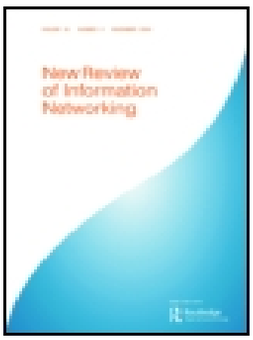

New Review of Information Networking

ISSN: 1361-4576 (Print) 1740-7869 (Online) Journal homepage: https://www.tandfonline.com/loi/rinn20

# Citizen Archive: My Precious Information

Anssi Jääskeläinen & Liisa Uosukainen

**To cite this article:** Anssi Jääskeläinen & Liisa Uosukainen (2018): Citizen Archive: My Precious Information, New Review of Information Networking, DOI: 10.1080/13614576.2018.1537800

**To link to this article:** https://doi.org/10.1080/13614576.2018.1537800

NEW REVIEW OF INFORMATION NETWORKING
https://doi.org/10.1080/13614576.2018.1537800

Routledge
Taylor & Francis Group



# Citizen Archive: My Precious Information

Anssi Jääskeläinen and Liisa Uosukainen

Digitalia Research Center, South-Eastern Finland University of Applied Sciences, Mikkeli, Finland

**ABSTRACT**

The trend toward personal archiving is rising, but no official archives are interested in the "precious" information possessed by average citizens. Consequently, citizen data is stored on cloud drives, USB devices, and optical media that cannot be considered to be reliable or trustworthy. This article describes the Citizen Archive solution, which aims to be the place where citizens can preserve their precious data. Furthermore, we discuss some previously implemented, functional, and tested solutions considering e-mail preservation workflow and PDF splitting workflow. The experiences from pilot users of the Citizen Archive are also reviewed in the text.

**KEYWORDS**

Personal archiving; e-mail; PDF; long-term preservation; workflow; usability; metadata; SaaS application

## Introduction

Personal archiving as well as digital materials possessed by the average citizen are both under-researched and underrated. The answers to a few simple questions will highlight this. Is personal data safe in a cloud? Can ordinary citizens get their precious photographs, contracts and documents inside the shelter of a national archive? Are business archives interested in personal materials? Are there any other true digital repositories that would accept citizen's material? It would be surprising if even one of the aforementioned answers was "yes."

It is necessary to be politically or otherwise important to get materials into official repositories, and most citizens are not. Currently, a citizen's options for storing their 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, cloud storage or portable devices are not digital archives. Even though, many lay people do consider these to be archives, 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 the average

**CONTACT** Anssi Jääskeläinen anssi.jaaskelainen@xamk.fi Digitalia Research Center, South-Eastern Finland University of Applied Sciences, P.O.Box 68, Mikkeli 50101, Finland.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/rinn.

citizen is increasing rapidly (Hawkings). If a 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 would be challenging or even impossible with current tools. Does the digital revolution mean we are surrounded by information that cannot be coupled together, stored, preserved, or applied later?

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? There is a need for a professional quality digital archive service that offers the kind of user experience the public are used to. Citizen Archive aims to be the solution for all who require more than portable devices or cloud drives can offer.

## True digital archive

What is the definition of a true digital archive? For archival people, it is probably an archive that follows the OAIS reference model and uses IP (Information Package) packages (ISO 14721). More generally, a true digital archive considers multiple aspects that a plain cloud drive does not. These can include: the legal aspects of stored and shared information, the possibility to share usage rights, metadata according to some known metadata standard such as METS or Dublin Core, searching using metadata, guarantees that the data remains safe and inside the country perimeters, data and file format migration, suitable preservation formats, and, finally, machine readability.

## User experience point of view

Citizens require ease of use, transferability, online access, device independency, and so forth. During the Digitalia[1] project that we are representing, a need to edit and modify the digital material was raised. 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 and usability will be more user friendly. Furthermore, imagine a situation where a multi-gigabyte E-mail container (.pst file) is exported from Outlook. There is an important contract inside the file but the user does not have Outlook anymore. This solution takes the usability and accessibility of the E-mails to the next level by transforming E-mail containers with their original attachments and metadata into searchable PDF/A-3b files.

