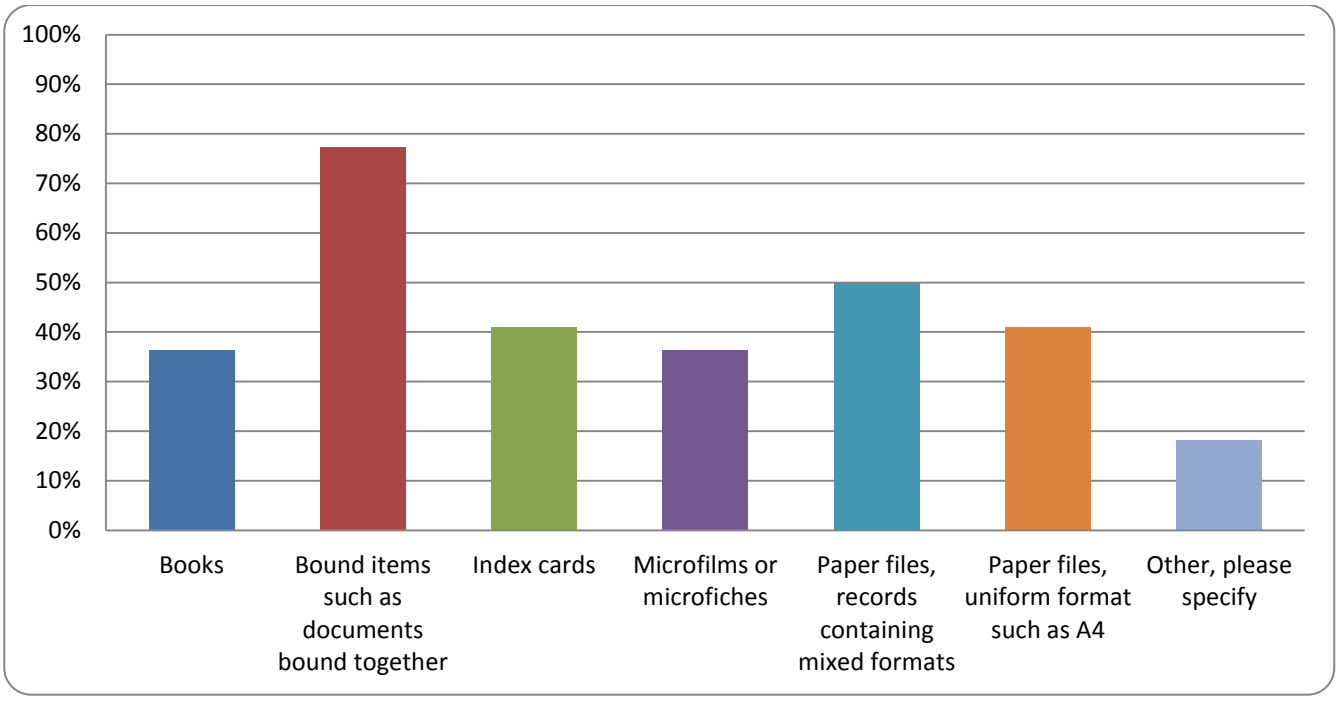


Graph 5 Digitized materials

Four institutions had a policy that allows disposal of analogue originals after digitization. In addition, six institutions were drafting criteria as their national legislation allows destruction of analogue records after digitization. None of the respondents had implemented disposal of records on large

[6] Two respondents did not answer the question.

scale. A number of answers were sceptical of destruction of analogue originals after digitization due to possibility of loss of information, because of quality failures in digitization. Three respondents stated that they would use much stricter quality control in order to guarantee integrity of the digitized records, if they would dispose original records.



Graph 6 Materials digitized in large scale projects

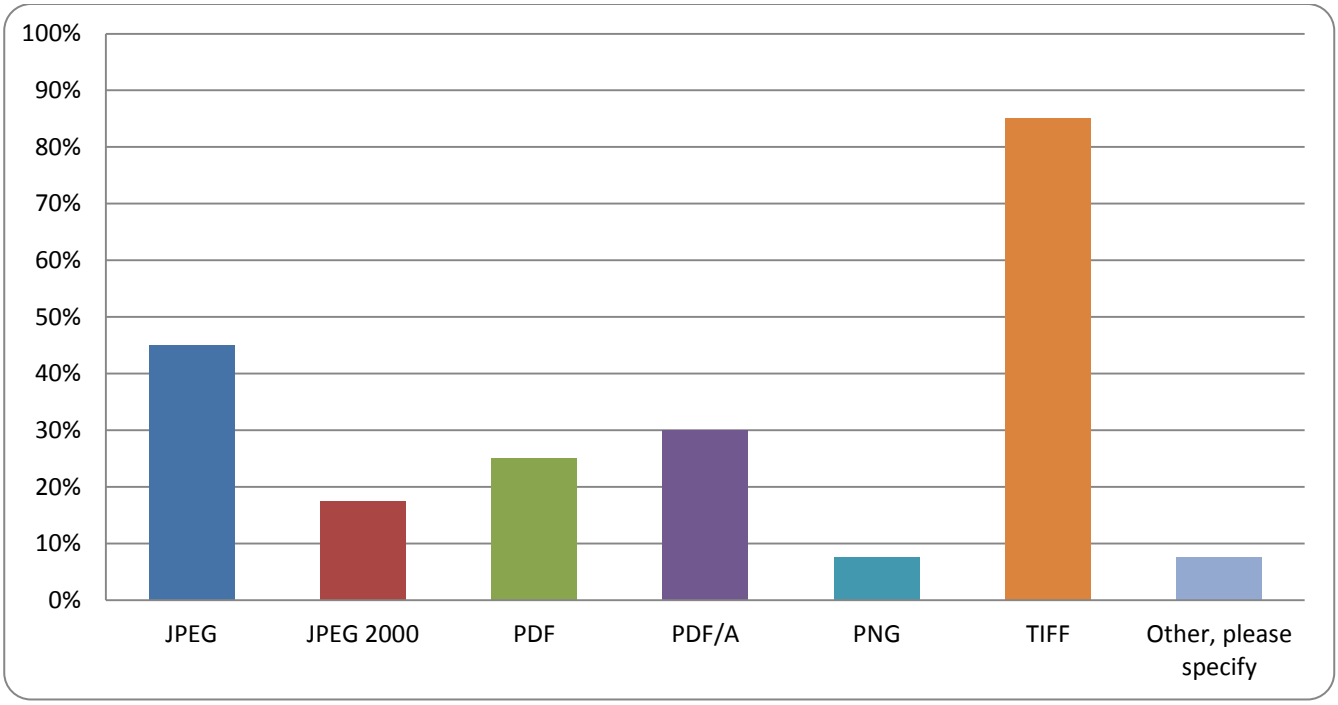
3.3 Technical Specifications for Digitization

3.3.1 File formats and Scanning Specifications

Scanning specifications and master file format, e.g. file used for preservation or as a basis for creation of access files, were homogeneous amongst the respondents with only slight variation.[7] The most used format for master file is TIFF. Some four fifths (85%) of the respondents use it. There is

[7] See Appendix 3 for scanning criteria and preferred file formats.

some variation in format of master file, for example, a number of institutions have also accepted JPEG files as master files, but they account slightly less than half of the respondents (45%). When one looks at the scanning specifications the desired master file is an uncompressed TIFF file. A number of respondents stated that the reason for choosing uncompressed TIFF files was the need to make sure that no information was lost during the digitization process and migrations of the files.

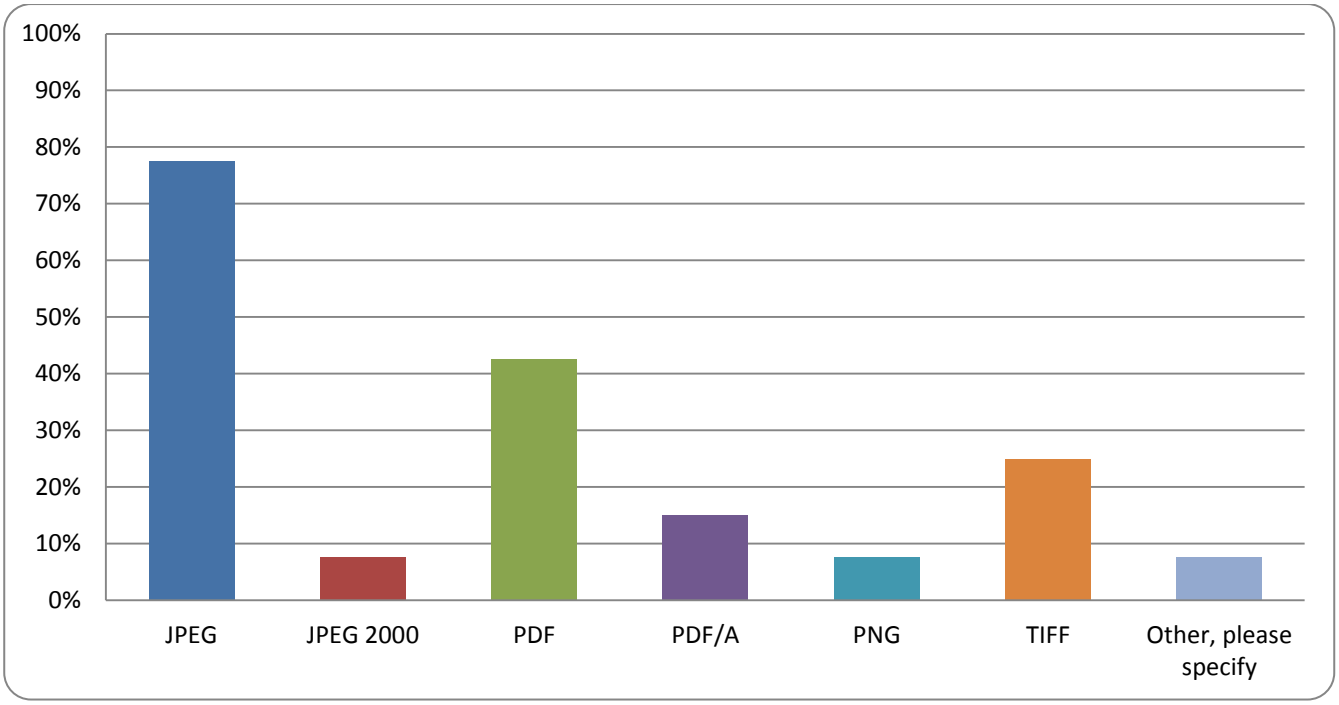


Graph 7 Formats used for master or preservation file

Only one respondent stated that the storage space can be an issue due to the size of TIFF files. The National Archives of Finland stated that they have been considering using LZW compression or JPEG (100%) for digitization of modern A4 documents due to their size. Only a few of the respondents used JPEG2000 with lossless compression as a master format. It is more common among the national libraries, where it has been adopted, at least, for some collections amongst the British Library, the Danish Royal Library, the Library of Congress, the National Library of Finland

and the Swedish Royal Library to name a few. Some respondents reported that they have previously accepted PDF (25%) or PDF/A (30%) files.

