THE RISE AND FALL OF MULTIMEDIA AUTHORING

Lloyd Rutledge and Lynda Hardman
INS2: Multimedia and Human-Computer Interaction Theme
CWI (Centrum voor Wiskunde en Informatica)
P.O. Box 94079
NL-1090 GB Amsterdam, The Netherlands
{FirstName.LastName}@cwi.nl

ABSTRACT

For much the past decade, our research group has focused on developing multimedia authoring tools. We developed a research multimedia format and were involved in developing a multimedia standard. Our lab developed a research prototype of a multimedia editor and then spun off a company that turned it into a commercial editor for this standard. With several research concepts now having reached full maturity in the form of this standard and product, our lab is back where it started: at the beginning of the research idea life cycles, with new visions and hopes for what's next.

Keywords: Multimedia, Multimedia Authoring, Structured Multimedia, SMIL

1. INTRODUCTION

Our journey began with the multimedia format CMIF (CWI Multimedia Interchange Format) and an editor for creating presentations in it called CMIFed (CMIF editor) [5]. Both were research prototypes that started simple and were added to piece by piece, idea by idea. They exercised such concepts as linking in context, user-centered adaptivity and the use of temporal hierarchies in the authoring interface.

After several years, these ideas seemed ready for standardization. We joined the W3C's SYMM (Synchronized Multimedia) Working Group, and SYMM developed SMIL (Synchronized Multimedia Integration Language) [11], the W3C's recommendation for multimedia on the Web. Key concepts from CMIF emerged in SMIL as well, and were introduced to the world.

With the release of SMIL and the subsequent presence of standardized multimedia on the Web, it was certain that a market would develop for playing and creating this multimedia. Thus, our research prototype CMIFed became a potentially profitable commodity. To test these research ideas on the market, our lab spun off the company Oratrix, and Oratrix turned CMIFed into GRiNS (GRaphical Interface to SMIL) [1], a GUI editor for putting large-scale SMIL presentations on the Web. Thus, our research ideas reached the apex of their existence, having been implemented into a commercial product.

But, of course, research begets more research, and multimedia authoring research is starting a new wave. Having established that authors can directly craft complex multimedia presentations, new ideas of what goes into making multimedia presentations are being explored. These include further fragmentation of media objects, more automatization in generating multimedia presentations, and the use of semantics in this process. Some of these ideas may appear in standards and industry in the coming years.

In this paper, we discuss the life-cycle of these ideas. First, we describe how they emerged and the initial work that was performed on them. After that, we describe our experience with involving these ideas in the standardization process. We then describe our experience with turning these ideas and the emerging standard into a business and a commercial product. In the following section, we discuss the problems that remain as these solutions become adopted. Finally, we describe current research directions that attempt to solve these problems.

2. BIRTH

When "multimedia" presentations first appeared, they were simple composites of multiple visual, and perhaps audio, media objects. When computers began rendering these presentations, the effect was still the same. Visuals and sounds would be played at particular moments. Images and videos would appear at particular locations on the screen. Multimedia editing systems came along, but by offering only flat timelines and simple screen placement, no additional power was given to the presentations they created. The result could be captured on videotape, and playing the tape would be the same as playing the multimedia presentation. Of course, much research began in many groups to progress beyond that.

Starting in early 1991, our CMIFed [5] project was one research exploration increasing the power of the author in creating multimedia presentations. CMIFed was an authoring environment for our multimedia format CMIF, which specifies what media objects are to be used in a presentation, and what structure along which they are to be integrated. CMIFed provided the author with an intricate interface to the *temporal hierarchical structure* with which to organize the authoring process. CMIFed also provided constructs for *adaptivity* to varying user and system characteristics. Finally, CMIFed implemented *linking in context*, enabling only parts of the presentation to change as the user navigates through it.

The CMIFed document structure was based primarily on temporal composition. The two main types of composition were *parallel* and *sequential*. Sub-presentations contained in a parallel composite were played at the same time. Sub-presentations contained in a sequential composite were played one after the other, in order.

