

Accession of Unstructured File Collections

Alexander Herschung, Marit Kleinmanns

startext GmbH, Dottendorfer Straße 86, 53129 Bonn, Germany
Alexander.Herschung@startext.de, Marit.Kleinmanns@startext.de

Abstract. Unstructured file collections will keep archives increasingly busy, as the digitisation of the workplace as well as the private correspondence proceeds. But the archiving of this kind of data requires a lot of time and personnel. startext COMO is a software tool that helps to organize and evaluate unstructured file collections.

Keywords: Digitisation, Digital Long Term Preservation, Unstructured File Collections.

1 Introduction

The transfer of file collections poses a new and partly unknown challenge for the archival work. There are solutions for the adoption of structured digital content, but what about unstructured file collections? With the digitisation of everyday life both job-related and personal, unstructured file collections will keep the archives increasingly busy in the near future. Whether it is the data collection of an administrator or the image collection of a photographer – the material is seldom structured and arranged in proper order for prompt accession into the archive. The archiving of these unstructured file collections does not only take time and personnel to first assort and then group similar or discard individual files, it is also prevalently impeded by a significant amount of data and by only fragmentary metadata.

It is a challenge to evaluate a digital file collection and then create Archival Information Packages (AIPs) from unstructured deposits. While paper files can be quickly examined by hand by a practiced person – and thus entire file accessions can be handled relatively fast and without great effort – the handling of digital content requires the help of suitable tools.

With regard to this difficulty, the idea was to engineer a simple and effective solution. startext COMO is a transfer editor which is platform-independent and can be operated without installation. It is even possible to start the tool directly from a USB flash drive. A flexible application is thus guaranteed for all working environments. Additionally, startext COMO is intuitive to use, inasmuch it is primarily intended for an easy access to the topic of digital archiving.

2 Related Works

The topic of structuring unstructured data files is currently being actively discussed in administrations and companies: How should such collections be handled and are there possible measures to structure and sort this collection of data already in their formation? Also a structural containment of the data masses before their accumulation is discussed. Archives are well aware of this issue and are interested in a professional solution to the problem.

In this context, an informal user-group was formed in 2011 and consists of archives using HP and SER as a service provider in Germany. The main objective of this group, whose members are e.g. the Bundesarchiv (Federal Archives), the Landesarchiv NRW (State Archive North Rhine-Westphalia) and the Historisches Archiv der Stadt Köln (Historical Archive of the City of Cologne), is, among other things, to increase communications and the creation of technical and economic synergy effects in system development (Schmidt, 2015). In cooperation, this user-group developed a Pre-Ingest Tool (PIT), which is designed to solve the problem of structuring and appraising unstructured file collections. This tool, however, is only accessible to members of the HP/SER user-group.

To address this issue, startext has decided to develop an own tool, which can be used by any archive without great financial or human resources and without belonging to a specific user group or using a specific archival software.

3 About startext Company

startext is a German software company, located in the city Bonn, established in 1980. startext develops software mostly for cultural heritage organizations as archives and museums. The company produces software products, such as inventory software for archives and museums and a full featured digital long term preservation software, compliant to the international standard model OAIS (Open Archival Information System).

startexts digital archive software covers all aspects of OAIS, is delivered with a ready-to-use configuration of ingest (including virus-scan, format recognition, -validation and -migration) and targets also small and medium sized organizations for allowing them to start with professional digital archiving at a reasonable price.

The company acts also as a partner for customer-specific software development projects. One example: On March 3rd, 2009 the municipal archive of Cologne was completely destroyed when the whole building collapsed into an opening hole in the ground. About 40km of archival goods were affected, damaged, destroyed. Within two weeks startext developed a new software to support the salvage operation. This software has since then been continuously enhanced – according to the changing requirements during the salvation process. Today it covers the full management and documentation of all efforts spent on restoring the damaged archival goods.

4 Regarding the Problem of Unstructured Data

Many documents once written on paper are now saved digitally: correspondence – memos – reports – (research) drafts; just to name a few. With the immense growth of the internet and to go along with this a massive amount of digital data the issue of archiving unstructured digital data becomes more and more prevalent.

