Using SIVA XML and SMIL for Interactive Non-linear Videos: a Comparison

Emanuel Berndl
Chair of Distributed Information Systems,
University of Passau, Germany
berndle@fim.uni-passau.de

Britta Meixner
Chair of Distributed Information Systems,
University of Passau, Germany
meixner@fim.uni-passau.de

Harald Kosch
Chair of Distributed Information Systems,
University of Passau, Germany
harald.kosch@uni-passau.de

ABSTRACT
With recent technologies, it is possible to create appealing multimedia presentations or extended videos with a high level of interactivity. Standards like SMIL provide extensive structures to describe metadata for timing and spacing of single media elements which then form a presentation. While multimedia presentations are viewed mainly in a linear manner, provide interactive and non-linear videos a much higher level of interactivity and navigational possibilities. In this work, we examine the usability of SMIL to describe temporal and spatial relationships of videos and annotations, as well as interaction and navigational elements. We therefore compared SMIL with the SIVA XML. We tried to find ways to express SIVA XML structures with SMIL attributes and elements. After that, we compared the DTD/XSD of SMIL and SIVA XML using XML metrics.

Author Keywords
SMIL, Interactive Video, Non-linear Video, XML, Video Annotations, Multimedia Document, Metrics

ACM Classification Keywords
I.7 Document and Text Processing: Document Preparation—Hypertext/hypermedia, Multi/mixed media, Standards

INTRODUCTION
Nearly every web page provides multimedia contents today. These reach from animated images to sounds and embedded videos. With recent technologies and increasing Internet bandwidths, appealing combinations of different types of media and various forms of user interaction are possible. Two main fields of research can be found in this area, namely "multimedia presentations" and "hypervideos". The latter ones are very similar to "interactive non-linear videos with additional information". An applicable definition of multimedia presentation is given by Nimmagadda et al. as follows: "Multimedia presentations are collections of different media files [...] like text, images, videos, and animations with different resolutions, durations, and start-times. [...] The layout of multimedia presentations is defined by the locations and the start times of the objects” [11]. In contrast, interactive non-linear videos are defined by us (extended from [6]) as follows: “ [...] An interactive non-linear video is a digitally enriched form of video materials arranged for an overall concept. It presents additional information beyond the original content. Furthermore, it offers new forms of influence and navigation in the video and additional contents” [10].

Watching multimedia presentations, the viewer is rather passive, but basic interaction and navigation may be possible. The viewer is elicited from his passivity viewing interactive non-linear videos in contrast. This form of video consists of video scenes (“main video”) and additional information which enhance the scenes. Timeline and control bar are extended with additional functions. These provide control on the flow of the video and give hints on when additional information is displayed. Decision elements in the video allow the selection of a certain branch of the video instead of watching it in a linear way. Furthermore, additional information which may be any type of medium, like text, image, or video, is added as an annotation to the main video. We proposed an XML format for this type of video in [9]. It is customized for interactive non-linear videos while using SMIL [14] may be possible up to a certain point, but leads to problems and work-arounds in some areas. In this work, we try to show the advantages and disadvantages of using SMIL for the description of interactive non-linear videos compared to our XML format1. Thereby, this work makes the following research contributions: Requirements were identified (see section ) and both formats were checked for their suitability to implement these requirements (see section ). Metrics were used to compare the complexity of the SIVA XML schema and SMIL, see section .

REQUIREMENTS
An analysis of usage scenarios like e-learning, virtual tours, mobile help systems, or sport events revealed several requirements according to timing and spacing of media elements in interactive non-linear videos with additional information. Needed functions and elements are as follows (see also [9]):

• Media, main video, and annotations: As specified for interactive non-linear videos, always one video is displayed as the main video. Additional information called “annotations” the remainder of this work may be shown with

---

1For a more detailed description of the SIVA XML schema see [9]. SMIL is described in [4] and [14]
this video. Therefore, several different types of media like images, audio-files, videos, and text should be usable. It should be possible to handle them differently during playback, for example should a subtitle be positioned automatically.

