Running Head: SUBTITLES DIGITAL LIBRARY

Metadata Cataloging, Storage, and Retrieval of Multilingual Motion Picture Subtitles: An XML Digital Library

Helena Marvin and Kimmy Szeto

Graduate School of Library and Information Studies, Queens College, City University of New York

Abstract

The popularity of motion pictures in digital form has seen a dramatic increase in recent years, and the global entertainment market has driven demands for subtitles in multiple languages. This paper investigates the informational potential of aggregating a corpus of multilingual subtitles for a digital library. Subtitles are extracted from commercial DVD releases and downloaded from the internet. These subtitles and their bibliographic metadata are then incorporated in an XML-based database structure. A digital library prototype is developed to provide full-text search and browse of the subtitle text with single- or parallel-language displays. The resulting product includes a set of tools for subtitles acquisition and a web browser-based digital library prototype that is portable, extensible and interoperable across computing platforms. The functionalities of this prototype are discussed in comparison to another subtitles corpus created for computational linguistics studies. Several informational potentials of this digital library prototype are identified: as an educational tool for language learning, as a finding aid for citations, and as a gateway for additional temporal access points for video retrieval.

Keywords: metadata, subtitles, digital library, cataloging, XML, SRT, motion pictures

Metadata Cataloging, Storage, and Retrieval of Multilingual Motion Picture Subtitles: An XML Digital Library

The popularity of motion pictures in digital form has seen a dramatic increase in recent years due to the widespread availability of production equipment and the pervasiveness of multimedia delivery networking technology. Each year, approximately 9,000 hours of motion pictures are released professionally (Wactlar, 2001). Unlike text based formats, the graphical and temporal contents of video introduce additional dimensions of information, complicating their cataloging, indexing, and methods for search and retrieval. Beyond the video content itself, DVDs contain more information than the previous generation of Betamax and VHS video media. Auxiliary information available on DVDs often includes special features, video and audio commentaries, dubbed audio and subtitles, all of which overlay the video during playback.

The global entertainment market has fueled demands for subtitles in multiple languages. These subtitles are frequentlly professionally produced and included in commercial DVD releases, as well as independently created by users who make them available on the World Wide Web. As a result, web sites containing these user-created subtitles, as well as professional subtitles extracted from DVDs, have emerged and grown into large databases. These repositories of subtitle texts have become fertile ground for research. Tiedemann (2007) has found the following in his Opus Project:

> There is a huge demand for subtitles on the Internet and users provide them to others in various languages via download services on-line. They are available in form of

> plain text files … subtitle databases provide a unique multilingual resource with various kinds of valuable information encoded in the texts (p. 1-2).

Tiedemann (2007) finds subtitles often translated using idiomatic language, and that freely available subtitles are a convenient source for creating a corpus for studies in translation and computational linguistics. Linguists, however, are not the only group to benefit from the empirical evidence of real language data in this corpus (Hockey, 1999). The same corpus of subtitles can reveal other informational potentials for libraries and information science.

Objectives

The objective of this project is to create a multilingual motion picture subtitles digital library prototype, a proof of concept digital library of motion picture subtitles, along with accompanying documentation. This project itself consists of three parts: this document, the software tools needed to create this digital library, and a prototype of the digital library. This document contains a review of the related literature, a description of methodology, comparisons of related projects and technologies, and discussions of the potential future developments of this concept. A workflow chart detailing the procedures to create this digital library is included for libraries interested in developing a similar product. The prototype of this digital library developed and presented here is a proof of concept, consisting of the English, Spanish, and French subtitles of 10 motion pictures.

Specification and Functionality

The software tools included for this project serve to extract subtitle information from DVDs and convert this information into a format for the subtitles database and the digital

library. The digital library includes these subtitles as well as bibliographic information on the films. Users of this digital library are able to browse and search the full text of all the subtitles in any language. The browse and search results for subtitle texts are displayed in a single language or in parallel languages, with titles and time stamps. By utilizing this time stamp, users can also browse subtitles at any point during the film. All external software tools are free and/or open source, and the underlying technologies for the resulting digital library are to be small, portable, extensible, and interoperable across multiple computing platforms.

Intended Audience

This digital library is designed for a broad audience, from language students to film scholars, from the casual web surfer to cinephile historians. The inclusion of title and temporal information set this repository apart from existing film script and subtitle databases. In academic and research areas, the time stamps can be a handy reference for citation; casual users and scholars alike can utilize the time stamp to locate clips of interest within a film. The metadata format developed here also aids in the synchronization of subtitles for parallel language displays, a useful tool for language learners.

Organization of This Paper

This paper includes five sections: this introductory section, literature review, methodology, analysis and conclusion. The literature review surveys the data structure and content standards for cataloging and digital library databases; the methodology section discusses how the particular technologies and data standards are selected and the process of their development; the analysis section presents an overview of the resulting digital library

prototype and a comparison to another online multilingual subtitles corpus. The paper concludes with a discussion of the implications of this project on present and future media library practices, and suggests where additional investigation is needed.

