Extended Date/Time Format (EDTF) in the Digital Public Library of America's Metadata: Exploratory Analysis

Oksana L. Zavalina
College of Information,
University of North Texas,
1155 Union Circle #311068, Denton, TX 76203
Oksana.Zavalina@unt.edu

Mark E. Phillips
University of North Texas Libraries,
Digital Libraries Division,
1155 Union Circle #305190, Denton, TX 76203
Mark.Phillips@unt.edu

Daniel Gelaw Alemneh
University of North Texas Libraries,
Digital Libraries Division,
1155 Union Circle #305190, Denton, TX 76203
Daniel.Alemneh@unt.edu

Hannah Tarver
University of North Texas Libraries,
Digital Libraries Division,
1155 Union Circle #305190, Denton, TX 76203
Hannah.Tarver@unt.edu

Priya Kizhakkethil
College of Information,
University of North Texas, 1155 Union Circle
#311068, Denton, TX 76203
PriyaKizhakkethil@my.unt.edu

ABSTRACT
Considering the value of dates in the life cycle of the digital resource, capturing and storing dates metadata in a structured way can have a significant impact on information retrieval. There are a number of format conventions in common use for encoding the date and time values; the Extended Date/Time Format (EDTF) is one of the most expressive. This paper presents results of an exploratory analysis of representation of dates in over 8 million metadata records from one of the largest digital aggregators, Digital Public Library of America (DPLA), and compares it to EDTF specifications. This benchmark study provides empirical data – at both the individual provider level and the group level (content hubs or service hubs) – about the overall level and patterns of application of date metadata in DPLA metadata records in relation to EDTF.

Keywords
Metadata aggregations, dates, EDTF, metadata evaluation

INTRODUCTION AND BACKGROUND
Arising out of a vision of a United States national digital library, Digital Public Library of America (DPLA) which was launched in April 2013, provides a single platform and portal for open access digitized cultural heritage if the Unites States. DPLA not only hosts and preserves digital information but also provides Application Profile Interfaces (APIs) and maximally-open data to software developers, researchers, and others for building discovery tools. Functioning on a distributed network model, DPLA consists of a group of national partners providing content and services and categorized as content hubs and service hubs (Ma, 2014). Content hubs constitute large libraries, museums, archives and other digital repositories (e.g., ARTstor, California Digital Library, The U.S. Government Printing Office, etc.) which provide metadata records which represent digital objects from their own collections and also maintain and enhance these records as needed. Service hubs are state, regional, or other collaborations (e.g., Connecticut Digital Archive, The Portal to Texas History, etc.) which bring together digital objects from multiple cultural heritage institutions and provide metadata records to the DPLA through a single data feed such as OAI-PMH. The DPLA data model is based on the Europeana data model and the Resource Description Framework (RDF) and employs JavaScript Object Notation-based serialization for Linked Data (JSON-LD), which is disseminated via API output. The descriptive metadata standard employed by the DPLA is the Dublin Core (Mitchell, 2013). Since the primary goal of DPLA is the compilation of harvested metadata to augment the discovery of the digital resources, metadata gathered from providers is stored along with metadata generated or extracted during the data collection process.

ASIST 2015,November 6-10, 2015, St. Louis, MO, USA.

The metadata aggregated and normalized by DPLA is in the public domain and can be harvested via the OAI-ORE standard.

The DPLA, which currently holds over 8 million of unique metadata records from thousands of cultural heritage institutions, is a unique source of rich data for research into metadata quality. One of the areas of metadata quality which has important implications for information retrieval is representation of dates. As a number of dates are associated with events in the life cycle of the digital resource, capturing and storing date information in a structured (and machine accessible) way facilitate access and retrieval of resources. In the DPLA metadata application profile, Date property is a "recommended" metadata element – an element that should be included in metadata record if the information is available (DPLA, 2015a). In the metadata used internally by institutions that serve as DPLA hubs (e.g., Qualified Dublin Core and local metadata application profiles based on it, MODS, VRA Core) it might be common for metadata records to include more than one instance of a Date element to represent dates of creation, issuance, modification, copyright, etc. However, in the process of normalizing native metadata and harvesting it from content hubs and service hubs into DPLA, only one instance of Date element is used – the Date Created of the Qualified Dublin Core metadata scheme (DPLA, 2015a).

