Topic Map Aided Publishing – A Case Study of Assembly Media Archive

Aki Kivelä[1], Olli Lyytinen[1]

[1] Grip Studios Interactive Oy
office@gripstudios.com

Abstract. Contemporary web publishing requires efficient technologies and tools to build and maintain large and dynamic sites. This paper introduces a Topic Map application used to publish documents during Assembly'04 event. Motivation of this paper is to describe the application and reveal success points and real-life problems encountered while applying Topic Maps on document and knowledge publishing.

1. Introduction

Over the last decade WWW has become a stable and vivid platform to publish content. First publications were human coded and interlinked HTML documents. Although the simplicity of publishing certainly empowered the explosion of WWW, production and management of large sites was difficult and great number of technologies and applications has been developed to make the publishing process more efficient. In the advent of semantic web technologies this development is expected to be taken one step further although buzzwords and hype make the view obscure.

This paper reviews a real-life application of Assembly Media Archive, using semantic web technology, namely Topic Maps to maintain knowledge about published documents and discusses success points and problems faced while building the application.

1.1 Assembly'04 Event

Assembly is an intense festival for young computer enthusiasts. During the four-day festival participants compete in graphics, music and demos, play games and meet like-minded. Assembly'04 was arranged in Helsinki, Finland during 5th – 8th August 2004 and gathered over 5000 computer demo programmers, media artists and gamers from over 20 countries [2].

Assembly'04 involved extensive media production. For example, AssemblyTV broadcasted live footage almost 24 hours a day in digital television and internet. Objective of the Assembly Media Archive reviewed in this paper was to publish selected media documents produced during the Assembly'04 event.

2. Assembly Media Archive

Assembly Media Archive is a WWW publishing platform tailored to publish images and videos produced during the Assembly'04 event. The Media Archive was integrated to publishing processes and systems of Assembly'04. Publishing environment of the Media Archive included *VideoStore*, *Elaine* and two content production teams (See fig. 1).

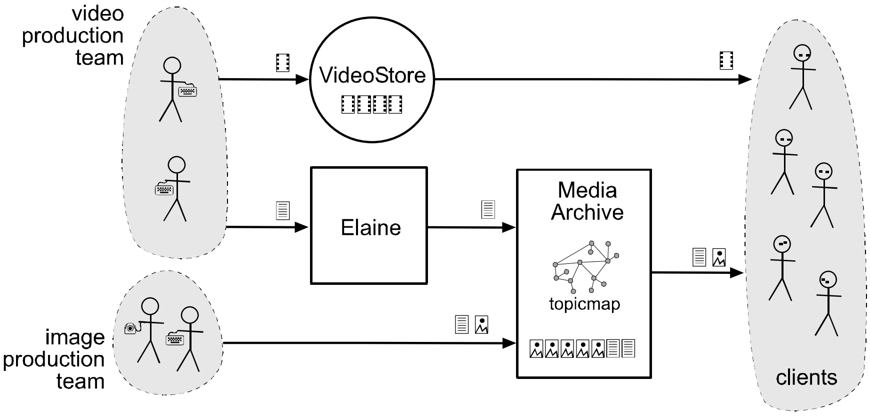

Fig.1. Media Archive's publishing environment included content production teams for videos and photographs, AssemblyTV's broadcast database Elaine, WWW repository VideoStore and clients. Arrows represent content and data streams. Video production team stored video files to VideoStore and information about video files to Elaine. Media Archive received information about video files from Elaine but never directly accessed video files. Image production team used Administration application and picture maintenance tool to store image files and related knowledge to Media Archive. Media Archive maintained all internal knowledge in single topic map. Template engine was used to construct Media Archive's public interface used by clients.

3. Data and Document Sources

Assembly Media Archive acquired data and documents from external and internal sources. These sources were internal in the case of ontologies and external in the case of images and videos. Next chapters discuss these sources more detailed.

3.1 Ontologies

Term *ontology* refers in this paper to a simple predefined classification of concepts relevant to the Assembly Media Archive. The Assembly Media Archive had two internal and predefined ontologies. The system ontology was not visible to public and contained information for the administration interface and the Media Archive server. The keyword ontology contained all the predefined keywords, organized into a tree hierarchy to make the client browsing more convenient. Subclass and super-class relations were used to build the tree hierarchy.

Table 1. Example of keyword category Atmosphere. Atmosphere has four subclasses specifying the nature of addressed atmosphere.

Atmosphere	
	Drowsy
	Excited
	Frustrated
	Joyful

Ontologies were expected to require minimal maintenance during the Assembly'04 event and it was decided that no specific administration interface is needed for ontology management. Ontologies were stored as plain XTM topic maps and edited with general topic map editor included in Administration application.

