NoTube

Networks and ontologies for the transformation and unification of broadcasting and the Internet

FP7 – 231761

D1.4 Vocabulary alignment service of NoTube vocabularies v.2

1.4 Coordinator: Dan Brickley (VUA)

Original Coordinator (D1.3): Ronald Siebes (VUA)
With (v1.3) contributions from:
C. van Aart, D. Brickley, B. Schopman (VUA)
M. Riethmayer (IRT)
Neil Benn, S. Dietze, (OU)
Jaesuk Ahn (KT)

Quality Assessor: Ronald Siebes
Quality Controller: Libby Miller

Document Identifier:	NoTube/2011/D1.4
Version:	1.0
Date:	31/10/2011
State:	Final
Distribution:	PU

EXECUTIVE SUMMARY

This revision of the WP1 deliverable is, following advice received in the 2011 annual review, a very minimalistic update to the previous 1.3 edition. Rather than producing a new and lengthy deliverable document we have focussed on establishing the sustainability of this work, and planning for "life after the project". These developments are summarised in a new leading section of the document (section 1.). Readers familiar with the earlier report will find the remainder of the document largely unchanged from the previous version, with the exception of minor fixes. Readers unfamiliar with D1.3 can treat this document as its replacement; there is no need to read both.

Workpackage WP1 of NoTube is concerned with the 'background datasets' that inform and inter-link all of our data-driven work. Throughout the project, we have emphasised the importance of adopting and promoting widely-used Web standards, rather than proprietary and ad-hoc vocabularies. As NoTube draws to a close, this strategy is demonstrating its worth. For much of the data aggregated and integrated in WP1, life-after-the-project is assured already, since the data is already managed externally. In this deliverable we examine some additional strategies for ensuring that internally-sourced datasets remain available to the wider community after NoTube has completed. At this stage, this includes reporting on some ongoing conversations with external parties about transferring maintenance. As these options are evaluated and final decisions are made, we will notify the wider community through the public NoTube blog at http://notube.tv/. See section 1 below for full details of the post-project maintenance policy for WP1.

This document summarizes the main content of deliverable D1.4 (a prototype deliverable) which regards the implementation of alignments and making them available via the Web. Every source is available via a SPARQL endpoint, but we also created some extra services to provide facetted browsing and resolvable identifiers

We first describe the two most important finalized SKOS alignments:

- an alignment between the Dutch Cornetto Wordnet and W3C Wordnet
- alignments between many different genre vocabularies (e.g. BBC, IMDB, TVA)

After that we describe ongoing work on proposing a layered approach like FRBR as a model for choosing and relating reusable identifiers that point to several types of video descriptions.

DOCUMENT INFORMATION

IST Project Number	FP7 - 231761	**Acronym**	NoTube
Full Title	Networks and ontologies for the transformation and unification of broadcasting and the Internet		
Project URL	http://www.notube.eu/		
Document URL			
EU Project Officer	Francesco Barbato		

Deliverable	**Number**	1.4	**Title**	Vocabulary alignment service of NoTube vocabularies v.2
Work Package	**Number**	1	**Title**	NoTube Models and Semantics

Date of Delivery	**Contractual**	M 30	**Actual**	31/10/2011
Status	version 1.0		final ☑	
Nature	prototype ☑ report □ dissemination □			
Dissemination level	public ☑ consortium □			

Authors (Partner)	VUA, BBC, OU, KT, IRT			
Responsible Author	**Name**	Dan Brickley	**E-mail**	danbri@danbri.org
	Partner	VU	**Phone**	+31 6 34 03 68 02

Abstract **(for dissemination)**	This document summarizes the main content of deliverable D1.4 (a prototype deliverable) which regards the implementation of alignments and making them available via the Web. Every source is available via a SPARQL endpoint, but we also created some extra services to provide facetted browsing and resolvable identifiers We first describe the two most important finalized SKOS alignments: - an alignment between the Dutch Cornetto Wordnet and W3C Wordnet - alignments between many different genre vocabularies (e.g. BBC, IMDB, TVA) After that we describe ongoing work on proposing a layered approach like FRBR as a model for choosing and relating reusable identifiers that point to several types of video descriptions. As the second edition of this report, we additionally (in section 1) address questions of long-term, post-project maintenance for vocabularies, mappings and alignments produced and managed by NoTube WP1.
Keywords	Alignment Service, SKOS, Identifiers, Access policies, Security, Service Ontology, Layered identifier abstractions, Preservation, Maintenance.

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**
31/10/11	1.0	Dan Brickley	Revised and updated from D1.3 to form D1.4.

PROJECT CONSORTIUM INFORMATION

Participants		Contact
Vrije Universiteit Amsterdam	vrije Universiteit	Guus Schreiber Phone: +31 20 598 7739/7718 Email: schreiber@cs.vu.nl
British Broadcasting Corporation	BBC	Libby Miller Phone: +44 787 65 65 561 Email: Libby.Miller@bbc.co.uk
Pronetics	pro-netics	Marco Bruni Phone: +39 06 45472503 Email: marco.bruni@pro-netics.com
Engin Medya Hizmetleri A.S.	Engin	Ron van der Heiden Phone: +31 6 2003 2006 Email: ron@engin.tv
Institut fuer Rundfunktechnik GmbH	Institut für Rundfunktechnik IRT	Christoph Dosch Phone: +49 89 32399 349 Email: dosch@irt.de
Ontotext AD	ontotext	Atanas Kiryakov Phone: +35 928 091 565 Email: naso@sirma.bg
Open University	The Open University	Stefan Dietze Phone: +44 1908 858 217 Email: s.dietze@open.ac.uk
RAI Radiotelevisione Italiana SPA	Rai	Alberto Morello Phone: +39 011 810 31 07 Email: a.morello@rai.it
Semantic Technology Institute International	STI · INTERNATIONAL	Lyndon Nixon Phone: +43 1 23 64 002 Email: lyndon.nixon@sti2.org
Stoneroos B.V.	Stoneroos Interactive Television	Annelies Kaptein Phone: +31 35 628 47 22 Email: annelies.kaptein@stoneroos
Thomson Video Networks	THOMSON	Raoul Monnier Phone: +33 2 99 27 30 57 Email: Raoul.monnier@thomson-networks
Polymedia, SpA	Polymedia	Tullio Pirovano Phone: +39 02 25771 1 Email: tullio.pirovano@polymedia.it
KT Corporation	olleh kt	Myoung-Wan Koo Phone: +82 2 526 6347 Email: mskim@kt.co.kr

Table of Contents

Maintenance and post-project Planning

This introductory section is the main new content in D1.4, building upon the earlier version of this work published as D1.3 in 2010. In addition to the vocabularies and alignments described in the earlier edition, the project has made broad and varied use of a number of data vocabularies and related datasets. We do not attempt to catalogue and itemise those individually here. Instead, this final WP1 deliverable focuses exclusively on the broader issue of life-after-project continuation for all NoTube-related datasets.

Overview of NoTube datasets for preservation

During the working life of the project, WP1 has drawn together a variety of TV-related datasets. Wherever possible, we build upon and contribute to existing work. Sometimes we create additional works, either vocabularies, RDF versions of existing formats, or alignments and mappings that link things together. So for example at http://www.cs.vu.nl/~ronny/notube/mappings/ we collect many of the genre, format and identifier mappings developed during WP1 and by the project at large.

The question therefore arises, "what becomes of these works after NoTube completes in spring 2012?". The answer comes in several parts, corresponding to the different kinds of contribution we have made, and the opportunities that are available in the wider environment around NoTube.