Although there are a number of date and time format conventions in common use, the widely adopted best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and follows the YYYY-MM-DD format. A number of aggregators such as, for example, California Digital Library (CDL) develop and apply normalization protocols to date elements in harvested metadata to fit the ISO 8601 standard (Tennant, 2004; Loy & Landis, 2005). Extended Date/Time format (EDTF) (http://www.loc.gov/standards/datetime/pre-submission.html) is a specification by the Library of Congress which defines the features to be supported in a date/time string and which are considered useful for a wide range of applications. EDTF that can be looked upon as a profile of or an extension to ISO 8601 international standard aims to increase consistency in features of ISO 8601 through providing a standard syntax for the representation of date and time. EDTF also includes features which are not present in ISO 8601. EDTF structure comprises of three levels (0-2); level 0 supports basic features while levels 1 and 2 provide more options for flexibility and complexity while also complying with preceding levels ("EDTF introduction", loc.gov, 2012). Tarver and colleagues (Tarver & Phillips, 2013; Tarver, Waugh, Alemneh, & Phillips, 2015), analyze importance of dates metadata in describing serial publications and discuss the advantages that implementation of EDTF at University of North Texas (UNT) Libraries provides through the availability of formats that meets all the date requirements for items in the organization's digital collections and whose adoption could prove to be advantageous for other cultural heritage institutions through the provision of flexible date/time formats not provided by other standards. They also note that the standardization through the usage of EDTF for complex dates can improve metadata interoperability and result in complete and consistent information provision to users.

In the DPLA metadata application profile documentation, EDTF is listed as a suggested syntax schema for a context class, which contains further information about time spans referred to by a resource, for two properties: beginning and ending dates of a time span (DPLA, 2015b, p.16).

A number of research projects have evaluated the quality of metadata in digital libraries and aggregations of different size and subject scope which use a variety of metadata schemes: Dublin Core, MODS, GILS, NSDL, etc. Several of these research projects, among other metadata analyses, studied frequency of application of date-related metadata element (e.g., Bui & Park, 2006; Jackson et al., 2008; Kurtz, 2010; Park & Maszaros, 2009; Weagley, Gelches, & Park, 2010; Zeng, Subrahmanyam, & Shreve, 2005). The findings of these studies show the level of application of Date element to range between 86% and 100% for different repositories.

Application of date-related metadata elements has been considered as one of the indicators of metadata quality. For example, Jackson and colleagues (2008) found that metadata creators often erroneously map publication dates to Dublin Core's Publisher metadata element or Coverage metadata element which is intended to hold the dates and time periods that the information object is about not the dates related to the lifecycle of information object. Similarly, Park and Maszaros (2009) observed that publication dates were erroneously mapped to the *placeName* subelement of the *originInfo* top-level element of Metadata Object Description Scheme (MODS) metadata scheme -- instead of the *date* subelement. These types of incorrect mappings result in problems with metadata accuracy which has been identified by various researchers (Bruce & Hillman, 2004; Moen, Stewart, & McClure, 1998; Park & Tosaka, 2010; Zavalina, 2011, etc.) as one of the most important metadata quality criteria. Another metadata quality problem – directly related to date encoding standards such as ISO 8601 and EDTF – observed with date metadata (e.g., Barton et al., 2003; Dushay and Hillman, 2003; Shreeves et al., 2003, 2005) is the lack of consistency in the format of its data value – particularly when, in addition to the year, the month and/or the day were involved, as in collections of digitized newspapers, etc. Often, within the same digital library, some metadata records were found to use words for month names while other records used numbers, the order of month and day – expressed in two-digit numbers each – was found to vary, sometimes the year was expressed in two digits instead of four, etc.

Our review of the literature demonstrates the lack of empirical studies into metadata quality in digital libraries with the specific focus on date and time metadata. The exploratory study reported in this paper is one of the first attempts to systematically evaluate metadata related to date and time, and the first one to use a very large aggregator such as Digital Public Library of America as its target.

METHODS

The research question that guided this exploratory study is: How are the dates of creation of information objects represented in metadata records across content hubs and service hubs of the Digital Public Library of America (DPLA)? To address the research question, we applied the exploratory quantitative content analysis research method, which relied on basic descriptive statistics such as counts, percentages, etc. Unlike many previous studies of metadata in large-scale digital libraries that analyzed a sample of metadata records, the authors of this study took a "big data" approach that analyzes the whole dataset and therefore avoids sampling errors and produces statistically-valid results. We used DPLA's Bulk Download (http://dp.la/info/developers/download/) to harvest the entire DPLA metadata dataset of over 8 million records in JSON-LD RDF-based serialization which was then parsed into individual item records that contained both the original metadata submitted by DPLA hubs and the normalized metadata conforming to the DPLA Metadata Application Profile. Each record was processed to extract the date values which were then indexed using the Solr full-text indexer (http://wiki.apache.org/solr) and processed with the Python ExtendedDateTimeFormat module (https://github.com/unt-libraries/ExtendedDateTimeFormat). Our workflow made use of the statsComponent and built-in faceting functionality in Solr indexer to work with the 8 million records by aggregating, grouping and performing statistical methods during this exploration.

