Open Source Archive

Liisa Uosukainen
Mikkeli University of Applied Sciences
Patteristonkatu 2
P.O.Box 181, 50101 Mikkeli
liisa.uosukainen@mamk.fi

Anssi Jääskeläinen
Mikkeli University of Applied Sciences
Patteristonkatu 2
P.O.Box 181, 50101 Mikkeli
anssi.jaaskelainen@mamk.fi

ABSTRACT
It is inevitable that business and industrial world will soon meet one of their biggest challenges so far. How to govern a cross platform distributed information which physical location is unknown? Currently the best practice has been via restrictions and rules but this method is inoperative with modern users. This paper describes how we at the Mikkeli University of Applied Sciences have started to resolve this dilemma.

Categories and Subject Descriptors
H.1.2 [**Models and principles**]: User/Machine Systems– human factors, human information processing.

H.3.4 [**Information storage and retrieval**]: Systems and Software–Distributed systems.

General Terms
Management, Design, Human Factors

Keywords
Open source, digital archive, user orientation.

1. INTRODUCTION
The utilized storage technology renews regularly. 5.25, 3.5, HDD, SSD and optical media. These represent the local storage technologies that are in use or have been used in home environment. Currently, the movement has been towards social media and clouds, so when something is created it won't be stored locally. Ipads' can be linked with the DropBox, Android devices use GoogleDrive as a backup place and so forth. From the user point of view this behavior is very welcome, thus it releases user from doing things manually. Furthermore, the content in a cloud is platform independently accessible virtually anytime and anywhere.

While the end users generally love this new way of sharing and distributing information, the IT management of a company will not. They have been working hard to keep the hardware and software structure homogeneous. It is only a matter of time before the cloud era will inevitably take possession of the business and industrial world as well. This has already began and the IT management will meet one of their biggest conundrums ever. How to govern a cross platform distributed unstructured information which physical location is unknown?

2. DISTRIBUTION, A FRIEND OR A FOE
Even though the IT management is struggling with the introduced problem, from the author's point of view, this is a wrong way to start solving this dilemma. The problem has already taken place since workers are using services, applications and clouds that are not maintained or supported by the IT management. Therefore the question that should be asked is: "Why our workers use cloud drives or personal devices to manage their information?"

There can be numerous reasons for this behavior including defective operation, slowness, missing support, usability issues, a fuzzy user interface (UI), illogical workflow or just general bad user experience [1]. Furthermore, if a user that is comfortable with single sign-on and modern web UI:s is forced to use some awkward old system, there will be resistance and disobedience. For example, the IT management of our university does not support the usage of GoogleDrive, but everybody is still using it.

Utilization of personal devices or third party services shouldn't be seen as a troll. Instead it offers many possibilities. Bring your own device mentality can for example reduce data administration costs, number of commercial licenses, hardware costs as well as maintenance costs [6]. Furthermore, users are happy when they can use devices, services and applications that they are comfortable with.

In spite of the benefits, when something is changed radically the resistance to change will be an issue. Therefore the change should be made with light steps that users have time to adapt. When users feel that they had a possibility to affect, they are more open towards changes [4]. This is something that needs to be in mind when the information management policies are changed or new technological tools are introduced to the end users.

3. SUGGESTED SOLUTION, OPEN SOURCE ARCHIVE
The problem of divergent information is enormous, modern users prefer clouds and use those in spite of policies while older generations like to keep things as they were. Our solution is to enhance the existing digital archive system so that it still relies on rules from the older generation but modernizes the utilization and ideology with open source and novel search methods.

Although we speak for open source, we realize that it is not an answer to everything. It cannot for example change the information security policies or established best practices. However, it offers a possibility to show that things can be done differently e.g. with half the price and higher customer satisfaction. Eventually, these observations can be the initiating force that leads to a big policy changes on the company level.

Open Source Archive (OSA) project, which is a basecamp for the solution is an ERDF project that started in May 2012 and will run until the end of December 2014. The focus of the OSA project is to identify and develop solutions that provide value and new kind of archival user experience for the current and future users of digitally archived data. The principal aim of the OSA-project is to develop user oriented archive solution for preserving, managing, and providing access to digital content. The project is carried out by Mikkeli University of Applied Sciences with multiple partners such as archives, software vendors, service providers and educational institutes. Users inside the partner organizations have been the source for the end user wishes and needs. Figure 1 presents the

simplified structure of the OSA solution and the area of our development focus.