Preservation strategies: pragmatism and partners

The notion of preservation can make sense in various ways. For RDF vocabularies developed within NoTube we have taken care to use domain names that are expected to survive. Most of the WP1 work uses the widely-adopted OCLC PURL system <see http://en.wikipedia.org/wiki/Persistent_Uniform_Resource_Locator>. This allows us to identify an RDF vocabulary in an indirect way, using http://purl.org/* identifiers, and have the PURL server redirect to the current home of the content. Long term management of the NoTube-related PURLs will be handled by VUA, or if this ever becomes unsustainable. Since many other VUA projects face similar issues, it is expected that NoTube PURLs will redirect to **.vu.nl/** addresses for the foreseeable future. However our adoption of PURL means that these can be moved elsewhere if needed.

Other workpackages within NoTube have declared RDF vocabularies in collaboration with others, and in some cases are using other domain names. So for example, http://xmlns.notu.be/aair/ maps the non-NoTube Activity Streams specification into RDF, and http://xmlns.notu.be/wi/ shows a simple model for describing weighted interests. In both cases, the dependency on a domain name is noted and there are conversations underway about continuing the vocabulary development work in another post-NoTube venue, such as W3C's recently announced "Community Groups" environment.

The situation of a small project-based RDF vocabulary needing to make long-term preservation plans is not unique to NoTube. During 2010 members of the NoTube team were part of an collaborative initiative between the Dublin Core Metadata Initiative and the FOAF project, to establish a precedent for simple, lightweight collaboration towards vocabulary preservation. An agreement was reached in which Dublin Core will help underwrite and monitor the long-term stability of the FOAF namespace (see http://dublincore.org/documents/dcmi-foaf/). We are exploring the extension of such models for EU project outcomes, such as the 'aair' and 'wi' vocabularies in the notu.be domain.

Other NoTube work has different natural homes. For example, a small study exploring the mapping of Archive.org identifiers to DBpedia/Wikipedia identifiers (see http://danbri.org/words/2011/02/01/658) gave rise to a modest but useful dataset that links video content on Archive.org with RDF data derived from Wikipedia entries. The data files produced currently reside in temporary hosting within NoTube-related Web sites. For their long-term survival they should be archived into Web sites that have longer-term guarantees of availability, and licensed with liberal (CC0) terms to encourage re-use and ongoing exploitation of the data. To achieve this, we expect to post these files on Archive.org, and ideally also in the W3C Semantic Web Interest Group space on www.w3.org. Additionally, for Wikipedia/DBpedia-related content it is possible to arrange for data to be loaded into the DBpedia service itself, so that it is directly available as a layer of "linked data" to all DBpedia applications. This brief example is offered here as an exemplar of our approach. Rather than offer a single monolithic approach to life-after-project we are working through on a case-by-case basis and ensuring that each piece of work is considered individually, and partnerships and collaborations identified that maximise the chance of continued post-project availability.

Another example of unusual NoTube data enrichments comes from the WP3/WP7c collaborations, where we have been combining recommender systems and mining user preference data. Specifically in WP3/WP7c we have had privileged access to user viewing / rating data from the BBC. This data is very private, and cannot ever be shared. However we have in our NoTube work been computing derived 'similarity measures' between content items in our demos, based in large part on user ratings. We are investigating the possibility of sharing highly aggregated, non-identifying derived data of this kind, and anticipate that by the close of NoTube it may be possible to share some of this as Linked Data for subsequent projects to exploit further. As above, it is likely we will use archive.org hosting for any such dataset.

We will not attempt a case by case enumeration within this report. Instead, readers are directed to the NoTube blog at http://notube.tv, as we will list each 'final home' per dataset / vocabulary / mapping in blog post, during the final weeks of the project. Finally, regarding the persistence of the blog itself, several of the project partners (and team, as individuals) have expressed commitments to maintain the blog on an ongoing basis post-project. It has become a hub for many researchers in the Web/TV research community, and as such will be a useful avenue to ensure uptake and preservation of NoTube datasets as the project moves into its final phase.

2. Rationale for the choice of our alignment efforts

Next to extending the linked data cloud, the most important reason within this project for creating alignments is to provide input for content-based recommendation algorithms.

Recommendation algorithms play an important role in the NoTube project. Take for example the use case WP7b from the technical annex on personal TV guides.

The objective of this use case is to illustrate the design and development of a Personalized TV Guide recommending to the viewer TV programs and proposing him/her additional content and services including advertising material. We intend to experiment with various types of users in a multilingual setting, with a range of context identification technologies (e.g. RFID/NFC tags, sensors), as well as with different control interfaces (e.g. traditional TV remote control, Web-enabled remote control). The project will consider only using existing consumer electronic devices with integrated RFID and NFC tags and readers, e.g. mobile phones, laptops and TV remote controls. For a high accuracy of the recommendations produced by the electronic program guide we will also experiment various recommendation algorithms developed in WP3. The TV guide content used in this use case will be enriched according to the strategies in WP4. Here we are only dealing on a metadata level and not in the TV program stream.

In order to provide a personal recommendation, some algorithms compare between users the context that describe when, how, where and/or what they have been watching. When two users are 'similar' in someway regarding the preferences on video content, the preferences could be shared between them as recommendation. For example, if Chris and Guus like the same movies, then the movies of Guus that are not seen by Chris are recommended to Chris. Traditional collaborative filtering algorithms compare only the identifiers of the content. Content-based collaborative filtering also incorporates meta-information of the content in order to differentiate between different contexts. For example, it could be that Guus likes both Science-Fiction movies and Musicals, and Chris only Science-fiction, then only the Science-Fiction movies of Guus should be recommended to Chris. For the latter type of recommendation algorithms we identified three different types of content where alignments are useful.

1. **Genres**. Genres are an interesting ingredient for content-based filtering algorithms. Like the above example, if Chris watches only movies that are classified as Science-fiction, it could be a good idea to propose some new Sci-fi movies that are broadcasted the coming week on BBC via his new TV-subscription. His old TV-subscription only had Dutch TV-channels, and the broadcasters provided the genres in Dutch. Obviously, an alignment between the BBC-genres and Dutch TV channels would make it possible to compare the historical profile of Chris with the content classification of the new broadcasts on BBC.
2. **Wordnet.** When the content of a movie, like the *plot* is described in natural language, automated Named-entity extractors like Alchemy[1] allow recommendation algorithms to compare the content based on identified concepts. For example, if Balthasar often watches programs that contain the automatically identified concept 'London', then he could be interested in more programs that contain that concept. W3C wordnet[2] is a standard conversion of Princeton WordNet to RDF/OWL. The automated entity extractors are mainly written for the English language, and therefore, when a program is described in another (less popular) language like Dutch, the tools will not be useful. An alignment with 'wordnets' in other languages will allow a precise translation between concepts.
3. **Identifiers.** Popular programs like 'the Simpsons' are broadcasted on many different channels and also available via other sources like DVDs or YouTube. Now imagine that Bob watched the same episode as Fritz, but Bob via a BBC broadcast and Fritz via a downloaded torrent. Needless to say that recommender systems need to know that the content of the two videos is the same in order to function. Also it is important to recognize that a recommender system may regard two sources with the same content still not as similar, because it could (for example) take the technical capabilities of the computer into account. For example, if the recommender system 'knows' that the machine of the user is not able to play content in DivX encoding, then a torrent containing such a file should not be recommended (even if the content itself would be very matching). Section 3 gives a description of ongoing work to model these different identification layers. The layered approach allows us to specify that two identifiers point to 'similar' content on one level but not on another level. For example, when we take the MD5 hash of the DivX file containing the episode of the Simpsons as identifier 1 and the URL to the broadcast information on the BBC of the same episode, then these two identifiers are *similar* regarding the *content*, but not regarding the *format*. Alignments that take into account these layers will make it easier for recommendation algorithms to determine the relevance of the content for the users.

