EBU

OPERATING EUROVISION AND EURORADIO

TR 019

EBU MIM SEMANTIC WEB ACTIVITY REPORT

SOURCE: MIM

Geneva
August 2015

Page intentionally left blank. This document is paginated for two sided printing

Contents

Executive Summary **5**

1. Introduction **5**

2. Definitions **6**

3. Guiding principles of Semantic Web Technology **7**

3.1 Semantic Web and LOD 101 7

3.2 What should I do next? 9

3.3 More basic considerations 9

3.4 What effort for large scale deployment? 10

4. Current Implementations **10**

5. Conclusions **10**

6. References **11**

Annex 1: Semantic Middleware - Linked Data: RTBF "GEMS" prototype **13**

A1.1 The technological challenge of information handling 13

A1.2 Description of the RTBF/GEMS prototype: 16

Annex 2: Use of Linked Data at the BBC **19**

A2.1 Introduction 19

A2.2 A short history of the BBC's Semantic Web activities 19

A2.3 Future uses of Linked Data 20

Annex 3: EBU activities **21**

A3.1 Introduction 21

A3.2 EBU activities on Semantic Web since 2008 21

A3.3 EBU activities on Semantic Web since 2013 22

A3.4 EBU Ontologies and tools 22

Annex 4: VRT **23**

Annex 5: RAI **25**

A5.1 Media Contract Ontology 25

A5.2 Rights statements should be "machine readable" 25

A5.3 MPEG-21 MCO resulting from subdivision of MPEG-21 CEL 25

A5.4 Brief tutorial 26

Annex 6: ABC Australia **33**

A6.1 Using Ontologies for the Development of a Data Model 33

A6.2 Generating Ontologies from Databases. 33

A6.3 Generating Ontologies from XSD. 33

A6.4 Integration of Ontologies 34

Annex 7: IASA-OK **35**

Annex 8: MediaMixer Project **37**

A8.1 About Media Mixing 37
A8.2 Scenario: Re-use of Media Fragments from Video Footage 38
A8.3 Media Fragment Creation 38
A8.4 Media Fragment Description 39
A8.5 Media Fragment Rights 39
A8.6 Media Fragment Management 40
Annex 9: YLE's Meta-API: Improving the Findability of Web Content with Semantic Tagging .. 43
Abstract 43
A9.1 Introduction 43
A9.2 Tagging the Content with Vocabularies and Content Objects 43
A9.3 Creating the Tags Manually and Automatically 45
A9.4 Meta-API: Making the Tagging Data Available for Applications 47
A9.5 Front-end Applications 49
A9.6 Mapping the Vocabularies 51
A9.7 Manual Linking of Cross-Media Content 54
A9.8 Lessons learned and future work 56
Acknowledgements 56

EBU MIM Semantic Web Activity Report

EBU Committee	*First Issued*	*Revised*	*Re-issued*
TC	2013	2015	

Keywords: Semantic Web,

Executive Summary

This document provides an introduction to Semantic Web technologies and provides several use cases in the broadcast environment (**Annexes 1 - 9**).

The purpose of this report is to raise awareness on the importance and high potential of Semantic Web technologies now rapidly developing from initial conceptual prototypes to services in real production also in the broadcasting and media domain. Several successful applications of these technologies now exist for media archives. Others are being considered to enrich second-screen applications or search engines.

The MIM Strategic Programme has developed this report with contributions from ABC Australia, BBC, Titan Asbl, Perfect Memory, RAI and RTBF as well as reports from various international activities such as IASA-OK, the MediaMixer Project and YLE. These reports show that Semantic web technologies can in fact be used at all broadcast stages from commissioning through production, archiving and distribution.

Based on these findings, this report is an invitation to explore Semantic Web technologies and to investigate what they can bring to the broadcast business.

The MIM Strategic Programme is looking forward to more updates on further implementations by EBU members and will continue to study and promote Semantic Web technologies and their applications in media and broadcasting.

1. Introduction

The concept of the Semantic Web has its foundation in 1998 when Tim Berners-Lee proposed to form a consistent logical web of data [1] ("in some ways like a global database"). The reasons for such proposals were in the intrinsic heterogeneity of web resources that hindered the possibility of fully exploit the meaning, or the semantics, of the concepts and objects that the resources were about. Original theoretical foundations of the Semantic Web are also to be identified with some early works by Nicola Guarino [2], who pointed out a series of well-defined formal rules to implement ontologies.

These initial efforts were followed by an extensive process of software and standards development, mostly carried out within the W3C, the most recent of which include semantic enrichment and Linked Open Data (LOD) or complementary representation formats such as RDfa and Microdata, or JSON-LD.

Nowadays, all these technologies have an important relevance for media and broadcasting since

they contribute to the implementation of a layer of machine-readable information about the semantics of content which can be directly exploited by applications and systems.

This report is an introduction to the guiding principles of Semantic Web, illustrated by implementations from several different sources. It is an invitation to explore these technologies and what they can bring to the broadcast business. Semantic web technologies can be used at all broadcast stages from commissioning through production, archiving and distribution.

Contributions collected to assemble this document show that these technologies are being implemented and provide results meeting the expectations.

The MIM Strategic Programme is looking forward to more updates on further implementations.

2. Definitions

Links to related web resources, tutorials and specifications are provided at the end of the report.

Semantic Web	W3C. Aims at converting the current web dominated by unstructured and semi-structured documents into a "web of data". The Semantic Web stack builds on the W3C's Resource Description Framework (RDF) [3].
Linked Open data	W3C. Extends technologies such as HTTP and URIs to share information in a way that can be processed by computers. This enables data from different sources to be connected and queried. It is an extension of the Semantic Web [4] [5] [6] [7] [8].
RDF	W3C Resource Description Framework. RDF is a standard model for data interchange on the Web [9].
OWL	W3C Web Ontology Language [10]. OWL 2 ontologies provide classes, properties, individuals, and data values and are stored as Semantic Web documents. OWL 2 ontologies can be used along with information written in RDF, and OWL 2 ontologies themselves are primarily exchanged as RDF documents.
Schema.org	Provides a collection of schemas [i.e., html tags using the syntax of RDFa or Microdata] that webmasters can use to mark-up their pages in ways recognized by major search providers [11]. Search engines including Bing, Google, Yahoo! and Yandex rely on this mark-up to improve the display of search results, making it easier for people to find the right web pages. A user and developer group has been established in W3C.
TV Radio Schema for schema.org	A joint BBC-EBU proposal for extending schema.org to address TV and Radio Programmes and associated publication events [12].
W3C Media Annotation	Based on a core set of properties which covers basic metadata to describe media resources. It defines syntactic and semantic level mappings between elements from existing formats. It has been developed to describe media resources on the Web [13].
W3C Media Fragment	Specifies the syntax for constructing media fragment URIs (URL combined with start time and duration) and explains how to handle them when used over the HTTP protocol [14].
W3C Web & TV	Provides a forum for Web and TV technical discussions, to review existing work, as well as the relationship between services on the Web and TV services, and to identify requirements and potential solutions to ensure that the Web will function well with TV [15].

3. Guiding principles of Semantic Web Technology

What is this section about?

This section is intended to reassure you that working with the Semantic Web is EASY.

It is not intended for experts but for newcomers who want to discover these technologies. Although the word "ontology" is often associated with the Semantic Web, no philosophical references will be used and the intention is demystification.

3.1 Semantic Web and LOD 101

In a nutshell, the Semantic Web is about presenting information using simple phrases or statements. If you are able to expose your model simply and logically, you are good to go.

Let's take the book analogy. If someone were to describe what a book is, he may say:

- A book has a title.
- A book is organised in chapters.
- A chapter has a number.
- A chapter contains paragraphs.
- A paragraph contains sentences
- A sentence has a subject.
- A sentence has a verb.
- A sentence has a complement.

The same following the Semantic Web approach:

Class	ObjectProperty	Class	DataProperty	Datatype/Value
Book			hasTitle	*value (string)*
Book	isOrganisedIn	Chapter		
Chapter			hasNumber	*value (integer)*
Chapter	contains	Paragraph		
Paragraph	contains	Sentence		
Sentence			hasSubject	*value (string)*
Sentence			hasVerb	*value (string)*
Sentence			hasComplement	*value (string)*

Another key aspect of the Semantic Web is the idea that URIs can identify things, not only web pages. In our example a particular book could have a URI such as <http://mybookrepository.com/books#1>. As explained below, such a URI can also be associated with an ISBN number.

We could actually stop here. From a pure technical perspective, the Semantic Web is not more complicated than that and the same principle applies whether the model is simple or complex.

But let's take this opportunity to go one step further in some of the concepts and definitions:

- Triple: it is a statement like "Book isOrganisedIn Chapter" made of a subject, a verb (or predicate) and a complement (or object or resource or value)
- Class: classes represent the main objects / resource of your model. The

example speaks of a Book and Chapters, but it could equally be a Programme and Clips. Class are uniquely identified by Unique Resource Identifiers in the form of a URL (Uniform Resource Locator) or URN (Uniform Resource Name)

- ObjectProperties: these properties are used to establish relations between classes / resources.
- DataProperties: these properties are used to give values corresponding to simple datatypes (strings, dates, integers, URIs, etc.)

If we now want to instantiate this model, we are going to identify a particular book using for example its ISBN Number 1438886851 (dummy). We won't delve into the syntax but it basically means:

- About ISBN-1438886851 hasTitle "my dummy example"
- About ISBN-1438886851 hasChapter ISBN-1438886851-C1
- About ISBN-1438886851-C1 hasNumber "1"
- About ISBN-1438886851-C1 hasParagraph ISBN-1438886851-C1P1
- Etc.

Using a good consistent identification scheme is vital. It is the "glue" that allows machines to reconstitute / infer / derive / automatically reverse engineer your model without even needing to know what the model or an ISBN number is about. As mentioned earlier, an ISBN number can be used and extended, which can be further associated with a URL or URN.

However such a consistent identification scheme is not enough. The notion of Linked Data introduces the notion of 'dereferencable' URIs.

The URI <http://mybookrepository.com/books#ISBN> should allow accessing more information about this book, including, for example, links to its chapters. If each chapter is then in turn identified by a URI, e.g. <http://mybookrepository.com/chapter#ISBN-11> more information is provided about theach chapter. Following the same approach more and more information can be aggregated about that particular book.

We can also use Linked Data by adding a new "isRelatedTo" property to our book example:

- Book isRelatedTo Book

Or

- About ISBN-1438886851 isRelatedTo ISBN-1439996422

This can be extended at will to define all sorts of relations linking to all sorts of relevant resources, which can either user friendly (a web page) or machine friendly (RDF/OWL or an alternative machine readable representation of the information). As an example, one could think of a new property like "hasEditor" pointing to the webpage of an Editor.

- Book hasEditor Editor

Or

- About ISBN-1438886851 hasEditor http://www.mydummyeditorwebpage.com

Linked Open Data offers a long awaited solution to the dereferencing of classification schemes or controlled vocabularies like for Genre. This can be illustrated by the following example using SKOS.

- - Book hasGenre Genre

Or

- - About ISBN-1438886851 hasGenre http://www.ebu.ch/metadata/ontologies/skos/ebu_ContentGenreCS.rdf#_3.1.6.1

While dereferencing this Genre term through its URI, a machine would be then able to automatically get the associated preferredLabel, in this case "applied sciences". http://www.ebu.ch/metadata/ontologies/skos/ebu_ContentGenreCS.rdf is a valid link.

But the notion of relation is a good preliminary step in what makes semantic modelling so interesting: reasoning an inference. In other words, a good ontology will help highlighting / deriving new relations (in the form of additional simple statements) from the constituted knowledge base. This can be simply expressed as follows:

Paragraph1 isPartOf Chapter1, Chapter1 isPartOf BookA

then a reasoner will infer and reflect in a simple statement that

Paragraph1 isPartOf BookA

While building more expertise, one of your goal will be to explore the possibilities offered by the Semantic Web and Linked (Open) Data to define better models and enrich your data.

3.2 What should I do next?

As you can see, we have been able to go through the main guiding principles of the Semantic Web and Linked (Open) Data without writing a single line of RDF (Resource Descriptive Framework), OWL (Ontology Web Language), TTL (Turtle for Terse RDF Triple Language), RDFa or Microdata used to embed/hide triples into HTML webpages for the attention of search engines, as promoted by schema.org (an initiative from Google, Bing, Yahoo! And Yandex). The good thing is that you can continue your investigations without these languages.

Early developers had to hard-encode their ontologies, which they would test using early versions of validators that usually returned cryptic error messages. Editors are now available that allow a user to concentrate more on the model and its semantics rather than on the syntax. *Protégé* is one of these user-friendly editing tools (http://protege.stanford.edu/download/download.html) that is especially suitable for beginners. Another tool, aimed at developers is TopQuadrant's TopBraid Composer (http://www.topquadrant.com/). Examples and documentation can be found on both websites.

Your model will improve as your expertise grows but the approach will actually remain the same.

3.3 More basic considerations

We hope this report will encourage you to further investigate the Semantic Web technologies. These bring the agility and flexibility that metadata developers and users have long been looking for, while developing XML common schemas for more interoperability.

What makes the Semantic Web offer more in terms of interoperability is its format. All you need use are triples. XML data from legacy silos can be converted into triples and combined/associated with new ontologies by defining appropriate relations (also expressed as new triples). The model will not break.

All this doesn't mean that XML is doomed; XML still remains very powerful in terms of validation. You need no longer hesitate in converting XML data into triples to bring data together in your organisation.

It is recommended that you do not develop automatic translations from XML schemas into RDF. Keep your model in mind! Step back and think how you can best express your data model using classes and simple statements that include semantically rich relations between classes.

You may also want to consider the use of tools to convert the data of your relational databases (good definition of well identified classes and well defined relations between tables) into RDF for certain applications.

3.4 What effort for large scale deployment?

Of course, there is no such thing as a free lunch!

Large scale developments require:

- The acquisition of expertise in these technologies.
- Investing in developing a sound data model for your domain of application
- Developing new ontologies or deriving new models from legacy data structures
- Studying what ontologies are already available that solve part of your problem, and understanding how to reuse them

L(O)D has its own requirements:

- L(O)D doesn't mean zero cost, in contrary it is a significant investment
- You need to know why you want to use L(O)D for
- Linked Data doesn't have to be Open
- Persistence, i.e. the property of data being accessible and stable indefinitely over time, is an issue and some users actually aspire data they link to for backup
- Disambiguation is an issue (Paris refers to "Paris, Capital of France", "Paris, Texas" or "Paris Hilton"?). This means that you cannot get rid of some minimal level of supervision.
- The editorial quality of linked data is important.
- There is not any zero-cost certification mechanism available.

However, you first ambition will be to link your internal data.

4. Current Implementations

Annexes 1 - 9 show what implementers do with the Semantic Web and Linked (Open) Data technologies. Each Annex goes further into the technical and theoretical meanders of the Semantic Web and L(O)D. They demonstrate the potential of the Semantic Web as a pervasive framework supporting all aspects of modern media data management: archives, rights, publication, production and programme exchange. LOD is not only a matter of knowledge management; it is a powerful tool that gives broadcasters a means to better expose and value their content and know-how. Today, the broadcaster's challenge is to connect each link of the audiovisual chain to their own L(O)D.

5. Conclusions

This updated report aims at presenting Semantic Web technologies to EBU Members through a detailed disclosure of their use in some practical use cases related to media and broadcasting.

Semantic Web technologies represent a revolution, rather than an evolution, of traditional ways of managing data and metadata. This is due to the introduction of a few but nevertheless very powerful tools, all based on the W3C's RDF (Resource Description Framework) standard. This

standard was first issued in 1998 as a working draft and subsequently standardized in 2004. It enables the realisation of unprecedented scenarios.

Though grounded on very rigorous theoretical bases, Semantic Web Technologies have an unexpected friendly face, as they appear as natural language sentences with a subject, a predicate and an object. Even this quite simple data model (equivalent to a graph) is a really powerful interoperability machine, capable of modelling many of the data structure artefacts employed in information systems, such as XML trees and relational tables.

The adoption of Semantic Web technologies is a big opportunity to boost the exploitation of metadata in media organizations by making data "alive" and "linked" with world-wide knowledge through the Semantic Linked Open Data (LOD).

As with any new technology, the Semantic Web also brings along some issues. These are specifically related with the paradigm shift that it introduces with respect to more established and enduring ways of managing metadata, such as, for example, XML Schemas.

This implies that engineers working in the domain of metadata in their organizations need to acquire a new perspective of their work by studying the key elements of the Semantic Web and starting to apply them in their new projects rather than trying to perform a blind data mapping from XML. Once acquired, this way of working produces outstanding results both in increasing modelling productivity and in the expressiveness of results. The impact of the Semantic Web is probably only comparable to that resulting from the introduction of XML in early 2000.

6. References

[1] http://www.w3.org/DesignIssues/Semantic.html

[2] N. Guarino, C. Welty, "A Formal Ontology of Properties", 12th International Conference on Knowledge Engineering and Knowledge management, 2000.

[3] http://en.wikipedia.org/wiki/Semantic_Web

[4] http://www.w3.org/standards/semanticweb/data

[5] http://open-data.europa.eu/open-data/linked-data

[6] http://linkeddata.org

[7] http://en.wikipedia.org/wiki/Linked_data

[8] http://www.oclc.org/research/activities/linkeddata.html

[9] http://www.w3.org/RDF/

[10] http://www.w3.org/TR/owl2-primer/

[11] http://schema.org/

[12] http://www.w3.org/wiki/WebSchemas/TVRadioSchema

[13] http://www.w3.org/TR/2010/WD-mediaont-10-20100608/

[14] http://www.w3.org/TR/2012/PR-media-frags-20120315/

[15] http://www.w3.org/2011/webtv/wiki/Main_Page

Page intentionally left blank. This document is paginated for two sided printing

Annex 1: Semantic Middleware - Linked Data: RTBF "GEMS" prototype

Roger Roberts, RTBF R&D – Knowledge management (rro@rtbf.be)
Steny Solitude – Perfect-Memory (http://www.perfect-memory.com)
Guy Maréchal – Memnon (http://www.memnon.eu)

A1.1 The technological challenge of information handling

Broadcast data types and applications can be classified in several functional domains': data model (information object) versus data set (data bases), live, dynamic and fixed data. The requirements of each of these "functional domains'' are such that each requires distinct technological solutions and exploitation rules.

Assets are of different nature and IT equipment and strategies have distinct requirements in each case:

- a dynamic process is always involved in the life cycle of the assets. The data are built in real time, from transactions or through creative work. But at some moment, most of them have to be fixed.

