Search, Discovery and Harmonization of Diverse Digital Contents

Mikko Lampi
Mikkeli University of Applied Sciences
Patteristonkatu 3 D
50100 Mikkeli
+358504364161
mikko.lampi@mamk.fi

Aki Lassila
Disec Ltd.
Sammonkatu 12
50101 Mikkeli
+358400869955
aki.lassila@disec.fi

Timo Honkela
University of Helsinki, Department of Modern Languages
National Library of Finland, Centre for Digitization and Preservation
Nykykielten laitos, Kieliteknologia
PL 24 00014 HELSINGIN YLIOPISTO
+358504480953
timo.honkela@helsinki.fi

ABSTRACT
This paper provides an overview of search, discovery and harmonization of diverse digital contents. Each concept is described in detail and illustrated with use cases and examples. The requirements and drivers are studied in order to make the harmonization possible. The process and technologies for harmonization are also discussed. Information extraction and indexing are presented as the foundation for these concepts. Emphasis is also put on search and access with strong use case examples. Analysis and advanced discovery are reviewed from a scientific point of view in contrast to some use cases. Three use cases are used throughout the paper: Open Source Archive, Capture and Finna projects. The paper represents views from each project in combination of pragmatic experience and developments and scholarly approach.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: H.3.1 Content analysis and Indexing, H.3.3 Information Search and Retrieval, H.3.4 Systems and Software, H.3.5 Online Information Services, H.3.7 Digital Libraries

General Terms
Management, Performance, Experimentation, Standardization, Languages, Theory.

Keywords
Information, Metadata, Digital Content, Extraction, Indexing, Harmonization, Search, Discovery, Access, Information Analysis, Digital Archive, Digital Repository

1. BACKGROUND
Information governance in distributed environment is challenging due to the complexity and diversity of contents, data sources and standards (W3C 2009). However, harmonization enables discovery, access and analysis for the information in digital information systems. Therefore, the availability of harmonized data promotes the discovery of useful information and relations within the data that otherwise might remain undetected. In addition, knowledge discovery can also take place based on unstructured data. Unsupervised data and text mining techniques can be used to automatically find key phrases and taxonomies that can be used in metadata construction and harmonization. Such techniques provide support for interoperability in cases when a full harmonization of conceptual content is difficult or impossible due to theoretical or practical reasons. The theory of conceptual spaces (Gärdenfors 2014) provides a rigorous framework for the description of data in cases where logical formalization may fall short. Furthermore, harmonization can also be achieved by utilizing ontologies, vocabularies and other linked data, transformations and mappings.

As is generally known, information can be indexed and further analyzed. Language and entity identification supports natural language processing and understanding about language specific properties. Finnish language processing is used as a use case in this paper. Indexing provides fast access and additional ways, such as faceting and geospatial analysis to discover, access and visualize the information. Harmonized metadata can be linked and exposed as public or private linked data. The usage of native linked data technologies enables the efficient exploitation of information and open data (Bizer et al. 2009).

Searching has become more than just using a web search engine like Google or Bing. Searching is now associated with discovery platforms with full-text and natural language search possibilities which also include features such as visualizations, facets and mashups. In addition, usability and user experience are very important factors in search and access. The platforms should support complete machine-readability and data interoperability. The trustworthiness of data sources is another important aspect.

As our use cases in order to demonstrate the concepts and technologies, this paper provides three projects: Open Source Archive, Capture and Finna. Open Source Archive (OSA) is a project executed by Mikkeli University of Applied Sciences (MAMK) and funded by European Region Development Fund. The project started in June 2012 and ends in December 2014. The primary objective of OSA project is to find and develop open source tools and solutions for digital archives. Its key features include archival materials and lifecycle management, long-term preservation, ingest, search and access. OSA software is based on well-known open source software. Later in this paper, OSA is used to refer to the digital archive software unless stated otherwise. The OSA project is based on Capture, which was a data modeling and digital archive definition project by Central Archives for Finnish Business Records (ELKA) and MAMK. It was executed during 2011 - 2012. The primary deliverables from

Capture were a concept of a harmonized metadata model and the specifications for a modern and flexible digital archive system.