Literature Review

Metadata, broadly speaking, stand for information about information. The next subsection introduces the concept of metadata and the definitions of associated terminology. The next several subsections focus on the development of metadata in the context of the history of cataloging, and include a brief survey of current metadata standards, examples of metadata standards created for managing multimedia formats and video subtitles, and the use of metadata in the area of information retrieval services. The last subsection examines two projects that utilize auxiliary video contents as their primary contents of study.

Review of Metadata

The concept of metadata. The catalog is a collection of metadata of the materials described in the catalog. In this sense, the concepts and techniques of metadata creation well predate the advent of the internet. The term "metadata" first appeared in the 1960's in the context of database and management: "1969 Proc. IFIP Congr. 1968 I. 113/2: There are categories of information about each data set as a unit in a data set of data sets, which must be handled as a special meta data set" (Oxford English Dictionary, 2009). The Merriam-Webster codified the term in its 1983 edition with the definition "Data that provides information about other data" (Merriam-Webster, 2009). The term has been further refined in today's field of information science to denote "structured information that describes, explains, locates, or

otherwise makes it easier to retrieve, use, or manage an information resource" (NISO, 2004). For library cataloging, the Association for Library Collections and Technical Services Committee on Cataloging: Description; Access formed a task force in 1998 to examine the role of metadata. About a year later, the group submitted a working definition: "Metadata are structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities" (Committee on Cataloging, 2000). The Dublin Core Metadata Initiative, which has been widely adopted by libraries, defines metadata as "data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation" (Woodley, 2007).

Components of metadata and metadata schemas. Due to the broad applicability of metadata, numerous metadata standards have been developed for a wide range of applications. Metadata consists of indivisible, atomic pieces of informational elements that form the building blocks of the data set. The term "metadata element" refers to a specific, singular concept: for example, duration. A "metadata element instance," generally simply called an "element," is a specific example of the concept.

A "metadata schema" is a prescribed set of elements that are used for resource description. A "metadata schema instance" is a specific example, such as Dublin Core, but the usage of the term is generally not differentiated from "metadata schema," and sometimes it is called "metadata standard" or "markup language." It is a set of rules that govern how a document is encoded electronically. A "markup" of a text includes the text itself as well as

annotations to the text. Following such a standardized encoding enables a uniform system to process such annotations. The objective is to designate content—to distinguish text in a way that is otherwise syntactically indistinguishable during machine processing. In other words, standardized encoding facilitates the electronic implementation of metadata schemas. These standards are often created for specific aspects of metadata description, including data structure (e.g. Dublin Core), data content (e.g. AACR2), data values (e.g. LCSH) and data exchange (e.g. MARC). Metadata sets often employ a combination of metadata standards for various aspects of the data.

The highest in the metadata hierarchy within the scope of this paper is the "schema language," such as the Extensible Markup Language (XML), the set of rules that govern the creation of metadata schemas. These rules provide a generalized framework to simplify the process of combining multiple metadata standards into a single product, and to ensure interoperability among metadata standards.

Metadata in the Context of Cataloging

ISBD for bibliographic markup. International Standards for Bibliographic Description (ISBD) punctuations is one example of markup language. Published by the International Federation of Library Associations since 1971 to regularize form and content of bibliographic descriptions, ISBD prescribes punctuations delineate the different parts of the bibliographic description. The most common punctuation includes ". -- " (full-stop, space, dash, space) precedes a new area of description; " / " (space, dash, space) precedes the statement of responsibility; " : " (space, colon, space) and " ; " (space, semi-colon, space) precedes the

second and third elements within an area of description, respectively (IFLA, 2007). The standardized grammar provided by ISBD enables users to recognize the various areas of description without the need to understand the language of the material.

AACR for bibliographic data encoding. While ISBD provides the interoperability of bibliographic description across multiple languages, the Anglo-American Cataloging Rules (AACR) provides the standard for encoding the description. Following the "Paris Principles" set forth at the International Conference on Cataloging Principles in Paris in 1961, AACR and its subsequent version (AACR2) and revisions provide structure and control for bibliographic description. They also designated access points that are indexed for searching and browsing.

MARC for data exchange. As library catalogs migrate to electronic form with network access, a markup language, Machine-Readable Cataloging (MARC), was developed to enable computer manipulation of data and online sharing of cataloging records. Instead of using punctuation, MARC uses tags to delineate the various fields of description and subfields of related components. Fields are designated with a three-digit number from 000 to 999, and subfields are marked by the subfield symbol, followed by a one-character identifier, a-z and 0-9. The large number of possible fields makes the MARC markup highly flexible. It allows for a high degree of granularity, because new fields and subfields can be added. In fact, MARC has been revised to accommodate not only print materials, but also multimedia and electronic objects. However, while highly extensible, MARC records are not designed for human readability or interoperability with other online resources.

Dublin Core for the World Wide Web data structure. The Dublin Core Metadata Initiative began in 1995 to provide a general purpose, interoperable metadata standard for a broad range of applications. The Dublin Core Metadata Element Set consists of 15 elements that form the core of the Dublin Core vocabulary. These 15 elements cover metadata for content description of the work (such as title and subject), intellectual property information (such as creator and publisher), as well as instantiation (such as language and format). These elements reflect an interdisciplinary consensus about the basic element sets that can be applied to a wide range of real and virtual works and objects. Dublin Core has since been extended and refined, but this 15-element set remains the core standard, and has been endorsed by international and professional standards (ISO 15836, NISO Z39.85-2001 and IETF RFC 5013) (Woodley, Clement and Winn, 2009).

