# Open Source for Policy, Costs and Sustainability

*Mikko Lampi and Osmo Palonen, Mikkeli University of Applied Sciences, Mikkeli, Finland*

## Abstract

*This paper gives the reader a brief presentation of a decade-long experience of MUAS developing services and applications for digital archiving. Based on this experience it is shortly explained why Open Source was not the option a decade ago and why it will be the future of digital archiving. The results of the recent survey made in project OSA (Open Source Archive are also explained. It covers topics like which are the selected tools and environments that will be used in the project when building IT-systems for archiving.*

## Background: Digitalization of archive

A decade ago many archivists thought that they have to have the data in-house the same way as the paper documents. To get the operational model selected in MUAS accepted in the archiving community was not a piece of cake on that time. The question was not only to understand the new model of IT service at the same time the new paradigm of archiving was taking place: archives are not for preservation only – more important is to provide information for their customers.

The pressure was rising also from the change of the customer profile in the archives: Most of the visitors in the archive are no more academic researchers like historians. What is important, they do not have the training to understand how the archives were structured and build; many of them did not care, they just Google. Instead, they want everything on-line acting as virtual visitors.

In this advanced service requested, the often used phrase "Digitize or die" did not mean only to digitize content. Instead the archives had to start digitizing the whole process and operational environment including the IT in archiving. The database-based finding aids or catalogues – the great developments of 1990's – were really not what the customers were asking for.

At the beginning of this digitizing process the archivist were trying to get the IT-systems that supports the new requirements. In many cases, these systems developed in a waterfall project model and based on the specification misunderstood by both participants,the system user and programmer were not what was really the target. The IT-developers, who like to use their standard tools and developing methods, and archivists who had an idea how they like to use the new system, could not communicate on the same level and understand each other. As an example the term "long-term" had quite a different meaning between these two pools of development projects. The archivists thought it is a period of centuries, IT specialists quarters or years.

During this first development cycle of archiving applications there were not many people who were thinking about the sustainability of the systems or what happens when the components used in the system eventually get out of use and the market. This kind of understanding has woken up not earlier than the first generation of archival systems are going to be replaced.

## MUAS as developer and service provider

MUAS has been a decade a major software and service developer for digital archiving and digitization in Finland. During these years, the IT environment has changed. However, it has not affected the serious archiving that much. In this development, MUAS have managed to be forerunners in certain areas. A novelty today is the archive in the cloud – MUAS started a private cloud model of operation in 2004 without knowing that the principle will be revolutionary. We shared resources for multiple customers; most of those services are running in a shared software instance. Only for the medical data the servers and software environment are separated. Our partners ELKA (Finnish business archives) and forest giant UPM have had their audiovisual archives "in a cloud" since 2006 and their contract has been based on currently popular acronym SAAS (Software as a service).[1]

Within these ten years, many things have changed in the MUAS, of course. When the archiving software development started in co-operation with partners like ELKA, we knew that we have to use well-selected open standards to keep the contents alive for decades, but we could not find proper open source tools to build reliable and sustainable services. Instead, we tried to select widely used solutions which were cost-effective. As an example we selected MS SQL Server instead of Oracle and were bound to use Windows Servers at least for the database. On that time, Microsoft solutions were not expensive either, you could simply buy the per server license and not need to pay support, updates or clients. By then nobody could estimate the mess Microsoft licensing is today. Another proprietary selection was made using new Finnish development Profium Metadata Server as RDF-based metadata engine. With that, we got the first bit of open source in use: Profium used Lucene as full text search within the metadata.[2]

## Open standards – not yet open source

When not able find open source tools (in 2004 MySQL or Postgre were not very extensive), the developers in MUAS concentrated on open standards to manage the contents. In archiving the content is the most important. If you can use open standards with the content formats, it is not a great mistake to use some proprietary tools to distribute it. Ten years ago the team in MUAS selected proven commercial products as the corner stones to ensure the reliability of the services and results of the projects. The archival content was preserved using XML, RDF, BWF, MPEG-2/AVI and international and national metadata standards. When it became available in practice, the video format was changed to MXF/Motion JPEG2000 and the document format to PDF/A. The preservation system was build based on OAIS even when it was developed more from an internal idea than knowledge of the standard.

In 2004-2005 the developers in the MUAS understood that in principle open source would have been a better option. However,

there were no communities and only a few components or applications to select. Since then quite many developments have been made in the content management community.

## Good reasons for open source

Today it is the time to prefer open source. There are several reasons why to select open source as the basis when building IT-services for archives and other memory organizations. These reasons are generated from both principled and practical issues as well as long and short term effects.

Money is a consultant in the short term. Most of the memory organizations are struggling with the combination of diminishing public funding and increasing costs. Even when the hardware prices compared with the capacity have gone down, IT is important part of cost expansion: the prices of commercial operating systems and database applications have multiplied compared with 2005.

