Identifying Spatial and Temporal Media Fragments on the Web

Raphaël Troncy[1], Lynda Hardman*[1], Jacco van Ossenbruggen[1], and Michael Hausenblas[2]

[1] CWI, Amsterdam, The Netherlands, {FirstName.LastName@cwi.nl}
[2] JOANNEUM RESEARCH, Graz, Austria, {Michael.Hausenblas@joanneum.at}

1 Introduction

Semantic descriptions of non-textual media available on the Web can facilitate retrieval and presentation of media assets and documents that contain them. Semantic Web languages can represent controlled vocabularies and shared annotations of media content on the Web. By identifying concepts to consider, Uniform Resource Identifiers (URIs) are the building blocks of the Semantic Web. RDF subject-predicate-object triples provide the mortar by specifying relations between them.

Often, particular regions of an image or particular sequences of a video need to be localized (anchor value in [1]) and uniquely identified in order to be used as `subject` or `object` resource in an RDF annotation. However, the current Web architecture does not provide a means for uniquely identifying sub-parts of media assets, in the same way that the fragment identifier in the URI can refer to part of an HTML or XML document. Actually, for almost all other media types, the semantics of the fragment identifier has not been defined or is not commonly accepted.

The URI specification defines the general meaning for `scheme#fragment`, and, for example, when the scheme is `http`, the RFC2616 specifies that a `HTTP GET` has to be performed to find out what the `fragment` is, yielding a certain `Content-Type` in the response (i.e. the mime-type of the fragment). The mime-type registry[3] specifies what the fragment means within a document depending on its type. For example, for `'text/html'` the RFC2854 defines that the fragment is actually a part of the document identified by an anchor.

Providing an agreed upon way to localize sub-parts of multimedia objects (e.g. sub-regions of images, temporal sequences of videos or tracking moving objects in space and in time) is fundamental[4] [2, 3, 4, 5, 6]. The requirements for expressing and processing these fragments have been studied [7]. This position paper describes several ways for identifying fragments of multimedia content on the Web using W3C recommendations, ISO standards or RFC.

* Lynda Hardman is also affiliated with the Technical University of Eindhoven.

[3] `http://www.iana.org/assignments/media-types/`

[4] See also the related discussion in the W3C Multimedia Semantics XG `http://lists.w3.org/Archives/Public/public-xg-mmsem/2007Apr/0007.html`.

2 Identifying Spatial Fragments

Imagine that Nathalie, a history student, wants to create a multimedia presentation of the major international conferences and summits held in the last 60 years. Her starting point is the famous “Big Three” picture[5], taken at the Yalta (Crimea) Conference, showing the heads of government of the United States, the United Kingdom, and the Soviet Union during World War II (figure 1). Nathalie could either use an automatic face detection and recognition web service, or draw manually the bounding boxes around Winston Churchill, Franklin D. Roosevelt and Josef Stalin. She would like then to link the face regions to detailed textual information about these characters.

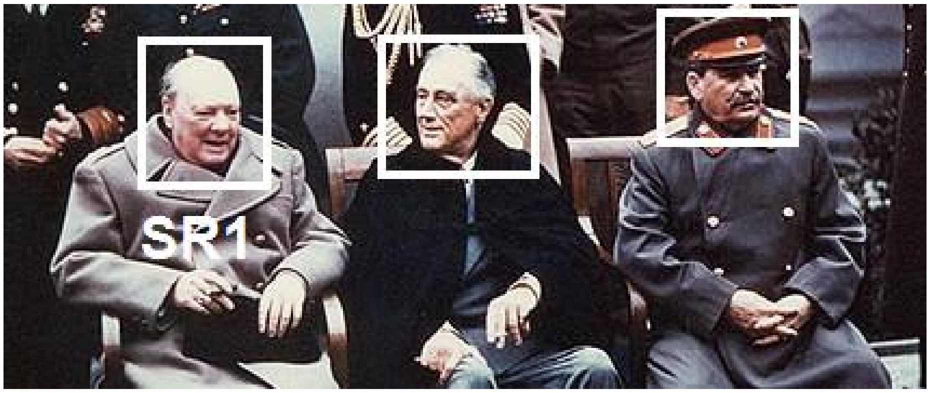

Fig. 1. The “Big Three” at the Yalta Conference (Image adapted from Wikipedia)

