Access Management and Openness in Digital Archives and Repositories

Mikko Lampi & Johanna Räisä; Mikkeli University of Applied Sciences; Mikkeli, Finland

Abstract

This paper discusses the access management in digital archives and repositories. Access management consists of technical procedures, solutions and policies such as privileges management, access control, user management, privacy management and metadata. In addition, information ethics regarding sensitive materials needs to be taken into consideration. All of these aspects have an impact on access and management of the archival materials during their lifecycle. Access management becomes more important when both contents and information systems are online.

Another hot topic, or even an emerging megatrend, is openness. In this context, it often means open access to the contents. Perhaps they are published via open APIs which are built into open source digital archive systems; and the archival contents could be enriched with open data.

Both linked data and open data are examined from a pragmatic point of view. The concept of open data, linked data and open linked data are introduced. They are discussed as both technology and best practices. This paper then focuses on the context of digital archives and repositories. In addition, a compact overview on the challenges and possibilities is presented. The paper covers the data consumer and provider perspectives. It features a couple of examples for practical benefits.

Two projects carried out by the Mikkeli University of Applied Sciences are used to demonstrate the developments and the challenges: Open Source Archive and The Karelia Database.

Digital archives and repositories in the age of digital revolution

Digital revolution - a massive digitalization of organizations and the society - is happening everywhere, even in the archives and repositories. This creates contrast between the contents and the information technology. The contents are changing slowly or not at all. Still, the content is required to be preserved over long periods of time. Technology, on the other hand, is rapidly evolving and the trends come and go fast.

Digitization of analogue content is only the beginning. When sufficient amount of the contents is born digital or digitized a new world of possibilities opens before us. The archives, repositories and other memory organizations have huge assets in their hands as the keepers of data. The data is available to be used, linked and shared via web based services and interfaces. Nevertheless, at the same time it should be trustworthy and meet its original function; preserving information. However, it is easy to question preservation without access.

Today, we live in a connected world. Everyone can access the Internet via computer, tablet, smart phone and other devices. One can access more data and services easier and faster than before. Some memory organizations are adapting to the change already. For example, libraries offer electrical books and the archive catalogs and metadata are often in information systems exposed via search interfaces.

Providing online access to materials is not always a straightforward or easy task. There can be various restrictions by laws and regulations. Materials can be placed behind paywalls because of business models. Content or even some of its metadata can be sensitive or otherwise confidential. However, the benefits can overcome the challenges given enough time and support. It is both a philosophical and a practical decision. Especially, contents by organizations operating with public funding should be transparent and provide open access to justify their operations and funding. From practical perspective it will strengthen the information ecosystem and even introduce new revenue models. Open access or at least online access ensures that the content is actually usable and findable. It is a common misconception that online access would jeopardize the access management. Today, people can manage even their finances, taxation, healthcare and voting online.

Openness is a megatrend of this decade. The modern open movement began with open source in universities such as MIT in early the 1970's. However, it really gained popularity with the Internet revolution and software such as Linux and Apache Foundation projects. Open access and science are increasing in popularity as well. Briefly, open data is freely available with similar licensing and ideology as open source. Open data is more commonly published by public or governmental organizations rather than private companies. Often the content is delivered via an open API (application programming interface). Open API is an interface that is conforming to open standards such as W3C REST. The basic idea is that regardless of the openness of the content, the technology and documentation for the interface is openly available. It is especially useful for contents which are meant to be shared and compatible, like open data or linked data. The whole open movement supports the concepts of crowdsourcing, democracy and decentralization.

Challenges

During the development done in the use cases, a couple of challenges were identified. First, the access control metadata needed to be added to the archival materials. In logical sense, it was not archival metadata but management metadata. Therefore, policies needed to be defined for audit trail creation, metadata standards et cetera. The information about the users, groups and roles is required to link the access control metadata into the actual users and systems. LDAP or similar directory service can be used to provide this information. However, the issue of portability of the access management rules and definitions still remains. Access control is usually a compromise between performance and mutability speed.

Due to the relatively short but fast paced history of information technology and in that sense digital archiving lots of

contents and metadata is scattered in different formats and systems. Often the old information is kept alongside the migrated data to maintain compatibility or just in case. This has led to challenges in modeling, cleaning, manipulating and publishing preserved information.

As said by Hooland and Verbogh [3], it should be recognized that designing and building an inclusive information architecture is not feasible objective. A comprehensive model or architecture could perhaps be built for just one organization which has a coherent metadata scheme. The model would not probably fit other types of contents and it could mean losing the ability to describe the content in full detail. While designing the Capture [4] metadata model, it was decided that preserving the materials is a top priority. What is preserved can be described but what is lost due to strict regulations is worth nothing.