Parallel and sequential composites could be contained in each other, making a temporal hierarchy. A composite itself would be considered to end when the last of its children ended, providing an upwards temporal inheritance. At the leaf node level, continuous media objects would end when they were finished playing. With this, intricate synchronization relationships and timelines could be set up without the author entering a single numeric timestamp – all timing information could be inherited up from the natural timing of the media itself. Not only did this provide many timing shortcuts for the author, but it also provided a natural hierarchy with which to organize the document and how it is authored.

CMIFed also provided *adaptivity* in multimedia presentations. Alternatives to certain media components and subpresentations could be established, and in each case the best alternative chosen for given user and presentation environment. The construct for providing this was the *channel*, a virtual device on which media items are played. Each channel could be turned on or off based on player menu settings or user interaction. For example, a CMIFed presentation could have a separate channel for each language of audio. All languages are timed as playing in the temporal hierarchy, but only one language channel is turned on during playing: that of the user. Having channels separate from the temporal hierarchy enables different media objects to have shared presentation properties, such as by having all audios of a given language be assigned the same channel.

Finally, CMIFed made presentations even more versatile with *linking in context*, the ability to have only part of the presentation's appearance and timing change as you navigation through it. For example, with linking in context, you can cause the main part of the presentation to change while a side menubar stays constant, similar to how frames are often used in HTML. Also, this feature allows you to link forward or backward in a presentation while keeping the background music playing without skipping. Linking in context is provided with a third hierarchical temporal composite called the *choice node*. When a sub-presentation contained in a choice node is the destination of user navigation, any other subpresentations in the choice stop, and that one is played from its beginning. Any sub-presentations playing in parallel to the choice node as a whole keep playing uninterrupted.

By developing and promoting the ideas of temporal composition, adaptivity and linking in context, we hoped to help break the confines of the flat timeline "play as video" model and make multimedia more dynamic. But, of course, for these ideas to have an impact, they have to be used in more places than just one research lab. It was time to show them to the world and see how they fared.

3. RISE

In early 1997, the W3C formed the SYMM Working Group to develop a multimedia format for the Web. We joined up, hoping to contribute our ideas in these areas and to learn from the contribution from the other SYMM members. This began a still continuing process of emails, telephone conferences and face-to-face meetings across the world, in which members propose and debate the means of putting multimedia on the Web, and in which each member's dreams and perceptions for the future of multimedia get intertwined with everyone else's. The result of this process is the Synchronized Multimedia Integration Language (SMIL).

SMIL is the W3C language for multimedia on the Web. Version 1.0 of SMIL was released in 1998 with the basic foundation for distributed multimedia [11]. SMIL 2.0 is expected to be released soon as a recommendation [13], defining state-of-the-art Web-based multimedia with many new features such as event-based timing, animation and transitions. Some of our research topics found their equivalents in SMIL when it was released, though typically much changed through discussion, debate and integration with other contributions.

Of these research ideas, CMIFed's temporal composition has the most direct equivalent in SMIL. The main component of the XML-define syntax is a temporal hierarchy, whose root is at the <body> element. It contains an XML content tree consisting mainly of <par> (parallel) and <seq> (sequence) elements. This gives SMIL the same time specification shortcuts that were in CMIFed, and the same authoring structure model. Fellow SYMM members from INRIA made one of their research impacts on SMIL by applying their work on temporal constraints to SMIL timing for referential timing [7].

Adaptivity has received much attention in multimedia research, and so naturally has a place in SMIL. The adaptivity model used in SMIL 1.0 was not much like our channel-based model. Instead, it is selects between different sub-trees of the temporal and content hierarchy.

Adaptivity is provided by SMIL 1.0 with the <switch> element. It appears side-by-side with <par> and <seq>. Among the components of a switch element, only one is chosen. The browser goes through its child elements one at a time and selects the first, if any, it determines is appropriate for including in the presentation. In a multi-lingual presentation, for example, a clip of speech in each language is put within a switch, and the one whose language matches that of the speaker is selected.