- **Event-based timing model**: Main video and annotations may be time dependent or time independent. For this reason, an event-based timing model is preferred to a structured timing model due to the high level of interactivity mixed with fixed points in time where annotations are displayed or hidden. By keeping timing issues as local as possible, synchronization is realizable more easily.

- **Temporal relationships between main video and annotations**: Temporal relationships in form of start and end point or durations of display need to be defined between the main video (scene) and each of the annotations.

- **Spatial relationships between videos and annotations**: A positioning of main video and single annotations or groups of annotations needs to be defined. Annotations may be displayed statically in areas around the video or as an overlay over the video. Furthermore, dynamic annotations may move on a path on the video canvas. Automated arrangement of annotations in defined areas facilitates the authoring process.

- **Decision elements at forks in video flow**: The playback of interactive non-linear videos includes different strands of scenes. Button panels and quizzes are needed to select the next scenes which are displayed to the viewer.

- **Table of contents**: One way of extended navigation is provided by a table of contents which has to be defined and linked with single scenes.

- **Keyword reference list**: A second way of extended navigation is implemented with a keyword search. Keywords need to be linked with scenes or annotations in order to find information more quickly.

- **Extensibility**: The structure of the XML format has to be extensible in case of new ways of interaction that should be mapped into the model. Furthermore, changes in the XML file should be kept as local as possible in the structure without changing bigger parts of the existing file.

**DESCRIPTION OF SIVA XML SCHEMA AND SMIL DTD**

The SIVA XML schema and the SMIL DTD show several differences in structure and scope. While SMIL tries to cover many different areas of application, the SIVA XML schema is exactly tailored to the needs of interactive non-linear videos with additional information. We now give a short overview over the formats before we compare them based on the requirements we determined in the previous section.

**SMIL DTD**

SMIL stands for Synchronized Multimedia Integration Language and it is a standard for interactive multimedia presentations released by the World Wide Web Consortium (W3C). Design goals of SMIL were to define “an XML-based language that allows authors to write interactive multimedia presentations. Using SMIL 3.0, an author may describe the temporal behavior of a multimedia presentation, associate hyperlinks with media objects and describe the layout of the presentation on a screen. [Furthermore, it should allow] reusing of SMIL 3.0 syntax and semantics in other XML-based languages, in particular those who need to represent timing and synchronization” [14]. Used media files are images, text, audio files, videos, animation, and textstreams which are linked to an internal graph structure. Navigation is possible in a presentation but not in single continuous media files. Furthermore, it is possible to define hotspots for navigation or to display additional information. With the usage of the elements and attributes from the timing modules, “time can be integrated into any XML language” ([4], S.117). It is possible to define start and end time, duration, persistence, repetition, and accuracy of objects and relations between those objects ([4], S.117). The layout of a presentation is defined by the “relative placement of (multiple) media objects”, but SMIL does not concern the internal formatting of media objects ([4], S.149). SMIL is based on CMIF [3] and the AHM [7]. The final version of this standard is the SMIL 3.0 Recommendation which was published on December 01, 2008 [14]. Previous versions of this standard were SMIL 1.0 released in 1998, SMIL 2.0 released in 2001, and SMIL 2.1 released in 2005 [4]. SMIL 3.0 consists of 12 modules of elements and attributes (Animation, Metainformation, Content Control, Structure, Layout, Timing and Synchronization, Linking, Time Manipulations, Media Objects, Transition Effects, smilState and smilText) described as a DTD. The Timing and Synchronization part is the most important one [4]. Furthermore, five profiles are built which use the enlisted elements and attributes, namely the SMIL 3.0 Language Profile, the SMIL 3.0 Unified Mobile Profile, the SMIL 3.0 DAISY Profile, the SMIL 3.0 Tiny Profile, and the SMIL 3.0 smilText Profile [14]. These profiles may limit the elements and attributes of the standard or extend it with functionality from other XML languages [4].

