My precious information – how to preserve it?

Anssi Jääskeläinen, Miia Kosonen, Liisa Uosukainen, South-Eastern Finland University of Applied Sciences, Mikkeli, Finland

Abstract

Do you think your information remains safe inside a cloud? Do you have another truly trustworthy place where you can store all your precious information? These questions lead us to the basic problem behind this paper: None of the official instances are interested in materials possessed by average Joes and Janes. You will have to be politically or otherwise important person to get your personal life story into official digital repositories. We at the Digitalia[1] (Research Center on Digital Information Management) at South-Eastern Finland University of Applied Sciences, believe that there is a strong need for a digital preservation service that would give ordinary citizens the right to decide what to do with their personal information. It is not right that common folks must rely on cloud drives with dubious terms and conditions or unreliable portable or optical devices to store their precious digital information. This article describes an initiative of a low cost full-scale digital archive solution that will be available to common people.

Motivation & problem

People are increasingly interested in documenting their personal life and being able to capture its most valuable artifacts. At the same time, the amount of digital information produced by an average citizen has increased exponentially [4]. This information is spread over a variety of online services, it is difficult to find, may be destroyed accidentally, or may become inaccessible over time [12]. Hence, there is a need for a trustworthy service allowing users to collect information from a variety of data sources, while allowing them to search, present, share and enrich such information resources easily. As Pimminger et al. [12] note in their description of the Themis platform, many current tools have limited scope, restricted to backup, encryption or simply exchanging messages. Collecting, sharing and inheriting personal digital information is an entirely new territory. The reason may be twofold. Firstly, it can be a challenging task for developers to integrate all these aspects into one manageable service, especially if the basic principles of user centered design are followed [13]. Secondly, professional archivists and archival institutions have focused on the most influential representatives of the society and cultural heritage, and personal archiving practices and solutions for ordinary citizens have simply gained less attention. From the citizen or offspring point of view, this "culturally non-important" information can yet have a price above rubies.

Therefore, the authors present a justified 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? We maintain the answer 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. In practice this means that all references to archival field, such as terminology as well as practices must be fully transparent to the user who seeks modern functionalities such as drag-and-drop or intuitive search.

Typical historical sources of personal information range from print documents, letters and photos to analog videos. An example: letters from the war and other important mails used to be stored in a "secure" cardboard boxes and further on the upper shelf of the wardrobe. These boxes in a wardrobe were, and in many cases still are, the family archives. When the "archival masters" are getting older and older, many of these important documents are being digitized in order to preserve valuable knowledge for the future generations. However, the trustworthy place to store and share these digitized artefacts in many cases, is missing which leads to preservation on e.g multiple usb sticks or cloud.

In contrast, modern digital media tools and social media accounts allows everyone to share the aspects of their life story easily. The classic art of storytelling has even been rephrased as digital storytelling. Digital images, YouTube videos, MyData, such as measurement data, and online communication like e-mails or social-network exchanges may form the current "family archives" which are scattered around the internet. People produce the rich sets of digital collections, growing every day and making every individual as a culture himself or herself [5] – these are not just sets of digital data but "a set of artifacts that has the potential to chronicle your life". In line with MyData principles, personal digital archiving needs to ensure managing coherent descriptive metadata and access rights while also ensuring privacy and usefulness.

People should have a right to obtain data about themselves, use it freely, and to share, donate or do whatever they like with my data. This however, leads to many questions and unsolved issues mainly due to legislation. Digitalia has also considered these issues in co-operation with information law specialists, but suggestions and solutions are in the scope of future papers. The fact however, is that these need to be solved before e.g. personal healthcare information can be shared via archives or portals. Personal information also links various users and communities together, such as in the case of family archives. Particularly in the digital world, information is not created in isolation but connected with others in that content ecosystem [5].

To sum, people may easily produce stories by collecting information fragments around the web content, but how to "keep these found things found"? [14]. Does the digital revolution mean we are surrounded by dynamic information that cannot be combined, stored, preserved and applied later? Do we simply have to "donate" all the information about ourselves to large businesses and officials without having the right to use it? This, although is the current situation, does not seem reasonable nor justified. Information is the currency of democracy, as Thomas Jefferson used to say.

