

Interactive Maps for a Digital Video Library

Michael G. Christel

*CS Dept. and HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
+1 412 268 7799
christel@cs.cmu.edu*

Andreas M. Olligschlaeger

*Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213 USA
+1 412 268 5137
olli@cs.cmu.edu*

Abstract

The Informedia Digital Video Library contains over 1200 hours of video. Through automatic processing, descriptors are derived for the video to improve library access. A new extension to the video processing is the extraction of geographic references from these descriptors. The operational library interface shows the geographic entities addressed in a given story, highlighting the regions discussed at any point in the video through a map display synchronized with the video playback. The map can also be used as a query mechanism, allowing users to search the terabyte library for stories taking place in a selected area of interest.

1. Introduction

The Informedia Project at Carnegie Mellon University investigates the utility of speech recognition, image processing, and natural language processing techniques for improving search and discovery in the video medium. Since 1994, the project has been digitizing, in MPEG-1 format, news video from CNN as well as documentary/educational video from the British Open University, QED Communications, NASA, the Discovery Channel, and numerous other United States government agencies such as the National Park Service and U.S. Geological Survey. The resulting digital video library now contains over 1200 hours of video, and continues to grow at a rate of 10 hours per week [1].

The sheer volume of the video data reveals new issues for interfaces to digital video libraries of the future. A good query engine is not sufficient because often the candidate result sets grow in number as the library grows. Interfaces for browsing both the library and defined library subsets such as the results from a query become increasingly important.

Users are interested in quickly finding the set of video stories or segments relevant to their needs. When the library was on the order of a hundred hours, a statistical word query engine adequately provided this focus. Users entered text queries, and a small set of segments was returned sorted by the query engine's relevance score. Alternate representations of the video that could be viewed in less time were presented to aid the user in deciding which of these segments was worth a full viewing [2]. This style of library interface is shown in Figure 1.



Figure 1. Informedia interface with images representing 6 video segments, and text headline for the second segment

When the library grew to a thousand hours, queries returned hundreds of segments, overwhelming users

much like Web search engines can return lists whose length and default ordering no longer meet the needs of the user. An information visualization interface was developed to let the user browse the whole result space without having to resort to the time-consuming and frustrating traversal of a list of results [3]. The employed visualization techniques allowed the user to browse and retrieve video from the Infromedia library based on date (i.e., “when”) and word occurrences (i.e., “what”). We realized that a potentially rich vein of information was being overlooked in our corpus, however. Many documentaries and most news stories deal with location information (i.e., “where”). This information dimension could also be used in presenting overviews of the video content, summarizing multiple video segments, and as a query mechanism to find segments dealing with a particular region of interest. The remainder of this paper discusses the use of interactive maps with the Infromedia Digital Video Library.

2. Extracting Video Geographic References

The transcript of the narrative is the greatest source of geographic reference information for the videos in the Infromedia library. The Carnegie Mellon University Sphinx speech recognition engine is used to transcribe the content of the video material, with word error rate proportional to the amount of processing time devoted to the task [1]. If closed-captioned text exists for a video, it is integrated with the output of the recognizer. The final text transcript is synchronized at a word level to the video through Sphinx processing. Hence, if the narrative mentioned “Heavy snows in Switzerland caused...”, the video time when “Switzerland” was mentioned would be captured by this process.

While the transcript provides the primary source of geographic references, it is not the sole source. Often a location name and perhaps a person’s name are overlaid on the video, especially for news. The Infromedia Video OCR (VOCR) process [4] identifies video frames containing probable text regions, in part through horizontal differential filters with binary thresholding. VOCR then filters the probable text region across the multiple video frames to improve the quality of the image used as input for OCR processing. Commercial OCR software converts the final filtered image of alphanumeric symbols into text. The VOCR-produced text is another potential source of geographic references. For example, a video segment discussing volcanic activity included shots of lava with the overlaid text stating “Mount Etna, Italy.” While the transcript text was associated to video times through Sphinx speech alignment, the VOCR text is associated to video times

through image processing which identifies the frames containing the probable text regions.

Other descriptors for the Infromedia library contents, i.e., metadata, include production notes, automatic topic identification, and user annotations whereby the user can type or speak comments pertaining to a specified portion of video [1]. These additional sources of text may also include location information, such as a user comment about “Add Los Angeles to my itinerary.”

