https://doi.org/10.2352/issn.2168-3204.2019.1.0.30

Challenges in the Cloud of Personal Archives

Hugo Quisbert, ArkivIT, Stockholm, Sweden.

Abstract

In this paper we address some challenges in the development of personal archives. Personal archives can be seen as a collection of archival holdings connected to one specific individual during its lifetime. The collection can be the result of personal activities in private life or part of activities in within an organisation. In order to develop easy-to-use and low-cost archival services for individuals, any company need to carry out a pre-scanning over the challenges that the company may confront during development. Some of the challenges we could identify in advance where connected to the following areas: Characteristics of personal archiving services, the user, underlying technical infrastructure, portability, security, pricing and the market itself.

Introduction

In this paper, which describes work in progress, I address some challenges in the development of personal archives. Personal archives can be seen as a collection of archival holdings connected to one specific individual during its lifetime. The collection can be the result of personal activities in private life or part of activities in within an organisation [2]. Personal archives and personal information management has being paid more attention lately, however the field is still under-researched. [10] Personal archive services (PAS) are not established as the primary option when individuals approach to archive their personal holdings but rely in popular cloud drives like Dropbox or Google Drive for securing the longevity of their holdings. One of the first questions we confronted was how to make PAS low cost, attractive and better option than like the mentioned and other free of charge services? During a series of workshops within our company we could identify a set of challenges connected to the design, development, deployment and marketing of PAS. Some of the identified challenges are connected to the following areas: Characteristics of PAS, the user, underlying technical infrastructure, portability, security, pricing and market. When discussing the challenges, we also tried to envisage (to a quite low extent) some kind of "solution" to every challenge.

Personal Archives

What is the main difference between traditional archives and personal archives? Within organisations archives are created as part of the organisations collective memory and reflects the processes of the organisation. Traditionally, archive creators have been organisations in the public administration. In Sweden the Archival Act (Arkivlag (1990:782)) [2] regulates the creation, appraisal and long-term archiving of records created within public administration domains. There is a long tradition in the archival practice to deal with paper archival holdings; however even archiving electronic records is rather new phenomenon, archivists are getting familiar with archiving electronic records.

Personal archives on the other hand, reflects the individual, its work and its own style of organising the holdings or as McKemmish describes [8] as "Evidence of me …" Because of the digital revolution, individuals have the possibility of creating relatively huge amounts of data, which need to be organised, documented, preserved, accessed and even shared or redistributed. In most cases, the creation, documentation and organisation of personal holdings is ad hoc and not regulated by laws, neither standards are supposed to be used. Moreover, personal archives tend to be private and may be accessed by a small number of stakeholders. One more and important difference is that every single item within the set of holdings may be subject to deliberate alteration. That means that the holdings are alive during and perhaps beyond the individual lifetime. The collections of individuals, might be kept because of affection and not because the strict following of regulations, as is the case in traditional archiving.

Problems with data created and collected by individuals

As stated, the digital revolution has caused a huge creation of data and thus even created unexpected problems related to digital holdings of individuals. I see two major categories of problems in this regard. I call them the *accumulation problem* and the *evaluation problem*. The accumulation problem is caused by using backups as "archives", the process of mass-copying files from one computer to a newer one, the replication of "valuable" files in movable media (as USB-drives), the retentions of e-mails including attachments as an archiving activity, and even the preservation of old computers to access specific files using very special software. Some of these issues are described in [1] Marshall [7] states several problems connected with the long-term archiving of personal data: (1) people find it difficult to evaluate the worth of accumulated materials; (2) personal storage is highly distributed both on- and offline; (3) people are experiencing magnified curatorial problems associated with managing files in the aggregate, creating appropriate metadata, and migrating materials to maintainable formats; and (4) facilities for long-term access are not supported by the current desktop metaphor. The problems (1) and (3) comprises the evaluation problem. However, I also encounter a *preservation problem*; which comes attached to problem (4).

The Challenges

As stated in the introductory section of the paper, we identified challenges in the following areas: Characteristics of PAS, the user, underlying technical infrastructure, portability, security, pricing and market. Naturally there are much more challenges, but for now we focus in the mentioned.

Characteristics of PAS

Our prospective services can be seen analogically as the set of applications in MS Office 365; interconnected and comprising a bigger wholeness. Bearing that in mind, then it is necessary to think about PAS in terms of information systems. That means current properties and features of information systems need to be attributed to PAS. That is the easy part of it. What is the challenge in this case? To answer that question we borrow some inspiration from the Personal Information Management field. In [3] are some design and development principals proposed as follows: (a) The *subjective classification principle* stating that all information items related to the same subjective topic should be classified together regardless of their technological format; (b) The *subjective importance principle* proposing that the subjective importance of information should determine its degree of visual salience and accessibility; and (c) The *subjective context principle* suggesting that information should be retrieved and viewed by the user in the same context in which it was previously used. The authors claim that a user-subjective approach should help the user find the information item again, recall it when needed and use it effectively in the next interaction with the item. Why we chose this perspective? The way we analyse this issue, leas us to concentrate in the creativity of the individual thus, the major issue is then to retain the individual style of organising the holdings in such a way he or she might recognise themselves in our commercial services.