In the end financial reasons are only one part of the problem, the practical side. More important is the principled side of the issue. To preserve records and documents for centuries have been possible by developing rules and practices in the archiving community. There are guides and regulations for the environment where the documents are kept, as well as methods controlling the access. In the digital archiving world, the control of the contents must be similar: the archives have to have all rights to develop the tools and practices controlling the digital environment by themselves. Preservation cannot be made under the unstable nature of commercial IT suppliers. Open source is based on communities and international cooperation. Building those archival IT communities is the key for the preservation and access.

## How we see the open source ecosystem

As part of OSA project we conducted a brief survey of the open source field. We focused on the archive systems and primary technologies like search tools, databases and infrastructure. However, we noticed the overall change that has been taking place over the past few years.

We found out that open source is now a serious line of business. In addition to being an ideology and a policy on software industry, it is a feasible business model. Open source has developed much from the early days, when it was more or less like a bunch of communities for pioneers and other enthusiasts. Adopting open source required deep knowledge and substantial amount of time and other resources.

Now open source has achieved certain maturity and attracted wide enough audience to be a working ecosystem. A number of major software companies like IBM, Red Hat and Oracle have invested in developing and using open source in their business. This adds credibility to what would otherwise be a scattered field of competing projects; like it used to be, for instance with the Linux distributions. The availability of professional services like consulting and support is a key requirement, and allows organizations to utilize open source without investing in the technical know-how. Open source can completely be outsourced as well. There are plenty of service providers today.

With the Internet, social media and good community hubs available it is easy to become part of the various ecosystems. Open source communities are usually interested in companies and people making good use of their products. Of course, when working with the communities, value should always be given back. It keeps the ecosystem healthy.

## OSA project

OSA is an abbreviation of Open Source Archive. The aim of the project is to find and develop solutions for digital archiving. We have two main priorities and a few other objectives:

- Dark archive solution for long term preservation of digital content
- Service oriented archive solution, which provides value and services for the archived data and physical archives
- Provide knowledge and technologies to higher education
- Participate in and create community to continue the work
- Find partners to develop open source storage in future projects
- Create the final report and project website

The project timeframe is from May 2012 to the end of June 2014. It is funded by European Regional Development Fund granted by South Savo Regional Council. OSA is administrated and carried out by Mikkeli University of Applied Sciences. As partners we have archives, software vendors, service providers and educational institutes. We also co-operate with the National Archives of Finland and others.

## OSA archive software

We have an ambitious vision of project results: two systems which will handle long-term archiving and support the core processes and the business requirements of the archives and other memory institutions. There will also be an open source platform to host the systems. The platform will provide scalability and reliability, and capacity for future developments and services. As usual in real life software development, no project can create a final product that can be declared complete. We do not aim for production readiness but for a stable and operational pilot. Our goal is to find or create a community that would continue the development and make use of the software. It could be the complete system, infrastructure or just some parts and tools.

The dark archive solution is based on DAITSS, which is developed by Florida Center for Library Automation (FCLA). The first version has been in production since 2005, and the second since 2010. DAITSS is used to archive and preserve OAIS packages over a long period of time. It has no access system or interfaces to publicly available websites. We have no plans to drastically change DAITSS or do other customizations; other than the system configuration. The decision was made also to maintain the full compatibility and upgrade path. DAITSS is SOA based and developed with modern tools and programming languages. The architecture and design principles make it easy to extend later, if found necessary.

The DAITSS servers are going to be built with only common hardware. This is to ensure full independence from hardware vendors and proprietary drivers and such. The cost of the hardware is also very manageable and we can scale the environment by adding more servers to the farm. The system will also be mirrored in geographically wide area.

