Flexible data model for linked objects in digital archives

Mikko Lampi, Mikkeli University of Applied Sciences, Mikkeli, Finland
Olli Alm, Central Archives for Finnish Business Records, Mikkeli, Finland

Abstract

In this paper is presented a pragmatic data model for creating and managing linked contextual and material objects, the design process and the reasons why it was developed. The model was created to meet the needs of private archives and digital archive services. The design goals were flexibility, cost-efficiency and suitability for daily use. The basis of the model is to use contextual objects to describe the archive objects; should they be text documents, moving image, audio or any other type of content to be preserved. By linking the contextual objects such as events, places, agents and actions, the redundancy of archive data can be minimized. The objects gain more informational value because the linking forms a network of connected objects. The data model was based on the national and international metadata standards as well as best practices from private archives. It was implemented using Fedora Commons' object modeling features as a part of the ongoing Open Source Archive project.

Background

Preservation and managing archives, both the analogue and the digital, is primarily practical work such as ingesting, describing and organizing the preserved content. Standards, organized methods and software systems are created to help the process and keep the archived data accessible as required. Development of these tools and standards is essential in modern archives.

Central Archives for Finnish Business Records (ELKA) is operating in the private sector, which has generally been more streamlined environment than the governmental and cultural field. ELKA's approach has been more pragmatic than scholarly or scientific. This is also because ELKA has been able to define and shape its own methods. Mikkeli University of Applied Sciences (MAMK) Department of Electrical Engineering and Information Technology provides digital preservation services. In MAMK, a similar operational model was found suitable. However, the selected approach requires data models and tools supporting and guiding the preservation processes. It has to be kept in mind that the models and processes should comply with national and international standards even though there are some divergent requirements. ELKA and MAMK have a history of co-operation in digital archives, so it was practical to have a joint effort project to solve the common problems.

Linked data is nothing revolutionary and has existed for some time already. The semantic web is based on its principles and has already been proven to be an effective method in adding value to isolated data. Projects like Europeana Open Data Pilot have inspired this work [8]. There was a need to build a linked data model that would support both ELKA's and MAMK's digital archive services. From these starting points two projects were initiated: Capture and Open Source Archive.

Capture and OSA project continuum

The Capture project was carried out by ELKA in co-operation with MAMK during 2011-2012 and funded by the European Regional Development Fund (ERDF). Its purpose was to create a specification documentation of a modern archiving system. The system was designed to include the functionalities for managing analogue material reference data, digitized materials and born-digital materials. The project included a pilot phase for testing new methods to catalogue and store the analogue and the digital materials obtained, for example, from the Finnish Broadcasting Company YLE. Due to the limitations of the ERDF funding, it was not possible to include an actual implementation of the system in the project.

Capture implementation is being done as part of the ongoing Open Source Archive (OSA) project which is administrated and executed by MAMK department of Electrical Engineering and Information Technology. The project started in the summer of 2012 and will continue until the end of 2014 and is funded by ERDF. The goals of the project are to develop a service archive system which would meet also the Capture specifications and to search for and test a dark archive solution for long-term preservation. The focus and key values are open source, sustainable solutions and user centered development. After the project is completed, the results will be migrated with MAMK's digital archive services and their current archive platforms.

Agile digital preservation

Both ELKA and MAMK are looking for agile methods in their digital archive services. The basis is to provide commercial services based on the needs of the private enterprises, city archives and such. While the customer base of ELKA and MAMK varies, the requirements are close to each other which motivates to find and develop more flexible ways of modeling data. Flexible in this context means ways to make ingesting and describing data easier, faster and more automatic. In case of a manual ingest with web forms or client software it could mean better user interfaces with auto-completion with metadata fields, the lazy creation of contextual objects or inheriting metadata from the archive hierarchy. Lazy describing allows creating contextual objects just in time while describing an object, if it should require additional information by linking it to another object that does not exist in the system yet. Automation can be achieved by linked ontologies or machine readable interfaces to document management systems or other archives.

Another method is to allow organizations to define their own metadata models suitable for their data. Private enterprises are not always interested in archiving their content with fixed and predetermined standards. The problem was solved by allowing them to choose which fields they use and then link them to a larger

©2014 Society for Imaging Science and Technology

compound metadata model, which could be crosswalk into compatible standards. When operating in the private sector, the archive services cannot dictate how the customers want to use the archive, but have still to offer conversion tools, guidance and best practices.