Paper documents are filed into - usually labelled - binders and thus can be evaluated quickly and without difficulties. In contrast, the digital system of saving documents is often much more complex. If the files are not named elaborately and not collected in precisely named folders, these data sets are very laborious to assess as the archivist has to open each file and skim through the content.

The aim of startext was to develop an easy-to-use software (see Fig. 1.) which would simplify the accession of unstructured file collections immensely. And this is exactly what startext COMO does!

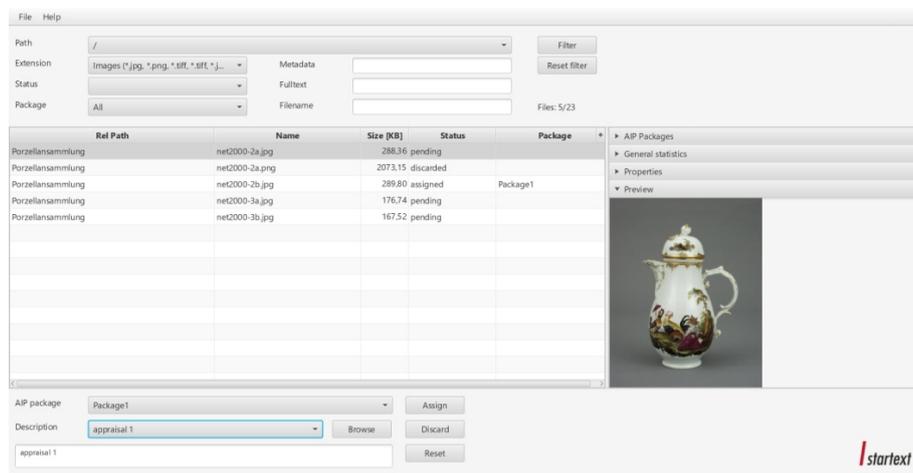


Fig. 1. startext COMO user interface with imported data

4.1 From Accession to the Formation of the Transfer Packages

When it comes to structuring and sorting the data, the question is on which parameters are the contents filtered by startext COMO.

There are two stages where the software helps with selecting dispensable data from files that should be archived: First, the software enables a breakdown of the structure of the file collections into easy to collocate components, e.g. folder structure, (technical) metadata or full text. Thereafter the files can be filtered and sorted according to multiple parameters, e.g. path, status, package allocation or file name components. Based on the selections with these parameters, all files can be sorted and assigned to a package or be discarded (see Fig. 2.). startext COMO then generates transfer packages in form of .zip-files that contain the structure of the folders, the files and in addition a

METS- XML file (METS = Metadata Encoding & Transmission Standard) which contains a protocol of all folders/files (this includes the discarded ones, too) and of the evaluation process.

The software follows a structure of six simple steps to structure, sequence, and value the files:

1. Import of primary data
2. Extraction of technical metadata
3. Creation of transfer packages
4. Sort and filter
5. Appraisal decision and allocation to a transfer package
6. Creation of transfer packages

Rel Path	Name	Size [KB]	Status	Package
Porzellansammlung	net2000-2a.jpg	286.36	pending	
Porzellansammlung	net2000-2a.png	2073.15	discarded	
Porzellansammlung	net2000-2b.gif	353.92	discarded	
Porzellansammlung	net2000-2b.jpg	289.80	assigned	Package1
Porzellansammlung	net2000-3a.jpg	176.74	assigned	Package1
Porzellansammlung	net2000-3b.jpg	167.52	assigned	Package2
startext-ACTApro-Information\ACTAproDokumenta...	Anmerkungen zur Nutzung der Exce...	286.81	assigned	Package2
startext-ACTApro-Information\ACTAproDokumenta...	Dokumentation ACTApro Benutzun...	1597.68	pending	
startext-ACTApro-Information\ACTAproDokumenta...	Dokumentation Desk Magazin.pdf	2823.78	pending	
startext-ACTApro-Information\ACTAproDokumenta...	Schnelleinstieg ACTApro Desk.pdf	294.89	pending	
startext-ACTApro-Information\ACTAproDokumenta...	Schnelleinstieg ACTApro Magazin.pdf	332.51	pending	
startext-ACTApro-Information\ACTAproDokumenta...	Dokumentation ACTAproBatch.doc	898.50	pending	
startext-ACTApro-Information\ACTAproDokumenta...	Dokumentation ACTAproBatch.pdf	762.78	pending	
startext-ACTApro-Information\ACTAproDokumenta...	Dokumentation FormDesigner.docx	881.41	pending	

