3. codecs and standards

without compression delivery is virtually impossible

learning objectives

After reading this chapter you should be able to demonstrate the necessity of compression, to discuss criteria for the selection of codecs and mention some of the alternatives, to characterize the MPEG-4 and SMIL standards, to explain the difference between MPEG-4 and MPEG-2, and to speculate about the feasibility of a semantic multimedia web.

Without compression and decompression, digital information delivery would be virtually impossible. In this chapter we will take a more detailed look at compression and decompression. It contains the information that you may possibly need to decide on a suitable compression and decompression scheme (codec) for your future multimedia productions. We will also discuss the standards that may govern the future (multimedia) Web, including MPEG-4, SMIL and RM3D. We will explore to what extent these standards allow us to realize the optimal multimedia platform, that is one that embodies digital convergence in its full potential. Finally, we will investigate how these ideas may ultimately lead to a (multimedia) semantic web.

1

3.1 codecs

Back to the everyday reality of the technology that surrounds us. What can we expect to become of networked multimedia? Let one thing be clear

compression is the key to effective delivery

There can be no misunderstanding about this, although you may wonder why you need to bother with compression (and decompression). The answer is simple. You need to be aware of the size of what you put on the web and the demands that imposes on the network. Consider the table, taken from Vasudev and Li (1997), below.

media	uncompressed	compressed
voice 8k samples/sec, 8 bits/sample	64 kbps	2-4 kbps
slow motion video 10fps 176x120 8 bits	5.07 Mbps	8-16 kbps
audio conference 8k samples/sec 8bits	64 kbps	16-64 kbps
video conference 15 fps 352x240 8bits	30.4 Mbps	64-768 kbps
audio (stereo) 44.1 k samples/s 16 bits	1.5 Mbps	128k-1.5Mbps
video 15 fps 352x240 15 fps 8 bits	30.4 Mbps	384 kbps
video (CDROM) 30 fps 352x240 8 bits	60.8 Mbps	1.5-4 Mbps
video (broadcast) 30 fps 720x480 8 bits	248.8 Mbps	3-8 Mbps
HDTV 59.9 fps 1280x720 8 bits	1.3 Gbps	20 Mbps

You'll see that, taking the various types of connection in mind

(phone: 56 Kb/s, ISDN: 64-128 Kb/s, cable: 0.5-1 Mb/s, DSL: 0.5-2 Mb/s)

you must be careful to select a media type that is suitable for your target audience. And then again, choosing the right compression scheme might make the difference between being able to deliver or not being able to do so. Fortunately,

images, video and audio are amenable to compression

Why this is so is explained in Vasudev and Li (1997). Compression is feasible because of, on the one hand, the statistical redundancy in the signal, and the irrelevance of particular information from a perceptual perspective on the other hand. Redundancy comes about by both spatial correlation, between neighboring pixels, and temporal correlation, between successive frames.

The actual process of encoding and decoding may be depicted as follows:

codec = (en)coder + decoder

signal $\rightarrow$ source coder $\rightarrow$ channel coder (encoding)

signal $\leftarrow$ source decoder $\leftarrow$ channel decoder (decoding)

Of course, the coded signal must be transmitted accross some channel, but this is outside the scope of the coding and decoding issue. With this diagram in mind we can specify the *codec design problem*:

> *From a systems design viewpoint, one can restate the codec design problem as a bit rate minimization problem, meeting (among others) constraints concerning: specified levels of signal quality, implementation complexity, and communication delay (start coding – end decoding).*

2

compression methods

As explained in Vasudev and Li (1997), there is a large variety of compression (and corresponding decompression) methods, including model-based methods, as for example the object-based MPEG-4 method that will be discussed later, and waveform-based methods, for which we generally make a distinction between lossless and lossy methods. Hufmann coding is an example of a lossless method, and methods based on Fourier transforms are generally lossy. Lossy means that actual data is lost, so that after decompression there may be a loss of (perceptual) quality.

Leaving a more detailed description of compression methods to the diligent students' own research, it should come as no surprise that when selecting a compression method, there are a number of tradeoffs, with respect to, for example, coding efficiency, the complexity of the coder and decoder, and the signal quality. In summary, the follwoing issues should be considered:

tradeoffs

- *resilience to transmission errors*
- *degradations in decoder output – lossless or lossy*
- *data representation – browsing & inspection*
- *data modalities – audio & video.*
- *transcoding to other formats – interoperability*
- *coding efficiency – compression ratio*
- *coder complexity – processor and memory requirements*
- *signal quality – bit error probability, signal/noise ratio*

For example, when we select a particular coder-decoder scheme we must consider whether we can guarantee resilience to transmission errors and how these will affect the users' experience. And to what extent we are willing to accept degradations in decoder output, that is lossy output. Another issue in selecting a method of compression is whether the (compressed) data representation allows for browsing & inspection. And, for particular applications, such as conferencing, we should be worried about the interplay of data modalities,in particular, audio & video. With regard to the many existing codecs and the variety of platforms we may desire the possibility of transcoding to other formats to achieve, for example, exchange of media objects between tools, as is already common for image processing tools.

compression standards

Given the importance of codecs it should come as no surprise that much effort has been put in developing standards, such as JPEG for images and MPEG for audio and video.

Most of you have heard of MP3 (the audio format), and at least some of you should be familiar with MPEG-2 video encoding (which is used for DVDs).

Now, from a somewhat more abstract perspective, we can, again following Vasudev and Li (1997), make a distinction between a *pixel-based approach* (coding the raw signal so to speak) and an *object-based approach*, that uses segmentation and a more advanced scheme of description.

- *pixel-based* – MPEG-1, MPEG-2, H3.20, H3.24
- *object-based* – MPEG-4

As will be explained in more detail when discussing the MPEG-4 standard in section 3.2, there are a number of advantages with an object-based approach. There is, however, also a price to pay. Usually (object) segmentation does not come for free, but requires additional effort in the phase of authoring and coding.

MPEG-1 To conclude this section on codecs, let's look in somewhat more detail at what is involved in coding and decoding a video signal according to the MPEG-1 standard.

MPEG-1 video compression uses both *intra-frame analysis*, for the compression of individual frames (which are like images), as well as. *inter-frame analysis*, to detect redundant blocks or invariants between frames.

The MPEG-1 encoded signal itself is a sequence of so-called I, P and B frames.

frames

- I: intra-frames – independent images
- P: computed from closest frame using DCT (or from P frame)
- B: computed from two closest P or I frames

Decoding takes place by first selecting I-frames, then P-frames, and finally B-frames. When an error occurs, a safeguard is provided by the I-frames, which stand on themselves.

Subsequent standards were developed to accomodate for more complex signals and greater functionality. These include MPEG-2, for higher pixel resolution and data rate, MPEG-3, to support HDTV, MPEG-4, to allow for object-based compression, and MPEG-7, which supports content description. We will elaborate on MPEG-4 in the next section, and briefly discuss MPEG-7 at the end of this chapter.

example(s) – *gigaport*

GigaPort[1] is a project focussing on the development and use of advanced and innovative Internet technology. The project, as can be read on the website,

[1] www.gigaport.nl/info/en/about/home.jsp

focuses on research on next-generation networks and the implementation of a next-generation network for the research community.

Topics for research include:

GigaPort

- optical network technologies - models for network architecture, optical network components and light path provisioning.
- high performance routing and switching - new routing technologies and transport protocols, with a focus on scalability and stability robustness when using data-intensive applications with a high bandwidth demand.
- management and monitoring - incident response in hybrid networks (IP and optical combined) and technologies for network performance monitoring, measuring and reporting.
- grids and access - models, interfaces and protocols for user access to network and grid facilities.
- test methodology - effective testing methods and designing tests for new technologies and network components.

As one of the contributions, internationally, the development of optical technology is claimed, in particular *lambda* networking, networking on a specific wavelength. Locally, the projects has contributed to the introduction of fibre-optic networks in some major cities in the Netherlands.

research directions– *digital video formats*

In the online version you will find a brief overview of *digital video technology*, written by Andy Tanenbaum, as well as some examples of videos of our university, encoded at various bitrates for different viewers.

