The Use of Preservation Tools among Dutch Heritage Organizations

20 May 2020

By Ania Molenda

dutch digital heritage network

Contents

Acknowledgements

The findings described in this report are based on the results of a survey conducted between March and April 2019 among twenty-seven heritage organizations in the Netherlands that work with digital collections.[1] The survey was conducted as part of the project Preservation Tools (*Preserveringstools*) within the activities of the Sustainable Digital Heritage domain (*Digitaal Erfgoed Houdbaar*) of the Dutch Digital Heritage Network (*Netwerk Digitaal Erfgoed*). It documents the current use of preservation tools among the organizations that in 2018 reported that they have a digital repository[2].

I would like to thank all of the institutions and colleagues who engaged in preparing this survey. Especially Joost van der Nat, Marcel Ras, Sam Alloing, Pepijn Lucker, Walter Swagemakers, Jacob Takema, Remco van Veenendaal, Ernst van Velzen, and Erwin Verbruggen. I would also like to thank all the respondents who kindly shared their knowledge in the context of this research. Your contribution was invaluable to better understand the current situation and research possible ways to address it.

[1] The Regional Historical Centres were not included in this survey.

[2] Joost van der Nat, Marcel Ras, 'Survey Digitale Archieven in Nederland', 2018. Available online at: https://www.netwerkdigitaalerfgoed.nl/wp-content/uploads/2018/05/20180507-Netwerk-Digitaal-Erfgoed-Survey-Digitale-Archieven.pdf

Executive Summary

This report is based on the results of a survey 'Digital Tools in the Netherlands' conducted in 2019 as part of the project Preservation Tools (*Preserveringstools*) of the Dutch Digital Heritage Network (*Netwerk Digitaal Erfgoed*). It was carried out among the twenty-seven heritage organizations, which in 2018 reported that they have a digital repository. It focused on understanding how those institutions use automation for processing digital collections at the level of pre-ingest[3] and ingest.[4] The survey looked at the current state of tool usage, but also at future needs, the development of workflows, the use of standards, and the influence of the quality of acquired collections on pre-ingest workflows.

From the research it is clear that these heritage organizations face challenges at different levels. Many of them are still developing and implementing their digital preservation systems. Some are ready and willing to improve the systems and workflows they have in place to process bigger and more complex collections faster, but at the moment only a few can implement advanced improvements such as chain-automation and quality control. Several institutions recognize the need for pre-ingest in their workflows and report that automating parts of pre-ingest and ingest is relevant to them. That is the case not only for institutions with large and/or complex collections, but also for small institutions that do not have the manpower to process even relatively small collections without the help of preservation tools. It is also clear that digital collections are changing both in scope and level of complexity. For more than half of the interviewed organizations, digital collections will change in the coming two years and will include a bigger array of complex digital objects such as social media, email, and three-dimensional digital objects.

Despite many differences, there are areas of overlap in needs and a great potential for knowledge exchange, with best practice indicated as the most sought-after form of collaboration. While more than half of the tools currently pending implementation are new to the community, about half is already in use at other organizations. It means that there is plenty of space for both knowledge exchange and experiment when it comes to introducing new preservation tools.

The differences between collections and institutions are big and will remain so because of their varied character, legal obligations, and relationships with depositors. Most of the interviewed organizations receive their collections from external sources and most of them have only partial influence on how the material is deposited. Only about a third of the interviewed heritage organizations can set requirements and thus influence how the collections they receive are prepared

[3] Pre-ingest as defined by the Digital Preservation Coalition (DPC) in its Open Archival Information System (OAIS) Community wiki is a stage currently not included in the OAIS model, but often applied in archiving practices. It is a stage in which the received material can be checked on aspects that are fundamental for deciding whether it should be accepted into the repository as is, whether it should be rejected, or whether certain actions are required prior to its acceptance. (https://wiki.dpconline.org/index.php?title=Pre-ingest)

[4] As defined by the DPC glossary, ingest is a part of the OIAS. It is the process of turning a Submission Information Package (SIP) into an Archival Information Package (AIP), i.e. putting data into a digital repository. (https://www.dpconline.org/handbook/glossary)

and delivered. This confirms that for many of the respondents, using pre-ingest tools will remain vital to properly process their collections. The majority reports that pre-ingest plays an important role in their workflow. Even though they might call it differently, 73 percent make a distinction between pre-ingest and ingest. At the same time, for many of them workflow descriptions remain in development. For 77 percent, workflow descriptions are not yet in use or are in use partially. Out of those, 45 percent is currently working on developing one or more workflow descriptions, which shows that there is considerable dynamism in the creation of workflow descriptions among the interviewed organizations.

Organizations use a large number of tools, on average between one and ten, and this number seems to be growing. In total 111 tools were mentioned as currently used in production among eighteen respondents, and more than half of them said that they are interested in testing or implementing new tools. A lot of the tools currently used are proprietary (41 percent); however, with 68 percent indicating open source tools as interesting for implementation in the future, there is substantial interest in introducing more open source tools among heritage organizations.

Looking at different steps in the workflows, not many organisations use preservation tools for acquisition, but most of them use tools for transfer and processing. Most respondents report that they currently use tools for fixity check, virus scan, and the transfer of digital-born material (between 60 percent and 70 percent); the normalization or conversion of files, extraction of technical metadata, file format identification and validation (70 percent). Nevertheless, there still seems to be a significant interest in tools related to file format identification and validation, technical metadata extraction, as well as normalization and conversion. At the other end of the spectrum, the least mentioned actions include deduplication (11 percent), encryption detection and dependency analysis (17 percent), which calls for attention and further analysis.

Introduction

The survey 'Digital Tools in the Netherlands', which forms the basis of this report, was conducted between March and April 2019 as part of the project Preservation Tools (*Preserveringstools*) of the Dutch Digital Heritage Network (*Netwerk Digitaal Erfgoed*). It was carried out among twenty-seven different heritage organizations including archives, libraries, museums and other types of institutions. The goal of this survey was to gain more insight into the different types of digital collections currently held by various institutions in the Netherlands, and into the types of tools and workflows that are used to prepare them for long-term preservation at the pre-ingest and ingest stages. It focused on investigating which preservation tools are being used for what actions or types of digital objects. Can we identify processes that lack appropriate tools? Are there tools that could meet those needs? Could we better understand where knowledge exchange and collaboration are most urgently required?

The survey also investigated the context in which preservation tools are used, such as the profiles of the institutions and digital collections, the relationships with archive creators and their influence on the way pre-ingest and ingest workflows are organized.

Due to a large proliferation of preservation tools and varying definitions of what is considered a preservation tool, for the purpose of this survey and report 'a tool' was not narrowly defined. So, a tool could range from a script, software or wrapper, to a digital preservation system.

Focus on preservation tools for pre-ingest and ingest

Preservation tools play an important role in the automated processing of digital collections. Without automation, processes such as file identification, validation, deduplication, or encryption detection are cumbersome, and with large collections practically impossible to manage. These processes are crucial for the collecting institutions to gather sufficient information about their collections in order to ensure long-term preservation. Therefore, it is important to gain more insight into the use of preservation tools and how they differ between institutions.

The stages of pre-ingest and ingest are labour-intensive and require large amounts of time and human resources. According to research about the cost of long-term digital preservation,[5] the early stages of the preservation workflows such as selection, ingest, and processing are some of the most expensive. This research also points to the fact that staff costs related to the processing stage can amount to nearly 42 percent of all staff costs. One of the conclusions seems to suggest that better preparation of collections could lead to cost reduction at this stage. However, as will be further

[5] H. Uffen, B.A. Wiendels,T. Kinkel, C. Groeneveld, E.J. Krupe, Onderzoek naar de kosten digitale duurzaamheid, BMC onderzoek, Nationale Coalitie Digitale Duurzaamheid (NCDD), 2017 (online at: http://www.ncdd.nl/wp-content/uploads/2016/02/20170421-DE-Houdbaar-rapport-kostprijsmodel-duurzame-toegang.pdf, last accessed 8 October 2019).

discussed in relation to this research, as many as 70 percent of interviewed institutions do not have any influence on how the collections are deposited. In this context it might not be possible to reduce costs in this way. While it is certainly relevant for some organizations, for many others it remains out of reach, and therefore other opportunities in addressing this issue should also be explored. Could Dutch heritage organizations benefit from improvements in the use of preservation tools, which could potentially diminish the workload at the early stages of digital preservation? What could those improvements be and what role can the Dutch Heritage Network play in supporting organizations in managing pre-ingest and ingest more effectively? This survey focused on outlining the way in which preservation tools are used to identify aspects where such improvements could be made.