Extensions for SMIL can be found in the work of Cazenave et al. [5], Pihkala and Vuorimaa [12], and Vaisenberg et al. [13]. These works add a table of contents, a search function, and a bookmark function [13], “location information, tactile output, forms, telephoning, and scripting” [12], and the option to publish multimedia documents on the web using HTML5, CSS, and SMIL Timesheets [5] to SMIL. We do not consider these extensions, because they are not implemented in standard players. In the following sections only elements and attributes from the SMIL 3.0 specification are used.

**SIVA XML Schema**

The SIVA XSD\(^2\) was designed during the projects “Interaktives Video Editierungstool zum netzwerkbasierten Wis- senstransfer (ivi-Pro)\(^3\)” and “ivi-Pro 2.0 - Interaktives Video

---

\(^2\)The XSD file can be downloaded from [http://siva.uni-passau.de/sites/default/files/downloads/sivaPlayer.xsd](http://siva.uni-passau.de/sites/default/files/downloads/sivaPlayer.xsd)

\(^3\)“Interactive video editing tool for network-based knowledge transfer (ivi-Pro)”
im Zeitalter von Mobilität und Kollaboration". Major design goals were an easy expandability and a slim format which exactly fitted our requirements as well as existing and potential future scenarios without too many limitations. We decided to implement some logic into the player to avoid repetitive definitions in the XML file. Besides a main video, usable media files are images, audio files, videos, richtexts and subtitles. These can be displayed as “global annotations” during a whole video, or as “local annotations” during a single scene. It is possible to define a non-linear structure of scenes, where each decision element provides selection panels or quizzes to viewers. Other navigational elements are a table of contents and a keyword search. Hotspots in the video trigger the display of additional information. The timing is kept local, thus synchronization is only necessary for a scene and not for a whole video. The SIVA XML consists of six parts represented by six main elements below the root element: <projectInformation>, <sceneList>, <resources>, <actions>, <tableOfContents> and <index>. These elements are linked by ID/IDREF attributes which are checked by constraints for their consistency. A more detailed description of the XML format can be found in [9].

FEASIBILITY ANALYSIS
We already stated what an interactive non-linear video is and what requirements need to be fulfilled in order to satisfy all needs of such a video. The following part shows how feasible an implementation of an interactive non-linear video is with both of the given XML languages SMIL and SIVA XML. Therefore we first present the feasibility of every requirement in regard of both languages. We also propose how extensive the implementation is and what features can or cannot be realized. Afterwards we conclude our feasibility results in section . Examples used in this section are adapted from [2].

Media, Main video, and Annotations
The entire presentation of an interactive video consists of a main video with the addition of annotations. Annotations are multimedia elements, that are supposed to enhance the interactive feeling and can be used to give more information about the topic of the video. Annotations can be triggered (invoked for display) by user interaction, established by a click on certain defined portions of the video, or by reaching a specified point of time. The placement is a fixed point or a path, resulting in a moving annotation.

Both XML languages support the full variety of media annotations, but differences are met in terms of the placement. When the editing of a SMIL presentation is finished, the placement of all its elements is set. This can result in overlapping annotations, for example pictures, when they are placed in the same area. The SIVA XML is usually interpreted by a player which supports an automated placing function, that will arrange the annotations next to each other. Another weakness of SMIL is found concerning the pathing. In order to achieve the exact demanded movement, the element of the annotation would require four <animate> for every step of the path.

Event-based Timing Model
The SIVA XML is fully designed to fulfill the requested timing model that an interactive non-linear video needs. Scenes are built modularly and do not have to be processed in a linear order as in SMIL. Annotations are started by defined triggers during a scene.

In contrast, SMIL makes use of an interval-based timing model. Although the whole functionality of an interactive non-linear video could be implemented, there are slight disadvantages with this model. Every relation between main video and annotation is bound together as a result of the SMIL element structure. It is not as modular and as local as in the SIVA XML.

Temporal Relationships between Main Video and Annotations
Temporal relationships can be implemented well in both languages. The modularity of the SIVA XML makes it possible that every temporal relationship can be modeled by local XML constructs which are linked by ID/IDREF attributes.