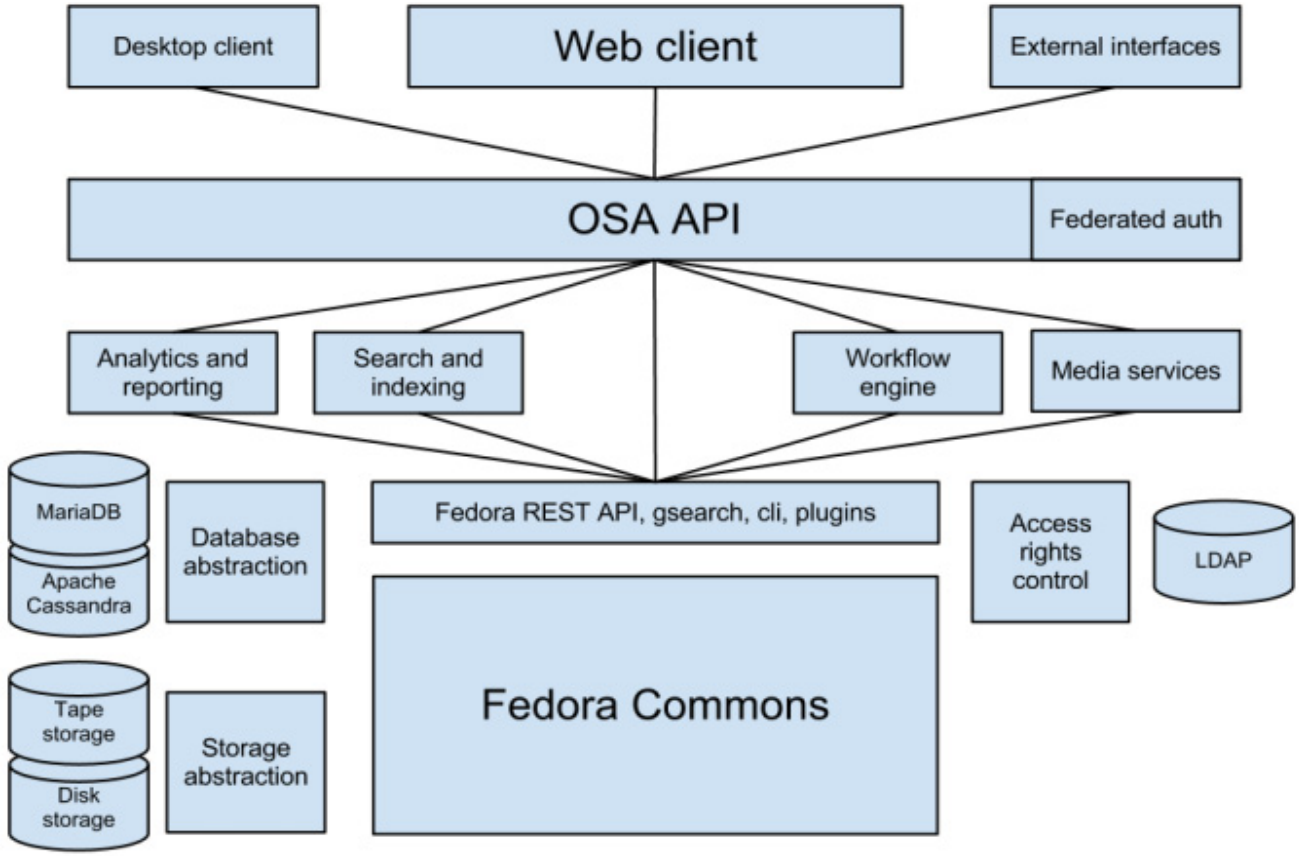


*Figure 1. High-level light archive system architecture.*

Service oriented archive, or a light archive, is a loosely coupled set of software usable via a common API. We will implement a reference web client which consumes the web services, and provided user friendly access to the features. We have an in-house developed desktop client which could be modified to be used with the light archive. Though, it is only for the Windows platform, and therefore not part of the project.

As a core, we use Fedora Commons. It is repository software which can be used like a framework to build services. Fedora is not a ready out-of-the-box solution but instead requires adapting the system to data and processes not vice versa. Before OSA was launched we had a specification project, in which we defined and researched the ideal data models, services and other features of a light archive. Fedora was the only system which could handle the requirements without greatly modifying the software or the internal structure. The other condition was set by our proficiency with Java.

We use SOA based approach to integrate other software with Fedora. REST and other similar interfaces are easy to work with and are not as complex as ESB solutions. Java was chosen as base technology because of its wide spread status, maturity and our existing experience with it. Fedora is made with Java so we can customize and extend it more easily, if required. There was a discussion about other technologies because we were aware of newer more agile languages which are used with several successful Fedora based projects. There is nothing which prevents us from using components made with other tools and technologies as well.

A modern relational database for the system was required for configuration and management data. MariaDB was chosen based on its reputation as a successor to the well-known MySQL. It is more advanced and fits open source better than now Oracle owned MySQL. There is a movement of various open source project migrating from MySQL to MariaDB; like Wikipedia for instance. MariaDB developers have also worked on integration with Apache Cassandra, a NoSQL database engine. We are looking into NoSQL as a more native alternative to storing objects and metadata.

For other services, there are plans to implement well established and stable open source software including:

- Solr for indexing and searching
- Planets or Droid for file characterization
- OpenWMS for ingest
- Eclipse Birt for online reporting
- Nagios for monitoring
- Apache Directory for user management

We have performed research on what tools and projects are being used and which are most actively developed. We hope to be

able to provide back some value to the communities: feedback, bug fixes or new features, like integration packages.