FINDINGS

Our analysis revealed that 83% of the 8,012,390 DPLA metadata records included date information. The percentage of metadata records that contained date information in the DPLA dataset varied widely across the DPLA hubs, from as high as 100% of records to as low as 21% of records. Comparison of the overall level of inclusion of date information between the two DPLA hub types demonstrates no major difference between the content hubs and service hubs. Content hubs had 83.4% of records with and 16.6% of records without date information while service hubs had 80.9% with and 19.1% of records without date information.

Our findings also suggest that 51% of the date values were valid according to the Extended Date/Time Format (EDTF) specification. We observed a wide variability across DPLA hubs in proportion of dates that were valid EDTF date strings, from as high as 100% to as low as 1.8%. As a group, service hubs exhibited substantially higher consistency in the percentage of EDTF-valid date strings than content hubs. All but one service hub had EDTF-valid date strings in most of populated Date fields in their records. To the contrary, over a third of content hubs had non-EDTF-valid date strings in most of their records that contained Date information.

The vast majority of EDTF-valid date strings in the DPLA metadata records (99.1%) comply with EDTF Level 0 specification which includes standard dates such as years (yyyy, as in 1900); years and months (yyyy-mm as in 1900-03); years, months, and days (yyyy-mm-dd as in 1900-03-03); full dates with time (yyyy-mm-ddT followed by time in the format of hh:mm:ss, as in 2014-03-03T13:23:50), and intervals in which the beginning and ending dates in any of the above formats are separated by the "/" (e.g., 2004-02/2014-03-23).

The use of Level 1 EDTF features (e.g., uncertain/approximate dates, unspecified dates, extended intervals, years exceeding four digits and seasons) was much lower overall. It was observed in 2.6% of all metadata records with EDTF-valid date strings. Most (12 out of 14) of DPLA content hubs and over a half (5 out of 9) of DPLA service hubs were found to make use of Level 1 EDTF features to greater or lesser extent.

The use of Level 2 EDTF features (e.g., partial uncertain/approximate dates, partial unspecified dates, sets, multiple dates, masked precision and extensions of the extended interval, years exceeding four digits, seasons) was observed in only 0.3% of all metadata records with EDTF-valid date strings. One service hub and one content hub make use of Level 2 features, with EDTF-Level2-valid date strings present in 1.8% and 4.8% of their DPLA metadata records respectively.

DISCUSSION AND CONCLUSION

Our analysis of over 8 million of Digital Public Library of America (DPLA) metadata records revealed that slightly over half of date strings in Date fields are EDTF-valid date strings, mostly conforming to EDTF Level 0 features. Only two hubs used EDTF-Level2-valid date strings in a small percentage of their metadata records. It is worth noting that date strings created according to some of the common date formats – ISO and W3C – that are widely used both within and outside metadata creation context (e.g., 1999, or 2000-04-03) are valid EDTF date strings despite the fact that these date strings may not have been created with the idea of supporting EDTF. Many of the "valid EDTF" dates in the DPLA fall into this category. However, this is an important finding as it means that institutions or DPLA hubs that might want to convert date strings in their date and time metadata to a machine-readable format may already have large portions of their collections in conformance with EDTF if they have many dates in the format that matches the EDTF Level 0 date format specifications.

We observed some similarities between the two groups of DPLA national partners – content hubs and service hubs – in application of Date metadata. The overall level of use of

Date element in DPLA metadata records was high and very similar for content hubs and service hubs. The differences were observed between two groups in the proportion of hubs with EDTF-valid date strings conforming to Level 1 EDTF features. EDTF-Level1-valid date strings were observed in metadata records of a higher proportion of content hubs.

The date is one of the most important metadata elements that can influence the full understanding, sharing and use of digital content. Despite the fact that EDTF is still a relatively new date specification which has not yet been formally adopted as either a standard of its own or as an extension of an existing standard, a number of institutions and initiatives are interested in using the EDTF specification as a way of representing complex date strings found in cultural heritage collections' metadata records. Due to exploratory nature of the study reported in this paper, its limitations include using a single study target (DPLA) and a single data analysis approach exploratory content analysis – to address its research questions. Future studies will need to combine comparative content analysis date representation in metadata records of multiple similar aggregators such as DPLA, Europeana, Canadiana, with analysis of date and time metadata guidelines in hubs' metadata creation policies and survey of hub representatives' opinions about factors affecting their date metadata creation practices.

REFERENCES

Barton, J., Currier, S., & Hey, J.M.N. (2003). Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. In DC-2003: *Proceedings of the International DCMI Metadata Conference and Workshop.* [United States]

Bruce, T.R., & Hillmann, D.I. (2004). The continuum of metadata quality: defining, expressing, exploiting. In Hillman, D. and Westbrook, L. (Eds.), *Metadata in Practice*. Chicago: American Library Association, pp. 238-256.

