

Citizen Archive: My Precious Information

Anssi Jääskeläinen¹, Liisa Uosukainen¹

¹ South-Eastern Finland University of Applied Sciences, Digitalia Research Center,
P.O.Box 68, 50101 Mikkeli, Finland
{anssi.jaaskelainen, liisa.uosukainen}@xamk.fi

Abstract. The trend of personal archiving is rising, but no official instance is interested in “precious” information possessed by average citizens. Cloud drives, USB devices and optical medias cannot be considered as reliable or trustworthy. This paper describes the Citizen Archive solution which aims to be the place where citizen can preserve their precious data. Furthermore, we discuss some already implemented, functional and tested solutions considering email preservation workflow and PDF splitting workflow. The experiences from our pilot users of the Citizen Archive are also considered in the text.

Keywords: personal archiving, email, PDF, long-term preservation, workflow, usability, metadata, SaaS application.

1 Introduction

Personal archiving as well as digital materials possessed by an average citizen are both very underrated areas. Harsh thing to say, but let us explain by asking a few simple questions. Do you think your precious data remains safe in a cloud? Are you able to get your precious photographs, contracts and documents inside the shelter of a national archive? Has any business archive shown interest towards your materials? Are there any other true digital repositories that would accept your material? We, as the developers of a Citizen Archive are truly surprised if even one of the above answers was yes.

The fact is that one needs to be politically or otherwise important person to get materials into the official repositories, most of the citizens aren't. Currently, citizen options for storing precious materials are portable USB devices, optical media or clouds such as Dropbox, Google Drive and OneDrive. These are good for storing backup copies, especially if multiple simultaneous methods are used. However, by all means even with goodwill cloud storage or portable devices cannot be considered as digital archives. Even though, most people do, until the disaster strikes and the data is either gone or unreadable.

At the same time, average citizens are increasingly interested in documenting their personal lives and being able to capture the most valuable artifacts. The amount of digital information produced and possessed by an average citizen raises rapidly [3]. If

users' intention is to build a personal digital life story that would cover and combine all the necessary aspects of information but nothing extra, it is challenging or even impossible with current tools. Does the digital revolution mean we are surrounded by information disposables that cannot be coupled together, stored, preserved and applied later? Information is the currency of democracy, as Thomas Jefferson used to say.

The ultimate driver behind this development work is a question: Is it right that the citizen must rely on cloud drives with dubious terms and conditions, portable USB drives or unreliable optical drives to preserve their precious information? Our answer to that questions is clear: there is a need for a professional quality digital archive service that offers the kind of user experience common folks are used to. Citizen Archives is aiming to be the solution for all who require more than portable devices or cloud drives can offer.

1.1 True digital archive

How can one define a true digital archive? For most of the readers it is probably an archive that follows the OAIS reference model and uses IP (Information Package) packages [9]. More generally, a true digital archive considers multiple aspects that plain cloud drives does not. E.g. legal aspects of stored and shared information, possibility to share usage rights, metadata according to some known metadata standard such as METS or Dublin Core, findability by using metadata, guarantee that the data remains safe and inside the country perimeters, data and file format migration, suitable preservation formats, etc.

1.2 User experience point of view

Citizens require ease of use, transferability, online access, device independency etc. During the Digitalia project that we are representing, also a need to edit and modify the digital material in order to get it more usable, aroused. For example, if a large PDF file with hundreds or even thousands of pages is automatically split into multiple smaller files according to the subject, content or keywords, the search results as well as using will be more user friendly. Furthermore, imagine a situation where you possess a multi gigabyte email container (.pst file) exported from Outlook. You know that the important contract is inside that file but you don't have Outlook anymore. Our solution takes the usability and accessibility of the emails to the next level by transforming email containers with their original attachments and metadata into searchable PDF/A-3b files.

2 The Citizen Archive

South-Eastern Finland University of Applied Sciences, Digitalia research center has been developing an archiving application which is based on an open source code

during the past few years. The implementation of the application was started in 2013 in Open Source Archive (OSA) project [5]. The first version of the archive solution, a service oriented solution suitable for a long time preservation, was developed and launched in this project. The OSA application has since been applied by civil sector organizations and non-profit associations and is now being modified to accommodate personal archives.