What is the situation? For traditional television, there are three standards. The american (US) standard, NTSC, is adopted in North-America, South-America and Japan. The european standard, PAL, whuch seems to be technically superior, is adopted by the rest of the world, except France and the eastern-european countries, which have adopted the other european standard, SECAM. An overview of the technical properties of these standards, with permission taken from Tanenbaum's account, is given below.

system	**spatial resolution**	**frame rate**	**mbps**
NTSC	704 x 480	30	243 mbps
PAL/SECAM	720 x 576	25	249 mbps

Obviously real-time distribution of a more than 200 mbps signal is not possible, using the nowadays available internet connections. Even with compression on the fly, the signal would require 25 mbps, or 36 mbps with audio. Storing the signal on disk is hardly an alternative, considering that one hour would require 12 gigabytes.

When looking at the differences between streaming video (that is transmitted real-time) and storing video on disk, we may observe the following tradeoffs:

item	**streaming**	**downloaded**
bandwidth	equal to the display rate	may be arbitrarily small
disk storage	none	the entire file must be stored
startup delay	almost none	equal to the download time
resolution	depends on available bandwidth	depends on available disk storage

So, what are our options? Apart from the quite successful MPEG encodings, which have found their way in the DVD, there are a number of proprietary formats used for transmitting video over the internet: Quicktime, introduced by Apple, early 1990s, for local viewing; RealVideo, streaming video from RealNetworks; and Windows Media, a proprietary encoding scheme fromMicrosoft. Examples of these formats, encoded for various bitrates are available at Video at VU.

Apparently, there is some need for digital video on the internet, for example as propaganda for attracting students, for looking at news items at a time that suits you, and (now that digital video cameras become affordable) for sharing details of your family life.

Is digital video all there is? Certainly not! In the next section, we will deal with standards that allow for incorporating (streaming) digital video as an element in a compound multimedia presentation, possibly synchronized with other items, including synthetic graphics. Online, you will find some examples of digital video that are used as texture maps in 3D space. These examples are based on the technology presented in section **??**, and use the streaming video codec from Real Networks that is integrated as a rich media extension in the *blaxxun* Contact 3D VRML plugin.

comparison of codecs A review of codecs[2], including Envivio MPEG-4, QuickTime 6, RealNetworks 9 en Windows Media 9 was published januari 2005 by the European Broadcast Union[3]. It appeared that The Real Networks codecs came out best, closely followed by the Windows Media 9 result. Ckeck it out!

3

[2] www.ebu.ch/trev_301-samviq.pdf
[3] www.ebu.ch/trev_home.html

3.2 standards

Imagine what it would be like to live in a world without standards. You may get the experience when you travel around and find that there is a totally different socket for electricity in every place that you visit.

Now before we continue, you must realize that there are two types of standards: *de facto* market standards (enforced by sales politics) and committee standards (that are approved by some official organization). For the latter type of standards to become effective, they need consent of the majority of market players.

For multimedia on the web, we will discuss three standards and RM3D which was once proposed as a standard and is now only of historical significance.

standards

- XML – eXtensible Markup Language (SGML)
- MPEG-4 – coding audio-visual information
- SMIL – Synchronized Multimedia Integration Language
- RM3D – (Web3D) Rich Media 3D (extensions of X3D/VRML)

XML, the *eXtensible Markup Language*, is becoming widely accepted. It is being used to replace HTML, as well as a data exchange format for, for example, business-to-business transactions. XML is derived from SGML (Structured Generalized Markup Language) that has found many applications in document processing. As SGML, XML is a generic language, in that it allows for the specification of actual markup languages. Each of the other three standards mentioned allows for a syntactic encoding using XML.

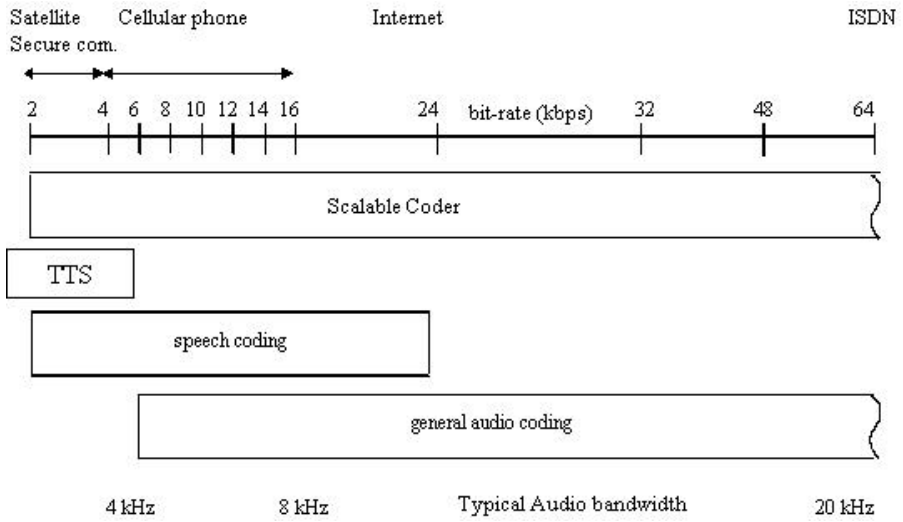
MPEG-4 aims at providing "the standardized technological elements enabling the integration of production, distribution and content access paradigms of digital television, interactive graphics and multimedia", Koenen (2000). A preliminary version of the standard has been approved in 1999. Extensions in specific domains are still in progress.

SMIL, the *Synchronized Multimedia Integration Language*, has been proposed by the W3C "to enable the authoring of TV-like multimedia presentations, on the Web". The SMIL language is an easy to learn HTML-like language. SMIL presentations can be composed of streaming audio, streaming video, images, text or any other media type, W3C (2001). SMIL-1 has become a W3C recommendation in 1998. SMIL-2 is at the moment of writing still in a draft stage.

RM3D, *Rich Media 3D*, is not a standard as MPEG-4 and SMIL, since it does currently not have any formal status. The RM3D working group arose out of the X3D working group, that addressed the encoding of VRML97 in XML. Since there were many disagreements on what should be the core of X3D and how extensions accomodating VRML97 and more should be dealt with, the RM3D working group was founded in 2000 to address the topics of extensibility and the integration with rich media, in particular video and digital television.

remarks Now, from this description it may seem as if these groups work in total isolation from eachother. Fortunately, that is not true. MPEG-4, which is the

most encompassing of these standards, allows for an encoding both in SMIL and X3D. The X3D and RM3D working groups, moreover, have advised the MPEG-4 commitee on how to integrate 3D scene description and human avatar animation in MPEG-4. And finally, there have been rather intense discussions between the SMIL and RM3D working groups on the timing model needed to control animation and dynamic properties of media objects.



4

MPEG-4

The MPEG standards (in particular 1,2 and 3) have been a great success, as testified by the popularity of mp3 and DVD video.

Now, what can we expect from MPEG-4? Will MPEG-4 provide *multimedia for our time*, as claimed in Koenen (1999). The author, Rob Koenen, is senior consultant at the dutch KPN telecom research lab, active member of the MPEG-4 working group and editor of the MPEG-4 standard document.