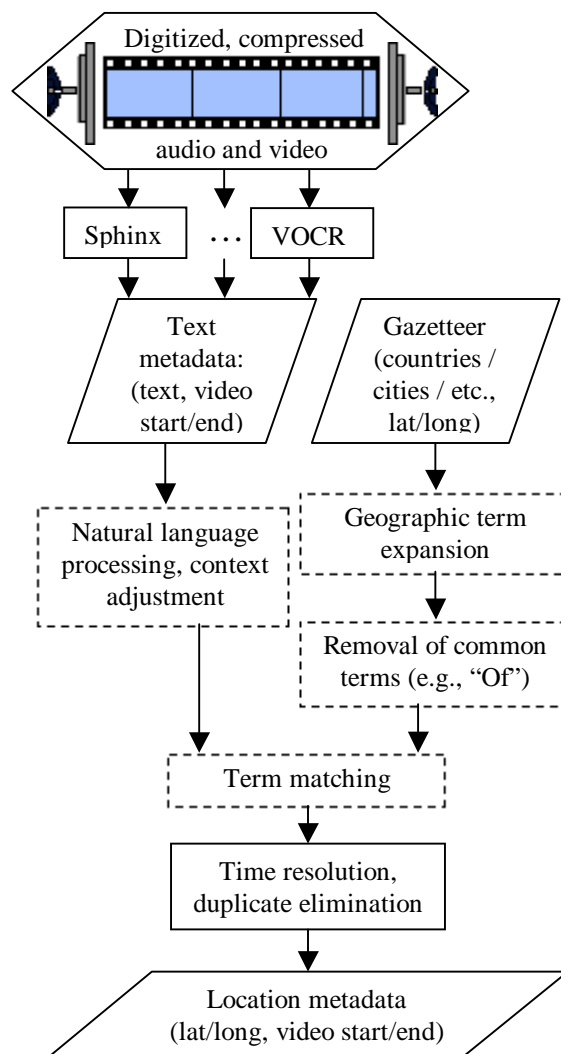


Figure 2. Adding location information to video

Figure 2 shows the process for adding geographic references to video segments. Matching addresses, or in this case, named places, to their spatial coordinates is known as geocoding [5]. Geocoding video begins by using the text metadata as the source material to be processed. A known set of places along with their spatial coordinates, i.e., a gazetteer, is chosen to resolve geographic references. The gazetteer used with the

Informedia library is derived from a subset of data contained in the world gazetteer from Environmental Systems Research Institute (ESRI) [6], consisting of approximately 300 countries, states, and administrative entities and 17,000 major cities worldwide. A post-processing step is added to expand the gazetteer to include related terms, e.g., “Canada” will identify that country but so will “Canadian.” Further post-processing removes common terms from subsequent matching, such as the city of “Of.” The text metadata, which associates text with video times, is then matched to terms in the geographic “code book” which maps geographic text terms to latitude and longitude. The end result is the tagging of video sequences with latitude and longitude.

Geocoding addresses in fixed field format, i.e., with known address components such as street number, street name, and city, has been well researched and documented [5]. However, extracting named places from free-form text such as video transcripts is a relatively new idea and far more complicated than geocoding addresses. Addresses tend to be unique, e.g., there is only one place called 123 Main St. in Anytown, and, providing they are in a consistent format, can simply be matched against a known set of addresses whose coordinates are also known. With free-formatted text, on the other hand, it is not known a priori which portion of the transcript contains references to places.

Thus, it is necessary to parse the text metadata to extract candidate portions of sentences that may represent places. Currently this is done simply by eliminating those words that are most commonly used, such as “and” and “the”, and examining the remainder of the text. The dotted items in Figure 2 indicate portions of the process that can be greatly improved with the addition of new knowledge, including this parsing step.

If we know that “Of” is part of the phrase “Of, Turkey” in the text metadata, then we can use that contextual information to resolve “Of” to be a city, where “of” in the phrase “things of the past” would not resolve to any location information. Similarly, context can help with the term matching. The proper noun “Washington” could refer to a western state in the United States, its capital city, or could be a person’s name. Contextual cues such as “Seattle, Washington”, “Washington, D.C.” and “since the time of George Washington” distinguish these different meanings. Through contextual analysis on the source metadata, the system will be able to better classify proper nouns as persons’ names or places, and more accurately assign location information.