The recommender algorithms developed within this project also take into account many other meta-data, for example channel-id's, actors etc. For those there are no alignments available and therefore are not in the scope of this document. However, it may well be that these other types of metadata may be as, or even more, useful. For more information about the recommendation strategies in NoTube, we like to refer to workshop article at SDOW2009 [4].

In the next sections we provide some information about the realized alignments and a model for aligning different identifiers, and how to use the online services that provide several ways access to these alignments.

[1] http://www.alchemyapi.com/
[2] http://www.w3.org/TR/wordnet-rdf/

3. Cornetto Wordnet <--> W3C Wordnet

The Cornetto project [2] is funded by the Nederlandse Taalunie in the STEVIN framework (www.taalunieversum.org/stevin). It is a collaborative project with also the VU Amsterdam as partner. The goal was to build a lexical semantic database for Dutch, covering 40K entries, including the most generic and central part of the language. Cornetto combines the structures of both the Princeton Wordnet and FrameNet for English, by combining and aligning two existing semantic resources for Dutch: the Dutch Wordnet (Vossen 1998) and the Referentie Bestand Nederlands (Martin et al 1999). To create the initial Cornetto database, the word meanings in the Referentie Bestand Nederlands (RBN) and the Dutch WordNet (DWN) have been automatically aligned. The final Cornetto Database can be divided into two subsets:

- Core Cornetto - the part that has been manually checked, edited and confirmed after an initial set of automatic alignment procedures (appr. 8,500 lexical units and 3,000 words).
- Extended Cornetto- the part of Cornetto that is not manually checked and covers approximately 89,000 words and 110,000 lexical units.

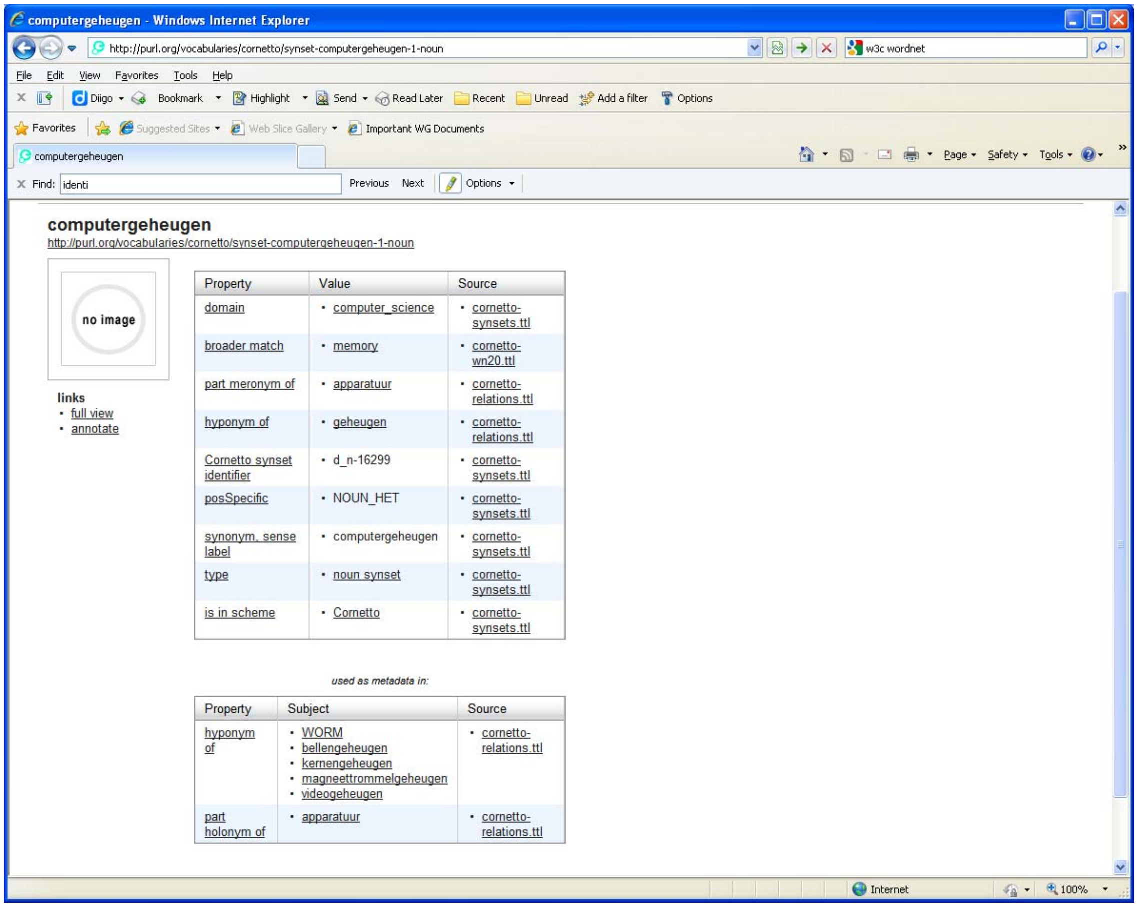
Currently we are transforming the alignments to RDF, meaning that the relations between the concepts *within* each wordnet *and* the relations *between* the wordnets are transformed. We take W3C Wordnet in RDF as the format for creating the identifiers of the Dutch Cornetto concepts.

We also have chosen to introduce identifiers for the instances of classes Synset, WordSense and Word. We use the base uri + a locally unique ID. Three kinds of entities need a URI: instances of the classes Synset, WordSense and Word. Instead of using the unique Cornetto IDs we have tried to use IDs derived from information in the source and also tried to make them human-readable, similary as the identifier construction of W3C wordnet. Because the IDs have distinct syntactic patterns, it is easy to identify the type of the resource (Synset, WordSense or Word) by examining the URI. The patterns are described in Primer to using RDF/OWL WordNet.

We use two different namespaces: one for the schema and one for the instances. This makes it possible to manage the schema separately from the instances.

For example http://purl.org/vocabularies/cornetto/synset-computergeheugen-1-noun, is an identifier for the Dutch synset word “Computergeheugen” (computer-memory), which is the first meaning of the noun in Cornetto. In order to make the identifiers permanently resolvable and to have a distinction between the resolving server and naming server, we chose to create PURL urls. We use the Cliopatria server[3] to host the created RDF alignments and the relations within the Wordnets. Figure 1 provides a screenshot of resolving http://purl.org/vocabularies/cornetto/synset-computergeheugen-1-noun.

[3] http://e-culture.multimedian.nl/software/ClioPatria.shtml


computergeheugen - Windows Internet Explorer
http://purl.org/vocabularies/cornetto/synset-computergeheugen-1-noun
w3c wordnet
File Edit View Favorites Tools Help
Diigo Bookmark Highlight Send Read Later Recent Unread Add a filter Options
Favorites Suggested Sites Web Slice Gallery Important WG Documents
computergeheugen
Page Safety Tools
Find: identi
Previous Next Options
no image
Internet
100%

computergeheugen

http://purl.org/vocabularies/cornetto/synset-computergeheugen-1-noun

links
- full view
- annotate

Property	Value	Source
domain	computer_science	cornetto-synsets.ttl
broader match	memory	cornetto-wn20.ttl
part meronym of	apparatuur	cornetto-relations.ttl
hyponym of	geheugen	cornetto-relations.ttl
Cornetto synset identifier	d_n-16299	cornetto-synsets.ttl
posSpecific	NOUN_HET	cornetto-synsets.ttl
synonym, sense label	computergeheugen	cornetto-synsets.ttl
type	noun synset	cornetto-synsets.ttl
is in scheme	Cornetto	cornetto-synsets.ttl

used as metadata in:

Property	Subject	Source
hyponym of	WORM bellengeheugen kernengeheugen magneettrommelgeheugen videogeheugen	cornetto-relations.ttl
part holonym of	apparatuur	cornetto-relations.ttl

Figure 1: Screenshot of the alignment results between the dutch word « computergeheugen » and W3C Wordnet

Below in figure 2: an example of one of the alignments in RDF. For readability, we skipped the namespace definitions.

```
<wn20s:NounSynset rdf:about="&cornetto;d_n-39464">
  <cornetto:domain rdf:resource="&cornetto;furniture"/>
  <cornetto:domain rdf:resource="&cornetto;person"/>
  <cornetto:eqNearSynonym rdf:resource="&wn20i;synset-dishwasher-noun-1"/>
  <cornetto:eqNearSynonym rdf:resource="&wn20i;synset-dishwasher-noun-2"/>
  <cornetto:hasHyperonym rdf:resource="&cornetto;d_n-32445"/>
  <cornetto:posSpecific xml:lang="en">NOUN_MASCULINE</cornetto:posSpecific>
  <skos:inScheme rdf:resource="&cornetto;"/>
  <wn20s:senseLabel xml:lang="nl">glazenspoeler</wn20s:senseLabel>
</wn20s:NounSynset>
```

4. Program Genre alignments

Given the nature of the NoTube project (integrated project), we focus on creating alignments between the data provided by the content partners in the project (RAI, BBC, KT, Engin, etc) in order to allow interoperability and adhere to the multi-lingual requirements. We created a set of manually aligned genre identifiers for most of the relevant (video) content. Due to its quality, the rich vocabularies and elaborate schemas, we decided to use TV-Anytime[4] as the standard schema in NoTube for annotating the various aspects of the programmes. TV-Anytime has an extensive genre vocabulary, which Jean-Pierre Evain also made available[5] in SKOS. Figure 3 shows an example of the 'adventure' genre with the SKOS schema in RDF/XML format:

```
<rdf:Description rdf:about="http://www.ebu.ch/metadata/ontologies/skos/ebu_ContentGenreCS.skos.xml#3.4.6.1">
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept" />
  <skos:prefLabel xml:lang="en">Adventure</skos:prefLabel>
  <skos:definition xml:lang="en">e.g. Tomb Raider</skos:definition>
  <skos:inScheme rdf:resource="http://www.ebu.ch/metadata/cs/skos/ebu_ContentGenreCS.skos.xml" />
<skos:broader>
<rdf:Description rdf:about="http://www.ebu.ch/metadata/ontologies/skos/ebu_ContentGenreCS.skos.xml#3.4.6">
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept" />
  </rdf:Description>
  </skos:broader>
...
```

Figure 3: example of a TV-Anytime program genre and its SKOS relations

Another candidate for being the schema to describe the content is the BBC Program Ontology[6]. The reason for choosing TV-Anytime is because the BBC Program Ontology does not cover many technical aspects, like broadcast formats and types of streaming and because it is has a less formal status.

At the moment we made the following alignments between the several Genre-vocabularies in SKOS[7]:

BBC	<==>	TV-Anytime
DVB-SI (German)	<==>	TV-Anytime
Engin (Turkish)	<==>	TV-Anytime
Uitzending-gemist (Dutch)	<==>	TV-Anytime
BBC	<==>	Youtube
BBC	<==>	IMDB

And we plan to have

RAI (Italian)	<==>	TV-Anytime
KT (Korean)	<==>	TV-Anytime

An example of one of the alignments in SKOS schema and RDF-Turtle format can be found in figure 4.

[4] http://www.etsi.org/website/technologies/tvanytime.aspx

[5] http://www.ebu.ch/metadata/ontologies/skos/ebu_ContentGenreCS.skos.xml

[6] http://www.bbc.co.uk/ontologies/programmes/

[7] for an up-to-date overview of the available alignments, please visit http://www.cs.vu.nl/~ronny/notube/mappings/

```
@prefix skos:<http://www.w3.org/2004/02/skos/core#>.
@prefix yt-genre:<http://purl.org/identifiers/genres/youtube.com/genres/>.
@prefix bbc-genre:<http://purl.org/identifiers/genres/bbc.co.uk/genres/>.

# program genre alingments between Youtube genres and TVA genres
# Balthasar Schopman and Ronald Siebes, VU University Amsterdam
# September 2009

yt-genre:Autos skos:broadMatch bbc-genre:factual_carsandmotors .
yt-genre:Comedy skos:closeMatch bbc-genre:comedy .
yt-genre:Comedy skos:narrowMatch bbc-genre:childrens_entertainmentandcomedy .
yt-genre:Education skos:closeMatch bbc-genre:learning .
yt-genre:Education skos:closeMatch bbc-genre:factual .
yt-genre:Education skos:narrowMatch bbc-genre:drama_historical .
yt-genre:Education skos:narrowMatch bbc-genre:factual .
yt-genre:Entertainment skos:closeMatch bbc-genre:entertainment .
yt-genre:Entertainment skos:narrowMatch bbc-genre:childrens_entertainmentandcomedy .
yt-genre:Film skos:closeMatch bbc-genre:drama .
yt-genre:Howto skos:broadMatch bbc-genre:learning .
yt-genre:Music skos:closeMatch bbc-genre:music .
yt-genre:Music skos:narrowMatch bbc-genre:childrens_music .
yt-genre:Music skos:narrowMatch bbc-genre:comedy_music .
yt-genre:News skos:closeMatch bbc-genre:news .
yt-genre:News skos:narrowMatch bbc-genre:childrens_news .
yt-genre:People skos:closeMatch bbc-genre:comedy_character .
```

Figure 4: part of program genre alignments between Youtube genres and TVA genres

The namespaces for the genre vocabularies are chosen to be on Purl.org, because we want resolvable identifiers which can be reused also after the project. The format of the URI is

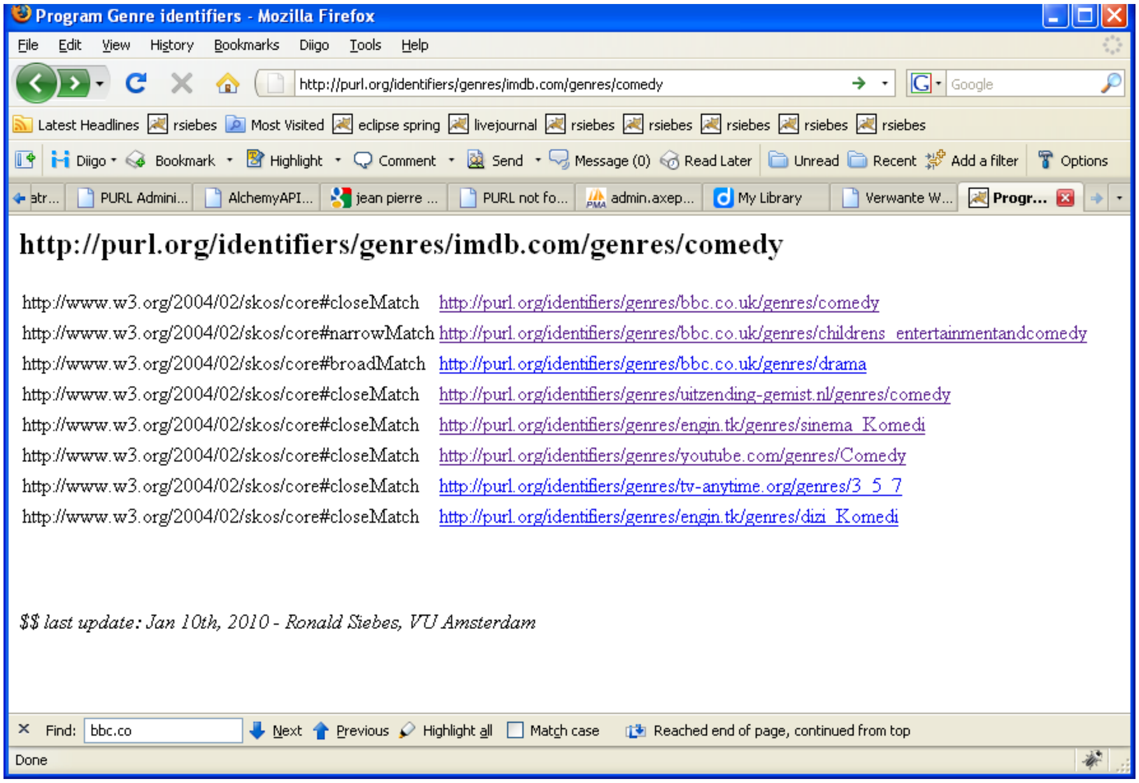
http://purl.org/identifiers/genres/*DOMAIN*/*GENRE-ID*