JPEG (78%) followed by PDF (45%) were the most common formats used for providing access to digitized images, which is expected as they are ubiquitous file types and supported by all web browsers. Somewhat surprising was that quarter of respondents had also used TIFF files for providing access to the images, as the format needs a dedicated viewing solution for most browsers.



Graph 8 Formats used for access or dissemination file

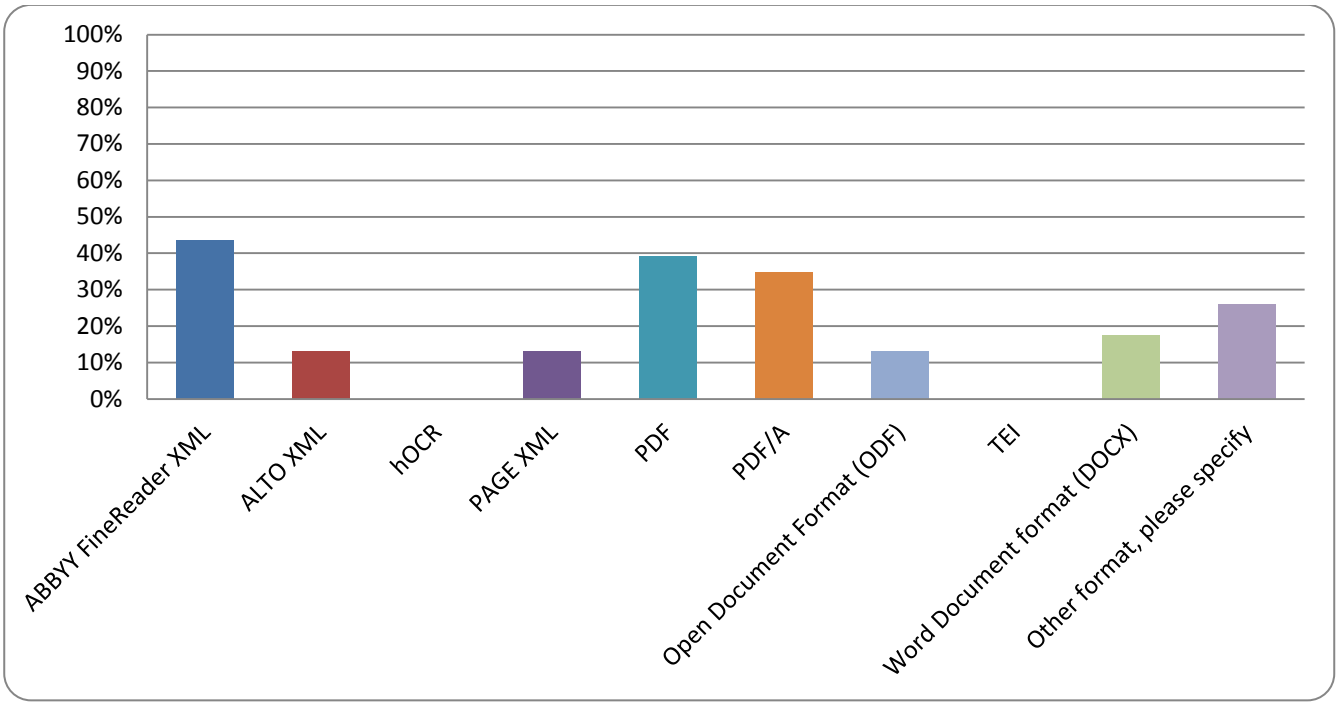
Technical criteria for scanning are quite similar across the respondents. Most used resolution for imagining was 300 dpi. Only three respondents stated that they used a lower resolution: 200 dpi was used by one respondent for digitization of microfilms, and 150 dpi and 100 dpi were by one respondent each. Higher resolutions ranging from 400 to 600 dpi were used for detailed objects

such as maps or when the documents contained very small characters. Photographs were scanned in different resolutions depending on the size of the original. Colour images were preferred formats in the written criteria, and they were scanned as 24 bit RGB images. One respondent stated that they used 8 bit RGB colour profile, while one other stated that 8 bit colour profile could be used in addition to 24 bit RGB. The RGB profile was not usually stated, but Adobe RGB (1998), ECI-RGB version 2, Pro Photo RGB and sRGB were mentioned. 8 bit grayscale was used most often for scanning microfilms. Only two respondents stated they used 16 bit grayscale for some material. In both cases imagining was done at higher resolution than 300dpi. Outsourcing a digitization project has not affected the acceptable scanning criteria among the respondents. Only one respondent stated that they had different criteria for outsourced digitization, which affected mainly granularity of multipage PDFs and metadata. On the other hand one respondent stated that had had to do much more quality assurance than typical for one outsourced project, because of quality issues in digitization.

3.3.2 OCR

Slightly more than half (55%) of the respondents have used OCR techniques for digitized materials. There is no strong correlation between use of OCR and materials respondents have digitized. As proportion different types of materials digitized was not surveyed, it can only speculated that whether digitization of large quantity of books, newspapers or relatively modern documents with easily machine readable typeface could have been one reason for employment of OCR. The absolute quantity of digitized material cannot be seen as a factor for use of OCR as the National Archives of

Finland and the Danish National Archives, both of whom have digitized over fifty million files, have not used OCR.



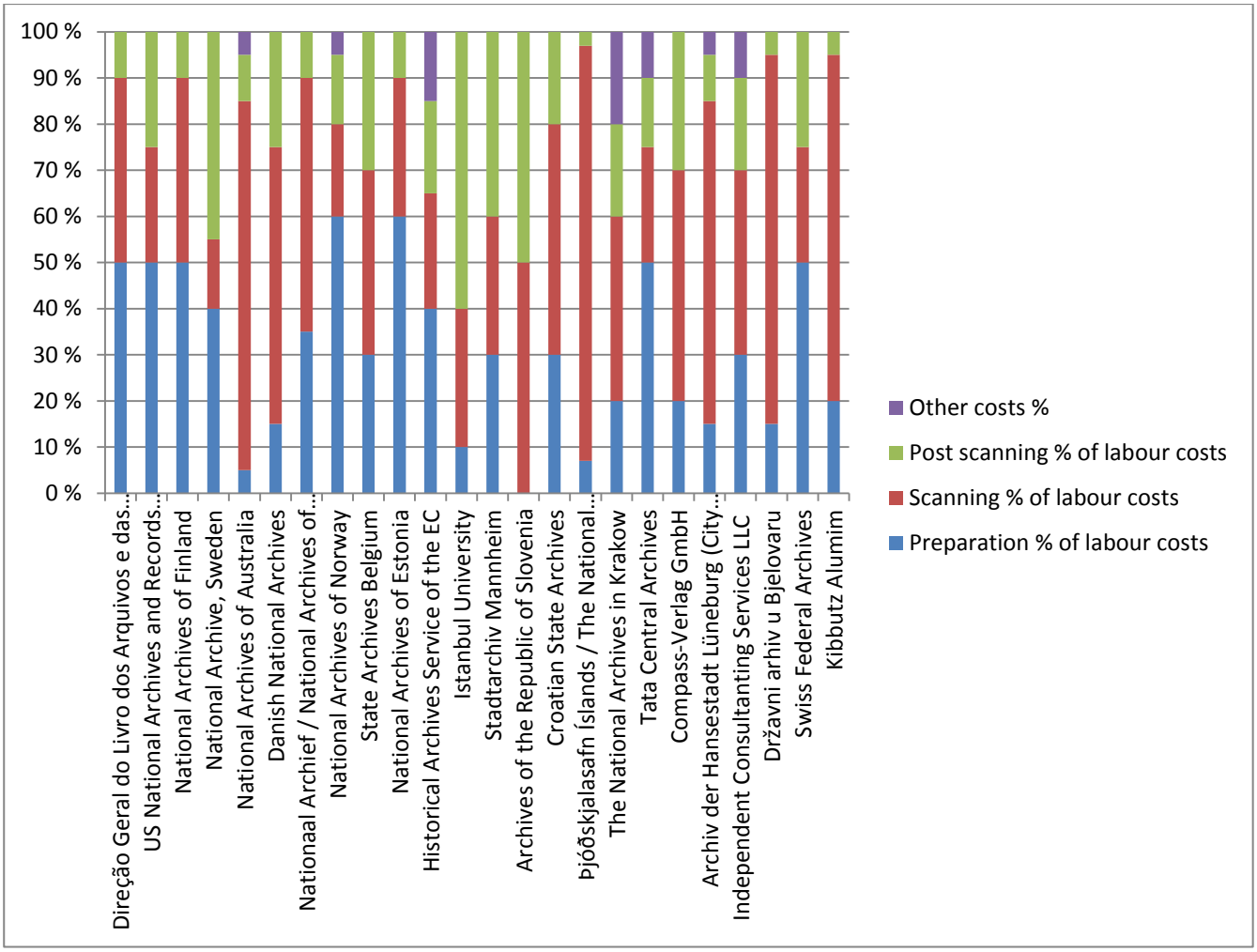
Graph 9 OCR file formats

A variety of formats have been used for saving text captured by OCR. No single format dominates the responses. Abbyy FineReader XML is the most common, but it has been used by less than half of the respondents that have used OCR (44%), which means that one fourth all respondents have used Abbyy FineReader. PDF (38%) PDF/A (35%) are the next most common OCR file formats. There is not much correlation between use of OCR and employment of any single format for saving the files.[8] All in all XML or PDF are the most common formats for saving OCR data.