The information assets could be of any nature such as of cultural, scientific, social, political, medical historical interest:

- some of the assets are generated and made accessible through a DBMS (Data Base Management System) and are called "STRUCTURED".; i.e. that the assets is global and the individual elements are embedded in the structure expressed in the DBMS.
- the other assets (called "UNSTRUCTURED") are managed by independent semantic items. Each of these items has its structure within its instance of representation and is mostly 'floating' in the computers environment.

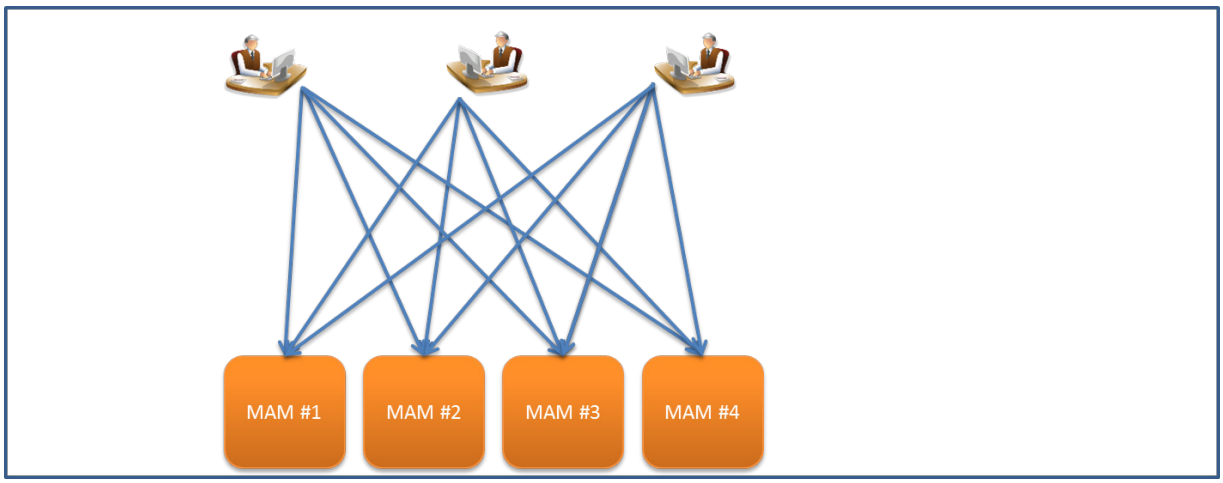


Figure A1.1: the actual state of the art (network perspective)

All of these solutions require custom integration of many discrete hardware and software components, as well as application development, which generally lead to proprietary solutions that do not extend through the ages. Moreover, the database technology requires a lot of human resources to encode and update the audio-visual material.

It has been a long time that users (producers, distributors, consumers, rights holders, etc.) of audiovisual, textual and iconographic content expect a computer open architectural framework, where the various management and exploitation functions of audio-visual materials are fully integrated.

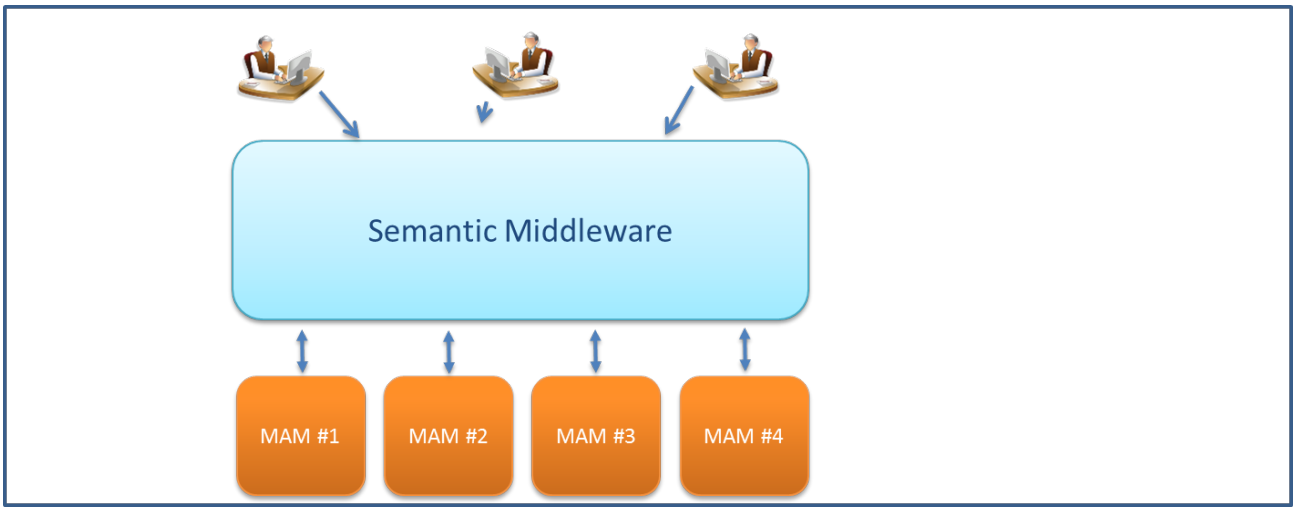


Figure A1.2: the user's perspective (MediaMap architecture)

For the last couple of years, partial answers appeared in the Internet world (Linked Data) and a semantic middleware built upon the W3C standards (RDF, OWL, SKOS, ...) processes the relationship between a data (its representation) and the information (the meaning) so that applications and Databases can be operated through common interfaces.

Based on this vision, the Eureka Celtic MediaMap Project (intended to improve the collaborative production of audio-visual subjects between amateurs and professionals) has designed and developed a middleware based on semantic standards.

For years, the industry has developed formats to encapsulate metadata models in AV wrappers able to manage the essences but less adapted for information handling. Using a radically new approach, the project has reversed the roles conferring to the information to encapsulate the essences! The partners have constructed a logical/physical model that encapsulates the media, a production intranet that interconnects different databases (internal and external) and built a customization of the view inside the system.

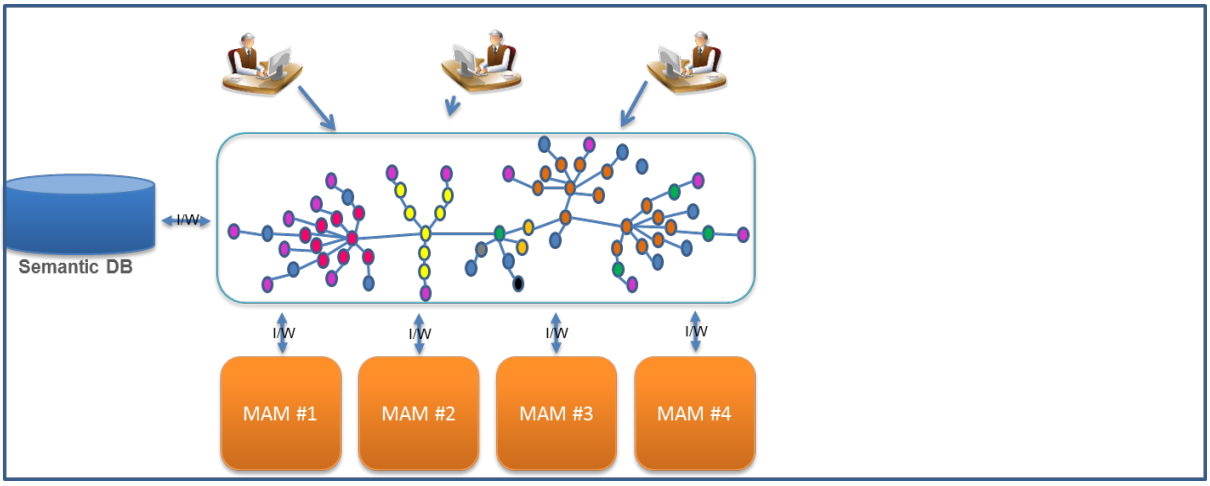


Figure A1.3: the MediaMap architecture: USE - OSB - IW and VIEW

The MediaMap middleware relies on four concepts: the production and annotation ontologies, the media wrapper (USE: Unique Semantic Entity), the network (OSB: Open Semantic Bus - Interoperability Windows), and finally the personalized vision (VIEW).

- The audio-visual production ontology is a conceptual model that describes in terms of products, contributors, roles and rights any object (Editorial Object - Annotation - Temporal Object - Physical Object.) This ontology is interoperable with others (EBUCore) and constructs the access and views to AV projects for each Member involved in the collaborative process. Just like the EBUCore, it structures the information inside and outside of the middleware.
- The wrapper packages any content to store or exchange: its wrapping is semantically described on the basis of the ontology, which can be defined as an entity that "hooks knowledge to knowledge". One can say that the package is "autonomous" in the sense that it includes not only the instances of the classes but also the definition of the classes.
- The network is based on a controlled distributed architecture, which manages the semantic messages flow. The spatial and temporal interoperability with each application connected is provided through an interoperability window (IW).
- The middleware allows the design, the consultation, and the editing of the entities through workmanship oriented interfaces.

The semantic middleware allows:

- to set properties/relationships that shares explicit metadata (machine process-able)
- to add machine-readable metadata to existing content so that information can be analyzed, questioned, shared, reused, ...
- and thus to enable the identification of new relationships by machine reasoning and deductions (inferences).

In addition, the fact of having explicit identified objects and relations enables automatic reconciliation of an uncontrolled information source, facilitating enrichment, research and information processing. All these external resources can be represented through a distributed and federated data knowledge graph which is instantiated as a network of occurrences of the classes of the graph.

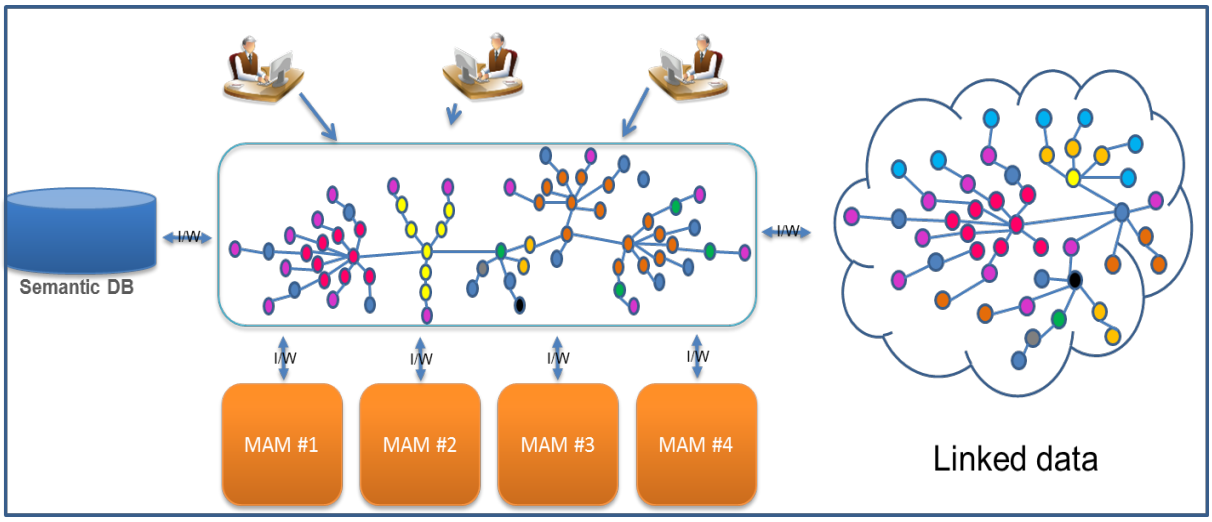


Figure A1.4: Linked data: the internal and external interoperability

From such a semantic level, it becomes possible to use the full power of the computer so that the user can manipulate the data at the information level. Because of the semantic web standardization this will be the case for any information produced inside or outside of an organization.

Benefits of a semantic environment:

- Applications and websites can handle information and not data anymore
- Search tools are able to display the results of the more relevant information in the result of semantic browsing facilities.
- The future big data mashers will be likely to combine information from different sources to create new relationships and serendipitous search.

A1.2 Description of the RTBF/GEMS prototype:

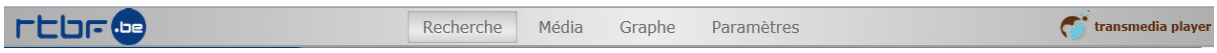
In order to validate all of these new concepts, the RTBF decided to conduct a POC (proof of concept) implementing all mechanisms designed by the MediaMap project. From the functional point of view, the semantic middleware has been set at:

- the RTBF Knowledge Base level that provides all users with the available instances of the transmitted Profile.
- the RTBF Profile which integrates the Ontology and the Knowledge Base. External links (linked data) belongs to the Profiles mobilized by the AV project.

Perfect Memory has deployed the semantic environment (including 4 relational databases: Netia - Dalet - Tramontane audio – Tramontane video) for the enrichment of publishing and cross-media promotion programmes and services and the RTBF documentation content-based interoperability with archiving tools. Memnon has provided sound services analysis for the semantic media enhancement.

The RTBF has provided newscasts and news magazines content which have been semantically ingested in conformance with the USE concept. AV files have been encapsulated in a logical structure defined by the audio-visual production ontology. A speech-to-text analysis tool is activated during the ingest process. It generates representative tags (each qualified with its classified topics) synchronized with the stream.

The developed application provides four interfaces manipulated through following tabs "search", "media", "graph" and “parameters”.



Recherche – Search

On the search page, nine "categories" are semantically described, namely:

- date	- programme	- text
- physical person	- location	- activity
- object	- organization	- abstract (concept)

Each of these elements and their combinations can be subject to multiple manipulations! Other categories can be created by a user and added to the existing database. The results are rich multimedia content that can be viewed directly in the interface.

Media : Content visualisations

The visualization of the content offers:

- The presentation of all information on the programme from which the segment is extracted, at the bottom of the page, a compass that indicates the position of the segment in the editorial object (programme)

- The presentation of semantic "entities" deemed significant for the segment/editorial object. This displays the metadata's that were semantically treated.
- The structure in chapters of the editorial object
- The presentation of the time-coded semantic annotations of the media (physical person, location,..) extracted (automatically or not) of the editorial object (full programme).

All the metadata handled in the network are first subject to an alignment with semantic data repositories (explicit metadata). On this semantic status, the metadata can be linked with others and so enriched. The extracted metadata corresponding at the name of a politician is enhanced as being the instantiation of the concept "physical person". From this semantic statement the middleware collects all the information related to this explicit term in connected knowledge bases.

Knowledge graph

The platform provides for each segment an explicit representation, a knowledge graph consolidated in the RTBF semantic database.

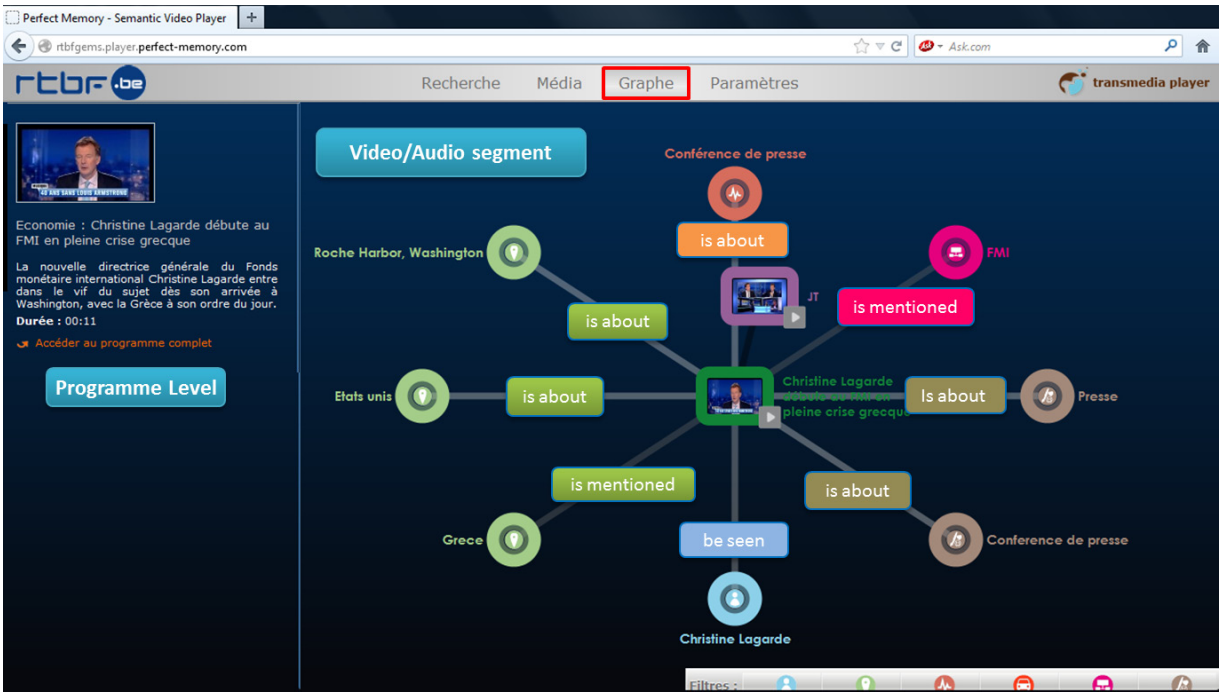


Figure A1.5: The Semantic Video Player Interface: contextualization of the content

Indeed, the contextualized entities of each segment/editorial object offer an explicit navigation in the structure of the information displayed. Each entity can be subject to interrogations, and the tool restructures the representation of information based on the new collected data's. In fact, the developed tool offers a dual information modelling: the objects constructed by the audio-visual production chain but also the knowledge built around these AV objects.

Each category is displayed with the characterized relationship (Greece is mentioned in the subject, Christine Lagarde is visible, it's a Press Conference, etc.). It goes without saying that each segment/editorial object may be display in the Graph interface!

Parameters

The last page contains the interface parameters (current, active filters by default).

Lessons learnt from the project:

The establishment of a semantic middleware and linked data's offer the following advantages:

- Interoperability between databases and unstructured content
- A well designed robust architecture
- An automatic low cost collection of information and enrichment
- Traceability of media (human, location) treatments
- Tools for 360° Publication, multiplatform content monetization

The new Eureka Celtic MediaMap+ project:

- The opening of the audio-visual production chain raises a number of conditions including the clarification of the prescription, the rights management, of a content in structured and standardized languages still handled today by humans but tomorrow by machines.
- The MediaMap+ project should offer to the AV industry a transparent access to all the dedicated resources, structured and standardized by open languages. This should allow new, fasten and low cost processes for rich publishing on multiple devices!