Other current metadata standards. Other current standards include the Text Encoding Initiative (TEI), a set of rules for marking up electronic texts; the Encoded Archival Description (EAD), developed as a tool for creating finding aids for archival inventories, registers, and indexes; and the Metadata Object Description Standard (MODS), a descriptive standard derived from MARC expressed in XML.

XML: a schema language for metadata creation. XML stands for Extensible Markup Language, a schema language created by the World Wide Web Consortium (W3C). It is a restricted form of the Standard Generalized Markup Language (SGML, ISO 8879) that "provides a mechanism to impose constraints on the storage layout and logical structure" (W3C, 2008). This structure serves as a means to implement metadata standards. Among its

many goals, XML aims to provide a simple, generalized, and human-readable platform on which text-based documents can be encoded, transmitted, and processed over the internet. Numerous metadata schemas are encoded according to XML specification. Examples include MARCXML (an XML implementation of MARC), MPEG-7 (a multimedia metadata standard described later), and Dublin Core.

Metadata in Multimedia Resource Cataloging

MPEG. Metadata schemes for multimedia resources often contain both the metadata and the content. The range of applicable objects is broad, including audio, video, and 3-dimensional object models. These media objects are not necessary interdependent, so they can be used individually, and, therefore, described separately. However, they may also be combined to form multimedia objects. For example, playback of a DVD with video, Dolby Digital surround sound, and subtitles requires the synchronization of the video stream, six separate audio streams, and a text stream. The predominant encoding standards are developed by the Moving Picture Experts Group (MPEG)—MPEG-2 for motion pictures on DVD and digital television, and MPEG-7 for multimedia content description.

The MPEG-2 (ISO/IEC 13818) specification handles the encoding of the multimedia object itself, and is primarily concerned with the storage and transmission of video and audio contents and their synchronization and multiplexing (ISO, 2000). MPEG-7 (ISO/IEC 15938), unlike MPEG-2, is independent of the actual encoding of the video and audio signals. It is intended to complement other MPEG standards that encode the actual multimedia objects through metadata descriptions of the multimedia content. Built on XML, the MPEG-7

standard includes structural description about the content, provides search functionality, allows for indexing, and informs how individual objects are synchronized. The building block of the MPEG-7 standard is the descriptor, which represents a single, low-level feature, such as a color or a sound. Description scheme is the next level of metadata that govern the relationships between descriptors, such as regions and events. These description schemes are defined according to the description definition language, which is built into the MPEG-7 specification. Multimedia descriptions are achieved when a description scheme combines descriptors of various media with temporal descriptors (a temporal interval, or a relative time point from a base point, or incremental time points). This metadata structure allows for free text annotation of description schemes, which allows any event or object to be accompanied by any text string. This text string can be a caption or subtitle to be displayed on the screen, or it can be classification terms used for subject searches. The text can even be tagged with semantic and syntactic information to allow for the possibility of on-the-fly, automatic processing and composition of text (ISO, 2004).

ID3. ID3v1 is a metadata container often used for audio objects, such as the MP3 format. Its initial version is a 128-byte position-dependent metadata tag that stores information about title, artist, album, genre, year, etc. A new standard called ID3v2 was created in 1988, and has no relation to ID3v1 (Nilsson, 2000). The ID3v2 standard provides for multiple free-text containers, called “frames,” for different metadata elements, including title and artist, as well as lyrics and URL data. After several versions, ID3v2 begins to allow tagging of specific locations, called chapters, within an audio file. This capability is intended

for a synchronized display of text and/or images. By extension, rapid, multiple occurrences of these synchronized sound, images, and text can result in a video-like multimedia object with subtitles. Another interesting development of the ID3v2 standard is its embedded image extension, which treats images as metadata for audio files: for example, album covers and artist's portrait. ID3v2, however, uses a non-XML tagging scheme, and requires some kind of a crosswalk (for example, Bosma, 2007) to achieve interoperability.

Subtitle Metadata Formats

Subtitles and captions. Text that appears at the bottom of a video to provide additional or interpretive information comes in two main varieties: subtitles or closed captioning. The two differ in intended purpose as well as in technology. Closed captioning is designed for viewers who have difficulty with hearing or comprehension. Words are displayed verbatim on the screen, along with the identities of the speakers and the sound effects. In NTSC videos (used in the United States and Canada), captions are encoded directly into the same data stream as the graphical content. Closed captioning often appears as black-boxed white text scrolling at the bottom of a screen. Closed captioning is generally not provided in a multiple languages, unlike subtitles, for which many variations of the same script may be available. Additionally, it can be very difficult to find timing information for closed captioned text.

*Principal features of a subtitle format***.** There is a plethora of subtitle formats in use by different media players use for different video data formats. The differences in features are summarized in Table 1.

Projects Utilizing Auxiliary Content as Primary Content

Subtitle information is auxiliary to the primary content. It is provided to expand the audience base, and it is meant to be used in combination with the primary content. However, there are also a number of projects which have taken subtitles, or have generated subtitles themselves, for use separate from the primary media.