Capture model

The data model developed in the Capture project is based on the insight that traditional one-dimensional archival description does not function effectively today when organizations are continuously changing, merging, diverging and entering into joint processes with other organizations. It was found that, a model of contextual description is more flexible and quicker from the point of view of cataloguing if the material itself is made into an entity of its own and the metadata closely related to the description, such as activity, agents, places and events, are made entities of their own. In Capture model, these four entities are called collectively the 'contextual entities' as opposed to the materials entity.

Traditional archival description also tries to represent contextual information relating to the creation and use of the documents, but that is done in a one-dimensional multilayer fashion. The Capture model takes more determined stance to separate the description of archival materials from the description of contextual entities. Figure 1 presents a simple case of the Capture model.

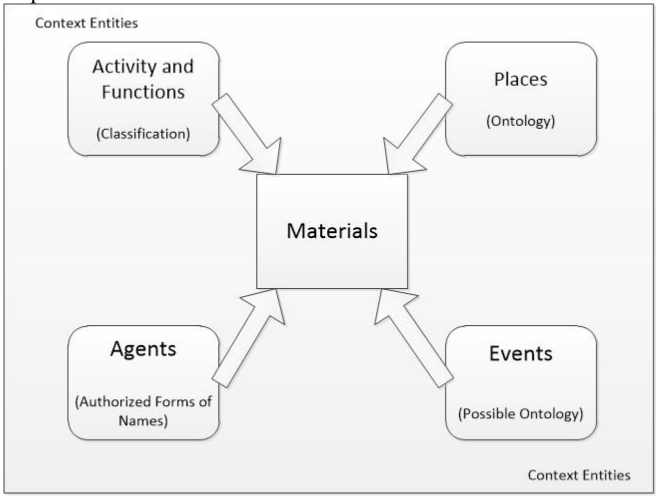

***Figure 1**. Context entities linked into a material entity in Capture model.*

Flexibility and efficiency are the result of the fact that the contextual objects need to be described only once, not separately for each unit as in the traditional method. As an example, consider a document which is a travel account of the 1932 Olympic Games in Los Angeles: the materials entity describes only the data directly associated with the document. We can then link contextual objects to it, such as the place (Los Angeles), event (the 1932 Summer Olympics in Los Angeles), activity (e.g. recreational) and agent (the author of the travel account). Ideally, the contextual entities are obtained directly from a classification or ontology lookup service, reducing the need for manual cataloguing even further.

The metadata model is designed so that the system is able to receive materials in several different metadata formats. It supports at least the MoReq2010, Dublin Core and several national formats, such as SFS-ISO 5914, JHS 143 and SÄHKE 2. In addition to these metadata elements related to records management and archiving, the metadata model includes many technical metadata elements as well as those needed in the cataloguing of special materials, such audio and video recordings. In figure 2, an example of the object network is shown.

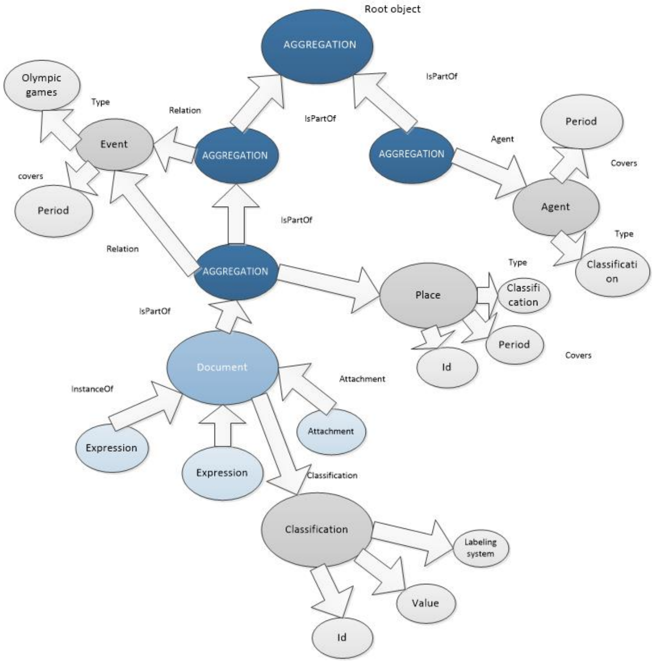

***Figure 2.** Example of object network modeled with the Capture model.*

Implementing the Capture model

Variety of technologies and platforms were evaluated for the implementation as part of the OSA project. In addition to the Capture data model, the digital archive system should meet with the other requirements defined in the Capture project and superset the features in MAMK's YKSA archive service. The research and analysis on the technologies was introduced in last Archiving conference in 2013 at Washington D.C [16].