In other words, our services have to show a high grade of *trust* and *transparency* towards the user.

We see implementing specific characteristics of PAS as a major development challenge and we urge to take a deeper look into this matter.

The user

This we consider a real challenge due to is almost impossible to know in advance who is going to be the customer and what needs or requirement she or she may have. Nevertheless, we envisage two main categories of users: the *average user* and the *sophisticated user*: *The average user*, is the one who might have holdings comprised of family photos and videos, document as wage slips, vouchers, receipts, invoices, etc.

The *sophisticated user* has more special needs and has a special interest in creating and collecting some specific type of holdings. In this category we identify some recurrent profiles: *The photographer,* who may need to archive large quantities of photos. *The designer* who produces drawings and sketches. *The writer* (whom might be a journalist as well), who produces extensive texts, blog posts; for the journalist the production may include photos. *The musician*, whose production may comprise large amount of sound files. *The filmmaker* is the one whose production may contain large quantities of films, video clips and other moving picture files. On profile within the sophisticated category that we identify as a bit different, is *The University Professor*, whose production may comprise of different kind of publications, books, presentation, educational material; everything created as a part of a professional career within different educational organisations.

Technique

In the technical track, several challenges need to be addressed such as the integration to other daily tools such as internet banking, social media like Facebook, etc. This requires secure and approved and allowed API:s at all these tools. In the case of Facebook, our company is carrying out a project, which aim is to harvest social activities of public organisations made in Facebook. The results (unfortunately could not be presented here) are promising and will considered as part of the implementation of our PAS.

However, managing the files of the holdings will be a challenge due to the fact there are no standards ruling the use of specific formats. In fact the holdings may contain quite rare file formats, difficult to convert to standarised and preservation friendly formats. Nevertheless, there is a quite extensive research and guide lines from the field of digital preservation when it comes to the recommendations of choosing preservation friendly file formats. We will follow established guidelines and thus making the digital preservation transparent for the user. Wes summarize a set of file formats that we intend to use in our services. When selecting file formats for archiving, the formats should ideally be: Non-proprietary, unencrypted, uncompressed. Some preferred file formats:

- Containers: TAR, GZIP, ZIP
- Databases: XML, CSV
- Geospatial: SHP, DBF, GeoTIFF, NetCDF
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII, DTA, POR, SAS, SAV
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Tabular data: CSV
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Web archive: WARC

The subject of standards requires a bit of attention. Because on one hade we could show standards connected to file formats, but on the other hand the standardised development of PAS, from a information systems perspective needs a bit of attentions. Since we are addressing archiving services with emphasis in “archiving” we need to focus on standards for the development of archiving systems, and naturally we have to take the OAIS Reference Model into account. [6] This leads us into the question, how we address personal archives in the light of the OAIS reference Model? As we see it, the difference is the designated community, which is comprised of the categories of stakeholders. One is the producer is that at the same time is also the consumer, and the other is a small amount of stakeholder that are granted access by the producer.

Connected to the managing of files and their formats, managing metadata and aggregated metadata that form information objects to be archived is another issue. In the area of technique, we also conceive the ingest process and the access as challenging due to users nowadays use different platforms (i.e., smart phones, tablets, etc.). Nonetheless, there is a need of testing towards more integrated IT-systems (like ERMS) and huge amount of file types.

Portability

What happens if a customer wants to move to another supplier of archive services? In that case a set of export, migration routines and API:s needs to be developed. However, in writing moment is difficult to address the issue since there are no actors in the market and we don’t know about other actors API:s or technical infrastructure.

Security

Since cloud technology is assumed for the services, a public and publicly-known product (consumer product) is at greater risk

of targeted attacks than traditional (internal) e-archive platforms. That means, in order to our company provide PAS, we need to develop our services bearing in mind the recurrent set of thearts that exist for clouds nowadays, including Data security, Network security, Data locality, Data integrity, Data segregation, Data access, Authentication and authorization, Data confidentiality, Web application security, Data breaches, Virtualization vulnerability, Availability, Backup, Identity management and sign-on process. One natural question in this matter is, whis security issue is more important than other when it comes to PAS? Naturally, we start from an archivist's point of view there there undesired alteration of data is a major issue. That means, Data integrity is crucial, which of course depends on how good access, authentication and authorization has been implemented. Moreover, other security issues may include age requirements for users (minimum age), GDPR-requirements, management of user rights.

Summing up, the area of security is very extensive and we don't see this field as a major issue, however this needs to be taken into consideration when developing PAS, and as we know, security issues has its own complexity.