The Informedia Digital Video Library (2008) project at the Carnegie Mellon School of Computer Science is exploring techniques to aid users in searching for and retrieving information from video. The overarching goal of the Informedia project is to develop a computer system that can create automatic, computer-generated visualization and summarization of film and video contents, as well as search and retrieve information embedded in the video medium.

One branch of the Informedia project involves building a “text segmenter” that provides auxiliary content to accompany broadcast news programs. This research uses speech recognition technology in combination with the closed caption information to generate a searchable text of the news program, complete with timing information. However, this text is auxiliary to the main focus of the research; the text is generated to serve as input to the segmenter, so that the segmenter can analyze and divide an entire news program into individual stories (Hauptmann, Wang, Hu, and Chang, n.d.).

A similar project utilizing, if not generating subtitles as a primary rather than secondary source of information, is OPUS OpenSubtitle Database by Jorg Tiedemann (2009). Aggregating a subtitles database of over 20,000 files in over 29 languages from www.opensubtitles.org, Tiedemann (2009) creates a publicly available parallel corpus of film

subtitles. The goal of this database is to provide a sizable body of text to test the scene segmentation and line alignment algorithm being developed by Tiedemann's team. The online search interface provides search capability for words and phrases in any of the 29 languages and displays software-aligned text in parallel languages. However, the display lacks bibliographic information about the film from which the subtitles come, as well as the time stamps for the retrieved subtitles.

Methodology

Creating a prototype digital library of subtitles necessitates many choices. These considerations include how to acquire the content for the library, what technologies to utilize for creating the library, and what technologies would be utilized for procuring and storing the resulting content.

Subtitle Acquisition

Subtitles are available from several different sources. For the purposes of the digital library prototype developed here, subtitles are gleaned from two sources: DVDs owned by the authors of this paper, and the online subtitle database www.opensubtitles.org.

Original and Copy Cataloging Conceptual Parallel

There are two distinct types of cataloging: original cataloging and copy cataloging. Copy cataloging can be summarized as "ready-to-use records" (Chan and Hodges, 2007). Acquiring subtitle files from www.opensubtitle.org is akin to gleaning existing bibliographic records, while extracting subtitles from original DVDs is more like original cataloging. Although time ought to be taken to verify that a subtitle is high-quality before acquisition

from an online database, the acquisition still takes considerably less time than the extraction process. The digital library prototype presented here contains a mixture of “original” and “copied” subtitles. Upon acquisition, all subtitles are verified and then assembled into the database.

Subtitles Format Selection

SRT was selected from among many subtitle metadata standards. The goal of this project is to convert subtitles from information primarily for display and playback, into information suitable for search and retrieval, in a format easily incorporated into a library system. The subtitle texts and the time information are the focus of this project. Other information, for example, font, font size, color, and screen location, is extraneous. In order to simplify processing further, formats without any inline tagging are preferred (SRT, n.d.; CoreCodec, n.d.).

SRT is chosen because it meets the requirements of this project (see Table 1), and it is also the most popular format on www.opensubtitle.org, where 66.36% of all subtitles are in this format (see Table 2).

Evaluation of Metadata Standards

As distilled from the definitions of metadata quoted previously, a metadata standard should provide the structure that facilitates description, preservation, discovery, identification, retrieval, and management. These six criteria will be the guide for evaluating the metadata implementation of this project.

For the metadata standard, evaluation criteria will be based on the six “Metadata Principles” from NISO (2007). These six principles are: conformity to community standards, interoperability, use of authority control for collocation, inclusion of intellectual property rights, support of long-term curation and preservation, and ability to provide metadata and its contents in a single object.

MARC vs. XML. The digital library prototype presented here is a standalone digital library, where an XML schema is the standard choice for implementation. However, the technology described here is designed to be integrated with existing library systems, where MARC is widely used. Which implementation should be used for providing full-text search of subtitles and their associated temporal information?

The MARC data structure marks up data into fields and subfields. While the length of data is variable, each field is expected to be relatively short. This characteristic of the MARC data structure enables manipulation and combination of various small subsets of information (such as the controlled access points), and also enables efficient indexing and sorting of data elements, two crucial computational concerns in the early days of computing that are mostly resolved by today’s computing power. XML is also designed to encode documents of any length. The tag and attribute structure give XML its unique characteristics: it lacks a built-in directory for indexing and sorting, but has the ability to impose a hierarchical structure on an otherwise linear document. These characteristics make XML suitable for documents with data elements of great length.