where DOMAIN represents the source of the identifiers and GENRE-ID the identifier of the genres. We plan to set up a generic authentication mechanism where each domain can maintain its own identifiers. For example, imdb.com is allowed to alter the identifiers for http://purl.org/identifiers/genres/imdb.com/ The advantage is that such a distributed mechanism is more scalable and gives the authority to the domain owners.

Currently we have the following domains for the different genre identifiers:

http://purl.org/identifiers/genres/bbc.co.uk -> BBC
http://purl.org/identifiers/genres/imdb.com -> IMDB
http://purl.org/identifiers/genres/engin.tv -> ENGIN
http://purl.org/identifiers/genres/youtube.com -> Youtube
http://purl.org/identifiers/genres/tv-anytime.org -> TV-Anytime
http://purl.org/identifiers/genres/uitzending-gemist.nl -> Uitzending Gemist

The purl domains currently redirect to a server hosted at the VU (domain http://eculture2.cs.vu.nl:8080/ntsrv/) where a Java Servlet running on Tomcat provides the SKOS alignment with other genre identifiers. Figure 5 shows a screenshot for http://purl.org/identifiers/genres/imdb.com/genres/comedy


Program Genre identifiers - Mozilla Firefox
File Edit View History Bookmarks Diigo Tools Help
http://purl.org/identifiers/genres/imdb.com/genres/comedy
Google
Latest Headlines rsiebes Most Visited eclipse spring livejournal rsiebes rsiebes rsiebes rsiebes rsiebes
Diigo Bookmark Highlight Comment Send Message (0) Read Later Unread Recent Add a filter Options
atr... PURL Admini... AlchemyAPI... jean pierre ... PURL not fo... admin.axep... My Library Verwante W... Progr...
Find: bbc.co Next Previous Highlight all □ Match case Reached end of page, continued from top
Done

http://purl.org/identifiers/genres/imdb.com/genres/comedy

http://www.w3.org/2004/02/skos/core#closeMatch	http://purl.org/identifiers/genres/bbc.co.uk/genres/comedy
http://www.w3.org/2004/02/skos/core#narrowMatch	http://purl.org/identifiers/genres/bbc.co.uk/genres/childrens_entertainmentandcomedy
http://www.w3.org/2004/02/skos/core#broadMatch	http://purl.org/identifiers/genres/bbc.co.uk/genres/drama
http://www.w3.org/2004/02/skos/core#closeMatch	http://purl.org/identifiers/genres/uitzending-gemist.nl/genres/comedy
http://www.w3.org/2004/02/skos/core#closeMatch	http://purl.org/identifiers/genres/engin.tk/genres/sinema_Komedi
http://www.w3.org/2004/02/skos/core#closeMatch	http://purl.org/identifiers/genres/youtube.com/genres/Comedy
http://www.w3.org/2004/02/skos/core#closeMatch	http://purl.org/identifiers/genres/tv-anytime.org/genres/3_5_7
http://www.w3.org/2004/02/skos/core#closeMatch	http://purl.org/identifiers/genres/engin.tk/genres/dizi_Komedi

$$ last update: Jan 10th, 2010 - Ronald Siebes, VU Amsterdam

Figure 5: screenshot of genre browser in SKOS

5. FRVR: layered approach for content identifier alignments

In this section we describe work in progress on making a model for the different layers of descriptors for video content.

5.1. Video context abstractions

The several case-studies in WP7 mention a rich variety of situations where people watch video content. For example, programs broadcasted on televisions, DVD's played via media-players, watching online Youtube videos, etc. As already mentioned before, the different types of recommender algorithms require specific types of contextual information. This contextual information comes from different sources, and we need a mechanism to link these different contexts in order to express that there is a relation between them. Such a mechanism basically deals with two issues:

- How to find a uniform way of choosing/creating identifiers
- How to link these identifiers and express the type of relation

These sources independently provide their own different types of annotations, and have their own expertise and incentives to do this. For example:

- Movie makers like Warner Bros, Colombia etc give content descriptions like the *plot*, *keywords*, *actors* etc. in order to gain interest of the public
- Broadcasters provide the schedule and channel information in order for people to tune in

- Torrent creators provide technical information about the type of encoding in order not to disappoint people after downloading it
- Users may (or their players like Boxee) provide the timestamps of watching it and feedback in order to get good recommendations
- Etc

The contexts vary in which a NoTube user can experience video content

Think about the following examples:

- Watching the uncut version of an episode of the "adams family", remastered 40 years later, having Italian voice-over, downloaded via a torrent in DivX format projected on a beamer.
- Watching a repetition of the latest "BBC ONE late news", on "the BBC world channel", on the hotel tv, in a hotel room in Dubai.
- Watching a trailer of the "pigs in space" movie on a Real Audio stream, on a mirror site of the Warner Bros trailer portal, on a opensource player in a Google Android phone.

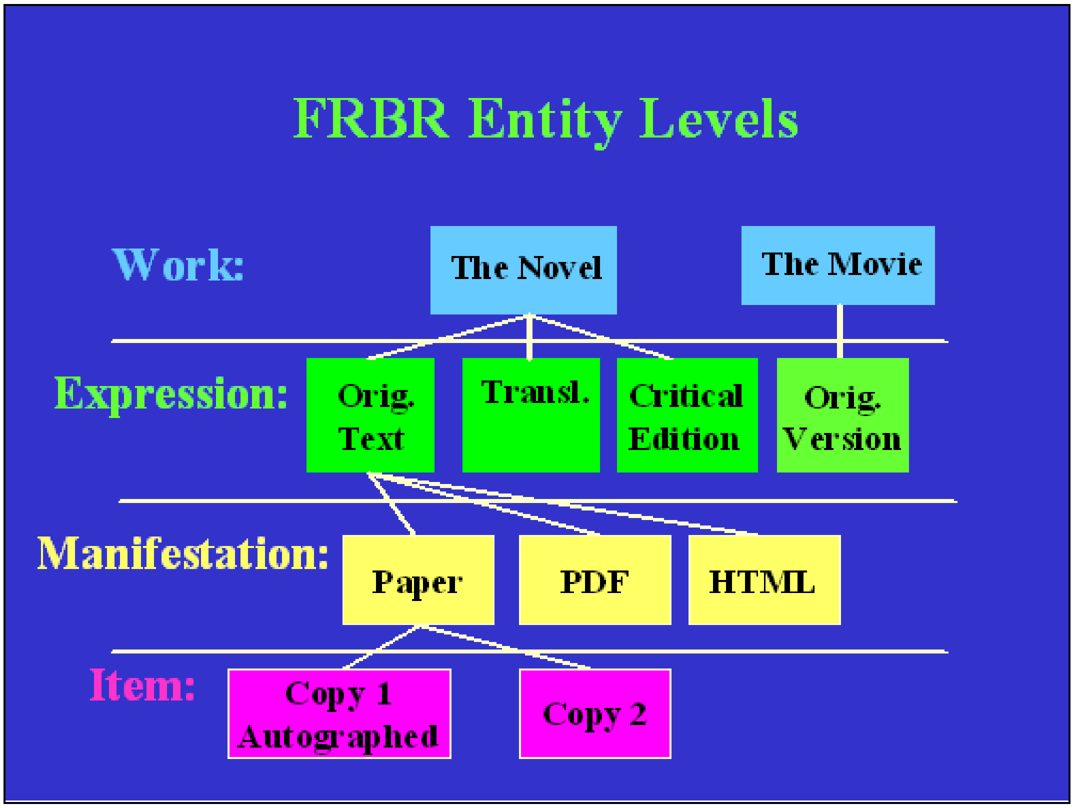
Every example can be clustered via the following features (which all may be relevant, for example to the recommendation algorithms)

