Video Requirements for Web-based Virtual Environments using Extensible 3D (X3D) Graphics

Don Brutzman and Mathias Kolsch
Web3D Consortium
Naval Postgraduate School, Monterey California USA
brutzman@nps.edu kolsch@nps.edu

28 November 2007
Submitted to W3C Video on the Web Workshop

Abstract. Real-time interactive 3D graphics and virtual environments typically include a variety of multimedia capabilities, including video. The Extensible 3D (X3D) Graphics is an ISO standard produced by the Web3D Consortium that defines 3D scenes using a scene-graph approach. Multiple X3D file formats and language encodings are available, with a primary emphasis on XML for maximum interoperability with the Web architecture. A large number of functional capabilities are needed and projected for the use of video together with Web-based virtual environments. This paper examines numerous functional requirements for the integrated use of Web-compatible video with 3D. Three areas of interest are identified: video usage within X3D scenes, linking video external to X3D scenes, and generation of 3D geometry from video.

X3D Background. Extensible 3D (X3D) is a Web-based standard for 3D graphics, enabling real-time communication using animation, user interaction and networking. X3D capabilities are proposed, implemented, evaluated and approved by members of the nonprofit Web3D Consortium (www.web3D.org). X3D is an open, royalty-free standard that is rigorously defined, published online, and ratified by the International Organization for Standards (ISO). Multiple commercial and open-source implementations are available. Web3D has a long-standing liaison relationship with W3C to ensure that X3D maximizes interoperability with the Web Architecture. [X3D 2007] [Brutzman Daly 2007]

The X3D specifications are a highly detailed set of technical documents that define the geometry and behavior capabilities of X3D. Three functionally equivalent file-format encodings are defined: Extensible Markup Language (XML) (.x3d extension), ClassicVRML (.x3dv extension), and Compressed Binary Encoding (.x3db extension). Application programming interface (API) bindings are similarly provided for the Ecmascript (aka Javascript) and Java programming languages, with a third language binding implemented (and likely to be standardized) for C++. Figure 1 illustrates these relationships. [X3D Specifications 2007]

X3D has a long history of successful development, starting with the Virtual Reality Modeling Language (VRML) as early as 1994, integrating a large number of additional advanced 3D graphics and networking capabilities along the way. Backwards compatibility with the VRML97 ISO standard has been successfully maintained. Numerous VRML/X3D importers, exporters and translators are available for the plethora of different 3D graphics formats available.

X3D uses a scene graph to model a virtual environment. The scene graph is a tree structure that is directed and acyclic, meaning that there is a definite beginning for the graph, there are parent/child relationships for each node, and there are no cycles (or loops) in the graph. This scene graph collects all aspects of a 3D scene in a hierarchical fashion that properly organizes geometry, appearance, animation and event routing.