Our lab argued later that channel-based adaptivity had some advantages over temporal- and content-based adaptivity [2]. However, it is more complex to author and thus arguably not suited for the intentionally introductory SMIL 1.0. SMIL 2.0 adds behavior similar to CMIFed's channel selection by letting <switch> appear with layout-defining constructs, thus allowing adaptation to choose between different layout designs as well as different content and temporal sub-trees.

Finally, linking in context is provided with SMIL 2.0's <excl> element, offering a close equivalent to CMIFed's choice node. As with CMIFed's choice, the <excl> element makes sure that only one of its children sub-presentations is active at any one time. The <excl> element goes further, however, allowing timed events as well as user interaction to activate its children, and providing many alternatives for how the timing within activated children is handled.

During the many frequent changes and additions that entered SMIL as the members made their contributions, we kept modifying CMIFed to keep it in line with SMIL and produce SMIL output from its authoring interface. The result was a multimedia standard created from a wide variety of influences and perspectives, and a prototype tool that could play and produce it. As SMIL came closer to its version 1.0 release, and received a lot of attention, it turned into a viable product as well, for our lab and for others.

4. GLORY

In January of 1999, our institute, CWI, spun off a company named Oratrix for the purpose of commercializing CMIFed. The result is a product called GRiNS, which provides an extensive graphic user interface for creating and maintaining SMIL presentations of all sizes. It is essentially a polished-up and more robust update of CMIFed, with a sleeker interface for the general public, and heavily debugged for publicly-acceptable performance stability. Beta versions were available since the standard was released, and the 1.0 version of the GRiNS player and editor was released in January 2000.

Oratrix and GRiNS have had good company. From the release of SMIL 1.0, SYMM partner RealNetworks has had a SMIL 1.0 player available, and millions of copies of it have been downloaded. Several other non-profit SMIL 1.0 players were also released soon after the standard was. The multimedia authoring ideas of us and our SYMM partners were being adopted more and more widely across the Web.

The development of SMIL 2.0 has spurred on further industrial involvement. Microsoft implemented a prototype of one of SMIL 2.0 child languages, XHTML+SMIL, in Internet Explorer 5.5. Oratrix continued updating GRiNS to keep up with the growing SMIL 2.0. A beta-release of the GRiNS player was made with the release of SMIL 2.0 last call draft that played that version of SMIL. Oratrix plans to release a SMIL 2.0 version of the GRiNS player and editor on or near the release date of SMIL 2.0 itself. RealNetworks has continued heavy involvement in the development of SMIL 2.0, as have other major multimedia companies.

In the GRiNS interface, our research topics took on more life beyond that which SMIL gave them. The structural view provides a zooming two-dimensional graphic interface of the temporal composition, vizualizing the <par> and <seq> elements' impact on the presentation's timing. This same view shows the selective <switch> and <excl> elements, and their placement of alternative parallel mutually exclusive possibilities within the temporal structure. This provides the author much power in perceiving and manipulating all these aspects of temporal structure on a large scale. Other research has also had an impact on GRiNS interface to timing – GRiNS's timeline view uses interface techniques that were explored in the Madeus project of SYMM partner INRIA [7].

The GRiNS interface also adds the user group facility from its pulldown menu to further enhance the authoring of adaptivity. It assists the SMIL 2.0 constructs in setting up different user types and controlling how presentations adapt to these types. CMIFed adaptivity features that were not used in SMIL were able to still apply, but providing the user with an interface modeled on those ideas that translated those ideas to their SMIL equivalents. Another adaptivity-related GRiNS interface is its network traffic emulator. With this, the author can understand how a presentation's behavior varies with connectivity to media servers, and thus be better able to write a SMIL presentation that adapts well to varying bandwidth.