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

Case: E-mails

One aspect that raises a particular interest is e-mails. Even though some might wonder why such an old-fashioned way of communication was chosen, we point out that according to internet live stats[2], every single second about 2.56 million e-mails are sent. According to the EMC Digital Universe Study[3], the amount is even suspected to grow in the future. The number of e-mails is huge in comparison with 7500 tweets that are sent every second, or 59 000 Google searches that are conducted every second. In addition, as the *de facto* tool of online communication e-mails provide researchers and communities valuable information about knowledge-sharing patterns, organizational power structures, and personal life stories. From the wide variety of digital data, it is of the particular essence to be able to capture and preserve the valuable parts of e-mail exchanges. They can be re-used to track the past and to estimate the future.

Proprietary formats are a big issue in the world of digital information and preservation. This is also the case with e-mails. National archives tend to have their own recommendations for preservation formats, but their solutions are mostly based on preserving the original file as OAIS (Open Archival Information System) AIP (Archival Information Package) packages or XML structure. From an average end user point of view these are not options. Even if producing such information packages or XML structure would be an easy task, the correctly formatted object alone is not enough, a truly reliable and sustainable place to store these objects is also required. This is something citizens don't have thus cloud drives, portable USB devices or optical media cannot be considered as reliable nor trustworthy. If an acceptable solution cannot be found, we are in danger of losing more than 20 years of history in personal communication and, in broader terms, personal digital cultural heritage.

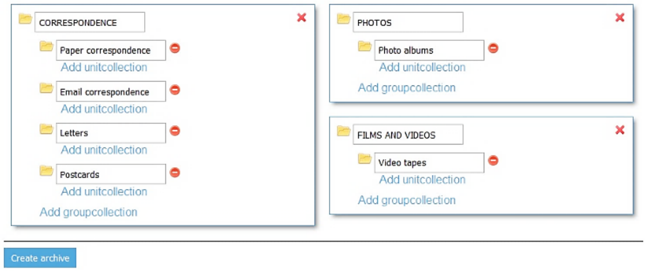
Citizen archive

In Digitalia, we believe in a brighter future where people are able to manage and preserve their precious personal information with easy-to-use and low-cost tools. Yet it requires developing software clients to better match the need for ordinary citizens. We have started to design the citizen archive together with actual end users according to their needs and wishes. Naturally a decision needs to be make when to listen to users and when to not, this important issue is also emphasized by Lowdermilk [13]. For example, we received a complaint from a user who was not able to access his repository. Logs revealed that he was trying to use an outdated browser which did not support our new security certificate. Therefore a principal decision was made that we won't be supporting outdated browsers and it will be user responsibility to update their browsers to meet the minimum requirements.

The benefits of the citizen archive development are twofold. Firstly, from archiving point of view, it offers a reliable long-term preservation solution. Secondly, in line with MyData principles it allows citizens to better manage their personal information by supporting its collection, access, presenting, sharing and control over such information.

Naturally, the concept of the citizen archive is not only about the technical aspects of development. It covers a wide range of fields from legislation to socio-cultural issues such as supporting communities, groups and families in preserving their shared digital heritage. At the current stage, we have conducted a report on the legal rights and responsibilities concerning the collection and preservation of personal digital data. The report serves as the basis for further development and supports us in outlining the potential business models for the citizen archive. This report is available in Xamk theseus[4], it is written in Finnish but the abstract on page 8 is in English.

Previously developed and enhanced OSA (Open Source Archive) solution is working as a base of the citizen archive [8]. A citizen can register his or her own archive in the OSA service to preserve personal digital documents, letters, photos and videos with appropriate metadata. Part of this archival creation process is presented in Figure 1. In addition to the personal digital material, the owner of the archive can create entities to describe the places, events and people [7]. Using these kind of entities when adding the metadata for the personal digital material links all the material together. In this way a consistent metadata and accurate searches can be provided. In order to facilitate the ingest process, technical and manually added metadata is extracted automatically from files as well, when uploading them to the archive. The OSA based archive provides secure preservation. The archive owner manages its own material and can add or remove archive users. It is possible to define individually which material is available for the user in the archive and naturally the material can also be marked as public.



CORRESPONDENCE
Paper correspondence
Email correspondence
Letters
Postcards
PHOTOS
Photo albums
FILMS AND VIDEOS
Video tapes

Figure 1: Part of the archival creation process

Although the OSA environment works already well in the archival context it needs to be mentioned that it was designed and developed in co-operation with archives and archivists. Archivist have their own professional terms, ways of working etc. due to which, lots of development work needs to be done before it can fulfill the requirements of ordinary citizens.