SVG approach

For spatial location, one can use an SVG [8] code snippet that defines the bounding box coordinates of specific regions. For example, the code below defines a rectangle on the image jpg resource:

```
<svg  xmlns:svg="http://www.w3.org/2000/svg"
      xmlns="http://www.w3.org/2000/svg"
      xmlns:xlink="http://www.w3.org/1999/xlink">
  <g id="layer1">
    <image
      xlink:href="http://upload.wikimedia.org/wikipedia/commons/1/1b/Yalta_Conference.jpg"
      x="-0.34" y="0.20" width="400" height="167" id="image_yalta" />
    <rect
      x="14.64" y="15.73" width="146.98" height="147.48" id="sr_churchill"
      style="opacity:1;fill:none;fill-opacity:1;fill-rule:nonzero;stroke:#ff0000;stroke-opacity:1"/>
  </g>
</svg>
```

[5] http://en.wikipedia.org/wiki/Yalta_Conference

MPEG-7 approach

Alternatively, one can use a MPEG-7 [9] snippet code for defining the same region:

```
<Image id="image_yalta">                             <!-- whole image -->
  <MediaLocator>
    <MediaUri>http://upload.wikimedia.org/wikipedia/commons/1/1b/Yalta_Conference.jpg</MediaUri>
  </MediaLocator>
  [...]
  <SpatialDecomposition>
    <StillRegion id="sr_churchill">                  <!--    still region     -->
      <SpatialMask>
        <SubRegion>
          <Box>14.64 15.73 161.62 163,21</Box>
        <SubRegion>
      </SpatialMask>
    </StillRegion>
  </SpatialDecomposition>
</Image>
```

However, both approaches require an indirection. Assuming these XML (MPEG-7 or SVG) descriptions are identified by a URL, an RDF annotation will be *about* a fragment of this XML document that *refers* to the multimedia document (i.e. the image).

3 Identifying Temporal Fragments

Nathalie would then like to describe a recent video from a G8 summit, such as the retrospective *A history of G8 violence* made by Reuters[6]. She would like to describe precisely each sequence in the news report. She could either do the video decomposition manually in her favorite authoring environment or use again an automatic segmentation tool for detecting the seven main sequences of this 2’26 minutes report: the various anti-capitalist protests during the Seattle (1999), Melbourne (2000), Prague (2000), Gothenburg (2001), Genoa (2001), St Petersburg (2006), Heiligendamm (2007) World Economic Forums, EU and G8 Summits (figure 2).

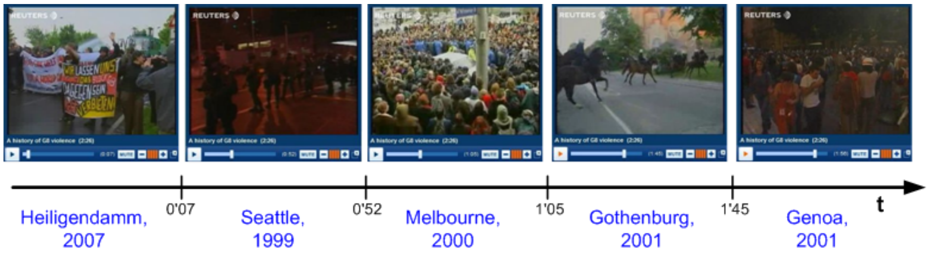

Fig. 2. A history of G8 violence (©Reuters)

[6] http://www.reuters.com/news/video/summitVideo?videoId=56114

SMIL approach

For temporal location, one can use the following SMIL [10] code:

```
<smil xmlns="http://www.w3.org/2001/SMIL20/Language">
 <head>
  <layout>
   <root-layout width="640" height="480"/>
   <region id="video_G8"/>
  </layout>
 </head>
 <body>
  <seq>
   <video src="http://int1.fp.sandpiper.net/reuters/t_assets/20070608/85c5b86bd03020c63e8976db406a3aa1efe8c1f3.flv"
          region="video_G8" clipBegin="3" clipEnd="9"/>
   <video src="http://int1.fp.sandpiper.net/reuters/t_assets/20070608/85c5b86bd03020c63e8976db406a3aa1efe8c1f3.flv"
          region="video_G8" clipBegin="44" clipEnd="55"/>
   [...]
  </seq>
 </body>
</smil>
```

MPEG-7 approach

Alternatively, one can use a MPEG-7 [9] code snippet for defining the same video sequences:

```
<VideoSegment id="video_G8">                        <!-- whole video -->
  <MediaLocator>
    <MediaUri>
http://int1.fp.sandpiper.net/reuters/t_assets/20070608/85c5b86bd03020c63e8976db406a3aa1efe8c1f3.flv
    </MediaUri>
  </MediaLocator>
  [...]
  <TemporalDecomposition gap="true" overlap="false">
    <VideoSegment id="seq_1">                       <!--    sequence 1     -->
      <MediaTime>
        <MediaTimePoint>T00:00:03:0F30000</MediaTimePoint>
        <MediaDuration>PT00H00M06S26116N30000F</MediaDuration>
      </MediaTime>
    </VideoSegment>
    [...]
  </TemporalDecomposition>
</VideoSegment>
```