[1] https://www.xamk.fi/en/rdi/digitalia-research-center-digital-information-management/.

## The citizen archive

The South-Eastern Finland University of Applied Sciences, Digitalia research center has been developing an archiving application which is based on open source code. The implementation of the application was started in 2013 in Open Source Archive (OSA) project (Jääskeläinen and Uosukainen). The first version of the archive solution, a service-oriented solution suitable for long term preservation, was developed and launched in this project. The OSA application has since been applied by organizations in the civil sector and non-profit associations and is now being modified to accommodate personal archives.

Digitalia, the Research Center on Digital Information Management, has developed the next version of the OSA archiving application called the Citizen Archive. The archiving solution has been piloted 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 (Kausalainen and Uosukainen). 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 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 on tape drives.

An embedded workflow engine enables the design of extra tasks to be utilized in the Citizen Archive workflow. For example, the E-mail archiving procedure, described thoroughly in the Implemented Features section, has been created with the workflow engine. The E-mail archiving procedure is a distributed micro-service in the pre-ingest workflow and is illustrated in Figure 1.

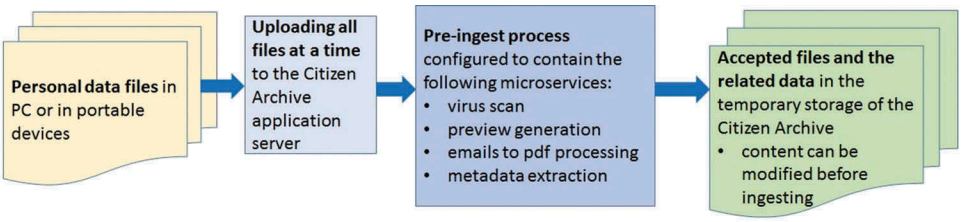


**Figure 1.** Ingest workflow of the Citizen Archive.

Personal archivists determine the usage and the privacy of their archival content. Each archive is able to develop its own policies; such as who can access the archive and how to use the archive material. In practice, the administrator of the archive determines the roles for the archive that will be assigned to the user accounts. In the 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-based access management system provides a way to determine access rights at a collection level or at a digital document level if needed.

## 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 (User-Centered Design) practices, it was possible to create user interfaces for self-registration online and for 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. As a result, 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 consists of a wide variety of material. The Citizen Archive supports the most common types of the personal material, such as documents, e-mails, pictures, video, audio, 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), and the Finnish national standard for electronic records management (SÄHKE2).

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. people, families, communities). An entity, such as the event of an archivist's sixtieth 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 Figure 2.

The pilot case study showed that collections created according to the generations of the family are a suitable way to arrange and preserve family