Annex 2: Use of Linked Data at the BBC

Yves Raimond, BBC (Yves.Raimond@bbc.co.uk)

A2.1 Introduction

Our use of Linked Data at the BBC can be split in three main categories.

- **Publishing Linked Data**: to make our content more findable (e.g. by search engines) and more linkable (e.g. via social media or by other Linked Data publishers using the same vocabularies and identifiers). In particular, we worked with the EBU and Google to write a schema.org extension for TV and Radio in order to improve search results around broadcast content. We publish Linked Data around all our programmes through our bbc.co.uk/programmes automated programme support platform, as well as through a variety of other sites (e.g. bbc.co.uk/music and bbc.co.uk/nature);
- **Consuming Linked Data:** to "borrow" additional context for our content where we don't have existing data and want to cut content by specific domains (music, nature, food, sport). The Linked Open Data that we use also helps give us additional links between domains.
- **Managing data internally as Linked Data:** to maximize the use we get out of editorial input by propagating editorially added links across data graphs; to make more links between otherwise siloed sites; through the use of the BBC's Linked Data Platform.

A2.2 A short history of the BBC's Semantic Web activities

It is difficult to pinpoint an exact moment when the BBC first started to use Semantic Web technologies. It was more something we have evolved toward from a shared approach and shared philosophy. We have been thinking in Linked Data terms for seven or eight years without necessarily using specific technologies. A rough chronology would be:

- **2004:** Around 2004, work started on PIPs (programme information pages), which aimed to create a Web page for every radio programme broadcast by the BBC. This began our approach of using one page (one URL) per thing and one thing per page (URL).
- **2005:** Tom Coates published "The Age of Point-at-Things," a blog post filling out some of the thinking behind giving things identifiers and making those identifiers HTTP URIs. Also in 2005, BBC Backstage was launched as an attempt to open BBC data and build a developer community around that data.
- **2006:** Work began on */programmes*, a replacement for PIPs covering both radio and TV. Around the same time we bought — in bulk — copies of Eric Evan's "Domain Driven Design" which influenced the way we designed and built websites to expose more of the domain model to users. Building on Backstage, we added data views to */programmes* (JSON, XML, YAML, etc.).
- **2007:** In 2007, we started work on rebuilding */music* as a way to add music context to our news and programmes. Because we didn't have our own source of music metadata we looked for people to partner with and settled on MusicBrainz because of their liberal data licensing. Previously we had silo'ed micro-sites for programmes and music. By stitching MusicBrainz artist identifiers into our playout systems we linked up these silos and allowed journeys between /programmes and /music. At the same time as we started to consume open data, we also started to publish Linked Open Data, creating the Programmes Ontology and adding RDF to both */music* and */programmes*. At the time, we found it much easier to develop separate but related applications in a loosely coupled fashion by dogfooding our own data: /programmes uses data views from */music* and vice versa.

- **2008:** We rebuilt more of bbc.co.uk (/*nature* and /*food*) according to domain-driven design and Linked Data principles, publishing a Wildlife Ontology and RDF for /*nature*. Again we borrowed open data to build a framework of context around our content: this was the start of us using the web as our CMS and the web community as our editors.
- **2010:** Published the World Cup website using a BigOWLIM triple store (a triple store is a database that stores RDF data). News articles were tagged with entities in the triple store and inference used to propagate those tags to all relevant entities through the graph.
- **2011:** Rolled out the World Cup approach across the whole of BBC Sport.
- **2012:** Rolled out the Olympics site using the same model as BBC Sport. Start of the BBC's **Linked Data Platform**, providing a rich set of controlled vocabularies and ontologies, which can be used to categorise content in a wide range of domains and to drive a number of features on the BBC website.

A2.3 Future uses of Linked Data

We are currently exploring various other uses of Semantic Web technologies within BBC R&D. In particular we're looking at ways in which Linked Data can be used to help search and discovery of archive content. We have been working on automatically identifying the topics and the contributors for BBC programmes from their content, using a combination of Linked Data, signal processing, speech-to-text and Named Entity Recognition technologies, which we have been talking about in various places, such as the Linked Data on the Web workshop and at WWW '2012 . The automatically generated links from programmes to entities described in the Linked Data cloud might be incorrect in places, so we are also exploring how users can validate or correct those links, and how this feedback can be taken into account within our automated interlinking workflow. We wrote about our experiments on the BBC R&D blog:

- The World Service archive prototype;
- Developing the World Service archive prototype;
- Developing the World Service archive prototype: UX.

We are currently annotating quite a lot of our content with Linked Data URIs to drive a number of aggregations on our site, but we are making little use of the connections between all these URIs. So far, we have only been using those in our automated tagging tools, to disambiguate between candidate identifiers. There is a big opportunity in using those connections for storytelling purposes — using paths in that graph of data to help tell stories around our content. It becomes even more of an opportunity if we start describing the content of individual programmes in more details, such as describing the narrative structure of dramas, for example. We started some investigation in that area in our Mythology Engine project, but there is much more that could be done.

The Linked Data Platform will continue to explore the use of geo-spatial and temporal aggregation of content, with an expanding range of BBC content. This project also aims to provide a public API, giving powerful new ways to access our data.

Annex 3: EBU activities

A contribution from EBU Technical Development and Innovation
Jean-Pierre Evain (evain@ebu.ch).

A3.1 Introduction

EBUCore has been working on Semantic Web for several years now with the main purpose of raising awareness within the EBU metadata expert community.

A3.2 EBU activities on Semantic Web since 2008

- Early investigation on the value of SW technologies started in 2008 by studying the main specifications, identifying tools and developing early implementation such as the RDF/SKOS representation of EBU controlled vocabularies.
- http://tech.ebu.ch/semanticweb_ebu is a page of the EBU Innovation and Development department with collection of important information about Semantic Web.
- In 2009, an EBU paper was presented at IBC: "Is Semantic Web part of the broadcasting future". The intention of the presentation made during the conference was educational with a 101 walk through SW, followed by examples of implementations combining EPG and news ontologies (how to bind a news programme description into an overall EPG metadata flow). (http://tech.ebu.ch/docs/metadata/ibc2009_JPE_SemanticWeb.pdf)
- Another paper was published in the EBU Technical review on "Semantic TV" (http://tech.ebu.ch/docs/techreview/trev_2009-Q3_SemanticWeb_Evain.pdf)
- The EBU is a W3C member and joined the W3C's Media Annotation Working Group (http://www.w3.org/2008/WebVideo/Annotations/). "The mission of the Media Annotations Working Group, part of the Video in the Web Activity, is to provide an ontology and API designed to facilitate cross-community data integration of information related to media objects in the Web, such as video, audio and images." In this framework, the EBU has developed several mappings including for EBUCore, TV-Anytime and NewsML-G2. The EBU has also been directly involved in the authoring of the RDF media-annotation ontology (ma-ont). This resulted in the W3C "Ontology for Media Resources" (http://www.w3.org/TR/2012/REC-mediaont-10-20120209/).
- Early 2012, the EBU - as a member of the IPTC - has participated in the work on the rNews ontology although it slightly off scope of EBU with a focus on press website publication.
- More recently, the EBU worked with the BBC and Google on an extension of schema.org for TV and Radio to improve search results around broadcast content (refining the concepts of TV and Radio Programmes and introducing Services and Publication Events).
- The EBU has joined the EUScreen European project as "technology provider" bringing EBUCore as a reference schema. EUscreen also used EBUCore RDF to generate and submite data to Europeana as Linked Open Data. EUScreen has stopped in 2012 and will be replaced EUScreenXL starting in March 2013. In the framework of the EUSCreenXL project, a profile of EBUCore will be mapped to the Europeana's EDM ontology: (http://www.europeana.eu/schemas/edm/).
- In 2012, the EBU also published the Class Conceptual Data Model (an RDF ontology) representing important elements of the broadcasting operation from commissioning to distribution. This work has been done in collaboration with the experts of the EBU MIM and MIM-MM community, and in particular ABC Australia and VRT. The VRT has adapted CCDM in two projects on archiving and system integration.

A3.3 EBU Semantic Web activities since 2013

Since the first version of this report was published in 2013:

- The EBUCore RDF ontology has been updated.
  - The technical properties of EBUCore are used by the Europeana project (Europeana Data Model Profile for sound).
  - The EBUCore and PBCore communities are actively working on the harmonisation of their semantic metadata sets around EBUCore RDF.
  - EBUCore is listed as a linked data vocabulary (LOV) - http://lov.okfn.org/dataset/lov/vocabs/ebucore.
  - At the initiative of Penn State University, EBUCore is listed as RDF-Vocab for Ruby developers, which is connected to developments on Hydra and Fedora 4 - https://github.com/ruby-rdf/rdf-vocab.
  - The use of EBUCore RDF is being investigated by the Dwerft project in Germany.
- The EBU has started developing an ontology for sport called EBUSport, originally tested for athletics and now tested against 30 other sports with different data structures (events and results).
- The EBU CCDM ontology is being updated.

A3.4 EBU Ontologies and tools

EBUCore

The EBUCore ontology can be accessed from: http://www.ebu.ch/metadata/ontologies/ebucore/.

In order to comply with rules of Linked Open Data, it is important that users can dereference URIs to definitions of the ontology. This can be done through accessing the HTML documentation or directly the RDF representation of the ontology. The EBU server has been setup to accept rdf and html requests following recipe 3 of the "Best Practice Recipes for Publishing RDF Vocabularies" (http://www.w3.org/TR/swbp-vocab-pub/), which can be tested using the "Vapour Linked Data Validator" (http://validator.linkeddata.org/vapour).

The EBUCore ontology can be accessed from: http://www.ebu.ch/metadata/ontologies/ebucore/ or http://www.ebu.ch/metadata/ontologies/ebucore/ebucore or http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#title

These links take you by default to the documentation of the ontology.

However the default could be changed to access of the RDF file, which can be opened from http://www.ebu.ch/metadata/ontologies/ebucore/ebucore.rdf.

CCDM - Class Conceptual Data Model

The CCDM ontology can be downloaded from http://www.ebu.ch/metadata/ontologies/ccdm/20120915/CCDM_Core.owl . This will soon be updated to comply with the "Best Practice Recipes for Publishing RDF Vocabularies" .

NRK and VRT are collaborating on an update of CCDM.

SKOS - EBU Classification Schemes and controlled vocabularies

All EBU controlled vocabularies are now available in RDF/SKOS and accessible from:

https://www.ebu.ch/metadata/ontologies/skos/.

Annex 4: VRT

Mike Matton, VRT (mike.matton@vrt.be)

Recently, the VRT has adopted the EBU CCDM standard into its production projects. Two projects have started up which use CCDM as underlying data model: namely the new archival system, and the new integration layer.

It is planned to replace the current archival system at VRT. The tender corresponding with the replacement included a reference data model has been devised based on CCDM. The proposed model is not binding, but it is a clear benefit if the supplier is conformant with this data model. The diagram of the reference data model is shown in Figure A4.1. The implementation has not yet started at the time of writing this report.

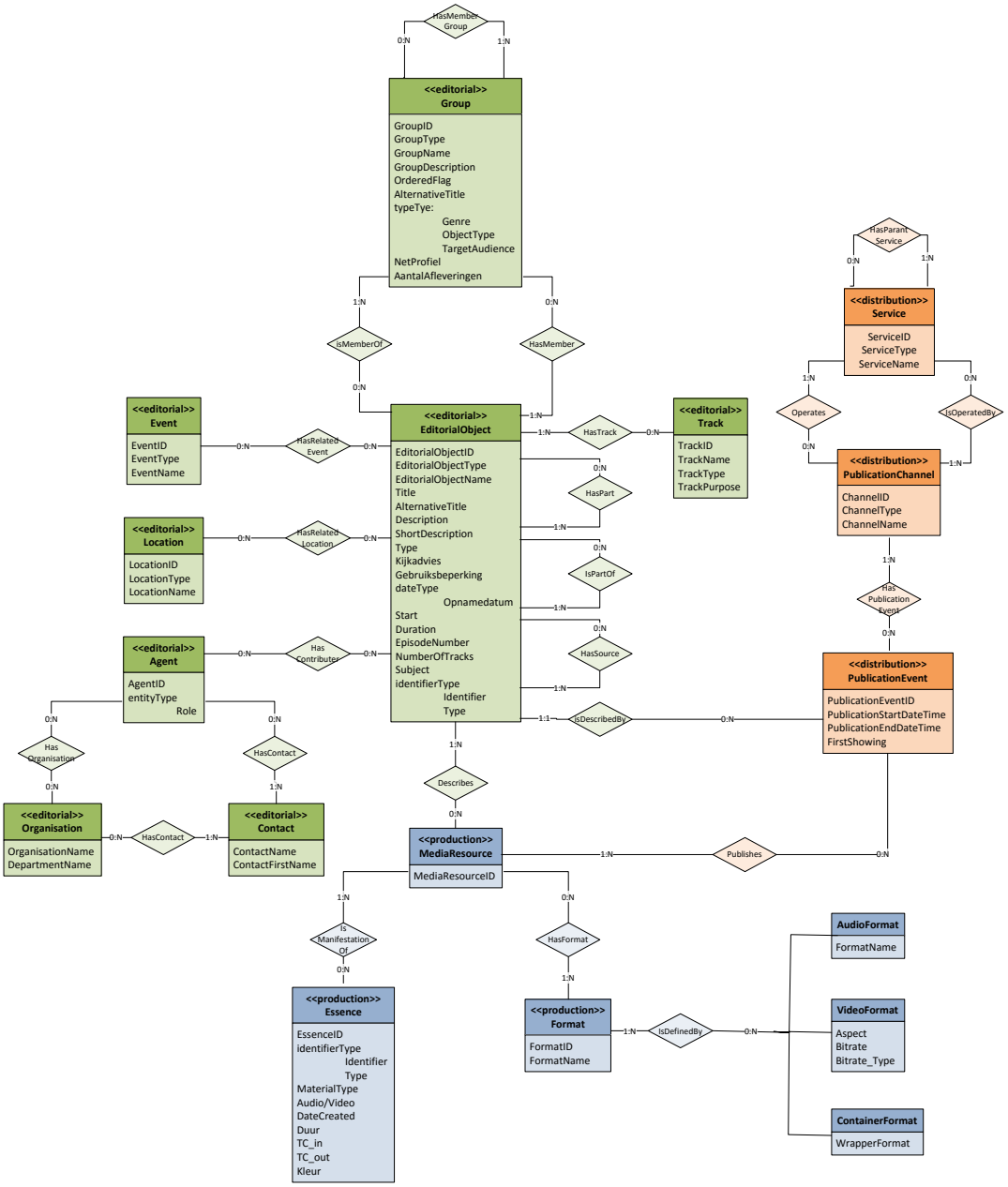


Figure A4.1: Reference archive data model (VRT)

Secondly, the architecture of the integration layer combining different media services is being redrawn. For this purpose, a new team called MIG (media integration group) has been formed. The

task of this team is to develop integrations between different components in the VRT infrastructure. In order to successfully build these integrations, a central data model is necessary. The current data model is based on the EBU CCDM standard. It is shown in Figure A4.2. Only the components required for current integrations have been implemented. It will be extended as needed for future integrations. Current integrations are (among others) looking into integrating subtitles and loudness in the workflows.

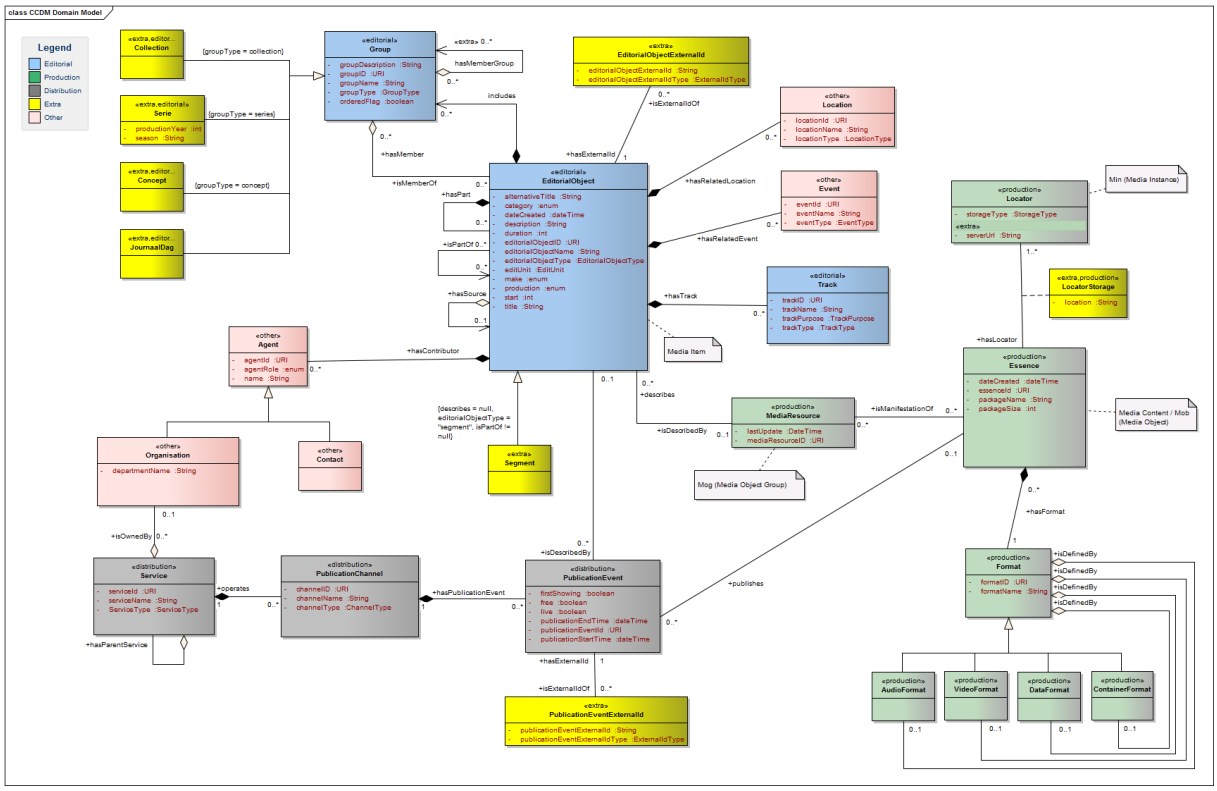


Figure A4.2: VRT media integration data model

Annex 5: RAI

Laurent Boch (laurent.boch@rai.it)
Annarita Di Carlo (annarita.dicarlo@rai.it)

A5.1 Media Contract Ontology

In the last years RAI have been strongly engaged in activities related to audiovisual rights, in particular in the framework of the European funded project PrestoPRIME[1] and including involvement in standardisation, within the MPEG-21 framework, that has resulted in a Media Contract Ontology (MCO). The scope of such activities addressed real contract terminology, rights modelling, standard format for representing rights, software tools and solutions for rights management

A5.2 Rights statements should be "machine readable"

The text of contracts is not in general easy reading. However professional persons in the legal domain are capable of understanding the various agreed terms, discerning between important and flyweight aspects, and deciding on apparent inconsistencies. In other words the narrative contract text is not good input for Natural Language Processing (NLP) tools to take automated decisions, because even a single word can make the big difference.