As a core framework Fedora Commons and its technology stack was selected. It was supplemented with other database technologies such as Apache Solr, MariadB and MongoDB; each matching a specific need and type of data.

Development challenges

Now that the design and specifications for the data model were done in the Capture project, it needed to be implemented in a machine readable form. Fedora Commons provides a content model architecture (CMA) and a compound digital object model. The CMA was not utilized in full because of the extra complexity it would introduce. The object model and content model approaches were used but the service definitions and deployments were removed. The Fedora Commons repository was designed to be the backend repository and the master data store. Our Fedora Commons implementation is discussed in more detail later. For linked data, the compound model and CMA provided a good starting point. The most time consuming task was to search for documentation and use cases on how other projects had solved the same challenges we had.

Providing fast natural language searching and indexing was a challenge because it meant a compromise in between flattening the data and the performance. Apache Solr was the preferred searching and indexing solution but its data model is a flat record based index. The solution was to include required information about objects' location in the archive hierarchy, known as isPartOf relation, and perform the required combining of data and building the compound objects back together in the search application. In the OSA platform, the web client is a front-end for the loosely coupled modular backend platforms.

Another challenge was to provide the mapping and crosswalk from various ingest processes into the Capture model. The best solution was to decouple the data presentation and the mappings from the actual data model. Fedora content models hold the information about the Capture model and its schemas. The workflows, user interfaces and APIs hold the knowledge about how to map data and how to present it. Related to the issues caused by diverse data and processes, was also the issue that different organizations had very variable amount of metadata. The cardinality had to be set on the organizational level and the data model level. The organizations have to be able to define what metadata is important regarding their data and processes. However, at the same time the archive software must keep the data usable and compatible with the standards. Generally, the Dublin Core is considered a minimum requirement.

Due to nature of the preservation services, the data can be ingested once and then preserved, or it can be ingested and continuously enriched. The implementation had also to support that the network of linked objects could change and expand. Meaning, that the relations cannot be flattened and the objects cannot contain hard-coded values from other objects. It is more a system design challenge. For example, if a records creator or a root collection is updated or relocated the child objects, such as groups, collections and records, should not be required to be updated. They should instead be linked by URIs or similar unique identifiers.

Content modeling and data stores

Fedora Commons was chosen to be the core repository and master data storage. However, its default CMA had extra complexity and to reduce it, the existing content models and other Fedora based system were examined. Islandora provided a well-documented and tested base for our work. However, we wanted to even more simplify the model. The data presentation related information, such as user forms and client side rules were moved into a client configuration in the accessing application instead of the repository itself.

The Fedora compound object model helped us to preserve the digital objects without reducing the relationships or compromising the idea of the Capture model. Both the material objects and contextual objects were ingested as independent and complete entities in the repository. The Capture metadata model was added as a mandatory data stream and the Dublin Core record was automatically mapped from it.

Relations to other objects, either in the repository or in linked ontologies, were stored in the RELS-EXT stream, which Fedora automatically stores in an embedded RDF database. With the current version of Fedora Commons, Mulgara is shipped by default. During the OSA project, we will consider other RDF stores when migrating to Fedora Futures during the summer and early fall. The triple model and SPARQL provides a natural way to access and store the relations of linked data. With the relational database based solutions used in earlier versions of archive software, the relations were always a compromise between the normalization and the natural presentation. Some relational databases have since provided features to overcome the limitations.

Considering the challenge with variable data sources, we decided to test how schema-less solutions would fit storing the metadata records. Functions such as reporting and accessing large amounts of data would require something else than requesting the objects from Fedora's disk based storage. MongoDB is a document database suitable for web applications. It stores data as JSON or binary JSON which is a near native format for metadata of any kind. Mongo matches a different use case than other data stores used in the project. Primarily, it is used to access data via web interfaces, such as RESTful web services, and to temporarily store ingested records before the actual ingest, which allows the enrichment and curation of the data before preserving it. A common scenario is curating a network disk or an USB disk. However, the need for another data store will be critically considered before making it a part of the final product.