> *"Perhaps the most immediate need for MPEG-4 is defensive. It supplies tools with which to create uniform (and top-quality) audio and video encoders on the Internet, preempting what may become an unmanageable tangle of proprietary formats."*

Indeed, if we are looking for a general characterization it would be that MPEG-4 is primarily

MPEG-4

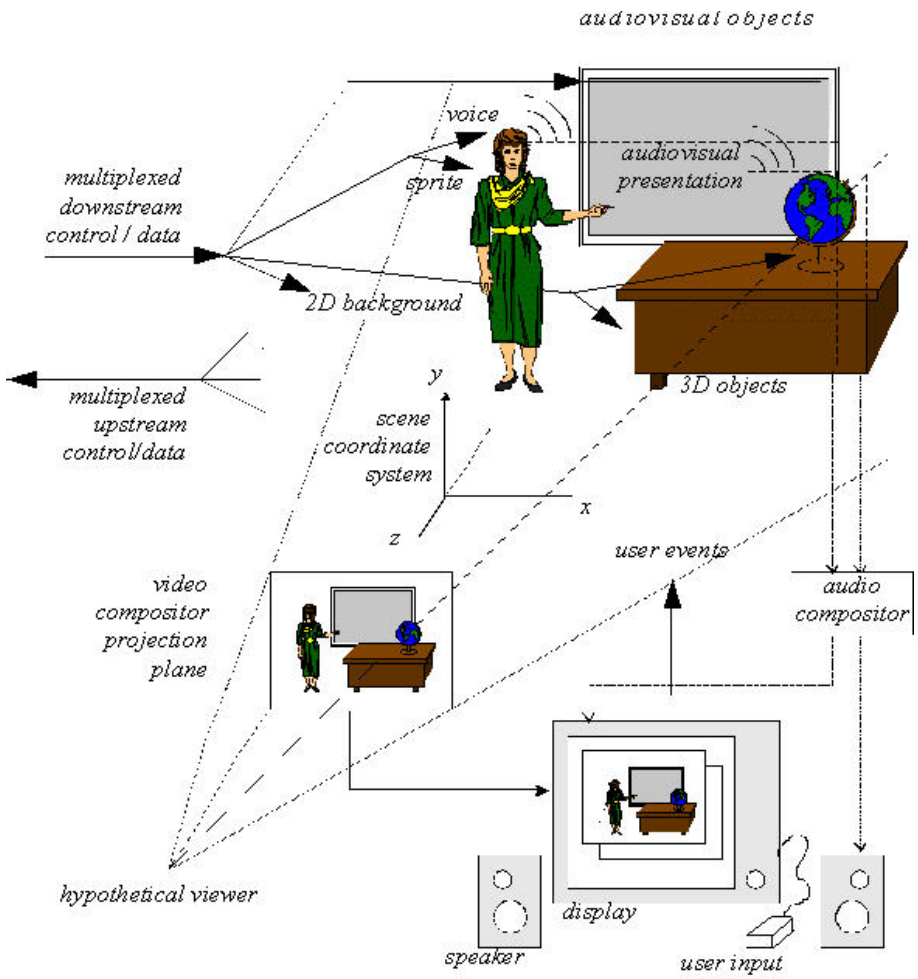
a toolbox of advanced compression algorithms for audiovisual information

and, moreover, one that is suitable for a variety of display devices and networks, including low bitrate mobile networks. MPEG-4 supports scalability on a variety of levels:

scalability

- *bitrate* – switching to lower bitrates
- *bandwidth* – dynamically discard data
- *encoder and decoder complexity* – signal quality

Dependent on network resources and platform capabilities, the 'right' level of signal quality can be determined by selecting the optimal codec, dynamically.



media objects It is fair to say that MPEG-4 is a rather ambitious standard. It aims at offering support for a great variety of audiovisual information, including still images, video, audio, text, (synthetic) talking heads and synthesized speech, synthetic graphics and 3D scenes, streamed data applied to media objects, and user interaction – e.g. changes of viewpoint.

Let's give an example, taken from the MPEG-4 standard document.

example

> *Imagine, a talking figure standing next to a desk and a projection screen, explaining the contents of a video that is being projected on the screen, pointing at a globe that stands on the desk. The user that is watching that scene decides to change from viewpoint to get a better look at the globe ...*

How would you describe such a scene? How would you encode it? And how would you approach decoding and user interaction?

The solution lies in defining *media objects* and a suitable notion of composition of media objects.

media objects

- *media objects* – units of aural, visual or audiovisual content
- *composition* – to create compound media objects (audiovisual scene)
- *transport* – multiplex and synchronize data associated with media objects
- *interaction* – feedback from users' interaction with audiovisual scene

For 3D-scene description, MPEG-4 builds on concepts taken from VRML (Virtual Reality Modeling Language, discussed in chapter 7).

Composition, basically, amounts to building a *scene graph*, that is a tree-like structure that specifies the relationship between the various simple and compound media objects. Composition allows for placing media objects anywhere in a given coordinate system, applying transforms to change the appearance of a media object, applying streamed data to media objects, and modifying the users viewpoint.

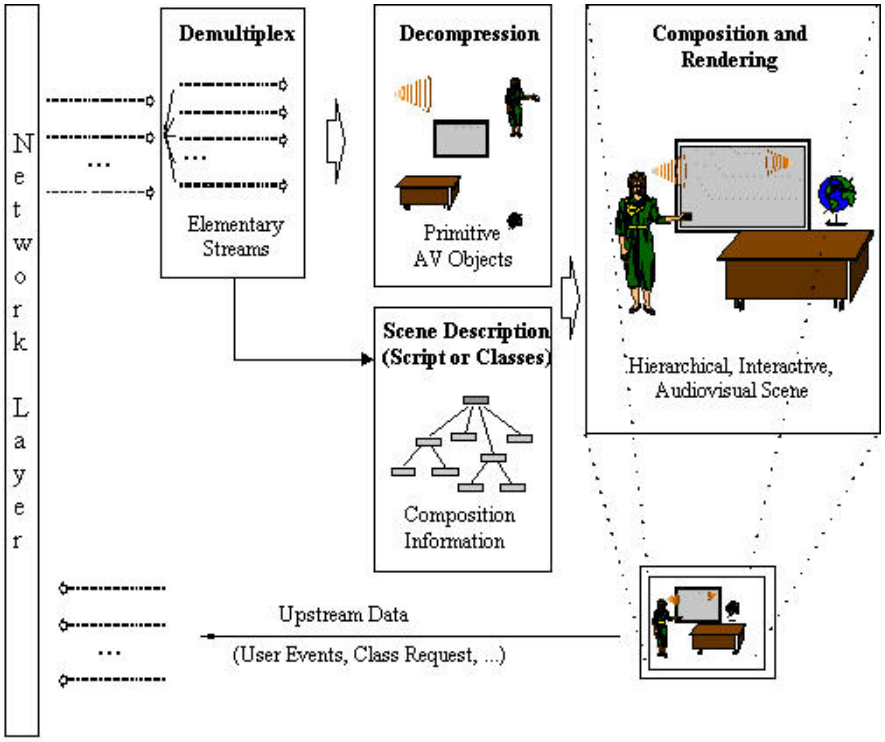
So, when we have a multimedia presentation or audiovisual scene, we need to get it accross some network and deliver it to the end-user, or as phrased in Koenen (2000):

transport

> *The data stream (Elementary Streams) that result from the coding process can be transmitted or stored separately and need to be composed so as to create the actual multimedia presentation at the receivers side.*

At a system level, MPEG-4 offers the following functionalities to achieve this:

- BIFS (Binary Format for Scenes) – describes spatio-temporal arrangements of (media) objects in the scene
- OD (Object Descriptor) – defines the relationship between the elementary streams associated with an object
- *event routing* – to handle user interaction

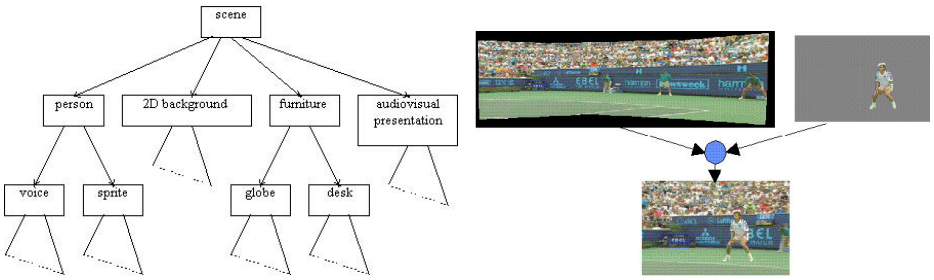


6

In addition, MPEG-4 defines a set of functionalities For the delivery of streamed data, DMIF, which stands for

Delivery Multimedia Integration Framework

that allows for transparent interaction with resources, irrespective of whether these are available from local storage, come from broadcast, or must be obtained from some remote site. Also transparency with respect to network type is supported. *Quality of Service* is only supoorted to the extent that it ispossible to indicate needs for bandwidth and transmission rate. It is however the responsability of the network provider to realize any of this.