Another challenge from the information technology point of view is to avoid its pitfalls. Technology is never an end, only means. Right now open data and linked data are hyped and any implementation should be made carefully to avoid the loss of data, context or provenance.

Use cases at Mamk

This paper refers to two use case projects which were both carried out by the Mikkeli University of Applied Sciences (Mamk). The first project, Open Source Archive (OSA), has been featured in the previous Archiving conferences multiple times. The project started in May 2012 and closed at the end of December 2014. It was funded by European Regional Development Fund (ERDF) granted by South Savo Regional Council. The objectives in OSA project were to develop a digital platform for digital archives and repositories and to search for and test a dark archive solution for long-term preservation. OSA was primarily a development project which focused on finding well-known and widely used open source components and solutions. Furthermore, previously designed Capture data model was implemented in OSA project. Capture is a flexible data model suitable for linked objects in digital archives and repositories. The technical platform for OSA was Fedora Commons. The dark archive was built with DAITSS.

The other use case is the modernization of Karelian Database. The work was done in the Structures of the Applied Research of the eServices project which was carried out during 2011-2014. The project was about strengthening the digital services innovation, knowledge structures and connections with the regional organizations and the public sector, especially in the area of eGovernance and digital preservation. The Karelian Database is a research register and an archived database. It contains digitized and decoded demographic information about people in the ceded Karelia region on the 19th and 20th centuries. The modernization included migration from the old database to a modern relational database, building sophisticated search and indexing features and developing better access to the information.

Access management

In this paper, access management is discussed as an umbrella term to cover different aspects of managing access to contents and metadata and managing who has the privileges to perform operations on them. The approach is pragmatic and based on the use cases described in the previous chapter. The basic principle is that digital archives, repositories and such are connected to the Internet and available online. This paper does not cover the access management of dark archives.

Basically, access management is about granting authorized users the rights to use all the services and access all the contents they are entitled to, while denying these privileges from non-authorized users. These principles have been implemented well in the existing web applications and content management systems. During the development of both OSA and Karelian Database these solutions and practices were researched and applied. The technical implementation is different with the two example projects but the principles are the same: portability, modularity and loose coupling of users, privileges and the contents.

Access management in OSA is based on the Fedora Commons' data architecture. The atomistic architecture supports different metadata levels and components which are called streams in Fedora Commons. The initial challenge was to separate the archival metadata from the management metadata. It was solved by using object's isPartOf relation to identify the object itself and determine its location in the archival hierarchy. The hierarchy implementation was adopted from the previous developments done at Mamk and based on best practices in the Fedora community. Identifying an object is based on its unique persistent identifier (PID). The PID is to be replaced with an URI in the next version of the software. The PID cannot be changed once the object has been created and the isPartOf chain is seldom changed assuming the hierarchy is designed properly. Access management is based on the chained isPartOf information. The logical root of the chain is the organization itself. The rest of the chain is based on the organizations records management plan. The platform itself does not limit or dictate it. This information is stored in the object's descriptive metadata.

The isPartOf chains are connected to access rights with patterns matching them. The principle is simple and efficient. A pattern can be an exact match or match a beginning of a chain. For example, a pattern can match single collection and nothing else. This kind of pattern could be used to set access rights for just that collection. Alternatively, the same pattern could end in a wildcard; meaning the match would be that collection and all objects which are part of it. A rule consists of a pattern and an associated right. A right can permit or deny a specific access right level to the objects that it is linked to. The link is dynamic and matched during the run-time, to use computer science terminology. Finally, the level determines what kind of privileges are granted to the object. The lowest possible level is -1 which means basically no rights. Increasing the level affects the numeric value and the rights: read metadata, read contents, write metadata, add contents, manage the object and full rights. The idea of the numeric value is to make the level easy to compare programmatically. The higher level includes all the previous rights. Access right rules are not stored within the archived objects. Instead they are stored in a standard LDAP based user directory service. These rules, called roles in the user directory, can be archived if required.

More fine-grained access options can be formed with the publicity class metadata. In OSA, it can set the object as public, restricted or confidential. These can be combined with the access right levels to create more sophisticated rules. The publicity class information is stored within the object's metadata. For example, public objects can be automatically published as open data while

restricted materials could require a login and confidential materials therefore require explicit access rights.