Here is an example of how two lines of subtitles are cataloged in MARC and XML. To provide full-text search capability, the content of the subtitles will need to be stored in the record in its entirety, along with the index number and the time code. A cataloger will most likely follow AACR2 rules using the 505 field, which is designated for formatted content. The way the subtitle text, speaker and time code can be recorded is to use the $t and $g. For example:

```
505 1# $g 1 $t Hello! $g 00:02:06,000 -- $g 2 $t Nice to see you! $g 00:02:08,500
```

With an XML implementation, the three pieces of information reside in a record, in which the index number is stored in the ID attribute type. This index number stores a unique identifier, while the title and the time code are tagged in separate elements. For example:

```
<subtitle IDnum=S0001>
     <caption>Hello!</caption>
     <timeCode>00:02:06,000</timeCode>
</subtitle>
<subtitle IDnum=S0002>
     <caption>Nice to see you!</caption>
               <timeCode>00:02:08,500</timeCode>
</subtitle>
```

In MARC, field 505 subfield $t is not a controlled access point, but is indexed for title keyword search. However, the text in 505 $t is indistinguishable from any other title text in fields 24X, 7XX and 8XX, thus only attaining a low level of precision. Field 505 subfield $g, on the other hand, is not indexed. This makes the presence index numbers moot, and searching by time becomes impossible. As for the time codes, which are also stored in subfield $g, a script can be written to display the time value immediately after the title search string. However, if a user discovers an error and needs to insert a line, the user will need to find the location manually where the new line is to be inserted.

In an XML implementation, subtitles can reside in their own element class, enabling searches within subtitle texts. The index numbers can be referenced through the XML built-in IDref attribute type, should the database later expand to include associated lines with subject classification, with a speaker, or with any other information. Inserting a line is also a matter of adding a new <subtitle> tag, since ID numbers do not need to be consecutive. The time codes can also be indexed and searched as text strings, or as numerical values.

XML has demonstrated its suitability and flexibility in all six criteria, while MARC fails in some aspects of preservation, discovery, and management. Thus, an XML implementation is chosen for this project. In fact, many developers of SGML/XML, like McCallum (1996), Gaynor (1996), and Lam (2001), have pointed out suitability of XML in the online computing environment. There are several additional benefits of using XML. First, it is designed for usability over the internet, and the product here is a digital library accessible on the internet. Second, XML uses plain text Unicode characters, which enhance human

readability as well as language compatibility. Third, an increasing number of file formats and applications are based on XML. Microsoft Office's Office Open XML, OpenOffice's OpenDocument, as well as Apple's iWork have all adopted XML as the basis for the file format; also based on XML are the Resource Description Framework (RDF) and the Web Ontology Language (OWL), both of which are considered to be major players in the future conception, description, and modeling of information.

Database data format. The data format for the digital library prototype allows subtitle and temporal contents to be defined and stored within an XML-based schema. Furthermore, this format is to be compatible with any pre-existing XML-based database, and can be added to that pre-existing database as another module without any structural modification to the existing database. The challenge in defining a new format is not determining how to incorporate features and properties from existing subtitle formats designed for display, but determining the minimal set of features that is needed for the specification. The reasons for restricting the scope of the database format are:

1. Most of the display and formatting capabilities do not need to be included
2. This database should be able to stand alone, and at the same time easily implemented into existing databases. The implementation should not rely on extensive standard-specific processing facility
3. This database format should be expandable to accommodate features without having to redefine the base model

The following section describes the details of the database model, beginning with consideration of the overall design goals, then describing the metadata features of the format.

Design goals and non-goals. The general design goal of this database format is to define a simple, portable format that can also be embedded into one or more XML-based host databases, or can be expanded to incorporate other features. The following features describe the basic design goals:

1. Facilitation of the storage of text and its temporal relationship to the video
2. Support for retrieval based on plain text search phrases
3. Support for retrieval based on timing and language constraints
4. Support for chronological browsing, with optional language constraints
5. Support for simultaneous browsing of texts in multiple languages
6. Display of browse and search results, including video of origin, subtitle text, and temporal information

The following features are not part of the design goals:

1. Information on the styling, formatting, or positioning of text
2. Information on the duration of text to be displayed in the video
3. Any conditional display or removal of text from the display screen
4. Any labeling or tagging of specific lines of text or of the video as a whole (i.e., all texts are treated equally)

However, any of the above functionalities can be added through additional XML metadata and/or styling modules. In fact, for the prototype developed here, Dublin Core description has been added for each video from which the subtitles originate.

The result of these goals and non-goals is the definition of an XML document format that is simple, portable, searchable and, browsible, and at the same time interoperable and extensible with any other XML document formats.

Base model elements and attributes. The elements and attributes are summarized in Table 3. The base model of the database includes text and time. These two data sets fundamentally belong to different types. Text can contain any characters allowable in the XML specification, which is currently 16-bit Unicode or its subsets, and each line of text contains a variable number of characters. In contrast, temporal information has a fixed data length of hours, minutes, and seconds (sometimes milliseconds), and uses numeric characters only. Because of this fundamental difference in type, the two sets of data are treated differently in the base model.

Subtitle texts are contained as data within the <subtitle> element. Each line of text is contained in an individual tag to reflect the line-by-line structure of subtitles. The temporal information, due to its fixed data format, is stored in an attribute associated with the <subtitle> element “subID.” Since the temporal information is unique to each line of subtitle text, it is an attribute of the line. In order to facilitate embedding and expansion of this base model, a unique identifier is needed to link each line of subtitle text to other pieces of information.

