Multimedia Data Integration on the Semantic Web

Mark Flores

Abstract

This paper will review the state-of-the-art for multimedia integration into the Semantic Web by reviewing and evaluating leading technologies SMIL, SVG, VoiceXML and MPEG-7. An assessment of these technologies will be done based on Semantic Capability, User Friendliness, and Cross Platform Compatibility. This paper provides some insight into which approach is currently leading the way for integrating multimedia data into the Semantic Web as well as identifying areas that require further development.

1. Introduction

With the Internet’s massive growth, the retrieval of specific information has become difficult with current web searches and queries often returning out of date material or unrelated topics all together. One goal of the Semantic Web is to solve these issues and to enhance problem-solving on the Internet to another level. However, most if not all of the Semantic Web development completed to date has focused on text-based web content. Semantic Web support for multimedia-based content is only just emerging as an area of investigation. Multimedia data, such as images, 3D models, audio, and video will play a very important role in the Semantic Web, and there is a growing need to have multimedia data machine processable, similar to what Semantic Web applications are currently doing with textual based data [5]. Support for multimedia on the Semantic Web will allow some degree of machine interpretation of audiovisual information. There are several languages/protocols that show potential for attaining this goal. In this paper, we will review the current state-of-the-art for multimedia on the Semantic Web today and perform an assessment of the technologies to determine their potential for leading the way in multimedia semantics.

2. The Semantic Web

The “Text Based” Internet has been described as entering into its third generation of advancement, the Semantic Web [7]. The Semantic Web is described as an extension of the current web in which information is well defined to better enable people and computers to cooperate together. The Semantic Web is based on the idea that web content should be “machine readable” to better retrieve and automate various services to bring the Internet to its full potential. The Semantic Web will establish an infrastructure that will enable web pages, databases, PDAs etc. to all work together to process or generate data on the web. Semantic Web Agents are the starting point/portal for all users which both interface with the Semantic Web but also stores personal information and preferences locally. The World Wide Web Consortium is currently hard at work establishing an open source standard for the community [13].

While rapid progress is being achieved in incorporating Semantic Web standards for text-based information, with the recent release of the Resource Description Framework (RDF) and OWL Web Ontology Language on February 10, 2004 by W3C (World Wide Web Consortium, w3.org) [13], there has been far less progress in developing Semantic Web for multimedia data. There in lies the need to establish a method for extracting “machine processable” semantics from multimedia data. For instance, an ideal query on the Semantic Web would have both a Travel Agency’s Semantic Web Agent and a Tourist’s Semantic Web Agent retrieve the same Canyon picture without relying on the picture to be labeled, “Grand Canyon”.

In a 2002 W3C presentation, Charles McCathieNevile gives a discussion entitled “Semantic Web and Multimedia” [6], which was essentially a high level discussion on the status of multimedia incorporation into the Semantic Web. SMIL, SVG and VoiceXML (all W3C approved standards) were briefly discussed as emerging technologies to aid in multimedia semantics, along with XML and RDF being the foundations for Semantic Web. This W3C approved presentation remains the “latest” overview discussion given and provides the basis for the investigation discussed in this paper. In addition to SMIL, SVG and VoiceXML, MPEG-7 was also included due to its highly utilized MPEG format in the Internet today. We provide an overview of these technologies in the next section.

3. Overview of Multi-Media Technologies

Currently accepted W3C standards, SVG, SMIL and Voice will help integrate various multimedia formats into the web via XML and RDF. Another developing technology, MPEG-7, also utilizes XML to annotate multimedia files with text descriptions. Each of the four technologies has the ability to extract or annotate information from various multimedia formats which will aid in the next generation, multimedia rich, Semantic Web. In this section, we review the four technologies to be assessed.

3.1 SVG (Scalable Vector Graphics)

Created by the W3C in 1999, SVG is a language for generating 2-dimensional graphical applications, similar to Macromedia FLASH™, via XML [11]. SVG 1.1 is currently a W3C Recommendation and is the basis for SVG development. SVG 1.2 is currently in development. There is also an SVG format for PDAs, or resource limited devices, which are SVG Basic and SVG Tiny. There is also an SVG Print protocol to produce final-form documents in XML for archiving and printing.

SVG has many advantages over other image formats, such as JPEG and GIF, which are [10]:

- **Scalable** – SVG is a vector format, which means unlike GIF or JPEG, images can be printed with high quality and at any resolution distortion.
- **Zoomable** – Straightforward, you can zoom in and out of images without degradation.
• **Searchable and Selectable Text** – Unlike bitmapped and JPEG images, text in SVG remain selectable and searchable. From a Semantic Web standpoint, now a portion of the SVG image is “machine readable”!

