The Sonzogno Digital Library Project

Sanghee Oh
School of Library and Information Studies,
Florida State University
142 Collegiate Loop, FL 32306
shoh@cci.fsu.edu

Wonchan Choi
School of Library and Information Studies,
Florida State University
142 Collegiate Loop, FL 32306
wc10d@my.fsu.edu

Silvia Valisa
Department of Modern Languages and Linguistics – Italian Division,
Florida State University
625 University Way, FL 32306
svalisa@fsu.edu

ABSTRACT
This poster reports on a work-in-progress project of building a digital library for metadata records of books published by Sonzogno between 1800 and 1943. Metadata records from seven major libraries that have Sonzogno collections were harvested and used for building a unified collection for this project. Omeka, open-source software for building digital libraries, was used to host, manage, and provide the public with access to the collection. Currently, a sample set of the metadata of books published in 1885 is available from the digital library (http://sonzogno.cci.fsu.edu/). This poster reports on the current status of the project, explaining the process and challenges associated with building a digital library of the metadata collection (metadata harvesting and crosswalk) using Omeka.

Keywords
Digital library, metadata scheme, metadata harvesting, metadata crosswalk, Dublin Core, MARC, UNIMARC.

INTRODUCTION
The Sonzogno Digital Library Project is a collaborative research project between the School of Library and Information Studies (SLIS) and the Department of Modern Languages and Linguistics (MLL) at Florida State University which intends to build a digital library that will provide universal access to the bibliographic/metadata records of books published by Sonzogno. Founded in 1861, the Società Sonzogno embodies the entrepreneurial spirit of the 19th century European media industry, and a commitment to undermine the distinction between high and low culture in post-Unification Italy.

Despite its’ essential commitment to the Italian publishing history, there has been little research on Sonzogno until recently because Sonzogno’s archives were destroyed by bombs in 1943 during WWII. Due in part to this loss of basic data, access to its publications has been limited and decentralized. Our digital project reconstructs its historical catalogue collection and makes it accessible online. The goal is to provide a new investigation tool with which to re-examine the cultural and editorial life of Italy in the last two centuries, and to foster curiosity and understanding of a firm that was, and is, a protagonist in the cultural life of modern Italy.

The purpose of this project is to provide centralized access to the metadata records of Sonzogno publications, which are scattered across countries. The seven major libraries of the Sonzogno collections in Italy and the U.S., namely, (1) the National Central Library of Florence (BNCF), (2) the National Library Braidense, Milan, (3) Polo SBN Regionale Lombardia (4) the Library of Congress (LC), (5) WorldCat, (6) Harvard Library, and (7) Brown University Library, were identified and their metadata were harvested using a semi-automatic method. Most of the metadata schemas of these libraries are either Machine-Readable Cataloging Record (MARC) or Universal MARC (UNIMARC). However, Dublin Core (DC) was chosen as the metadata schema of the Sonzogno Digital Library in order to provide better access to the collection on the Web. Also, consideration is being given to future developments of the digital library that would provide access to full-texts of the publications and the web documents. Therefore, metadata crosswalk was performed, mapping MARC or UNIMARC to DC. In order to manage and to provide public access to the collection, Omeka, open-source software of digital library building (Omeka, 2013), was used.

This project intends to host metadata records of Sonzogno collections published between 1800 and 1943. At the current stage of the project, a sample set of the metadata records of books published in 1885 is available from the digital library website (http://sonzogno.cci.fsu.edu/). This poster reports on the current status of the project, explaining the process and challenges associated with building a digital library with the metadata collection (metadata harvesting and crosswalk) using Omeka.

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.
LITERATURE REVIEW

Metadata Harvesting and Crosswalk

Metadata is the structured information describing and explaining an information resource, which is useful to facilitate easy retrieval, use, and management of information objects (NISO, 2000). As proposed in this Sonzogno project, after harvesting metadata from the dispersed collections, metadata crosswalk, was carried out. This is a process in which relationships among the different schemas are specified, and a unified schema is developed for the purposed collection. It is one of the important tasks, in terms of building “semantic interoperability,” among collections and making the new digital library meaningful and unique (St. Pierre & Laplant, 1998; Woodley, 2000).

Among several different cases of metadata crosswalk, such as EAD-MODS (Bountouri & Gergatsoulis, 2009), MARC-TEI-DC (Kurth, Ruddy, & Rupp, 2004), DC-VRACore-REACH-EAD (Greenberg, 2001), etc., this project is most interested in mapping metadata schemas between MARC/UNIMARC and DC (Development and MARC Standard Office, 2008; Walsh, 2011) since they are the most popular and common metadata schemas for presenting bibliographic metadata records of books, such as those in Sonzogno publications. Previous studies regarding mapping and crosswalks report that the most frequently observed issues include incorrect mappings, misuse of metadata elements, confusion in descriptive metadata and administrative metadata, and lost information (Jackson, Han, Groetsch, Mustaoff, & Cole, 2008). Due to the varying degrees of depth and complexity, the crosswalks among metadata schemas may not be necessarily equally interchangeable (Chan & Zeng, 2006; Development and MARC Standard Office, 2008). Considering these challenges and issues, this project basically adopted the LC metadata crosswalk framework between MARC and DC and applied variations according to the nature of the Sonzogno collection.