SMIL on the other hand supports a broad range of elements to satisfy the temporal needs of an interactive non-linear video. By making use of the basic temporal elements <seq> and <par> combined with more complex ones like priorityClass and their timing attributes start and end, every relationship between and inside the parts of an interactive video can be implemented.

Spatial Relationships between Main Video and Annotations
The SIVA XML shows advantages in spatial relationships compared to SMIL. All media and navigational elements of the interactive non-linear video can be placed specifically where they need to be, annotations can be arranged automatically, their paths and/or positions that are defined in the SIVA Producer will be fulfilled. If elements of a displayed panel (e.g. the table of contents) cannot be shown in its full size, the player can adapt to it by using techniques like scrollbars to supply the full range of accessibility for all elements.

SMIL on the contrary has some difficulties establishing these requirements. Every element has to be aligned exactly with its left-, top-, right-, bottom-, height-, and width-attributes in order to determine its position. If the given set, for example a list of links or buttons, is to large, it cannot be displayed entirely. The portion that is to large for the displaying area will be cut out. Furthermore, links can not be sized according to the width that their text needs, so consecutively you can activate a link by clicking into the “free” area that is residing to the right side of the link. The counterpart here would be to size the buttons to a fixed length, but if
a given text exceeds that boundary, the text will be cut off as well.

**Decision Elements at Forks**

Enhanced navigational features of the interactive non-linear video are forks. These are usually button panels and quizzes. At a button panel the viewer can pick one option how the main video shall proceed. In a quiz, a row of questions with multiple choice answers is posed. Every answer will give a certain amount of points. The achieved sum will determine the continuation after the quiz. For these functionalities, different elements like the button or answer panel are needed.

In the **SIVA XML schema**, all of this is supported entirely by defined complex types. The modular structure of the XML file makes the scenes accessible by triggers which are linked to the buttons of a choice panel. The quiz functionality specifies questions together with their answers. The correct answers are marked and obtainable points per question are set. Furthermore, point ranges for a whole quiz are defined for the selection of the following scene.

All of these features can be implemented in SMIL. Forks and their paths are supported by button panels, that suffer the spatial disadvantage already mentioned in section . Although complex in the XML structure, the quiz functionality is also realizable. But a problem arises from the fact, that a path after a fork may be an edge to an already played scene. Jumps inside the SMIL file are in need of a link element. In order to implement this, the viewer is confronted with a panel that has to be clicked so that the video can continue at an earlier point.

**Table of Contents**

The table of contents contributes to the non-linear character of the interactive video. When displayed (after clicking a button), a panel with links in a tree structure is presented to the user. By activating one of these links, the corresponding scene will be played. Here, the **SIVA XML** allows the addition of sub entries for every item of the panel. SMIL can only implement a list of links for this feature, while also suffering the spatial problems mentioned in section.

**Keyword Reference List**

A keyword search could not be established for the SMIL export as such functionality is not supported by the language. The **SIVA XML** supports all the requirements that are needed for a search. Keywords are linked with scenes or annotations. When the user selects a keyword of a scene, the scene starts at its beginning. Selecting the keyword of an annotation, the video starts play-back at the point where the annotation is displayed. They can be searched while the interactive non-linear video is played.

**Extensibility**

Regarding the possibilities to extend the given model for new features, both XML languages are capable of integrating additional sets of elements. On both sides, the DTD or XSD files have to be adapted as well as the interpretation of the resulting exported XML document at the player. But in terms of the possibilities of changing an XML document of both languages that already exists, the SIVA XML has slight advantages. Due to the modular structure, it is easy to add new scenes, keywords, or annotations to an interactive video. This process is more complicated in SMIL based on the nested composition of the elements. Especially the temporal structure needs to be kept correct. Adding one single element may have impact on different parts of the interactive video, altering it in a way that may not be intended by an author or an insertion algorithm.