The third case, Finna (www.finna.fi) was started in 2012 and it is part of the Finnish National Digital Library program which aims to "ensure that electronic materials of Finnish culture and science are managed with a high standard, are easily accessible and securely preserved well into the future." (National Digital Library, http://kdk.fi/en) Briefly, Finna is an online discovery service for all Finnish materials by libraries, archives, or museums. The items can be books, drawings, old advertisement brochures, scientific articles, etc. Finna's long-term objective is to provide information from each and every Finnish memory organizations' content in a meaningful way. Finna relies heavily on indexing, the harmonization of metadata and other issues discussed further in this paper.

The paper is organized as follows. The second section is about extracting and indexing information. Section three discusses metadata harmonization and some practical examples of it. In section four, search and access are reviewed via the use cases. Section five reviews discovery and analysis. This paper is concluded with discussion about the results and future research and development suggestions.

2. EXTRACTING AND INDEXING DIGITAL CONTENTS

The first step in harmonizing digital contents is to extract the metadata and the file content in machine-readable form. Extraction requires that each file format has a compatible parser. There are easily tens of formats for rich text documents, audio, moving image, pictures and other available and valuable digital contents. Each format requires a parser library which can extract its technical metadata, embedded descriptive metadata and the actual content. For archival usage, one must know the significant information for the specific format in order to correctly preserve it. Different tools provide different technical outputs which need to be mapped and processed before forwarding the information for harmonization and indexing. After the initial extraction data is in usable form but by no mean harmonized or normalized.

A widely used extraction solution is Apache Tika. It can be used to extract information from documents and detect the language automatically. Tika will identify the file and automatically select a suitable parser if known. Automation can be achieved by integrating Tika or other data extraction solutions with indexing engines. Tika will be implemented in the OSA project and is widely used in other archival software developed in MAMK.

Indexing is necessary for efficient access to huge amount of textual data, such extracted contents of rich text documents. Usually index itself is a binary format data store. It does not replace or make obsolete the original data but supports its usage. Indexing is required to enable feasible and efficient processing of time consuming tasks such as full-text search and certain analysis processes. Analysis and data mining are described in more detail later in this paper.

The basic principle in indexing is similar across different technical solutions. Databases and other data stores can be indexed for faster read operations and information retrieval. Write operations become slightly slower but the performance gain is multiple. This is because writes are usually done less often while reads are more or less continuous. Search engines use indexes to rapidly find relevant information based on the search terms and then return objects from the data store based. While most of the operations could be completed without indexes they would often be very inefficient. The performance difference is even more drastic if the data is read from a file system or disks instead from memory.

Furthermore, full-text indexing is very useful for unstructured digital contents. It enables full-text search which users are used to when using search engines like Google. Other benefits for full-text indexing include statistical information based on the indexed terms and their respective hit rates. Full-text search is discussed in more detail in section four in this paper.

One of the most used indexing solutions is Apache Solr, which was also used in OSA and Finna. In addition to indexing, Solr provides search features and tools for simple analysis. It can be extended with various plugins such as information extraction with Apache Tika.

Language processing is critical part of full-text indexing. It provides the accurate and valid identification of terms and entities. Some languages, such as Finnish, have inflected forms and thus require the basic forms of words to be determined. This can be very problematic without vocabularies. There are also other entities, such as proper nouns, which need to be detected and indexed correctly. In some cases specific entities need to be removed to protect privacy or confidentiality. In the OSA project, Apache Solr was used in combination with Voikko library for accurate Finnish language indexing and queries. Due to the nature of the Finnish language, Voikko includes also extensive vocabulary in addition to the grammatical rules. Voikko is an open source project and used in projects like LibreOffice. Integration of Voikko and Solr was developed as open source by The National Library of Finland as part of the Finna project (http://www.kdk.fi/en/public-interface/software-development).

In addition, indexed terms can be linked to ontologies or vocabularies for formal definitions and interoperability. For example, indexed Finnish place names could be linked with the national spatio-temporal ontology SAPO. The information would be more usable in a geospatial information system than unnormalized terms.