Digitalia, Research Center on Digital Information Management, has developed the next version of the OSA archiving application called the Citizen Archive. We have been piloting the archiving solution as a SaaS service with a private person who has a comprehensive collection of personal and family records. He has already prepared, digitized and arranged plenty of private and family collections. The material was stored in the personal computer and backed up in a cloud storage system. His intention was to share the material in the future with close relatives using USB flash drives [6]. Taking part in this project opened up new possibilities to use and share this material with immediate family members. Storing the data in a repository where the material is grouped and described with a sufficient metadata provides a reliable way to manage, maintain and share his own material. The Citizen Archive application uses Fedora as a core. Fedora is a robust and scalable open source repository, that stores the content and keeps all of the changes as former versions of the content as well. The Citizen Archive handles the backup processes on behalf of the personal archivist by storing the archived content into tape drives.

An embedded workflow engine enables the designing of the extra tasks to the archiving processes, such as pre-ingesting, updating and deleting archiving content, as the workflows. For example, the email archiving procedure, described in Chapter 3, has been implemented as a distributed micro-service in the pre-ingest workflow illustrated in the figure below.

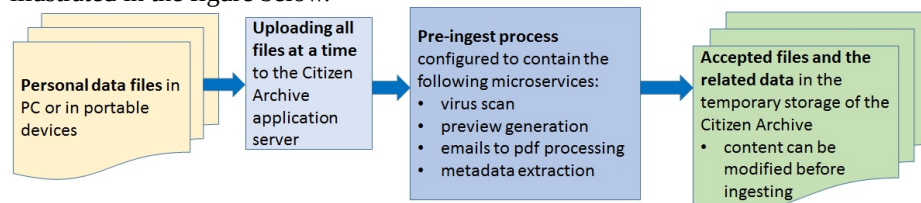


Fig. 1. Pre-ingest mass process in the Citizen Archive. Micro-services were developed as single units (i.e. jar files), that can be combined to the workflow processes in the workflow configuration file.

Personal archivists determine themselves the usage and the privacy of their archival content. Each archive is able to develop its own policies; who can access to the archive and how to use the archiving material. In practice, the administrator of the archive determines the roles for the archive which will be assigned to the user accounts. In case of the Citizen Archive, some default roles were created automatically during the archive registration process; a role for viewers, a role for browsers and a role for those who are allowed to modify the archival content. A role base access management system provides a way to determine access rights in a collection level or in a digital document level if needed.

2.1 Determining contents of a personal archive

The first steps to start digital preservation in the Citizen Archive, after collecting and arranging large personal collections, is to describe a suitable preservation plan for the personal collections. For a professional archivist this is a simple task but for the average citizen it is not. Therefore, by utilizing the UCD practices, we created user interfaces for the self-registration online and for the planning and defining the hierarchical structure of the collections. The basic Citizen Archive solution contains a few templates that the archive's owner can modify before creating the collections according to the selected template. The archive solution sets the default metadata properties for the retention period and access right according to the organization specific archiving rules automatically. So in the simplest case the user only needs to click less than five times to have a fully operational digital archive with an appropriate preservation plan.

The personal digital material is consisted of the wide variety of the material. The Citizen Archive supports the most common types of the personal material, such as documents, letters, emails, pictures, video and audio records, maps, etc. Each type of material has its own metadata model. The metadata fields used in the Citizen Archive conforms to the most common metadata standards like Dublin Core, the Finnish recommendation on document metadata JHS 143 [4], and the Finnish national standard for electronic records management SÄHKE2 [2].

In addition to the personal material, the Citizen Archive enables the description of contextual entities as the objects of their own. In the private archive the most useful entities would be places, events, and agents (i.e. persons, families, communities). An entity, such as the event like an archivist's 60-year birthday, is described unambiguously only once and it can be used when adding descriptive metadata to the archival material. The descriptive metadata definitions using contextual entities linking data inside the archive is illustrated in the figure below.

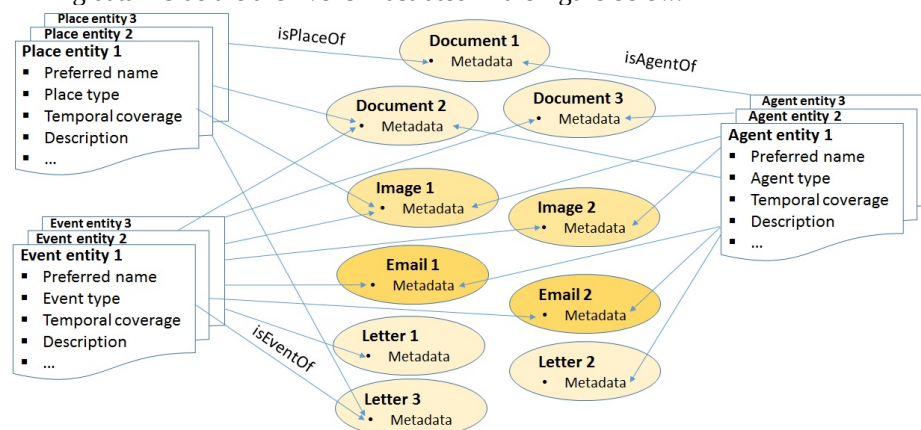


Fig. 2. Linking the archived content using contextual entities in the Citizen Archive.