[8] Pearson correlations for Abbyy FineReader (0,406), PDF (0,487) and PDF/A (0,452) are quite low. There is slightly stronger correlation between use of PAGE XML together with ODF formats (0,640), but there are only three observations for ODF in the survey data meaning that the correlation is dubious.

3.3.3 Labour Costs of Digitization

Question 26 asked rough estimates of labour costs for digitization excluding storage costs. Respondents were asked to estimate the proportion of labour costs for preparation of materials, scanning of material and post scanning costs such as quality assurance. As the type of materials being digitized can have a large effect on the costs, the numbers reported should be seen as rough averages for whole digitization of the respondents' organization.



Graph 10 Labour Costs of Digitization

Twenty four respondents were able to provide labour costs. There is quite large variance between institutions, some of which can be explained by the scale of digitization and, likely, difference between digitized materials. Means of labour costs are as follows:

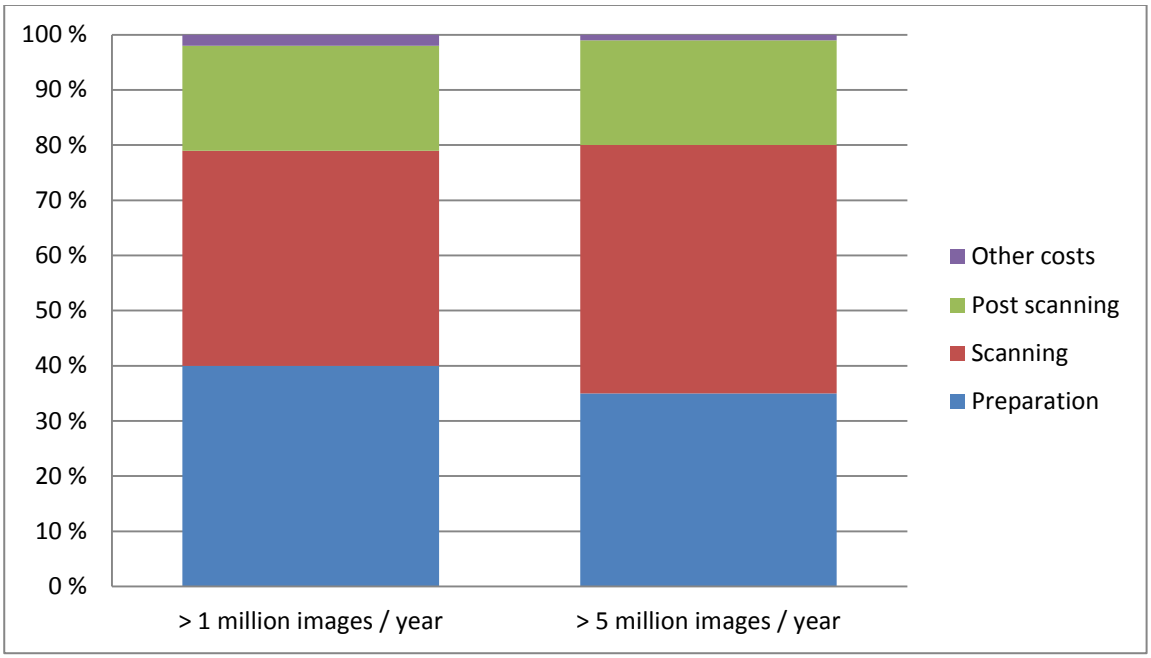
- Preparation 31 percent, standard deviation of 18.
- Scanning 45 percent standard deviation of 21.2.
- Post scanning 21 percent, standard deviation of 15.
- Other labour costs 3 percent, with a standard deviation of 5,5.

One respondent stated that most of their material had to go through conservation before digitization, which amount to 20 percent of costs, and was counted as other costs. Another one stated the conservation of material might drive the preparation cost up by a large amount. The conservation was counted as a preparation cost in the above distribution. Other costs were made of stock taking and administrative cost. Scanning costs had the greatest variance with minimum costs of 15 percent and maximum costs of 90 percent. The Swedish National Archives (15%), the Historical Archives of the European Commission (25%), Swiss Federal Archives (25%), Tata Central Archives (25%) and US National Archives and Records Administration (25%) had the lowest scanning costs. Preparation costs ranged from zero to 60 percent, and post scanning costs from 3 to 60 percent of all costs.

Costs for preparation and scanning do level out at circa 40 percent when taking in account institutions that digitize at more than million images yearly.[9] The group is made of eleven

[9] Preparation costs: mean 40%, standard deviation 17.5. Scanning costs: mean 39%, standard deviation 19.5.

institutions, ten of which are national archives.[10] The mean of post scanning costs is 19 percent with a standard deviation 11.4. Other costs are 2 percent.[11] The scanning cost increase again to a mean of 45 percent,[12] while preparation costs drop to 35 percent,[13] when at least five million images are digitized yearly. Post scanning costs remain at roughly the same level with a mean of 19 percent.[14] Other costs have a mean of 1 percent.[15] A 20% random sample of provides similar results[16].



Graph 11 Labour costs of digitization - largest digitizers

A number of respondents stated that the nature of material digitized can affect the preparation costs significantly, if it needs conservation or is time consuming to prepare for the scanners. There is

[10] They are the national archives of Australia, Belgium, Denmark, Estonia, Finland, the Netherlands, Norway, Portugal, Sweden, USA and the Historical Archives Service of the European Commissions.
[11] Standard deviation is 4.7.
[12] Standard deviation is 22.
[13] Standard deviation is 18.3.
[14] Standard deviation is 13.4.
[15] Standard deviation is 1.9.
[16] Preparation costs: mean 35%, standard deviation 21.8 scanning costs: mean 43%, standard deviation 24.5; post scanning costs: 18%, standard deviation 11.9; other costs: mean 3%, standard deviation 7.3. Australia was included in the random sample. It reported scanning costs at 80% level, but calculations without Australia produced results that were almost exactly the same as the ones reported above with a difference of less than 3%.

significant variation in the costs scanning, which can be partially explained by the small sample of twenty four responses and the fact that no data on the circumstances digitization or most commonly digitized materials was asked. The amount of digital images produced yearly do seem to affect somewhat the preparation costs with Finland and USA reporting that preparation covers roughly 50 percent while Sweden stats that preparation accounts to 40 percent of labour costs. Portuguese Direção Geral do Livro dos Arquivos e das Bibliotecas reported preparation costs as 30 percent, but as they also cover digitization cost for libraries they might not be easily comparable to digitization costs of archival material. National Archives of Estonia and Norway report the highest preparation costs of the institutions that produce more than million images per year. Their preparation costs amount to 60 percent of all labour costs.

Istanbul University (60%), Archives of the Republic of Slovenia (50%), the Swedish National Archives (45%) and Stadarchiv Mannheim (40%) had the highest proportion of post scanning costs. The level of quality assurance can affect significantly post scanning costs. Relatively high post scanning costs of the Swedish National Archives arise, at least partially, from human examination of all digital images destined for the National Archives.[17]

The very large variance in labour costs of digitization can be partially explained by the wide difference in scale of digitization amongst the participants of the survey. If digitization is relatively small scale, one can expect the scanning costs be higher due to unfamiliarity of scanner operators. On the other hand, scanning costs were over 50 percent of all labour costs amongst some of the

[17] The digitization centre of Swedish National Archives also produces images for other institutions, which do not have as strict quality assurance for their digitized files.

larger digitizers. The relatively high scanning costs amongst the participants might be caused by un-optimized digitization process. Post scanning costs amount to circa fifth of labour costs of digitization, which likely indicates that quality assurance and metadata ingestion is relatively low key. One explanation for this could be that material is digitized in order to provide digital access for customers and not for digital preservation, which would likely result in higher post scanning costs.

4 Summary

The digitization survey had a relatively good coverage with forty responses ranging from large national archives, to local archives, university libraries and private businesses. The answers were mainly from European organizations, but there were responses from Australia, India, Israel and USA. Majority of respondents had digitized a wide range of items from audio-visual media to more traditional textual archival material. Digitization is mostly organized as mixture of digitization programs and fixed duration projects. Many of the participants of the survey had outsourced digitization at some point. Based on the number of yearly produced images largest digitizers were US National Archives and Records Administration, Direção Geral do Livro dos Arquivos e das Bibliotecas of Portugal, the National Archives of Finland, the Swedish National Archives, the National Archives of Australia and Danish National Archives, all of which digitized more than ten million images per year.

Scanning parameters and the choice of file format of the preservation or master file of the digitized image were made in order to minimize loss of information during the digitization process. Majority of the participants had chosen TIFF as their master file format and used JPEG for dissemination of the images. The preferred type of scanned image is a 24 bit RGB colour image with resolution of 300 dpi. Depending on the type of digitized material the imagining resolution can be higher in order to capture all details.

The survey gauged labour costs of digitization process by asking estimates of relative costs of preparation of material for digitization, actual scanning and post scanning operations such as quality assurance and metadata ingestion. Twenty four participants could provide estimates on the labour costs. Averages of the costs were: preparation 35 percent, scanning 45 percent, post scanning 21 percent and other costs 3 percent. There was a wide range of variation amongst the responses regardless of the scale of digitization. Scanning costs are relatively high amongst most of the respondents, but some respondents stated that scanning amounts circa 25 percent of all costs. This might indicate that the scanning practices could be rationalized for more efficient digitization process.