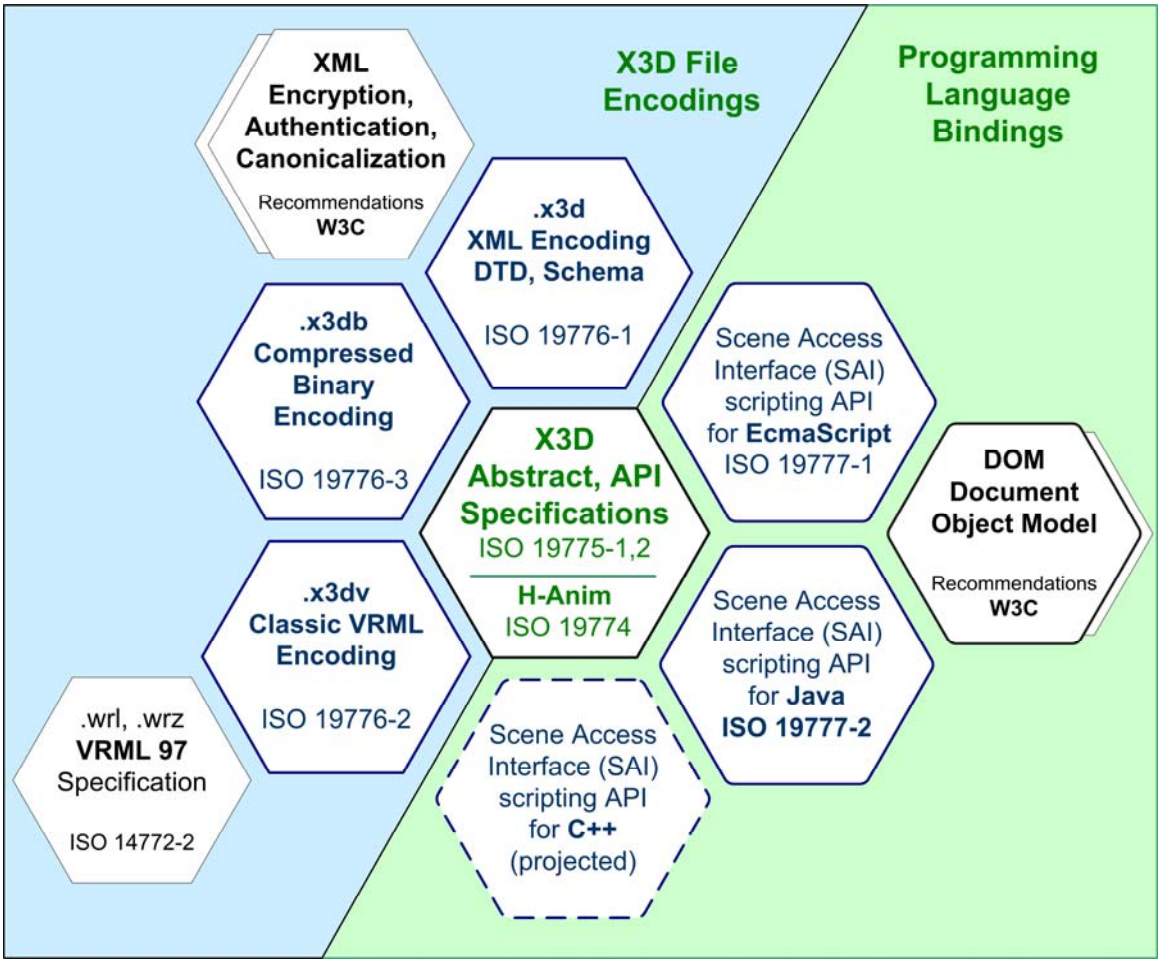

Figure 1. X3D file encodings and programming-language bindings equivalently implement a common abstract functionality.

Building scenes is more like a creating a Computer-Aided Design (CAD) model or authoring web-page content than programming. Using the .x3d encoding lets XML validation ensure that scene content remains free of errors, thereby enabling authors to focus on 3D modeling rather than syntax. This is more productive than worrying about how to achieve consistent results in different hardware and software environments.

Several key distinctions of interest between video and 3D graphics relate to camera viewpoint. In a recorded video, the camera viewpoint is fixed and unmodifiable. In a live video, camera position and orientation may be modifiable, but image data is only recorded for the given field of view. In 3D virtual environments, the surrounding scene is thoroughly defined and one or more cameras can be independently, simultaneously controlled. In effect, the viewpoint presented to the user is a virtual camera that can itself be animated and moved throughout the virtual environment.

Scope. Throughout this paper, various technical and social requirements are posed using the X3D graphics standard. Nevertheless, these principles also largely apply to a large variety of commercial and nonstandard 3D virtual environments (Linden Labs' *Second Life*, Forterra *OLIVE*, Sun *Wonderland*, etc.) [Sanders 2007].

Digital maps and 3D virtual worlds are also becoming more feature-rich and popular at a rapidly increasing rate. Many types of media can be integrated either automatically through sophisticated approaches and methods (local search in Google Maps and Google Earth) or through manual user interaction (photo placement in Microsoft Virtual Earth). NASA's World Wind integrates many different layers from a diverse set of geo-referenced data sources including weather data. This data is made available through online repositories, streamed over the web as the area in question comes into view at the client's site.