3. METADATA HARMONIZATION

Metadata harmonization is a process consisting of multiple steps, both technical and non-technical. The main drivers for harmonization are interoperability and feasibility (e.g. Nilsson 2010). While lots of entities described are subjective to humanistic sciences, for instance, rather than technical, the information systems require structured and machine-readable data. The results are new or better services for consumers and better understanding about the materials.

Different fields and industries have specified their own metadata standards to support their contents and activities. For example, MARC21 (http://www.loc.gov/marc/) is widely used in libraries and LIDO (http://lido-schema.org) in museums. Most of the standards have in common that they support well the specific metadata and objects but are not intended for managing information systems management or information exchange. LIDO, for example, covers all kinds of museum objects such as art, architecture, cultural history, the history of technology, and natural history. LIDO enables creation of normalized records for museum context. These records can be further enhanced by

providing ontology linking. Semantic records can then be shared with other systems and environments.

Metadata interoperability is one of the primary drivers for metadata harmonization. Interoperability requires that metadata records are machine-readable and compatible with each other. Dublin Core Metadata Initiative defines metadata interoperability as the ability of two or more agents, such as information systems and software components, to exchange metadata so that the interpretation remains consistent with the original context and information (Nilsson 2011).

Interoperability means that normalized records conform to metadata models which can then be mapped to ontologies, vocabularies and other metadata models, which can be internal models or metadata standards. Both Finna and OSA have adopted mapping as the primary method for harmonizing metadata (see Finna 2014). The basic principle is to map various input formats into an internal umbrella model. Finna creates a machine-readable index for materials originating from Finnish archives, libraries and museums. OSA harmonizes data first into a master data model which is then used to generate the index. Finna and OSA serve different purposes but the reasons for harmonization are the same. They both need to ingest diverse contents and provide access and management in a coherent manner. It is not feasible to implement different user interfaces, application logic and user experience for each kind of data.

During Capture project additional drivers for interoperability were identified. Firstly, it was confirmed that there is a need for digital archive and repository services, preferably hosted as SaaS. It was built as part of OSA project. This approach had lots of different files, metadata, standards and formats put into one system which all the tenants share. It has a single core repository, Fedora Commons, which manages the content and the metadata. Fedora can manage all the files and metadata formats as separate streams but it can end up in the complexity creep and hard to manage environment. It is more efficient to harmonize as much as possible. (Lampi & Alm 2014).

An umbrella metadata model, known as Capture model, was designed to tackle the harmonization challenges. The Capture model was designed to be compatible with several national and international metadata models such as Dublin Core, SFS 5914, JHS 143 and SÄHKE2 (Alm 2013). It can be extended to support other standards and custom metadata definitions as needed. Because of the extent of the unified model, a smaller piece of it can be defined as a content model for various content types. Each content type is fully compatible with the main model. Metadata values can be links to ontologies and vocabularies. Content described with Capture model form a linked data network which can be private, public or a hybrid. (Lampi & Alm 2014).

Furthermore, an important lesson learned is that a harmonized model cannot dictate too many restrictions. The umbrella model needs to support all kinds of needs and provide a coherent internal harmonization framework. Restrictions like cardinality and locale based settings need to be applied in interfaces pulling and pushing the data. In OSA, mappings and transformations are integral part of the architecture. Because OSA is a multi-tenant environment each organization has its own set of mappings which binds the data to user interfaces and APIs. Each mapping is also archived so that the original meaning and knowledge on how to read it are preserved. The mappings can be executed technically with any suitable transformation method such as XSLT. This way harmonization is a lossless and two-way process.

The harmonization process should be automatic, which means the data models and interfaces have to be machine-readable. It is achieved by providing sufficient technical information for processing the data models, metadata and contents. The data itself has to be structured or otherwise machine-readable. Finally there needs to be APIs for data operations. The APIs can be public or private and a public API can be used to deliver non-public content. More about open data and open API concepts can be read elsewhere.