Finally, Solr required schemas for indexing the metadata and contents of rich text documents. It is possible to use Solr without a pre-configured schema but it requires using dynamic declarations. Because the project is still in development, we are not sure about the impact on performance. Also, it would require additional services to describe the metadata model for any client software. So, the Capture metadata model was defined as a Solr schema and best practices were adapted from existing projects such as Finna, which is a search interface for Finnish archives, libraries and museums. Any customer specific metadata and other dynamic content can be added because of the flexible Fedora models and dynamic Solr schema additions.

Interoperability and extensibility

From the beginning, it was decided that the data model and its implementation should be interoperable and compatible with other metadata standards and software systems such as document management systems and digital archives. MoReq2010 was used as a general guideline. The approach used in YKSA service was further developed in OSA. The system should provide a mapping service for its metadata models. The schema information with the metadata elements and descriptions is preserved as part of the content models in order to make sure the data is understandable and machine-readable even if the services or the system should become obsolete. Fedora compound object model allows extending the models with any standard required. Of course, it requires adding mapping information and interfaces to manage the data but it is taken into consideration in designing the implementation.

A RESTful API will be built to provide access to the data and to expose an ingest interface. The API communicates with a workflow engine which can be extended with micro-services. The project will develop the core micro-services for transforming and managing the data collected with the pilot use cases. The transformations can be done using a standard XSLT or any other tool via the modular workflow system. The workflow engine is released as a standalone open source product by the end of the OSA project.

All the data that is ingested is validated against the Capture metadata model which is a superset of Fedora's requirements and some other metadata standards such as Dublin Core. It is, however, configurable if there should be some other master model than the Capture.

Use case: archive as a service

ELKA and MAMK both provide archive as a service. These services are not direct competitors but support each other. MAMK is more of a software and digital storage provider and ELKA operates as an archive. Because ELKA's core functions do not include the design, development and maintenance of software systems, ELKA purchases the services from an external operator, such as MAMK.

Waiting for the launch of the archive software from the OSA project, ELKA has been using other archive software provided by MAMK: ElkaD and starting by 2014 the YKSA archive service. ELKA's customer base consists of both the companies, who need the preservation services, and the researchers who wish to access the archive materials. Until now, the preservation services ELKA has offered to companies have included storage, sorting, cataloguing and digitization of only the analogue materials. In the future, these services will be extended to digital long-term preservation, which will be initially implemented using YKSA and later OSA.

Piloting Capture and digital archive services

As part of the Capture project, there were pilot projects on digital archiving. Based on the pilots and the customer surveys carried out among commercial companies, ELKA has researched which kind of digital preservation services their customers require. The services can be divided into three cases:

1. The company has photographs or some other historically valuable material which they wish to digitise. In this case, the service comprises the digitisation of the analogue materials, the permanent preservation of the analogue materials, the long-term preservation of the files generated in the digitisation and a data system service through which the company will be able to use the digitised materials.
2. The company has a disused data system, a lot of material on a network drive or a large amount of other electronic material which they wish to take charge of and subject to long-term preservation measures. In this case, the service comprises the sifting, sorting and mass storage of the file material. In such cases, there are hundreds of thousands of files.
3. The company actively uses a data system, such as a document management system or Microsoft SharePoint, from which it wishes to move materials to be permanently preserved to a separate archiving system. In this case, the service comprises an open interface through which the material can be transferred to an archive service system administered by ELKA.

In the Capture project, a client application was developed to easily discover and manage files and collect and add metadata before ingesting them into a digital archive. The client application proved to be effective tool with regard to services one and two. There are plans to integrate it into the OSA software as well.

Archive software as a service

Information technology services have been offered as a service for more than fifteen years now. It all started from emailing and reached new businesses and industries fast. The usual drivers have included the cost and the required resources for running the services. By purchasing the services from dedicated service providers, the companies could reallocate the resources to their core businesses.