Actually the main requirement for a rights representation format is to have "unambiguous" statements, so that rights information has to be "machine readable", i.e. a processing can be defined and implemented to check contexts, provide matching results, and even take decisions about acting an action or not.

A5.3 MPEG-21 MCO resulting from subdivision of MPEG-21 CEL

MPEG-21 part 21, i.e. the Media Contract Ontology (MCO) currently under ballot as FDIS (Final Draft International Standard), resulted from the subdivision from MPEG-21 part 20 (Contract Expression Language, CEL). So the latest version of CEL provides the XML structure representation of contract only, while MCO provides the OWL semantic representation of contract.

The scope of Media Contract Ontology is the whole contract, with the exception of the economical aspects.

- Identification of the contract and of possible relations with pre-existing contracts
- Identification of the parties and signatories (with signatures)
- Identification of the object of the contract
- Unambiguous representation of the agreed deontic[2] expressions
- Possibility to encrypt the whole contract or parts of it

The MCO document can be a Contract document; however MCO can still be used as in the PrestoPRIME archive context, to represent the rights owned by the content holder of a specific archival item. In such case new contractual events, or any other event implying modification of rights, can be used as input for keeping the rights record up-to-date.

---

[1] Project PrestoPRIME - FP7-ICT-2007-3 231161: http://www.prestoprime.eu

[2] Deontic logic is the field of logic that is concerned with obligation, permission, and related concepts.

A5.4 Brief tutorial

General deontic expression model

The MCO deontic expression model is just a generalisation of the MVCO permission model illustrated in Figure A5.1, where the class Permission can be replaced by Obligation or Prohibition, with the object property "permitsAction" replaced with "obligatesAction" or "forbidsAction" respectively.

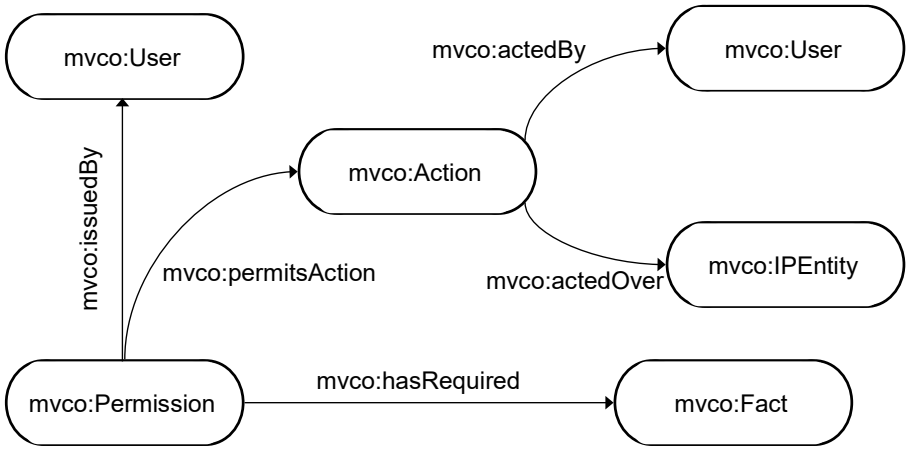


Figure A5.1: Diagram representing the Permission model of Media Value Chain Ontology

The Permission, issued by a User playing the licensor role, grant to another User playing the licensee role, the right to act an action over an intellectual property entity. The Permission to be valid requires a number (from 0 to unbounded) of facts to be true. This is the way to define conditions.

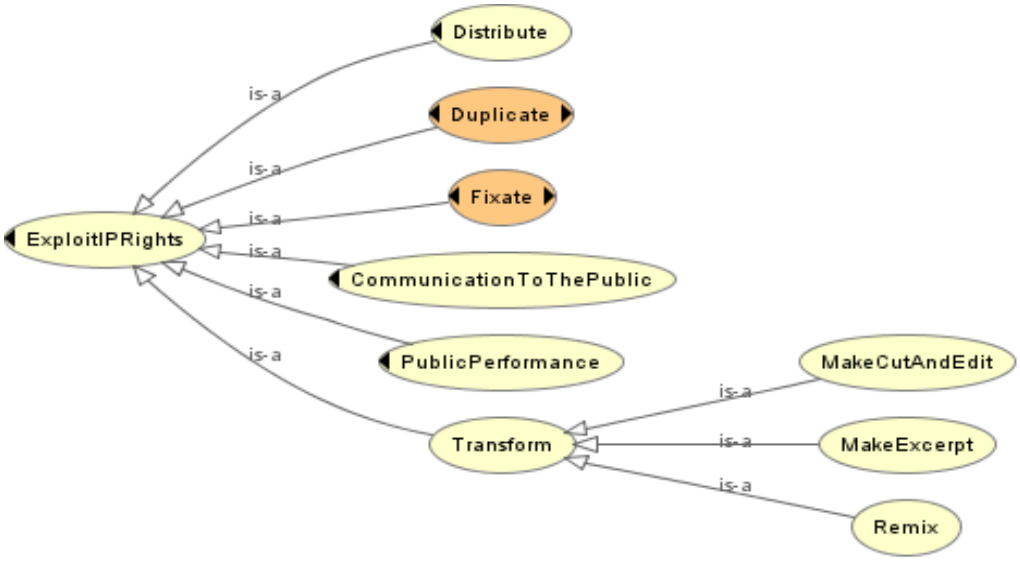


Figure A5.2: Hierarchy of Actions defined in Media Contract Ontology as subclasses of ExploitIPRights

MCO Action hierarchy

In addition to the number of Actions which were already defined in MVCO, MCO specifies a hierarchy of Actions, depicted in the diagram of Figure A5.2, above, clearly reflecting the

exploitations rights as defined by the common legal framework about the protection of intellectual property.

MCO Fact Composition and Fact hierarchy

Complex conditions can be expressed by means of fact composition with logical operators implemented by specific facts: FactIntersection (implements AND); FactUnion (implements OR); FactNegation (implements NOT[1]). Notice that the default operator for single Permission is AND (all the required facts must be true).

As an example, it is possible to represent the condition that a permitted broadcast occurs either by Satellite or Terrestrial means.

A hierarchy of Facts, given in Figure A5.3, overleaf, is defined in order to model the conditions found in real contracts, according to various contexts and dimensions.

[1] The FactNegation is redundant, as the negation can be represented by using a negative object property

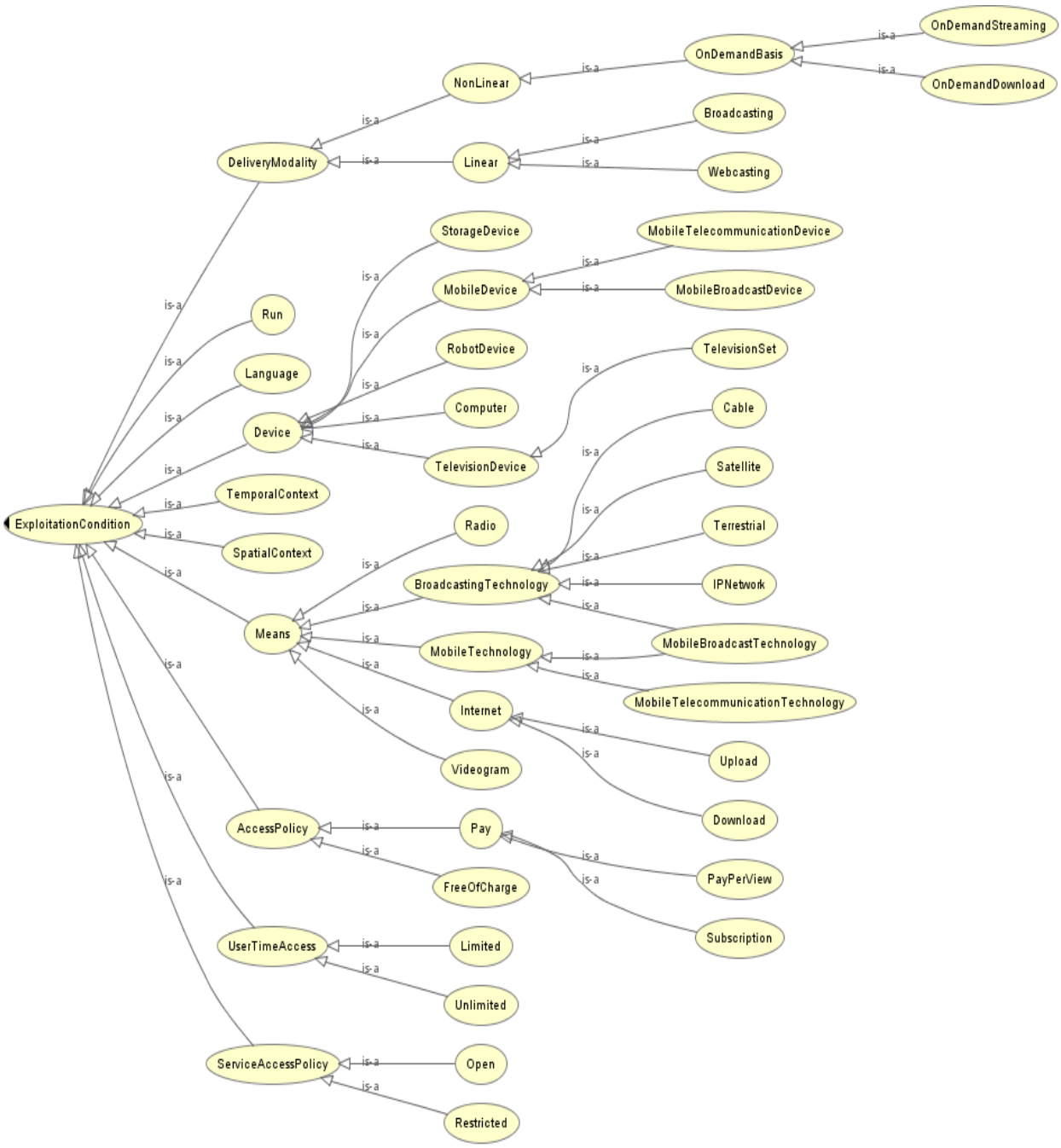


Figure A5.3: Hierarchy of Facts defined in Media Contract Ontology as subclasses of ExploitationCondition

Data Properties

While for a number of Facts, the complete expression of a condition is simply given by its relationship with Permission (for example “FreeOfCharge”); in other cases the conditions are fully specified only by the assignment of data properties. In the examples below the use of data properties will be shown for expressing “TemporalContext” (dates of license periods), “SpatialContext” (countries), “Language”, other temporal conditions, such as delays and validity, and eventually the conditions on “Runs”.

Also attributes of the Permission itself, such as the flags of exclusivity or sublicense, are expressed by means of data properties.

Graphical representation

The figures of the following examples make use of diagrams, automatically obtained from the OWL documents, which faithfully represent their content.

The following conventions apply:

- Ellipses represent individual of Classes. The class and individual IRIs (only suffix) are printed.
- The arcs represent Object Properties. Negative Object Properties are represented in red.
- The gray boxes, linked to ellipses, represent Data Properties, with their values (unless too large).

Simple example

A simple example is given in Figure A5.4, where it's depicted the case of a permission for RAI to act a "communication to the public" provided that: it is over free of charge service, with the delivery modality of broadcasting (linear with many simultaneous viewers), before the end of February 2014, within the territories of Italy, the Republic of San Marino and the Vatican City. Besides the Permission is "exclusive" (no permission compatible with this one is granted to anyone else).

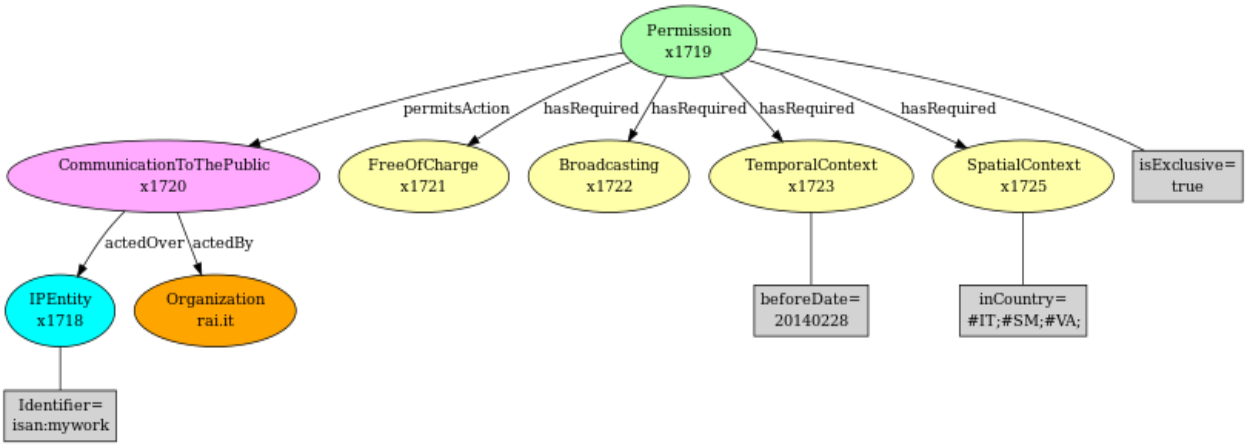


Figure A5.4: Diagram of a simple Permission in MCO with four conditions

More complex example

A more complex example is given in Figure A5.5 and Figure A5.6, depicted separately for aim of clarity. The peculiarities of the first permission, otherwise similar that above, are:

- There is a condition on the number of runs, bounded to four, and it is specified that within an interval of validity (seven days in the example), any repetition can be considered as the same run. This kind of condition is found in real contracts
- A condition on means is expressed by an OR, so that it doesn't matter if the broadcast happens on Satellite or Terrestrial means. The number of runs has to be computed on all the permitted means.
- whenever the permitted action starts, a particular Fact "ActionStarted" will get true and stay so for a validity time (15 days in the example). This is related to the second permission of Figure A5.6.

The second permission, represented in Figure A5.6, requires the condition of non linear delivery modality, more specifically by "on demand streaming", through the Internet. However the real peculiarity is that the Permission requires the "ActionStarted" fact, already depicted in Figure A5.5, to be true. This case is also found in real contracts. The second permission actually

depends on the occurrence of an action permitted by another/main permission, granted with some different conditions.

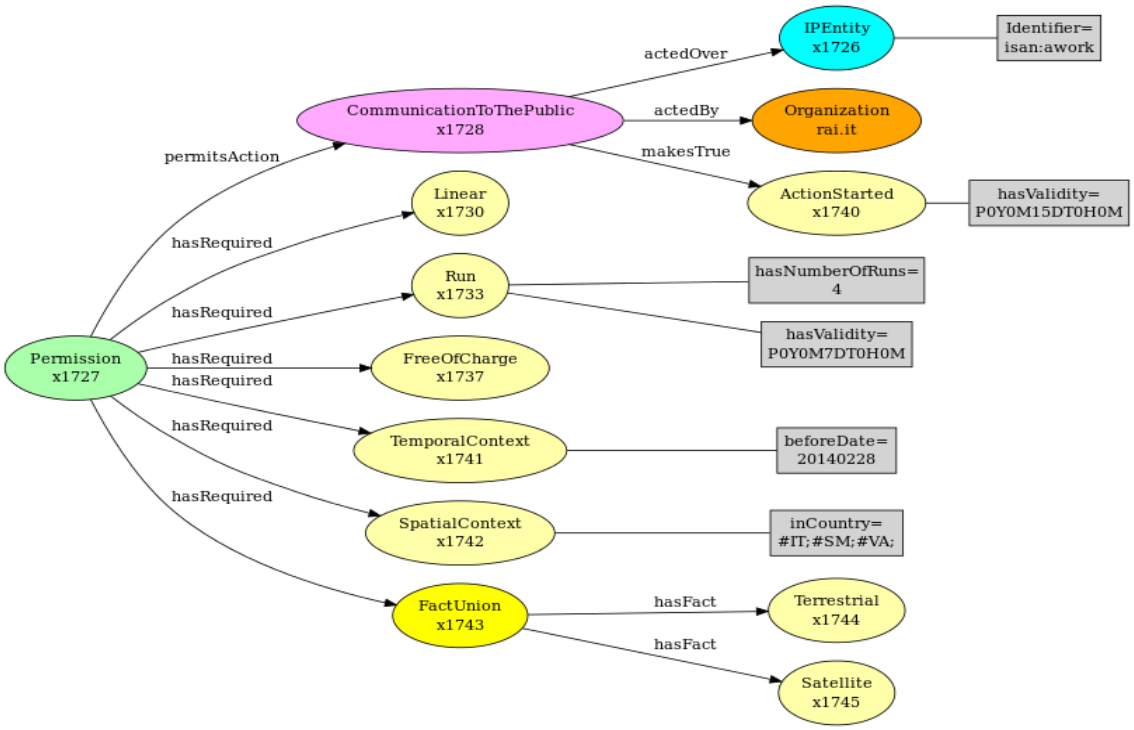


Figure A5.5: Diagram representing the "main permission" of MCO complex example

The case of the example is sometimes named "Catchup TV". A narrative text might explain that only when the main permission is exploited, through Satellite or Terrestrial means, the licensee is also granted to exploit the content over a kind of "video on demand service".