In partnership with the Open Geospatial Consortium (OGC), the Web3D Consortium's X3D Earth working group is producing high-fidelity global models using the X3D Geospatial Component. Scene authors are able to access and utilize a broad variety of geospatial data without restrictions. Furthermore, multiple globe-generating codebases are now available that can convert diverse terrain, bathymetry and imagery datasets into mutually compatible X3D Earth archives. Optimization of X3D Earth scene-graph design patterns and improved interoperability with OGC standards is ongoing. [X3D Earth 2007] [OGC 2007]

Current Functional Requirements. Support for a broad variety of technology and policy requirements is needed to make the widespread use of video together with 3D virtual environments both practical and effective. Demonstrated, existing capabilities are listed first.

- *Linkability*. Video content needs to be identified via a Uniform Resource Identifier (URI) to be referencable within an X3D scene. Ideally, addressing schemes for bookmarks or timestamps within a video clip can be compatibly provided as part of the URI. [Berners-Lee et al., ,January 2005]
- *Internal playability*. Video content is most often played internally, embedded within a virtual environment as a user-facing billboard or overlay.
- *External playability*. Video content might also be launched by an X3D browser into an external player as part of a multimedia presentation.
- *Interoperability*. While support for diverse image and audio formats is allowed and encouraged, X3D browsers are only required to support a small number of formats: PNG, JPEG and GIF for 2D images, WAV for audio, and MPEG-2 for video. (MIDI and MP3 are also recommended for audio support.) Because the X3D `url` field is an ordered list of URI addresses, authors can offer multiple formats to an end-user's browser, thus allowing format flexibility while retaining guaranteed interoperability among baseline formats. Because the integration of multimedia within an X3D scene is performance-intensive and requires large codec implementations, and also because licensing restrictions prevent the broad use of many commercial video encodings, X3D browsers cannot add arbitrary video encodings without significant cost. Adoption of imagery, audio and video formats used by W3C recommendations holds great interest to X3D software and content producers.
- *Animation*. Synchronization of video playback with animation behaviors between the event-passing model within an X3D scene is essential for consistent playability. An important action item is to correlate X3D timing structures with Synchronized Multimedia Interface Language (SMIL).
- *Metadata*. Numerous metadata requirements are shared between video recordings and geometry-based 3D scenes. Duration is needed when a video clip will be looped (repeated), although this parameter may be derived if the video file is fully available and not streamed. Recorded time and location may be needed for placing the video clip

properly within the virtual environment. Camera direction may also be needed if the video clip is being spatially aligned with geometry, or if geometry is being generated from the imagery. Other metadata associated with the video clip needs to be inspectable by the 3D browser in order to utilize special features or directly provide metadata to users.

- *Construction of 3D geometry*. Structured video can now be used to generate polygonal structures to populate a 3D scene. Shape-from-shading and computer-vision techniques can automatically produce meaningful shapes such as buildings, trees, etc. Typically camera position must be carefully registered in space throughout the video capture.

Additional Functional Requirements. A number of essential future capabilities for X3D and virtual environments are likely and also worth considering.

- *Security*. Scene authors often want to encrypt or authenticated X3D content. The X3D Compressed Binary Encoding (CBE) currently utilizes the ISO-approved Fast Infoset (FI) standard as the basis for non-geometric compression. X3D plans to transition the CBE to the Efficient XML Interchange (EXI) once that format achieves final approval as a W3C recommendation. In either case, use of the W3C Security recommendations is expected to become the primary mechanism for protecting content (via encryption) and authenticating ownership (via digital signature). [EXI 2007]
- *Streamability*. Historically, streaming capabilities have been provided in numerous 3D virtual environments. Network streaming capabilities for video highly desirable, presumably using SMIL or the Real-Time Streaming Protocol (RTSP).
- *Synthesis*. Applications such as motion capture, which precisely record the positions and orientations of marker points on human bodies, are used to generate comprehensive data files and filtered behaviors which are then used for geometry animation. The ability to synthesize correlated video and geometric information is expected to become an important future capability.
- *Transparency*. A valuable future feature will be the ability to embed transparency or masking information within video imagery, similar to transparency within individual 2D images. The option of embedded transparency within a video stream can permit superposition of imagery of interest, minus backgrounds or non-applicable imagery. Assigning masking colors as metadata may be an effective way to achieve video transparency without adding alpha (opacity) information to individual pixels.
- *Export*. A recent practice finding increasing use is end users recording their interactions within a virtual environment and publishing it as video. Establishing common metadata, linking and synchronization conventions can encourage geospatial and temporal consistency, and may encourage further interoperability among Web applications using video, hypermedia and 3D graphics.