(a) scene graph (b) sprites

7

authoring What MPEG-4 offers may be summarized as follows

benefits

- *end-users* – interactive media accross all platforms and networks
- *providers* – transparent information for transport optimization
- *authors* – reusable content, protection and flexibility

In effect, although MPEG-4 is primarily concerned with efficient encoding and scalable transport and delivery, the *object-based* approach has also clear advantages from an authoring perspective.

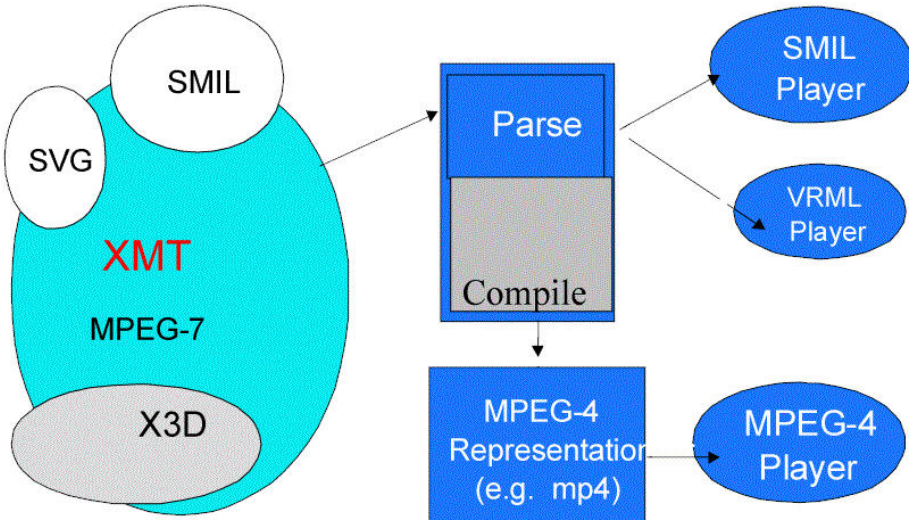
One advantage is the possibility of reuse. For example, one and the same background can be reused for multiplepresentations or plays, so you could imagine that even an amateur game might be 'located' at the centre-court of Roland Garros or Wimbledon.

Another, perhaps not so obvious, advantage is that provisions have been made for

managing intellectual property

of media objects.

And finally, media objects may potentially be annotated with meta-information to facilitate information retrieval.



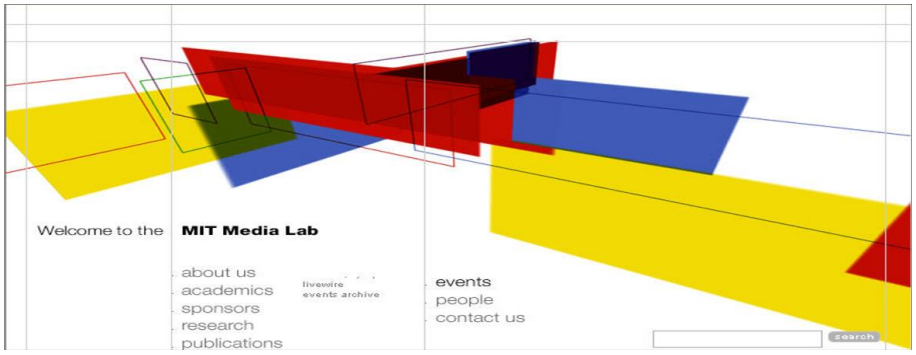
8

syntax In addition to the binary formats, MPEG-4 also specifies a syntactical format, called XMT, which stands for *eXtensible MPEG-4 Textual format*.

XMT

- XMT contains a subset of X3D
- SMIL is mapped (incompletely) to XMT

when discussing RM3D which is of interest from a historic perspective, we will further establish what the relations between, respectively MPEG-4, SMIL and RM3D are, and in particular where there is disagreement, for example with respect to the timing model underlying animations and the temporal control of media objects.



9

example(s) – *structured audio*

The Machine Listening Group[4] of the MIT Media Lab[5] is developing a suite of tools for *structered audio*, which means *transmitting sound by describing it rather than compressing it.* It is claimed that tools based on the MPEG-4 standard will be the future platform for computer music, audio for gaming, streaming Internet radio, and other multimedia applications.

The structured audio project is part of a more encompassing research effort of the Music, Mind and Machine Group[6] of the MIT Media Lab, which *envisages a new future of audio technologies and interactive applications that will change the way music is conceived, created, transmitted and experienced*,

SMIL

SMIL is pronounced as *smile*. SMIL, the Synchronized Multimedia Integration Language, has been inspired by the Amsterdam Hypermedia Model (AHM). In fact, the dutch research group at CWI that developed the AHM actively participated in the SMIL 1.0 committee. Moreover, they have started a commercial spinoff to create an editor for SMIL, based on the editor they developed for CMIF. The name of the editor is GRINS. Get it?

As indicated before SMIL is intended to be used for

TV-like multimedia presentations

The SMIL language is an XML application, resembling HTML. SMIL presentations can be written using a simple text-editor or any of the more advanced tools, such as GRINS. There is a variety of SMIL players. The most wellknown perhaps is the RealNetworks G8 players, that allows for incorporating RealAudio and RealVideo in SMIL presentations.

parallel and sequential

> *Authoring a SMIL presentation comes down, basically, to name media components for text, images,audio and video with URLs, and to schedule their presentation either in parallel or in sequence.*

Quoting the SMIL 2.0 working draft, we can characterize the SMIL presentation characteristics as follows:

presentation characteristics

- The presentation is composed from several components that are accessible via URL's, e.g. files stored on a Web server.
- The components have different media types, such as audio, video, image or text. The begin and end times of different components are specified relative to events in other media components. For example, in a slide show, a particular slide is displayed when the narrator in the audio starts talking about it.

[4] sound.media.mit.edu/mpeg4
[5] www.media.mit.edu
[6] sound.media.mit.edu

- Familiar looking control buttons such as stop, fast-forward and rewind allow the user to interrupt the presentation and to move forwards or backwards to another point in the presentation.
- Additional functions are "random access", i.e. the presentation can be started anywhere, and "slow motion", i.e. the presentation is played slower than at its original speed.
- The user can follow hyperlinks embedded in the presentation.

Where HTML has become successful as a means to write simple hypertext content, the SMIL language is meant to become a vehicle of choice for writing *synchronized hypermedia*. The working draft mentions a number of possible applications, for example a photoalbun with spoken comments, multimedia training courses, product demos with explanatory text, timed slide presentations, onlime music with controls.

As an example, let's consider an interactive news bulletin, where you have a choice between viewing a weather report or listening to some story about, for example, the decline of another technology stock. Here is how that could be written in SMIL:

example

```
<par>
   <a href=" #Story"> <img src="button1.jpg"/> </a>
   <a href=" #Weather"> <img src="button2.jpg"/></a>
    <excl>
        <par id="Story" begin="0s">
          <video src="video1.mpg"/>
          <text src="captions.html"/>
        </par>

        <par id="Weather">
          <img src="weather.jpg"/>
          <audio src="weather-rpt.mp3"/>
        </par>
    </excl>
</par>
```

Notice that there are two *parallel* (PAR) tags, and one *exclusive* (EXCL) tag. The *exclusive* tag has been introduced in SMIL 2.0 to allow for making an exclusive choice,so that only one of the items can be selected at a particular time. The SMIL 2.0 working draft defines a number of elements and attributes to control presentation, synchronization and interactivity, extending the functionality of SMIL 1.0.