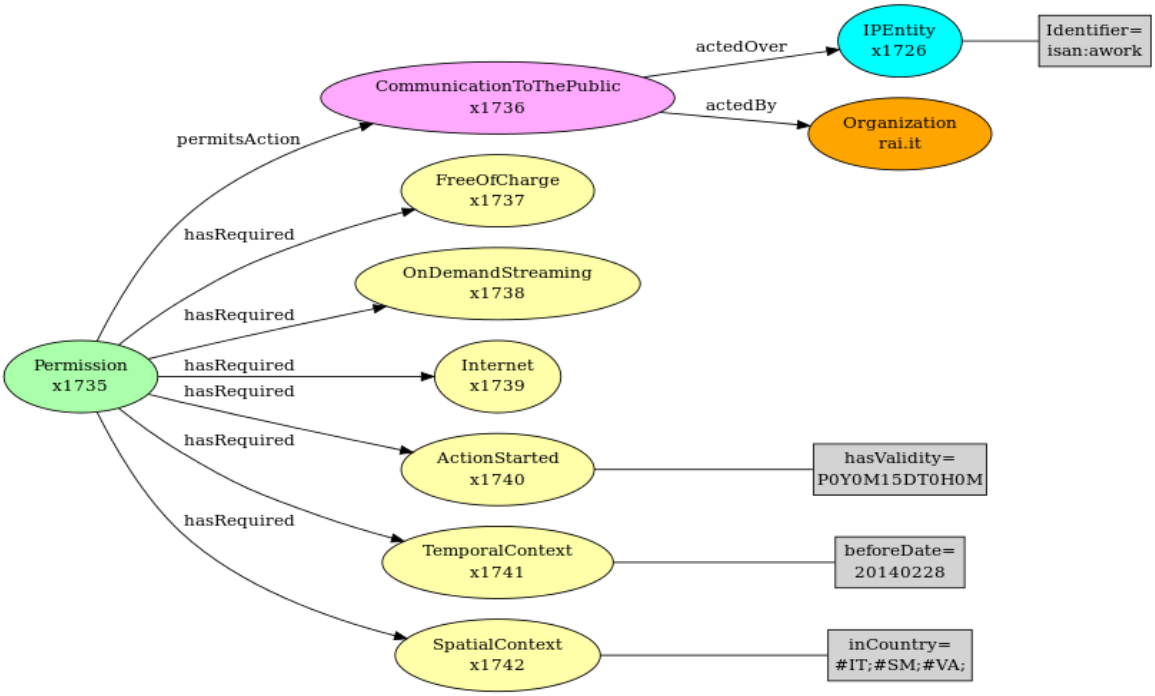


Figure A5.6: Diagram representing the "secondary permission" of MCO complex example

MCO Contract documents

The following example shows the use of MCO for a whole contract document. In this case the diagram representing a fictional simple contract between RAI and BBC is depicted in Figure A5.7. The Contract is well identified as the source of all the deontic expressions, one permission in this case. The roles of the users or organisations with respect to the contract (being a party, being a signatory) and with respect to the deontic expression (being the issuer/licensor, being the principal/licensee) are also very clear.

The objects of the contract, e.g. IPEntities, are simply the objects of the actions permitted (or obligated or forbidden) in the contract. It is possible to have a single contract dealing with multiple objects and it is also possible to have parties having multiple roles (licensor for one permission and licensee for another one).

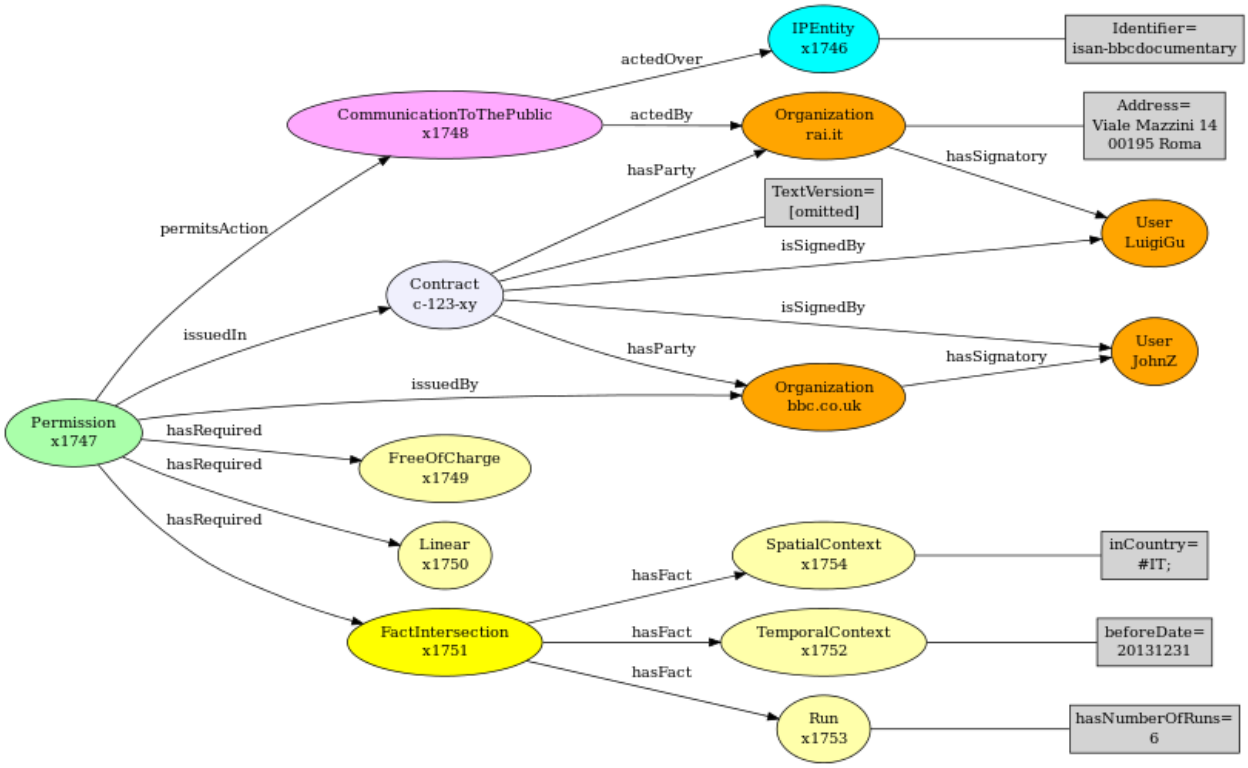


Figure A5.7: Diagram representing an example of MCO contract

Who may be interested

The broadcasters and their organisations, such as EBU, should definitely be interested, together with the companies or individuals active in television and cinema production, but also in audiovisual products in general, including fortuitous producers, such as for the so- called “user generated content”.

Audio and video archives, content distributors, providers of media services and platforms are clearly also interested. Besides, all the companies and organisations which are owner of the rights related to events which are at the origin of content, such as sports or festivals, should be considered. Eventually the providers of IT products and or services for the typologies of actors mentioned above should be concerned.

Page intentionally left blank. This document is paginated for two sided printing

Annex 6: ABC Australia

Lizbeth Moore, ABC (Moore.Lizbeth@abc.net.au)
Tris Hoyne, ABC (Hoyne.Trish@abc.net.au)

A6.1 Using Ontologies for the Development of a Data Model

To maintain business continuity and access to data from legacy television broadcast information systems ABC Television plans to undertake a project to develop a Television Broadcast Data Archive application. As part of the preparatory work for this project the ABC's metadata working group were approached to recommend on a data model that could be used to support data integration from both the legacy and current broadcast information systems.

The metadata working group had selected the EBU's Class Conceptual Data Model (CCDM) as the foundation of its common media information ontology. However, the project also needed to consider the data structures in both the current and legacy systems as well as other domain standards such as BXF, ISAN, TV Anytime and the ABC's own version of the BBC Programmes Ontology. To arrive at a recommendation that took all of these structures into consideration the metadata working group used an ontology mapping process. This process involved three steps:

Step 1 Generating ontologies from existing databases.

Step 2 Generating ontologies, where they were not available, to represent the information structures in the relevant domain standards.

Step 3 Integrating these ontologies into a single subject specific ontology for television programme broadcast information and using this to determine correspondences between classes via mappings to CCDM.

A6.2 Generating Ontologies from Databases.

This was a manual process and involved the mapping of database tables and columns to create classes, object and datatype properties in the legacy database ontologies. As a first attempt at this kind of work the direct mapping process was kept deliberately straightforward and focused on creating the Class and subClass hierarchy and properties while ignoring some of the more advanced features of OWL.

Although we did not capture the full extent of the semantics that could be expressed in the legacy databases, the process generated the basic semantics as well as providing us with some exposure to the issues that may be encountered in automating database to ontology mappings.

A6.3 Generating Ontologies from XSD.

As with step 1. this was a manual process and involved the mapping of XML schema structures to create ontologies for the BXF, ISAN and TV-Anytime standards. Again, the resulting ontologies were also simple and focused on expressing Class and subClass relationships and object and data properties rather than capturing the full set of semantics made possible in OWL.

Whereas the transformation of the database structures to OWL i.e. table to Class seemed intuitive, the translation of the XSD semantics to OWL seemed less straightforward and required a formal methodology. While a brief survey of the literature suggested a range of possible approaches (sometimes conflicting in their recommendations) we ultimately decided on a manual implementation of a subset of the XSD2OWL rules outlined in the **ReDeFer** project.

A6.4 Integration of Ontologies

This final step involved importing the generated database and standards ontologies, together with CCDM, into a single subject domain ontology and expressing the correspondences between classes using the axiom equivalentClass.

Using the Protégé ontology editing software this integrated ontology enabled the metadata working group and the technology staff working on the Television Broadcast Data Archive application to interrogate the data structures within the legacy databases, domain standards and CCDM from multiple perspectives. The resulting recommendation drew heavily on the BXF standard with relationships to the key standards and the ABC's upper level MIM ontology CCDM captured and documented for future development of this application solution. In undertaking the process the metadata working group made the first tangible relations between existing ABC data structures and CCDM as our upper level MIM ontology and gained the hands on experience in generating and working with ontologies for any future automated implementation of semantic technologies.

Annex 7: IASA-OK

Guy Maréchal – PROSIP (guy.noel.marechal@gmail.com)

IASA-OK: The International Association for Sound and Audio-visual archives has initiated a task-force for studying and promoting the applications of the semantic technologies in the archival sector. The first focus selected is the elaboration of an open interoperable format.

That pivot / axis format would be open, flexible and interoperable, suitable for constructing persistent archives, for enabling easy 360° publishing, for facilitating an effective interchange between independent systems or data bases and for empowering aggregation portals. That axis format would be based on a core semantic profile including mainly an upper-ontology (with associated upper-taxonomy/thesaurus/terminology/configuration management); a resolvable URI allocation and management protocols and hooks for domain and media oriented specific profiles; it will be autonomous in the sense that all the profiles involved in an instance of export and import would be included in the wrapped interchange data.

That IASA-OK initiative plans to collaborate with the EBU-MIM project and with the AXIS- CRM initiative of the Non Profit Association TITAN.

Page intentionally left blank. This document is paginated for two sided printing

Annex 8: MediaMixer Project

Roberto García, Universitat de Lleida, Spain (roberto.garcia@udl.cat)

Lyndon Nixon (STI Research, Austria); Vasileios Mezaris (CERTH, Greece); Benoit Huet, Raphaël Troncy (EURECOM, France); Rolf Fricke (CONDAT, Germany); Martin Dow (Acuity Unlimited, UK);

MediaMixer is an EU funded action to support organisations in enhancing their media contents to create greater value and extending reach across customers, consumers and the media value chain. MediaMixer promotes semantic technologies that enable the fragmentation of media items into distinct parts, which can be re-purposed and re-sold while managing the associated copyright.

A8.1 About Media Mixing

Media repositories happen to expose their individual media resources as atomic (complete) items. While consumers are often interested only in salient parts, which address their content, need. Media mixing is the process by which self-contained parts of media (fragments) are identified and exposed via media repository interfaces, so that consumers can access and re- use only the parts they are interested in. Media Mixing requires the application of new technologies for the creation, repurposing and reuse of media fragments across borders on the Web, which are integrated into media systems and workflows, like the one shown in Figure A8.1.

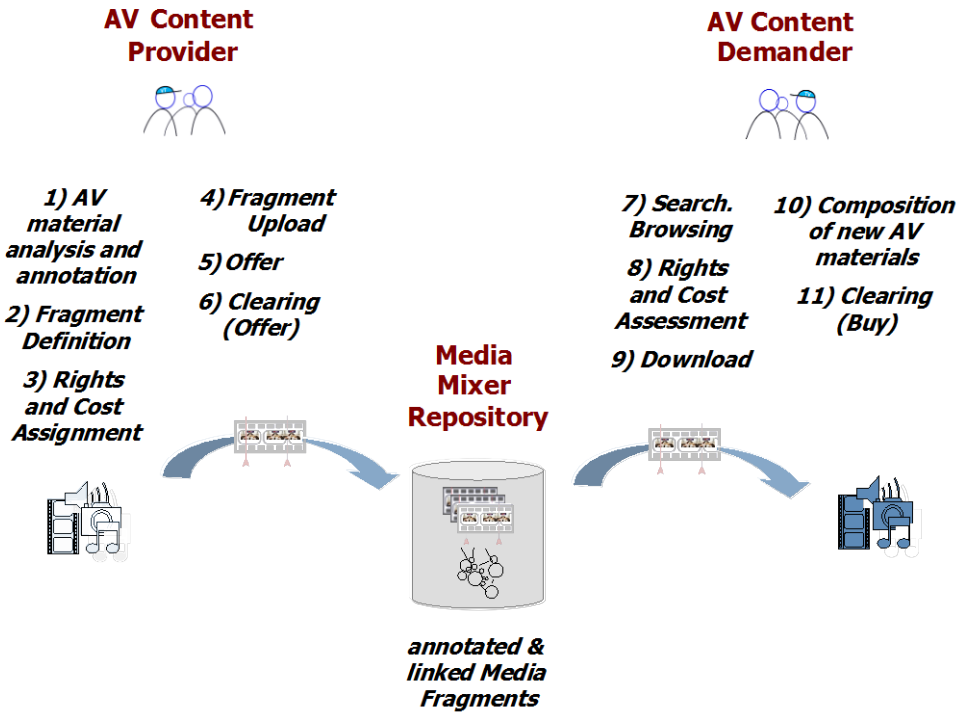


Figure A8.1: Media Fragments workflow example

A typical MediaMixer application will involve the fragmentation of the media assets (in terms of generating a fragment description), the storage of these descriptions in a repository (linked to the assets themselves), and exposing those descriptions to customers (for fragment level search and selection). Depending on the use case, rights information may be attached to fragments to control and manage the appropriate access to and re-purposing of fragments (alone and in combination).

Client side media fragment systems may enable linear pay-out of multiple fragments as a single media presentation, so that audio-visual content may be composed into a new resource, or

interactive non-linear browsing across fragments, so that media remixes are created.

MediaMixer vision is that, by enabling media owners to more flexibly and dynamically provide (sets of) media fragments to consumers, will support new usages of media repositories and help them uncover new value for their media.

A8.2 Scenario: Re-use of Media Fragments from Video Footage

As an example of media fragment re-use, in the current video production for news, commercials or magazines the editors re-use material from footage providers such as Getty Images, ITN Source, Video Clipdealer, Insertstock, iStockphoto, NBC archives and Thought Equity Motion.

They all offer complete "clips" with durations from a few seconds up to several minutes and with metadata such as source, title, time, place, persons and category (nature, technical, sports, etc.), rating and more. However, as editors manually define all metadata (a very costly process), only a small part of the available video content can be offered.

The MediaMixer project envisages exploiting a much wider range of video footage by offering fragments through automatic annotation by visual, textual and speech analysis as well as face recognition. Additional metadata is inherited for each fragment from previous or related shots, scenes or the complete video. The link to the fragment source usually allows determining crucial parameters, such as owner, price and creation time. Media fragments can be video snippets of any size from single frames or shots to several scenes.

The MediaMixer project envisages a complementary use of clip archives and media fragments retrieval: if the search in clip archives with a smaller number of clips was not successful, media fragments retrieval allows to browse through a much wider range of footage, but - due to the less complete annotations - in a more explorative way by using recommender algorithms, similarity search and personalization.

The re-use of video footage is supported by the tools offered from the MediaMixer community as well as by some other available tools, which can be employed in the user application environment to generate, retrieve and present Media Fragments. The user interfaces will be designed according to the needs of editors to re-use Media Fragments in their application environment such as news production, learning, advertisement, documentary and product presentation.

Resources

- "Let Google Index Your Media Fragments" reports how Google can now index media fragments and offer rich snippet preview for media fragments, http://eprints.soton.ac.uk/336529/1/devel2012.pdf.

A8.3 Media Fragment Creation

Starting from a video file, media fragment creation is the process of identifying different parts of this video (i.e., fragments) that each has some meaning by itself, and therefore could be re- used independently of the rest of the video.

Media fragments can be temporal, in which case we call them shots, scenes or stories, depending on how these fragments were defined and detected, or even spatiotemporal, e.g. corresponding to a specific object that appears in a video shot.

Typically, media fragment creation is achieved by applying a combination of analysis technologies to the video, which include feature extraction for video representation, feature transformation, and supervised learning as well as other machine learning techniques.

Resources

- State of the Art and Requirements for Hypervideo is a LinkedTV project document looking at current approaches to derive spatial and temporal fragments of media and related metadata about those fragments (for the Media Fragment Description), http://www.linkedtv.eu/wp/wp-content/uploads/2012/11/LinkedTV_D1.1_State-of-the-Art-and-Requirements-Analysis-for-Hypervideo.pdf
- Video of temporal fragment creation results (a LinkedTV project result), http://www.youtube.com/watch?v=fvAfIGiJGgY
- Video of spatiotemporal fragment creation results (a LinkedTV project result), http://www.youtube.com/watch?v=0IeVkXRTYu8

A8.4 Media Fragment Description

Semantic media fragment descriptions permit the connection of self-contained media fragments to the concepts (things, people, locations, events ...) they are perceived as representing.

Semantic technology is a means to describe media in a way that can be understood and processed by machines. Concepts can be unambiguously identified by URIs using Linked Data principles. Ontologies – which define permitted terms and how they relate to one another – are the basis for machine reasoning and automatic derivation of new knowledge about the media (e.g. a fragment that shows Angela Merkel is also showing the German Chancellor)

Semantic descriptions of the media can be derived from existing metadata generated in the media production process and augmented by tools provided within the media creation phase. The former case is handled by definitions of mappings from legacy metadata formats to the media fragment description format, and Media Fragment Creation tools handle the latter. Fragments are identified, and then linked to semantic descriptions, using the Media Fragments URI 1.0 (basic), a W3C Recommendation. It specifies the syntax for constructing media fragment URIs and that explains how to handle them when used over the HTTP protocol.