AIP package: Assign Discard Reset
 Description: Browse

Fig. 2. Index of files during the evaluation process.

Import of primary data

While developing this tool, it seemed important not to modify, move, or delete the original files, before the evaluation process is finished. For this reason, the original data are not imported into the tool. Rather, startext COMO creates a search index. By this, there is no risk of damage or loss of data. All decisions can be altered at any time. All changes made are recorded in the mentioned METS-XML file.

This approach also ensures a fast performance and very low memory usage; thereby there is no size limit on the file collection or the number of files it contains.

Within the tool, a folder can be selected and the individual files can be displayed directly into the tool. The individual files are displayed in a table view and can be sorted according to several criteria to ensure faster processing. The files can be opened directly

from the file store. While importing the files into the tool, an MD5 check sum is calculated for each file. By using an MD5 check sum, you can ensure that the files to be archived correspond to the delivered files. This check sum is matched in the last processing step of the tool when creating the individual transfer packages.

However, it is possible to open the individual files at this point to allow direct access to the contents. A preview image provides an initial insight. So there is the possibility of a first assessment decision. The preview image helps also to select which of the files need to be opened for more extensive inspections.

Extraction of technical metadata

A read-out of the technical metadata is indispensable for further processing and archiving. The technical metadata can help with the evaluation, if the files are to be archived or not and serve later on as an aid for the deeper archival description. That way, technical metadata can be used to determine the date or the file creator. startext COMO is also able to find duplicates of files and different versions of the same file.

In addition, storing the technical metadata is essential for compliance with the OAIS paradigm. Technical metadata form the basis for the data transfer into a long-term archive (ingest) and a later maintenance planning with format migration or emulation (preservation planning).

For indexing the technical metadata, startext COMO uses the search engine technology Apache Lucene, which is widely recognized and works very efficiently.

Creation of transfer packages

Prior to the actual processing, individual transfer packages, the SIPs (Submission Information Package), have to be defined in the tool. These SIPs form the basis for the so called AIPs (Archival Information Package) when the packages are subsequently transferred to an archive information system (AIS).

At this point, the titles of the SIPs can already be defined and later transferred to an archive information system.

Sort and filter

For a better overview, the files can be filtered according to certain criteria. For example, all files of a particular extension can be selected - alternatively also a group of extensions, for example the group text documents, in order to directly or completely accept whole file formats. By this means, all system files (they usually do not have any archival value) in a file storage can be located and discarded with one click. This simplifies the processing of large unstructured file collections.

Furthermore, the file storage can be sorted according to its path. Systematically, certain folder structures can thus be handled simultaneously or rather be incremented or disposed in its entirety.

The search engine Apache Lucene indexes the full text of each file and in this way allows a full-text search. As a result, contents of text or PDF files can be found very quickly, to enable more efficient evaluation decisions. The search function supports both lunar and phonetic search, as well as other functions, such as Boolean operators.

Appraisal decision and allocation to a transfer package

Are all the files sorted and filtered, the actual handling can start: the placing of appraisal decisions. Files can either be evaluated individually or as whole blocks. At this point, files can either be assigned to an already created SIP and thus be acquisitioned or be discarded. The number of files assigned to one SIP is unlimited; on which principle the SIPs are formed is therefore the responsibility of the archivist.