5 Appendices

5.1 Appendix 1 – Questionnaire and answers

The questions are followed by number of answers and their percentage. Written answers have not been included.

1. Name of the organization

40

2. Country

40

3. Name of contact person

40

4. Position in the organization

40

5. E-mail address

40

6. Do you wish your answers to be anonymised in the written report?

No	52,5%	21
Yes	47,5%	19

7. How do you organize your digitization?

Does your organization have an ongoing digitization progam or do you engage in more project based digitization? Do you provide on demand digitization for scholars? Please select all relevant fields.

Ongoing program	60,0%	24
Project based digitization (	60,0%	24
On demand digitization	65,0%	26
No digitization program currently	7,5%	3
Other, please specify	12,5%	5

8. How your organization digitizes its analogue materials?

Is the actual digitization process conducted in-house or do outsource it? Please select all relevant fields.

In-house digitization	92,5%	37
Outsourced digitization	65,0%	26
Collaboration	37,5%	15
Other, please specify	10,0%	4

9. Type of digitized materials

What kind of material your organization has digitized or is planning to digitize? Please select all relevant fields.

Audio media	52,5%	21
Books	50,0%	20
Bound items	90,0%	36

Individual records, small dossiers	90,0%	36
Modern A4 documents	65,0%	26
Maps and technical drawings	77,5%	31
Microfilms or microfiches	45,0%	18
Photographs	77,5%	31
Other types, please specify	30,0%	12

10. What is the approximate quantity of your digitized materials?

How much analogue material have you digitized in linear metres?

Please answer in whole numbers without punctuation or spaces, for example, 20000 = 20 km.

27

What is the approximate number of your digital images?

Please answer in whole numbers without punctuation or spaces.

36

11. How much material you digitize yearly?

In linear meters

Please answer in whole numbers without punctuation or spaces, for example, 20000 = 20 km

25

Digitized images

Please answer in whole numbers without punctuation or spaces.

34

12. Technical specifications for scanned digital images

What are your minimum or recommended specifications for scanned images? Please provide specifications you use such as image type (grayscale, colour, black and white),bit depth, colour mode, resolution (dpi), scanning ratio, compression etc.

Instead of typing the specifications you can upload them or provide an online link in the text field bellow.

34

13. Image formats for digitized records

What type of files your digitization process produces? You can select multiple formats.

Format of preservation file

Preservation file = the master file used for storage, conservation or as a basis for creation access files

JPEG	45,0%	18
JPEG 2000	17,5%	7
PDF	25,0%	10
PDF/A	30,0%	12
PNG	7,5%	3
TIFF	85,0%	34
Other, please specify	7,5%	3

Format of access file

Access file = the file used for distribution and consultation of digitized record

JPEG	77,5%	31
JPEG 2000	7,5%	3
PDF	42,5%	17
PDF/A	15,0%	6
PNG	7,5%	3
TIFF	25,0%	10

Other, please specify	7,5%	3

14. Has the analogue material affected your choice of image formats for preservation and access files?

For example when digitizing photographs you save the master file as TIFF and the access file as JPEG.

Yes	57,5%	23
No	42,5%	17

15. If you answered yes to the previous question, why and how the analogue material has affected your choice of image formats.

20

16. Have you used OCR or other text recognition methods for the captured digital images?

Yes	55,0%	22
No	45,0%	18

17. If you have used OCR, what file format you have used to save the text captured from the images? Select all relevant file formats.

ABBYY FineReader XML	43,5%	10
ALTO XML	13,0%	3
hOCR	0,0%	0
PAGE XML	13,0%	3
PDF	39,1%	9
PDF/A	34,8%	8
Open Document Format (ODF)	13,0%	3
TEI	0,0%	0
Word Document format (DOCX)	17,4%	4
Other format, please specify	26,1%	6

18. Does your organization have a policy to dispose analogue records after digitization?

Disposal is understood here to cover destruction, transfer of custody or alteration records after their digitization.

Yes	10,0%	4
No	90,0%	36

19. What the criteria or conditions analogue records have to fill in order for them to be eligible for disposal after digitization?

If you have written disposal criteria, you can provide an online link to them or you can upload them with the button bellow instead of typing.

12

20. Do you have different quality criteria for digitization when the analogue records are disposed?

For example, do you use stricter scanning and quality assurance than you normally employ, when the original analogue records are destroyed after digitization?

Yes	24,1%	7
No	75,9%	22

21. If you answered yes to the previous question, please specify how the criteria differ?

For example, if you create different type of image files or files, or do you need stricter quality assurance processes, when disposal is possible.

You can provide an online link in the text box or upload the criteria instead of typing.

7

22. Has your organization conducted or planned a mass scale digitization project?

Implemented a large scale project or on-going mass digitization	52,5%	21
Planned or in process of planning a mass digitization project	17,5%	7
No	45,0%	18

23. What is the scale of your mass digitization project?

Linear kilometres of digitized analogue records

Please answer in whole numbers without punctuation or spaces, for example, 20000 = 20 km

14

The number of digitized images

Please answer in whole numbers without punctuation or spaces.

19

24. If you have conducted a mass digitization project, what materials have you digitized?

Please select all relevant types

Books	39,1%	9
Bound items	78,3%	18
Index cards	43,5%	10
Microfilms or microfiches	39,1%	9
Paper files, records containing mixed formats	52,2%	12
Paper files, uniform format such as A4	43,5%	10
Other, please specify	17,4%	4

25. Have you unbound bindings before digitization in order to speed the scanning process?

For example, have you used paper guillotine cutter?

Yes	33,3%	13
No	66,7%	26

26. **How are labour costs divided between preparation of material, scanning and post scanning? Please enter approximate percentual proportions bellow. The costs should total 100%**

Preparation = costs due to removal of stables or binding from records, sorting material into stacks for scanner etc.

Scanning = operator costs of the actual scanning. The cost of hardware is not included.

Post scanning = costs due to validation of the scanned material, post processing.

Please answer in whole numbers without spaces.

Preparation % of labour costs	25
Scanning % of labour costs	25
Post scanning % of labour costs	25
Other costs, please state the type and % of labour costs	11

27. What is the single most important factor that affects throughput speed of scanning?

30

28. Have you outsourced or planned to outsource your mass digitization project?

Yes	31,4%	11
No	40,0%	14
Both in-house and outsourced digitization	31,4%	11

29. If you have outsourced your mass scale digitization project, what service providers you have used?

Please state, if you do not want to the names of service providers to be made public.

17

30. If you have outsourced digitization, have you used different digitization specifications from in-house digitization?

For example, use of less strict digitization specifications than when done in-house.

Yes	3,7%	1
No	96,3%	26

31. If you answered yes to the previous question, please specify how the specifications have differed.

You can also upload specifications for outsourced digitization bellow.