Resources

- Media Fragments URI 1.0 (basic), http://www.w3.org/TR/media-frags/.
- Open Annotation Model (future W3C recommendation) is promoting the use of media fragment URI for annotating any media, http://www.openannotation.org/spec/future/index.html.
- Media Fragments technology showcase reflects current implementations of the Media Fragments URI 1.0 spec, http://www.w3.org/2008/WebVideo/Fragments/wiki/Showcase.
- Multimedia Broadcasting and eCulture, by Lyndon Nixon, Stamatia Dasiopoulou, Jean-Pierre Evain, Eero Hyvönen, Ioannis Kompatsiaris, Raphael Troncy. Chapter in the book "Handbook of Semantic Web Technologies" has a section on semantic media vocabularies. Springer, 2011, ISBN 978-3-540-92912-3. http://www.eurecom.fr/~troncy/Publications/Troncy-swhandbook11.pdf.
- W3C Media Ontology provides for a common subset of media properties across typical metadata vocabularies and a mapping between them, http://www.w3.org/TR/mediaont-10/.

A8.5 Media Fragment Rights

Media Mixer rights management for fragments addresses current barriers to the exploitation of media fragments due to the lack of unambiguous, automatable and interoperable ways to represent and manage rights. This is already needed to make rights management scale to a web of media, but it is even more critical when dealing with fragments.

Furthermore, fragments enable a potentially enormous re-use market so rights management must be able to deal with issues beyond media access control. The objective is to provide means to manage copyright through the whole value chain, from creators to consumers without missing re-users, producers, distributors or any other participant.

Rights management for fragments is based on a rich, semantic model for the expression of content rights and licensing terms over media assets in terms of their fragments, to underpin the media discovery and usage negotiation process and complement the final acceptance of (legally binding) license agreements. The semantic model is based on ontologies, like the Copyright Ontology for the copyright domain. The creation model part of this ontology is shown in Figure A8.2.

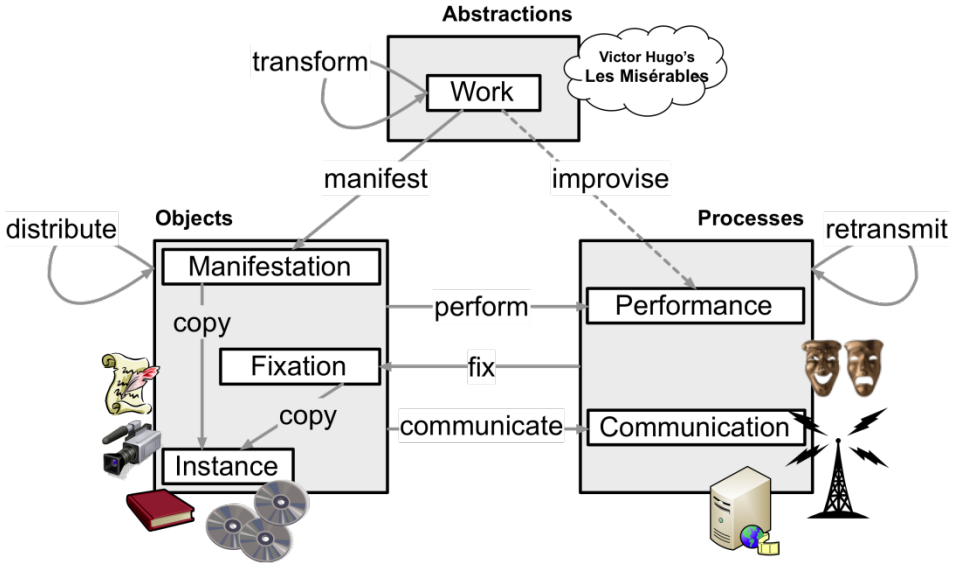


Figure A8.2: Copyright Ontology creation model

Media Mixer will provide best practices and guidelines around rights management for the maximal possible automation of functionalities for rights negotiation. Based on an explicit and interoperable semantic representation for the communication of rights, Media Mixer facilitates assessing the reusability of a given media asset fragment and eases bringing content onto this flourishing market. For instance, by interoperating with DDEX, ODRL or MPEG-21 rights data.

Resources

- García, R.: "A Semantic Web Approach to Digital Rights Management". VDM Verlag Dr. Müller, 2010. ISBN 978-3-639-15740-6. http://www.amazon.com/Semantic-Approach-Digital-Rights-Management/dp/3639157400.
- Copyright Ontology: an ontology that formalises copyright concepts and allows modelling and reasoning about copyright licenses, http://rhizomik.net/ontologies/copyrightonto/.

A8.6 Media Fragment Management

Media asset management (MAM) addresses the coordination of tasks that manage operations over media lifecycles, such as ingestion, annotation, cataloguing, storage, preservation, discovery, access control, retrieval and distribution of digital video assets. MediaMixer plans to promote the adaptation of existing media asset management systems to handle media fragments by building a reference system that extends an existing open source media asset repository.

MediaMixer envisages management of media fragments so that robust, persistent identifiers are maintained to meet needs around both HTTP URIs for web connectivity, as well as industry identifiers. Metadata for web resources addressable by HTTP URIs are an integral part of the

semantic web machinery, and MediaMixer envisages that semantic metadata for media fragments are an integral part of robust media asset management, future-proofing media assets for the web.

At the same time, industry identifiers are an integral part of industry schemes under development that address future rights trading and compliance requirements. MediaMixer envisages that media fragment management can create actionable policies with asset management systems that utilise semantic rights metadata, enabling deployment of ontologies such as the Copyright Ontology. This would assist automation of access control and compliance checking, and help simplify communication of terms of use to end-users.

Resources

- Fedora Commons Open source digital content repository framework; MediaMixer plans to use media implementation to promote general approaches to MAM adaptation, http://fedoracommons.org.
- Avalon Media System is an example of the implementation of Fedora Commons for audio-visual media, http://www.avalonmediasystem.org.
- Reference Model for an Open Archival Information System, Recommended Practice; This 2012 update is of interest to management of web media fragments within media archives; it further addresses distinctions between syntax and semantics and the use of access rights as envisaged for MediaMixer, http://public.ccsds.org/publications/archive/650x0m2.pdf.

Page intentionally left blank. This document is paginated for two sided printing

Annex 9: YLE's Meta-API: Improving the Findability of Web Content with Semantic Tagging

Pia Virtanen, Kim Viljanen, Mikael Hindsberg / The Finnish Broadcasting Company YLE
(mailto:firstname.lastname@yle.fi)

Abstract

A current challenge of the Finnish Broadcasting Company YLE is how to improve the online findability of YLE's TV, radio and web content on the web and in a multitude of different mobile and other applications. This article presents a metadata and Linked Data based approach where "semantic tagging" is used for describing the content with concepts from selected vocabularies (subject topics, people, places, events, etc) and for interlinking content to related other content (e.g. a TV programme to an article). In the article we present our automatic and manual tools for tagging content, examples on end-user applications using the metadata, the Meta-API for publishing our tagging data, the concept vocabularies and ontologies we are using for tagging the content, and our process for harmonizing (mapping) the vocabularies.

A9.1 Introduction

At the Finnish Broadcasting Company YLE[1] TV, radio and web content is produced in organizational and system silos to achieve efficiency by specialization in specific subject areas. The disadvantage of this silo structure is that when the content is published on the web, the content is not interlinked as much as it could be because the creators of individual pieces of content can not know well enough what else YLE has produced about the same topic recently or during the past decades.

For our audience this means that finding all relevant content that YLE has published about a specific topic is difficult. The customers should not be aware of YLE's different internal boundaries; what matters for them is to find interesting and relevant content no matter who or in which production system or in which format it has been produced.

A9.2 Tagging the Content with Vocabularies and Content Objects

To address this findability problem of related content we use the Linked Data[2] and semantic tagging[3] approaches in the following two ways:

- concept vocabulary based tagging of content and
- *content to content* tagging.

With semantic tagging we mean that a content item (e.g. an article or TV programme) is described by using concepts from selected vocabularies (including ontologies and classifications) which express the essential meaning or other aspects of the content. For example, a TV programme about the Eurovision Song Contest winner of year 2015 could be tagged with concepts such as "The Eurovision Song Contest" (the event), "Year 2015" and "Heroes" (the song).

[1] http://www.yle.fi

[2] http://en.wikipedia.org/wiki/Linked_data

[3] http://www.hedden-information.com/SemanticTagging.pdf

The concept "semantic" refers to the practice of Semantic Web[1] and Linked Data, where global identifiers (URIs) are used for identifying the concepts unambiguously compared to using human-readable literal terms. For example, identifiers intend to help avoiding mixing identical words with different meanings (homonyms), such as "bank" - a financial institution vs. a river bank - or persons and places with the same name (namesake). In addition, using web-wide global identifiers that are shared between organizations make it easier to interlink, share and combine content from different sources, which is the (distant future) vision of the Semantic Web.

YLE's content covers all aspects of life - for example global, national and local politics, sports, art, history, music, entertainment, recipes, and so on - and also many different cultures, such as the Finnish, the Finnish Swede, and the Sami. To describe the subject areas of all potential content in detail requires a very broad collection of concepts.

Currently we use primarily following publicly available vocabularies for describing the *subject matter* (Dublin core: subject[2]) of YLE's content:

- KOKO[3]: a collection of Finnish core ontologies merged together into one ontology, maintained by the Finnish national ontology service Finto at the Finnish National Library. KOKO contains mainly general concepts like *love, war* or *climate change* but also some individuals (proper nouns), such as *Finland*, *the European Union* and *Jupiter*.
- Freebase[4] by Google. Freebase is a knowledge base that contains individuals such as individual persons, organizations and places.

In addition we also use the following nonpublic vocabulary for describing the *subject* of YLE's content:

- Leiki: a proprietary ontology by the Finnish company Leiki[5], originally based on the subject codes of the International Press Telecommunications Council IPTC.
  Leiki contains both general concepts and proper nouns.

In addition to external vocabularies we also have some internal vocabularies, such as subject vocabularies created for the YLE Archives and different web services, and the YLE content classification. The internal vocabularies exist both due to legacy needs and because some concepts relevant for YLE might not be relevant outside YLE.

In addition to vocabularies, we also allow tagging the content items with any other YLE content item. This is used to represent the fact that two content items are (intellectually) related (Dublin core: relation[6]). That is, a member of the audience interested in one of the content items is most probably also interested in the related other content item.

[1] http://en.wikipedia.org/wiki/Semantic_Web

[2] http://purl.org/dc/elements/1.1/subject

[3] http://finto.fi/koko/en/

[4] http://www.freebase.com/

[5] http://www.leiki.com/

[6] http://purl.org/dc/elements/1.1/relation

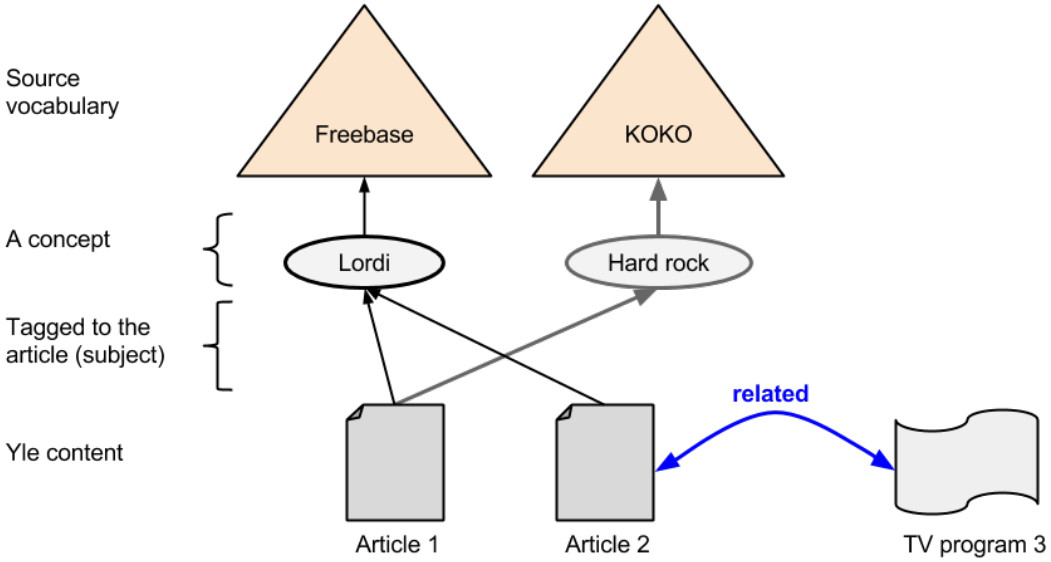


Figure A9.1: The subject of articles 1 and 2 is both "Lordi" (from Freebase) and "Hard rock" (from KOKO). In addition, article 2 is directly related to a specific TV programme

A general architectural principle of the Metadata project is to prefer using external vocabularies as sources for concepts instead of creating and maintaining a YLE's own vocabulary, if such vocabularies exists. By using external vocabularies we avoid dublicated work of creating and maintaining the vocabularies and improve the odds for making our content interoperable with other organizations that use the same vocabularies for describing their content.

However, to avoid dependencies on external services, we import all used external concepts to our internal Meta-API database: When new concepts are used from the source vocabularies, we add them into the Meta-API database, give them an YLE identifier (YLEId) and also save relevant other information about the concept such as labels and references to external source vocabularies. Some concepts also contain the type of the referred object, e.g. *person, organization* or *place*.

By keeping the original identifiers we will later have the possibility to connect our content with other content outside of YLE, if those other content providers use these same, or compatible, vocabularies for their content description. In addition, we have the possibility to enrich our Meta-API with additional relations and other information from the source vocabularies, when needed.

A9.3 Creating the Tags Manually and Automatically

The subject matter of a piece of content can be tagged both manually and automatically. In the following we present some of the ways YLE uses.

Many organizational units[1] of YLE use Drupal as their web publishing platform. To tag web content in these Drupal content management systems (CMS), YLE has created a tagging module called YILD (YLE Integrator for Linked Data)[2] for Drupal 7.

With the help of YILD, the user can search for suitable concepts from connected vocabularies and add these concepts as tags to the article at hand. (See image below.)

The module can easily be connected to any vocabulary that provides an open REST API for searching the vocabulary. YLE uses YILD with Freebase and KOKO, but YILD has also been tested against Wikipedia, Wikidata, DBpedia and Geonames. In addition, the Semantic web company built an

[1] The Swedish YLE unit, the Creative Content unit, the Media unit

[2] https://www.drupal.org/project/yild

extension module to Yield to connect it to their PoolParty Semantic Suite[1], thus making a proof of concept for connecting YILD to SPARQL endpoints, further extending its scope. YLE is planning to extend the modules compatibility to Drupal 8.

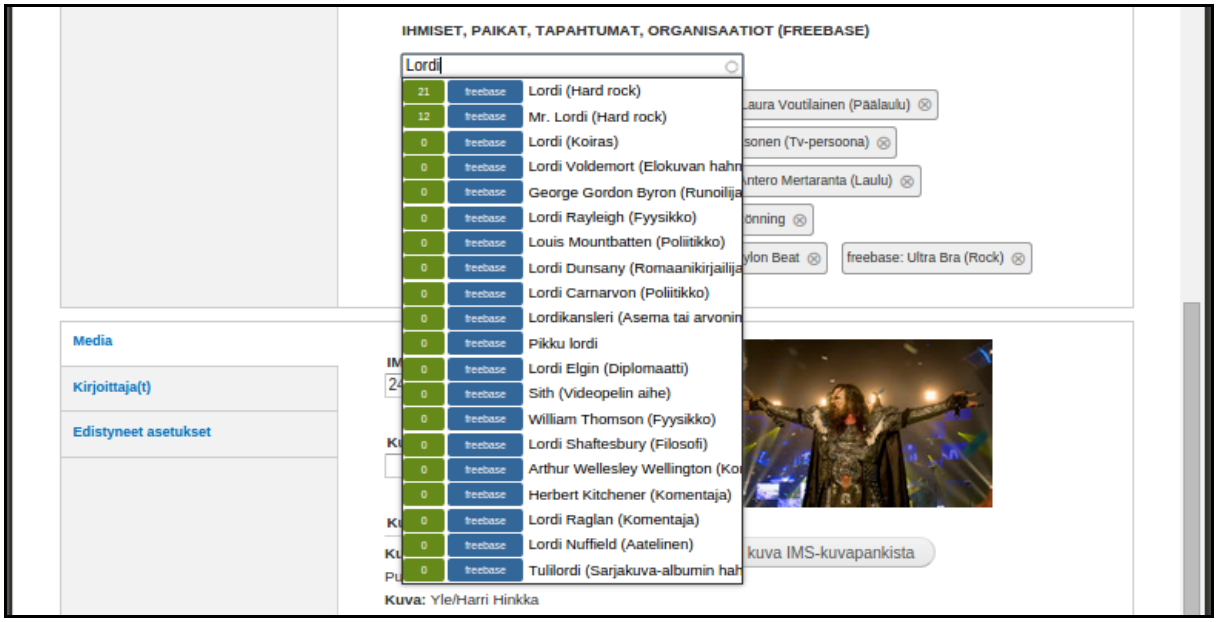


Figure A9.2: YILD in action

The user is annotating an article about the Finnish Eurovision Song Contest winner “Lordi”. When the user starts to write the keyword “Lordi”, the autocomplete search results in matching terms from Freebase such as the band “Lordi”, the artist “Mr. Lordi”, and “Lord Voldemort” (the arch enemy of Harry Potter).

We have also added a content tagging functionality to YLE’s TV and radio media asset management (MAM) system (Avid Interplay) which makes it possible to tag TV programmes. The functionality is similar to YILD but the implementation differs. Currently the MAM tagging functionality is tested with archive materials.

For automatic tagging of textual content YLE uses the Leiki system[2] which automatically analyzes the textual content of each article and produces tags that match the content. This is currently used for tagging YLE’s News and Sport content for their news application Uutisvahti (“News Watch”[3]). The CMS (Escenic) sends the article to the Leiki system which analyzes automatically the textual content and produces the matching Leiki tags. This happens fully automatically and does not require input from the journalist.

To improve the quality of automatic tagging, YLE is planning to bring the possibility to manually edit the automatically created tags. The machine would give the suggestions of tags and the journalist would then choose the suitable ones and complete the relevant missing ones manually.

We are also investigating in cooperation with research organizations the possibility to automatically tag video and audio content. These projects include speech and image recognition as well as exploitation of subtitles.

[1] https://www.drupal.org/project/yild_poolparty

[2] http://www.leiki.com/technology

[3] http://qvik.fi/en/portfolios/yle-uutisvahti/

A9.4 Meta-API: Making the Tagging Data Available for Applications

YLE's online service architecture is based on a collection of different APIs[1]. For example, the Programme's-API[2] contains scheduling and other metadata about the TV and radio programmes; the Articles-API contain web articles.