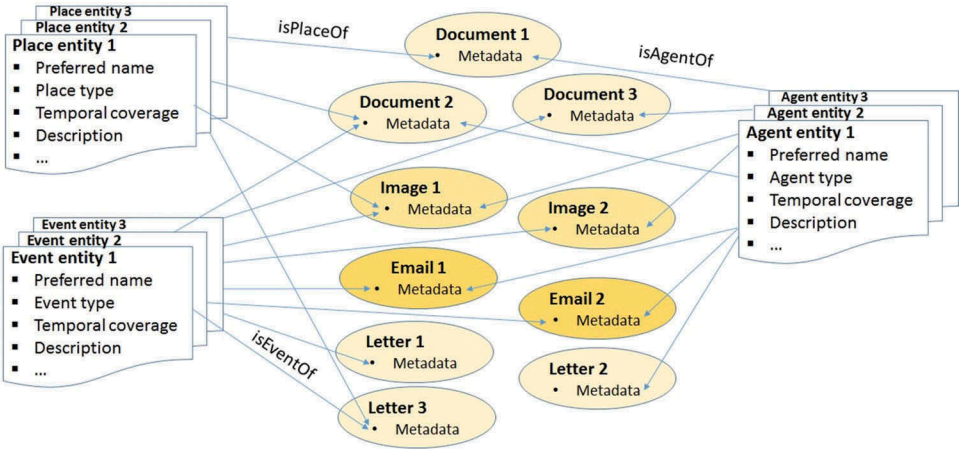


**Figure 2.** Linking data with contextual entities.

records. The Citizen Archive utilizes the Apache Tika metadata extraction tool to retrieve embedded metadata and text from files during the ingestion 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 (Riley). 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 content can be classified by people, places or events.

## *Benefits and challenges of SaaS personal archiving*

The pilot tests were carried out during the previous projects from January 1, 2015 to July 31, 2017. A few pilot users and organizations registered their own archives in the Citizen Archive platform and continuously provided free form feedback regarding their experiences with the system. According to the feedback, the Citizen Archive supported establishing numerous types of archives. It provided the pilot team members with a platform to manage and view family heritage data. However, the pilot users also wanted assurances about the reliability and continuity of the archiving service. The Citizen Archive made it possible to share data with the restricted users in a way that would otherwise not have been possible (Kausalainen and Uosukainen). 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. Unlike cloud services, the Citizen Archive offered various automatic processes to manage the archive material. Automatically moving the archive material to the disposal list at the end of the retention period is an example of this kind of service.

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 demanding and time-consuming task. This may be a potential impediment of using the archiving service. The archival terminology was perceived as difficult to understand as well. Therefore, the Citizen Archive grapical user interface (GUI) needs further development to make it a more user friendly, self-explanatory, and intelligent archive. More work is also needed in the automatization of the functions in the Citizen Archive. A service such as the Citizen Archive could open up new possibilities. In the future, the Citizen Archive could complement official national archives from which historians and genealogists could collect additional information (Kausalainen and Uosukainen).

## Implemented features

Without an archival feature set (ingest, archival storage, data management, preservation planning, access, and administration; ISO 14721: 2012), from the user perspective, the Citizen Archive would simply be an enhanced cloud drive. When an archive achieves the aforementioned list of functionalities, it can be considered a true digital archive. However, this is insufficient for average end users. For the average end user, a google-like user experience (UX) is the current minimum, where a drag and drop operation is expected to be working everywhere and backup functionality should happen automatically in the background. For these reasons, work has been carried out on UX and some features have been developed that will increase the usability and usefulness of the Citizen Archive. The first implemented feature tackles the issue with everyday formats versus archival formats. Generally speaking, this issue is too complicated to require the ordinary citizen to handle. The format migration part must be an automated part of the process when an electronic item is ingested into an archive. The first sub-chapter describes the studied and implemented process of transforming proprietary E-mail formats into a fully qualified archival format that contains all the original metadata and attachments of every individual E-mail. The second sub-chapter handles the issues with distributing, sharing, and managing large PDF files.

### *Proprietary E-mails into archive*

E-mail messages are already in a digital format; yet, many of the files are either in a proprietary format, incompatible with each other, potentially obsolete after few years, or just in an unreadable format for any modern software (Anderson). In the short term, 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. All national archives currently accept PDF/A format into their digital repositories. However, in the case of E-mail formats there are still national archives that have not yet even considered the plain format. The Finnish National Archives, for example, only accepts E-mails if those have arrived through the case management system as cases. If the recommendations from bigger national archives are studied, then the practices are quite coherent and thus most of them accept .eml, .mbox, .pst (.ost), and .msg.