While in recent years both international[6] and national[7] organizations have conducted various projects mapping out preservation tools, there is no comprehensive overview of the tools used within Dutch heritage institutions. Such an overview is necessary for defining the next steps which the Dutch Heritage Network can take to stimulate a more effective organization of pre-ingest and ingest and thus reduce the waste of time and resources. A more targeted use of tools could also enable institutions to process complex or unusual collections that have been delayed, because they do not have the scope within existing workflows and would require additional preservation tools. At the same time, for smaller organizations, gaining more knowledge and introducing tools in their workflow could enlarge their capacity to process digital collections generally.

Why pre-ingest and ingest?

Pre-ingest is not a part of 'end-to-end' solutions in digital preservation. Neither is it formally described as a part of the Open Archival Information System (OAIS), but in practice it does take place, which leaves it open to interpretation. How pre-ingest and ingest are defined can differ substantially between heritage institutions. For some (including research repositories, and broadcasting archives), pre-ingest is a part of a service provision agreement; for others, handling different types of collections can depend on the relationship with archive creators, or on a specific type of collection. A single institution may have varied approaches to different types of collections that can result in divergent workflows and use of tools. In most cases[8], however, digital collections are not supplied to the heritage organizations ready for direct ingest into a digital repository without any form of preprocessing. This is mostly due to the large variety of digital objects and the lack of standardization among archive creators (especially in the case of private collections). That is why pre-ingest is vital for ensuring sustainable long-term preservation. However, it requires either a lot of time and manpower or an automated solution and the know-how needed to implement and supervise it. Today for many Dutch institutions using tools to process collections prior to ingesting is both a necessity and a challenge. Looking into the future we can only expect that the complexity of this challenge will increase with the growing diversity of digital objects, increasing file sizes and the growing amount of information generally. Hence, it is important to investigate how pre-ingest and

[6] Important initiatives are listed in the Digital Preservation Handbook of Digital Preservation Coalition (DPC): Community Owned digital Preservation Tool Registry COPTR, AV Preserve tools list, Digital Curation Centre (DCC) tools and services list, DCH-RP registry, Inventory of FLOSS (Free/libre open-source software) in the cultural heritage domain, Library of Congress NDIIPP tools showcase, Preserving digital Objects With Restricted Resources (POWRR) Tool Grid.

[7] One of the most relevant projects recently conducted in the Netherlands was the *Digitaal Atelier* project of the National Archive (NA) and the Regional Historical Centres (RHCs). The project concluded in 2018. It resulted with a Program of Requirements describing what is needed to implement services for processing digital archives at RHC's and led to agreements between them and preservation tool providers on adapting selected tools for use through a shared facility.

[8] This can be as much as 73 percent of the cases, however, that includes cases in which institutions report to have partial influence on how collections are supplied.

ingest can be conducted more effectively and efficiently, so that preservation tools can provide crucial support, not only for large institutions dealing with extensive collections, but also for smaller organizations that lack manpower to work manually despite the relatively small scale of their digital collections.

This research creates an inventory of the tools used for different types of digital objects within digital archives, museum collections and research repositories in the Netherlands, but also other forms of information stored digitally, such as information about collections. In this research I investigated specific conditions within institutions, including differences and similarities in approaches, for instance towards collecting and ingest. The goal of studying these differences was to find areas where looking beyond individual domains is possible and to learn how to provide both knowledge and support for the use of tools at a national level.

Survey

In 2018, in a survey 'Survey Digitale Archieven in Nederland' investigating the facilities for storing digital collections in the Netherlands, twenty-seven institutions reported that they have a digital archive or repository.[9] In this survey, conducted between March and April 2019, these organizations were asked in detail about their digital collections and the way in which they deal with pre-ingest and ingest.

The survey was prepared in four steps. First, four interviews were conducted with organizations that later took part in the survey. The interviews served as a basis for identifying relevant questions and aspects to be researched. Secondly, a draft of the survey was prepared and refined based on the feedback from these four organizations. Thirdly, this draft survey was tested on a sample of four other respondents and filled in during a face-to-face meeting. This stage helped to formulate guidelines for other respondents, to clarify the terminology and unify language to make it more relevant for different types of organizations. Lastly the survey was shared digitally with all of the other institutions.

The survey was conducted in two parts consisting of an online questionnaire (Appendix 1) and a spreadsheet template (Appendix 2). The questionnaire gathered information about the collections, the way in which they are supplied, standards used, needs of organizations in terms of preservation tools and collaboration, and their current challenges. The matrix represents an inventory of tools and the actions for which they are used at different stages in the workflow.

Of the twenty-seven institutions, twenty-two responded to the questionnaire (81 percent), but not all of them filled in the matrix, finally delivering eighteen (67 percent) complete responses. Due to the relatively small sample, all available results were used in the final analysis. Hence, the difference in the total number of respondents between the two parts.

Questionnaire: twenty-two out of twenty-seven respondents (81 percent)

A questionnaire collected information about the following:

- type of digital objects in collection, now and in two years
- supply method for collections or archives
- the use of standards and their influence on the quality of supplied material
- definition of pre-ingest and ingest in practice
- workflows and workflow descriptions
- missing tools (content and function specific)
- challenges
- collaboration

[9] Joost van der Nat, Marcel Ras, 'Survey Digitale Archieven in Nederland', 2018. Available online at: https://www.netwerkdigitaalerfgoed.nl/wp-content/uploads/2018/05/20180507-Netwerk-Digitaal-Erfgoed-Survey-Digitale-Archieven.pdf

Matrix: 18/27 respondents (67 percent)

A matrix assembled information about the following:

- which tools are used?
- for which action and where in the process?
- are they in production?
- which tools are being tested?
- which tools are you interested in using?
- are they open source, freeware, proprietary or self-developed?
- what is the cost of the tools used?

Profile of organizations

The institutions participating in the survey can be categorized into seven different types: archive, audio-visual archive, library, university library, museum, knowledge institution, and other. About one third (27 percent) of the respondents represent archival institutions, 18 percent are knowledge institutions, 14 percent each are audio-visual archives, museums and other organizations, 9 percent are university libraries, and 4 percent public libraries.

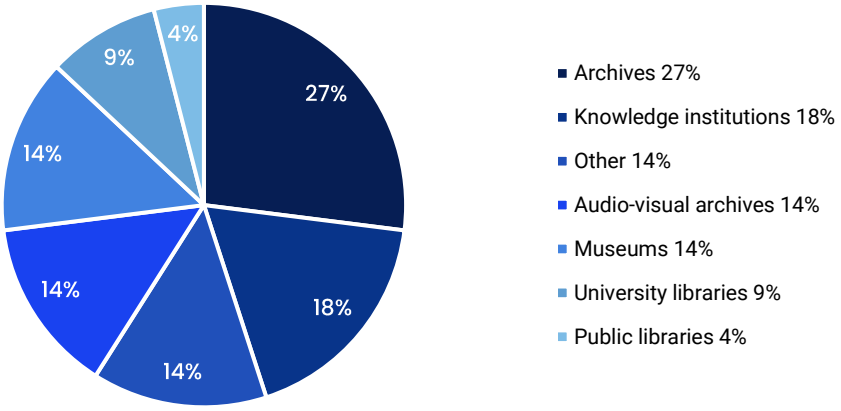

Figure 1: Types of organizations that participated in the survey

Profile of digital collections

While the survey 'Survey Digitale Archieven in Nederland'[10] investigated collection profiles in general, this survey looked at the profiles of their digital components. Understanding these profiles is important in the context of preservation tools, because some types of digital objects or file formats require specific tools to process them.

Currently the vast majority of digital collections comprise digital images (95 percent of respondents reported to have digital images in their collection), video (91 percent), text-based documents (91 percent) and audio (82 percent). More than half also includes spreadsheets (64 percent), databases (59 percent) and websites (55 percent).