Before discussing how the functionality proposed in the SMIL 2.0working draft may be realized, we might reflect on how to position SMIL with respect to the many other approaches to provide multimedia on the web. As other approaches we may think of *flash*, dynamic HTML (using javascript), or java applets. In the SMIL 2.0 working draft we read the following comment:

history

> *Experience from both the CD-ROM community and from the Web multimedia community suggested that it would be beneficial to adopt a declarative format for expressing media synchronization on the Web as an alternative and complementary approach to scripting languages.*
>
> *Following a workshop in October 1996, W3C established a first working group on synchronized multimedia in March 1997. This group focused on the design of a declarative language and the work gave rise to SMIL 1.0 becoming a W3C Recommendation in June 1998.*

In summary, SMIL 2.0 proposes a *declarative format* to describe the temporal behavior of a multimedia presentation, associate hyperlinks with media objects, describe the form of the presentation on a screen, and specify interactivity in multimedia presentations. Now,why such a fuzz about "declarative format"? Isn't scripting more exciting? And aren't the tools more powerful? Ok, ok. I don't want to go into that right now. Let's just consider a *declarative format* to be more elegant. Ok?

To support the functionality proposed for SMIL 2.0 the working draft lists a number of modules that specify the interfaces for accessing the attributes of the various elements. SMIL 2.0 offers modules for animation, content control, layout, linking, media objects, meta information, timing and synchronization, and transition effects.

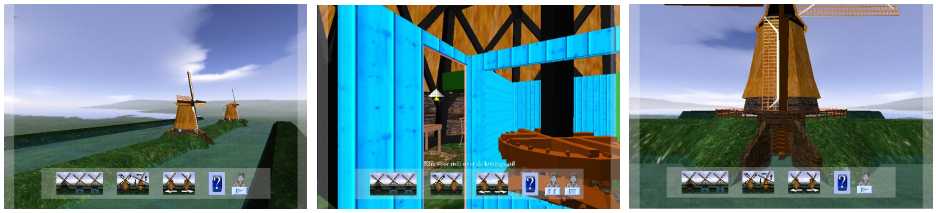
This modular approach allows to reuse SMIL syntax and semantics in other XML-based languages, in particular those that need to represent timing and synchronization. For example:

module-based reuse

- SMIL modules could be used to provide lightweight multimedia functionality on mobile phones, and to integrate timing into profiles such as the WAP forum's WML language, or XHTML Basic.
- SMIL timing, content control, and media objects could be used to coordinate broadcast and Web content in an enhanced-TV application.
- SMIL Animation is being used to integrate animation into W3C's Scalable Vector Graphics language (SVG).
- Several SMIL modules are being considered as part of a textual representation for MPEG4.

The SMIL 2.0 working draft is at the moment of writing being finalized. It specifies a number of language profiles topromote the reuse of SMIL modules. It also improves on the accessibility features of SMIL 1.0, which allows for, for example,, replacing captions by audio descriptions.

In conclusion, SMIL 2.0 is an interesting standard, for a number of reasons. For one, SMIL 2.0 has solid theoretical underpinnings in a well-understood, partly formalized, hypermedia model (AHM). Secondly, it proposes interesting functionality, with which authors can make nice applications. In the third place, it specifies a high level declarative format, which is both expressive and flexible. And finally, it is an open standard (as opposed to proprietary standard). So everybody can join in and produce players for it!



10

RM3D – not a standard

The web started with simple HTML hypertext pages. After some time static images were allowed. Now, there is support for all kinds of user interaction, embedded multimedia and even synchronized hypermedia. But despite all the graphics and fancy animations, everything remains flat. Perhaps surprisingly, the need for a 3D web standard arose in the early days of the web. In 1994, the acronym VRML was coined by Tim Berners-Lee, to stand for *Virtual Reality Markup Language*. But, since 3D on the web is not about text but more about worlds, VRML came to stand for *Virtual Reality Modeling Language*. Since 1994, a lot of progress has been made.

www.web3d.org

- VRML 1.0 – *static 3D worlds*
- VRML 2.0 or VRML97 – *dynamic behaviors*
- VRML200x – *extensions*
- X3D – *XML syntax*
- RM3D – *Rich Media in 3D*

In 1997, VRML2 was accepted as a standard, offering rich means to create 3D worlds with dynamic behavior and user interaction. VRML97 (which is the same as VRML2) was, however, not the success it was expected to be, due to (among others) incompatibility between browsers, incomplete implementations of the standards, and high performance requirements.

As a consequence, the Web3D Consortium (formerly the VRML Consortium) broadened its focus, and started thinking about extensions or modifications of VRML97 and an XML version of VRML (X3D). Some among the X3D working group felt the need to rethink the premisses underlying VRML and started the Rich Media Working Group:

groups.yahoo.com/group/rm3d/

> *The Web3D Rich Media Working Group was formed to develop a Rich Media standard format (RM3D) for use in next-generation media devices. It is a highly active group with participants from a broad range of companies including 3Dlabs, ATI, Eyematic, OpenWorlds, Out of the Blue Design, Shout Interactive, Sony, Uma, and others.*

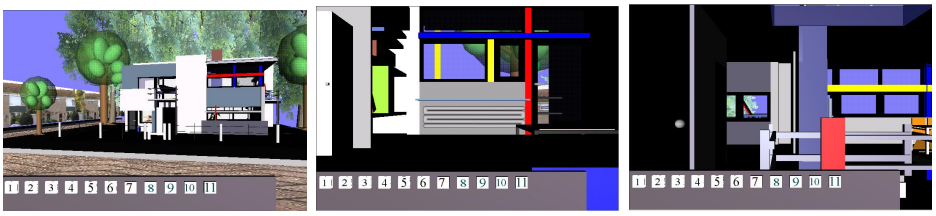
In particular:

RM3D

> *The Web3D Consortium initiative is fueled by a clear need for a standard high performance Rich Media format. Bringing together content creators with successful graphics hardware and software experts to define RM3D will ensure that the new standard addresses authoring and delivery of a new breed of interactive applications.*

The working group is active in a number of areas including, for example, multi-texturing and the integration of video and other streaming media in 3D worlds.

Among the driving forces in the RM3D group are Chris Marrin and Richter Rafey, both from Sony, that proposed *Blendo*, a rich media extension of VRML. Blendo has a strongly typed object model, which is much more strictly defined than the VRML object model, to support both declarative and programmatic extensions. It is interesting to note that the premisse underlying the Blendo proposal confirms (again) the primacy of the TV metaphor. That is to say, what Blendo intends to support are TV-like presentations which allow for user interaction such as the selection of items or playing a game. Target platforms for Blendo include graphic PCs, set-top boxes, and the Sony Playstation!



11

requirements The focus of the RM3D working group is not *syntax* (as it is primarily for the X3D working group) but *semantics*, that is to enhance the VRML97 standard to effectively incorporate rich media. Let's look in more detail at the requirements as specified in the RM3Ddraft proposal.

requirements

- *rich media* – audio, video, images, 2D & 3D graphics (with support for temporal behavior, streaming and synchronisation)
- *applicability* – specific application areas, as determined by commercial needs and experience of working group members

The RM3D group aims at interoperability with other standards.

- *interoperability* – VRML97, X3D, MPEG-4, XML (DOM access)

In particular, an XML syntax is being defined in parallel (including interfaces for the DOM). And, there is mutual interest and exchange of ideas between the MPEG-4 and RM3D working group.

As mentioned before, the RM3D working group has a strong focus on defining an object model (that acts as a common model for the representation of objects and their capabilities) and suitable mechanisms for extensibility (allowing for the integration of new objects defined in Java or C++, and associated scripting primitives and declarative constructs).

Notice that extensibility also requires the definition of a declarative format, so that the content author need not bother with programmatic issues.

The RM3D proposal should result in effective 3D media presentations. So as additional requirements we may, following the working draft, mention: high-quality realtime rendering, for realtime interactive media experiences; platform adaptability, with query functions for programmatic behavior selection; predictable behavior, that is a well-defined order of execution; a high precision number systems, greater than single-precision IEEE floating point numbers; and minimal size, that is both download size and memory footprint.