Hidden Markov models (HMMs) can be used to achieve this level of analysis. HMMs have proven effective for automatically tagging entities, including locations, in text output from speech recognizers, where

such text lacks punctuation cues [7]. We are currently exploring the use of HMMs for the dotted tasks of Figure 2. We expect the HMM approach will deliver more accurate geographic referencing than our current quick processing baseline which uses little context adjustment, removes all common and ambiguous terms from the gazetteer match set, and enforces strict term matching. Further studies will be needed to determine the accuracy and benefits associated with the additional processing.

3. Enabling Geographic Reference Use

The unit of information retrieval in the Informedia library is the video segment, which (when segmentation strategies work to perfection) contains a single story. On average, each hour of broadcast news consists of 20 segments. For each segment, a list is constructed during the geocoding process consisting of those places that are mentioned in a segment. A place may be named more than once in a segment, but it is represented only once in the segment’s list. The number of references is included in the entry for each place to enable subsequent interfaces to emphasize locations visually based on how frequently the places are mentioned.

The geocoding process establishes a relationship between the video and place names. For a given video time interval, the place names referenced in that interval can be identified, and for a given place, its time interval is quickly accessible. For places identified in transcript metadata, the sentence or sentences where each place is mentioned is tracked. If a place is only mentioned once, the beginning and ending times, in milliseconds from the beginning of the video, of the sentence in which it is mentioned are used as the time interval for that place. If a place is mentioned more than once, the time span from the start of the first sentence containing the place reference to the end of the last sentence containing a mention of the place is used. The timing of transcript words to the video is accessible via the Informedia Sphinx speech recognition processing [1]. If a place is identified from other text metadata such as from VOICR, then the start and end time associated with the text for that metadata is used. For VOICR, this time span approximates the duration when the overlaid text appears in the video. As with transcripts, if a place occurs multiple times then its time span extends from the start time for its first mention to the end time for its final mention in the video segment.

The resulting list is written to a database (currently in .dbf format), and converted to a shape file in ESRI format using the geocoding capabilities of ESRI’s ArcView and ESRI’s gazetteer. Finally, the shape file is indexed geographically in order to optimize spatial

searches. Table 1 shows a few entries and attributes as an example of the output from the geocoding algorithm.






Table 1. Example entries (not all rows nor columns shown) for results of geocoding

| Text | Type | X | Y | Admin. | Country | SegmentID | Start time (ms) | End time (ms) |
|-------------|-------|---------|--------|-----------|----------|-----------|-----------------|---------------|
| KENYA | CNTRY | 37.915 | 2.605 | EASTERN | KENYA | CWT0Z4 | 204704 | 366366 |
| INDIA | CNTRY | 82.410 | 20.605 | MADHY ... | INDIA | CWT0Z4 | 362095 | 362095 |
| NAIROBI | CITY | 36.804 | -1.270 | NAIROBI | KENYA | CWT0Z4 | 195329 | 375209 |
| CHINA | CNTRY | 108.986 | 36.628 | SHAANXI | CHINA | CWT1F16 | 1665665 | 1667667 |
| SOUTH KOREA | CNTRY | 127.772 | 36.711 | CH'UNG... | S. KOREA | CWT1F16 | 1738805 | 1738805 |
| JAPAN | CNTRY | 137.795 | 35.551 | CHUBU | JAPAN | CWT1F16 | 1635068 | 1721054 |
| NORTH KOREA | CNTRY | 126.538 | 39.105 | P'YONG... | N. KOREA | CWT1F16 | 1629530 | 1754087 |
| BRUSSELS | CITY | 4.368 | 50.837 | BRUXEL... | BELGIUM | CWV1031 | 3248748 | 3250250 |

4. Map Representation of a Video Segment

Figure 3 shows an example of how the geocoded information is incorporated into the Informedia interface. When a video segment is played back, an optional window pops up that contains a map displaying all of the places discussed in that segment. Functionality provided by the ESRI MapObjects library is used in creating this interface, written using Microsoft Visual Basic. A glance at this overview shows that the given segment covers the countries of Kenya and India and the city of Nairobi.

The user can interact with the map through the use of the toolbar icons, which from left to right support the operations of:

-  Zooming in to reveal more detail and less area
-  Zooming out to show more area and perhaps less detail
-  Panning to a different area at the same resolution
-  Selecting an area to search (discussed in the next section)
-  Returning to the full map extent which shows the countries of the world

Aside from the spatial search, these icons and their underlying operations are supplied directly by the MapObjects library. By integrating this functionality into a digital video library interface, the user can access detail relevant to the video contents.