Case study: synthesis of 3D geometry from video. For this position paper, it is interesting to consider the particular case of integrating pictorial information into virtual worlds at their geo-accurate locations. This scenario is quite different from more typical uses of Web video, exercises a number of X3D-related functional requirements simultaneously, and also hints at current research directions.

In the simple case for synthesizing 3D geometry, the photo or video is tagged with the approximate location at which it was taken and displayed upon demand in a full-screen fashion. More challenging, however, is the geo-referenced display in which the imaged position overlays

with the exact same virtual position. If the exact camera poses and lens characteristics are known, algorithms exist to perform this registration. This is rarely available, however. Image and video analysis methods will have to make up for the gap of knowledge about pose and camera calibration. This has been addressed in a number of research projects, namely Microsoft Research's PhotoSynth and UW's project on "Multi-View Stereo for Community Photo Collections". It is not clear, however, how to perform similar feats in online worlds and at a much wider scale of imaging data, many more diverse objects.

The following metadata is needed for geo-registration of video imagery:

- Camera position, in the same coordinate system as the virtual world,
- Camera orientation, in a world-centered coordinate system,
- Camera calibration, which includes focal length (changes with zoom!) and lens distortions.
- If photometric registration is required, we further need information about exposure times, lens intensity and color aberrations.
- If temporal registration, we obviously need the time the image was taken or the time each video frame was taken.
- If object-level registration is desired, automatic object detection and extraction must be performed. This is useful, for example, to create consistent representations of vehicles that are visible concurrently in multiple views.

Summary. The broad application area of online virtual environments requires a significant number of important technical and policy requirements pertinent to the goals of video on the Web. This point paper lists current and expected requirements, primarily divisible into usage of video within X3D graphics scenes, linkage to video in web-based applications external to X3D graphics scenes, and generation of 3D geometric content from spatially annotated video inputs. Royalty-free video capabilities are critical important to achieve essential requirements for interoperability and performance. Standards-based X3D requirements also appear to be representative of the needs presented by alternative proprietary multiuser virtual environments. Further integration and cross-fertilization is expected as Web-based virtual environments are constructed in X3D that freely utilize Web-based video content.

References.

Berners-Lee, T., R. Fielding, and L. Masinter, *Uniform Resource Identifier (URI): Generic Syntax*, IETF RFC 3986, January 2005, http://www.ietf.org/rfc/rfc3986.txt.

Brutzman, Don, and Daly, Leonard, *X3D: Extensible 3D Graphics for Web Authors*, Morgan Kaufmann Publishing, 2007. 468 pages, book website is http://x3dGraphics.com.

Open Geospatial Consortium (OGC) website, http://www.opengeospatial.org.

Sanders, Kent, *Creating a Persistent, Open-Source Mirror World for Military Applications*, Masters Thesis, Naval Postgraduate School, Monterey California, November 2007. Co-advisor Amela Sadagic.

World Wide Web Consortium (W3C) Efficient XML Interchange (EXI) working group, http://www.w3.org/XML/EXI.

Extensible 3D (X3D) Graphics website, http://www.web3D.org/x3d.

X3D Earth working group website, http://www.web3D.org/x3d-earth.

X3D Specifications, http://www.web3d.org/x3d/specifications.