- First, a Brand (e.g. Lost) has several series and/or seasons, which themselves are divided into episodes.
- An episode can have several versions in which the content is changed (for example an uncut or censored version)
- An episode could also be altered to improve the quality, like colouring in case when the original is in black and white.
- An episode can be modified to adjust it to the cultural context, for example voice over and subtitles
- The technical context of the intended audience varies a lot, for example encoded in DivX for torrents, or on HDTV for fancy broadcasts.
- Also, the physical experience of the episode could be identified depending on the context. For example, if the episode was on a legal DVD, it could be identified by the product number. In case of a Torrent it could be the torrent hash code. In case of a webcast, it could the URL of the stream. In case of a TV broadcast, it could be the program identifier (e.g. the BBC identifiers).

To summarize the above, we can cluster video content regarding to the *use-cases*, the *stakeholders* and the *properties*. Choosing the identifiers for the content is far from trivial. In the next section we describe how the FRBR model, originating from the bibliographic domain, can be used do separate the different context layers in a way that coincides with the different types of stakeholders.

5.2. The FRBR model

The Functional Requirements for Bibliographic Records (FRBR) is a conceptual model of the bibliographic universe, describing the entities in that universe, their attributes, and relationships among the entities. The FRBR Entity Levels express the different layers that are relevant for the bibliographic domain. For example, the book "Rant" by Chuck Palahniuk, is a "Work". It can have different "Expressions", for example a censored version, or a translation in French. It can have different "Manifestations" like PDF, and the local copy on your machine is an "Item". In the MultiMediaN project[8], where the VUA is involved, a simpler model is proposed for creating the distinctions in the domain of cultural heritage. There are only two layers: works and expressions, like the painting of the Van Gogh's Sunflowers, and a picture of that painting. The FRBR model appears (cf. Figure 6)to be more suitable which we will discuss in the next section.



FRBR Entity Levels

Work: The Novel | The Movie

Expression: Orig. Text | Transl. | Critical Edition | Orig. Version

Manifestation: Paper | PDF | HTML

Item: Copy 1 Autographed | Copy 2

Figure 6: FRBR example

One can argue that the choice of these levels of FRBR might be a bit arbitrary. For example, why not have different types of layout, or prints (paper-back or hard-cover)? Probably this is the most practical one that fits most needs. Davis et al. created an RDF version of FRBR[9], and provides a clear overview of the different properties in their schema. We will adjust this schema in order to match it with the domain of video content.

Figure 7 shows a snapshot of the RDF version.

```
<rdf:Description rdf:about="http://purl.org/vocab/frbr/core#Endeavour">
   <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
   <rdfs:label xml:lang="en">endeavour</rdfs:label>
   <skos:definition xml:lang="en">Any of the products of artistic or creative endeavour.</skos:definition>
   <rdfs:comment xml:lang="en">This class represents any one of the FRBR group one entities.</rdfs:comment>
   <rdfs:isDefinedBy rdf:resource="http://purl.org/vocab/frbr/core"/>
   <owl:disjointWith rdf:resource="http://purl.org/vocab/frbr/core#ResponsibleEntity"/>
   <owl:disjointWith rdf:resource="http://purl.org/vocab/frbr/core#Subject"/>
   <owl:equivalentClass rdf:nodeID="arc675eb4"/>
```

[8] http://www.multimedian.nl/nl/home.php
[9] http://vocab.org/frbr/core.html

```
  <dct:issued>2005-07-15</dct:issued>
  <skos:changeNote rdf:nodeID="arc675eb9"/>
  <skos:changeNote rdf:nodeID="arc675eb10"/>
  <skos:changeNote rdf:nodeID="arc675eb11"/>
</rdf:Description>

<rdf:Description rdf:nodeID="arc675eb4">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label xml:lang="en">the union of all expressions, items, manifestations and works</rdfs:label>
  <owl:unionOf rdf:nodeID="arc675eb5"/>
</rdf:Description>

<rdf:Description rdf:nodeID="arc675eb5">
  <rdf:first rdf:resource="http://purl.org/vocab/frbr/core#Expression"/>
  <rdf:rest rdf:nodeID="arc675eb6"/>
</rdf:Description>

<rdf:Description rdf:about="http://purl.org/vocab/frbr/core#Expression">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label xml:lang="en">expression</rdfs:label>
  <skos:definition xml:lang="en">A realization of a single work usually in a physical form.</skos:definition>
  <rdfs:comment xml:lang="en">This class corresponds to the FRBR group one entity 'Expression'.</rdfs:comment>
  <rdfs:isDefinedBy rdf:resource="http://purl.org/vocab/frbr/core"/>
  <rdfs:subClassOf rdf:resource="http://purl.org/vocab/frbr/core#Endeavour"/>
  <owl:disjointWith rdf:resource="http://purl.org/vocab/frbr/core#Work"/>
  <owl:disjointWith rdf:resource="http://purl.org/vocab/frbr/core#Item"/>
  <owl:disjointWith rdf:resource="http://purl.org/vocab/frbr/core#Manifestation"/>
  <dct:issued>2005-07-15</dct:issued>
  <skos:changeNote rdf:nodeID="arc675eb16"/>
  <skos:changeNote rdf:nodeID="arc675eb17"/>
  <skos:changeNote rdf:nodeID="arc675eb18"/>
</rdf:Description>

<rdf:Description rdf:nodeID="arc675eb6">
  <rdf:first rdf:resource="http://purl.org/vocab/frbr/core#Item"/>
  <rdf:rest rdf:nodeID="arc675eb7"/>
</rdf:Description>

<rdf:Description rdf:about="http://purl.org/vocab/frbr/core#Item">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label xml:lang="en">item</rdfs:label>
  <skos:definition xml:lang="en">An exemplar of a single manifestation.</skos:definition>
  <rdfs:comment xml:lang="en">This class corresponds to the FRBR group one entity 'Item'.</rdfs:comment>
  <rdfs:isDefinedBy rdf:resource="http://purl.org/vocab/frbr/core"/>
  <rdfs:subClassOf rdf:resource="http://purl.org/vocab/frbr/core#Endeavour"/>
  <owl:disjointWith rdf:resource="http://purl.org/vocab/frbr/core#Work"/>
  <owl:disjointWith rdf:resource="http://purl.org/vocab/frbr/core#Expression"/>
  <owl:disjointWith rdf:resource="http://purl.org/vocab/frbr/core#Manifestation"/>
  <dct:issued>2005-07-15</dct:issued>
  <skos:changeNote rdf:nodeID="arc675eb22"/>
  <skos:changeNote rdf:nodeID="arc675eb24"/>
  <skos:changeNote rdf:nodeID="arc675eb25"/>
  <skos:historyNote rdf:nodeID="arc675eb23"/>
  <skos:historyNote rdf:nodeID="arc675eb26"/>
</rdf:Description>
```

Figure 7: a snapshot of the RDF version of the FRBR model by Davids et al.