SMIL 2.0 and the products for it represent the current state-of-the-art for authoring the integration of media objects into adaptive and dynamic multimedia presentations. But of course, completing a task also provides insight into what was left undone, and what is needed to do next. Is there more to making multimedia than taking pre-selected media components and stringing them together (even with very complex string)?

5. FALL

With all the editing power of products like GRiNS and the representative power of formats like SMIL, some aspects of multimedia authoring remain difficult. While SMIL provides extensive tools for integrating media files and streams, few features are available for integrating fragments of these files and streams – they are typically taken in their entirety. Also missed in SMIL is that while it provides much power in integrating media objects, it is still takes much authoring effort to find the right media to integrate. And finally, there's the simple consideration that now that the multimedia authoring process has been formalized enough to form a format and editor interface, how much of it can be automated to further minimize human author effort? As authors grow accustomed to the presentation facilities of SMIL and the authoring facilities of editors like GRiNS, these unsolved problems will become more and more apparent.

6. REBIRTH

Fortunately, multimedia research is under way to provide solutions to these problems. The dividing up of media into fragments that can be used independently is provided by several emerging standards. Scalable Vector Graphics (SVG) is a W3C recommendation for encoding graphics in XML [4]. Having graphic display components structured in XML makes it easy for XML tools to break the display up along the lines of this structure. Referring to XML-defined components of SVG, and other XML documents, is provided by the emerging W3C format XPointer [3]. Once these are implemented, SMIL can refer to XPointer-defined portions of SVG graphics for integration.

Another emerging standard for fragmentation is MPEG-7 [6]. MPEG-7 defines annotations for continuous media in general. These annotations can split the media into fragments that can be referenced and located. As MPEG-7 gets implemented and used, it could potentially be used with SMIL 2.0 constructs like media markers to further the authoring of media fragments into SMIL presentations.

There is much active research on how to find the right media to use in a presentation. Much of it centers on the indexing of large collections of media items so that an author can find the one item among them best suited for a task. One example is the Acoi system being developed by our research neighbors at CWI [14]. Such systems would enable the author to enter a query describing what is being sought for in the desired media item, and then return the media item or items that best matches that query, greatly facilitating the media collection process that must precede the integration enabled by SMIL. However, media indexing is a very complex problem, and much work remains to be done before such systems become readily usable on a large scale.

One key component being explored for media indexing and retrieval is the use of semantics. This would enable authors to define a search with one or more keywords, and then find matches on media items that are annotated in ways that semantically match these keywords. This recent bridge between the artificial intelligence and Web communities has resulted in the W3C recommendation Resource Description Framework (RDF) [12], which provides an initial foundation for approaching solving this problem. The recently started standards effort DAML+OIL seeks to build up top of RDF a full ontology-based solution of both standards and tools for putting semantics on the Web [9]. We describe this problem and potential solutions for it in other work [10].

The final question when making authoring easier is wondering how to remove the need for human authoring effort altogether. Of course, some degree of human involvement will be needed for quite some time in making intelligible multimedia presentations. But research is being performed now in making the process at least semi-automatic. With media indexing and retrieval, you can enter a query and get back an appropriate media object for it. The next question is: can you specify a multimedia presentation you'd wish to see, and then not just have all the relevant media automatically fetched but also integrated and structured into a SMIL presentation on the topic? We have begun work in this area, trying to determine what kinds of input parameters would be meaningful, and how they can be interpreted into the intricate multimedia structure of the type defined by SMIL [8].

This work also involves how to generate presentations of this content with sensible rhetorical structure and narrative flow.

These research and standards efforts are building a Web-based infrastructure in which authors create media, annotations and meta-data instead of final presentations. Furthermore, end users of this vision can specify in more detail what presentation they wish to see and have it generated for them. The most appropriate media content available, down to the fine level of fragment-defined detail, will be retrieved and composed. And it will be integrated into a sensible presentation.