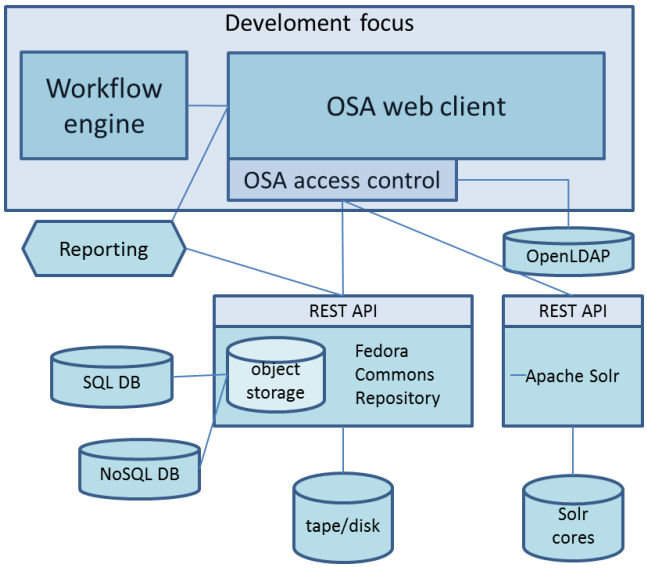

Figure 1. Simplified structure of the OSA solution

3.1 Preservation and data analysis

The key feature of the OSA archiving system is data lifecycle management. We intended to develop a solution that supports the core processes of archives and other memory institutions. Our focus was to develop a customizable solution to ensure that each organization, or even an individual user, can configure the archive to meet their particular needs. The predecessor of the OSA archive system has for example used successfully in creating a family archive [5] and OSA development partially relies on this success.

Metadata creation is another key feature of OSA. This is made flexible and configurable for all defined types of digital objects. Metadata creation is automated as far as possible in order to prevent users from doing irrelevant things or accidental mistakes. Naturally, a possibility to define metadata fields individually for different digital object types (collections, documents, images, audio recordings, moving images, etc.) is also present [3]. During the ingest process, the OSA system will capture the metadata of objects and runs normalizing procedures. Furthermore, the usage of entities (agent, place, event, action) is supported in describing the archival objects. The entity related metadata can be partly defined by using available glossaries or ontologies. The utilization of contextual entities in describing objects links digital objects to each other and therefore makes it easier to search information from the archive.

Automation, to some extend can be done in archives. For that purpose we used the workflow engine that has been developed as a bachelor thesis in this project [2]. Some processes were modeled as micro services and combined into workflows. A set of micro services were created for pre-ingest workflow and ingest-workflow covering virus checking, technical metadata capturing and normalizing, checksum checking, and preview generation. There is a possibility to trigger workflows automatically when files are uploaded to a certain directory. The workflow engine is expandable; new data processing functionalities are fast to create and add to organization's workflows. With the utilization of such a technology the actual user involvement during the archiving process can be minimized and thereby also the mistakes done by the user.

3.2 Search and access

The end users access the OSA system via web client. Search and index features visible to end users can be configured to meet their particular needs. For example search form, columns in the result layout and facet fields are all fully configurable. With the utilization of linked objects and faceted search, the ingested data is more accessible and easier to understand and reuse. Finally, OSA uses role based access control to ensure, that data is accessible only to permitted user according to the given rights.

3.3 Architecture

The OSA system is based on a service oriented archiving application written in Java. It provides user friendly access via web client to the implemented features. The application is a coupled set of software components used via common API (Application Programming Interface) [4]. As much focus as possible was given on sustainable software development and selected community-driven, open source components.

OSA is based on Fedora Commons (Flexible Extensible Digital Object Repository Architecture) repository module. Fedora Commons provides a framework for modelling digital data and to build archiving services. OSA-application utilizes other open source technologies and tools like MariaDB as a relational database, Apache Solr as a search platform, MongoDB as NoSQL store and OpenLDAP for authentication purposes. Results of the project will be released as open source at the end of this project.

Red Hat Enterprise Virtualization (RHEV) running on blade servers was chosen for building and managing cloud IaaS (Infrastructure as a Service) environment. RHEV is based on Kernel-based Virtual Machine (KVM) hypervisor and *oVirt* open virtualization management platform [7]. Virtualization is utilized to ensure scalability and capacity for future development and services. Finally, all original files of ingested objects will be stored into tape drives for long-term storage.

4. CONCLUSIONS

The work for OSA system is still under way and this paper described the current development phase. The most important aspect of this paper is the highlighted distribution problem and the positive feedback received considering the suggested OSA solution. We have utilized the information gained from the users to automate the ingest workflow as far as possible. We suggest that the end users should be brought into the process as soon as possible in spite of the area of development. The end users commonly have a better understanding of what they want to have than a bunch of regulators or designers sitting in the ivory tower. In generally speaking, from the authors' point of view there are only two ways to manage the inevitable change to 21st century, resist it to the bitter end or go with the flow.

5. REFERENCES

[1] Cooper, A. *The inmates are running the asylum.* Sams Publishing, 2004, USA

[2] Kurhinen, H., Lampi, M. 2014. Micro-services based distributable workflow for digital archives, *in Proceedings of Archiving 2014*, (Berlin, Germany, May 13-16, 2014)

[3] Lampi, M., Palonen, O. 2013. Open Source for Policy, Costs, and Sustainability, *In Proceedings of Archiving 2013*, Archiving 2013 (Washington DC, USA, May 13-16, 2013)

[4] Lowdermilk, T. *User-Centered Design*, O'Reilly, CA, 2013.

[5] Uotila, P. 2014. Using a professional digital archiving service for the construction of a family archive. *In Proceedings of Archiving 2014*, (Berlin, Germany, May 13-16, 2014)

[6] Zielinski, D. *Bring Your Own Device*, HRMagazine, Vol 57, 2, 2012.

[7] Chen, K. *Red Hat Enterprise Virtualization – White paper*, http://www.redhat.com/en/files/resources/en-rhev-idc-whitepaper.pdf