Representation of Recorded Knowledge and Extended Date/Time Format:

A Case Study of the Digital Public Library of America

Oksana L. Zavalina
Mark E. Phillips
Priya Kizhakkethil
Daniel Gelaw Alemneh
Hannah S. Tarver

A green light to greatness.

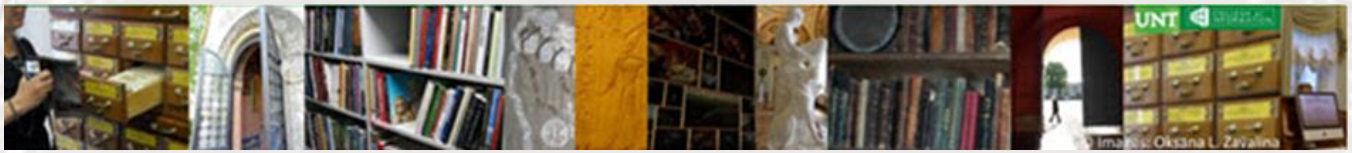
University of North Texas

UNT

ICKM 2015, Osaka, Japan

Background of the study: Date and time representation

- Metadata-based digital resource management in [digital] repositories



- Information retrieval is affected by structured representation of date/time of :
  - resource creation
  - publication/ issuance
  - digitization
  - validity
  - etc.

20+ metadata quality criteria:

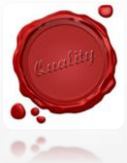
Most important (*Park & Tosaka, 2010*):

- Access
- **Accuracy**
- Availability
- Compactness
- Compatibility
- **Completeness**
- Comprehensiveness
- Content
- **Consistency**
- Cost
- Data structure
- Ease of creation
- Ease of Use
- Economy
- Flexibility
- Fitness for Use
- Informativeness
- Protocols
- Quantity
- Reliability
- Standard
- Timeliness
- Transfer
- Usability

Background of the study: Date and time representation and metadata quality

- **Date** metadata field is required if applicable in many repositories
- **Date** found in 86%+ of metadata records, e.g.:

completeness



- 86% in NSDL (*Zeng, Subrahmanyam, & Shreve, 2005; Bui & Park, 2006*)
- 86% in IMLS DCC aggregation (*Jackson et al., 2008*)
- 92% in OAIster (*Ward, 2003)*
- 96% in digital video collections (*Weagley, Gelches, & Park, 2010*)

Background of the study: Date and time representation and metadata quality (2)

accuracy

- Confusion of Dublin Core's **Date** and **Coverage** elements (*Jackson et al., 2008*)
- Mapping MARC **260 $c** (date of publication) to Dublin Core's **Publisher** or MODS **originInfo** subelement **placeName** (*Jackson et al., 2008; Park & Maszaros, 2009*)

consistency

- Lack of consistency in the format of data value (*Dushay & Hillmann, 2003; Shreeves at al., 2003, 2005*)
- Data normalization for **Date** metadata in repositories (*Tennant, 2004; Loy & Landis, 2005; Tarver & Phillips, 2013; Tarver et al., 2015*)

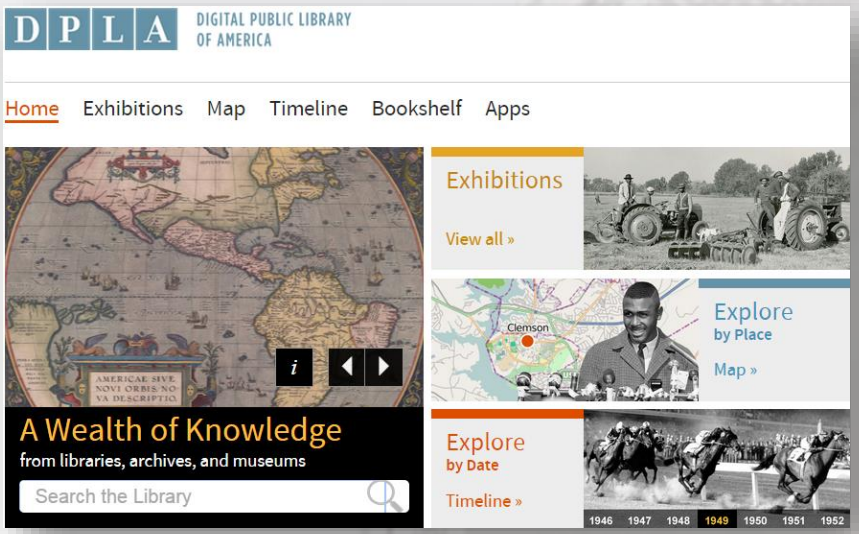
Background of the study: EDTF

http://www.loc.gov/standards/datetime/pre-submission.html#introduction

- A candidate for a single standard for encoding and normalization of dates and times in large-scale centralized databases
- Extension to ISO 8601 [W3CDTF]
- One of the most **consistent** and **flexible** specifications
- **Expressive** : allows for standardized representation of virtually all possible kinds of dates and date ranges