Finally, harmonization is not all about technology. A very important factor is communication between all involved parties. Understanding the context and meaning of the materials is essential in preserving it unaltered during the process. For example, in Finna project the harmonization work was from the beginning a mutual task with users and involved organizations. The understanding about the needs and usage of metadata have grown during the project and it is a continuous process.

4. SEARCH AND ACCESS

Search and access in this context is more than a textbox-based search engine like Google or Bing. It is a combination of a discovery portal, browsing catalog, recommendation and curation engine and technical platform. Search is a method of finding interesting records and objects from possibly huge data sets but selected sources.

Traditional search engines provide some information based on user entered search terms from various and unknown sources in some format depending on the source with very limited metadata. The algorithms and indexes are good but all else is just counting on luck. With digital archives, repositories and other kinds of collections, one cannot afford Google like results: if it is not on the first page, it probably won't be found; and if Google cannot find it, it doesn't exist at all.

Now, let's look at the differences on search and access features in OSA and Finna. Search in both systems provides a highly configurable search page. It includes a familiar full-text search and depending on system multiple additional search fields for boolean logic expressions, pre-fetched facets e.g temporal, spatial and content type searches and some visualizations for those. In OSA, it is possible to estimate the search results accuracy and count before rendering the results. Of course, full-text search can be used like Google or Bing search.

Next, search is performed against the sources. Finna and OSA are not web search engines. Instead they find contents in their indexes. Finna finds materials submitted by Finnish libraries, archives and museums. OSA is more complex since it has public contents as well as restricted and confidential content per organization. By default OSA searches materials based on the user information such as organization, roles and access rights. If no user information is found, it will search only public materials for a specific organization. OSA is not a portal like Finna. Each user interface is for a single organization only. Currently, there is no cross organizational search but it is technically possible to build. Put differently, Finna and OSA use reliable and selected sources. Full-text search can understand languages, identify words, synonyms and other entities. OSA will also search the contents of rich text documents such as PDFs.

The search results are returned with harmonized and standardized metadata and in easy-to-read and understandable format. Rich metadata enables a configurable result page and additional methods of refining the results. Finna uses a template-based, modular interface with some customization options. It provides a selection of the results with small thumbnails and nutshell information on the search. Then the search can be refined with facets or additional search terms. OSA has completely configurable results view which can automatically adapt to returned data. For example, if all the search results are pictures a thumbnail view can be shown. Each organization can define the significant metadata which is displayed automatically. The search view can show different amount of information based on if the user is logged in or not, and depending on the roles and access rights. Harmonization makes it possible to use common search terms and facets to search and filter digital contents. Both systems provide access to diverse metadata records and files in a coherent manner. They support storing the original metadata as additional information.

Metadata records, previews and such can be displayed for the search results. All the data available in the index can be used for searching and can be exposed as a facet. Facets are valuable before and after the search. Before search, facets can provide suggestion and completion features and help to choose search terms that will return meaningful results. After the search, they can help to profile the results and filter the records. OSA provides download, preview and management options according to roles. Due to the origins of its materials, Finna can also display additional information on findings e.g. whether they are available for lending like books on the libraries.

The technical solution under the hood in both systems is Apache Solr. The front-end and search logic is built on top of that with different technologies. In OSA, the front-end is based on earlier development done by MAMK's digital archive projects and services. Finna uses open source VuFind and custom-made back-end for management. Both projects have put lots of effort at usability. The development model in OSA and Finna is based on agile methodologies and emphasis is put on listening to feedback from participants.