Again, both approaches require an indirection. Assuming these XML (MPEG-7 or SMIL) descriptions are identified by a URL, an RDF annotation will be *about* a fragment of this XML document that *refers* to the multimedia document (i.e. the video). On the other hand, MPEG-7 can be used to specify very complex segments (masks) such as the union of not temporally connected sequences, or the tracking of moving regions over time.

Temporal URI approach

The TemporalURI is an RFC[7] that does not have this limitation as it specifies a generic URI syntax for identifying temporal fragments of video on the web. For example, the first sequence of the G8 video could be directly dereferenced by the following URI:

```
http://int1.fp.sandpiper.net/reuters/t_assets/20070608/85c5b86bd03020c63e8976db406a3aa1efe8c1f3.flv#npt:0:00:03-0:00:09
```

MPEG-21 approach

MPEG-21 specifies also a normative syntax to be used in URIs for addressing parts of any resource but whose media type is restricted to MPEG [11]. If the

[7] http://www.annodex.net/TR/URI_fragments.html

mime type of the video would have been MPEG, the following URI would also have identified the first sequence of the G8 video:

```
http://int1.fp.sandpiper.net/reuters/t_assets/20070608/85c5b86bd03020c63e8976db406a3aa1efe8c1f3.flv#ffp(item_ID=_seq1-video)*mp(/~time('npt','0:00:03','0:00:09'))
```

Both TemporalURI and MPEG-21 approaches do not suffer from the indirection problem explained above. However, the expressivity for representing complex fragments is reduced.

4 Conclusion

Providing a standardized way to localize spatial and temporal sub-parts of any non-textual media content is now urgently needed to make video a first class citizen on the Web. Any proposed solution should be compatible with the 'http-range-14' TAG finding and follow the "cool URIs"[8] good practices.

For many media types, one could define a simple fragment identifier syntax for direct URI reference. We believe the problem is mainly a social one: it just has not been done yet. The Web community is looking to the multimedia community to do it, but the multimedia community does not care enough. Given the amount of work already done in this area, we suggest that a W3C REC track should be straightforward. For the more complex localization, as stated above, it would most likely require an indirection. This should be further investigated in the Semantic Web activity for the possible consequences with the RDF model and semantics.

Acknowledgments

The research leading to this paper was partially supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content – K-Space.

References

[1] Halasz, F., Schwartz, M.: The Dexter Hypertext Reference Model. Communications of the ACM **37**(2) (1994) 30–39

[2] Ossenbruggen, J.v., Nack, F., Hardman, L.: That Obscure Object of Desire: Multimedia Metadata on the Web (Part I). IEEE Multimedia **11**(4) (2004)

[3] Nack, F., Ossenbruggen, J.v., Hardman, L.: That Obscure Object of Desire: Multimedia Metadata on the Web (Part II). IEEE Multimedia **12**(1) (2005)

[4] Geurts, J., Ossenbruggen, J.v., Hardman, L.: Requirements for practical multimedia annotation. In: Workshop on Multimedia and the Semantic Web. (2005)

[5] Arndt, R., Troncy, R., Staab, S., Hardman, L., Vacura, M.: COMM: Designing a Well-Founded Multimedia Ontology for the Web. In: 6^{th} International Semantic Web Conference (ISWC). (2007)

[8] http://www.w3.org//2001/sw/sweo/public/2007/cooluris/

[6] Jochum, W.: Requirements of Fragment Identification. In: International Conferences on New Media Technology and Semantic Systems, Graz, Austria (2007) 172–179

[7] Rutledge, L., Schmitz, P.: Improving Media Fragment Integration in Emerging Web Formats. In: The International Conference on Multimedia Modeling (MMM). (2001) 147–166

[8] W3C: Scalable Vector Graphics (SVG) 1.1 Specification. W3C Recommendation (2003)

[9] MPEG-7: Multimedia Content Description Interface. ISO/IEC 15938 (2001)

[10] W3C: Synchronized Multimedia Integration Language (SMIL 2.1). W3C Recommendation (2005)

[11] MPEG-21: Part 17: Fragment Identification of MPEG Resources. ISO/IEC 21000-17 (2006)