The acceptance of .eml and .mbox files is understandable because these are open file formats; however, the acceptance of the other file types such as .msg and .pst is less clear because of the general rule of long-term preservation is not to accept proprietary binary formats. Microsoft Outlook is a perfect example of a program that produces such formats (.Pst, .ost, and .msg are all created by the program 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, the older versions of Microsoft Outlook (97–2002) are not compatible with the newer ones. The main reasons for incompatibility are changed character encoding (ANSI versus UNICODE) and a renewed file format. The Outlook 97–2003 format only supports 2 GB files while 2010 and 2013 format defaults the limit to 50 GB. These changes might cause compatibility issues that are unmanageable for a non-technology oriented member of the public. Even if a particular Outlook data file could be imported into "Outlook 2026," the imported files are bound to that particular E-mail client. Consequently, multi-device utilization on different personal devices, for instance, is difficult. Therefore, placing proprietary E-mail formats into any archive is not recommended from the authors' perspective. The actions detailed in the following sections have been taken to solve this issue.

### *Manual conversion*

The first alternative, but obviously not a good one, is to let the user to do the conversion before uploading a file into an archive. Extensive information exists on what formats to use for long-term preservation, such as (NARA). Furthermore, virtually every office-, image manipulation-, or E-mail program is capable of producing some kind of archival format, generally PDF/A. However, from the user experience perspective, it can be an overwhelming job to save even a hundred important E-mails into an archival format, especially

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

| Original size | 10Mbps download | Average split size | 10Mbps download | Reduce in download time |
|---|---|---|---|---|
| 466.7 MB | 6min 31s | 38.9 MB | 32s | ~ 92% |
| 235.3 MB | 3min 17s | 9.6 MB | 8s | ~ 96% |
| 45 MB | 37s | 5 MB | 4s | ~ 86% |

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 growing list of different formats. The authors' opinion is that this would be highly problematic for an average end user to manage. Naturally, some third-party plugins exist, such as ImportExportTools for Thunderbird that simplify this task, but it is questionable how many average end users install additional plugins into the Thunderbird E-mail client.

### *Fully automated workflow*

The second, and better option, is to use an automated conversion workflow and this is what has been developed. This work has been performed solely using open source products, such as Linux, Python, and Java programming. Although the conversion utilizes multiple software, the end-user only needs to provide the source file (e.g., through the Citizen Archive UI).

This solution is aimed at PDF/A, which is the standard for archival documents and, furthermore, is readable and transferrable independently from devices. Finally, the use of PDF/A ensures that the document remains unmodified after it has been created as well as being rendered identically on all devices. Furthermore, the workflow also scans the metadata content of every single E-mail and produces a summative .csv file that can be fed into a network analysis software such as Gephi. The following list shows and briefly explains the programs that are executed during the conversion process:

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

This workflow is fully operational and produces valid PDF/A-3b files. The complete processing time for a 1 GB E-mail box with 10,800 E-mails is about 11 minutes with a 16-core server. Figure 3 demonstrates the transition from original E-mail into a valid PDF/A-3b file. At the current stage of the development, the format migration workflow supports Outlook, .Mbox, .eml, and .msg data files One of our primary tasks in the ongoing project is to extend this support to cover other E-mail formats as well.