For accessing the tagging data, we have created the Meta-API. When web content is tagged either automatically or manually, the information is stored in the various backend systems, and then made available to end-user applications via the Meta-API. The Meta-API contains the *subject* and *relation* tags originating from three different systems: the Drupal, Escenic[3], and YLE's MAM.

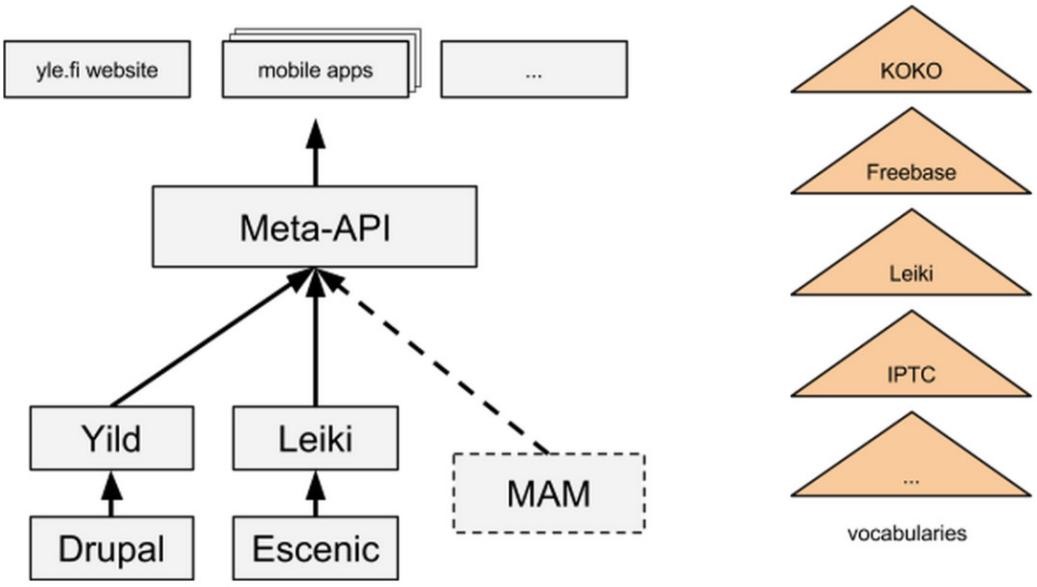


Figure A9.3: Meta-API construction

The tagging data created in different back-end systems manually or automatically is transfered to Meta-API and made available to the end-user applications.

The Meta-API provides a HTTP REST JSON interface for querying relations between concepts and content item (e.g. an article or programme), for searching concepts with string search, and for finding content-to-content relations. Below are two examples of the responses of the Meta-API.

[1] http://yle.fi/aihe/artikkeli/2015/05/06/road-open-apis

[2] https://tech.ebu.ch/docs/events/metadata14/YLE-Programs-API--Kim-Viljanen-at-EBU-MDN-2014---2014-06-03.pdf

[3] used by YLE's News, Sport and radio channels

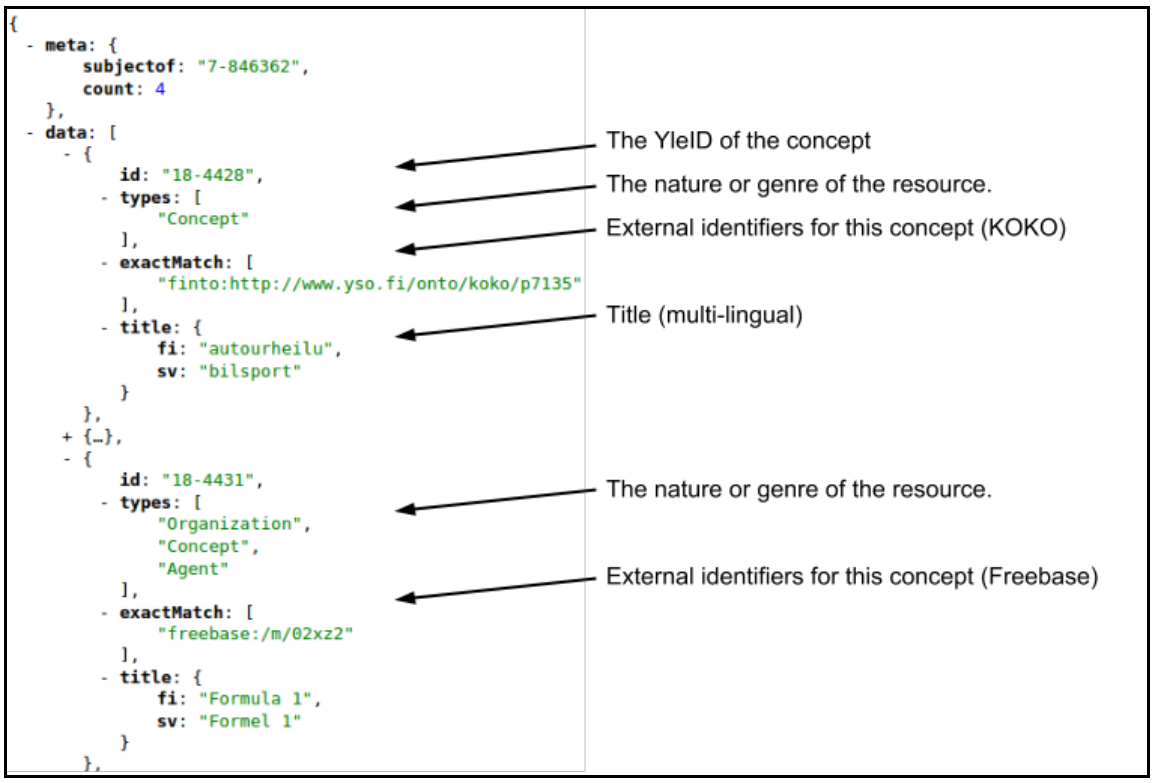


Figure A9.4: First example of a Meta-API output

The list of all concepts tagged to article identified with "7-846362". The Meta-API can also be queried for all content items (e.g. articles) that contain a specific concept.

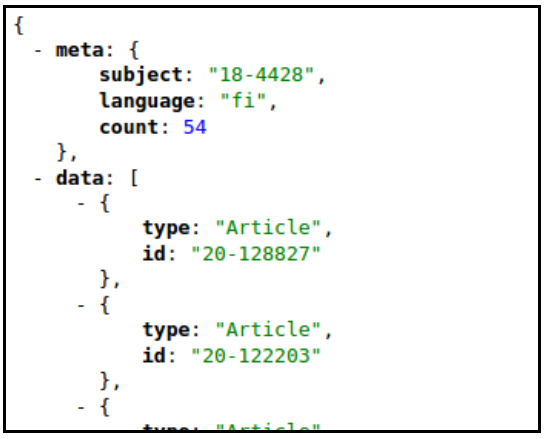


Figure A9.5: Second example of a Meta-API output

The list of all articles that are tagged with concept identified with "18-4428". (To access the articles, the respective identifier matches a specific article in the Articles-API.)

At the time of writing, Meta-API contains metadata about 1086552 articles and 211497 TV and radio programmes. For articles, we have both subject and relation (content to content) tags, for programmes almost only relation (content to content) tags.

Currently, the Meta-API contains 15408 concepts from KOKO, 20840 concepts from Freebase, 40945 concepts from Leiki and 3162 YLE internal concepts, giving a total amount of over 80 thousand concepts.

A9.5 Front-end Applications

With the help of the Meta-API, we are able to link content automatically on the web. We can already cross boundaries between organizational units (for example, the Swedish Content unit and The Creative Content unit), between the two main languages of our services, Finnish and Swedish, and between different systems (Drupal and Avid Interplay).

The following figures depict one way YLE uses the concept and relation tags on YLE's website.

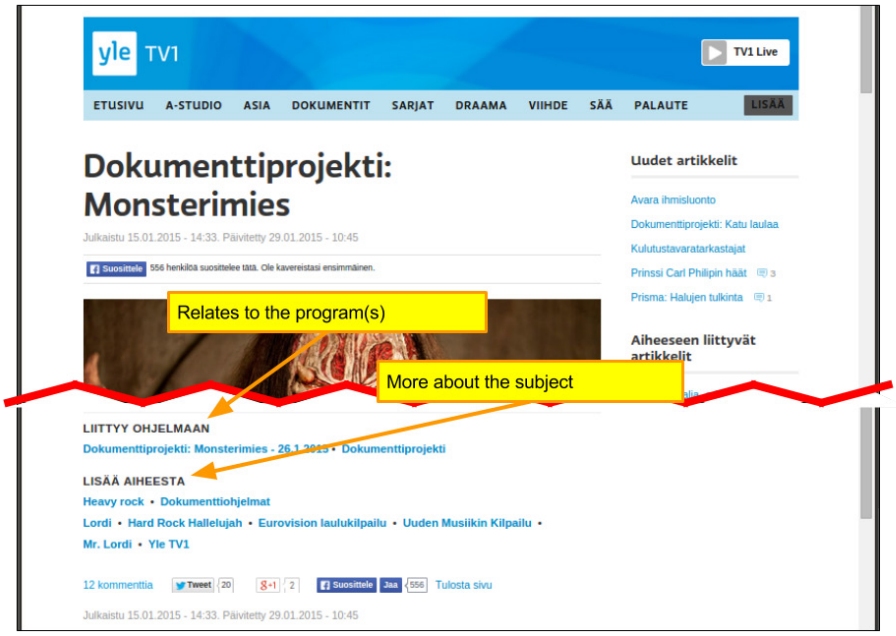


Figure A9.6: Concept and relation tags on YLE's website

This shows an article about the artist "Mr. Lordi". The article is linked to related programmes and to subject pages. By clicking on the tag "Mr. Lordi", the user sees a list of all YLE's content about this topic (see next figure).

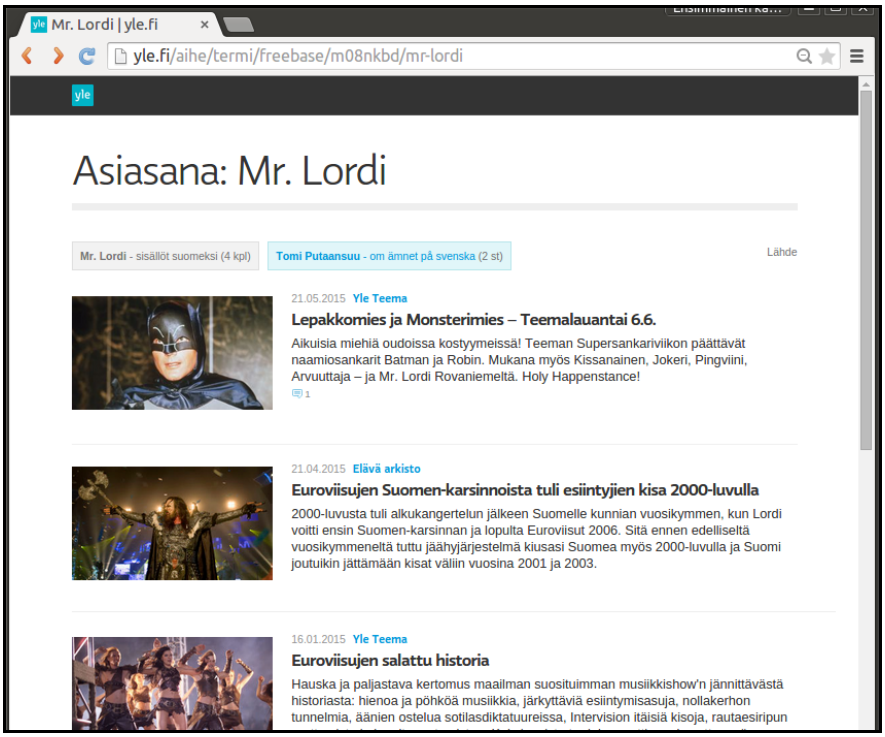


Figure A9.7: Latest articles with the subject "Mr. Lordi" in Finnish

On the top of the page (blue panel) there is a link to the corresponding list of Swedish articles about the same topic.

On many of YLE's webpages, the tagging metadata is embedded inside the HTML pages by using the RDFa[1] extension to HTML to improve the SEO of the pages. In addition to serve external web search engines, the tags are also indexed by YLE's own site search haku.yle.fi[2]and used to provide improved search and filtering functionalities. For example, the end-user of haku.yle.fi can search content that contain a certain concept (such as a name of a person) or filtering content based on the tags.

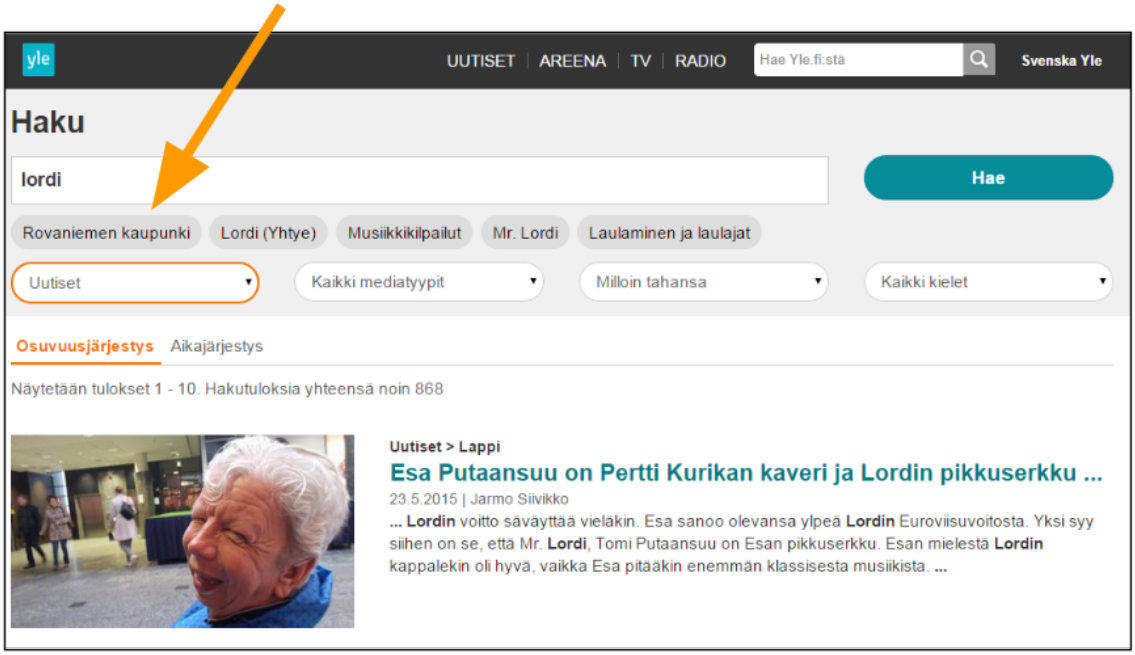


Figure A9.8. YLE's site search

The search result of keyword search "lordi" can be filtered with typically occurring tags, such as "Rovaniemen kaupunki" (City of Rovaniemi - the hometown of the Lordi band).

The YLE mobile news "Uutisvahti" application utilizes the (automatically created) Leiki tags that provide the end-user with a content profile functionality allowing his/her selection of topics (concepts) that are of interest and those that should be avoided. By providing feedback to the system, the application learns what the individual user is interested in and displays more relevant news for each user.

[1] https://en.wikipedia.org/wiki/RDFa

[2] http://haku.yle.fi

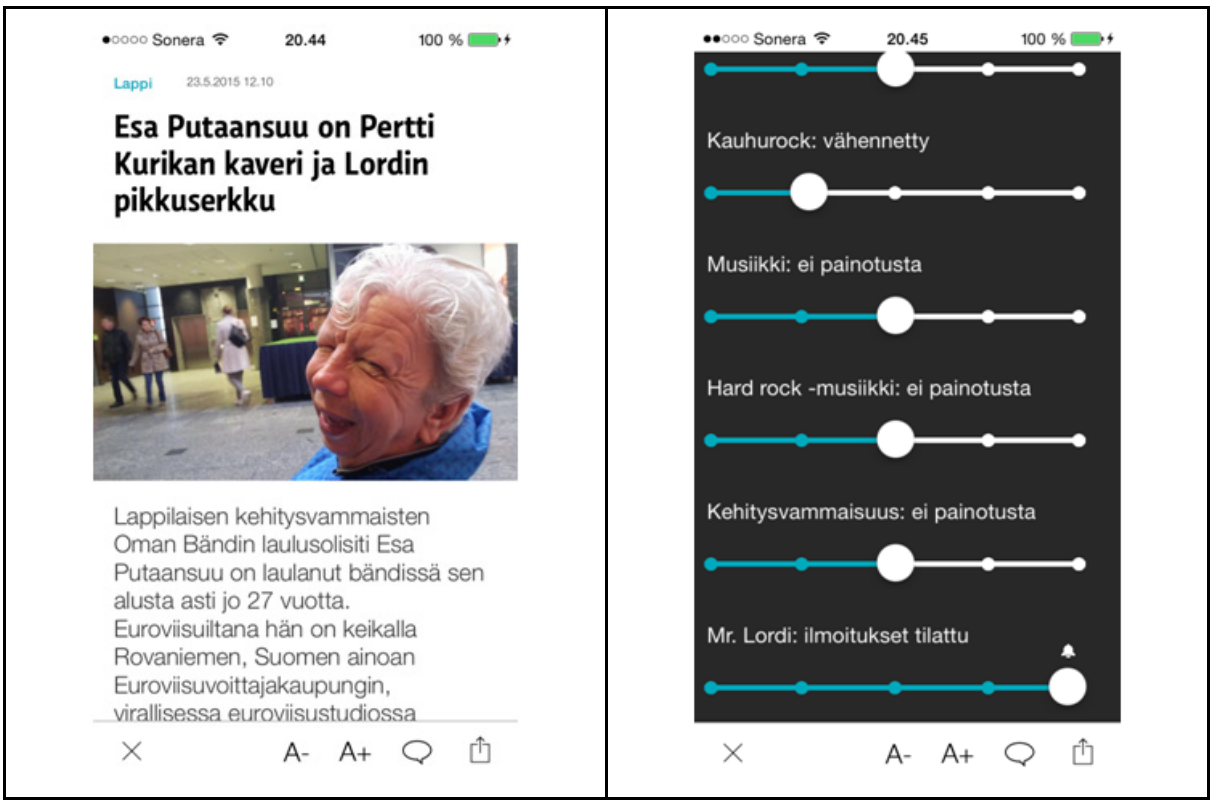


Figure A9.9. A screenshot of YLE's personalized News mobile application "Uutisvahti"

Each article (on the left) ends with a list (on the right) of Leiki tags found in the article. Using the sliders, the end-user can configure the weight of each topic in future. For example, if the user would like to completely avoid any articles about "Hard rock", the user would drag the Hard rock slider to the far left.