Using Omeka for Digital Library Building

Omeka is a free, open source web-publishing platform, developed as a project of the Roy Rosenzweig Center for History and New Media, George Mason University, to facilitate the display of and access to materials held in libraries, museums, archives, and scholarly collections and exhibitions (Omeka, 2013). As the developers intended, Omeka has been used by various types of institutions such as public libraries, hospital libraries, history archives, and art archives to post their collections to the Web. For instance, the Metropolitan New York Library Council (METRO) used Omeka to build a digital library for their collections. Their project team chose Omeka over other software platforms such as CONTENTdm and Wordpress to create and manage their digital collections on the Web. It was chosen for several reasons, specifically, that it is simple to install and configure, that it supports various plug-ins that allow users to customize their digital libraries, and that it has a flexible approach to metadata representations. Even though the project team took advantage of several of the strengths of Omeka, as mentioned, they also reported on some of the weaknesses, such as the difficulties in creating consistent detailed metadata, administrative interface issues, and search and retrieval-related weaknesses (Kucsmaj, Reiss, & Sidman, 2010). Other examples of systems using Omeka include the Northeast Georgia Health System (NGHS) Health Sciences and Resource Center Library (Bickett, 2013), the Modern Art Iraq Archive (Kansa, Shabout, & Al-Bahlooly, 2010), the Barcero History Archive (Nykolaiszyn, 2011), etc.

METHOD

Metadata Harvesting and Crosswalk

Both automatic and manual approaches were used for harvesting metadata, corresponding to the metadata management policies of each of the seven libraries. First, the targeted metadata records were located by limiting the search with two criteria, the publisher (Sonzogno) and publication years (1800-1943), in each library. Then, several programming scripts using Python and PHP were used to automatically harvest the metadata from Harvard Library, BNF, the National Library Braidense, Milan and Sormani Public Library, Milan. The metadata from LC, WorldCat, and Brown University Library were manually downloaded and used since each of them allows users to export the selected metadata as CVS files from the systems.

The original metadata entries of records from each library were harvested in order to preserve the initial presentation of the records. Most of them were structured using either MARC or UNIMARC, except for the Polo SBN Regionale Lombardia, Milan (This library does not apply a specific metadata schema for its collection). In order to perform the metadata crosswalk between MARC/UNIMARC and DC, all of the metadata schemas and associated elements applied to each of the seven libraries were thoroughly reviewed and then the major DC elements with which to present the collection of the Sonzogno digital library were identified.

Omeka Installation and Management

Omeka 2.0.3 was installed on the FSU CCI Web server in the Linux operating system with Apache, MySQL, PHP, and ImageMagick (http://sonzogno.cci.fsu.edu/, see Figure 1). A simple, default version of the interface is available from the website at this time. An introduction to the Sonzogno Digital Library is provided in the About page; it is written in two languages (English and Italian). Both Simple and Advanced Searches are usable. The browsing option allows users to view and sort the items alphabetically by title, creator, or date published.

By the end of the project, the digital library website will be upgraded by adding several useful plug-ins. Following are examples: the Record Relations plug-in facilitates creating relations across record types; Replace DC Title replaces DC title field in both public and admin views; Search By Metadata allows administrators to configure metadata fields to link to items with same field value (Omeka, 2013).
RESULTS
A total of 19,090 metadata records were collected from the seven libraries, i.e., 5,510 from BNCF, 5,382 from the National Library Braidense, 1,700 from Biblioteca Sormani, Milano, 4,704 from WorldCat, 876 from LC, 801 from the Harvard Library, and 117 from Brown University Library (Duplicates have not yet been cleaned in this stage of this project).

A total of thirty-seven unique MARC or UNIMARC elements were observed in the collected data set. They were matched to thirteen of the fifteen Unqualified DC elements, except Coverage and Rights (Table 1). Among 19,090 metadata records, 113 of them are the metadata of the books published in 1885 and are available from the current website of the digital library. Figure 2 shows an example of the metadata records in the Omeka Interface.