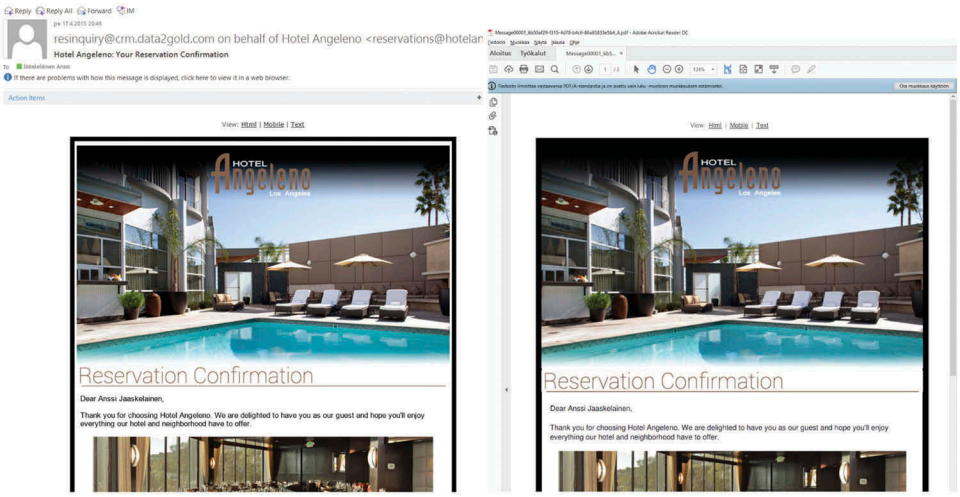


**Figure 3.** From Outlook email into archival PDF/A-3b file.

## *PDF splitter*

Imagine a situation where there are hundreds or even thousands of PDF documents many of which have more than 500 pages and have a file size of more than 50 MB. These files often need to be shared with end users who may be using older computers with slow internet connections. This situation was apparent at the Helsinki City Archives where the largest file was 466.7 MB and the longest file had almost 2,000 pages. For a modern computer with a fast internet connection, this is not problematic, but downloading a 466.7 MB file over a basic network connection (1–2 Mbps) will take over an hour and browsing a 2,000 page PDF file is challenging. For these reasons, an automated workflow was developed 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.

Today end users are used to instant responses (Lowdermilk). This means that download times must be as short 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 2,000 pages.

## *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, they 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 on a case by case basis. In our case, which was old city government records, the achieved accuracy with the keyword matrix was about 80%.

After the split points are defined, they are fed into an open source program called ghostscript that does the actual splitting. Handling 308 large PDF files took about two minutes and produced 5,917 smaller files that were then uploaded back into the Helsinki City Archives digital repository.

As an extended feature, 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 found page and row number are stored in a separate .csv file that can be used later to automate the anonymization process.

## Future development and conclusions

This article has illustrated the current development of a 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, which was implemented in the South-Eastern Finland University of Applied Sciences and is suitable for personal archiving. The Citizen Archive fills a major gap in the market by providing a solution for the reliable long-term preservation of citizens' digital material and maintains their personal life stories 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 E-mail migration tool will be extended to include other common E-mail data formats as well. Additionally, 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 (e.g., personal information can automatically be "black boxed" if required by the user). Also, the content analysis will be further developed, probably through utilization of NER (Named Entity Recognition) functionality and open data sources.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This work was supported by the European Regional Development Fund.

## Works cited

Anderson, David. "Preserving the Digital Record of Computing History." *Communications of the ACM*, vol. 58, no. 7, 2015, pp. 29–31. Print. doi:10.1145/2797100.

JHS 143, *Metadata in the description and administration of documents* Finland, (2012).

SÄHKE2, *SÄHKE2 Specification*. Finland, (2009).

Hawkings, Donald T. *Personal Archiving: Preserving Our Digital Heritage*. Medford, 2013.

Jääskeläinen, Anssi, and Liisa Uosukainen. "Mastering the Fuzzy Information in the "Cloud Era": Case Open Source Archive." Proceedings of the DLM Forum –7th Triennial Conference, Lisbon, 10–14 November 2014, pp. 89–91.

Kausalainen, Eero, and Liisa Uosukainen. (2017) "Kansalaisarkisto – Sukuyhteisön aarteet talteen digitaaliseen arkistoon." In M. Kosonen, *Digitaalinen tieto haltuun*, (40–47). Retrieved from http://urn.fi/URN:ISBN:978-952-344-020-3

Lowdermilk, Travis. *User-Centered Design a Developer's Guide to Building User-Friendly Applications*. O'Reilly Media, 2013.

NARA. "Revised Format Guidance for the Transfer of Permanent Electronic Records," *Bulletin 2014-04*, 2014, https://www.archives.gov/records-mgmt/bulletins/2014/2014-04.html

Riley, Jenn. *Understanding Metadata: What Is Metadata and What Is It For?* NISO, 2017.

International Organization for Standardization, Space data and information transfer systems - Open archival information system (OAIS) - Reference model. ISO 14721, 2012.