To guarantee the utmost transparency for all decisions made, the appraisal decision can be justified by a reason that can be given both via a free-text field as well as a freely configurable drop-down list. For not losing track, all files in the file collection can be easily edited by filtering the status - assigned or pending.

To avoid rash or unintended deletions of files, the discarded files are marked as such in the tool, but – as stated above – no original files are changed or deleted before the evaluation process is finished and the transfer packages are created. Thus, the appraisal decision can be changed at any time during the evaluation process.

Creation of transfer packages

When all files are processed, the transfer packages are created as .zip-files (see Fig. 3.). Within each package are all the files assigned to the package in their original folder structure. Structure points that do not contain files within this package are not included.

Name	Typ	Größe
Package1.zip	WinRAR-ZIP-Archiv	6.385 KB
Package2.zip	WinRAR-ZIP-Archiv	143 KB
Package3.zip	WinRAR-ZIP-Archiv	6.237 KB
Package4.zip	WinRAR-ZIP-Archiv	5.419 KB
Package5.zip	WinRAR-ZIP-Archiv	143 KB

Fig. 3. Transfer packages including assigned files.

Furthermore, startext COMO creates a METS XML file for each package. Therein all the information from the original file storage is saved. In addition to the details of the individual files and their technical metadata, this also includes the evaluation decision of each individual file. The information is kept for both acquisitioned and discarded files. By this, the METS XML file offers a culling history. The full text is also stored in the METS XML file so that it can be included into an archive information system.

The use of METS XML files is recognized as an archival standard, which is why the storage of information in this format offers the advantage of system-independent processing: the automated transmission of the information, including the title of the SIP, the technical metadata and the evaluation decisions, to archive information systems is ensured at this point.

4.2 startext COMO – Conclusion

Currently (as of June 2017), startext COMO is in the final phase of development, the release is planned for the following weeks. All startext clients who already work with the startext repository will receive one free license of the software.

Is it planned to include the opportunity of the basic description of the transfer packages via the data format Dublin Core, as well as the complete recording of the transfer data within the tool. startext COMO first will be released with a German user interface, but an English version will likely be offered soon. News about the transfer editor startext COMO will be published on the startext website: www.startext.de.

5 Discussion

As stated in the beginning, archivists and registrars are well aware of the subject and are trying to find answers to the challenge of unstructured data collections before the amount of data becomes overwhelming. Without any tool, unstructured file collections probably will either be transferred into the archives repository completely but unedited or be culled in their entirety. startext COMO offers a solution by allowing the archivist to structure the data according to specific standards and then to either cull or access the documents fast and without having to go through everything manually and with great expenditure of time.

At present, the program is restricted to maintaining the given or inherited folder structures and structuring them only according to certain aspects. Any reorganization or regrouping of the folder structure is not provided. Also, there is currently no opportunity to process the files contained in the file collection, i.e. it is not possible to check the format validity or even conduct a format change (e.g. from a .doc to a .pdf file). This has to be done during the ingest process to the digital archive repository.

References

- Hörl, L., Kraus, D., Massinger, T., Schmidt, A., Wanninger, S., & Wolz, A. (Mai 2017). Von Bearbeiternestern, Datenpaketen und Deduplikation. *Archivar*, S. 210-211.
- Sammler, J. (16. 03 2016). <http://fernweiterbildung.fh-potsdam.de>. Retrived on 10. 07 2017 Blog of the Department of Information Science at the University of Applied Sciences Potsdam: <http://fernweiterbildung.fh-potsdam.de/?p=1425>
- Schmidt, C. (08. 07 2015). <http://www.staatsarchiv.sg.ch>. Retrived on 10. 07 2017 at the Website of the National Archive St. Gallen, Switzerland: http://www.staatsarchiv.sg.ch/home/auds/16/_jcr_content/Par/downloadlist_1/DownloadListPar/download_1.ocFile/Schmidt_Zwischen_User_Group_un d_Entwicklungsgemeinschaft.pdf

Received: June 02, 2017

Reviewed: June 25, 2017
Final Accepted: July 11, 2017