[10] Ibid. v/d Nat, Ras

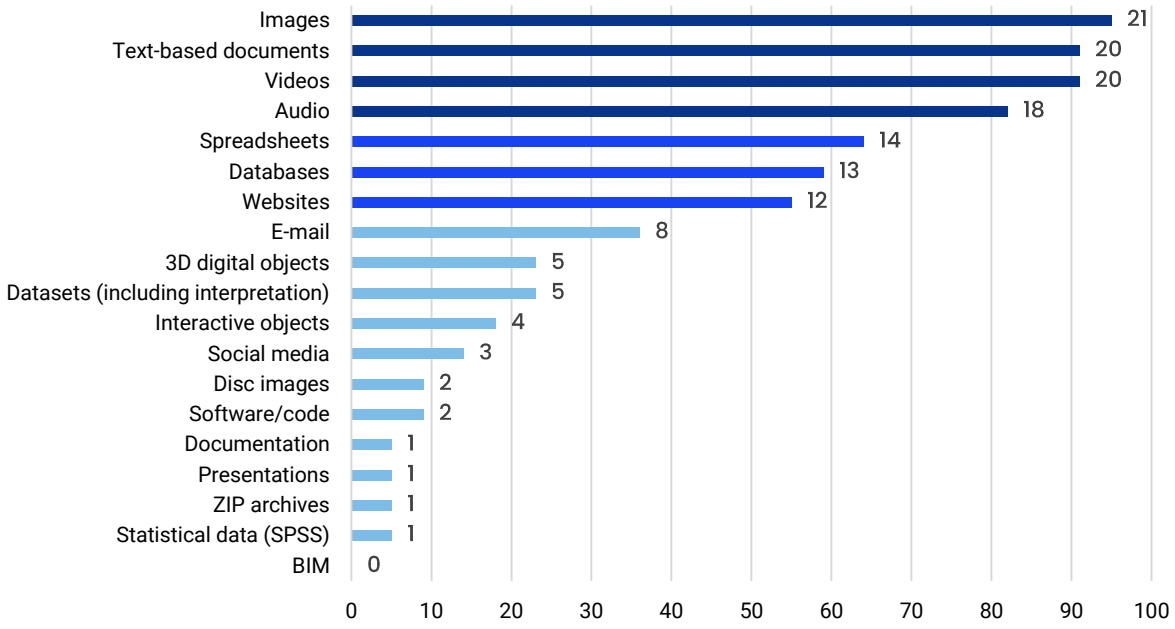

Figure 2: Types of digital objects currently included in collections. The numbers in the diagram refer to a total number of mentions by respondents, where this part of the survey returned a total of 22 responses. The horizontal scale shows the percentage of respondents who mentioned to have those types of objects in their collection.

Nearly half of the respondents (45 percent) does not expect major changes in their digital collection profile in the coming two years and expects to acquire more of the same types of digital objects as they have until now. **Over half of the respondents (55 percent), however, does expect to start collecting new types of objects not currently in their collection within the next two years.**

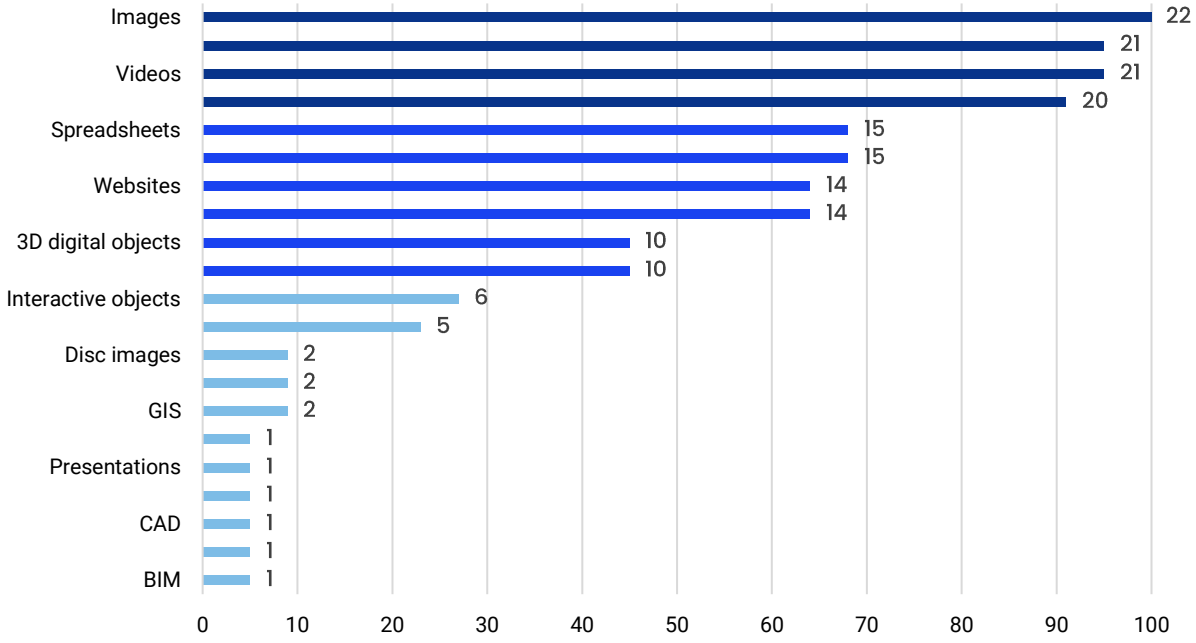

Figure 3: Types of digital objects expected in the collections in the coming two years. The numbers in the diagram refer to a total number of mentions by respondents, where this part of the survey returned a total of 22 responses. The horizontal scale shows the percentage of respondents who mentioned to have those types of objects in their collection.

The biggest increase is expected in the area of social media (from 14 percent to 45 percent), email (from 36 percent to 64 percent), three-dimensional digital objects (from 23 percent to 45 percent) and audio (from 82 percent to 91 percent). Furthermore it seems that more databases, websites and interactive objects are expected to appear in digital collections. Looking at new types of digital objects that have not yet been collected in the Netherlands, only Building Information Modelling (BIM) was mentioned as expected to be included in one collection in the coming two years.

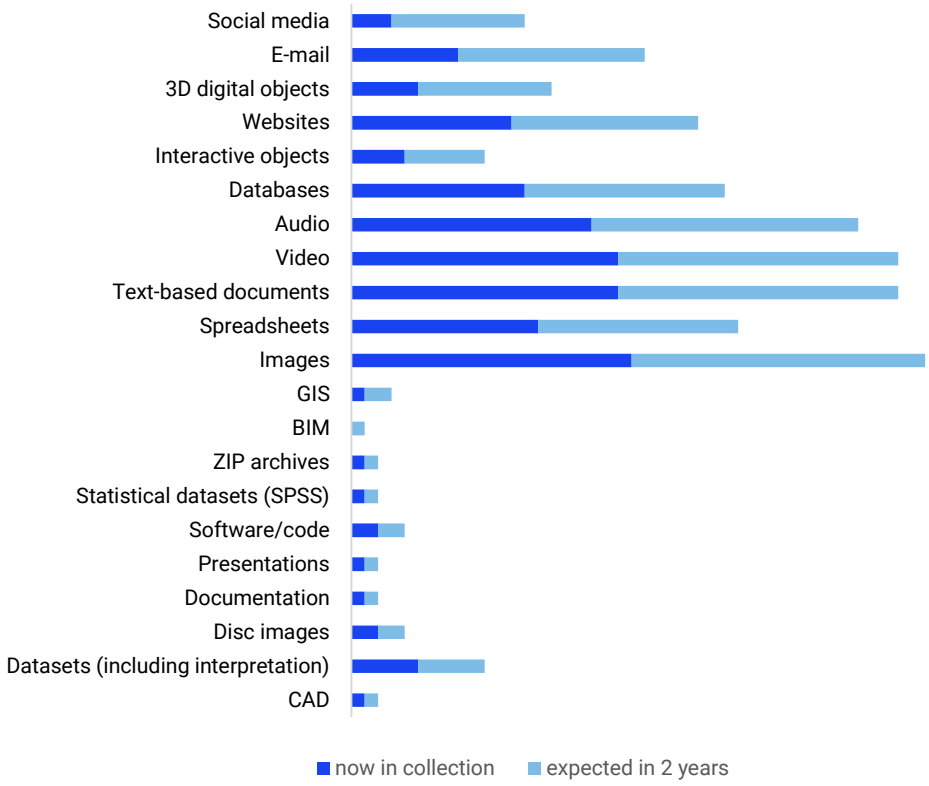

Figure 4: Comparison between digital objects in collections now and in two years' time

The relationship between archive creators, makers, suppliers and heritage institutions

The majority of the respondents receive their digital collections from external sources at least partially. Such sources can include archive creators (such as governmental institutions, other institutional or private actors), suppliers (for example publishers, broadcasting organizations or radio stations), makers (artists, researchers). In addition, collections include material that has been digitalized internally or externally on commission. Digitized material, however, does not seem to pose challenges for pre-ingest and ingest among the respondents, because it is created according to their requirements. Out of the twenty-two organizations, 64 percent receive digital material only from external sources, 27 percent receive collections both from external sources and digitalize internally, and 9 percent only has material that they digitalized themselves.