3 levels: each more expressive

Background of the study: DPLA



- Digital Public Library of America
- One of the largest digital repositories (launched in 2013)
- Distributed network model:
  - **Content hubs**
  - **Service hubs**

- RDF-based "data model" (a.k.a. metadata application profile)
  - **Date** is recommended field
  - Use of **EDTF** is recommended for Date field

- Metadata normalized at harvesting
  - All dates mapped to Dublin Core's **Date Created**

Problem statement & Research questions

Lack of systematic empirical studies of digital repository metadata with the focus on **date and time** metadata

- in very large aggregations (Hathi Trust, Internet Archive, DPLA, etc.)

How are dates represented in metadata records in DPLA? Is EDTF used?

- What are the differences and similarities in date metadata originating from **content hubs** and **service hubs**?

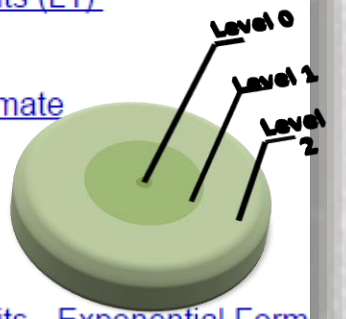
Data Collection

- Big Data approach
- DPLA Bulk Download
  http://dp.la/info/developers/download
  - over 8 million metadata records
  - Python ExtendedDateTimeFormatModule
    https://github.com/unt-libraries/ExtendedDateTimeFormat
  - Solr full-text indexer: Stats Component
    http://wiki.apache.org/solr/StatsComponent

Quantitative Explorative Content Analysis

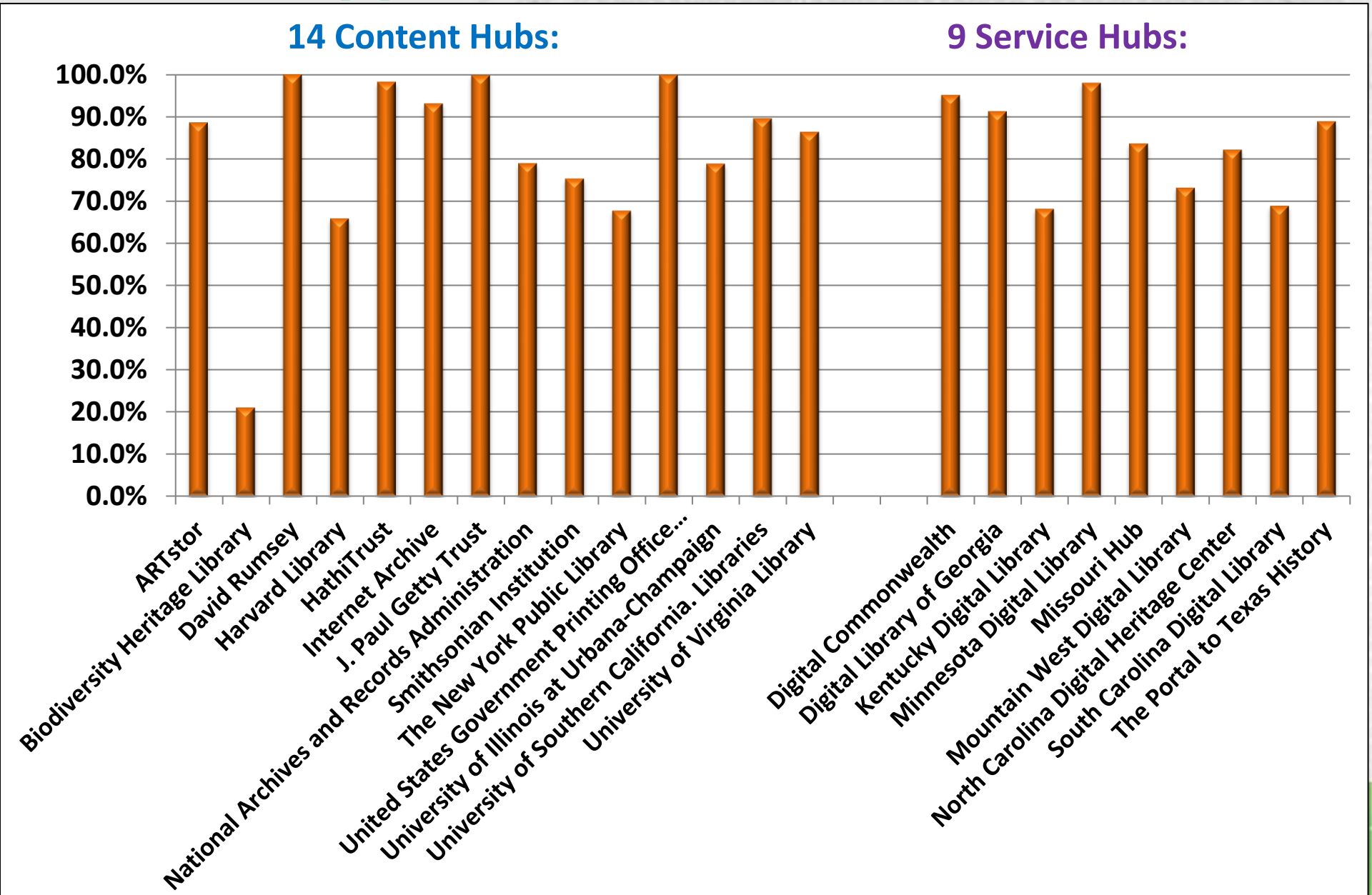
- Proportion of metadata records with **Date** metadata
  - EDTF-valid Date
    - Conforming to EDTF Level 0
    - Conforming to EDTF Level 1
    - Conforming to EDTF Level 2