The pilot case study showed that collections created according to the generations of the family is a suitable way to arrange and preserve family records. The Citizen Archive utilizes Apache Tika metadata extraction tool to retrieve embedded metadata

and text from files during the ingesting process. The automatic metadata extraction and the metadata inheritance according to the preservation plan facilitate the work of archivist but still a lot of work is needed. Family heritage material focuses heavily on descriptive information [10]. If contextual entities are enhanced when adding metadata to the documents in an archive, it enables the classification of the search results in different ways for example. The set of content can be classified along persons, places or events.

2.2 Benefits and challenges of a SaaS personal archiving

According to the results of the pilot testing, the Citizen Archive supported establishing many types of archives. It provided the pilot team members with a platform to manage and view family heritage data. On the other hand, they wanted to be assured about the reliability and continuity of the archiving service. The Citizen Archive made it possible to share the data to the restricted users in a way that would otherwise not have been possible [6]. This kind of archive offers opportunities to gather metadata collectively as well. The information is always up-to-date for all of the authorized users. It was also considered advanced, that unlike cloud services, the Citizen Archive offered various automatic processes to manage the archiving material. The procedure of moving the archiving material automatically to the disposal list at the end of the retention period is an example of this kind of service.

The archiving might be considered a simple task, but in practice collecting, digitizing, arranging, ingesting and describing the archiving material often turns out to be a very demanding and a time consuming task. This may be a bottleneck of using the archiving service. The archival terminology was perceived as difficult to understand as well. Therefore the Citizen Archive GUI needs further development to be more user friendly, self-explanatory and intelligent archive. More work is also needed in the automatization of the functions in the Citizen Archive. The service like the Citizen Archive could open up new possibilities. The Citizen Archive could in the future complement the official national archives where historians and genealogists could collect additional information from [6].

3 Implemented features

Without an archival feature set (ingest, archival storage, data management, preservation planning, access, and administration [9], the Citizen Archive would be just like an enhance cloud drive from the average citizen point of view. When an archive fills the above list of functionalities it can be considered as a true digital archive. However, for being suitable for average end users this is not enough. Google like UX is the current minimum, drag and drop operation is expected to be working everywhere and backup functionality should happen automatically in the background. For these reasons we have been working on UX and developed some nice features that will greatly increase the usability and usefulness of the citizen archive. The first implemented feature tackles the issue with everyday formats vs. archival formats.

Generally speaking, this issue is way too complicated and it is principally wrong to obligate the ordinary citizen to handle it. The format migration part must be an automated part of the process when an electronic item is ingested into an archive. The first subchapter describes the studied and implemented process of transforming proprietary email formats into a fully qualified archival format which contains all the original metadata and attachments of every individual email. The second sub-chapter handles the issues with distributing, sharing and managing very large PDF files.

3.1 Proprietary emails into archive

Email messages are already in a digital format; yet many of the files are either in a proprietary format, incompatible with each other, obsolete currently or after few years, or just in an unreadable format for any modern software [1]. In a short term usage, this is not a problem, but in a long term it will be.

When archival formats are decided, the rules and requirements from the national archives cannot be overlooked. We haven't encounter a single national archive that would not accept PDF/A format into their digital repository. However, in the case of email formats there are still national archives that haven't even considered the plain format yet. Finnish National Archives for example only accepts emails if those have arrived via case management system as cases. If the recommendations from bigger national archives are studied, then the practices are quite coherent thus most of them accept .eml, .mbox, .pst (.ost) and .msg.

We totally understand the acceptance of .eml and mbox files due to being open file formats. However, we won't understand the acceptance of the other file types in Table 1 since the general rule of long term preservation is not to accept proprietary binary formats for long term preservation. Microsoft Outlook is a perfect example of a program that produces such formats. Pst, .ost and .msg are all made by it and are proprietary binary formats.

One reason for acceptance of this format into national archives might be the fact that, this file format extension has existed for almost 20 years. Still, it is also a fact that the older versions (97-2002) are not compatible with the newer ones. The main reasons for incompatibility are changed character encoding (ANSI vs. UNICODE) and a renewed file format. Outlook 97-2003 format only supports 2Gb files while 2010 and 2013 defaults the limit to 50Gb. These changes might cause incompatibility issues that are unmanageable for a normal citizen. Even if a particular Outlook data file could be imported into "Outlook 2026" the imported files are bound to that particular email client. Therefore the multi device utilization on different personal devices, for instance, is difficult. So as a result, we don't recommend placing proprietary email formats into our archive, but instead of just mentioning this we have done something to solve this issue.