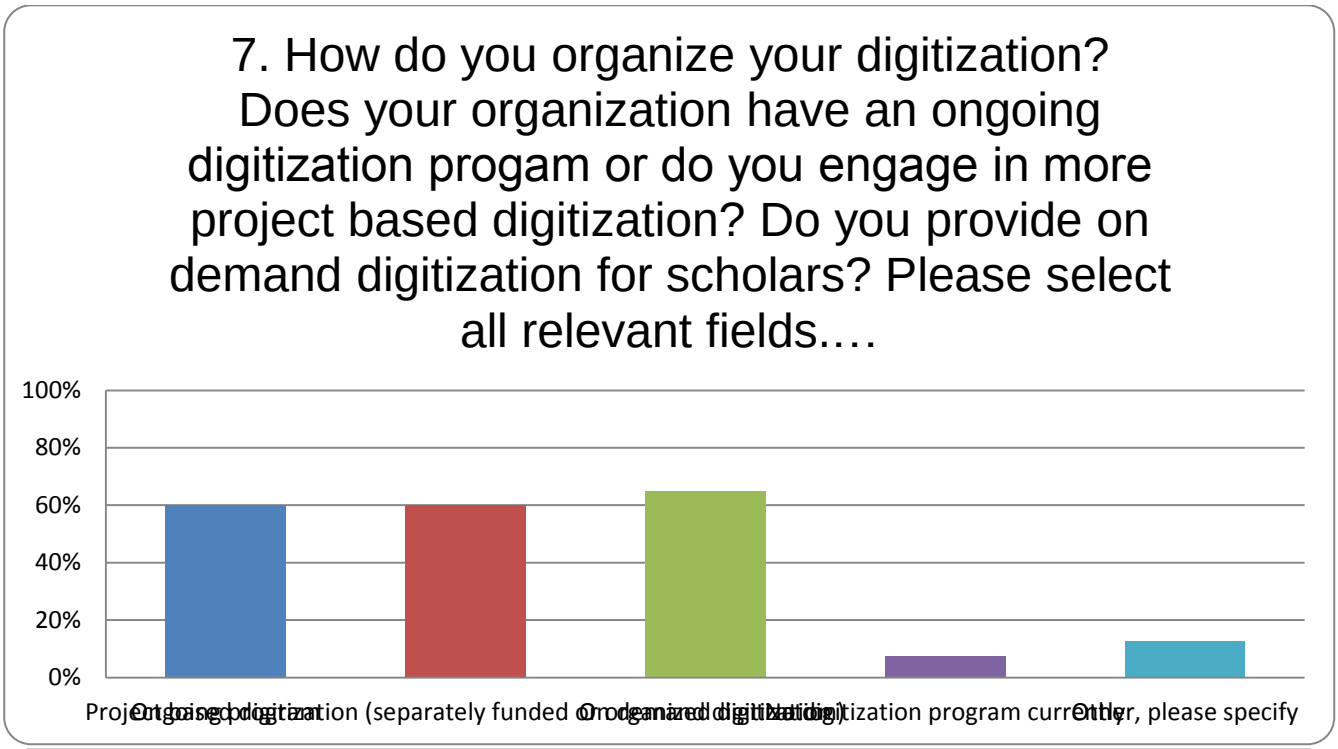
3

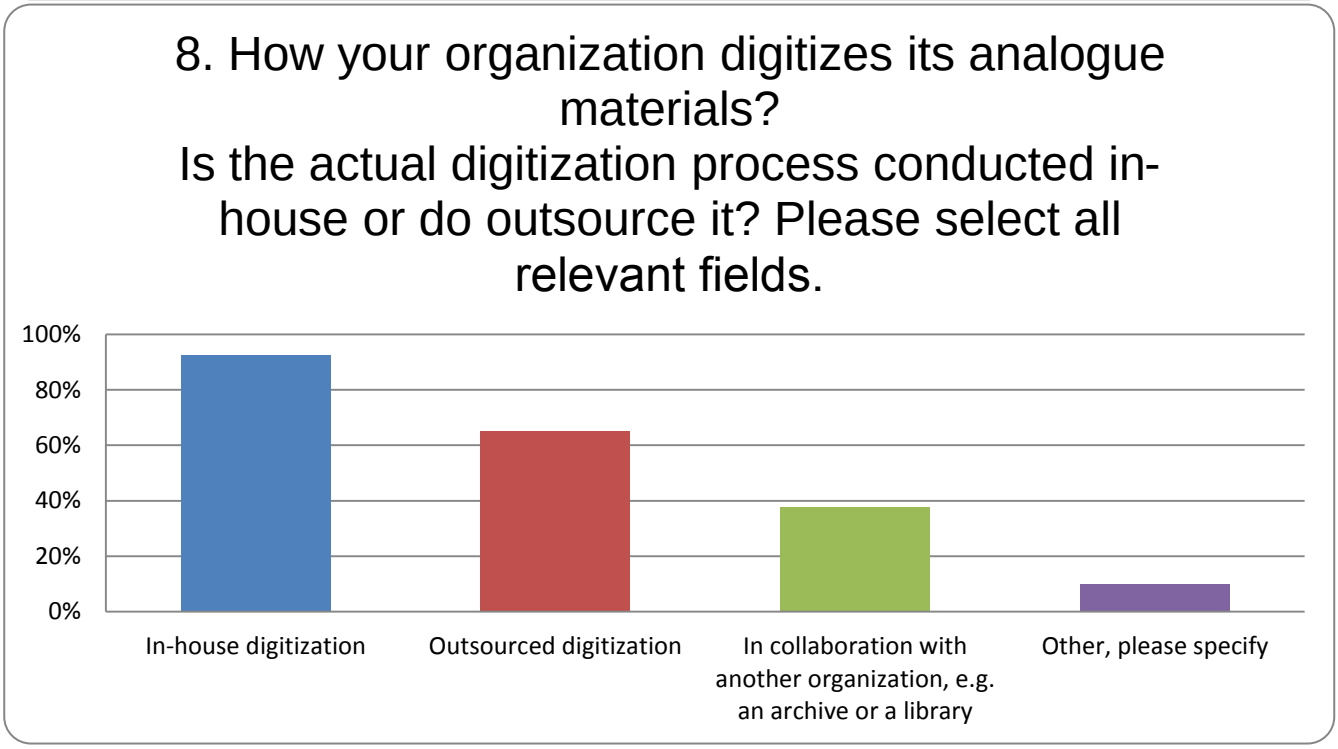
Thank you for taking time to complete this survey.

If you have anything to add or wish to specify something concerning digitization, you can use the field bellow.

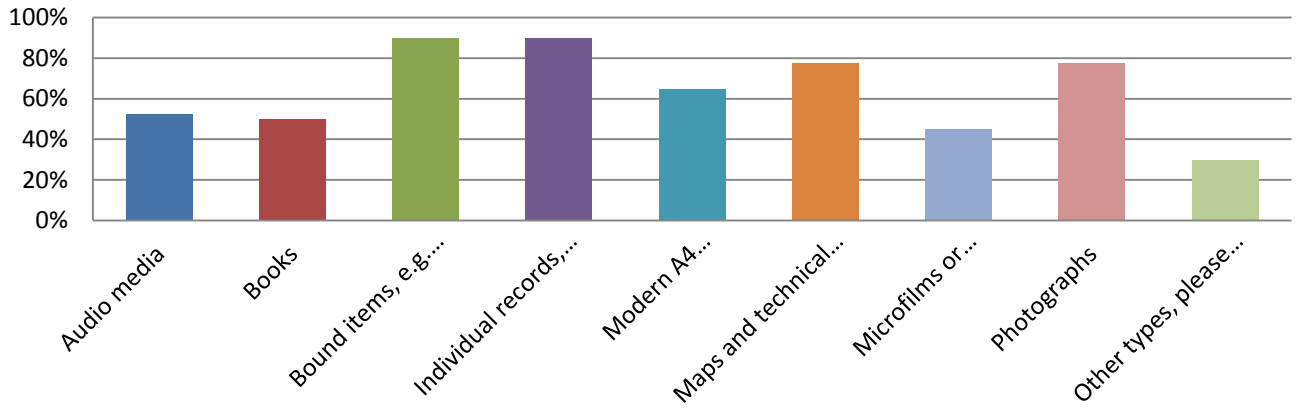
15

5.2 Appendix 2- Graphs



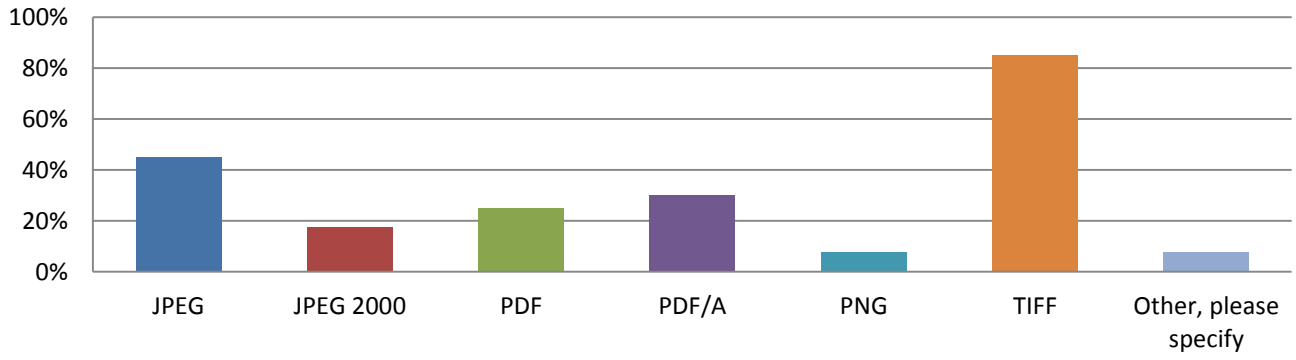


9. Type of digitized materials
What kind of material your organization has digitized or is planning to digitize? Please select all relevant fields.

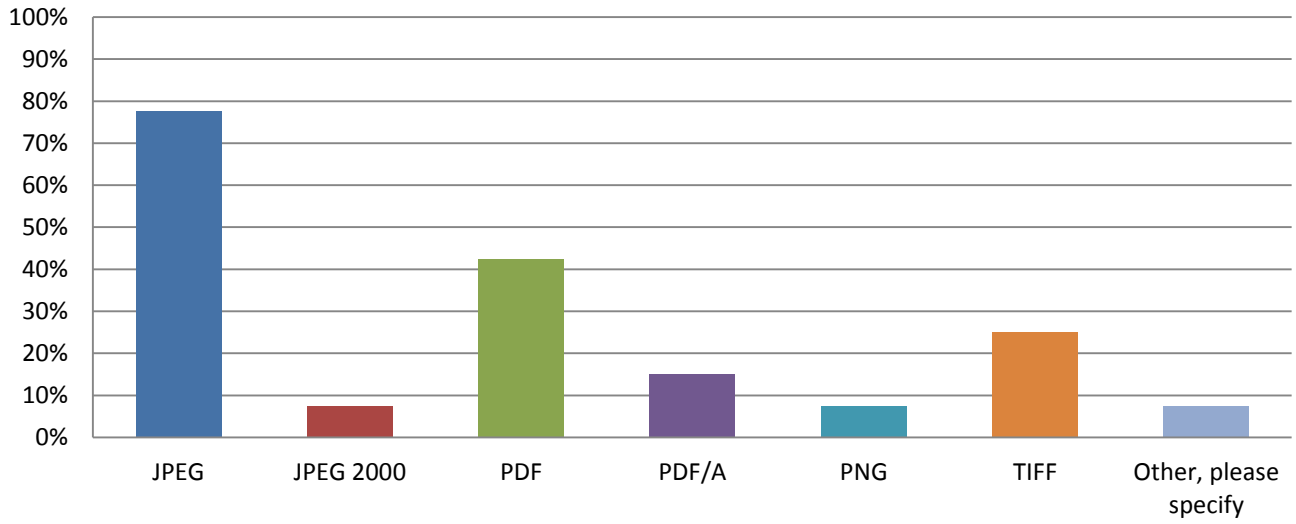


13. Image formats for digitized records
What type of files your digitization process produces? You can select multiple formats.