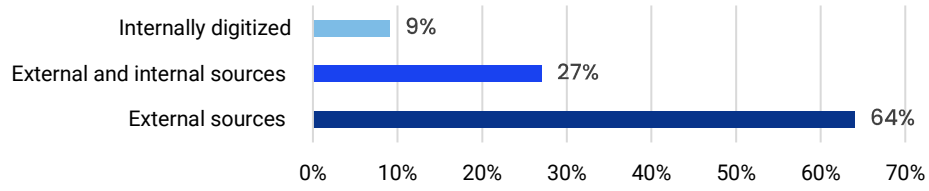

Figure 5: The provenance of digital collections.

Whether heritage institutions can set requirements is often related to the donors of their collections.

Only about a third of the interviewed heritage organizations can set requirements and therefore has influence on how the collections they receive are prepared and delivered. The ability to control how collections are prepared often diminishes the importance of pre-ingest. But as much as 64 percent of the respondents reported that they only have partial influence and cannot set hard requirements, and 9 percent reported that they are not in the position to set any requirements at all.

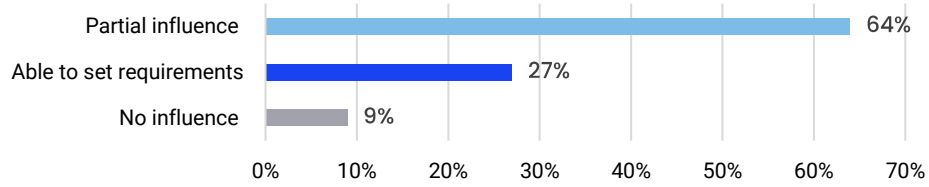

Figure 6: The ability of heritage institutions to set requirements for acquisition of digital collections

Of the different types of interviewed organizations, the knowledge institutions seem to have the biggest influence on how the archival material is delivered. As much as 75 percent of those organizations reported that they set hard requirements and would not accept material that does not meet their standards. Fifty percent of archives are also able to set such requirements. However, the vast majority of organizations including libraries, audio-visual archives, museums and the rest of the archives are often limited in their ability to do so. Especially those institutions that receive private collections or works of art are not in the position to reject what they are offered or to insist that depositors meet their requirements. They said that they make use of minimum requirements, wish lists or preferred formats, but it is not always possible to receive the material in a way that meets their standards. They reported that the way in which they receive collections largely depends on the archive creators and the kind of file types and formats they use. As a consequence, these institutions are forced to invest more time in preprocessing digital collections to prepare them for ingest and long-term preservation.

The use of standards

Slightly more than half of the respondents (54 percent) use more than one type of standard to ensure that their collections comply with requirements recognized outside of their own organization. **The standards they use are often a combination of legal standards and professional or industry standards** such as those used in the film industry or in archaeology. A third of the respondents use a single standard. In the majority of cases this would be an internally defined standard based on a combination of different standards. Fourteen percent reported that they do not use any standards. **However, it's important to note that among the respondents who did report using standards, many admitted that they are not always able to use them consistently.**

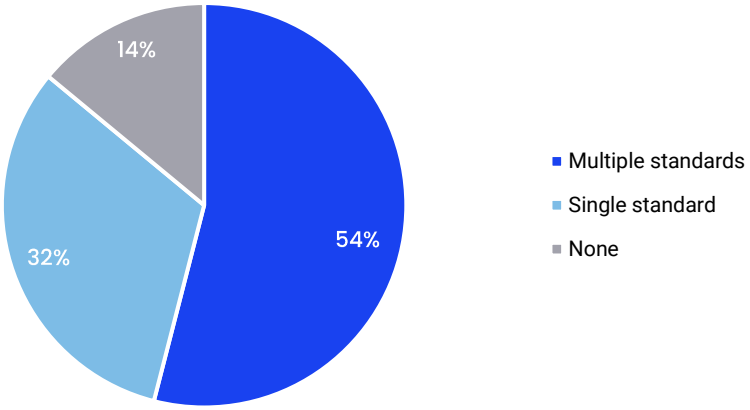

Figure 7: The use of standards among interviewed institutions

Since heritage institutions use many different standards or a combination thereof, I asked them to name the types of standards that are relevant to them. It is notable that 73 percent of organizations use internally defined standards. They reported that these are often based on standards defined by the industry or community, and/or the standards set out by national law or regulations. The graph below shows a breakdown of the different combinations of standards as currently used by heritage institutions.

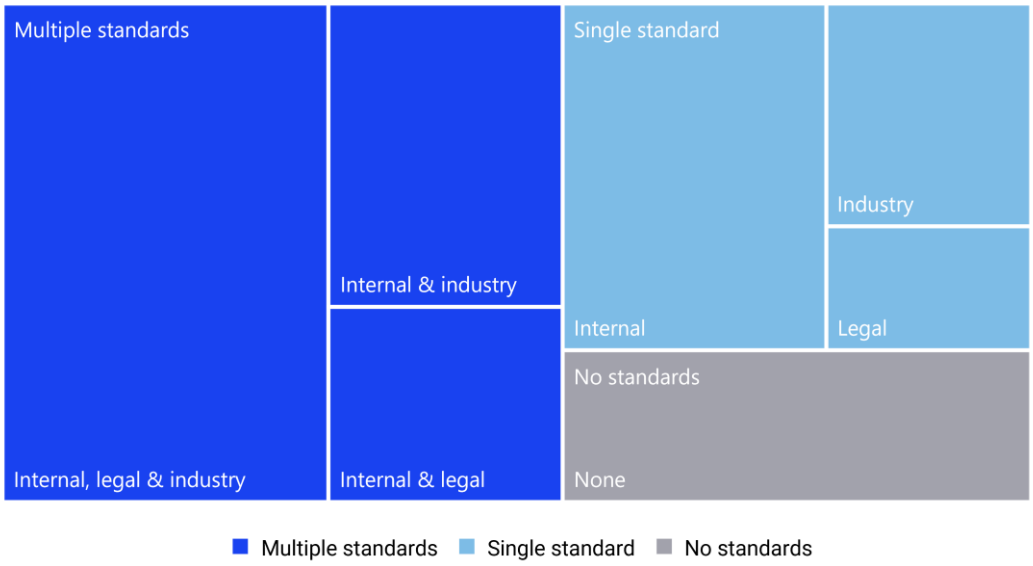

Figure 8: Breakdown of different combinations of standarts currently used as distributed among the respondents

Pre-ingest and ingest in practice

The vast majority (73 percent) of interviewed institutions make a distinction between pre-ingest and ingest even if they are not always defined as such.

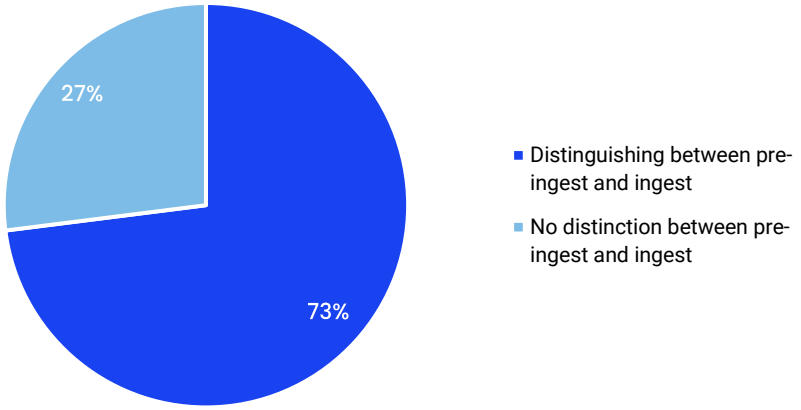

Figure 9: Organizations making a distinction between pre-ingest and ingest

The major differences between the way in which institutions define pre-ingest and ingest can be summarized in five groups of definitions provided by the respondents as described below:

1. Pre-ingest is a part of ingest.
2. The difference between pre-ingest and ingest is not clear.
3. Pre-ingest is a processing stage in a test environment, after which material is ingested into a production environment.
4. Pre-ingest leads to creation of Submission Information Packages (SIPs) that are subsequently ingested.
5. Pre-ingest is done by depositors, the deposit is equal to ingest.