Performance and mutability speed are were identified as challenges in the real life implementation of the access management. In OSA, it was solved by indexing the isPartOf chains and PIDs into Apache Solr based index. This way the lookup for access rights was efficient and took only milliseconds almost regardless of the amount of the contents. When performing search or some other operation it is not required to request to object from the content store, repository, archive or such to determine the level of access. Usually, these kinds of requests are not feasible with large amounts of objects e.g. because of I/O bottlenecks. The downside is the mutability speed. If there are major changes in the archive hierarchy, replicating the changes to the index and writing the objects' metadata is slow. However, updating the access rights rules does not require any changes to the objects. Likely the hierarchy is more or less permanent but the rights can vary.

User management is handled by the user directory. When the directory service contains both the roles and users, groups, organizational units etc. it is easy to map them together. LDAP is the de facto standard in user management. Depending on the technical solution it enables federated access. This way there is very loose coupling between the archive materials and the access rights.

Open Data

Open data refers to the unrefined information, accumulated by public administration, organizations, companies or private persons, which has been published online and licensed to be used freely and without any payment. This use includes also the users outside of the organization. However, public information is not always open data. Anything available via web pages, portals or services for everyone is public but its licensing and availability really determines if the data is really open or not.

The basis of open data for memory organizations should be the benefits and justification of preserving cultural history, society's memories and other valuable information. Open data should be public. Privacy cannot be compromise by opening metadata, records or contents. In case for example of private archives, the trade secrets and intellectual property needs to be protected.

Open data should be technically available in an open and machine readable format. People can easily read PDF documents and PDF is the preferred format for text based documents in repositories and archives as well. However, programs and machines cannot easily extract and parse the information. Even then, there is a risk of misinterpretation or loss of styles, context or other information. All digital archives should be able to produce data in XML or at least CSV format. The internal structure should be explained and documented. Proper metadata prevents misuse and misinterpretation.

The data published as open data should be available online free of charge. It lowers the barrier to explore the data and use it without bureaucracy involved in budgeting. Finally, the licensing should be permissive and allow all kinds of re-use. The license should be stated clearly and available in the same location as the data. Some archives still restrict the access to data even if it is published online. For example, it can be viewed only in the archive's premises or reading room.

For archives and memory organization opening data could mean more visits, clicks, page views, downloads or increase in other kinds of usage metrics which is sometimes linked to funding. Sharing e.g. cultural history is beneficial and should be public. Research and education would benefit from open resources which are relevant and trustworthy. It adds transparency in society and organizations which enforces democracy and trust. All in all, data is resource that does not lose value when it is shared, on the contrary.

In addition to publishing open data, the archives can utilize the existing open data. It can be used to enrich the metadata or create new value when combined with the archived materials. Examples include layering historical maps with Google Maps or curating content for visitors based on news and current events.

The Karelian Database approach

The Karelian Database is about information on people. It is mostly public but some of the personal information is very sensitive and not allowed for public access. The data includes records about diseases, crimes and deaths. It can be very colorful since priests in the 18th century did not have the same standards writing them as we have today. The European Union also has legislation for protecting personal data. The main concern is how different countries or individuals are handling the data. There are no common rules between countries. In this case, the Finnish legislation is very strict concerning personal data access and the web service needed to be built in a way that the data will be secured. Some of the data is about people who still live and this is also a reason why the data cannot be completely public yet. As the time goes on and the data becomes open for the public use, we have to think about the data ethics in a new way. People who work with the data have to have some respect for it and understand what it is about. Even though the data is old and there might be some interesting information there should be standards of some kind for using it. At least at the ethical point of view.

The reasons for opening the data for limited use at least are to enable research, analytics and data mining. New correlations could be found and for example more information about migrations, famine or diseases. Also, it provides information on how people lived during that time. Even if it isn't actually open data some records could be opened if the sensitive information would be cleared. Of course, there is a possibility for reverse engineer or combine the data to find out the sensitive information. Even worse scenario is to conclude fallacious information.

The Karelian Database case also has a historical and emotional point of view. The data is about a ceded region which was claimed by Russia because of the wars around the Second World War. It is still recent history and there are people alive from that time. This kind of archived history tends to have emotional impact. It has to be taken into consideration and cannot be solved with technology only.

Linked Data

Linked Data means connecting objects with meaningful and machine readable links together. This way, objects can gain contextual and derived metadata. Linked data is not a single technology or a standard but more like a collection of the best

practices and tools to create and publish structured and connectable data via the Internet using standard communication protocols such as HTTP. One technology to describe linked data objects is RDF triples. Basically, it is a very natural statement with the syntax of subject, predicate and object. RDF adds to that each part of the statement has an URI and links it to the network of entities.

