Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 147 (2014) 223 – 228

Procedia
Social and Behavioral Sciences

ICININFO

Data standardization in Digital Libraries: An ETD Case in Turkey

Özlem Şenyurt Topçu*, Tolga Çakmak, Güleda Doğan

Hacettepe University, Faculty of Letters, Department of Information Management, Ankara, 06800, Turkey

Abstract

Nowadays, data integrity and data standardization are significant topics for information retrieval systems and also for digital libraries. Although, many standards (such as VIAF, AACR2 and MARC) and institutional regulations developed for data standardization by the nature of library and information science field, consistency between different content resources is still a problem for today's information systems. It is also one of the most important steps for digital library projects especially for the ones who are planning to gather data from different content resources. As one of these projects, a digital library for electronic thesis and dissertations (ETD) for LIS in Turkey has been developed via a digital library project executed by PhD students of Hacettepe University Department of Information Management. In this direction, this paper aims to explain data standardization processes and limitations due to regulations, language and cultural characteristics of the country with examples from the indicated project. As a result of the project a standardized structure for ETD systems were created. In this context, author and advisor names, Turkish and English titles, keywords, access restrictions were determined as the main elements for standardization processes. In the end of the project, an authority file was created for advisors via VIAF, RDA and AACR2 in order to improve efficiency of access points by ensuring data integrity.

Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.
Keywords:Data integrity; data standardization; ETDs; digital libraries; Turkey

1. Introduction

Of late, the initiatives about digital libraries and digital archives have increased especially in universities and academic institutions. Furthermore, it is also seen that many digital archives have been developed with the aims of preserving cultural heritage and providing access to cultural heritage assets via recent technologies. Bibliographic

* Corresponding author. Tel.: +90-312-297-82-00; fax: +90-312-299-20-14.
E-mail address: ozlemsenyurt@hacettepe.edu.tr

Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.
doi:10.1016/j.sbspro.2014.07.157

description of information resources and cultural heritage assets covered by digital libraries and archives is an essential process in order to meet information needs of users. It also provides efficiency for information retrieval and makes digital library and archives usable for users by supporting decision-making processes. On the other hand, standardized presentation of data identified in digital library and archives is as important as description of metadata fields. It is an important requirement for the initiatives especially whose aim is to create a single platform for different content providers. Based on this information, this study addresses the data standardization phases in the case of a local digital archive project whose main objective is to develop an ETD platform specialized in Library and Information Science in Turkey by gathering data from different content providers.

2. Data Standardization, Consistency and Cleaning

Data standardization provides consistency between the content and format of data types represented in a system. Furthermore, data standardization facilitates efficient consistency for data mapping and data outputs (IBM, 2013). According to the literature data standardization, consistency and cleaning topics are covered by different fields like computer science, statistics and information science. In this context, different terms like normalization, data cleaning, noisy data are used to describe data standardization studies. It is indicated that data standardization processes improve information retrieval, prevent loss of data and provide unduplicated data (Normalleştirme, 2013; Tyagi and Samuel, 2013).

Data standardization is a determinative fact about data quality and high quality data should have some criteria. These criteria are accuracy, integrity, completeness, validity, consistency, uniformity and density data quality (Tyagi and Samuel, 2013). Data standardization provides internal consistency of data represented in digital libraries and archives. Through this process, each data type described in the digital library and archive have the same content representation properties (IBM InfoSphere Information Server, 2013). Standardization processes of data are referred as one of the essential processes in terms of user interaction and meeting information needs. Furthermore data standardization processes increase the clarity level of data to support users' interaction with digital library and archive systems in terms of accessing to required information (Tyagi and Samuel, 2013).

The necessity of synchronized structures of data presented in digital library and archives is discussed in studies. Accordingly many studies are conducted to provide data consistency during the implementation phases of digital library and archives. In this context data consistency is defined as a concept that summarizes validity, usability, integrity of applications and data in digital library or archive systems. Digital library or archive systems that have consistent data present a high quality and consistent platform for their users (Data Consistency, 2013).

Conceptually, data standardization is relevant with many components. One of these components is data cleaning processes. Data cleaning can be described as a set of operations carried out for detecting and removing errors and inconsistencies from data (Rahm and Do, 2000). Main aim of data cleaning is to weed out unsuitable or incorrectly entered data in the data set (Tekerek, 2011). Data cleaning is also stated as a process that increases data quality and solves data quality problems. According to Ramd and Do (2000, p.3), data quality problems consist of single-source problems and multi-source problems.

In their study, Ramd and Do (2000, p.3) divide data quality problems into two levels. These levels are metadata schema level and instance level. On the other hand, misspellings, redundancy of data, contradictory values are the essential problems that require data standardization and data cleaning operations. It is also expressed that data cleaning processes include completion of missing data, ensuring data consistency via identification of outliers (Oğuzlar, 2003).