Some institutions make no distinction between pre-ingest and ingest. For them pre-ingest is often seen as a part of ingest or it is not necessary. This might be because for some of them pre-ingest is seen as the responsibility of the depositor or archive creator. Those institutions receive material that is already highly standardized, which can be directly ingested.

For other institutions pre-ingest is a necessary step of preprocessing, which takes place after the deposit, but not all of them seem to agree which steps are part of pre-ingest, what its results are and where it takes place.

Workflows

Only 23 percent of the respondents are currently using workflow descriptions. **For the remaining 77 percent, workflow descriptions are not yet in use or they are in use only partially.** Out of those, 45 percent is currently working on developing one or more workflow descriptions, while for 32 percent it is either too early to create such descriptions or they have not been in use.

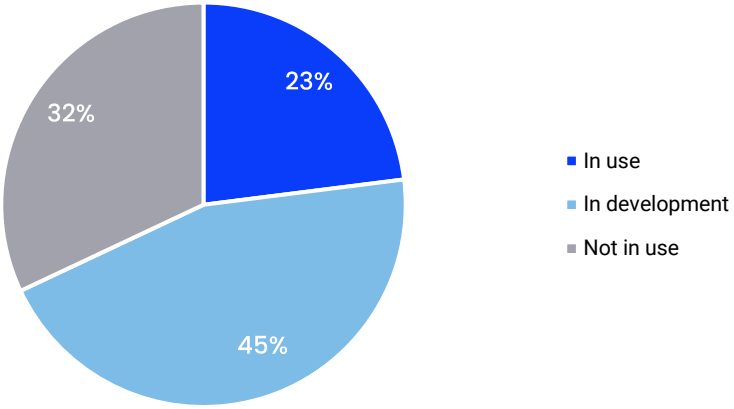

Figure 10: The use of workflow descriptions among the respondents

It seems that in the near future **45 percent of institutions aims to work with multiple workflow descriptions** for different types of collections (such as born-digital, digitized, internal, external), specific types of digital objects (AV, websites) or file formats, while **25 percent aims to work with one general workflow description.**

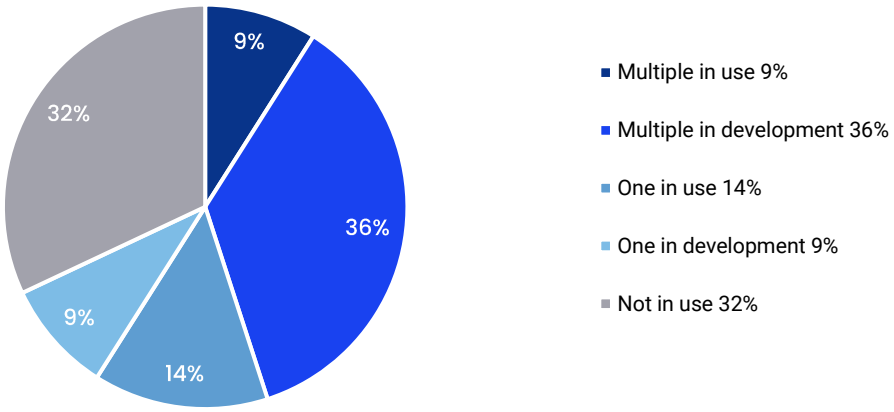

Figure 11: A detailed breakdown of workflow description development among the respondents

The use of tools

This research looked at the inventory of preservation tools for pre-ingest and ingest from two different points of view: what is currently being used and what is lacking.

Which tools are being used?

To create an overview of tools the respondents were asked to fill in a template where they could list which tools are currently in use at their organizations, for which parts of the collection, at which points in the process and for what type of action, the type of software license for those tools and their cost. In addition, the respondents were asked to fill in the same fields for tools that are not yet used in production and to include tools that they have been testing, developing or that they are interested in trying out.

Among the eighteen respondents who filled in this part of the survey, 111 tools were mentioned as currently used in production.[11] More than half of the organisations (56 percent) currently use between one and ten tools, 22 percent use between ten and twenty, and 11 percent use more than twenty, and 11 percent does not use any tools.

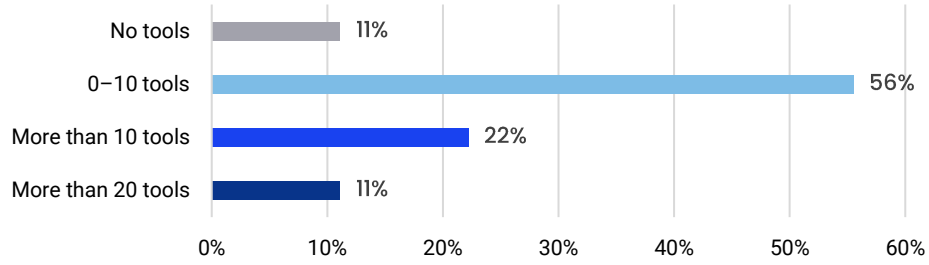

Figure 12: Number of tools used per organisation

Almost a third (28 percent) of the respondents is currently testing or running pilot programs for tool implementation, and most of them are doing this for several tools at a time. Also, more than half (56 percent) mentioned that they are interested in testing or implementing new tools. Among that group, the vast majority reported being interested in one to three tools. Together they mentioned twenty-three existing tools that they are likely to implement in the near future. Out of those tools, sixteen are currently being tested, while nine have been identified as potentially interesting.[12] Two organizations are looking for tools but have not yet identified which tools have the functionality they need. The scope of the functionality they mentioned is quite broad, but often concentrates on standard

[11] Some organizations said that they use more tools than what they reported in the survey, but that listing all of them would take too much time. Those organizations were asked to list the most important tools.

[12] There is a small overlap between the two numbers as two of the tools mentioned as interesting are already tested by other organizations.

functionalities such as file format identification and validation, technical metadata extraction, or normalization. It means that there is a rather large spectrum of organizations already using tools that meet those requirements, and that knowledge about them is available in the community. At the same time, over half of the tools that were listed as interesting are new to the community, with no precedent in use among other respondents. **It means that while there is a lot of room for organizations to help each other through knowledge exchange about the use of tools, there is an equally great need for organizations to experiment with using new tools.**

Tools pending implementation	Mentions	Users
Preservica	1	4
DROID	3	3
JHOVE	2	3
FITS	1	3
Ffmpeg	1	3
Archivematica	1	2
TopX-creator	1	2
Apache Tika	1	2
RM tool VHIC	1	1
IsoBuster	1	1
WebCurator	1	1
Not specified	2	0
Epadd	2	0
VeraPDF	2	0
Webrecorder	2	0
Bitcurator	1	0
Netarchivesuite	1	0
WARCit	1	0
FLAT	1	0
Libre Office	1	0
Image Histogram Comparison	1	0
SocialFeedManager	1	0
iromlab	1	0
JAI Image I/O Tools JRE Extension	1	0

Table 1: Tools to be implemented in the near future and/or currently being tested

What is the most common license for tools currently used?

Of the tools currently used in production, 41 percent is proprietary, 24 percent open source, 12 percent is self-developed, 10 percent is freeware, and 13 percent consists of other types of tools and scripts that are often not offered or licensed as standalone tools; they include mostly standard command line utilities and other small tools difficult to classify.

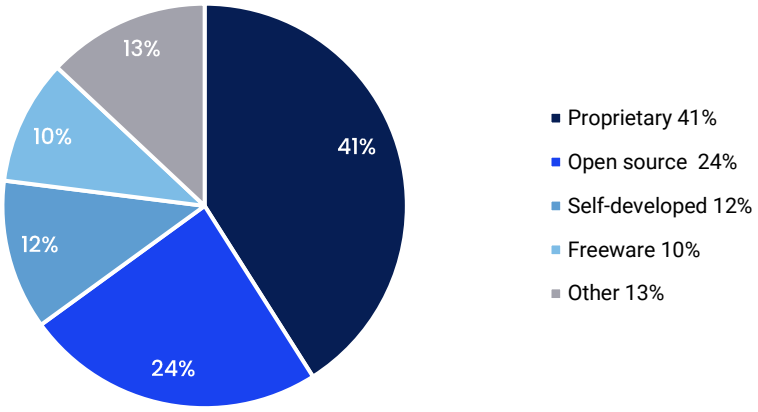

Figure 13: Breakdown of tools according to the type of software license

Among the tools that are not yet in production but have been mentioned as tested, developed or interesting for implementation, the majority (68 percent) is open source. Only 16 percent of tools mentioned as pending production is either proprietary or self-developed. This seems to indicate that the respondents are eager to use more open source software.