## OSA infrastructure

The third major part of the OSA project is open infrastructure. We found out that there is no point in deploying the software on servers or bare metal, as traditionally has been done. We needed to be able to extend our environment as well as have load balancing, good reliability and disaster recovery. The requirements could be fulfilled with virtualization. We had no resources or the will to use proprietary products. With open source we could rely on understanding the licensing and being able to upgrade components if some tools should become obsolete or there would be better options available in the future. As part of the project, we conducted a review of the available tools and built a test environment for discovered technologies. With community experiences and our own experiments, we chose KVM, OpenNebula and Sunstone.

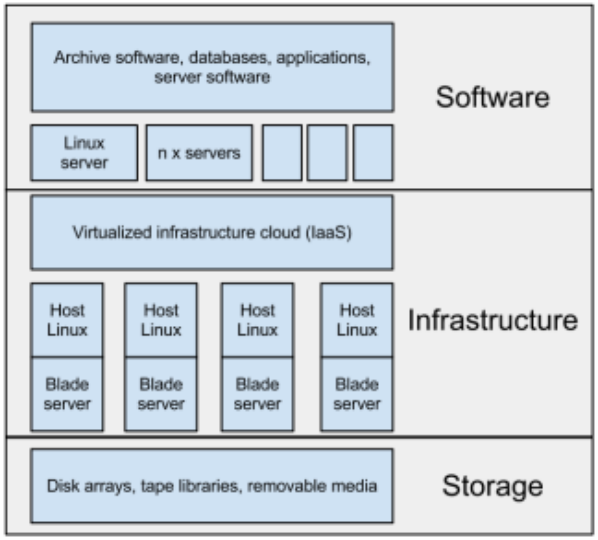


*Figure 2. OSA infrastrure overview.*

KVM was selected as the core virtualization technology. The benefits of KVM include good performance, availability of management tools, and compatibility with a wide range of host and guest systems. In addition, KVM based solutions can be used to virtualize proprietary systems like Windows servers, which makes it a good choice for data centers. IBM, SAP and other major vendors are using KVM and Red Hat has made a commercialized version of it. It has become a trend that open source projects no longer shun proprietary companies and software.

KVM can virtualize servers on one host. However, we required a virtualized infrastructure cloud as a service. For the purpose, we tested a few products and continued further with OpenNebula. It is a cloud computing toolkit, which can be used to build IaaS solutions like computing clusters and virtualized data centers. It can also manage multiple virtualization technologies on the same cloud. These include KVM, Xen and VMware. OpenNebula project has received funding from European Union's Framework Programme. For management purposes we installed Sunstone web interface to manage OpenNebula cluster. MariaDB is used as a backend database for the user management and configuration.

We wanted to make the virtualization hosts as light as possible but wanted full Linux distribution to work with. This way we had full control on the environment, instead of custom bare-metal hypervisor software. We did some experimenting with Ubuntu Server but the enterprise level support for hardware and software was not sufficient. Ubuntu was switched over to minimal install of Centos. Advantages of Centos include binary compatibility with RHEL, thus compatibility with the majority of enterprise software and stable software releases. Centos can be switched to commercial Red Hat Enterprise Linux if professional services and full support are required. In comparison of Centos, Ubuntu had a much newer kernel and software in general. However, Ubuntu community is more oriented for desktops and personal use while Centos is more focused on servers and enterprise use.

## Future developments

Because OSA covers topics from such wide field, there is plenty of room for future developments, and much is left open for new ideas and advancing technology.

One fundamental issue in open source archiving is storage technology. Currently, it is mainly closed and proprietary. We have plans to investigate if there are organizations interested in opening the technology. The ultimate goal would be creating an open software interface for any tape and disk storage. Other parts of the virtual infrastructure can also be worked in future projects: cloud capabilities, open source appliances, performance, compatibility and such.

Archives can contain vast amounts of metadata, relationships and other information. Big data methods and tools could be used to analyze and add value to it. Data visualization is another important aspect to archives. It could attract wider audience to make use of archive data. Especially, businesses could be interested in analyzing the data and providing new services or improving existing operations.

In future, there hast to be a community which will continue the work done in this project. It can be comprised of the original communities, OSA service providers, the users or the like.

## References

[1] ElkaD Loppuraportti (Final report), Mikkeli University of Applied Sciences, Mikkeli (2006).

[2] Palonen-Sirviö: Aton Projektin loppuraportti (Final report of Project ATON). Mikkeli University of Applied Sciences, Mikkeli (2007)

## Author Biography

*Mikko Lampi is a lead designer and system architect in OSA project at Mikkeli University of Applied Sciences. He has a BEng in information technology. He is working with agile development and open source technologies like cloud services, software frameworks and system integrations.*

*Osmo Palonen has been working with Mikkeli University of Applied Sciences since 2003 in developing and leading R&D for information management, digital archiving and digitization. He has BA and MA in history at the University of Tampere. He is also the editor of Faili, the only Finnish professional publication for records management and archiving.*