3. Data Standardization and Metadata

The need for standardization of data practices has step up with the growing of digitization technology in library related operations. The creation of digital collections using digital library technologies, either in the form of 'born-digital' or migrated into digital form, is now an important part of the activities for most of the higher education institutions. Using of these digital collections effectively is dependent on the metadata quality. Furthermore, management of digital resources requires standardized and high quality data (Gartner, 2008, p. 4).

After describing metadata fields for a digital archive/library, before the data entrance processes, data need to be standardized. As a core element of information retrieval systems metadata describes data quality and determines maintenance and preservation of a digital library. Data standardization is crucial for ensuring consistency with metadata for such reasons. Standardized data utilizes efficient discovery, access, transfer and use of common terms, common definitions, etc. (Gartner, 2008, p. 5; Why, 2013). Accordingly, standardized metadata enables users to access data effectively and efficiently by using a set of terminology (GeoNetwork, 2008, p. 32). Standardization of data and so metadata provide users finding data they need effectively and efficiently (Xie and Shibasaki, 2013).

With the advancements in technology, there are many digital library and archive systems implemented by the academic institutions. It is seen that many digital library and archives have the same metadata fields however they have different content data. In this direction, ISO/IEC 11179 (Specification and standardization of data elements) standard is stated as the most important outstanding standard for data representation in information systems by the aim of providing understandable and shareable representation of data stored in information systems. Moreover there are many rules and regulations developed by the libraries and library related communities and associations. These rules can be listed as: Cutter's rules, ALA Rules in 1908 and 1941, AACR (1967), AACR2 (1978), ISBD, AACR2R and RDA. Beside these developments, with the web archiving approaches, many metadata element sets were developed for data representation. In this context it is stated that Dublin Core based systems provide many advantages for data representation and interoperability of the systems (Caplan, 2003, p. 40, 55, 85). Additionally, there are many national and international data standardization projects especially for authority files for library automation systems. In this context, an international authority file project titled Virtual Authority File (VIAF) was created by OCLC (VIAF, 2013). As another project titled ORCID, it is aimed to provide a persistent digital identifier that distinguishes researchers from other researchers with a unique id (ORCID, 2013). This project can also be regarded as a data standardization effort for the identification of researchers and authors.

4. The Case of Library and Information Science Electronic Thesis and Dissertations Archive in Turkey

In this part of the study, data standardization processes of a digital archive project titled Turkey Library and Information Science Departments Thesis and Dissertations Archive are described. There are 12 Library and Information Science (LIS) departments in Turkey. Only four of them have master and PhD education. In this respect, thesis and dissertations that are used for the project were gathered from four universities. These universities are Ankara University, Istanbul University, Hacettepe University and Marmara University (BBY Haber, 2013; Düzyol, 2011, pp. 4-5).

The main aim of the project is to provide a platform that contains all thesis and dissertations completed in LIS departments in Turkey. In parallel with this aim, objectives of the project are:

- Creating a union catalog for ETDs completed in LIS departments in Turkey.
- Developing a digital archive that presents full texts and bibliographic descriptions of all ETDs in LIS departments in a single platform
- Identification of ETDs via interoperable and standardized structures.
- Increasing access and visibility of ETDs via a digital library platform that supports OAI-PMH standards and protocols, and provides an interoperable environment for search engines and crawlers of similar digital archives.

4.1. Content & Data Structure

Electronic collection presented in the digital archive consists of 436 post-graduate (masters and doctorate) theses that are completed in LIS Departments in Turkey by the end of 2012. Table 1 shows the contribution of the universities that are the content providers in the project.

Table 1. Content distribution of ETD platform.

	MA		PhD		Total	
Universities	N	%	n	%	n	%
Ankara University	62	18	29	30	91	21
Hacettepe University	126	37	36	37	162	37
İstanbul University	91	27	21	22	112	26
Marmara University	60	18	11	11	71	16
Total	339	100	97	100	436	100

As it is seen in Table 1, more than one-third (37%) of the collection presented in the digital archive is provided by Hacettepe University. It is followed by İstanbul University (26%) and Ankara University (21%). Moreover, It is also remarkable contribution that 30% of the PhD theses provided by Ankara University. Plus, two-third of all PhD theses is provided by Hacettepe and Ankara universities. As a consequence, data standardization processes were applied to 436 ETDs provided from four different content providers.

4.2. Data Standardization processes and data standardization work-flow

Data standardization is required to integrated presentation of ETD's that are situated in different sources with different structures/systems. Data presented under the metadata fields are standardized through various stages. Figure 1 displays the main data standardization workflow for digital library and archive initiatives. In this project, standardization workflow steps presented in Figure 1 were followed.