5.1 Level 0 Features
5.1.1 Date
5.1.2 Date and Time
5.1.3 Interval
5.2 Level 1 Features
5.2.1 Uncertain/Approximate
5.2.2 Unspecified
5.2.3. Extended Interval (L1)
5.2.4 Year Exceeding Four Digits (L1)
5.2.5 Season
5.3 Level 2 Features
5.3.1 Partial Uncertain/Approximate
5.3.2 Partial Unspecified
5.3.3 One of a Set
5.3.4 Multiple Dates
5.3.5. Masked Precision
5.3.6 Extended Interval (L2)
5.3.7 Year Exceeding Four Digits - Exponential Form
5.3.8 Season - Qualified

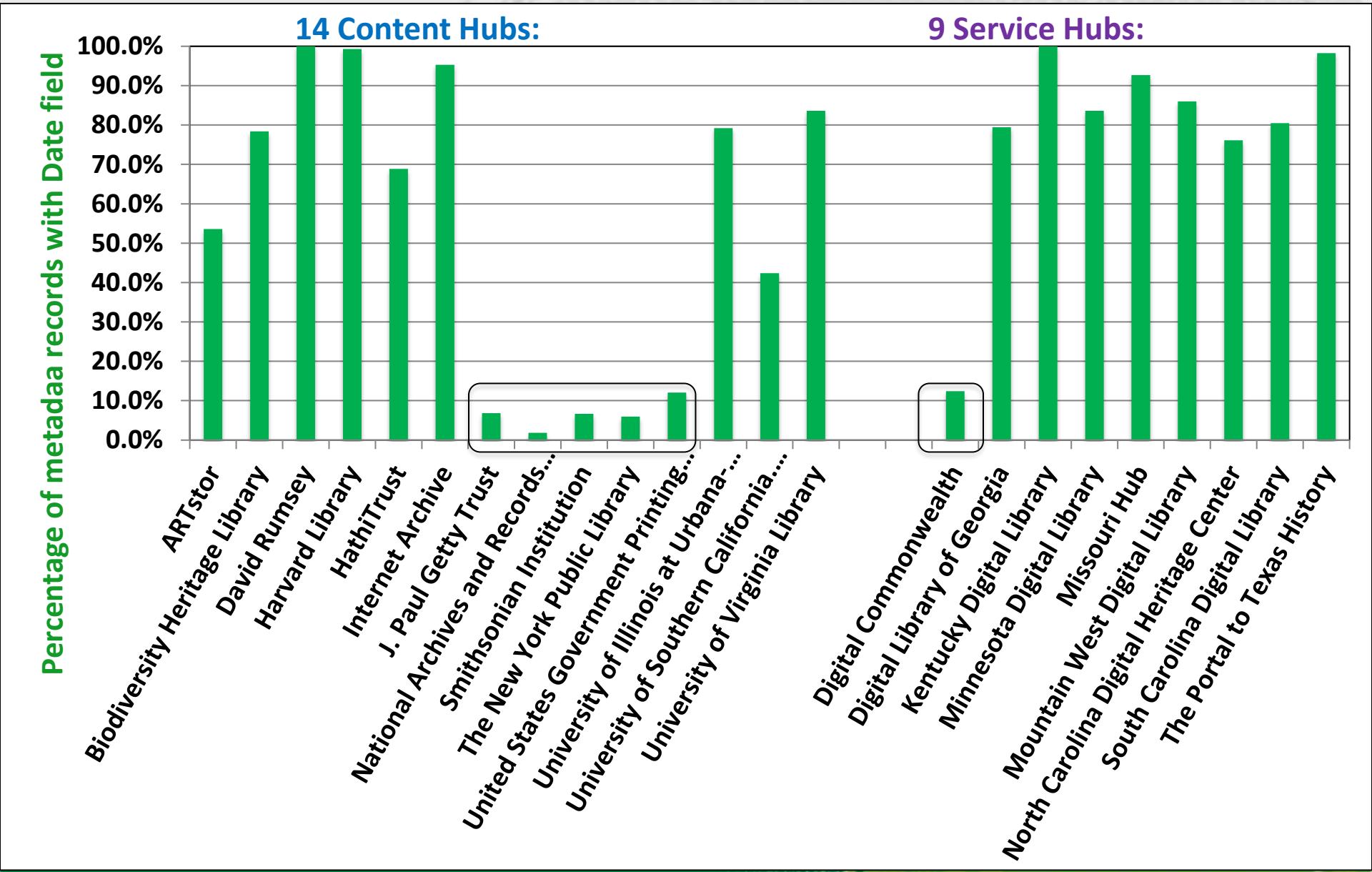


- Common **Date** data value patterns?

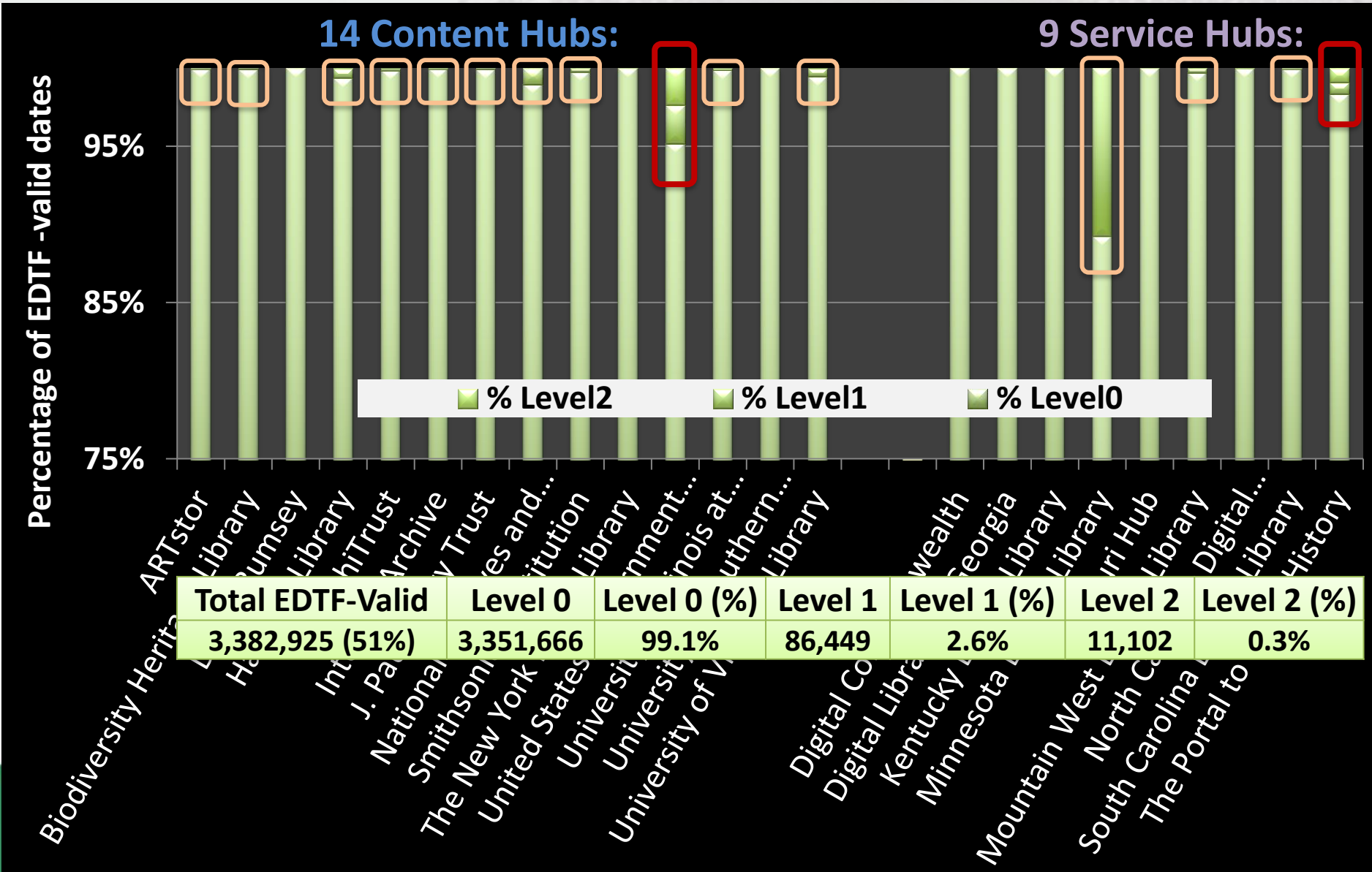
Level of application of Date metadata field