At which stage in the process are the tools being used?

It is challenging to describe a workflow that would be universal enough for a broad spectrum of organizations to identify with, but in order to understand and compare how Dutch heritage organizations are using tools, I decided to draft such a generic workflow framework. It is based on three steps distinguished by the relationship of the institution with the depositor: acquisition, transfer, and processing. The first step focuses on understanding the collection prior to acquisition. The second step describes different ways of transferring the Submission Information Package (SIP) or harvesting collection material, such as in the case of websites and social media. The last step brings together all actions that could take place after the transfer and before creating the Archival Information Package (AIP). Because of the focus of this survey on pre-ingest and ingest the access and further preservation aspects were not taken into account.

Each of the steps (acquisition, transfer, processing) was divided into a set of actions, which were derived from standard functionalities of popular digital preservation systems and preservation tools, a study of existing registries of tools such as Community Owned Digital Preservation Tool Registry (COPTR) and POWRR (Preserving digital Objects With Restricted Resources), as well as a close reading of different types of workflows available online via Community Owned Workflows (COW), Library Workflow Exchange and OSSArcFlow. This list was edited based on interviews and feedback from four organizations participating in the survey.

Because workflows differ greatly among organizations, respondents were provided with a detailed glossary (p. 3 of Appendix 1) of the steps and actions and were able to add their own actions within any of the three steps to indicate where in their workflow tools are used and what for.

Of the tools used in production the most are said to be used at the processing stage (38 percent), slightly fewer are used for transfer (34 percent) and the smallest number of tools is used at the acquisition stage (28 percent).

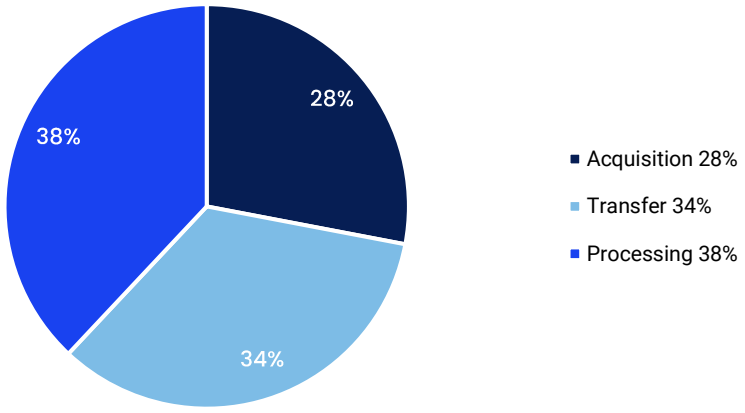

Figure 14: Tools as used at different steps in the workflow

Acquisition:
impact analysis, content visualization, SIP creation
Not all organizations are using tools at the acquisition stage. Even though 61 percent has mentioned that they use tools for acquisition it seems that the highest number of organizations (50 percent) using tools at this stage does so for SIP creation. A third uses tools for content visualization, 28 percent uses tools for other types of actions, which they identified as relevant to their own workflow, and 22 percent uses tools for impact analysis. The added actions include digitalization, making disc images, reading and checking disc images, capturing social media, web archiving, generating checksums and uploading files.

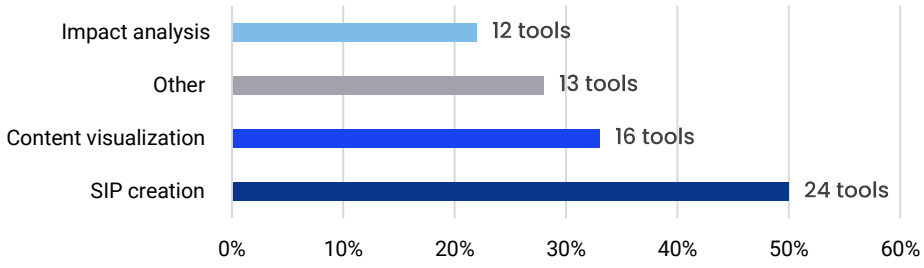

Figure 15: Percentage of organizations using tools for acquisition. The number of tools is mentioned in brackets. Please note that several organizations mentioned the same tool, so the total number of mentions is not equal to the unique number of tools used.

In total forty-six tools are currently used for acquisition with little overlap between organizations, and eight tools were mentioned as interesting or pending implementation. DROID, mentioned three times, made the top of the list for acquisition, followed by eleven tools that were mentioned only twice. The list includes diverse types of software ranging from end-to-end digital preservation systems to Adobe Premiere and MagicISO to MyGeodata Converter, which suggests that needs at this stage can vary a lot and depend on the type of material being acquired and the way in which material is deposited. (See full list on p. 1 of Appendix 3.)

Transfer:
online harvest, transfer, file extraction, compliance with requirements, fixity check, virus scan
Almost all of the interviewed organizations (89 percent) are using tools for transfer of files from depositors. Between 60 and 70 percent uses tools for fixity check, virus scan, and the actual transfer of digital-born material. About 30 to 40 percent stated that they use tools to check the material for compliance with their own requirements, to extract packaged contents and for the online harvest of content such as websites.[13] In addition, 17 percent uses tools for steps that were added by respondents including malware check, metadata transfer, and monitoring changes on websites that are part of the collection.

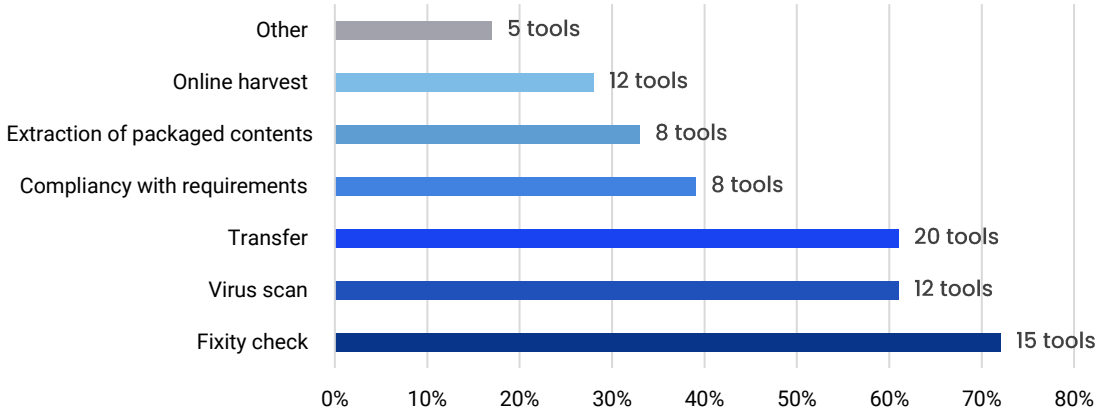

Figure 16: Percentage of organizations using tools for transfer. The number of tools is mentioned in brackets. Please note that several organizations mentioned the same tool, so the total number of mentions is not equal to the unique number of tools used.

In total fifty-five tools are said to be used at this stage, and three tools are mentioned as interesting or pending implementation. The tool mentioned most frequently is Archivematica (ten mentions for different actions by three organizations); TopX Creator had seven mentions by two organizations and individually scripted solutions had seven mentions by three organizations. They were followed by Preservica (with five mentions by three organizations) and MD5sum (four mentions by four organizations). (See full list on p. 2 of Appendix 3.)

[13] Some organizations saw this step as a part of their acquisition workflow, this is why this step appears both in acquisition and transfer.

Processing:
normalization and conversion of files, extraction of technical metadata, file format identification, and validation

Almost all of the interviewed organizations (83 percent) use tools for steps related to processing files at the level of pre-ingest and ingest. About 70 percent uses tools for the normalization or conversion of files, extraction of technical metadata, file format identification, and validation. Half of them reported that they use tools to create AIPs, and half said that they use tools for creating SIPs. Similarly, half of the respondents use tools to enrich metadata prior to ingest and to perform quality control. It's interesting that the step reported by two organizations (11 percent) is deduplication, and that only three organizations (17 percent) mentioned encryption detection and dependency analysis as part of their workflow.