*Construction of the subtitle tag identifier***.** The subtitle tag identifier needs to be constructed carefully, and several constrains need to be observed. First, the identifier needs to be unique among all possible subtitles. Second, the XML syntactic constructs limit the first character to letters, ":" (colon) and "_" (underscore) (W3C, 2000). The temporal information is unique to each line of text within each set of subtitles for a video. As a result, the combination of video title, language, and time can provide such a unique identifier. Since the identifier cannot begin with a number, the identifier cannot begin with the time code. The identifier should not begin with a film title either, because it is possible for a film title to begin with a number. Consequently, the language code must appear first in the tag identifier. The three-letter language code (ISO 639-2) is chosen because it is more human readable than two-letter version (ISO 639-1). The language code is then followed by the video title and the time code. The time code is in the form of a 9-digit number representing hours (2 digits), minutes (2 digits), seconds (2 digits), and milliseconds (3 digits). These three segments of the tag identifier need to be separated by a character, and the underscore is chosen because it does not exist in the language code or time code, and it appears extremely rarely in video titles. For example,

<subtitle subID=eng_Wedding Singer_000022469>Look at the happy couple.</subtitle>

Handling of duplicated tag identifiers. For commercial films, it is highly unlikely for more than one set of subtitles be created in the same language. Even if this is the case, it is even less likely for lines of subtitles in the same language to have a time code identical at the millisecond level. The only possible situation where tag identifiers might be duplicated is

from user-created subtitles where different texts replace existing subtitle texts, while all other aspects, including the time codes, remain unchanged. To handle this situation, an optional 2-digit extension, such as “F1” or “F2,” can be added after the tag identifier to denote user-created subtitles based on an existing release.

Subtitle Extraction Tool

When researching the various subtitle extraction tools, two criteria are of the utmost importance: the tool must extract subtitles to the SRT format, and the software must be a freeware. Among the numerous tutorials existing for extracting subtitles, the site WeetHet provides a comprehensive comparison of various free options. Of the various options discussed by Luijten (2003), the two contenders that meet the two criteria are VobSub and SubRip. Both VobSub and SubRip are freely available, open source software; both extract subtitles to SRT format. These two programs are installed, and a test subtitle extraction is performed. SubRip is chosen for its intuitive interface and ease of use.

Selection of Digital Library Technologies

The decision to create a custom digital library back-end for the database comes after evaluating several open source digital library software options. Goh et al. (2006) find Greenstone most consistently fulfills the criteria for what makes “good” open source digital library software. Greenstone has good “user interface customization [that] allows different [digital libraries] to create interfaces that suit the needs of its stakeholders, while automatic tools simplify content management and acquisition ... [and that] There is a wealth of online documentation and tutorial available on the Greenstone Web site.”

After an initial review and testing of the Greenstone software, the software package is ultimately decided to be too powerful and cumbersome for the purposes of project strives for. Instead, a custom implementation is developed using a simple XML container with Dublin Core bibliographic metadata, and the search engine is based on JavaScript, HTML and CSS. Extensible Stylesheet Language (XSL) would have been a better choice, but its latest version 2.0 is incompatible with the current version 3.5 of Firefox, a widely used web browser. For this digital library prototype, English, Spanish, and French subtitles from 10 films are included. A flowchart detailing the subtitles acquisition process, metadata creation, and digital library construction can be found in Figure 1.

Analysis

Functionalities of the Digital Library Prototype

The digital library prototype successfully presents the result of incorporating extracted and acquired DVD subtitles into an interoperable metadata structure. The user may begin a search by entering a search phrase, or may click for a list of all films in all available languages. The user is then presented with the option to browse any line on the search results screen, or browse from the beginning of the film from the list of all films. On the browse page, five lines of subtitles are displayed. Using the time stamps, the user can then browse to a particular time in any one or two language options. If the user chooses two languages, five lines of text in each language will be displayed on the screen concurrently. All lines of subtitles on display are accompanied by their respective film title and time stamp.

Comparison to Opus

While Tiedemann (2009)'s parallel corpus, OPUS, can be searched by word and phrase and display search results in context, it does not provide title or temporal information. The link to these two pieces of information sets the digital library prototype apart from OPUS and from the other online subtitle and script databases. The titles and time stamps are important components of a digital library as they can serve as a citation reference for scholars and reviewers, and enable all users to locate segments within a film.

Another two features available with the digital library prototype but not in the OPUS interface are the options to explore film subtitles by time, and the ability to scroll within the film. These are basic functionalities of a digital library, and the ability to view multiple translations concurrently is invaluable to language learners.

OpenSubtitle XML Corpus as a Starting Point

The main reason for the difference between this digital library prototype and Tiedemann (2009) is that the database content in OPUS is auxiliary to the multilingual text alignment algorithm research, resulting in radically different search and display interface. Despite this difference, both OPUS and this project utilize the same subtitle archive, www.opensubtitle.org, to aid in the creation of database content, and this content is essentially identical. Therefore, the OPUS database can serve as a potential source for a massive production-grade implementation of the present digital library prototype.

Subtitle Databases and Copyright Law

One of the concerns facing both Tiedemann (2009)'s OPUS and this project is the dubious legality of subtitles used without permission from the film copyright holders. Hatcher (2005) captures this issue succinctly:

> International copyright treaties such as the Berne Convention, state that its signatories (such as the United States and Japan) should grant authors the exclusive right to translation. In the United States—the frame of reference for most online discussion of fansub [user-created subtitles] legality—copyright law construes translations as "derivative works." Derivative works are any work "based upon one or more preexisting works." Fansub groups also infringe the right to reproduction by copying the original source material. By distributing the work to others, fansub groups violate the right to distribution. Each fansub therefore represents at least three violations of copyright law. (p. 521-522)

Yet OPUS has been online for more than five years and has faced no legal challenges. As for this project, it falls under educational fair use in its current form. A production grade implementation, however, will require revisiting this legality issue.