• **Scripting and Animation** – SVG enables a dynamic and interactive graphical display that is more sophisticated than a static images or even FLASH images.

• **Works with Java Technology** – SVG is compatible with JAVA’s high end graphic engine, Java 2D API.

• **Open Standard** – Perhaps one of the more important features, SVG is an open recommendation developed by a cross-industry consortium and is NOT proprietary.

SVG is an XML based language which means SVG graphics can easily be generated by servers today. With SVG, you can easily manipulate your graphics using standard XML tools. Below is a simple example of the SVG code, which draws a rectangle with the famous phrase, “Hello World” superimposed.

![SVG code example](image)

**Figure 1. Sample code SVG.**

SVG can generate high quality, low bandwidth graphics. SVG is also interoperable with existing JAVA applications/scripts with the capability of importing and exporting data back and forth with a Graphic 2D SVG Generator.

### 3.2. SMIL – Synchronized Multimedia Integration Language

Web multimedia has long been accessible via proprietary formats or JAVA programs for some time now and has been largely inaccessible to most web authors. With SMIL, an HTML-like language, it will once again make web multimedia readily accessible to the “common web author” and have an impact similar to the impact HTML had for general text [9]. The W3C has been developing SMIL since 1997 as a language for providing synchronized presentations where audio, video, text and graphics are combined in real-time. The following SMIL example shows the simple HTML-like language which pulls up two images and cycles between the two with a three second delay [12].

![SMIL code sample](image)

**Figure 2. SMIL code sample, W3C [12].**

SMIL is written in XML and is typically used for multimedia rich presentations where streaming audio and video are integrated and synchronized with images, text or any other media type. Since it is an HTML-like language, many SMIL languages are written in a simple text editor.

### 3.3. VoiceXML – Voice, Voice Browser

Another W3C approved language, VoiceXML 2.0 is currently a “Proposed Recommendation”, submitted in February 2004 [8]. VoiceXML is being pursued because the W3C is looking to expand the access of the web to allow people to interact via keypads, spoken commands, listening to prerecorded speeches, synthetic speech and music. VoiceXML would enable the telephone to access web pages and be an aid for people with disabilities [8].

VoiceXML is a web-based markup language for representing human-computer dialog, just like HTML. Where HTML assumes a graphical web browser, computer monitor, keyboard and mouse, VoiceXML uses a voice browser, with audio clips (computer synthesized or pre-recorded), and audio input (voice or phone keypad tones) [8]. Figure 3 shows VoiceXML interface and sample code with the equivalent “old” HTML way/interface on the right. Note that instead of a computer being needed to access the web page, a telephone is used instead. Both methods output the message “Have a nice day.” to the user.

Further work is anticipated in enabling use with other W3C markup languages such as XHTML, XForms and SMIL. All of this work will be done in conjunction with the other working groups.
3.4. MPEG-7 – Multimedia Content Description Interface

MPEG-7 is a multimedia content description standard, which focuses on how humans will interact with computer systems in the next generation with its rich descriptions that reflect Semantic Web expectation [2]. MPEG-7 is a standard that describes multimedia content so users can search, browse, and retrieve content more efficiently than they could with today’s mainly text based search engines.

The Moving Picture Coding Experts Group (MPEG), a group of the Geneva-based ISO/IEC standards organization, is developing MPEG-7. MPEG’s charter is the development of international standards for compression, decompression, processing, and coded representation of moving pictures, audio and a combination of both. MPEG is also the same committee that developed the Emmy Award-winning standards known as MPEG-1 and MPEG-2, and the 1999 MPEG-4 standard.

MPEG-7 will address a wide variety of media types including: still pictures, graphics, 3D models, audio, speech, video, and combinations of these. It should be noted however, that MPEG-7 will NOT automatically extract multimedia descriptions/information, it will instead rely on the web author to append the descriptions/information to the multimedia file. MPEG-7 working group also makes it clear that their standardization will not specify which search engines can make use of the MPEG-7 descriptions. The MPEG group will leave this portion up to the Search Engine companies to integrate MPEG-7 into their individual systems [2], a deliberate strategy which has historically served them well.

MPEG-7 uses XML schema as the language of choice for content description. MPEG-7 will therefore be interoperable with other leading standards. Although MPEG is not a W3C developed standard, with its XML based approach, it should also be compatible with most W3C XML based languages.

Now that we have reviewed the leading technologies for integrating multimedia into the Semantic Web, we will now establish the criteria (Semantic Capability, User Friendliness, and Cross Platform Compatibility) which will be used to assess each approach.

4. Evaluation

In this section, we describe the criteria to be used to evaluate the capability of each technology to integrate multimedia into the Semantic Web. We also provide a general formula that combines weighted values for each of the criteria.