Several steps towards this goal have already been taken. Jääskeläinen studied the meaning of user experience in digital preservation [1]. According to these results, users in general require ease of use, transferability, online access, device independency etc. aspects. Secondly, the OSA archival control UI was redesigned to be more human by utilizing novel methods such as service blueprints and customer profiling [2]. Currently the

[2] http://www.internetlivestats.com/

[3] https://www.emc.com/leadership/digital-universe/index.htm?pid=landing-digitaluniverse-131212

[4] http://www.theseus.fi/handle/10024/121274

functionality has been extended with drag and drop operation and other features that are trying to make the citizen archive less archive-like. One of these features is automatic migration of Outlook e-mail data structures into valid PDF/A-3 files.

From proprietary format into PDF/A-3

In theory it sounds like an easy task to convert e-mail into PDF. Outlook makes it possible to just select and print directly into PDF files. One of the authors tried this functionality with 405 Mb inbox. Core i7 laptop with 16 GB of memory was completely frozen and finally managed to pop out an "out of resources" error message. With 20 emails the functions managed to complete but all selected emails were print to the same file. Furthermore, this file did not contain the original metadata of the emails. So obviously this is not an option for archives.

E-mail messages are already in a digital format; yet many of the files are either in a proprietary format, incompatible with each other, obsolete or obsolete after a few years, or just in an unreadable format for modern software [1]. The issue is the same with virtually any everyday formats which are not archival graded. Lots of instructions on how to save a certain document into an archival format can be found for example from the web pages of national archives. As an example NARA [10] and Library of Congress [6] both offer very good guidelines. Furthermore, virtually every office-, image manipulation- or e-mail program is capable of producing some kind of archival format, generally PDF/A. However, it is always the native format that will be selected by default and it is the users' responsibility to recognize and pick the true archival format from the large list of different formats. From the authors' opinion, there is no option that an average Joe or Jane will read hundreds of pages of technical specifications just to be able to save file in a true archival format with correct metadata structure.

This issue with everyday formats vs. archival formats is too complicated and it is principally wrong to obligate the ordinary citizen to handle it. The format migration part should be automatic when the electronic item is ingested into an archive. This case describes the studied and implemented process of transforming proprietary e-mail formats into a fully qualified archival format which contains all the original metadata and attachments of every individual e-mail. This work has been done at the Digitalia by utilizing solely open source products, Linux, Python and Java programming.

There are many online e-mail providers, such as Google and Microsoft. However, the focus of this paper is in the dominant of the business side, Outlook. Later on the functionality will be extended to cover other providers as well. Outlook directly supports the saving of individual e-mails in .msg or .htm format for example, but it also support the exporting of complete folders or the whole e-mail account as .pst file. POP3, IMAP as well as web-based e-mail accounts rely on .pst files while Exhance uses .ost files to store Exhance account as an offline working copy. Also if Outlook.com account has been connected to Outlook, the contents of the online account are synchronized with a local .ost file [9].

Workflow

The whole conversion workflow is based on publicly available open source products, which are bound together by utilizing Python scripting and Java. Although, the conversion utilizes multiple software, the end user only need to provide the source file and the rest will happen automatically. Conversion itself is handled by Python which will utilize the required open source programs to complete the conversion. Next listing shows the programs that are executed during the conversion process.

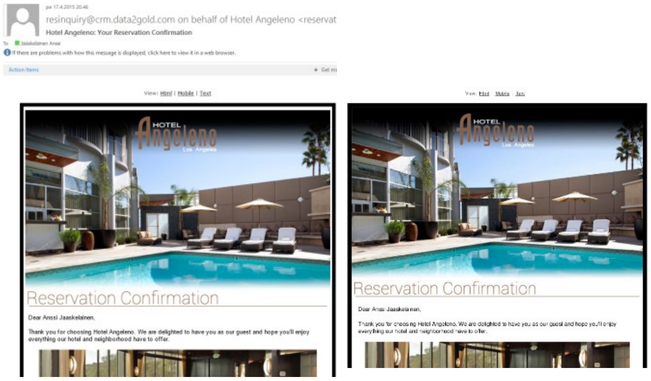
1. pffexport
2. ted / txt2html
3. wkhtmltopdf
4. gs
5. convertmetadata.jar
6. verapdf