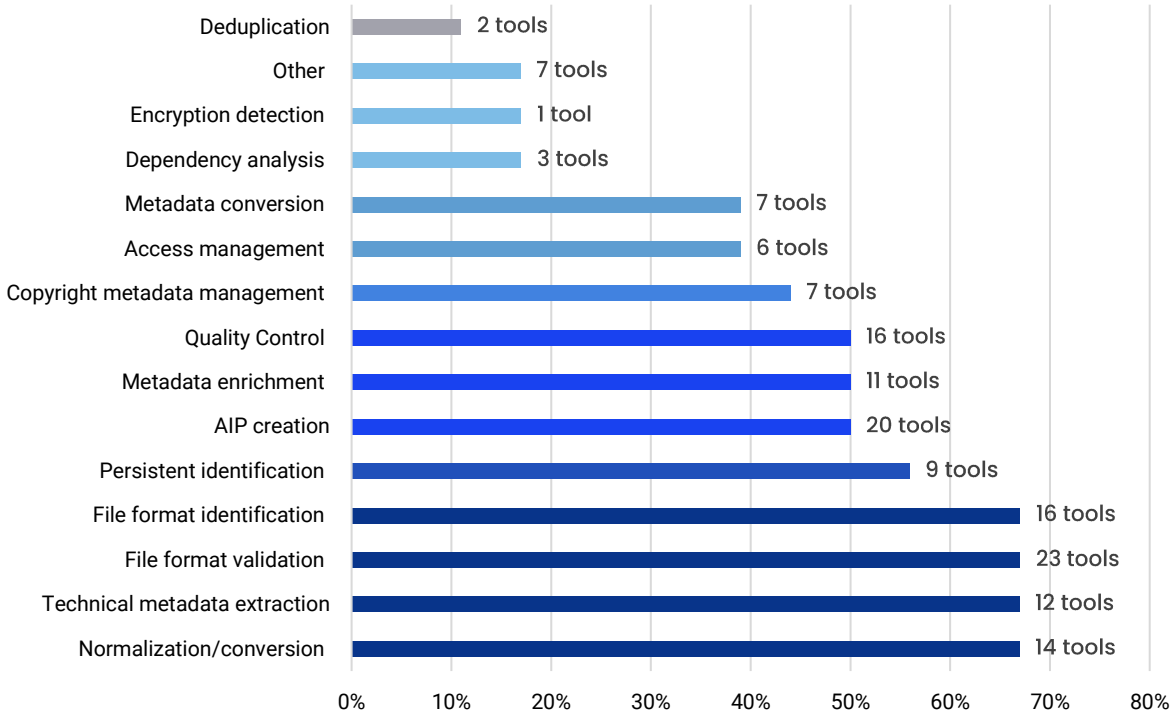

Figure 17: Percentage of organizations using tools for processing. The number of tools is mentioned in brackets. Please note that several organizations mentioned the same tool, so the total number of mentions is not equal to the unique number of tools used.

Looking more closely at the status of tool implementation there seems to be a significant interest in testing or implementing tools related to file format identification (eight organizations indicated that they are looking at six tools) and validation (seven organizations mention six tools). The two lists largely overlap and include DROID, Apache Tika, Archivematica, TopX-creator, FITS, JHOVE and VeraPDF.[14] Several organizations have also shown interest in tools for normalization and conversion (four organizations mention five tools) such as Archivematica, Ffmpeg, JAI Image I/O JRE Extension, Libre Office, and WARCit.

In total sixty-one tools were reported as used in production for processing digital files during pre-ingest and ingest, and five tools were mentioned as interesting or pending implementation. The top

[14] The list also included FLAT, a self-developed digital preservation system pending implementation at one of the interviewed organizations. Because it is of low to no relevance for other organizations this tool was omitted.

mentioned tools are customized internal scripts (eighteen mentions by two organizations)[15], Archivematica (seventeen mentions by three organizations), Preservica and TopX Creator both with eleven mentions by three and two users respectively, and Islandora (nine mentions by one organization). As noted before, two of the respondents are interested in finding several tools for processing digital collections with various functionalities (fourteen mentions).

Apart from the end-to-end solutions mentioned above the most mentioned tools are: JHOVE, Ffmpeg, FITS, MediaInfo, DROID, Apache Tika, and RM tool VHIC. (See full list on p. 3 of Appendix 3.)

What kinds of tools are missing?

To understand which tools they currently lack, I asked the respondents to identify if there are any specific types of content for which they need tools or if there are specific functional aspects for which they do not yet have tools in use. It seems that for a lot of organizations it is easier to say which functional aspects they lack, but for a large part it is still difficult to define what exactly they need. When asked for what types of content or file formats they lack tools, only 41 percent gave clear answers, while 59 percent did not give any answer or they indicated that it's too early for them for formulate such specific requirements.

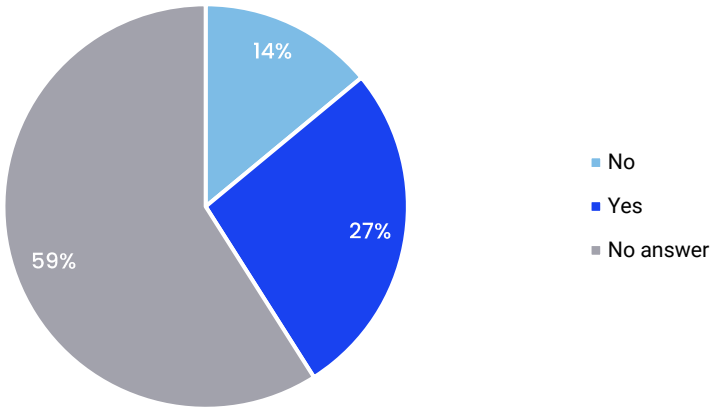

Figure 18: Are you missing tools for processing of specific digital objects?

Those organizations that did indicate types of content for which they do not have solutions gave the following answers (all were mentioned only once): databases, websites, video, social media, Microsoft Office files (conversion), software-based artworks, xml (validation), proprietary file formats (validation), e-mail (validation). This indicates that **while some of the answers relate to archiving specific types of digital objects, multiple point to the validation of specific file formats.** This corresponds with the answers obtained from the matrix where the most sought-after tools are for file format identification and validation.

[15] It seems that interviewed research repositories make use mostly of customized solutions. All four scientific research repositories, which took part in the survey have a self- developed digital repository system and make use a number of smaller internally developed solutions that work with those systems. However, two other organizations mentioned that they use individual scripted solutions at other stages of the process.

The situation looked different when organizations were asked to identify functional aspects they lack. Here 64 percent gave clear answers, out of which 50 percent indicated that they do miss function-specific tools.

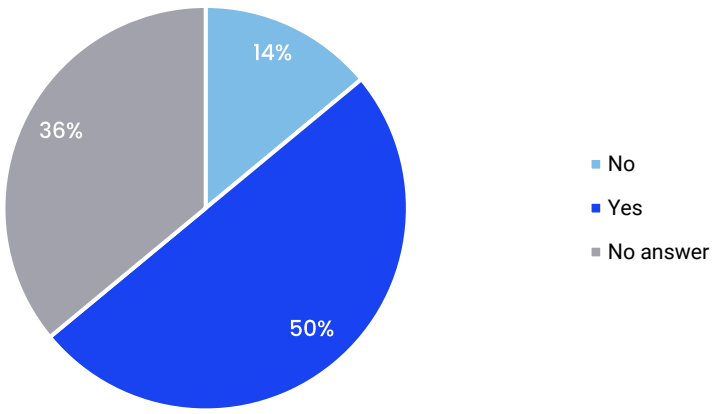

Figure 19: Are you missing tools for specific types of content in your workflow?

Looking at the breakdown of the functions, **technical metadata extraction and file format validation seem to be most sought for.**

Along with that, deduplication, encryption detection, file format identification and content management system (CMS) related improvements/development were mentioned more than once. Compared with the other part of the survey, deduplication and encryption detection were actions with the fewest active users, which means that they are likely to become a point of attention for the follow-up research.