Linked data started as the standardization of terms, metadata schemes and controlled vocabularies in the 1970s and 1980s when the use of databases in digital archives was increasing. Linked data development truly got speed after the rise of the Internet. Ever since, the trend has been the same. The budgets are decreasing and the amount of metadata and contents is increasing. Now the use of technology is mandatory to automate processes, decrease the costs and minimize the duplication of data. These are also the reasons why linked data was experimented in Open Source Archive project.

Linked data puts heavy emphasis on the quality of metadata. For example, if an ontology is used in describing an archival object, it needs to correct and trustworthy. The data source must be carefully selected and evaluated. However, when content is ingested in a digital archive or stored in a repository, it can automatically gain metadata based on the known links. There can be metadata about the links which adds a new layer of metadata. This is something that is already used in web applications such as recommendation engines and social media. Another issue with linked data and especially ingest time metadata enriching is the provenance. The origin of the metadata must be known and trusted. When new layers of metadata are added the complexity increases really fast.

Experiences in Open Source Archive

The OSA project explored the utilization of some low hanging fruits in linked data. The approach was very pragmatic and the objective was to encourage the future developments. Open linked data was used to add descriptive power to the platform. Finnish thesaurus and ontology service Finto was linked with a couple of metadata fields. When an object is ingested or described in the archive, it could be described with a term based on the ontology such as YSO (General Finnish ontology) and SAPO (Finnish Spatio-Temporal ontology). The term would be indexed to provide search capabilities and local translations but the URI of the linked entity would be stored in metadata.

Private linked data was used internally. The metadata model and the whole object model, which built on top of Fedora Commons' architecture, was based on linked data concepts. It was first designed in the previous Capture project by Mamk and the Central Archives for Finnish Business Records. The only shortcoming was the use of Fedora Commons' PIDs instead of full URIs. However, that is acknowledged in the development roadmap of the OSA platform.

When an object is ingested in OSA, it can automatically inherit metadata from its ancestors based on the isPartOf and the records management plan. Also, the basic principle in Capture model is that the archival material is the center piece and all additional descriptive and contextual entities are linked to it. These entities include activity and functions, places, events and agents. Each of them is described as an individual entity and can then be linked with any other entity with multiple kinds of relations. The end result is a flexible model which greatly decreases metadata duplication and speeds up the description process. It also allows new kind of discovery based on the metadata, links and the distance of various entities.

Conclusions

Now that the both OSA and Karelian Database projects are completed is time to look forward. At the edge of digital revolution, the world is becoming more data-centric. Archives, repositories and memory organizations hold valuable data. Some data would benefit from being opened and shared, some would benefit from being added value by enriching linked data and open data.

Regarding digital archiving, new opportunities are emerging. My data is personal data created by and about a person. It can be for example health records, official documents or personal memories such as photos, videos and writings. My data is valuable as a personal legacy and sometimes it records the picture of the period or public figures. In addition, preserving personal data, metadata and its context could become a feasible business model. It is something archives could achieve, if they could productize their services and develop them more user friendly.

Mamk is also investing in future of open data, linked data and data management in memory organizations. Mamk is about to establish a Research Center for Digital Information. The center is planned to be a permanent research institution and has two connected lines of action: information management and utilization; and digital long-term preservation.

References

[1] J. Räisä & M. Loponen, The modernization, migration and archiving of a research register. (2014).

[2] J. Lotarski & J. Sueters, Data protection in the archives world - fundamental right or additional burden? (2014).

[3] S. van Hooland & R. Verbogh, Linked Data for Libraries, Archives and Museums. (2014).

[4] M. Lampi & O. Alm. Flexible data model for linked objects in digital archives. (2014).

[5] B. Haslhofer, A. Isaac, The Europeana Linked Open Data Pilot. (2011).

[6] Europeana Linked Open Data, Europeana Labs. http://labs.europeana.eu/api/linked-open-data/introduction/.

[7] Helsinki Region Infosare. http://www.hri.fi/en/open-data/.

[8] C. Bizer, T. Heath & T. Berners-Lee, Linked Data - The Story So Far.

Author Biography

Mikko Lampi is the research manager in Information Management and Digital Archives at Mikkeli University of Applied Sciences. He has a B.Eng. in information technology. Mikko is interested in digital information management, open movement and developing information society.

Johanna Räisä has a Bachelor of Business Administration in information technology. She has knowledge of developing applications and social media. She is currently interested in mobile, open source and interface development. Johanna worked with the Karelian Database at Mikkeli University of Applied Sciences.