There are four Library and Information Science departments in Turkey that provide Master and PhD education. Therefore, these four departments are content providers for the data presented in the digital archive and the archive contains all theses and dissertations completed in these four departments by the end of 2012. Theses and dissertations that form digital archive are in PDF format. Finereader OCR (Optical Character Recognition) program were used just for the theses that cannot be copied because of the image based PDF files. In this framework, firstly data set was determined and created via various supplementary resources. These resources are a master thesis completed in 2011 (Düzyol, 2011), National Theses Center of The Council of Higher Education, institutional repositories and library catalogs and databases.

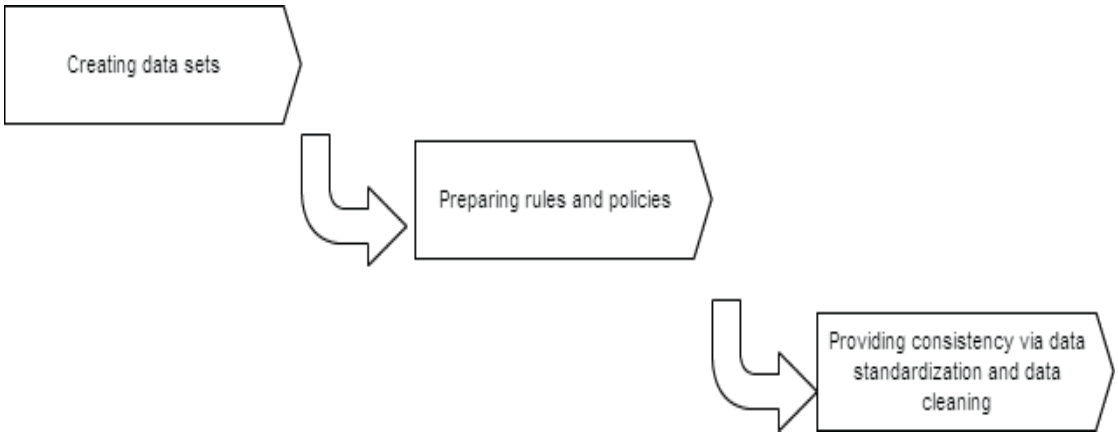

Fig. 1. Standardization workflow

In the second stage of the project, rules and norms that will be used for the standardization processes were determined. Policies were also created according to metadata fields and information resource characteristics of thesis and dissertations. Some metadata fields were also qualified with effect of language and other cultural characteristics of Turkey. AACR2 rules, VIAF structures and "ETD-MS v1.1: an Interoperability Metadata Standard for Electronic Theses and Dissertations" published by Networked Digital Library of Thesis and Dissertation (NDLTD) are the main indicators for data standardization and presentation processes in the project. Best practices and related national and international projects were also reviewed in this stage. In the light of these resources and indicators, qualified metadata fields can be listed as ID, Name, Surname, Type, Year, Proper Title, Alternative Title, Advisor, Link for full-text, Summary in Turkish, Summary, Access Restriction of National Theses Center of The Council of Higher Education, Keywords in Turkish, Keywords. In the end of the workflow, data integrity were provided, and false and flawed data were corrected. Data standardization processes that were carried out according to metadata fields are listed and summarized below:

ID: Every thesis and dissertations has an ID number that makes them unique in the digital archive. ID numbers were used to match bibliographic records of resources with their PDF based objects stored by the digital archive. Besides the ID numbers, handle server system was also created via handle.net and every record defined by a prefix nominated to Hacettepe University who is the official owner of the project. In this context, Registry of Open Access Repositories (ROAR) system registry was also completed.

Authorities: Virtual Authority File (VIAF) were used for author names as a directive resource. Besides, records were displayed in the form of last name, name within the framework of author entries specified by AACR2. Longest versions of the names were used for author and advisor names (i.e. Yaşar Ahmet Tonta instead of Yaşar Tonta or Yaşar A. Tonta). Current names preferred for the women authors and advisors that are married/divorced (i.e. Güleda Doğan instead of Güleda Düzyol). Only first letters of the titles are capital as it is in AACR2 except proper names in the titles.

Title: Titles and alternative titles of the thesis and dissertations were written in the form of only first letter capital unless there is a proper name as described in AACR2.

Date: Publication year information was identified in the form of month-day-year. The same format was applied for the thesis that has embargo date.

Keywords and subject headings: Capital letters were only used for the first letters of keywords unless there is a proper name as in titles.

Summaries: OCR and typo sourced errors in summaries were reviewed and corrected according data identification forms provided by the digital archive.