DISCUSSION
One of the challenges we faced at the early stages of the project was metadata harvesting. It had been planned to take advantage of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), a protocol widely used by digital libraries and archives, and institutional repositories for exchanging metadata. We had intended to collect metadata from the seven libraries, the OAI-PMH providers, automatically using a plug-in from Omeka, OAI-PMH Harvester. However, several libraries among the seven, especially those in Italy, do not support OAI-PMH. There is also a problem with OAI-PMH Harvester since it does not allow collecting metadata under the parameters we had decided to impose, i.e., by publishers (Sonzogno) and

### Table 1. Crosswalk of DC, MARC, and UNIMARC

<table>
<thead>
<tr>
<th>DC</th>
<th>MARC</th>
<th>UNIMARC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Uniform Title (240$a); Title Statement (245$s$a$b$c)</td>
<td>Title and Statement of Responsibility (200$a); Other Titles (517$a)</td>
</tr>
<tr>
<td>Subject</td>
<td>Subject Added Entry (600, 610, 611, 630, 650, 653)</td>
<td>General Note (500$a)</td>
</tr>
<tr>
<td>Description</td>
<td>General Note (300$a)</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>Index Term: Genre/Form (655)</td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>Original Version Note (534$s$f); Location (852); Holding Institution (850)</td>
<td>Original Source (801$a$b$c); Holding, Localization, Inventory No. (960$a$b$c$d$e)</td>
</tr>
<tr>
<td>Relation</td>
<td>Series Statement (490$a$v); Main Series Entry (760$a); Edition Statement (250$a); Other Edition Entry (755$b)</td>
<td>Title and Statement of Responsibility (200$v); Series (225$a)</td>
</tr>
<tr>
<td>Creator</td>
<td>Main – Personal Name (100$a$d)</td>
<td>Personal Name (700$a$b$c)</td>
</tr>
<tr>
<td>Publisher</td>
<td>Publication, Distribution, etc. (260$s$a$b)</td>
<td>Publication, Distribution, etc. (210$a$c); Place Access (620$d$)</td>
</tr>
<tr>
<td>Contributor</td>
<td>Added Entry- Personal Name (700$a)</td>
<td>Personal Name- Secondary Intellectual Responsibility (702$a$b)</td>
</tr>
<tr>
<td>Identifier</td>
<td>LCCN (010$a); Other Standard Identifier (024)</td>
<td>Record Identifier (001)</td>
</tr>
<tr>
<td>Language</td>
<td>Language Code (041); Language Note (546)</td>
<td>Language of the Item (101$a)</td>
</tr>
</tbody>
</table>

Figure 1. Homepage of the Sonzogno Digital Library

Figure 2. Example of a metadata record in the Sonzogno Digital Library
Another challenge in this project was the degree of complexity involved in the mapping between the different metadata schemas obtained from the seven libraries. The depth of presentation about editorial, series, and collection-level statements in MARC/UNIMARC had to be simplified into one element, as it is defined in DC. For example, the distinction between Creator and Contributor in DC should be defined before applying the several elements associated with personal names in MARC/UNIMARC. Even among the metadata records using the same MARC/UNIMARC, different levels or depths of the content of metadata elements were often observed, which seemed to hinder the “semantic interoperability” of the metadata. Other issues found in the project included the variations and inconsistencies in the format of the metadata records, such as notations and punctuation (e.g., variations of publication years, such as -1932, [1932], p193).

The installation and management of Omeka were accomplished without major problems. It was easy to use for both users and administrators. The options for designing the interfaces (named Themes) and launching searching and browsing functions were flexible; thus enhance the use of the Sonzogno Digital Library. Social features and associated plug-ins, such as social bookmarking, and online submission, allow building and developing a user community around the digital library. Technically, however, Omeka needs to be further developed to fully support the functions and features needed for digital libraries. For example, OAI-PMH Harvester would need to allow for refining the searches and for harvesting metadata under specific conditions.

CONCLUSION
Despite the challenges and issues, the collaborative efforts, discussion, and negotiation among the project team members in the two fields of Library and Information Studies and Modern Languages and Linguistics was fruitful as they worked on the digital library and found the most appropriate metadata schema for the Sonzogno publications. The Sonzogno Digital Library is the first digital library, which provides centralized access to the Sonzogno publications. The data available from the digital library can be used for research by the groups of people who are interested in studying the publishing and commercial development of Sonzogno in Italian publishing history.

In the future, we plan to collect and digitize the original books or other materials of the Sonzogno collections and make them available through the digital library along with the bibliographic records refined in this project. For example, a follow-up project proposal focusing on providing access to full-texts of the newspapers published by Sonzogno in collaboration with the Florida State University Library is under development. The project team will continue to build a comprehensive collection of the Sonzogno publications and to develop interfaces, which are the most appropriate in order to make the collection usable and accessible through the Sonzogno Digital Library. An evaluation of the use of Sonzogno Digital Library in the user community will also be conducted.

ACKNOWLEDGMENTS
This project is funded by a 2012 CRC Planning Grant from Florida State University.

REFERENCES