Manual conversion.

The first alternative, but obviously not so good one, is to let the user to do the conversion before uploading a file into an archive. For this purpose there exist a lot of

instructions on what formats to use if doing long term preservation, NARA [8] for example. Furthermore, virtually every office-, image manipulation- or email program is capable of producing some kind of archival format, generally PDF/A. Still, from the user experience point of view it is an overwhelming job to save e.g. even hundred important emails into an archival format especially when the native format of an application is always the default selection for saving. It is the users' responsibility to recognize and pick the true archival format from the big list of different formats. From the authors' opinion, there is no way that an average Joe or Jane is able to manage in this task. Naturally there exist some third party plugins, such as ImportExportTools for Thunderbird that simplify this task, but still how many average end users install additional plugins into Thunderbird email client?

Fully automated workflow.

The second, and a much better option is to use an automated conversion workflow and this is what we have been developing. This work has been done at the Digitalia by utilizing solely open source products, Linux, Python and Java programming. Although the conversion utilizes multiple software, the end user only need to provide the source file e.g. via Citizen Archive UI.

This solution is aiming at PDF/A which is the standard for archival documents and furthermore readable and transferrable device independently. To be more precise we are aiming at PDF/A-3 – it will make it easy to store original email attachments as they were. Finally, with the utilization of PDF/A, it can be ensured that the document remains unmodified after it has been created as well as is rendered identically on all devices. Furthermore, our workflow also scans the metadata content of every single email and produces a summative .csv file which can be fed into network analysis software like Gephi. Next listing shows and briefly explains the programs that are executed during the conversion process.

1. pffexport: Extracts the folder structure from the provided .ost or .pst file
2. metadata converter: Java application that converts metadata from pffexport format into a ghostscript format. Our own Java implementation.
3. abiword: Converts .txt and .rtf files into PDF format
4. wkhtmltopdf: Converts .html files into PDF format
5. Libreoffice writer & ImageMagick: Converts attachments into an archival format
6. gs: Ghostscript converts PDF files into PDF/A-3b files and adds the original metadata and attachments
7. VeraPDF: Validates the final output

This workflow is fully operational and in most cases will produce perfectly valid PDF/A-3b files. We still have an issue in defining file attachment with pdfmark in a way those are fully validated but other than that everything is working as expected. Complete processing time for about one GB email box with 10800 emails is about 11 minutes with a 16-core server. Figure 3 demonstrates the transition from original

email into a valid PDF/A-3b file. At the current stage of the development, the format migration workflow only supports Outlook data files. One of our primary tasks in the forthcoming project is to extend this support to cover other email formats such as mbox as well.

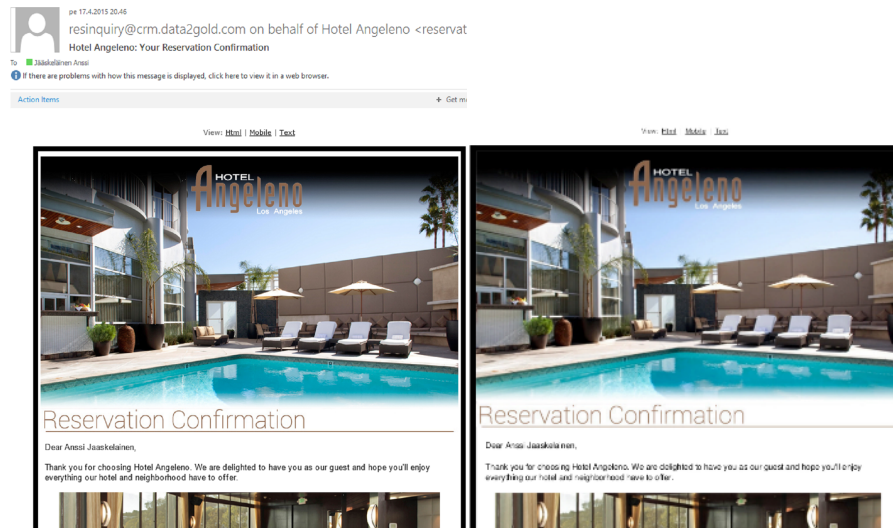


Fig. 3. Original email in Outlook (left) and the converted in Adobe Acrobat (right).