Format of preservation file
Preservation file = the master file used for…

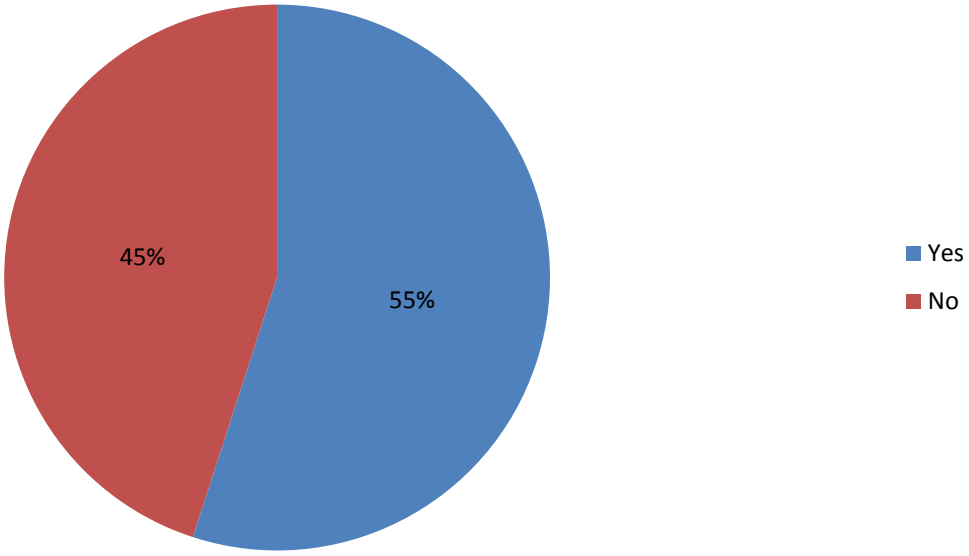


Format of access file
Access file = the file used for distribution and consultation of digitized record

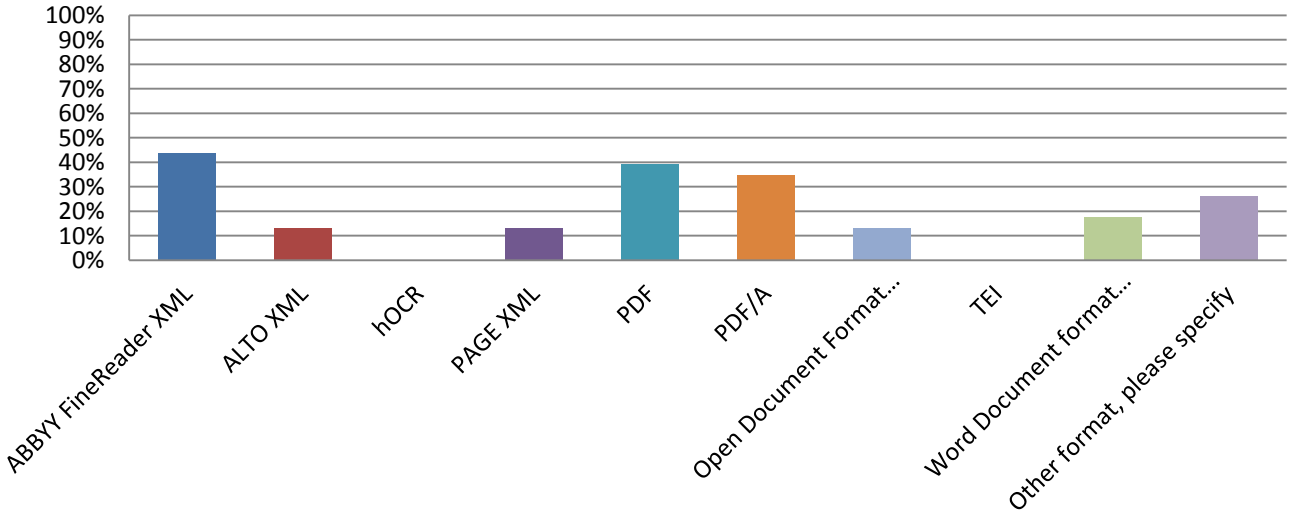


14. Has the analogue material affected your choice of image formats for preservation and access files?
For example when digitizing photographs you save the master file as TIFF and the access file as JPEG.

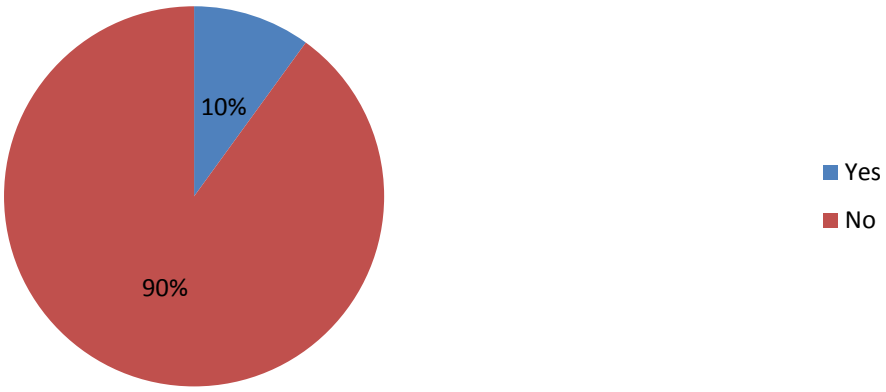
16. Have you used OCR or other text recognition methods for the captured digital images?



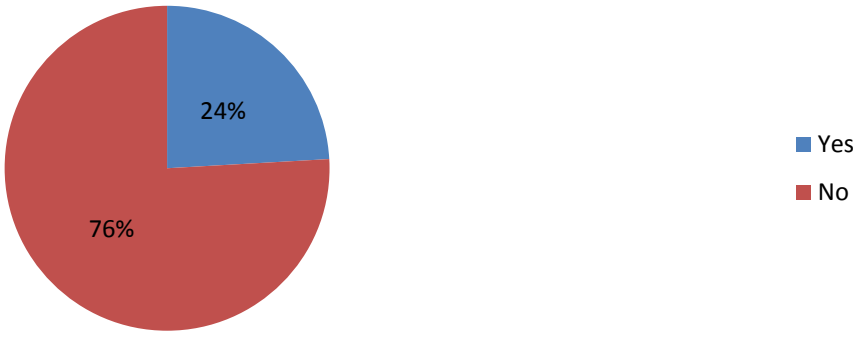
17. If you have used OCR, what file format you have used to save the text captured from the images? Select all relevant file formats.



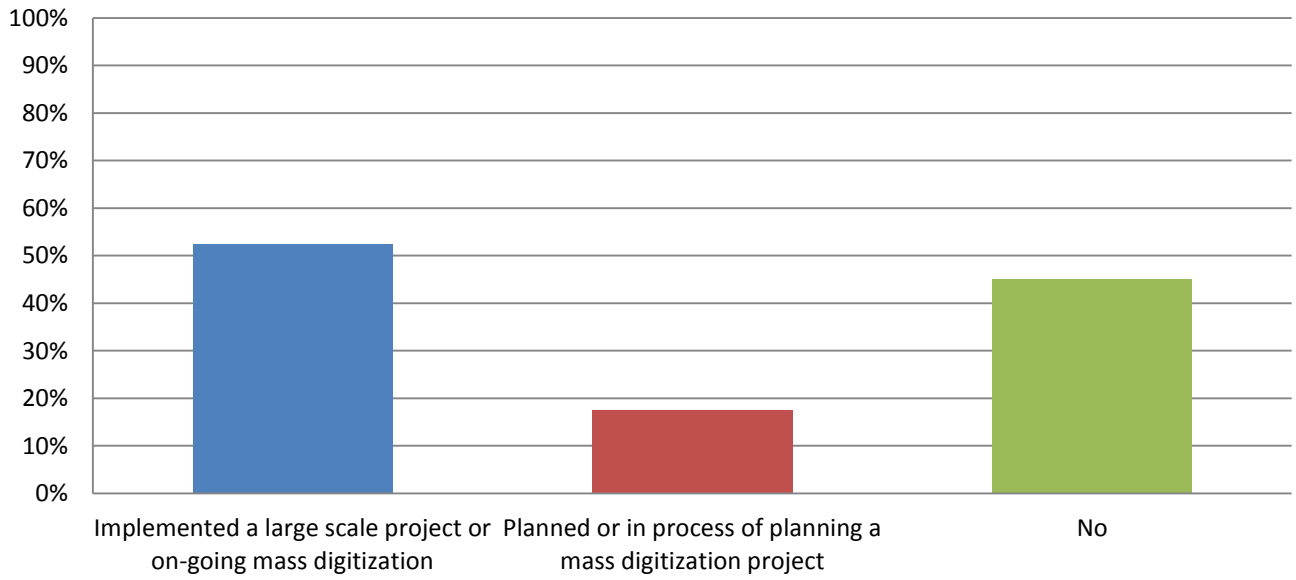
18. Does your organization have a policy to dispose analogue records after digitization? Disposal is understood here to cover destruction, transfer of custody or alteration records after their digitization.



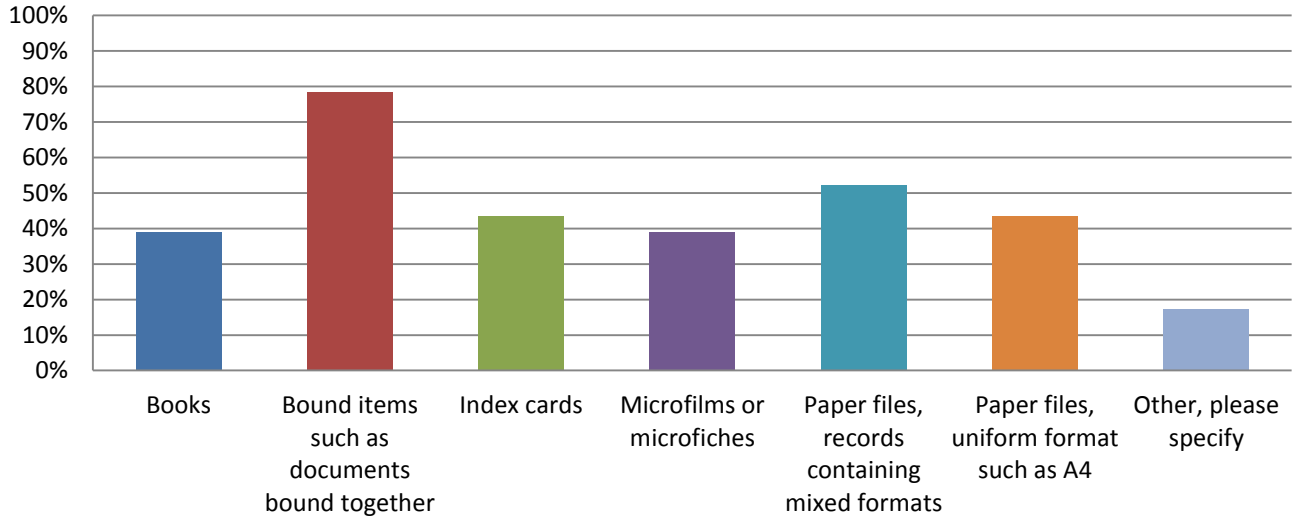
20. Do you have different quality criteria for digitization when the analogue records are disposed?
For example, do you use stricter scanning and quality assurance than you normally employ, when the original analogue records are…