For example, the video segment is paused in Figure 3 at a point where the CNN footage shows a map, illustrating that the producer of the CNN video recognizes the importance of geographic detail. This same detail can be automatically extracted and displayed through the geocoding process of Figure 2, generic map functionality such as that provided by MapObjects, and user interaction. In Figure 3 the user has checked the box to enable “tips” text to be displayed based on mouse movement over the map. When the mouse is paused over

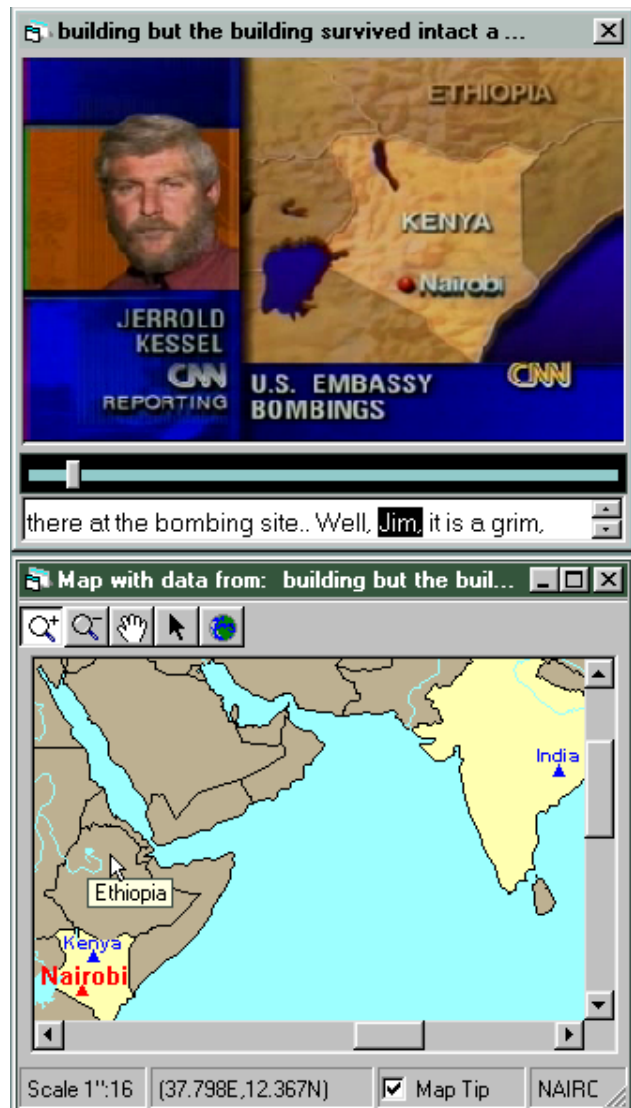


Figure 3. Map overview of location references for a video segment

Ethiopia, that country's name appears in a "tips text" window. The resulting map display looks very similar to the produced CNN shot.

Of course the real value to the geocoding and map interface is in displaying location information for video segments for which the producer has not previously added a map within the video data. Every news story does not have an embedded map that becomes part of the broadcast, but through our geocoding maps can be automatically produced to reflect the areas discussed within each story. Another benefit is that the user can interact with the interface map using the toolbar icons to get additional detail, whereas no such interaction is possible with an image of a map encoded as part of the video stream. For example, the user could zoom into the Kenya area of Figure 3 to see a map display with other Kenyan details such as the port city of Mombasa.

The maps accompanying videos are not static displays. They are animated in synchronization with video playback. As places are discussed, they are highlighted on the map. For countries and administrative areas such as states or provinces, the areas contained within their respective polygon boundaries are highlighted by changing the color with which they are shaded. Areas covered at some time during a video segment are colored yellow; when the video frames during which an area is discussed are played, that area is then colored orange. For cities and other places, the marker color is changed from blue to red and the label is shown. A glance at the map can then show the areas of current focus, e.g., Figure 4 shows a story where the current focus is on North Korea and Japan, with China, South Korea and Russia mentioned elsewhere in the story. The highlighting changes over time to show the story flow within the video segment.