Bui, J., & Park, J. (2006). An assessment of metadata quality: a case study of the National Science Digital Library metadata repository. In Moukdad, H. (Ed.), *Proceedings of CAIS/ACSI 2006*, pp.13.

DPLA. (2015a). An introduction to the DPLA metadata model. Retrieved May 29, 2015 from http://dp.la/info/wp-content/uploads/2015/03/Intro_to_DPLA_metadata_model.pdf

DPLA. (2015b). Metadata application profile: Version 4.0. Retrieved May 29, 2015 from http://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf

Dushay, N., & Hillmann, D.I. (2003). Analyzing metadata for effective use and re-use. In DC-2003: *Proceedings of the International DCMI Metadata Conference and Workshop.* [United States]: DCMI.

Extended Date/Time format (EDTF). (2012). Retrieved May 29th, 2015 from http://www.loc.gov/standards/datetime/pre-submission.html#introduction

Jackson, A.S., Han, M., Groetsch, K., Mustafoff, M., & Cole, T.W. (2008). Dublin Core metadata harvested through OAI-PMH. *Journal of Library Metadata*, *8*(1), 5-21.

Kurtz, M. (2010). Dublin Core, DSpace, and a brief analysis of three university repositories. Information Technology & Libraries, 29(1), 40-46.

Loy, D., & Landis, B. (2005). *Date normalization utility (DNU) documentation*. Retrieved May 29, 2015 from http://www.cdlib.org/services/access_publishing/dsc/projects/docs/datenorm_documentation.pdf

Ma, H. (2014). Techservices on the Web: DPLA: Digital Public Library of America. *Technical Services Quarterly*, *31*(1), 83-84. doi: 10.1080/07317131.2014.845013

Mitchell, E.T. (2013). Three case studies in linked open data. *Library Technology Reports*, *49*(5), 26-43.

Moen, W.E., Stewart, E.L, & McClure, C.R. (1998). *The Role of Content Analysis in Evaluating Metadata for the U.S. Government Information Locator Service (GILS): Results from an Exploratory Study*. Retrieved May 29, 2015 from http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm

Park, J. (2006). Semantic interoperability and metadata quality: An analysis of metadata item records of digital image collections. *Knowledge Organization, 33* (1), 20-34.

Park, J., & Maszaros, S. (2009). Metadata Object Description Schema (MODS) in digital repositories: An exploratory study of metadata use and quality. *Knowledge Organization, 36* (1), 46-59.

Park, J. & Tosaka, Y. (2010). Metadata quality control in digital repositories and collections: criteria, semantics, and mechanisms. *Cataloging & Classification Quarterly*, *48* (8), 96-715.

Shreeves, S.L., Kaczmarek, J., & Cole, T.W. (2003). Harvesting cultural heritage metadata using the OAI protocol. *Library Hi Tech, 21*, 159–169.

Shreeves, S., Knutson, E., Stvilia, B., Palmer, C., Twidale, M., & Cole, T. (2005). Is "quality" metadata "shareable" metadata? The implications of local metadata practices for federated collections. In Thompson, H.A. (Ed.). *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries*, pp. 223-237.

Tarver, H., Waugh, L., Alemneh, D.G., & Phillips, M.E. (2015). Managing serials in a large digital library: case study of the UNT Libraries Digital Collections. *The*

Serials Librarian. Retrieved May 29, 2015 from http://digital.library.unt.edu/ark:/67531/metadc406384/

Tarver, H., & Phillips, M. E. (2013). Lessons learned in implementing the Extended Date/Time Format in a large digital library. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 60-70. Retrieved May 29, 2015 from http://digital.library.unt.edu/ark:/67531/metadc174739/

Tennant, R. (2004). *Specifications for Metadata Processing Tools*. Retrieved May 29, 2015 from http://roytennant.com/metadata_tools.pdf

Weagley, J., Gelches E., & Park, J. (2010). Interoperability and metadata quality in digital video repositories: a study of Dublin Core. *Journal of Library Metadata, 10*(1), 37-57. DOI: 10.1080/19386380903546984.

Zavalina, O.L. (2011). Contextual metadata in digital aggregations: Application of collection-level subject metadata and its role in user interactions and information retrieval. *Journal of Library Metadata, 11*(3/4), 104-128.

Zeng, M., Subrahmanyam, B., & Shreve, G.M. (2005). Metadata quality study for the National Science Digital Library (NSDL) metadata repository. *Lecture Notes in Computer Science, 3334*, 339-340.