The Swedish Content unit also uses Meta-API for recommending audio or video content which is originally embedded in other articles. For those recommendations across different media types they traverse the Meta-API graph, assuming that a video or an audio clip inherits the same tags as the article it is embedded in. In practise this have given rather accurate recommendations considering the relatively weak semantic connection.

A9.6 Mapping the Vocabularies

As described earlier, YLE uses currently more than one source vocabulary for tagging the content. The vocabularies are partially overlapping which means that the same concept can separately be found in more than one vocabulary. For example, "Switzerland" is found in all our main source vocabularies KOKO, Freebase and Leiki, or "Mr. Lordi" in Freebase and Leiki.

The consequence of these unlinked concepts with same meaning is that finding and bringing together all YLE's content about a certain topic is difficult. This created a hinder for providing YLE's audience a combined view of all YLE's news content produced by different units and tagged with different vocabularies (Leiki, KOKO, Freebase), we started to "map" the equivalent concepts between KOKO, Freebase and Leiki in spring 2015. The process we used for mapping the equivalent concepts is outlined in the following.

First we automatically generated a list of potentially matching concepts by comparing the keywords of Leiki, KOKO and Freebase. The output was a long list (see following table) where next to each Leiki concept the matching KOKO and/or Freebase concept was listed (if such matches were found). The automatic mapping was made with string matching of the concept labels (the name of the concept) by searching for complete and partial matches.

Then the domain experts from YLE processed manually the automatically generated list by confirming each suggested mapping to be either true or false. A concept was mapped if they were considered to be "enough" equivalent from the audience's perspective. For example, Leiki concept "Giorgi Armani" (Leiki ID: 1091) and Freebase concept "Giorgi Armani" (Freebase ID: 024htv) are both referring to the same person (as judged by the domain expert), the concepts were marked to be mapped.

Table A9.1: Example of an automatically generated mapping table

Leiki		KOKO match				MAP?	Freebase match				MAP?
ID	keyword	ID	keyword	Broader concept	match	Leiki= KOKO	ID	keyword	Notable name	match	Leiki= Freebase
1091	Giorgio Armani						freebase: 024htv	Giorgio Armani		1	y
1840	Kobe Bryant						freebase: 01kmd4	Kobe Bryant	Koripallo	1	y
3109	Provinssi rock	koko: p2719	Provinssi-rock	Rockfesti vaalit	0	y	freebase: 02q8klt	Provinssi-rock	Musiikki-festivaali	0	y
3946	Tieto (Tieto Enator)	koko: p34386	Tieto	Yhteiskun nalliset tuotokset	0	n	freebase: 02jcc	Tietoteoria	Tutkimusala	0	n
24855	Rintaliivit	yso: p25544	Rintaliivit	Alusvaat teet	1	y	freebase: 01gmv2	Rintaliivit	Vaate	1	y
48558	Juliste	yso: p18786	Julisteet	julkaisut	0	y	freebase: 01n5jq	Juliste	Keräilyluokka	1	y
122751	Sotilas helikopteri	koko: p64649	sotilas helikopterit	helikopterit	0	y	freebase: 09ct_	Helikopteri	Keksintö	0	n
253042	Listat (musiikki)	koko: p32273	Listat	rakennus tarvikkeet	0	n	freebase: 017cc0	Vihreä liitto	Poliittinen puolue	0	n
288211	Mr. Lordi						freebase: 08nkbd	Mr. Lordi	Hard rock	1	y
393715	Sillat (rautatie verkko)	koko: p39849	rautatie-sillat		0	y	freebase: 0dr89x	Hiljaiset sillat	Draama	0	n
457945	British Tabloids						freebase: 064j1jl	British Tabloids	Henkilö	1	n

In the above example[1] of the automatically generated mapping table where the domain expert has filled the values to the "map?" columns (y=yes, n=no). If two keyword strings matches completely, the value of the "match" column is 1 and for partial matches 0.

To help the domain expert in making the mapping decisions, the automatically generated list also contains additional information about the suggested concepts, such as the broader concept (e.g. the "helicopter" is the broader concept of "military helicopters") and the Freebase's notable name information (e.g. the person "Kobe Bryant" is known for being a basketball player). If the domain expert required even more information to be able to disambiguate the meaning of a certain concept, the user interfaces or APIs of the respective vocabularies were used.

In some cases the automatic table generator missed some concepts because the labels of the concepts were not matching, such as synonyms and spelling variants. In those cases, the domain experts added manually new concepts to the mapping table.

[1]To abbreviate the example, the following namespaces are used in the figure to shorten the concept identifiers: koko=http://www.yso.fi/onto/koko/, yso=http://www.yso.fi/onto/yso/ (part of KOKO), freebase=http://freebase.com/m/

Finally, the mapping table was loaded into the Meta-API to merge the mapped concepts.


```
meta: {
    q: "mr_lordi",
    language: "fi",
    count: 2
},
data: [
  - {
        id: "18-48943",
      - types: [
            "Concept"
        ],
      - exactMatch: [
            "leiki:focus100k_na_288211"
        ],
        contentHits: 35,
      - title: {
            fi: "Mr. Lordi",
            sv: "Mr. Lordi"
        }
    },
  - {
        id: "18-23881",
      - types: [
            "Concept",
            "Person",
            "Agent",
            "MusicArtist"
        ],
      - exactMatch: [
            "freebase:/m/08nkbd"
        ],
        contentHits: 6,
      - title: {
            fi: "Mr. Lordi",
            sv: "Tomi Putaansuu"
        }
    }
]
```

```
meta: {
    q: "mr_lordi",
    language: "fi",
    count: 1
},
data: [
  - {
        id: "18-27156",
      - types: [
            "Concept",
            "Person",
            "Agent",
            "MusicArtist"
        ],
      - exactMatch: [
            "leiki:focus100k_na_288211",
            "freebase:/m/08nkbd"
        ],
        contentHits: 12,
      - title: {
            fi: "Mr. Lordi",
            sv: "Tomi Putaansuu"
        }
    }
]
```


Figure A9.11: The output of Meta-API when searching for "Mr. Lordi"

On the left, "Mr. Lordi" is found multiple times. After the mapping of concepts, "Mr. Lordi" is represented as a single concept on the right. (Note that the contentHits on the right is wrong due to incomplete data in the testing environment. Also the original concept ids are missing.)

At the moment, 30531 concepts are linked to each other (Leiki 13614, KOKO 8825, Freebase 7946 and 146 from internal vocabularies). There are still hundreds of concepts which could be mapped to each other (especially KOKO and Freebase concepts with no equivalent in Leiki), and as new concepts are added to Meta-API, additional mapping is required. Methods and tools for continuous mapping will be needed.

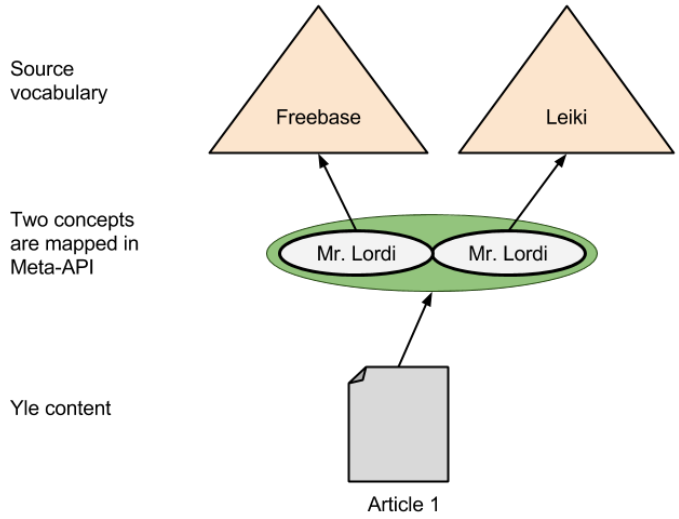


Figure A9.12: linking of concepts

Two concepts, "Mr. Lordi" (Freebase) and "Mr. Lordi" (Leiki), are mapped. When the concepts are

mapped, they act as single concepts both in end-user applications and YLE's internal tagging and other applications.

A9.7 Manual Linking of Cross-Media Content

Many of YLE's TV and radio programmes are enriched with articles and other web content to extend the viewing/listening experience and to promote programmes. However YLE's web content is produced and managed with completely different systems than TV and radio content, which makes it difficult to interlink the content.

To support cross-media publications, it is also possible to define *related* links by tagging content with some other content item. The link to a web content object and a TV or radio content object is created in the web content management system (Drupal or Escenic). After this, the link can be represented as a hyperlink for example next to the TV or radio programme in the VOD service.

The relatively simple functionality of creating a link between two web pages turns out to be rather complicated because at YLE the web content and the TV content (and the radio content) are all maintained in completely different systems. Also the content is displayed using separate front-ends. For example, YLE's web articles are displayed with a completely different front-end systems than the ondemand video-on-demand / audio-on-demand service. (See the Figure A9.below.)

This cross-silo interlinking functionality has turned out to be a very important promotional tool for directing the audience between programmes and related web content.

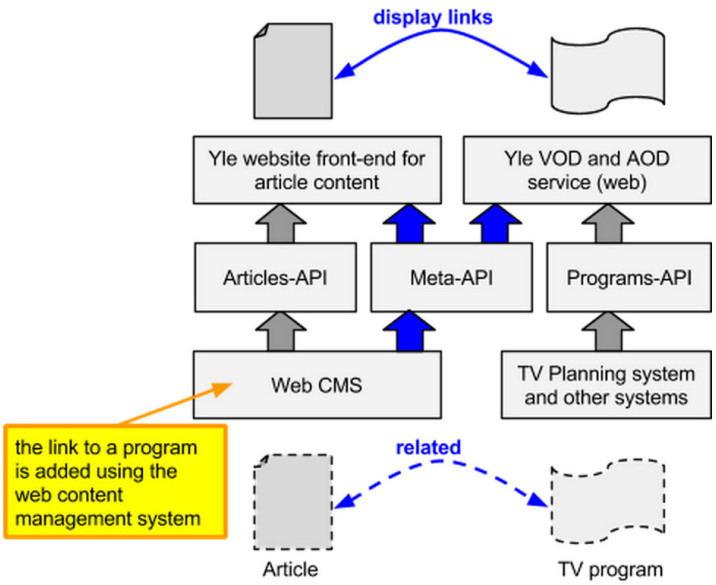


Figure A9.13: Relation between an article and a TV

The relation between an article and a TV programme is defined in the back-end systems by tagging the TV programme to the article in the CMS. The link data is made available via the Meta-API to the front-end applications which display the link.

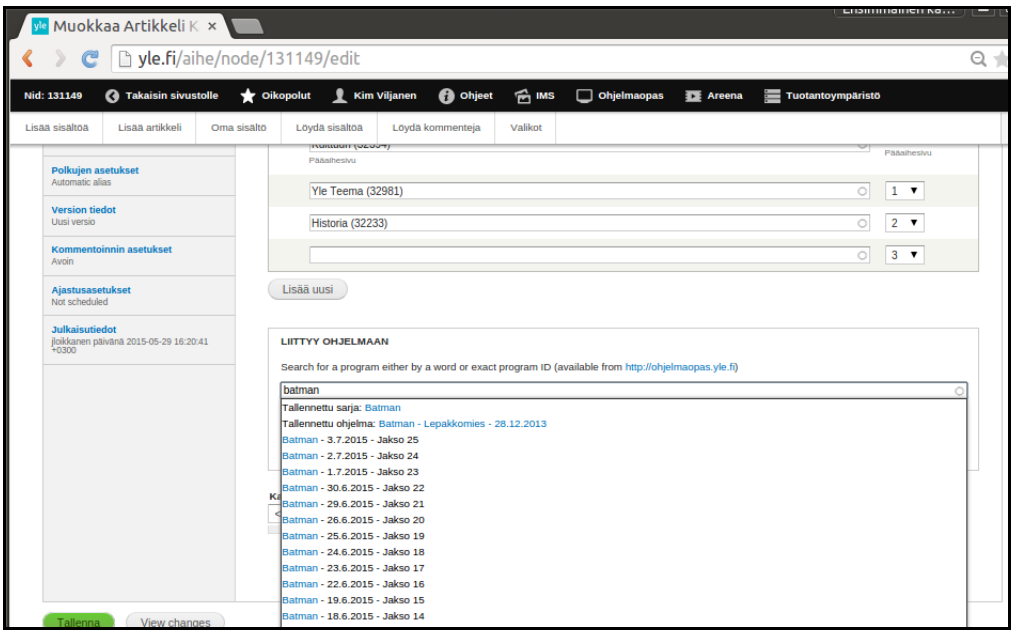


Figure A9.14: Interlinking in YLE's Drupal based CMS

In YLE's Drupal based CMS it is possible to interlink an article with any TV or radio programme. The linking is done by searching for the name of the programme (e.g. "Batman") and the result is shown as an autocomplete list of matching programme series and episodes. The user then selects one or several programmes that the article is related to.

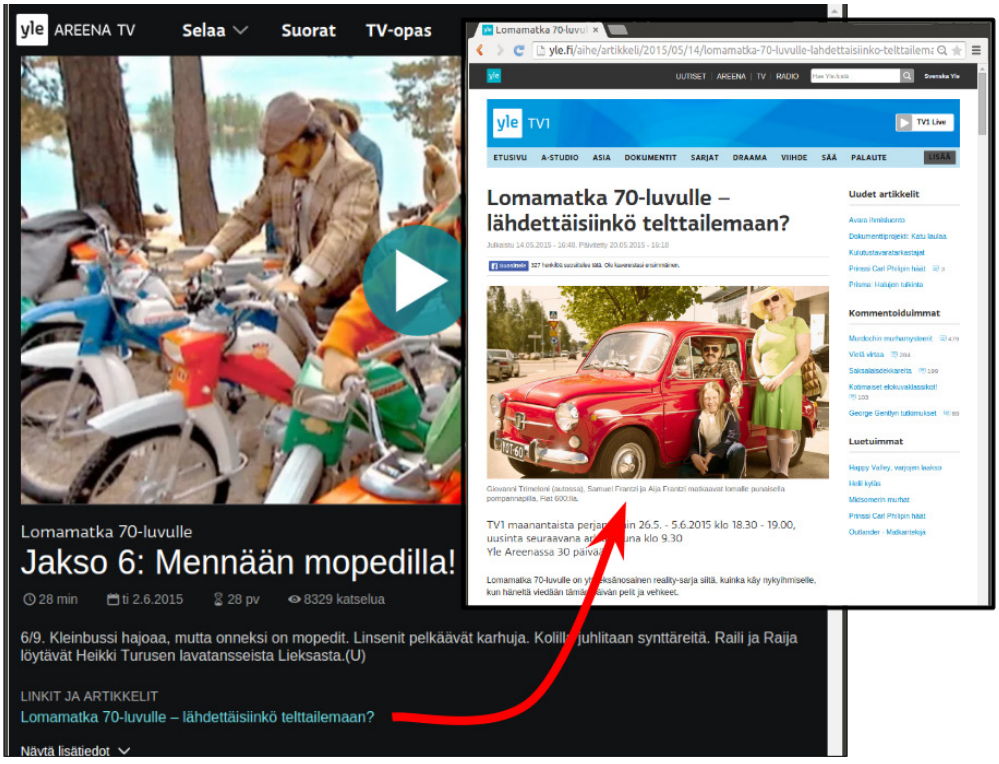


Figure A9.15: On-demand programme linking to a relevant article

The programme is linked in YLE's on demand TV and radio service "Areena" (at left) to the relevant article (at right). This link is based on the link entered in the web content management system when writing an article about this programme.

A9.8 Lessons learned and future work

The metadata creation process and the different applications described in this annex are in daily production use at YLE. Over the past two years Meta-API has been extended with new content, vocabularies and integrated to new front-end and backend systems. The vocabulary in Meta-API is already quite comprehensive, with over 80 thousand concepts, and it is growing as new concepts are required by creators to describe their content.

The demand for a company-wide tagging and relational metadata service is growing inside YLE, because this improves the visibility and the interoperability of all YLE's content. Also, by using the company-wide system, individual service developers avoid duplicated work with the Meta-API.

One of the big questions to be addressed is if we also want to start tagging TV and radio programmes. This would make it possible to bring audio and video automatically to the same context with articles. To test this, we have experimented with tagging archived content.

To decide how to focus our scarce development resources, we needed more user data; how customers consume our content, what kind of problems they have with findability, what kinds of presentation of our content they would like to have, etc. We now have good data to answer these questions and next we need to analyze it further and plan the next steps.

Improving the quality of metadata is not only a technical development; it simultaneously involves both training and process development. We instruct journalists in the process of tagging content and we make teaching materials (videos etc.) to improve the quality of content description.

In Dec. 2014 Google announced it will close Freebase and migrate its data to Wikidata. We shall probably follow suite; the final decision will be taken when details are made known by Google.

Although the (KOKO, Freebase, Leiki) source vocabularies contain hundreds of thousands of concepts, they can never contain all the concepts that are needed to describe all YLE's content. For example, historical persons and events relevant to the Finnish audience may not be relevant for the global audience. Some specific terms or subject areas might be politically difficult, which means that YLE must be able to make journalistic decisions regarding which terms YLE uses. In addition, some terms might only be relevant to YLE itself (e.g. YLE's classifications for content). So we need to technically enable the creation of new concepts directly into Meta-API.

We also plan to enrich our data with information from other source vocabularies to include labels in English, more and new concept types and further ontological relations between concepts. At the moment the concepts are not directly linked together, but indirectly via content tagged with the same concept or in the source vocabulary.

There are still vocabularies at YLE that might be added to the Meta-API in future. For example, YLE Archive (the "YLE Elävä arkisto / YLE Arkivet" service) and Knowledge & Learning (YLE Oppiminen) still have their own vocabularies.

YLE has recently started to open up its APIs[1] and the Meta-API is going to be opened in future, thus enabling third party developers to make services that improve the findability of our content.

Acknowledgements

We thank Sami Mattila for his comments and suggestions. We also thank all developers and other people that have participated in the creation of Meta-API, YILD, and the other systems and end-user applications described in the annex.

[1] http://yle.fi/aihe/artikkeli/2015/05/06/road-open-apis, http://developer.yle.fi/tutorials.html