Now, one may be tempted to ask how the RM3D proposals is related to the other standard proposals such as MPEG-4 and SMIL, discussed previously. Briefly put, paraphrased from one of Chris Marrin's messages on the RM3D mailing list

> *SMIL is closer to the author* and *RM3D is closer to the implementer.*

MPEG-4, in this respect is even further away from the author since its chief focus is on compression and delivery across a network.

RM3D takes 3D scene description as a starting point and looks at pragmatic ways to integrate rich media. Since 3D is itself already computationally intensive, there are many issues thatarise in finding efficient implementations for the proposed solutions.

12

timing model RM3D provides a declarative format formany interesting features, such as for example texturing objects with video. In comparison to VRML, RM3D is meant to provide more temporal control over time-based media objects and animations. However, there is strong disagreement among the working group members as to what time model the dynamic capabilities of RM3D should be based on. As we read in the working draft:

working draft

> *Since there are three vastly different proposals for this section (time model), the original <RM3D> 97 text is kept. Once the issues concerning time-dependent nodes are resolved, this section can be modified appropriately.*

Now, what are the options? Each of the standards discussed to far provides us with a particular solution to timing. Summarizing, we have a time model based on a spring metaphor in MPEG-4, the notion of cascading time in SMIL (inspired by cascading stylesheets for HTML) and timing based on the routing of events in RM3D/VRML.

The MPEG-4 standard introduces the *spring metaphor* for dealing with temporal layout.

MPEG-4 – spring metaphor

- duration – minimal, maximal, optimal

The spring metaphor amounts to the ability to shrink or stretch a media object within given bounds (minimum, maximum) to cope with, for example, network delays.

The SMIL standard is based on a model that allows for propagating durations and time manipulations in a hierarchy of media elements. Therefore it may be referred to as a cascading modelof time.

SMIL – cascading time

- time container – speed, accelerate, decelerate, reverse, synchronize

Media objects, in SMIL, are stored in some sort of container of which the timing properties can be manipulated.

```
<seq speed="2.0">
   <video src="movie1.mpg" dur="10s"/>
   <video src="movie2.mpg" dur="10s"/>
   <img src="img1.jpg" begin="2s" dur="10s">
              <animateMotion from="-100,0" to="0,0" dur="10s"/>
   </img>
   <video src="movie4.mpg" dur="10s"/>
</seq>
```

In the example above,we see that the speed is set to *2.0*, which will affect the pacing of each of the individual media elements belonging to that (sequential) group. The duration of each of the elements is specified in relation to the parent container. In addition, SMIL offers the possibility to synchronize media objects to control, for example, the end time of parallel media objects.

VRML97's capabilities for timing rely primarily on the existence of a *TimeSensor* thatsends out time events that may be routed to other objects.

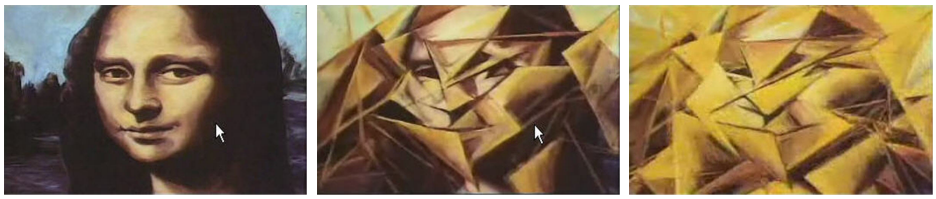
RM3D/VRML – event routing

- *TimeSensor* – isActive, start, end, cycleTime, fraction, loop

When a *TimeSensor* starts to emit time events, it also sends out an event notifying other objects that it has become active. Dependent on itsso-called *cycleTime*, it sends out the fraction it covered since it started. This fraction may be send to one of the standard interpolators or a script so that some value can be set, such as for

example the orientation, dependent on the fraction of the time intercal that has passed. When the *TimeSensor* is made to loop, this is done repeatedly. Although time in VRML is absolute, the frequency with which fraction events are emitted depends on the implementation and processor speed.

Lacking consensus about a better model, this model has provisionally been adopted, with some modifications, for RM3D. Nevertheless, the SMIL cascading time model has raised an interest in the RM3D working group, to the extent that Chris Marrin remarked (in the mailing list) *"we could go to school here"*. One possibility for RM3D would be to introduce *time containers* that allow for a temporal transform of their children nodes, in a similar way as grouping containers allow for spatial transforms of their children nodes. However, that would amount to a dual hierarchy, one to control (spatial) rendering and one to control temporal characteristics. Merging the two hierarchies, as is (implicitly) the case in SMIL, might not be such a good idea, since the rendering and timing semantics of the objects involved might be radically different. An interesting problem, indeed, but there seems to be no easy solution.



13

example(s) – *rich internet applications*

In a seminar held by *Lost Boys*, which is a dutch subdivison if Icon Media Lab[7], *rich internet applications* (RIA), were presented as the new solutions to present applications on the web. As indicated by Macromedia[8], who is one of the leading companies in this fiwld, *experience matters*, and so plain html pages pages do not suffice since they require the user to move from one page to another in a quite unintuitive fashion. Macromedia presents its new line of *flash*-based products to create such *rich internet applications*. An alternative solution, based on general W3C recommendations, is proposed by BackBase[9]. Interestingly enough, using either technology, many of the paricipants of the seminar indicated a strong preference for a backbuuton, having similar functionality as the often used backbutton in general internet browsers.

[7] www.iconmedialab.com
[8] www.macromedia.com/resources/business/rich_internet_apps/whitepapers.html
[9] www.backbase.com

research directions– *meta standards*

All these standards! Wouldn't it be nice to have one single standard that encompasses them all? No, it would not! Simply, because such a standard is inconceivable, unless you take some proprietary standard or a particular platform as the defacto standard (which is the way some people look at the Microsoft win32 platform, ignoring the differences between 95/98/NT/2000/XP/...). In fact, there is a standard that acts as a glue between the various standards for multimedia, namely XML. XML allows for the interchange of data between various multimedia applications, that is the transformation of one encoding into another one. But this is only syntax. What about the semantics?

Both with regard to delivery and presentation the MPEG-4 proposal makes an attempt to delineate chunks of core fuctionality that may be shared between applications. With regard to presentation, SMIL may serve as an example. SMIL applications themselves already (re)use functionality from the basic set of XML-related technologies, for example to access the document structure through the DOM (Document Object Model). In addition, SMIL defines components that it may potentially share with other applications. For example, SMIL shares its animation facilities with SVG (the Scalable Vector Graphics format recommended by the Web Consortium).

The issue in sharing is, obviously, how to relate constructs in the syntax to their operational support. When it is possible to define a common base of operational support for a variety of multimedia applications we would approach our desired meta standard, it seems. A partial solution to this problem has been proposed in the now almost forgotten HyTime standard for time-based hypermedia. HyTime introduces the notion of *architectural forms* as a means to express the operational support needed for the interpretation of particular encodings, such as for example synchronization or navigation over bi-directional links. Apart from a base module, HyTime compliant architectures may include a units measurement module, a module for dealing with location addresses, a module to support hyperlinks, a scheduling module and a rendition module.

To conclude, wouldn't it be wonderful if, for example, animation support could be shared between rich media X3D and SMIL? Yes, it would! But as you may remember from the discussion on the timing models used by the various standards, there is still to much divergence to make this a realoistic option.

3.3 a multimedia semantic web?