3.2 Images and Related Knowledge

Image production team photographed and entered pictures and related descriptions to the Media Archive. It was important that these photographs could be added real time to the archive. A separate picture maintenance tool was built to allow efficient input of pictures. With the tool photographer only needed to write a name and a caption of the image and select appropriate keywords from the predefined list.

Maintenance tool created small topic maps from the entered data and merged these into the internal topic map of Media Archive. Tool also created image thumbnails and uploaded images to the Media Archive server. Images were stored in Media Archive.

3.3 Videos and Related Knowledge

AssemblyTV team had an existing broadcast database system, called Elaine, to handle their video broadcasts and clipping into separate files to allow video-on-demand (vod). The vod files were published in Assembly Media Archive. Media Archive polled Elaine frequently for available video files. If Elaine's video list had changed Archive updated internal topic maps with the information provided by Elaine. Elaine's video lists were simple XML documents the Archive transformed into XTM topic maps using XSL style sheet before internal use.

Example of a XML document of video related knowledge produced by Elaine

```
<?xml version="1.0" encoding="iso-8859-1"?>
<videos>
    <video id="18" length="00:30:00" aired="">
    <name lang="*">Assembly memories</name>
    <description lang="fi">Vetiskelkää kanssamme
        Assembly muistejen lomassa. Tapahtuman yli 10
        vuotinen taival esitellään tiukassa paketissa.
        Takuukamaa kiinnostuneille. Sisältää kuvamate-
        riaalia vuodelta 1995 lähtien!</description>
    <description lang="en">More than 10 years of As-
        sembly is the topic of the show that´ll bring
        out the highlights of the partys throughout the
        decade. Don´t miss it! Contains video material
        beginning from the party 1995. </description>
    <resource type="video"
        src="http://www. videostore.com/
             Assembly_Memories.avi"
        start="00:00:00"
        length="00:13:48"
        preview="http://www.videostore.com/previews/
             Assembly_Memories.mpg"
        thumbnail="http://www. videostore.com/thumbs/
             Assembly_Memories.jpg" />
    </video>
</videos>
```

Video files were stored externally in VideoStore due to the required high bandwidth. VideoStore was a simple WWW site where video files had unique addresses. These addresses were entered to Elaine and passed to Assembly Media Archive in the XML video data described above.

4. Topic Maps

Media Archive used XTM topic maps [3] to store knowledge about media documents. Topic map is a standard[1] used to describe files, internet pages or actually pretty much anything. As the name suggests a topic map consists of topics. Each topic is an entity that we want to describe. Topics can be interlinked with associations and may have occurrences, which are either textual data or an URI pointing to an external resource. Detailed representation of topic map standard is beyond of this paper. Excellent topic map introduction is [4] for example.

[1] ISO/IEC 13250:2000. See references.

Topic map standard specifies a well defined operation to merge multiple topic maps into a single large topic map. Assembly Media Archive uses the merge feature to aggregate information from all different sources discussed in previous chapter into a single topic map. Merge is very powerful feature but involves also problems. One of the drawbacks of merging is that after a successful merge it is very hard or even impossible to solve where a specific piece of information came from or find all information coming from a specific topic map. This is not a problem if merge operation is repeated from scratch every time something changes. However the operation requires much computing resources and may be too slow for real time systems.

Thus we decided to keep the information from different sources separate and create a virtual topic map that would look like a regular topic map but actually presented the merged information from the underlying separate topic maps. When one of the underlying topic maps was changed, the virtual merged topic map changed automatically and no new merge was required.

Keeping topic maps separate had several advantages. First, as described above it allowed us to update information efficiently. Also, because all the editing was targeted at the underlying topic maps, management of concurrent editing became easier. If all the information was kept in a single topic map, everybody would edit the same topic map, but when we had different topic maps, people would not accidentally overwrite each others changes. It also protected information from accidental changes, for example all the system topics were kept in a topic map of their own.

4.1 Reduced Topic Maps

Assembly Media Archive used a reduced topic map model containing only a subset of the standard topic map features. However a reduced topic map is still a valid standard topic map. The reason for reduced implementation of standard was mostly because of experiences from previous topic map systems built in Grip Studios Interactive[2]. Many of the standard features of topic maps had little or no use in systems like Assembly Media Archive and getting rid of these features enabled the topic map implementation to be faster and more memory efficient.

The most notable differences between the standard and reduced topic maps are[3]:

1. Subject identifiers and subject locators are never resolved. This means that they are simply used as identifiers and it has no effect at all whether or not an URI used in a subject identifier contains something, for example another topic in the same or another topic map.
2. Only one base name is allowed per topic and base names cannot have a scope.

[2] Grip Studios Interactive had built three similar systems using Topic maps to store knowledge before Assembly Media Archive.

[3] Every listed difference defines reduced topic maps.