**Feasibility Conclusion**

Table 1 represents an overview of our analysis by listing every requirement and its feasibility for both languages. The feasibility is ranged on a scale from "very bad or not at all" (denoted as "--"), "partly feasible" (denoted as "--") over "neutral" (denoted as "0") to "feasible with some drawbacks" (denoted as "+") or "meets all requirements" (denoted as "++").

<table>
<thead>
<tr>
<th>Requirement</th>
<th>SIVA</th>
<th>SMIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media, Main video, and Annotations</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td>Event-based Timing Model</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td>Temporal Relationships</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>Spatial Relationships</td>
<td>++</td>
<td>0</td>
</tr>
<tr>
<td>Decision Elements</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>++</td>
<td>0</td>
</tr>
<tr>
<td>Keyword reference list</td>
<td>++</td>
<td>- -</td>
</tr>
<tr>
<td>Extensibility</td>
<td>++</td>
<td>++</td>
</tr>
</tbody>
</table>

The feasibility analysis shows, that the **SIVA XML** is very well adapted to the requirements of an interactive non-linear video. **SMIL** is able to realize many of the requirements or more precisely the needed features as well, but it lacks in certain details. As table 1 demonstrates, in terms of temporal relationships or the extensibility, both languages are suited very well. The lack of a keyword feature sets SMIL back in that requirement, while facets like the problems in the spatial relationships force an inferior evaluation compared to the SIVA XML in other categories.

**COMPARISON/METRICS**

In order to have a closer look and a numerical comparison of the two metadata formats, we will make use of the following XML metrics: Size, Structure Complexity, Depth, Fan-In and Fan-Out. For a detailed description see [8].

With these five values for a XML description format you are able to make statements about the complexity, comprehensibility, reusability, and convertibility of it. For example the higher one of the metric values is, the more complex is the possible resulting XML file. A high Fan-Out value makes it harder to alter a format because changes in single entities or elements may have an impact on multiple locations in the file. We have calculated the results for both, the SIVA XML and SMIL, to be able to compare them. For the first one, the SIVA XSD was converted into a DTD to be able to compare it with the SMIL DTD. Knowing that such a conversion will usually have an impact on the accurateness of the file, it does
not have an impact on the evaluated metrics. The DTD for SMIL can be found online as well [1]. As it contains the elements for the whole language, we have created a profile to model a DTD that only supports the modules and elements needed in the export for interactive non-linear videos. SMIL was also contemplated in two different ways: with and without the \textless metadata\textgreater element, which has a great impact on the Fan-Out value. This is caused by the fact that the element can become a child node of every SMIL 3.0 element. The use of this element is the wrapping of structured meta information. It contains an own XML tree as content and is not processed by the player at all. As we do not make use of this element in our export, it is still contained in the generated DTD of our interactive video profile. Therefore we differentiate between an analysis of SMIL with and without the \textless metadata\textgreater element. Our results are presented in table 2.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|}
\hline
  & Size & SC & Depth & Fan-In & Fan-Out \\
\hline
  SIVA XML & 58 & 67 & 5 & 12 & 8 \\
  (w/o meta) & 40 & 430 & $\infty$ & 21 & 16 \\
  SMIL & 41 & 507 & $\infty$ & 22 & 38 \\
\hline
\end{tabular}
\caption{Comparison of SMIL and the SIVA XML (SC = Structure Complexity, $\infty$ = unbounded)}
\end{table}

Regarding the first two entries in table 2, one can see that SMIL gets by with less elements than the SIVA XML, but its complexity is much higher. This is caused by the fact that many SMIL elements are used recursively. The high Fan-Out value is applicable for many of the occurring elements. The potential depth for SMIL is unbounded because the temporal elements \textless par\textgreater and \textless seq\textgreater can be boxed repeatedly. In some depth analyses, the recursion is ignored to not achieve a depth that is unbounded. We do not take this into consideration because in fact there can be an apparent endless potential depth by nesting forks. The Fan-In and Fan-Out metrics also state higher values in the SMIL-DTD and therefore indicate superior complexity.