To finish this chapter, let's reflect on where we are now with 'multimedia' on the web. Due to refined compression schemes and standards for authoring and delivery, we seemed to have made great progress in realizing *networked multimedia*. But does this progress match what has been achieved for the dominant media type of the web, that is text or more precisely textual documents with markup?

web content

- *1st generation* – hand-coded HTML pages
- *2nd generation* – templates with content and style
- *3rd generation* – rich markup with metadata (XML)

Commonly, a distinction is made between successive generations of web content, with the first generation being simple hand-coded HTML pages. The second generation may be characterized as HTML pages that are generated on demand, for example by filling in templates with contents retrieved from a database. The third generation is envisaged to make use of rich markup, using XML, that reflects the (semantic) content of the document more directly, possibly augmented with (semantic) meta-data that describe the content in a way that allows machines, for example search engines, to process it. The great vision underlying the third generation of web content is commonly refered to as the *the semantic web*. which enhances the functionality of the current web by deploying knowledge representation and inference technology from Artificial Intelligence, using a technology known as the *Resource Description Framework* (RDF). As phrased in Ossenbruggen et. al. (2001), the semantic web will bring

structure to the meaningful content of web pages,

thus allowing computer programs,such as search engines and intelligent agents, to do their job more effectively. For search engines this means more effective information retrieval, and for agents better opportunities to provide meaningful services.

A great vision indeed. So where are we with multimedia? As an example, take a *shockwave* or *flash* presentation showing the various musea in Amsterdam. How would you attach meaning to it, so that it might become an element of a semantic structure? Perhaps you wonder what meaning could be attached to it? That should not be too difficult to think of. The (meta) information attached to such a presentation should state (minimally) that the location is Amsterdam, that the sites of interest are musea, and (possibly) that the perspective is touristic. In that way, when you search for touristic information about musea in Amsterdam, your search engine should have no trouble in selecting that presentation. Now, the answer to the question how meaning can be attached to a presentation is already given, namely by specifying meta-information in some format (of which the only requirement is that it is machine-processable). For our *shockwave*or *flash* presentation we cannot dothis in a straightforward manner. But for MPEG-4 encoded material, as well as for SMIL content, such facilities are readily available.

Should we then always duplicate our authoring effort by providing (meta) information, on top of the information that is already contained in the presentation? No, in some cases, we can also rely to some extent on content-based search or feature extraction, as will be discussed in the following chapters.

15

Resource Description Framework – the Dublin Core

The Resource Description Framework, as the W3C/RDF[10] site informs us *integrates a variety of applications from library catalogs and world-wide directories to syndication and aggregation of news, software, and content to personal collections of music, photos, and events using XML as an interchange syntax.* The RDF specifications provide, in addition *a lightweight ontology system to support the exchange of knowledge on the Web.*

The Dublin Core Metadata Initiative[11] is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models.

What exactly is meta-data? As phrased in the RDF Primer[12]

meta data

> Metadata is data about data. Specifically, the term refers to data used to identify, describe, or locate information resources, whether these resources are physical or electronic. While structured metadata processed by computers is relatively new, the basic concept of metadata has been used for many years in helping manage and use large collections of information. Library card catalogs are a familiar example of such metadata.

The Dublin Core proposes a small number of elements, to be used to give information about a resource, such as an electronic document on the Web. Consider the following example:

Dublin Core example

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns #"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:dcterms="http://purl.org/dc/terms/">
```

[10] www.w3.org/RDF
[11] dublincore.org
[12] www.w3.org/TR/rdf-primer