5.3. FRVR: Functional Requirements for Video Records

The FRBR model will be used as inspiration to identify the different layers in NoTube regarding video content. First, we have to think about the purpose of identifying different kind of layers and then make a practical choice. With practical we mean that it should be easy to understand, it should cover most examples mentioned in the previous section in order to match the requirements of the recommendation algorithms developed within NoTube.

A recommendation has to match several types of requirements: technical, financial, geographical, time-wise, content-wise. The recommendation algorithms that make suggestions based only on the content, only need higher-level descriptions of the content. For example, similar to Amazon's algorithm: people who read these books, also read those, applies to people who watched these movies also watched those regardless of the availability, formats, locations and time-points these users watched these movies. So, for that a simple set of IMDB identifiers is enough. When the recommendation algorithm also takes technical aspects into account, for example only movies that can be played on an iPhone, we will need for to know more about the file like the torrent id which can be looked up in the torrent search engines. We need only those instances of the movies that are in the appropriate format.

Layers

The layers that we suggest for video content are: **Brand**, **Episode**, **Version** and **Item**, where item is an abstract class that has three different kind of instances "TV-broadcast-item", "TOD-item", "Web-item", "file-item". Every item has an 'encoding-type' e.g. mpg-stream, or divX, bitrate and an encryption-type.

- A TV-broadcast-item, is bounded by the moment of broadcasting, has a channel, a duration and perhaps some extra properties like price and advertisement moments.
- A TOD-item (TV on demand), is bounded by the moments the user can choose to watch it, the channel, the duration and perhaps extra properties like advertisement moments.
- A Web-item, is bounded by the moments the user can choose to watch it, the URLs of the streams, file-type, encryption, encoding etc
- A file-item is not time-bound, because it is assumed to be saved on a device owned by the user. For example a torrent, a DVD etc.

The TV-Anytime schemas and vocabularies already covers most of the properties needed to annotate each identified layer. Figure 8 shows the four layers and gives some examples.

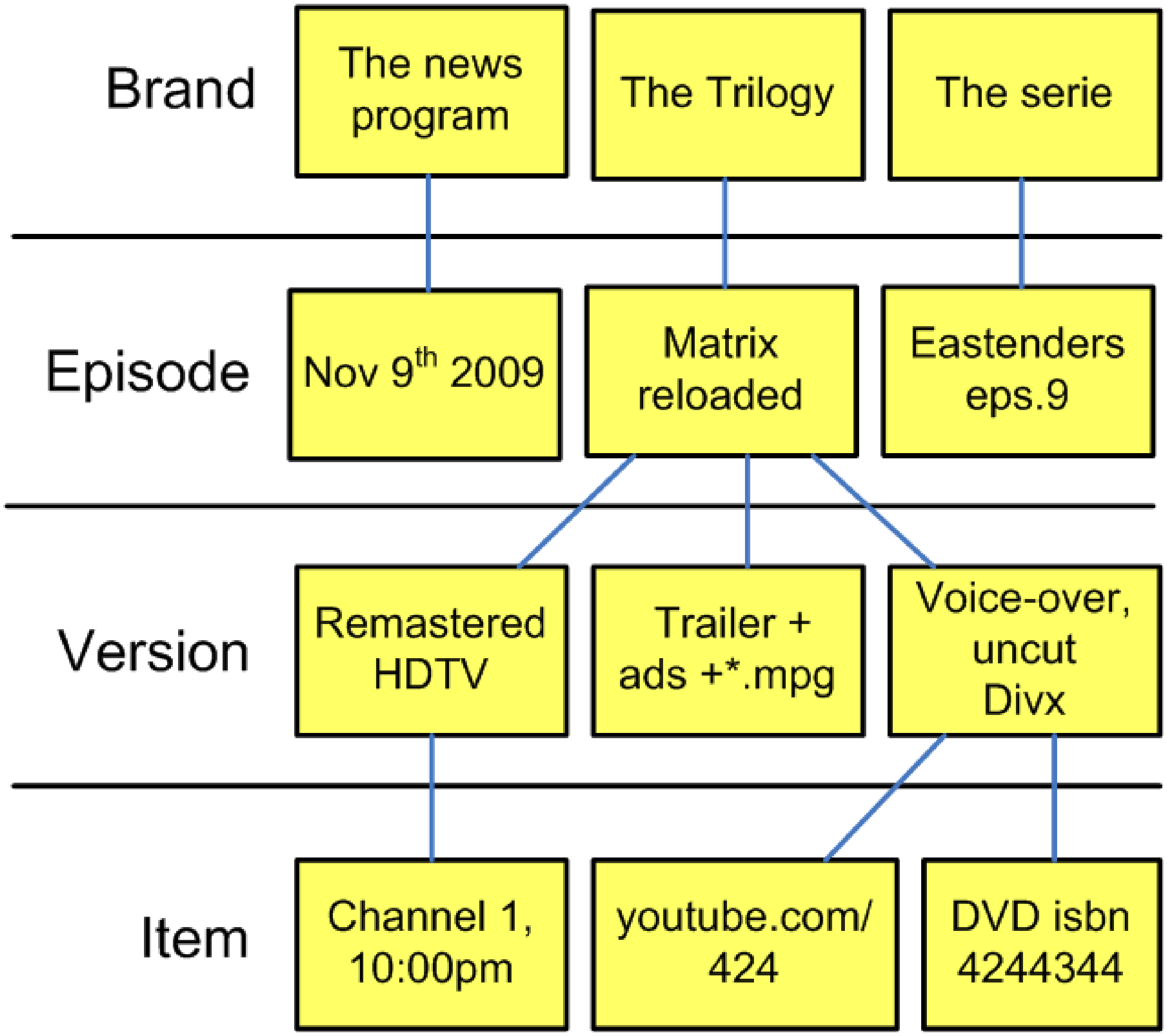


Figure 8: examples of the four FRVR layers

Identifiers

In order to achieve wide applicability of the services developed within Notube, like the enrichment services, it is advisable to re-use existing well-known identifiers as much as possible. Preferably, these identifiers need to be resolvable. This is pragmatic because people recognize them, and they are already used (e.g. IMDB, DBPedia, BBC-programid, Youtube-link, ISBN number, Torrents, MD5 file hashes etc). It is important that the identifiers are pointing to content that matches the level of abstraction belonging to the respective layer in our NoTube 'FRVR' stack. For example, content described at the 'Brand' layer can point to the generic IMDB description or DBPedia page. Similarly, for the 'Episode' layer. Finding identifiers for content described by the 'version layer' is a bit less trivial. Some movies, may have been re-mastered, and subtitled and in wide-screen format. In that case, the identifier preferably points to a description of this version. If that does not exist, we can create one and make it resolvable and registered as persistent urls on purl.org. For the item layer we can use references to online EPG data, a specific torrent id, an isbn number, or an MD5 hash of the content.