Pricing

How much will it cost? I writing moment is very difficult to give a complete price model and list of the services. Nevertheless, one idea is to build a progressive pricing model in which services are graded from simple (for instance storage), moderate to sophisticated services and the price progresses with the grade of complexity of the services. We got some inspiration from the work of Laatikanen & Ojala [5], especially the described value-based pricing. The difference may be we transfer some internal pricing factors (such as personell skills) to how valuable are the holdings for the customer. Anyhow, taking into account the plethora of cloud storage services, we are prone to outline a pricing range from 3 € to 40 € per month, bearing in mind that our services may offer more technical features than competitors even in simpler services.

Market

As stated in the introduction of this paper, a big challenge is to compete with free of charge services in popular cloud drives. A low cost and easily accessible product is more likely to compete if it provides clear benefits for the user. Since our market will be the primarily the Swedish market, we need to focus on and even get inspired by presumptive competitors. There are some few providers of "digital post boxes" as free services such as Kivra (www.kvira.se), eBooks (www.ebooks.com), Digimail (www.digimail.se), Min myndighetspost (www.minmyndighetspost.se). The mentioned services are used by individuals to receive and store e-mail from all kind of authorities and some banks and insurance companies. To compete with that kind of services may force our company to provide same kind of service for free or a free service attached to other simple service.

There are a plethora of cloud storage services, both free and paid. However, we imagine that automatic and seamless transcoding to the latest version of a selected set of file formats, may give some market advantage towards other services that offers just storage and there it is up to the user to convert its holding to newer file formats. Naturally, the use of PDF and PDF/A should be default in our services.

In writing moment we cannot find a competitor providing an interconnected set of services as our envisioned in the current market, rather a plethora of services or applications specialised in certain features of personal information management and storage of personal holdings.

Nonetheless, in order to be market competitive, our planed services need to be tested on a wide user base to address a good level of market acceptability, thus addressing the right customer base is of significate importance.

Discussion and Ideas for the Future

One crucial question that needs to be answered is whether our approach would solve the problems described in earlier section. An examination of the problems and the challenges we address show that the answer is yes, but just partially.

First of all, the services will provide real archiving features and thus the "backup issue is eliminated. The replication problem may also be mitigated, if the user uses our services as the only source of storage. For now, there are no plans for providing performance services (i.e., emulation) for rare types of file formats. Since our services will be designed and assessed by professional archivist, the services will provide mechanisms for metadata aggregations, making the evaluation of the holdings easier. We can also affirm that the preservation problem will be solved due the use of established best practices in digital preservation.

Addressing challenges requires a bit of thinking in advance in order to be successful in the long run. The discussions within our group working with this topic are guided by the lemma "what is the benefit for the customer?" Obviously, the long-term access to the individual holdings of the customer, by a low price, is the natural answer to the lemma. One recurrent issue is the question how to make the archive attractive. Our thinking goes to the notion of a place of "story telling" mixed with natural interaction (sound, touch), similar to the Living Box metaphor [9]. Moreover, we have seen as good contribution to our work the set of recommendations of the service model for personal archives as in [7].

References

[1] Allegrezza, Stefano. "THE FUTURE OF OUR PERSONAL DIGITAL MEMORIES: IT'S TIME TO START THINKING ABOUT IT." *Atlanti*+ 29.1 (2019): 55-65.

[2] Arkivlagen (1990). Swedish Archival Act. 1990:782. < https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/arkivlag-1990782_sfs-1990-782>

[3] Bergman, Ofer, Ruth Beyth-Marom, and Rafi Nachmias. "The user-subjective approach to personal information management systems." *Journal of the American Society for Information Science and Technology* 54.9 (2003): 872-878.

[4] Kim, S. "Personal digital archives: preservation of documents, preservation of self." Doctoral dissertation, University of Texas. (2013).

[5] Laatikainen, Gabriella, and Arto Ojala. "Pricing of digital goods and services." *Information Systems Research Seminar in Scandinavia*. IRIS Association, (2018).

[6] OAIS Reference Model, ISO 14721. www.oais.info, 2019.

[7] Marshall, C. C., Bly, S., & Brun-Cottan, F. "The long term fate of our digital belongings: Toward a service model for personal archives". In Archiving Conference (Vol. 2006, No. 1, pp. 25-30). Society for Imaging Science and Technology. (2006).

[8] McKemish, S. "Evidence of me …". Archives & Manuscripts, Vol. 24, Issue 1. (1996).

[9] Stevens, M. M., Abowd, G. D., Truong, K. N., & Vollmer, F. "Getting into the Living Memory Box: Family archives & holistic design". Personal and Ubiquitous Computing, 7(3-4), 210-216. (2003).

[10] Williams, Peter, Jeremy Leighton John, and Ian Rowland. "The personal curation of digital objects: A lifecycle approach." *Aslib Proceedings*. Vol. 61. No. 4. Emerald Group Publishing Limited, 2009.

Author Biography

Hugo Quisbert received 2009 his PhD in Computer and Systems Science from the Luleå University of Technology in Sweden. His research interest covers Digital Preservation, Electronic Archives, Open Data and Information Systems in general. Since 2016 he is coordinator of Research & Development at ArkivIT in Stockholm, Sweden, and working as an expert consultant within in electronic archives and even as Data Protection Officer.