```
<rdf:Description rdf:about="http://www.dlib.org/dlib/may98/miller/05miller.html">
 <dc:title>An Introduction to the Resource Description Framework</dc:title>
  <dc:creator>Eric J. Miller</dc:creator>
  <dc:description>The Resource Description Framework (RDF) is an
   infrastructure that enables the encoding, exchange and reuse of
   structured metadata. rdf is an application of xml that imposes needed
   structural constraints to provide unambiguous methods of expressing
   semantics. rdf additionally provides a means for publishing both
   human-readable and machine-processable vocabularies designed to
   encourage the reuse and extension of metadata semantics among
   disparate information communities. the structural constraints rdf
   imposes to support the consistent encoding and exchange of
   standardized metadata provides for the interchangeability of separate
   packages of metadata defined by different resource description
   communities. </dc:description>
 <dc:publisher>Corporation for National Research Initiatives</dc:publisher>
  <dc:subject>
   <rdf:Bag>
    <rdf:li>machine-readable catalog record formats</rdf:li>
    <rdf:li>applications of computer file organization and
     access methods</rdf:li>
   </rdf:Bag>
  </dc:subject>
  <dc:rights>Copyright  1998 Eric Miller</dc:rights>
  <dc:type>Electronic Document</dc:type>
  <dc:format>text/html</dc:format>
  <dc:language>en</dc:language>
 <dcterms:isPartOf rdf:resource="http://www.dlib.org/dlib/may98/05contents.html"/>
 </rdf:Description>
</rdf:RDF>
```

Items such as *title*, *creator*, *subject* and *description*, actually all tags with the prefix *dc*, belong to the Dublin Core and are used to give information about the document, which incidentally concerns an introduction to the Resource Description Framework. The example also shows how *rdf* constructs can be used together with the Dublin Core elements. The prefixes *rdf* and *dc* are used to distinguish between the distinct namespaces of respectively RDF and the Dublin Core.

The Dublin Core contains the following elements:

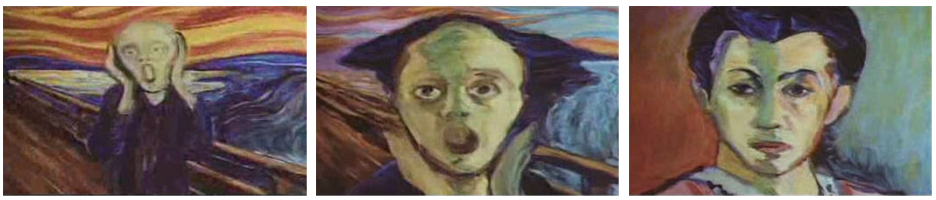
Dublin Core[13]

- *title* – name given to the resource
- *creator* – entity primarily responsible for making the content of the resource
- *subject* – topic of the content of the resource
- *description* – an account of the content of the resource
- *publisher* – entity responsible for making the resource available

[13] dublincore.org/documents/dces

- *contributor* – entity responsible for making contributions to the content of the resource
- *date* – date of an event in the lifecycle of the resource
- *type* – nature or genre of the content of the resource
- *format* – physical or digital manifestation of the resource
- *identifier* – unambiguous reference to the resource within a given context
- *source* – reference to a resource from which the present resource is derived
- *language* – language of the intellectual content of the resource
- *relation* – reference to a related resource
- *coverage* – extent or scope of the content of the resource
- *rights* – information about rights held in and over the resource

In section 10.3 we discuss an application in the domain of cultural heritage, where the Dublin Core elements are used to provide meta information about the information available for the conservation of contemporary artworks.



16

research directions– *agents everywhere*

The web is an incredibly rich resource of information. Or, as phrased in Baeza-Yates and Ribeiro-Neto (1999):

information repository

> *The Web is becoming a universal repository of human knowledge and culture, which has allowed unprecedented sharing of ideas and information in a scale never seen before.*

Now, the problem (as many of you can acknowledge) is to get the information out of it. Of course, part of the problem is that we often do not know what we are looking for. But even if we do know, it is generally not so easy to find our way. Again using the phrasing of Baeza-Yates and Ribeiro-Neto (1999):

browsing & navigation

> *To satisfy his information need, the user might navigate the hyperspace of web links searching for information of interest. However, since the hyperspace is vast and almost unknown, such a navigation task is usually inefficient.*

The solution of the problem of *getting lost in hyperspace* proposed in Baeza-Yates and Ribeiro-Neto (1999) is information retrieval, in other words *query & search*. However, this may not so easily be accomplished. As observed in Baeza-Yates and Ribeiro-Neto (1999), The main obstacle is the absence of a well-defined data model for the Web, which implies that information definition and structure is frequently of low quality. Well, that is exactly the focus of the semanics web initiative, and in particular of the Resource Description Framework discussed above.

Standardizing knowledge representation and reasoning about web resources is certainly one (important) step. Another issue, however, is how to support the user in finding the proper resources and provide the user with assistance in accomplishing his task (even if this task is merely finding suitable entertainment).

What we need, in other words, is a unifying model (encompassing both a data model and a model of computation) that allows us to deal effectively with web resources, including multimedia objects. For such a model, we may look at another area of research and development, namely *intelligent agtents*, which provides us not only with a model but also with a suitable metaphor and the technology, based on and extending object-oriented technology, to realize intelligent assistance, Eliens (2000).

For convenience, we make a distinction between two kinds of agents, *information agents* and *presentation agents*.

information agent

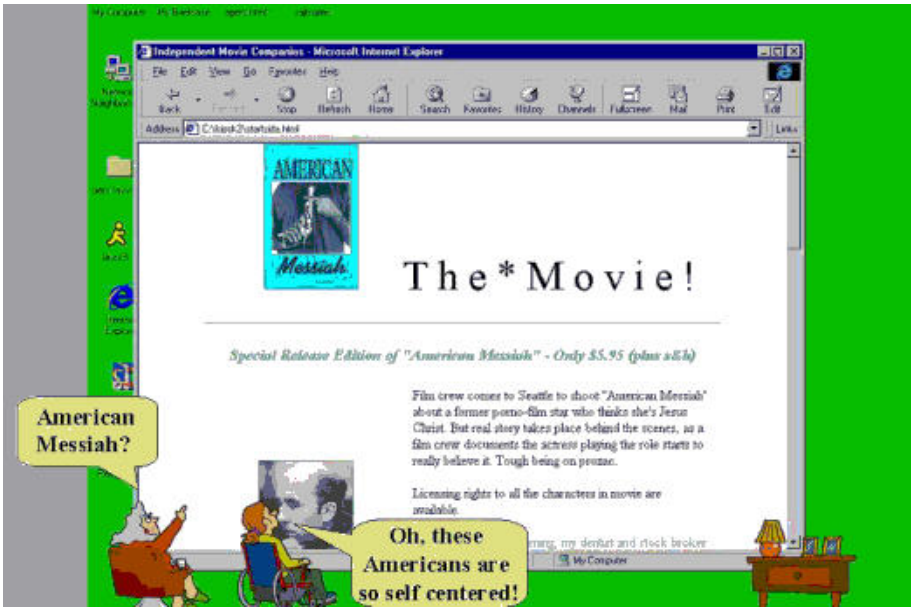
- gather information
- filter and select

Information agents are used to gather information. In addition, they filter the information and select those items that are relevant for the user. A key problem in developing information agents, however, is to find a proper representation of what the user considers to be relevant.

presentation agent

- access information
- find suitable mode of presentation

Complementary to the information agent is a *presentation agent* (having access to the information gathered) that displays the relevant information in a suitable way. Such a presentation agent can have many forms. To appetize your phantasy, you may look at the vision of *angelic guidance* presented in Broll et. al (2001). More concretely, my advice is to experiment with embodied agents that may present information in rich media 3D. In section **??**, we will present a framework for doing such experiments.



17

navigating information spaces Having *agents everywhere* might change our perspective on computing. But, it may also become quite annoying to be bothered by an agent each time that you try to interact with with your computer (you know what I mean!). However, as reported by Kristina Höök, even annoyance can be instrumental in keeping your attention to a particular task. In one of her projects, the *PERSONAS* project, which stands for

> *PERsonal and SOcial NAvigation through information spaceS*

the use of agents commenting on people navigating information space(s) is explored. As a note, the plural form of *spaces* is mine, to do justice to the plurality of information spaces.

As explained on the *PERSONAS* web site, which is listed with the acronyms, the *PERSONAS* project aims at:

PERSONAS

> *investigating a new approach to navigation through information spaces, based on a personalised and social navigational paradigm.*

The novel idea pursued in this project is to have agents (*Agneta* and *Frieda*) that are not helpful, but instead just give comments, sometimes with humor, but sometimes ironic or even sarcastic comments on the user's activities, in particular navigating an information space or (plain) web browsing. As can be read on the *PERSONAS* web site:

Agneta & Frieda

> *The AGNETA & FRIDA system seeks to integrate web-browsing and narrative into a joint mode. Below the browser window (on the desktop) are placed two female characters, sitting in their livingroom chairs, watching the browser during the session (more or less like watching television). Agneta and Frida (mother and daughter) physically react, comment, make ironic remarks about and develop stories around the information presented in the browser (primarily to each other), but are also sensitive to what the navigator is doing and possible malfunctions of the browser or server.*

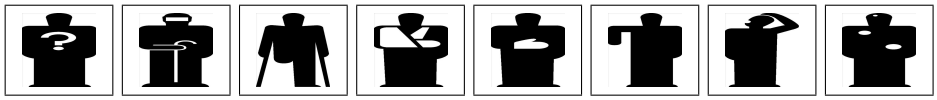
In one of her talks, Kristina Höök observed that some users get really fed up with the comments delivered by *Agneta* and *Frieda*. So, as a compromise, the level of interference can be adjusted by the user, dependent on the task at hand.

Agneta & Frieda

> *In this way they seek to attach emotional, comical or anecdotal connotations to the information and happenings in the browsing session. Through an activity slider, the navigator can decide on how active she wants the characters to be, depending on the purpose of the browsing session (serious information seeking, wayfinding, exploration or entertainment browsing).*

As you may gather, looking at the presentations accompanying this *introduction to multimedia* and Dialogs, I found the *PERSONAS* approach rather intriguing. Actually, the *PERSONAS* approach is related to the area of *affective computing*, see Picard (1998), which is an altogether different story.

The *Agneta* and *Frieda* software is available for download at the *PERSONAS* web site.



18

questions

codecs and standards

1. (*) What role do standards play in *multimedia*? Why are standards necessary for compression and delivery. Discuss the MPEG-4 standard and indicate how it is related to other (possible) standards.

concepts

2. What is a *codec*?
3. Give a brief overview of current multimedia standards.
4. What criteria must a *(multimedia) semantic web* satisfy?

technology

5. What is the *data rate* for respectively (*compressed*) voice, audio and video?

6. Explain how a *codec* functions.
7. Which considerations can you mention for choosing a compression method?
8. Give a brief description of: XML, MPEG-4, SMIL, RM3D.

projects & further reading

As a project, you may think of implementing for example JPEG compression, following Li and Drew (2004), or a SMIL-based application for cultural heritage.

You may further explore the technical issues on authoring DV material, using any of the Adobe[14], mentioned in appendix E. or compare

For further reading I advice you to take a look at the respective specifications of MPEG-4 and SMIL[15], and compare the functionality of MPEG-4 and SMIL-based presentation environments. An invaluable book dealing with the many technical aspects of compression and standards in Li and Drew (2004).

the artwork

1. costume designs – photographed from *Die Russchische Avantgarde und die Buhne 1890-1930*
2. theatre scene design, also from (above)
3. dance Erica Russel, Wiedermann (2004)
4. MPEG-4 – bits rates, from Koenen (2000).
5. MPEG-4 – scene positioning, from Koenen (2000).
6. MPEG-4 – up and downstream data, from Koenen (2000).
7. MPEG-4 – left: scene graph; right: sprites, from Koenen (2000).
8. MPEG-4 – syntax, from Koenen (2000).
9. MIT Media Lab[16] web site.
10. student work – *multimedia authoring I*, dutch windmill.
11. student work – *multimedia authoring I*, Schröder house.
12. student work – *multimedia authoring I*, train station.
13. animation – Joan Gratch, from Wiedermann (2004).
14. animation – Joan Gratch, from Wiedermann (2004).
15. animation – Joan Gratch, from Wiedermann (2004).
16. animation – Joan Gratch, from Wiedermann (2004).
17. Agneta and Frieda example.
18. signs – people, van Rooijen (2003), p. 246, 247.

Both the costume designs and theatre scene designs of the russian avantgarde movement are *expressionist* in nature. Yet, they show humanity and are in their own way very humorous. The dance animation by Erica Russell, using basic shapes and rhythms to express the movement of dance, is to some extent both solemn and equally humorous. The animations by Joan Gratch use *morphing*, to transform wellknown artworks into other equally wellknown artworks.

[14] www.adobe.com/tutorials
[15] www.w3c.org/AudioVideo
[16] medai.mit.edu