Conclusion

The Multilingual Motion Picture Subtitle Digital Library Prototype project successfully implements a number of features. English, Spanish, and French subtitles for 10 motion pictures and associated metadata for each film are arranged into an XML database. The digital library facilitates the retrieval of plain text phrases and supports all three languages. Browse results are constrained by timing and can be constrained by language

selection. Chronological browsing of subtitles, as well as the simultaneous browsing of subtitles in multiple languages, is supported.

There are many further possibilities for this prototype. For instance, it might be expanded to include titles not originally in English. Additional classification information, such as speakers and scene descriptions, could be added and linked to individual subtitles or groups of subtitles. On a large scale, this could become a comprehensive video search engine, akin to a text-based version of the video classification and retrieval studies of the Informedia Project (2008). On interoperability and cross-platform compatibility, the present prototype has only been tested on the latest versions of Firefox and Internet Explorer, and, although these browsers are free and cross-platform, expanding support to other browsers could make the project accessible to more users. Ideally, the search and display engines for the digital library should be rewritten in XSL, which is designed for use with the digital library's underlying XML metadata format.

This project also shows how a digital library of subtitles can become a reference source and a finding aid that provides meaningful textual and temporal access points to events within a film. The prototype provides a glimpse of a way librarians might utilize a freely available auxiliary information source they already own to create new tools and new resources. By cultivating auxiliary content, subtitles hold great potential to serve as a foundation of these new tools and new sources of information.

References

Bosma, A. (2007). Tag frame reference. Retrieved from http://age.hobba.nl/audio/tag_frame_reference.html

Chan, L. M., & Hodges, T. (2007). *Cataloging and classification: An introduction* (3rd ed.). Lanham, MD: Scarecrow Press.

Chiu, C. (2006). Subtitle format comparison. Retrieved from http://www.annodex.net/node/8

Committee on Cataloging: Description and Access. (2000). Task force on metadata: final report. Retrieved October 10, 2009, from http://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html

CoreCodec, Inc. (n.d.). SRT Subtitles. In *Matroska*. Retrieved from http://www.matroska.org/technical/specs/subtitles/srt.html

Dublin Core Metadata Initiative. (2008). Dublin Core Metadata Element Set, Version 1.1. Retrieved from http://dublincore.org/documents/dces/

Gaynor, E. (1996). From MARC to markup: SGML and online library systems. Published at http://www.lib.virginia.edu/speccol/scdc/articles/alcts_brief.html; retrieved from http://xml.coverpages.org/gaynorMARC96.html

Goh, D. H., Chua, A., Khoo D. A., Khoo, E. B., Mak, E. B., and Ng, M. W. (2006). A checklist for evaluating open source digital library software. *Online Information Review, 30*(4), 360 -379.

Hatcher, J. S. (2005). Of Otakus and fansubs: A critical look at anime online in light of current issues in copyright law. *Script-ed, 2*(4), 514-542. doi: 10.2966/scrip.020405.514

Hauptmann, A., Wang, Z., Hu, N., & Chang, J. (n.d.). Text segmentation in the Informedia Project. Pittsburgh, PA: Carnegie Mellon University School of Computer Science. http://www.cs.cmu.edu/~cjc/course/15781-report.htm

Hockey, S. (1999, May 20). Why work with corpora? In *Introduction to the use of computer corpora in linguistics*. Retrieved December 5, 2009, from http://www.humanities.ualberta.ca/Susan_Hockey/Intro_to_Corpora/Why_work_with_corpora.htm

Informedia Project. (2008). Informedia Digital Media Library. Pittsburgh, PA: Carnegie Mellon University School of Computer Science. Retrieved from http://www.informedia.cs.cmu.edu

International Federation of Library Associations and Institutions. (2007.) International Standards for Bibliographic Description (ISBD). Preliminary consolidated ed. Retrieved from http://www.ifla.org/files/cataloguing/isbd/isbd-cons_2007-en.pdf

ISO/IEC JTC1/SC29/WG11/N. (2000, October). Short MPEG-2 description. Geneva, Switzerland: International Organisation for Standardisation. Retrieved from http://www.chiariglione.org/mpeg/standards/mpeg-2/mpeg-2.htm

ISO/IEC JTC1/SC29/WG11/N6828. (2004, October). MPEG-7 overview (version 10). Geneva, Switzerland: International Organisation for Standardisation. Retrieved from http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

Lam, K. T. (2001). Moving from MARC to XML: Introduction. Retrieved from http://ihome.ust.hk/~lblkt/xml/marc2xml_1.html

Luijten, H. (2003). How do I rip subtitles of a DVD-movie? In *WeetHet - Video*. Retrieved September 5, 2009, from http://www.weethet.nl/english/video_ripsubtitles.php