\textbf{PROPOSAL FOR EXTENSION}

The extensibility of SMIL allows the addition of new elements that could be used to generate a module that is more adapted and able to provide support for interactive non-linear videos. The ideas for these elements arose from the problems that were encountered while modeling the elements for interactive non-linear videos in SMIL. In combination with these elements, SMIL could achieve a more dynamical structure and be more feasible referring to interactive non-linear videos. Possible useful additions might be the following.

\textbf{Jumps in the XML File}

The elements \textless goto\textgreater and \textless end\textgreater may change the flow of the SMIL presentation by jumping to another position of the same document when reached. While the former cause a jump to a given ID supported by a \texttt{to} attribute, the \textless end\textgreater element is supposed to bring the presentation to an end. They can be used like most of the timing elements in SMIL in terms of nesting as well as their attributes. Caution has to be paid, because loops and abrupt endings of the presentation can be built very easily.

\textbf{Choices at Forks}

To satisfy the requirement for fork elements more easily, these two additions could help. When a \textless fork\textgreater element is started, it composes a standard choice panel (that could be altered by its attributes in terms of shape etc.) which contains buttons to start one of its \textless choice\textgreater children nodes. One of these contains elements that are supposed to be played once its path has been chosen. In combination with the preceding elements \textless jump\textgreater or \textless end\textgreater, different continuations after a path can be established. This could also be done by a specific attribute for the \textless choice\textgreater element. If neither of these possibilities is supported, the \textless choice\textgreater as well as its corresponding \textless fork\textgreater element will be ended and the succeeding one will be started. Listing 1 shows an example for the use of \textless fork\textgreater and \textless choice\textgreater elements combined with the above mentioned jump elements \textless goto\textgreater and \textless end\textgreater. The decision ranges from line 4 to 21 and contains three different choices. By selecting one of them, the code inside the \textless choice\textgreater element will be started. The button panel itself is designed by the attributes to show round buttons with a size of 20 pixels. For 30 seconds the panel will be displayed. If no path is chosen in that timespan, the defined default path (in this example \texttt{path1}) will be played. The second path from line 9 to 14 contains a \textless goto\textgreater element in line 12 inside a sequential container with the effect, that after the other elements in the row, a jump from line 12 to line 2 (according to the given to element). The third path in lines 15 to 20 consists of a parallel node containing some code and then a \textless end\textgreater element in line 18, which also owns a \texttt{begin} attribute with the value \texttt{20s}. This structure of elements induces the behavior, that no matter what the content of the \textless par\textgreater element is, the \textless end\textgreater element will be started after twenty seconds, which causes the presentation to terminate.

\begin{lstlisting}[language=XML]
<body>
  <seq xmlns:id="start"/>
  <!-- Any SMIL content -->
  <fork shape="circle" size="20"
    region="main_region"
    dur="30s" defaultPath="path1"
    xml:id="fork">
    <choice xmlns:id="path1" after="#fork">
      <!-- Any SMIL content -->
      </choice>
    <choice xmlns:id="path2">
      <seq/>
      <!-- Any SMIL content -->
      <goto to="#start"/>
      </seq>
      <choice>
        <choice xmlns:id="path3">
          <par/>
          <!-- Any SMIL content -->
          <end begin="20s"/>
        </par>
      </choice>
    </choice>
  </fork>
</body>
\end{lstlisting}
The SIV A Suite

linear videos. Therefore, SMIL could be ex-

A new set of extensions would make SMIL more usable for

A further benefit of the SIVA XML is revealed by the analysis of the underlying DTDs of both

The research we did here was based on the standard SMIL 3.0

An important detail to mention is that SMIL is not meant or

The SIVA Suite\textsuperscript{6} contains a production software (SIVA Producer) that enables the user to design and create interactive

The SIVA Suite\textsuperscript{6} contains a production software (SIVA Producer) that enables the user to design and create interactive

REFERENCES

1. SMIL 3.0 DTDs.


\textsuperscript{6}The SIVA Suite can be downloaded from: \textit{http://www.siva.uni-passau.de/?q=node/66}