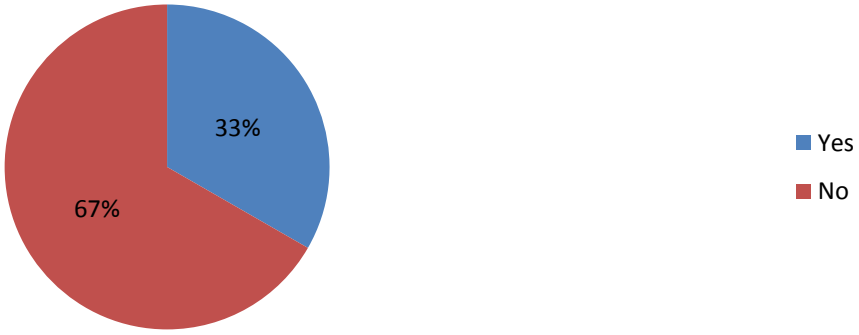
22. Has your organization conducted or planned a mass scale digitization project?



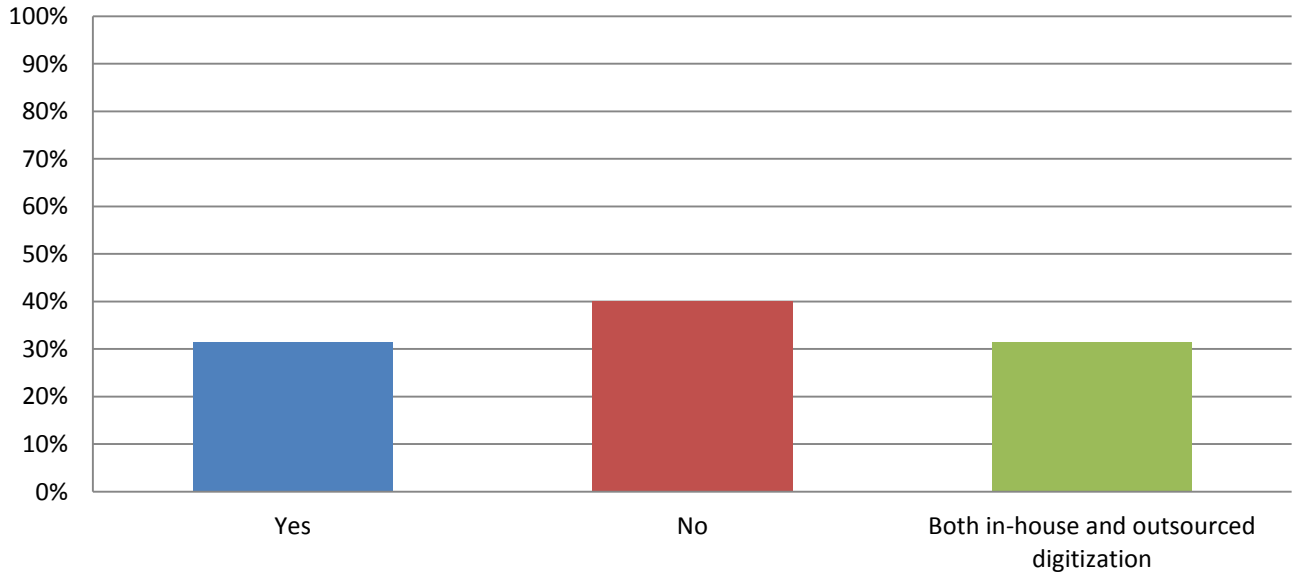
24. If you have conducted a mass digitization project, what materials have you digitized? Please select all relevant types



25. Have you unbound bindings before digitization in order to speed the scanning process? For example, have you used paper guillotine cutter.



28. Have you outsourced or planned to oursource your mass digitization project?



30. If you have outsourced digitization, have you used different digitization specifications from in-house digitization?
For example, use of less strict digitization specifications than when done in-house.

5.3 Appendix 3 – Tables

Table 1 Digitized images and yearly digitization

Country	Organization	Digital images	Yearly digitized images
Australia	National Archives of Australia		10164683
Austria	Österreichisches Staatsarchiv - Austrian State Archive		
Austria	Compass-Verlag GmbH	1200000	50000
Austria	Wiener Stadt und Landesarchiv	500000	
Belgium	State Archives Belgium	25000000	2500000
Belgium	Historical Archives Service of the EC	6000000	1200000
Belgium	City Archive of Leuven	800000	85000
Bulgaria	Archives State Agency	250000	45000
Croatia	State Archives in Zadar		100000
Croatia	Croatian State Archives	5000000	210000
Croatia	Državni arhiv u Bjelovaru	60000	100
Croatia	University of Zadar	10000	
Czech Republic	National Archives	10000000	900000
Denmark	Danish National Archives	55000000	10000000
Estonia	National Archives of Estonia	18000000	1500000
Finland	National Archives of Finland	60000000	18000000
Finland	Suomalaisen Kirjallisuuden Seura	468000	25000
Finland	Pohjan paikallishistoriallinen arkisto	1800	50
Germany	Stadtarchiv Mannheim	4000000	500000
Germany	Kreisarchiv Siegen-Wittgenstein	20000	1000
Germany	Archiv der Hansestadt Lüneburg (City archives of the Hanseatictown of Lueneburg)	20000	6500
Germany	University of Bielefeld	1167	500

Germany	Gemeindearchiv Schöneiche bei Berlin	0	0
Greece / Italy	ARCH / Associazione Raffaello Sanzio	200000	50000
Iceland	Þjóðskjalasafn Íslands / The National Archives of Iceland	300000	160000
India	Tata Central Archives	600000	80000
Israel	Kibbutz Alumim	7300	
Norway	National Archives of Norway	48000000	4300000
Poland	The National Archives in Krakow	1200000	150000
Portugal	Direção Geral do Livro dos Arquivos e das Bibliotecas	1750000000	30000000
Portugal	National Archives of Portugal	25168544	5248443
Portugal	Archive of the Anglican Church of Portugal	542	542
Slovenia	Archives of the Republic of Slovenia	1255000	350000
Sweden	National Archive, Sweden	100000000	11400000
Switzerland	Swiss Federal Archives		
Switzerland	State Archives of the Canton of Zurich	5000000	
The Netherlands	Nationaal Archief / National Archives of the Netherlands	30293518	6300000
Turkey	Istanbul University	259500	1000000
UK	University of London: Senate House Library	1500000	1
USA	US National Archives and Records Administration	390000000	30000000
USA	Independent Consultanting Services LLC	6100	500

Country	Type of document	File format	Resolution	Image type	Bit depth	Colour mode	Compression	Notes
Australia: National Archives of Australia	Maps or drawings, colour	TIFF	400	colour / grayscale	16	Pro Photo RGB or Gray Gamma 2.2	No	Also over A3 paper records
Australia: National Archives of Australia	Manuscript or typescript material	TIFF	300	colour	24	AdobeRGB (1998)	No	
Australia: National Archives of Australia	Film (negatives, dias, glass plates)	TIFF	2000 / 5000	colour / grayscale	16	Pro Photo RGB or Gray Gamma 2.2	No	5000 ppi for 35mm and under, 2000 ppi for over 35mm
Australia: National Archives of Australia	Photographs	TIFF	600 / 900	colour / grayscale	16	Pro Photo RGB or Gray Gamma 2.2	No	900 ppi for under 10 x 15 cm, 600 ppi for 10 x 15 cm to A3
Austria: Compass-Verlag GmbH	Manuscript or typescript material	TIFF	300	colour				
Austria: Compass-Verlag GmbH	Manuscript or typescript material	JPEG	300	colour				
Austria: Wiener Stadt und Landesarchiv	Manuscript or typescript material	TIFF, JPEG, PDF		colour / grayscale				
Austria: Österreichisches Staatsarchiv - Austrian State Archive	Manuscript or typescript material	TIFF	300	colour	24	RGB	No	On demand JPG and upto 600 dpi
Belgium: City Archive of Leuven	Manuscript or typescript material	TIFF	300	colour	24		No	
Belgium: Historical Archives Service of the EC	Manuscript or typescript material	TIFF	300	grayscale	8	Grayscale	No	Also bitonal PDFs
Belgium: State Archives Belgium	Manuscript or typescript material	TIFF	300	grayscale			No	Documents without colour
Belgium: State Archives Belgium	Manuscript or typescript material	TIFF	300-400	colour			No	Iconographic colour documents
Belgium: State Archives Belgium	Maps and drawings, colour	TIFF	300-400	colour			No	
Bulgaria: Archives State Agency	Manuscript or typescript material	TIFF	300	colour	24	RGB	No	
Bulgaria: Archives State Agency	Maps and drawings, colour	TIFF	600	colour	24	RGB	No	400dpi for large maps
Croatia: Croatian State Archives	Manuscript or typescript material	TIFF, JPEG, JPEG2000, PNG	300	colour	24	RGB	lossless	400 dpi if small characters
Croatia: Croatian State	Microfilms and	TIFF,	300	grayscale		Grayscale	lossless	400 dpi if small