A classic research area in cartography studies the accurate and effective display of data on a map; animated maps are only just beginning to be addressed. Depending on the number of features shown, it is important to avoid "noisy" maps that show much detail, but are difficult to read. Thus, we are currently limiting the appearance of text labels for city and administrative area names only to those times when they are discussed in the accompanying video. Country labels are always displayed. In Figure 3, the video is paused at a point where Nairobi is being discussed, and so that label appears in red. If the portion of video being played in that segment no longer actively references Nairobi, as indicated by the times stored in the database (see Table 1), then the marker for the Nairobi location remains visible, but in a different color and without its text. This strategy is useful when a video segment has many city references across a broad area such that the labels for all references could not be drawn



Figure 4. Animated map that highlights as video plays (showing scrolling transcript as well)

without significant overlap.

Since the amount of information displayed for each video segment varies, we plan to allow the user to change the default settings of the map display in the future. These settings include the types of entities to mark (e.g., cities and countries), the symbols and labels to use, and

the colors and styles for marking both the overview (as in Figure 3) and the highlights for a given video time (as in Figure 4). In addition, the default settings themselves may well depend on what type of video is being shown. For example, broadcast news that tends to mention only major cities and countries may be well served by our current default settings. However, a corpus that contains only videos for a European soccer league may well have only city labels displayed by default.

5. Accessing Video through Spatial Queries

The maps in the Infromedia interface are not merely for presentation but also can be used to specify a location query. The arrow icon in the toolbar for the map window is used to drag a rectangular region on the map that serves to identify the user's region of interest.

In Figure 5 the user has selected the region encompassing the Netherlands, Belgium, and Luxembourg. Functionality provided by the ESRI MapObjects library is used to perform the query against the video library's associated geographic references. The Infromedia library currently contains around 40,000 segments, with geocoding producing nearly 20,000 location references, i.e., rows of the form shown in Table 1. Searching against this corpus, within a few seconds the results are displayed with headlines and representative thumbnail images, just as results for word queries are displayed (as shown in Figure 1). Feedback is provided on the map to indicate the locations within the specified query that actually produced results. In Figure 5 this feedback is shown as white circles for the three countries, indicating that each country was found somewhere in the result set of 46 video segments. The headline for the sixth result shows that it is a story on a Belgian roller coaster. If this video is subsequently played, the map would change to an image like that of Figure 4, with Belgium initially colored to show that it is mentioned somewhere in the story and with that color changing during the video frames when it is actually discussed.

Figure 4 originated with the spatial query shown in Figure 6. When users search with words, the timing for words is used to mark the matching word locations within the video [2]. These match locations are indicated with vertical lines drawn in the video scroll bar shown beneath the video playback area. When the video segment is found via a spatial query, the same match location reporting can be used. Numerous matching places may be found within a given rectangular query area, with each place having associated video times. These times are used to draw the match lines. In Figure 4, the three match lines on the video scroll bar



Figure 5. Using the map to request video segments which deal with a specified location

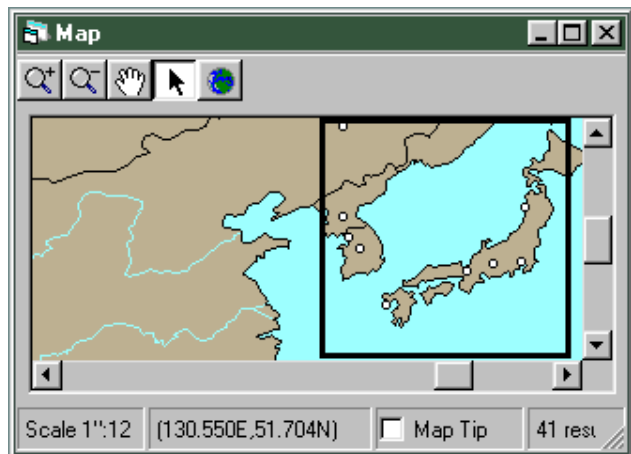




Figure 6. Spatial query which produced the video segment of Figure 4 as one result

correspond to the times when North Korea, Japan, and South Korea are mentioned in the segment, these three places being in the region of interest defined by the query shown in Figure 6. Through the use of additional

controls, e.g., the  and  buttons, the user can quickly seek to the video where the locations of interest are being discussed.