Below an informal example to illustrate the four layers

```
<rdf:RDF xmlns="http://www.ebu.ch/metadata/ontologies/EPG/EBU_EPG.owl#"
  xml:base="http://www.ebu.ch/metadata/ontologies/EPG/EBU_EPG.owl"
  xmlns:owl2xml="http://www.w3.org/2006/12/owl2-xml#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:ebu_epg="http://www.ebu.ch/metadata/ontologies/EPG/EBU_EPG.owl#"
  xmlns:dc="http://purl.org/dc/elements/1.1/#"
  xmlns:ebu_tva="http://www.ebu.ch/metadata/ontologies/TVA.owl#"
  xmlns:ebu_av="http://www.ebu.ch/metadata/ontologies/AV/EBU_OWL_AVAttributes_ontology.owl#"
  xmlns:FRVR="http://www.notube.tv#FRVR.owl"
  xmlns:owl="http://www.w3.org/2002/07/owl#">


<!— example of the BRAND layer -->
 <rdf:Description rdf:about="http://www.imdb.com/title/tt0106145/">
   <rdf:type rdf:resource="http://www.ebu.ch/metadata/ontologies/EPG/EBU_EPG.owl#ProgrammeGroup"/>
   <dc:title>"Star Trek: Deep Space Nine" </dc:title>
   <ebu_epg:hasProgrammeGroupSynopsisText>Orbiting the liberated planet of Bajor, a Federation space station guards the opening of a
stable wormhole to the far side of the Galaxy</ebu_epg:hasProgrammeGroupSynopsisText>
   <ebu_epg:hasProgrammeGroupGenre>
http://www.ebu.ch/metadata/ontologies/skos/ebu_ContentGenreCS.skos.xml#3.4.2</ebu_epg:ProgrammeGenre>
   <ebu_epg:hasScheduleEvent rdf:resource="http://www.bbc.co.uk/programmes/b006m86d"/>
 </rdf:Description>


   <!-- example of the EPISODE layer -->
 <rdf:Description rdf:about="http://www.imdb.com/title/tt0708505/">
   <rdf:type rdf:resource="http://www.ebu.ch/metadata/ontologies/EPG/EBU_EPG.owl#Programme"/>
   <ebu_epg:programmeEpisodeOfSeries  rdf:resource="http://www.imdb.com/title/tt0106145/">
   <ebu_epg:ProgrammeDuration>02:20:30 </ebu_epg:#ProgrammeDuration>
   <dc:title>"Star Trek: Deep Space Nine" Behind the Lines (1997)</dc:title>
   <ebu_epg:hasProgrammeSynopsis>After forming an attack plan on the Dominion, Sisko relinquishes command of the Defiant to Dax
after accepting a promotion. On DS9, the resistance faces discovery when Odo links with another
changeling</ebu_epg:ProgrammeSynopsis>
   <ebu_epg:hasProgrammeGenre>
http://www.ebu.ch/metadata/ontologies/skos/ebu_ContentGenreCS.skos.xml#3.4.2</ebu_epg:hasProgrammeGenre>
   <ebu_epg:hasScheduleEvent rdf:resource="http://www.bbc.co.uk/programmes/b006m86d"/>
   <ebu_epg:hasOnDemandProgramme rdf:resource="http://www.bbc.co.uk/iplayer/episode/b00q0g8g/>
   <ebu_epg:ProgrammeDuration>02:20:30 </ebu_epg:#ProgrammeDuration>
 </rdf:Description>


   <!-- example of the VERSION layer -->
 <rdf:Description rdf:about="http://www.bbc.co.uk/programmes/b006m86d">
   <rdf:type rdf:resource="http://www.ebu.ch/metadata/ontologies/EPG/EBU_EPG.owl#ScheduleEvent"/>
   <ebu_epg:ScheduleEventPublishedStartTime>2010-04-29T13:25:00Z</ebu_epg:ScheduleEventPublishedStartTime>
   <ebu_epg:ScheduleEventPublishedEndTime>2010-04-29T13:25:00Z</ebu_epg:ScheduleEventPublishedEndTime>
   <frvr:hasBroadcastChannel rdf:resource="http://purl.org/vocabularies/tv-channels/bbc-uk/bbc-one">
   <nt:hasBroadcaster rdf:resource="http://www.bbc.co.uk/foaf.rdf"/>
 </rdf:Description>

 <-- example of the ITEM Layer -->
 <rdf:Description rdf:about="http://www.bbc.co.uk/broadcast/b006m86d">
   <rdf:type rdf:resource="http://www.ebu.ch/metadata/ontologies/EPG/EBU_EPG.owl#BroadcastEvent"/>
   <ebu_epg:hasRelatedScheduleEvent rdf:resource="http://www.bbc.co.uk/programmes/b006m86d"/>
 </rdf:Description>       <!-- the on demand programme via stream -->
 <rdf:Description rdf:about="http://www.bbc.co.uk/iplayer/episode/b00q0g8g">
    <rdf:type rdf:resource="http://www.ebu.ch/metadata/ontologies/TVA.owl#OnDemandProgramme"/>
    <ebu_tva:DeliveryMode rdf:about="#streaming"/>
    <frvr:location>http://www.bbc.co.uk/iplayer/episode/b00q0g8g.avi</frvr:location>
    <ebu_tva:hasOnDemandProgrammeBeginOfAvailability>2010-05-
29T13:25:00Z</ebu_tva:hasOnDemandProgrammeBeginOfAvailability>
    <ebu_tva:hasOnDemandProgrammeEndOfAvailability>2011-05-29T13:25:00Z</ebu_tva:hasOnDemandProgrammeEndOfAvailability>
    <ebu_av:hasAVAttributesStreamBitrateAverage>5600<ebu_av:hasAVAttributesStreamBitrateAverage>
    <ebu_av:hasVideoAttributesCodingDefinition>AVI</ebu_av:hasVideoAttributesCodingDefinition>
 </rdf:Description>

<!—- another example of the ITEM Layer -->
 <rdf:Description rdf:about="http://torrents.thepiratebay.org/3561324/Star_Trek_Deep_Space_Nine_S04.3561324.TPB.torrent">
    <rdf:type rdf:resource="http://www.ebu.ch/metadata/ontologies/TVA.owl#OnDemandProgramme"/>
    <ebu_tva:DeliveryMode rdf:about="download"/>
    <nt:location>http://www.bbc.co.uk/iplayer/episode/b00q0g8g.avi</nt:location>
    <ebu_tva:hasOnDemandProgrammeBeginOfAvailability>2010-06-
29T13:25:00Z</ebu_tva:hasOnDemandProgrammeBeginOfAvailability>
    <ebu_av:hasVideoAttributesCodingDefinition>DIVX</ebu_av:hasVideoAttributesCodingDefinition>
  </rdf:Description>

</rdf:RDF>
```

Figure 9: an informal example of instantiations of the four FRVR layers

In the coming period we will formalize the FRVR model and connect it to the use-cases in WP7 and their respective recommendation strategies.

6. References

1. Buchanan, G. 2006. FRBR: enriching and integrating digital libraries. In Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (Chapel Hill, NC, USA, June 11 - 15, 2006). JCDL '06. ACM, New York, NY, 260-269. DOI= http://doi.acm.org/10.1145/1141753.1141812
2. Vliet van der H., I. Maks, P. Vossen, R. Segers(fc.). "The Cornetto Database: Semantic issues in linking lexical units and synsets", in the proceedings of the 14th EURALEX International Congress, July 6-10, 2010, Leeuwarden, the Netherlands.
3. Guus Schreiber and Alia Amin and Mark van Assem and Victor de Boer and Lynda Hardman and Michiel Hildebrand and Laura Hollink and Zhisheng Huang and Janneke van Kersen and Marco de Niet and Borys Omelayenko and Jacco van Ossenbruggen and Ronny Siebes and Jos Taekema and Jan Wielemaker and Bob Wielinga: MultimediaN E-Culture Demonstrator. In the proceedings of the Internation Semantic Web Conference 2006 , p. 951--958.
4. Chris van Aart, Ronald Siebes, Vicky Buser, Lora Aroyo, Yves Raimond, Dan Brickley, Guus Schreiber, Michele Minno, Libby Miller, Davide Palmisano, Michele Mostarda. "The NoTube Beancounter: Aggregating User Data for Television Programme Recommendation". In the proceedings of the Social Data on the Web (SDOW2009), Washington DC (USA), October 25, 2009