5. DISCOVERY AND ANALYSIS

In addition to relying on metadata, it is possible to extract useful information from the text collections themselves using text mining. Text mining can be used to help in the formal description of the content through automatic term extraction (Paukkeri et al. 2008) or taxonomy learning (Paukkeri et al. 2012). Complex morphologies can also be modeled using a data driven approach that has been successfully implemented in the Morfessor method (Creutz & Lagus 2007). In conceptual modeling, a data driven approach is also possible. In an early work, term-document matrices were analyzed using the self-organizing map algorithm to create the maps of documents (Kaski et al. 1998). The similarity relations between the documents emerge based on the contents of the documents without any predetermined categorizations. Since then, this kind of topic modeling has become very popular (see Steyvers & Griffiths 2007, Brauer et al. 2014). Not only the relations of documents and their topics can be analyzed in a data driven manner, but using the relations and features of words can be analyzed using similar methodology (Honkela et al. 2010, Lindh-Knuutila & Honkela 2013). From the conceptual point of view, semantic modeling in these approaches takes place within the vector space model that has a long history in the information retrieval research (Salton et al. 1975). This idea has been systematically explored in the formulation of the theory of conceptual spaces (Gärdenford 2014).

OSA demonstrates discovery and analysis by utilizing the object network created by Fedora Commons. Each entity archived or stored in OSA is a compound object consisting of multiple data streams. Fedora Commons uses a specific stream to store each object's relation information in RDF/XML format. Relation information is then indexed to a resource index which is a RDF database. By default Fedora Commons 3 ships with Mulgara which can be queried e.g. with SPARQL language. Objects in RDF database form a linked data network. OSA supports relations of any kind between the objects but currently only Dublin Core relations and a content model definition are being used. The relations network enables analysis on how entities are related and how distant the relation is. Another use case is the archival hierarchy catalog which can be built automatically and dynamically from isPartOf relations.

Discovery was found useful in Capture project when planning how the existing object network could enrich new objects during the ingest and the description process. The basic concept is that an object gains partial or complete context from surrounding linked objects such as agents, places, events and actions. These contextual objects can be formalized via ontologies or vocabularies. (Lampi & Alm 2014). It improves the description speed and information duplication is minimized. Enrichment can take place during ingest or access depending on the need. The principle is the same regardless of the timing. The process can be automatic or controlled by a user. It can add the information to the object's metadata or just modify the index leaving the original object unaltered. These developments done in the OSA project are experimental and not in production use.

6. SUMMARY

Based on the experiences and lessons learned in three case projects, it can be said that harmonization is an integral part of search, discovery and access. Depending on the source materials harmonization can require extraction and normalization before indexing can be done.

The current trend in repositories and archives is towards digitalization which causes fast growth in amount and the diversity of digital contents. The experience and research done in memory organizations could help commercial companies. This is due to the fact that challenges and drivers are more or less similar with every kind of content regardless of the owner organization.

In addition, new tools related to big data, analysis and data mining could add value to existing data that is stored in the information systems of archives, museums and libraries. However, in order to utilize new technologies and methods the data must be in good conditions regarding usability. Regarding usability, there are different aspects in content usability: machine-readability, context awareness and user experience to name a few.

Furthermore, content analysis could be used in completely new applications such as data based leadership and decision making. Statistical information about index usage could prove useful in developing services which consume the harmonized content.

This paper covered a lot of development and research done in multiple organizations and projects. As seen, many of the concepts and topics are merging and creating new features and

adding value to existing applications. Projects like OSA and Finna are not completed when their initial projects come to an end. They require the constant development and evaluation of the latest research and tools in the field. This is the kind of dialog that has been going on during the past few years and it is also the right direction for future collaboration.

Still, there is plenty of room for future research and development. To identify a few: managing the information overload, automatic curation and preservation of important knowledge and experiences, as generations before us have done.

7. REFERENCES

[1] Alkula, R., & Honkela, T. (1992). Tekstin tallennus-ja hakumenetelmien kehittäminen suomen kielen tulkintaohjelmien avulla: FULLTEXT-projektin loppuraportti (Development of text storage and information retrieval methods with natural language processing components). Valtion teknillinen tutkimuskeskus, informaatiopalvelulaitos.

[2] Alm, O., Strömberg, J. (2013). Summary of Final Report for Capture Project. http://www.elka.fi/useruploads/files/Summary.pdf

[3] Bizer, C.; Heath, T. & Berners-Lee, T. (2009), 'Linked Data - The Story So Far', International Journal on Semantic Web and Information Systems 5 (3), 1--22 .