While a query region shaped as a rectangle is currently operational, we recognize the value in supporting more powerful spatial query mechanisms. For example, the user might click on a country, shift-click a number of countries, or use a bounding polygon to select a number of countries within a region. All matches to cities, political entities, countries, and other geographic references within that area would then be included in the returned set of video segments. This extension from rectangular searches to searches using more map information such as country boundaries, is provided in the features for many geographic information systems (GIS), including the ESRI MapObjects library.

There is also a hidden assumption that the user expects the drawn area to specify a request for all locations contained within that drawn area. In specifying an area request the user might expect an “overlaps” or “contains” relationship rather than the “contained by” relationship [8]. The work of the Alexandria Digital Library [8, 9] addressing this ambiguity in formulating spatial queries holds great promise for use with geocoded digital video libraries.

6. Future Work

The dotted boxes in Figure 2 show areas ripe for further work. Natural language processing and HMMs can more accurately extract place name candidates from text metadata, expand place aliases, resolve ambiguity, and expand the set of names that can be matched. The interface can be improved through better exposure and greater use of GIS functionality, leveraging from lessons learned from digital library projects that have focused on map information [8, 9]. Different symbols, colors, and sizes can be prototyped and empirically tested as to their effectiveness in communicating the information embedded within the video library.

The Informedia digital video library interface supports word query, image query [1], and now spatial query through the use of maps. One interesting area of work will be to enable mixed modal query, such as finding all the video segments mentioning “famine” for a specified area on the map, or finding all the faces like a given face for a specified country. The presentation of information can be improved by utilizing not only the information dimensions of date/time and topic, but also the dimension of location. Results from word queries could be visualized on a map, revealing interesting patterns for discovery, e.g. a search on volcanoes might show the

stories are concentrated in a ring around the Pacific Ocean. By empowering the user to manipulate all the metadata for the video library, time-based patterns could be revealed. For example, perhaps a search on terrorism and bombings will show hot spots of activity in one geographic area for early 1997 but new hot spots in different areas for late 1998. With the inclusion of geographic references, the Informedia interface can better serve users in their quest for search and discovery in the video medium.

7. Acknowledgements

This paper is based on work supported by the DARPA under SPAWAR contract No. N66001-97-D-8502. The support of Informedia partners and team members has been invaluable; a complete list can be found at <http://www.informedia.cs.cmu.edu/> along with further information on Informedia-related efforts.

8. References

- [1] H.D. Wactlar, M.G. Christel, Y. Gong, and A.G. Hauptmann, “Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library”, *IEEE Computer* **32**(2), 1999, pp. 66-73.
- [2] M.G. Christel, D.B. Winkler, D.B., and C.R. Taylor, “Multimedia Abstractions for a Digital Video Library”, *Proc. ACM Conference on Digital Libraries*, ACM, 1997, pp. 21-29.
- [3] M.G. Christel and D.J. Martin, “Information Visualization within a Digital Video Library”, *J. Intelligent Info. Systems* **11**(3), 1998, pp. 235-257.
- [4] T. Sato, T. Kanade, E. Hughes, and M. Smith, “Video OCR for Digital News Archive”, *Proc. Workshop on Content-Based Access of Image and Video Databases*, IEEE, Los Alamitos, CA, 1998, pp. 52-60.
- [5] A. Olligschlaeger, *Spatial Analysis of Crime Using GIS-Based Data: Weighted Spatial Adaptive Filtering and Chaotic Cellular Forecasting with Applications to Street Level Drug Markets*, Ph.D. Thesis, Carnegie Mellon University, Pittsburgh PA, 1997.
- [6] Environmental Systems Research Institute, Inc., home page, <http://www.esri.com/>
- [7] F. Kubala, R. Schwartz, R. Stone, and R. Weischedel, “Named Entity Extraction from Speech”, *Proc. DARPA Workshop on Broadcast News Understanding Systems*, Lansdowne, VA, February 1998.
- [8] K. Beard, T. Smith, and L. Hill, “Meta-information models for georeferenced digital library collections”, *Proc. Second IEEE Metadata Conf.*, IEEE, September 1997.

[9] T.R. Smith, D. Andresen, L. Carver, R. Dolin, C. Fischer, J. Frew, M. Goodchild, O. Ibarra, R.B. Kemp, R. Kothuri, M. Larsgaard, B.S. Manjunath, D. Nebert, J. Simpson, A. Wells,

T. Yang, and Q. Zheng, "A digital library for geographically referenced materials", *IEEE Computer* **29**(5), 1996, pp. 54-60.