3. Variant names are not base name dependent but are all considered to be variants of the single base name of the topic. Variant names can have scopes.
4. Only resource data occurrences. When a resource reference occurrence would be needed, it can be expressed by creating a new topic for the referred URI, using the URI as the subject locator of the new topic and then associating the two.
5. Occurrences can only have a scope of one topic. This is usually used to store the language of the occurrence.
6. Topic can contain only one occurrence per type and scope.
7. Topic can contain only one variant name per scope.
8. Associations cannot have scopes.
9. Associations can only have one player per member and cannot have two members with same role.

It should be noted that many of the features that were left out are very vaguely defined in the topic map standard. For example, the interpretation of scopes is almost entirely left to the application, as is the decision when to resolve subject identifier URIs.

In the Assembly Media Archive we used a memory implementation of the topic map. A database-stored implementation would also be possible and quite easy to implement.

5. Administration Application

We built an administration application used to edit topic maps. Application was a Java program executed locally on a network connected machine. Administration application downloaded the topic map fragment to be edited from the Media Archive server and edition was accomplished locally. This worked well as the entire topic map was not needed and network traffic remained sufficient low. After the administrator was ready the application sent the changes to the server. The changes immediately propagated to the merged topic map and to the WWW interface open for clients.

Using the administration application required technical knowledge about topic maps and additional tools was built to make it easier to perform some common tasks. One of such tools was an image importer that the photographer used to add photographs into the topic map. These tools were extensions to the administration application and principle of partial edition described above also applied to them.

6. WWW interface

WWW interface of the Assembly Media Archive was open for public. Interface was used with a WWW browser and consisted of dynamically produced HTML pages constructed with Apache Velocity[4] template engine.

As a thumb rule each HTML page represented single topic with associations as links to other HTML pages and topics. Navigation structure of the Media Archive's WWW site conforms closely to the topic map. This is important as it is possible to predict the navigation structure if topic map is known and vice versa. Ideal vision of one topic per one page did not apply to page construction. HTML page construction required wide access to topic map.

In order to support free text search Apache Lucene[5] was used to construct a search index from the topic map. Search engine enabled very powerful search operations and expanded the navigation structure significantly.

Representing information of the topic map in a clear way is challenging because a regular user rarely wants to be aware of the underlying topic map. The user doesn't want to see that "a picture has an association of type keyword with topic computer." Instead we should try to express this in a more natural way: "This picture has a keyword computer." In some cases it may be easy to construct such statements from the information contained in the topic map but in general it becomes quite hard. It is especially hard trying to do this in Finnish.

7. Conclusions

Topic maps can be used in document publishing to store knowledge about media documents such as images and videos. It is easy to transform XML documents into small topic maps using style sheets. Topic map standard offers merge rules to aggregate smaller topic maps into a single one. Although merge is a very powerful feature, it can cause problems as the information of topic sources are loosed during the operation. This is especially critical when the system is not able to repeat the merge from scratch when only a portion of topic map is changed. To enable partial and efficient updates the Assembly Media Archive introduced virtual topic map as a collection of merged topic maps.

We experienced some features of topic map standard unnecessary and confusing. One of these features is naming of the topics. Our experiences also suggest that existing topic map implementations are inefficient if the number of topics and associations increase enough. As a consequence a reduced topic map implementation was built for the Assembly Media Archive.

[4] http://jakarta.apache.org/velocity/index.html

[5] http://jakarta.apache.org/lucene/docs/index.html

Topic Maps can be visualized with HTML. Topic map offers an intuitive navigation structure for HTML page visualizations and navigation structure can be improved considerably using search engine to index all textual data stored in topic map.

Acknowledgments

We would like to thank all Assembly'04 personnel making the Assembly Media Archive possible. Especially we would like to thank Kim Viljanen for the initial idea of adapting our expertise on Topic Map applications to Assembly Media Archive. Kim Viljanen also wrote Elaine's XML output function. We would also like to thank Lauri Pitkänen who made a considerable work on video-on-demand production process and Jacqueline Kivimäki who wrote the keyword category for images and used the category successfully to tag all published images. Also we would like to thank photographer Oleg Hartsenko for taking excellent pictures for the Archive.

References

1. ISO/IEC 13250:2000 Topic Maps: Information Technology -- Document Description and Markup Languages, Michel Biezunski, Martin Bryan, Steven R. Newcomb, ed. (1999).
2. Assembly'04 Press Release 2nd August 2004. See http://media.assembly.org/asm04/Assembly2004_press_ENG.rtf
3. Steve Pepper, Graham Moore ed. XML Topic Maps (XTM) 1.0, TopicMaps.Org Specification. 2001. See http://www.topicmaps.org/xtm/index.html
4. Steve Pepper. The TAO of Topic Maps, finding the way in the age of infoglut. 2000. See http://www.gca.org/papers/xmleurope2000/pdf/s11-01.pdf