In determining the criteria, for the four approaches to integrate multimedia into the Semantic Web, we will use a modified ISO/IEC standard for Software Evaluation, ISO/IEC 9126-1. Modification was necessary since all of the languages to be evaluated are still in development phase and are not mature enough to evaluate based upon all the attributes [1]. We have defined evaluation criteria based on the ISO/IEC 9126-1:

- Semantic Capability: We define this criterion as the Suitability and Interoperability portions of the Functionality portion of the ISO/IEC 9126-1 standard. Suitability and Interoperability will show the amount of multimedia manipulation/recognition the language is performing with respect to multimedia data, while determining the complexity in the multimedia integration. Note that the Accuracy and Security portions are not included in the Semantic Capability criterion.
  - User Friendliness: The Usability characteristic of the ISO/IEC 9126-1 standard will be assessed as User Friendliness and all ISO/IEC attributes will be covered. User Friendliness is usually a reasonably good indication of how well the technology would be adopted if it is simple and straightforward as HTML is.
  - Cross-Platform Compatibility: This criterion is related to the ISO/IEC 9126-1 standard characteristic of Portability – which includes aspects of Adaptability, Installability, Conformance, Replaceability. This attribute WILL be assessed as “Cross Platform Compatibility”. All attributes except “Replaceability” and “Conformance” will be assessed.

Note: Reliability, Efficiency, and Maintainability characteristics, from the ISO/IEC 9126-1 standard, are not included as criteria in our assessment since they are “not applicable”. The reason for non-applicability is that SVG, SMIL, VoiceXML and MPEG-7 still remain in the developmental phase of software development and thus has not accumulated available data to review.

Now that we have established the criteria to be used in the evaluation, we will now discuss each criterion in further detail and show how points will be assessed.

4.1. Semantic Capability

Semantic Capability addresses the degree at which the multimedia data is “machine processable”. The investigation of Semantic Capability was the primary motivation for our work. This assessment, based upon this criterion established above, provides a clear picture of the status of various technologies with respect to their ability to support multimedia data integration in the Semantic Web. Since Semantic Capability is the main attribute we are interested in, it was weighted the highest with a maximum of five category points (Poor, Fair, Good, Very Good, Excellent) to be assessed in our assessment formula. All other attributes to be evaluated will have a maximum four category points (Poor, Fair, Good, Very Good) assessment value. The category evaluation is determined as such:

- Poor 1pt – Maximum of 1 to 2 lines of text can be directly annotated to multimedia file/data. This level of multimedia description annotation exists in HTML and XML today. Therefore, languages that demonstrate capabilities equal to current standard will receive a score of Poor.
- Fair 2pts – Greater then 3 lines of text/description can be annotated to multimedia file/data.
- Good 3pts – Audiovisual manipulation/presentation while maintaining smaller multimedia data components (i.e. Text over images superimposed or a presentation that pulls in multiple subcomponents).
5. Results

In this section, we list the summarized data results from the evaluation, Table 1, and provide an overview of the results with some insight on how the points were assessed for the various protocols/languages. We will also review and discuss the overall results.

Displayed in Table 1 is the summarized “raw data” from the evaluation. Table 1 shows the total points each leading technology received, point-to-rating definition, and “total scores” of the evaluation. Sections 5.1 through 5.4 will discuss each criterion’s results further.

<table>
<thead>
<tr>
<th>Semantic Capability</th>
<th>SVG</th>
<th>SMIL</th>
<th>VoiceXML</th>
<th>MPEG-7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>User Friendliness</td>
<td>Total</td>
<td>3</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td>Cross Platform Compatibility</td>
<td>Total</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

5.1 Semantic Capability Results

As seen in Chart 1 below, VoiceXML has scored the highest in this criterion. VoiceXML elevates sophistication and complexity to another level with speech recognition capabilities. VoiceXML shows signs of true multimedia semantics by the “machine’s” capability to recognize incoming audio data. VoiceXML completely replaces textual input and display with speech recognition and audio playback.

SVG and SMIL evaluated the same due to their ability to manipulate multimedia data while still maintaining sub-component multimedia data/definition. SVG additionally has the ability to incorporate user interaction, zoom, animation replay and still allow text to be selectable. Where SMIL has the ability to synchronize various multimedia data and formats into a single presentation, along with also allowing text to be selectable.

MPEG-7 scored the least with a rating of Fair having capabilities just slightly better than HTML. MPEG-7 uses the current method of multimedia description by annotating the descriptions to the multimedia file.
However, it can annotate well more than 2 lines of description to obtain a rating of Fair.