3.2 PDF splitter

Imagine a situation where you have hundreds or even thousands of PDF documents which most have more than 500 pages and have physical size more than 50Mb. Furthermore, what if you have to share these files with your end users that are most likely using older computers with slow internet connections. We had this situation with our co-operator Helsinki City Archives and the largest file was as big as 466,7 Mb and the longest file had almost 2000 pages. For a modern computer with a fast internet connection, these are not problems, but downloading 466,7 Mb file over a basic network connection (1-2 Mbps) will take over an hour and the browsing of 2k page PDF file is everything else than smooth. For these reasons we decided to develop an automated workflow that divides PDF files into smaller chunks while maintaining the original metadata. Table 1 shows the calculated average loading times with 10Mbit/s internet connection as well as the reduced loading times.

Table 1. Download times before and after the PDF splitting.

Original size	10Mbps download	Average splitted size	10Mbps download	Reduce in download time
466,7Mt	6min 31s	38,9Mt	32s	~92%
235,3Mt	3min 17s	9,6Mt	8s	~96%
45Mt	37s	5Mt	4s	~86%

Modern end uses are used for instant response and fast action[7]. This means that download times must be as small as possible and the viewer programs should instantly respond to user actions. This is not going to happen with a 466 Mb file which has more than thousand page

Technical workflow

At first phase the original metadata is read by using an open source program called pdftk. The first task is to identify bookmarks which are shown as BookmarkTitle, BookmarkLevel and BookmarkPageNumber tags. If bookmarks are found, those are used as split points. If the PDF file does not contain bookmark information, then the possible cut points are decided by using a keyword matrix. Naturally these keywords are context sensitive and need to be picked case by case. In our case, which was old city government records, the achieved accuracy with the keyword matrix was about 80%.

When the split points are defined those are fed to an open source program called ghostscript which does the actual splitting. Handling 308 large PDF files took about two minutes and produced 5917 smaller files which were then uploaded back into the Helsinki City Archives digital repository.

As an extended feature, a functionality that will aid in anonymization of the PDF content was implemented. This functionality reads the PDF files word by word and seeks Finnish fore and last names. If such words are found, the found name and document name with founded page and row number are stored in a separate .csv file which can be used later on to automate the anonymization process.

4 Future development and conclusions

This paper illustrated the current development of the personal archiving application for citizens. Starting the agile development of the Citizen Archive with a small pilot team, we intended to release a more user friendly version of the application, that have been implemented in South-Eastern Finland University of Applied Sciences, being suitable for personal archiving. The Citizen Archive fulfills a major gap by providing a solution for the reliable long-term preservation of citizens' digital material and keeping their personal life stories found for future generations. Development of the

Citizen Archive solution will continue and most likely there will be a migration to the latest product version of Fedora or some other suitable environment. During the next few years of operation the functionality of the email migration tool is extended to include other common email data formats as well. Secondly, the operability of the solution is further extended, e.g. user can pick the target format, inclusion or exclusion of attachments and so forth. Also data-analytics will be developed so that e.g. personal information can automatically be “black boxed” if the user requires that. Also the content analysis will be further developed, probably by taking benefit of NER (Named Entity Recognition) functionality and open data sources.

References

- [1] Anderson, D. Preserving the digital record of computing history. *Commun. ACM* 58, 7, 29–31. (2015)
- [2] Finnish National Archives, SÄHKE2 spesification. (2009)
- [3] Hawkings, D.T. *Personal Archiving: Preserving Our Digital Heritage*. Medford, NJ. (2013)
- [4] JHS 143 Asiakirjojen kuvailun ja hallinnan metatiedot. (2012)
- [5] Jääskeläinen, A., Uosukainen, L. Mastering the fuzzy information in the “cloud era”: Case Open Source Archive. In *proc. of Lisbon DLM Conference*. (2014)
- [6] Kauslainen, E., Uosukainen, L. Kansalaisarkisto – sukuyhteisön aarteet talteen digitaaliseen arkistoon in the publications of South-Eastern Finland University of Applied Sciences: *Digitaalinen tieto haltuun*, 40-47. (2017)
- [7] Lowdermilk, T. *User-Centered Design a Developer’s Guide to Building User-Friendly Applications*. O’Reilly Media. (2013)
- [8] NARA. Revised Format Guidance for the Transfer of Permanent Electronic Records <http://goo.gl/Vp2ntJ>. (2014)
- [9] OAIS Reference model, ISO 14721:2012. (2012)
- [10] Riley, J. *Understanding metadata: What is metadata and what is it for?* Baltimore, MD: National Information Standards Organization. (2017)