For archives, the software as a service offers savings in cost and more reliable data storage since the data can be preserved in consolidated and secure server rooms. Still, it requires that the service providers understand the special requirements of the archives and long-term data. With understanding the co-operation with software companies, or in this case with the MAMK Department of Electrical Engineering and Information Technology, can be very effective and provide mutual benefits. Because MAMK has a long history in digital preservation research, it can put lots of effort in IT services and storage functions and ELKA can concentrate on the content services and the customers.

Results

The Capture model is pilot tested during spring and summer 2014. It is completed during the OSA project and as a result it will offer a flexible way to use a digital archive for various kinds of content which will form a semantic network of linked objects. The data model is a core part of the OSA software. The contextual descriptions provide faceting for exploring the archive contents for researchers and the public audience. In future, it can even be used for linking the objects in other systems, like in Europeana pilots.

Once the OSA project is completed more in-depth results are published. The coming licensing for the data models and the software is still being planned but as much as possible will be published in open source. The Capture project results are available in ELKA's website in Finnish and the summary in English.

Conclusion

A proper and well-fitting data model is the core of the daily preservation work. It provides savings in both the cost and time. By being more flexible, the ingesting process can cover more material in less time and allow working as efficiently as possible. Imperfect material can be preserved and completed later. It is a much more likely scenario than losing the data because of too strict and limiting methods and standards. Like the name of ELKA's original project describes well, first aim of the digital archiving must be the capturing of the documents. If that goal is not achieved, then nothing else will not matter much either.

The data model should enable describing objects as they exist or existed in the real world: linked to each other. The model of contextual description gives a much better opportunity to take advantage of the existing ontology services and to open up archival materials and metadata to users as linked open data.

References

[1] Olli Alm, Janne Strömberg, Capture project final report (ELKA, Mikkeli, 2013).

[2] Olli Alm, Janne Strömberg, Summary of Final Report for Capture Project (ELKA, Mikkeli, 2013).

[3] Olli Alm, Capture-hankkeessa kehitetään palveluja (Faili 1/2011).

[4] Olli Alm, Elka uudistuu Capturen kautta (Faili 4/2012).

[5] Capture client questionnaire 2012, results published in Capture project final report (ELKA, Mikkeli, 2013).

[6] Ari Häyrinen, Open Sourcing Digital Heritage - Digital Surrogates, Museums and Knowledge Management in the Age of Open Networks (University of Jyväskylä, Jyväskylä, 2012).

[7] Osmo Palonen, ELKA and MUAS - Partners in digital archiving, Research Publication 2013 (MAMK, Mikkeli, 2013).

[8] Bernhard Haslhofer, Antoine Isaac, The Europeana Linked Open Data Pilot. (2011).

[9] DLM Forum Foundation, MoReq2010: Modular Requirements for Records Systems - Volume 1: Core Services & Plug-in Modules (2011).

[10] Mika Nyman, Kohti kolmannen sukupolven kulttuuriperinnön tietojärjestelmiä, Muistilla on kolme ulottuvuutta: Kulttuuriperinnön digitaalinen tuottaminen ja tallentaminen (MAMK, Mikkeli, 2012).

[11] Nicholas David, The Google Generation: challenges and changes for libraries, archives and museums (2010).

[12] Marta Nogueira, Archives in Web 2.0: New Opportunities (2010).

[13] Joy Palmer, Archives 2.0: If We Build It, Will They Come? (2009).

[14] Jaana Kilkki, Arkistot 2.0 mullistaa arkistojen tietopalvelua (2010).

[15] Outi Hupaniittu, Tutkijoiden ääni ja sähköiset aineistot, Selvitys muistiorganisaatioiden asiakkaitten digitoitujen aineistojen tarpeista ja saatavuudesta (2012).

[16] Mikko Lampi, Osmo Palonen, Open Source for Policy, Costs, and Sustainability (The Society for Imaging Science and Technology, Washington DC, 2013).

Author Biography

Mikko Lampi is the project lead for Open Source Archive at Mikkeli University of Applied Sciences. He has a BEng in information technology. Mikko is interested in open source, agile development and involving the community and users with the software development and digital archives.

Olli Alm is the Information Services and Development Manager at The Central Archives for Finnish Business Records. He is the head of Information Services and Collection Management sections. He was also the Project Manager in the Capture-project in 2011 – 2012. Olli has Master of Arts in History at University of Joensuu.