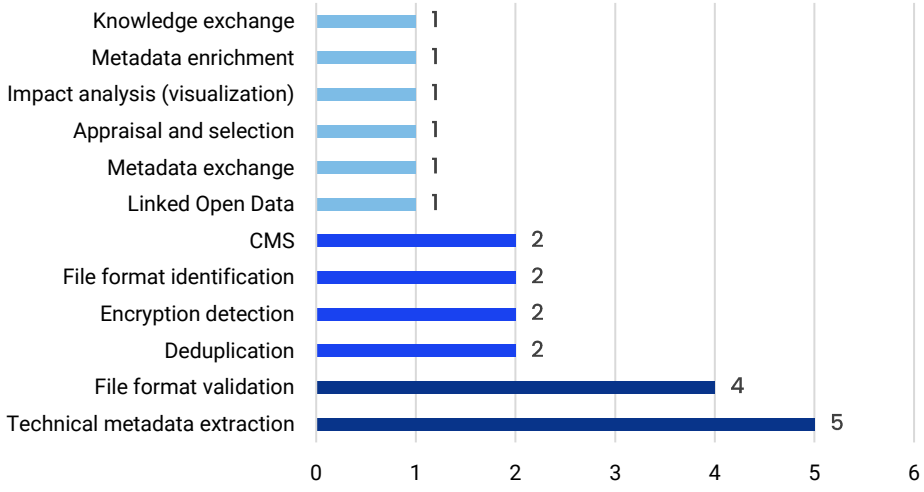

Figure 20: For which functions in your workflow are you missing tools?

Outlook

Challenges

The respondents were asked to identify the current challenges facing organizations concerning the automation of pre-ingest and ingest. The answers to this question show that Dutch heritage organizations have to deal with many different kinds of tasks that pose difficulties at various levels.

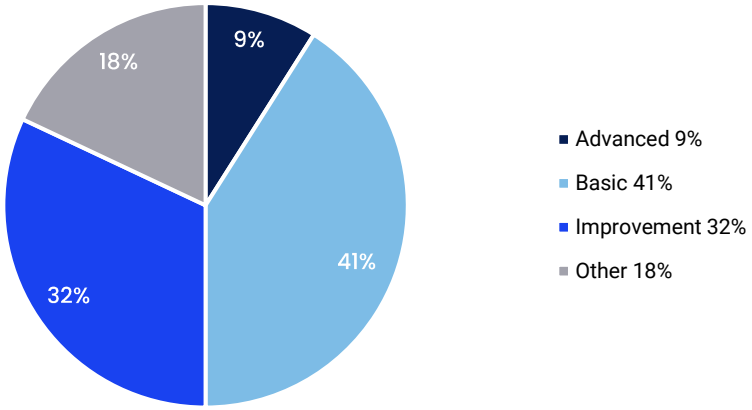

Figure 21: What are the challenges your organization is facing at the moment?

For many of them a challenge is to develop and implement digital preservation systems, develop interfaces, implement tools and start to automate. Forty-one percent of respondents fall in this group. Slightly more than 30 percent of organizations seem to be focusing on improving the automation they have in place either in terms of scaling up or speeding up. Many of them mention they would like to process more material at a higher rate. In this group it is also mentioned that difficulties include **dealing with non-standard file formats and legacy data not conforming with current standards.** For about 20 percent the challenges mentioned are on a rather advanced level, such as chain automation and quality control, and for about 10 percent the challenge lies somewhere else for instance in developing policies, workflows or communication regarding requirements that allow better ingest.

Need for collaboration

Regarding the need for collaboration the respondents agree that best practice is the most sought-after form of collaboration and knowledge exchange. Seventy-three percent mentions that this is how they would like to collaborate with others, while all others (36 percent) mention forms of collaboration including instructions for the use of tools, manuals, interpretation of errors, the development of new tools and other forms relevant to them.

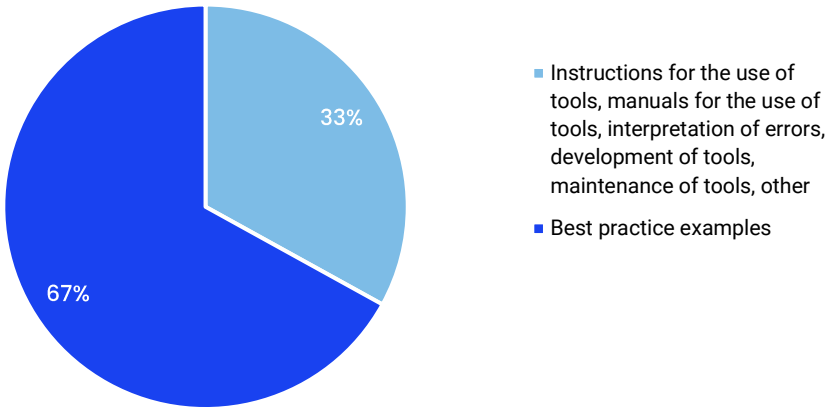

Figure 22: Where would you like to collaborate with others in the area of preservation tools?

In their suggestions, respondents mentioned the following points: knowledge exchange including specialist knowledge on specific types of digital objects, educating employees, the testing and prototyping of new tools, the use of the digital preservation system in general, data curation and expertise, metadata mapping, national storage and the maintenance of software.

Costs

As a part of the spreadsheet template, the respondents were asked to fill in the estimated cost of the tools they are using. While this part was mostly left blank by respondents, it seems clear that organizations find it difficult to determine the structural cost of digital preservation. Many of them mentioned that it's difficult to establish costs for tools that have been self-developed, for both their creation and maintenance. But for other tools, too, it was not possible to determine how much staff time is required to use and maintain tools, which also translates to developing knowledge on how to use tools generally and on addressing errors.

Conclusions

In order to use tools in a way that can be domain-transcending and relevant for organizations of various types, it is helpful to first understand the differences in needs and obstacles confronting heritage institutions, as well as the areas where they can help one another. How can the Dutch Heritage Network support individual organizations in the further development of automation of pre-ingest and ingest? And which tools can play a role in this in a sustainable way? These are questions raised by this research. Continuity and systematic maintenance are crucial for the sustainability of a tool. It is therefore vital that important tools are properly maintained, developed and supported. The outcomes show that there is a relevant overlap in the use of tools, which could be important for knowledge exchange between institutions. As part of a larger framework this research also contributes to the development of a national map of digital preservation facilities, which as a tool can support the coordination of assets and needs present in the network, including those related to the use of tools by different organizations. The type of information that could be shared and would be particularly helpful would have to be determined at a later stage.

The survey and its analysis lead to the identification of various needs and gaps in the use of tools for pre-ingest and ingest.

Needs include:

- tools for file format identification and validation, especially in relation to new file formats and digital objects that are expected in many collections in the coming two years (including email, websites, social media and three-dimensional digital objects);
- tools for normalization and conversion;
- tools for technical metadata extraction;

Gaps:

- Detected gaps in the use of tools for acquisition process: deduplication, encryption detection and dependency analysis.

It is worth noting that quite a few organizations use custom-made preservation tools and but a few off-the-shelf products. That seems to be driven by the fact that many of them rely on custom-built digital preservation systems. Nevertheless, it could be interesting to investigate whether some of those solutions could be relevant to other organizations.

During the following steps in the research within the currently running project of the Dutch Heritage Network on preservation tools (*Preserveringstools*), I will further analyse the areas of attention and related tools. This will lead to a creation of a roadmap for further development in selected area(s) of preservation tools for pre-ingest and ingest, which could include tool adoption, further development, popularization, or advice to organizations regarding knowledge exchange and development. For knowledge exchange, the next steps could focus on identifying and matching users with little expertise with those who have affinity with certain tools to stimulate knowledge exchange within the network, for example through the *gereedschapskist* (toolbox). The Dutch Heritage Network could also work towards popularizing knowledge about existing tools and best practice in the identified

areas, so that organizations could make more informed choices about tools best suited for their collections and workflows (for example through an already existing course in digital preservation, *Leren Preserveren,* or tailored workshops). The Dutch Heritage Network could also focus on further development or adoption for a specific tool or tools either within the network (following the example of the Open Preservation Foundation) or externally in collaboration with a supplier.

Colofon

I would like to thank:

Joost van der Nat, Marcel Ras (NDE)
Sam Alloing (The Royal Library)
Pepijn Lucker, Jacob Takema, Remco van Veenendaal (The National Archive)
Walter Swagemakers, Ernst van Velzen (Eye Film Museum)
Erwin Verbruggen (Sound and Vision)

All the respondents who kindly shared their knowledge in the context of this research.
And Heleen Schröder for copy editing.

About this publication

This report was published by the Dutch Digital Heritage Network (NDE) in May 2020. For further information, see: www.netwerkdigitaalerfgoed.nl/en/. If you have any queries or comments about the contents of this article, please feel free to email them to: info@netwerkdigitaalerfgoed.nl.

dutch digital heritage network