[4] Brauer, R., Dymitrow, M., & Fridlund, M. (2014). The digital shaping of humanities research: The emergence of Topic Modeling within historical studies. In Enacting Futures: DASTS 2014 Conference (Danish Association for Science and Technology Studies), 12–13 June 2014, Roskilde University, Denmark.

[5] Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing (TSLP), 4(1), 3.

[6] Finna (2014) Mappings from Different Formats to Finna's Index. https://www.kiwi.fi/display/finna/Kenttien+mappaukset+eri+formaateista+Finnan+indeksiin

[7] Gärdenfors, P. (2014). The Geometry of Meaning: Semantics Based on Conceptual Spaces. MIT Press.

[8] Hormia-Poutanen, K., Kautonen, H., & Lassila, A. (2013). The Finnish National Digital Library: a national service is developed in collaboration with a network of libraries, archives and museums. Insights: the UKSG journal, 26(1), 60-65.

[9] Honkela, T., Könönen, V., Lindh-Knuutila, T., & Paukkeri, M. S. (2008). Simulating processes of concept formation and communication. Journal of Economic Methodology, 15(3), 245-259.

[10] Honkela, T., Hyvärinen, A., & Väyrynen, J. J. (2010). WordICA—emergence of linguistic representations for words by independent component analysis. Natural Language Engineering, 16(03), 277-308.

[11] Kaski, S., Honkela, T., Lagus, K., & Kohonen, T. (1998). WEBSOM–self-organizing maps of document collections. Neurocomputing, 21(1), 101-117.

[12] Kettunen, K., Kunttu, T., & Järvelin, K. (2005). To stem or lemmatize a highly inflectional language in a probabilistic IR environment?.Journal of Documentation, 61(4), 476-496.

[13] Koskenniemi, K. (1984). A general computational model for word-form recognition and production. In Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics (pp. 178-181). Association for Computational Linguistics.

[14] Lampi, M., & Palonen, O. (2013, January). Open Source for Policy, Costs and Sustainability. In Archiving Conference (Vol. 2013, No. 1, pp. 271-274). Society for Imaging Science and Technology.

[15] Lampi, M., & Alm, O. (2014). Flexible Data Model for Linked Objects in Digital Archives. Archiving Conference Proceedings (Vol. 2014 pp. 174-178). Society for Imaging Science and Technology.

[16] Lindén, K., Silfverberg, M., & Pirinen, T. (2009). HFST tools for morphology–an efficient open-source package for construction of morphological analyzers. In State of the Art in Computational Morphology (pp. 28-47). Springer Berlin Heidelberg.

[17] Lindh-Knuutila, T., & Honkela, T. (2013). Exploratory Text Analysis: Data-Driven versus Human Semantic Similarity Judgments. In Adaptive and Natural Computing Algorithms (pp. 428-437). Springer Berlin Heidelberg.

[18] Nilsson, M. (2010). From Interoperability to Harmonization in Metadata Standardization: Designing an Evolvable Framework for Metadata Harmonization. Stockholm: KTH. http://kth.diva-portal.org/smash/get/diva2:369527/FULLTEXT02.pdf

[19] Nyberg, K., Raiko, T., Tiinanen, T., & Hyvönen, E. (2010). Document classification utilising ontologies and relations between documents. In Proceedings of the Eighth Workshop on Mining and Learning with Graphs (pp. 86-93). ACM.

[20] Paukkeri, M. S., Nieminen, I. T., Pöllä, M., & Honkela, T. (2008). A Language-Independent Approach to Keyphrase Extraction and Evaluation. In COLING (Posters) (pp. 83-86).

[21] Paukkeri, M. S., García-Plaza, A. P., Fresno, V., Unanue, R. M., & Honkela, T. (2012). Learning a taxonomy from a set of text documents. Applied Soft Computing, 12(3), 1138-1148.

[22] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

[23] Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7), 424-440.

[24] W3C (2009) Improving Access to Government through Better Use of the Web. W3C Interest Group Note 12 May 2009. http://http://www.w3.org/TR/egov-improving/