Archives	microfiches	JPEG, JPEG2000, PNG						characters
Croatia: State Archives in Zadar	Manuscript or typescript material	JPEG	300	colour				
Czech Republic: National Archives	Manuscript or typescript material	PNG	300	colour	24	RGB	No	
Czech Republic: National Archives	Manuscript or typescript material	JPG	300	colour	24	RGB	No	
Denmark: Danish National Archives	Manuscript or typescript material	TIFF	300	colour	24		No	
Estonia: National Archives of Estonia	Microfilms and microfiches	PNG	300	grayscale	8		No	
Estonia: National Archives of Estonia	Manuscript or typescript material	TIFF	300	colour	24		No	
Estonia: National Archives of Estonia	Manuscript or typescript material	TIFF	300	grayscale	8		No	
Estonia: National Archives of Estonia	Photographs, black and white	TIFF	600-3200	colour	24	RGB	No	
Estonia: National Archives of Estonia	Photographs, colour	TIFF	600-3200	colour	24	RGB	No	
Finland: Finnish Literature Society	Manuscript or typescript material	TIFF	300	colour	24	sRGB	No	
Finland: Finnish Literature Society	Manuscript or typescript material	TIFF	300	grayscale	8	Gray gamma 2.2	No	Type written materials
Finland: National Archives	Photographs, black and white	TIFF	300	grayscale	8	Gray gamma 2.2.	No	
Finland: National Archives	Photographs, colour	TIFF	300	colour	24	RGB	No	eciRGB v2, ProPhoto RGB, Adobe RGB,
Finland: National Archives	Maps or drawings, monochrome	TIFF	300	grayscale	8	Gray gamma 2.2.	No	
Finland: National Archives	Maps or drawings, colour	TIFF	300	colour	24	RGB	No	eciRGB v2, ProPhoto RGB, Adobe RGB,
Finland: National Archives	Manuscript or typescript material	TIFF	300	grayscale	8	Gray gamma 2.2.	No	
Finland: National Archives	Manuscript or typescript material	TIFF	300	colour	24	RGB	No	eciRGB v2, ProPhoto RGB, Adobe RGB, sRGB
Finland: National Archives	Microfilms and microfiches	TIFF	300	grayscale	8	Gray gamma 2.2.	No	
Germany: Archiv der Hansestadt Lüneburg	Manuscript or typescript material	TIFF	600	colour	24	AdobeRGB (1998)	No	Also PDF/A
Germany: DFG	Manuscript or typescript material	TIFF	300	colour	24	RGB	No	TIFF-LZW or JPEG2000 also possible; Official guideline of Deutschen Forschungsgemeinschaft

								(DFG)
Germany: DFG	Manuscript or typescript material	TIFF	300	grayscale	8	Gray gamma 2.2.	No	TIFF-LZW or JPEG2000 also possible; Official guideline of Deutschen Forschungsgemeinschaft (DFG)
Germany: Gemeindearchiv Schöneiche bei Berlin	Manuscript or typescript material	TIFF	300	colour	24		LZW	
Germany: Kreisarchiv Siegen-Wittgenstein	Manuscript or typescript material	TIFF, JPEG, PDF	100	colour			No	
Germany: Stadarchiv Mannheim	Manuscript or typescript material	TIFF	300	colour	24	RGB	No	TIFF-LZW or JPEG2000 also possible; Official guideline of Deutschen Forschungsgemeinschaft (DFG)
Germany: Stadarchiv Mannheim	Manuscript or typescript material	TIFF	300	grayscale	8	Gray gamma 2.2.	No	TIFF-LZW or JPEG2000 also possible; Official guideline of Deutschen Forschungsgemeinschaft (DFG)
Germany: University of Bielefeld	Manuscript or typescript material	TIFF	600	grayscale	16	Grayscale	No	
Germany: University of Bielefeld	Manuscript or typescript material	TIFF	600	colour	24		No	
Greece / Italy: ARCH / Associazione Raffaello Sanzio	Manuscript or typescript material	TIFF, JPEG, PDF	300-1200	colour				
Iceland: Þjóðskjalasafn Íslands / The National Archives of Iceland	Manuscript or typescript material	TIFF	400	colour	8	RGB	No	
India: Tata Group Central Archives	Manuscript or typescript material	TIFF	300	colour	24		No	True colour
Israel: Kibbutz Alumim	Manuscript or typescript material	PDF	150	colour				
Neatherlands: Nationaal Archief	Manuscript or typescript material	TIFF	300	colour	24	RGB, eci v2	No	
Neatherlands: Nationaal Archief	Manuscript or typescript material	JPEG	300	colour	24	RGB, eci v2	baseline 1:10	
Neatherlands: Nationaal Archief	Manuscript or typescript material	PDF/A-1b	300	colour	24	RGB, eci v2	No	tolerated sampling rate 2%
Norway: National Archives of Norway	Microfilms and microfiches	TIFF	300	grayscale	8	Grayscale	LZW	converted to PNG for long term storage
Norway: National Archives of Norway	Manuscript or typescript material	TIFF, JPEG, JPEG2000	300	colour	24	RGB	lossless, LZW	converted to JPEG 2000 for long term storage

Norway: National Archives of Norway	Photographs	TIFF	600	colour	24	RGB	LZW	Stored as TIFF
Poland: the National Archives in Krakow	Manuscript or typescript material	TIFF	300	colour	24	RGB	No	
Portugal: Archive of the Anlican Church of Portugal	Manuscript or typescript material	JPEG, PDF		colour				
Portugal: Direção Geral do Livro dos Arquivos e das Bibliotecas	Manuscript or typescript material	TIFF	300	colour	8 / 24		No	Metadata - according to ANSI/NISO Z39.87-2006 (R2011) Data Dictionary - Technical Metadata for Digital Still Images
Portugal: Direção Geral do Livro dos Arquivos e das Bibliotecas	Microfilms and microfiches	TIFF	200	grayscale	8	Grayscale	No	Metadata - according to ANSI/NISO Z39.87-2006 (R2011) Data Dictionary - Technical Metadata for Digital Still Images
Portugal: Direção Geral do Livro dos Arquivos e das Bibliotecas	Photographs	TIFF	600 / 800 / 1200	colour	24 / 48		No	Metadata - according to ANSI/NISO Z39.87-2006 (R2011) Data Dictionary - Technical Metadata for Digital Still Images
Slovenia: Archives of the Republic of Slovenia	Manuscript or typescript material	TIFF	300-600	colour / grayscale	24	RGB	No	
Sweden: National Archives of Sweden	Manuscript or typescript material	TIFF	300	colour	24		No	
Switzerland: State Archives of the Canton of Zurich	Manuscript or typescript material	TIFF	300	colour	16		No	Based on Metamorphoze Preservation Imaging Guidelines and the Guidelines Digitization of Photographic Materials with slight differences, for example resolution
Switzerland: Swiss Federal Archives	Manuscript or typescript material	TIFF	300	colour	24	sRGB	No	
Switzerland: Swiss Federal Archives	Film (negatives, dias, glass plates)	TIFF	≥4000		48			
Switzerland: Swiss Federal Archives	Photograps	TIFF	900		48			
Turkey: Istanbul University	Manuscript or typescript material	TIFF, JPEG	300	colour	24		No	
UK: University of London: Senate House Library	Manuscript or typescript material	JPEG, JPEG2000	300	colour	24			
USA: US National Archives and Records Administration	Manuscript or typescript material	JPEG quality level above 10	300	colour	8	sRGB or grayscale		Acceptable minimum; data from survey ->likely reproduction master

USA: US National Archives and Records Administration	Manuscript or typescript material	JPEG quality level above 10	300 / 400	colour / grayscale	24	sRGB or grayscale		Preferred; 400ppi at original size up to 34" x 55" originals, 300ppi at original size for originals between 34" x 55" and 46" x 73", 300 ppi at original size for originals larger than 46" x 73", Multipage OCR PDF file for textual materials
USA: US National Archives and Records Administration	Photographs	JPEG quality level above 10	400 / 600 / 800	colour / grayscale		sRGB or grayscale		Preferred; 400 ppi for 8"x10", 600 ppi 5"x7", 800 ppi for 4"x5" or smaller
USA: US National Archives and Records Administration	Photographs	JPEG quality level above 10	300	colour / grayscale		sRGB or grayscale		Acceptable minumum

Table 2 Labour costs of digitization

Organization	Preparation %	Scanning %	Post scanning %	Other costs %	Explanation
Direção Geral do Livro dos Arquivos e das Bibliotecas of Portugal	50	40	10	0	Conservation 20%, counted as preparation costs.
US National Archives and Records Administration	50	25	25	0	Depending on the nature of the materials and project, conservation costs could be significant but this will be project dependent.
National Archives of Finland	50	40	10	0	
National Archive, Sweden	40	15	45	0	
National Archives of Australia	5	80	10	5	5% administrative costs
Danish National Archives	15	60	25	0	

Nationaal Archief / National Archives of the Netherlands	35	55	10	0	
National Archives of Norway	60	20	15	5	Indexing and availability to users plus equipment
State Archives Belgium	30	40	30	0	
National Archives of Estonia	60	30	10	0	
Historical Archives Service of the EC	40	25	20	15	
Istanbul University	10	30	60	0	
Stadtarchiv Mannheim	30	30	40	0	
Archives of the Republic of Slovenia	0	50	50	0	
Croatian State Archives	30	50	20	0	
Þjóðskjalasafn Íslands / The National Archives of Iceland	7	90	3	0	

The National Archives in Krakow	20	40	20	20	10 - sharing scans; 10 - archiving scans
Tata Central Archives	50	25	15	10	Stock Taking 10
Compass-Verlag GmbH	20	50	30	0	
Archiv der Hansestadt Lüneburg (City archives of the Hanseatic town of Luenebur g)	15	70	10	5	5 (technical service)
Independent Consultanting Services LLC	30	40	20	10	
Državni arhiv u Bjelovaru	15	80	5	0	
Swiss Federal Archives	50	25	25	0	
Kibbutz Alumim	20	75	5	0	