McCallum, S. H. (1996). MARC data in an SGML structure: the future of communication formats. Published at http://www.acctbief.org/avenir/marcsgml.htm; retrieved from http://xml.coverpages.org/McCallumMARC.htm

meta-. (2009). In Oxford English Dictionary (Draft rev. Sept. 2009). Retrieved November 1, 2009.

metadata. (2009). In Merriam-Webster Online Dictionary. Retrieved November 1, 2009, from http://www.merriam-webster.com/dictionary/metadata

National Information Standards Organization. (2004). Understanding Metadata. Retrieved from http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Nilsson, M. (2000). ID3 tag version 2.4.0: Main structure. Retrieved from http://www.id3.org/id3v2.4.0-structure

NISO Framework Working Group. (2007). Metadata. In *A framework of guidance for building digital collections: A NISO recommended practice*. Retrieved from http://www.niso.org/publications/rp/framework3.pdf

OpenSubtitles.org. (2009). Retrieved from http://www.opensubtitles.org

SRT (SubRip) Subtitles. (n.d.). Retrieved from http://www.srt-subtitles.com

Tiedemann, J. (2007, January). Building a multilingual parallel subtitle corpus. Presented at CLIN 17: Computational Linguistics in the Netherlands, Leuven, the Netherlands. Retrieved December 13, 2009 from http://www.let.rug.nl/~tiedeman/paper/clin17.pdf

Tiedemann, J. (2009). OPUS: Corpus query (CWB). In *OPUS: An open source parallel corpus.* Retrieved from http://urd.let.rug.nl/tiedeman/OPUS/bin/opuscqp.pl?corpus=OpenSubtitles

Wactlar, H. (2001, November). The challenges of continuous capture, contemporaneous analysis, and customized summarization of video content. In *Defining a Motion Imagery Research and Development Program Workshop*. Presented at The National Center for

Supercomputing Applications (NCSA), Herndon, Virgina. Retrieved from http://www.cs.cmu.edu/~hdw/Wactlar-MotionImagery2001.pdf

Witten, I. H., & Bainbridge, D. (2003). *How to Build a Digital Library*. Amsterdam: Morgan Kaufmann Publishers.

Woodley, M. S., Clement G., & Winn, P. (2009). DCMI Glossary. Retrieved from http://dublincore.org/documents/usageguide/glossary.shtml

World Wide Web Consortium. (2000). Extensible Markup Language (XML) 1.0 (Second Edition), Section 2.3 Common Syntactic Constructs. Retrieved from http://www.w3.org/TR/2000/REC-xml-20001006

World Wide Web Consortium. (2008). Extensible Markup Language (XML) 1.0. 5th ed. Retrieved from http://www.w3.org/TR/xml/

Author Note

Kimmy Szeto, Graduate School of Library and Information Studies, Queens College, City University of New York; Helena Marvin, Graduate School of Library and Information Studies, Queens College, City University of New York.

Kimmy Szeto is now at the Queens Borough Public Library Long Island Division, Jamaica, New York.

Correspondence concerning this article should be addressed to Kimmy Szeto, Queens Library, Long Island Division, 89-11 Merrick Blvd., Jamaica, NY 11432, E-mail: ks287@columbia.edu and to Helena Marvin, E-mail: lena@lenamarvin.com.

Table 1. *Principle features of subtitle formats.*

		Needed for the database				Not needed for the database				
Subtitle format	**Extension**	**Subtitle Text**	**Multiple Lines**	**Timing precision** (in millisecond)	**Character encoding***	**Timing by frame**	**Screen positions**	**Moving text**	**Conditional timed text**	**Text styling**
For this Database		Yes	Yes	highest possible	UTF-8	No	No	No	No	No
MPlayer2	mpl	Yes	Yes	No	None	Yes	No	No	No	No
SAMI	smi	Yes	Yes	No	Windows-1252	Yes	Yes	No	No	Yes
subRip	srt	Yes	Yes	1 ms	Informally UTF-8	No	No	No	No	No
SubStation Alpha	ssa	Yes	Yes	10 ms	None	No	Yes	Yes	Yes	Yes
MicroDVD	sub	Yes	Yes	No	None	Yes	No	No	No	No
TMplayer	tmp	Yes	Yes	1000 ms	None	No	No	No	No	No
SubViewer	txt	Yes	Yes	10 ms	None	No	Yes	No	No	Yes

Note. Character encoding information is based on the report by Chui, 2006.

Table 2. *Distribution of subtitle formats on OpenSubtitles.org.*

Subtitle formats	Found on OpenSubtitles (Total: 672866)	percentage
srt	446532	66.36%
sub	175915	26.14%
tmp	30507	4.53%
mpl	11756	1.75%
txt	4051	0.60%
ssa	2859	0.42%
smi	1246	0.19%

Table 3. *Elements and attributes defined for the database and additional modules embedded into the prototype database.*

Features	Elements	Attributes
Subtitles	<subtitle>	
Temporal data		subID segment 3
Video of origin		subID segment 2
Language		subID segment 1
Disambiguation		subID segment 4 (optional)
Added module		
Video	<video>	
Bibliographic description	<metadata>	
	The Dublin Core core element set	xmlns pointing to the Dublin Core namespace definition

Figure 1. Workflow chart for metadata creation and digital library construction.