7. CONCLUSION

Research and development follows a life cycle of initial conception and research, modeling into standards and development into tools. Once this is accomplished, then the tools' functionality becomes commonplace enough that the problems left unsolved become obvious, and their solutions begin to be researched. The authoring of multimedia has just reached the end of a cycle. Formats and tools with which authors can integrate media components into complex presentations have now been developed. These leaves us with the problems of how better to acquire large amounts of media to put in these presentations, and how much further the human process of integrating these can be automatically facilitated. Those involved now in multimedia authoring have opportunities to help bring the standards and tools that will make this next wave possible, similar to how the authors were involved in the last wave of multimedia authoring standards and tools as described herein.

8. REFERENCES

[1] D. Bulterman, L. Hardman, J. Jansen, K. Mullender, and L. Rutledge. GRiNS: A GRaphical INterface for Creating and Playing SMIL Documents. In *Seventh International World Wide Web Conference*, Brisbane, Australia, April 14-18, 1998.

[2] D. C. Bulterman. User-Centered Abstractions for Adaptive Hypermedia Presentations. In *Proceedings of ACM Multimedia*, pages 145–150. ACM Press, November 1998.

[3] S. DeRose, E. Maler, and J. Ron Daniel. XML Pointer Language (XPointer) Version 1.0. W3C Candidate Recommendations are available at http://www.w3.org/TR, 8 January 2001.

[4] J. Ferraiolo. Scalable Vector Graphics (SVG) 1.0 Specification. W3C Candidate Recommendations are available at http://www.w3.org/TR, 2 November 2000.

[5] L. Hardman. *Modelling and Authoring Hypermedia Documents.* PhD thesis, University of Amsterdam, 1998. ISBN: 90-74795-93-5, also available at http://www.cwi.nl/∼lynda/thesis/.

[6] International Organization for Standardization/International Electrotechnical Commission. MPEG-7: Context and Objectives, 1998. Work in progress.

[7] M. Jourdan, N. Layaïda, C. Roisin, L. Sabry-Ismaïl, and L. Tardif. Madeus, an Authoring Environment for Interactive Multimedia Documents. In *Proceedings of ACM Multimedia '98*, Bristol UK, 1998.

[8] L. Rutledge, J. Davis, J. van Ossenbruggen, and L. Hardman. Inter-dimensional Hypermedia Communicative Devices for Rhetorical Structure. In *Proceedings of International Conference on Multimedia Modeling 2000 (MMM00)*, pages 89–105, Nagano, Japan, November 13-15, 2000.

[9] F. van Harmelen and I. Horrocks. Reference description of the DAML+OIL ontology markup language. http://www.daml.org/2000/12/reference.html. Contributors: Tim Berners-Lee, Dan Brickley, Dan Connolly, Mike Dean, Stefan Decker, Pat Hayes, Jeff Heflin, Jim Hendler, Deb McGuinness, Lynn Andrea Stein.

[10] J. van Ossenbruggen, F. Cornelissen, J. Geurts, L. Rutledge, and L. Hardman. Towards Second and Third Generation Web-Based Multimedia. In *The Tenth International World Wide Web Conference*, Hong Kong, May 1-5, 2001. To be published. This is a revised version of CWI technical report INS-R0025.

[11] W3C. Synchronized Multimedia Integration Language (SMIL) 1.0 Specification. W3C Recommendations are available at http://www.w3.org/TR/, June 15, 1998. Edited by Philipp Hoschka.

[12] W3C. Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendations are available at http://www.w3.org/TR, February, 22, 1999. Editied by Ora Lassila and Ralph R. Swick.

[13] W3C. Synchronized Multimedia Integration Language (SMIL) 2.0 Specification. Work in progress. W3C Working Drafts are available at http://www.w3.org/TR, 1 March 2001. Edited by Aaron Cohen.

[14] M. Windhouwer, A. Schmidt, and M. L. Kersten. Acoi: A system for Indexing Multimedia Objects. In *International Workshop on Information Integration and Web-based Applications & Services*, Yogyakarta, Indonesia, November 1999.