Usage restrictions: Usage restriction information of theses described by the Council of Higher Education was defined as authorized, unauthorized, restricted access options. Restricted access expression was used for the theses and dissertations that are not allowed to access for 1-2 or 3 years.

5. Conclusion

Digital libraries and archives are important structures for access to information resources and maintenance of the cultural heritage assets. These systems provide many opportunities for their owners and users. These opportunities are compliance with standards, interoperability with other systems, mostly supporting open access and scholarly communication. As essential components of these systems, Metadata fields and data represented in these fields provide effective information retrieval and support critical thinking processes of users. They are more important for the platforms that contain data from different resources and repositories. In this context, policies, rules/norms and studies based on providing data consistency have a vital role for digital library and archives in order to improve the effectiveness of information retrieval.

Although, there are five active departments, Turkey has a great potential with its new LIS departments for new thesis and dissertations. Additionally it would not be wrong to say that ETD initiatives in LIS field will potentially have a larger number of resources in the country. As the first union ETD initiative in LIS science, Turkey Library and Information Science Thesis and Dissertations Archive with its standardized and high quality data structures is a model for such attempts. In this regard this study reflects the importance of the efforts about data cleaning, data integrity and validity as well as the importance of compliance with standards, policies, rules and norms.

References

BBY Haber (2013). *BBY Bölümler.* Retrieved March 29, 2013, Available at http://www.bbyhaber.com/bby/bby-bolumler/

Caplan, P. (2003). *Metadata fundamentals for all librarians*. Chicago: American Library Association

Data Consistency. (2013). Retrieved July 16, 2013, Available at http://en.wikipedia.org/wiki/Data_consistency

Düzyol, G. (2011). *Türkiye Kütüphanecilik ve Bilgibilim literatürünün entellektüel haritasının çıkarılması: bir yazar ortak atıf analizi çalışması.* Unpublished Master Thesis. Hacettepe Üniversitesi, Ankara.

Gartner, R. (2008). *Metadata for digital libraries: state of the art and future directions*. JISC Technology & Standarts Watch. Retrieved June 17, 2013, Available at http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf

GeoNetwork Opensource (2008). *The complete manual*. Retrieved June 23, 2013, from http://apps.who.int/geonetwork/docs/Manual.pdf

IBM (2013). *Making data consistent through standardization*. Retrieved July 24, 2013, Available at http://pic.dhe.ibm.com/infocenter/iisinfsv/v8r5/index.jsp?topic=%2Fcom.ibm.swg.im.iis.qs.ug.doc%2Ftopics%2Fc_Conforming_output_data.html

IBM InfoSphere Information Server (2013). *Standardizing data.* Retrieved July 20, 2013, Available at http://pic.dhe.ibm.com/infocenter/iisinfsv/v8r5/index.jsp?topic=%2Fcom.ibm.swg.im.iis.qs.ug.doc%2Ftopics%2Fc_Conforming_output_data.html

Normalleştirme. (2013). Retrieved July 10, 2013, Available at http://yunus.hacettepe.edu.tr/~uras02/Hacettepe/3.sinif/Bilgisayar/access/MIS_Dersnotu.pdf

Oğuzlar, A. (2003). Veri ön işleme. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 21*(Temmuz-Aralık), 67-76.

ORCID (2013). *ORCID.* Retrieved June 29, 2013, from http://orcid.org/

Rahm, E. and Do. H. H. (2000). *Data Cleaning: problems and current approaches.* Retrieved June 10, 2013, Available at http://dc-pubs.dbs.uni-leipzig.de/files/Rahm2000DataCleaningProblemsand.pdf

Tekerek, A. (2011). Veri madenciliği süreçleri ve açık kaynak kodlu veri madenciliği araçları. paper presented at *Akademik Bilişim 2011 Konferans*ı. Malatya: İnönü University.

Tyagi, B.K. and Samuel, P.P. (2013). *Data consistency, completeness and cleaning*. Retrieved June 25, 2013, Available at http://www.inclentrust.org/resources/PPS%202-DATA_CONSISTENCY_Integrated.pdf

VIAF (2013). *Virtual Authority File*. Retrieved June 29, 2013, Available at http://www.oclc.org/viaf.en.html

Xie, R. and Shibasaki, R. (2013).Standardization framework for CEOP metadata development and application. *CEOP/IGWCO Joint Meeting*. University of Tokyo, Japan. Retrieved June 20, 2013, Available at http://jaxa.ceos.org/wtf_ceop/documents/CEOP_Metadata_Report_20th.pdf

Why standardize metadata? (2013). Retrieved June 21, 2013, Available at http://gep.frec.vt.edu/pdfFiles/Metadata_PDF's/3.0MD_Presentation-Section3.pdf