In summary, VoiceXML received the highest marks with 4 points (Very Good) due to its capabilities of speech recognition and audio interaction. SVG and SMIL came in second with their capability to manipulate multimedia files into larger presentations/visual display. MPEG-7 scored the lowest due to requiring the user to manually input all multimedia descriptions upfront.

5.2 User Friendliness Results

For User Friendliness, all of the languages received points for being an XML-based language and thus qualified for a minimum rating of Good by being compatible with the latest web browsers and able to be created within existing Internet infrastructure. However, VoiceXML was deducted a half a point because it can only be partially created via XML. VoiceXML also requires either speech synthesis or pre-recorded audio to be created and uploaded. In terms of server side support, all of the languages require support except for MPEG-7, which is completely XML-based in creation and retrieval, and thus scored a rating of Very Good, four points. Reference Chart 2 for the results of the User Friendliness evaluation.

The summary evaluation of the User Friendliness criterion is that MPEG-7 has the highest score due to its simplicity, which is undoubtedly related to its limited multimedia semantic capability aside from user input upfront. SMIL and SVG received a rating of Good for being able to work within the current Internet infrastructure. VoiceXML, which originally qualified for a rating of Good, received a slight deduction due to increased complexity since it is more geared towards being a call center with multiple phone lines or for the visually impaired and thus requiring additional server side support (hardware and software).

5.3 Cross Platform Compatibility Results

For Cross Platform Compatibility, all the XML based languages qualified for a minimum of a rating of Good, which means they are compatible with most operating systems. However, in terms of PDA compatibility, SVG showed the most extensive strength in ensuring compatibility with all popular PDAs available today. Therefore SVG received a rating of Very Good, four points. SMIL and VoiceXML both did not indicate PDA support, although PDA browser companies are actively looking to incorporate SMIL into their next release.

MPEG-7 also received a rating of Very Good for PDA support as well since it is again strictly an XML-based protocol and compatible across the board today. MPEG-7 was even described as being able to increase the performance of PDA Internet browsing by being able to retrieve low resolution images for incoming PDA requests on the fly. Reference Chart 3 for the results of the Cross Platform Compatibility evaluation.

5.4 Overall Results

Reviewing Chart 4, we see VoiceXML lost most of its points when it came to User Friendliness and Cross Platform Compatibility. This is understandable since VoiceXML is currently a very sophisticated technology that requires considerable server side support in hardware (multiple phone lines) and software (speech recognition). Mainly large corporations, such as Cingular, Verizon Wireless and IBM, use VoiceXML today for automated call centers.

The leaders of the Overall Evaluation were SVG and MPEG-7, as shown in Chart 4. Chart 5 is another illustration which shows how MPEG-7 scored the least in Semantic Capability but made up significant ground in User Friendliness and Cross Platform Compatibility. Where technologies such as SMIL ad VoiceXML showed more potential than MPEG-7 in Semantic Capability but lost ground in the other two categories to rank below MPEG-7 overall.
Multimedia subcomponents. Presentations on the fly while still maintaining the presentations on the fly while still maintaining the can only draw from extract semantics from multimedia files. SVG and SMIL lost some points since they do not possess the ability to undoubtedly play a vital role in the “multimedia rich” undoubtedly play a vital role in the “multimedia rich” can manipulate various multimedia formats into a single display for user interfaces and/or presentations and will undoubtedly play a vital role in the “multimedia rich” Semantic Web of the future. However in this evaluation, lost some points since they do not possess the ability to extract semantics from multimedia files. SVG and SMIL can only draw from digital archives (a database or archive of various multimedia formats) and make animations or presentations on the fly while still maintaining the multimedia subcomponents.

6. Conclusion

In summary, we have evaluated Semantic Capability of the SVG, SMIL, VoiceXML, and MPEG-7, and found that VoiceXML has rated the highest with SVG and SMIL coming in second. We next looked at User Friendliness where the results were the opposite with MPEG-7 scoring the highest with its simplistic XML based language and VoiceXML rating the least due to its higher level of complexity in incorporation. Finally we reviewed Cross Platform Compatibility where SVG and MPEG-7 rated very high due to high compatibility and support with operating systems and PDA devices, with SMIL and VoiceXML both rating Fair.

Each of the four approaches has various strong points however VoiceXML shows the most potential of true multimedia semantics, with its speech recognition capability. MPEG-7 ranked the highest overall with its simplistic XML based application and ability to be applied today without impacting current infrastructure. SMIL and SVG will also mostly likely play key roles in the Multimedia Semantic Web in responding to user requests by generating multimedia rich display/interfaces on the fly. Multimedia Semantic Web is rapidly approaching the widespread adoption. SVG, SMIL, VoiceXML and MPEG-7 will help further develop the Semantic Web into a fourth generation, which includes multimedia semantics via multimedia data recognition and processing.

References