- **Pffexport**[5] is a small tool that is included in Linux libpff library. The main purpose of this tool is to export items stored in Outlook personal folder files (.ost or .pst). This tool will browse through the provided mailbox file and extracts its content into original folder structure so that each individual e-mail is extracted finally into its own folder. Possible attachments are also added under this same directory structure. By default this tool will extract the e-mails as text but this can easily be changed to either .rft or .html to preserve the original look and feel of the e-mail. We are utilizing 'all' option to extract all possible e-mail versions and after the extraction, preferred (html → rtf → txt) is taken into the workflow.
- **convertmetadata.jar** is Digitalia made Java program that reads the metadata exported by pffexport and converts it into a format supported by Ghostscript. This same program creates .csv file which contains header information of every converted e-mail. This .csv can be fed into a visualization or analysis software such as Gephi.
- **Ted**[6] **/ text2html**[7] Ted is a full scale text processor but it also contains very good format conversion tools, in this case it is used to convert .rtf files into HTML format. Txt2html is a Perl written application that converts plain text to html format. This workflow utilizes one of these programs in case the pffexport did not directly produce .html formatted e-mail.
- **wkhtmltopdf**[8] program converts HTML files into PDF format by utilizing the Qt WebKit rendering engine. This program can be run in headless mode so it doesn't require display or display service.
- **oowriter and imagemagick**[9] are used when original e-mail attachments are converted into formats that are better suited for archiving. Imagemagick is an independent program while oowriter is part of OpenOffice or LibreOffice suites.
- **GhostScript**[10] is probably one of the best known PostScript processors in the world. It has been around for ages and therefore can be considered as reliable software. Versions after 9.19 contain support for PDF/A-3b, which is also our final target format. Ghostscript also writes the converted metadata back into the generated PDF/A-3b file as well as attaches the original e-mail attachments to PDF file.

[5] http://www.forensicswiki.org/wiki/Libpff
[6] https://www.nllgg.nl/Ted/
[7] http://txt2html.sourceforge.net/
[8] http://wkhtmltopdf.org/
[9] https://www.imagemagick.org/script/index.php
[10] http://www.ghostscript.com/

- **Verapdf**[11] is the final phase of the conversion. This software will validate the produced PDF/A-3b files against the latest standards

At the current stage of the development, this workflow can be run either directly from the command line or by using the citizen archive UI. Complete processing time for about one GB e-mail box with 10800 e-mails is about 11 minutes with a 16-core server. Figure 2 demonstrates the transition from original e-mail into a valid PDF/A-3b file.



pe 17.4.2015 20.46

resinquiry@crm.data2gold.com on behalf of Hotel Angeleno <reservat

Hotel Angeleno: Your Reservation Confirmation

To Jääskeläinen Anssi

If there are problems with how this message is displayed, click here to view it in a web browser.

Reservation Confirmation

Dear Anssi Jaaskelainen,

Thank you for choosing Hotel Angeleno. We are delighted to have you as our guest and hope you'll enjoy everything our hotel and neighborhood have to offer.

Reservation Confirmation

Dear Anssi Jaaskelainen,

Thank you for choosing Hotel Angeleno. We are delighted to have you as our guest and hope you'll enjoy everything our hotel and neighborhood have to offer.

Figure 2: Left, original e-mail in Outlook, right converted PDF/A-3b file

Some might wonder why we have not chosen html 5 or some other more manageable format. 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, the transferability of the document is superb. Finally, this solution maintain the original look and feel of the e-mail and makes possible the preservation of original metadata and attachments.

Future development

During the forthcoming months, the development process for both the citizen archive and the e-mail migration workflow will continue. We will also be extending the test user base and alter the functionality, usability and visual appearance according to the gained feedback.

At the current stage of the development, the format migration workflow only supports Outlook data files. Our primary mission is to extend this support to cover other e-mail formats as well. Secondly, it is our intention to produce an independent website for e-mail format migration before the conference that can be used for demonstration and testing purposes.

For the citizen archive, as part of our agile product development, we have collected feedback from our pilot customers continuously. According to feedback the future personal archivists generally prefer usability and availability. This requirement has been responded with web based UI, which is seen as a great advantage. It provides device-, browser-, time- and place independent easy access into the citizen archive. The content management, data life cycle management, and the ability to provide access to archived material improve both the usability and the distribution of the archived content. At the same time, digitized materials were seen as a way for protecting the fragile originals and help to distribute the content. The quality control of the digitized material before ingesting it to archive was also considered as an important future part of the workflow. It should be guaranteed that the file and metadata are accurate. The abilities to extract the archival content and transfer it to another archive were considered important as well as being able to import data from other storage systems. In some case the transformation of the ownership of the personal archive is needed when an archivist wants to pass the digital archive to heirs.

So far we have improved the search and the browsing properties of the material, because archive owners usually want to share collections by giving some people an access to its content. Actions to facilitate the usage of the archive are still needed. Remaining archival terms in UI should be replaced with common terminology. Separate help system and tooltip texts are needed as well to display informal aids in an application. Each archive owner i.e. an administrator preferred the customizable main view of the personal archive. It is the logical place to inform other users of interesting contents. Each archive and archive owner seems to have individual needs for user management. Citizen archiving application has so far certain user groups available, but they need to be editable. Archive owners want to determine what user interfaces are available at different user levels. Multitenant architecture used in the design of a citizen archive enables us to meet these requirements.

Conclusions

This article illustrated the current development of an OSA based personal archive for citizens. The long-term storage and maintenance of personal digital information brings social, technical, and legal challenges. We collaborate with leading Finnish specialists in information law and information security. We develop the Citizen archive platform together with our users, aiming at continuous improvement and a better user experience. In line with [11], we believe the needs of citizen archivists serve as a starting point for any development work in this field. By utilizing iterative development, it is simple to see how effectively the different versions of the citizen archive meet the user requirements. One important aspect that has been identified so far is automation. As we pointed out in our case example of citizens' e-mail archives, automated format migration is the only way to go, as citizens are not experts with electronic file formats.

In contrast to the original situation, a personal archive of tens of thousands of e-mails can now be transformed into a package of files than can be further analyzed and visualized by family archivists or researchers, e.g. by using social network analysis tools. Furthermore, even without the utilization of the citizen archive, it is simpler to share interesting PDF/A files than complete e-mail accounts or messages exported from the e-mail application. Principally it is also wrong to obligate the ordinary citizen to handle this issue. The format migration part should be automatic when the electronic item is ingested into an archive.

Overall, the resulting archive system fulfills a major gap by providing a solution for the reliable long-term preservation of citizens' digital material. In contrast to physical objects, the value

[11] http://verapdf.org/

of immaterial resources such as digital information is that one collection can be accessed from multiple places at one time, and remain connected to a certain person even after his or her death [5]. People have various digital collections of which they select one greater collection to be preserved as their digital heritage. The citizen archive keeps personal life stories found and accessible for future generations and perhaps in the future also publicly available for anyone who is interested in the story behind this precious information.

References

[1] A. Jääskeläinen, Rationalizing the concept of user experience in digital preservation, Proc. Archiving 2012, pg. 195. (2012)

[2] A. Jääskeläinen, T. Vuorikari, Smoothing away the relic of the past: Case archival control UI, Proc. Archiving 2014, pg 219. (2014)

[3] D. Anderson, Preserving the digital record of computing history. Comm. ACM 58, 7 (2015).

[4] Donald Hawkins, Personal Archiving: Preserving Our Digital Heritage (Medford, NJ, 2013)

[5] E. Carroll, J. Romano, 2011. Your Digital Afterlife. When Facebook, Flickr and Twitter Are Your Estate, What's Your Legacy? New Riders, Berkeley, CA.

[6] Library of Congress, 2016. Recommended Formats Statement 2015-2016.

[7] M. Lampi, O. Alm, Flexible data model for linked objects in digital archives, Proc Archiving 2014, pg. 174 (2014).

[8] M. Lampi, O. Palonen, Open Source for Policy, Costs, and Sustainability, Proc. Archiving 2013, pg. 271. (2013).

[9] Microsoft. Introduction to Outlook Data Files. DOI=https://goo.gl/6Cz8MB

[10] NARA. 2014. Revised Format Guidance for the Transfer of Permanent Electronic Records. DOI= http://goo.gl/Vp2ntJ[10]

[11] P. Uotila, Using a professional digital archiving service for the construction of a family archive, Proc Archiving 2014. pg. 188 (2014).

[12] S. Pimminger, J. Heinzelreiter, W. Kurschl, A. Lindley, Themis – Conserve Your Digital Life, FFH2015-CBS1-3.

[13] T. Lowdermilk, User-Centered Design (CA, O'Reilly Media, 2013)

[14] William Jones, Keeping found things found. The study and practice of personal information management (San Francisco, Morgan Kaufmann, 2008.)

Author Biography

Anssi Jääskeläinen has an IT MSc. (2005) from Lappeenranta University of Technology and a PhD (2011) from the same university. He has an extensive knowledge of user experience and usability. His current interests are in format migration and open source development.

Miia Kosonen holds a PhD from Lappeenranta University of Technology (2008). She is an experienced researcher and trainer in knowledge and innovation management, open communities, social technologies and social media. Her current interests are in the field of digital communication and preserving digital data.

Liisa Uosukainen has M.Sc. (Tech.) from Lappeenranta University of Technology (1994). She has years of experience in software development. Her current interests are in digital data and digital archiving.