EDTF-valid Date strings



EDTF-valid Date strings by EDTF Level



Total EDTF-Valid	Level 0	Level 0 (%)	Level 1	Level 1 (%)	Level 2	Level 2 (%)
3,382,925 (51%)	3,351,666	99.1%	86,449	2.6%	11,102	0.3%

10 most-used EDTF-valid Date patterns

Date Pattern	Number	Example Date String
0000	2,114,166	2004 Level 0
0000-00-00	1,062,935	2004-10-23 Level 0
0000-00	107,560	2004-10 Level 0
0000/0000	55,965	2004/2010 Level 0
0000?	13,727	2004? Level 1
[0000-00-00..0000-00-0]	4,434	[2000-02-03..2001-03-04] Level 2
0000-00/0000-00	4,181	2004-10/2004-12 Level 0
0000~	3,794	2003~ Level 1
0000-00-00/0000-00-00	3,666	2003-04-03/2003-04-05 Level 0
[0000..0000]	3,009	[1922..2000] Level 2

Top 5 EDTF-valid patterns by DPLA hub

Content Hubs:	Pattern 1	Pattern 2	Pattern 3	Pattern 4	Pattern 5
ARTstor	0000	0000-00	0000?	0000/0000	0000-00-00
Biodiversity Heritage Library	0000	-0000	0000/0000	0000-00	0000?
David Rumsey	0000				
Harvard Library	0000	00aa	000a	aaaa	
HathiTrust	0000	0000-00	0000?	-0000	00aa
Internet Archive	0000	0000-00-00	0000-00	0000?	0000/0000
J. Paul Getty Trust	0000	0000?			
National Archives and Records Administration	0000	0000?			
Smithsonian Institution	0000	0000?	0000-00-00	0000-00	00aa
The New York Public Library	0000-00-00	0000-00	0000	-0000	0000-00-00/0000-00-00
United States Government Printing Office (GPO)	0000	0000?	aaaa	-0000	[0000, 0000]
University of Illinois at Urbana-Champaign	0000	0000-00-00	0000?	0000-00	
University of Southern California. Libraries	0000-00-00	0000/0000	0000	0000-00	0000-00/0000-00
University of Virginia Library	0000-00-00	0000	0000-00	0000?	0000/0000
Service Hubs:					
Digital Commonwealth	0000-00-00	0000-00	0000	0000-00-00a00:00:00a	
Digital Library of Georgia	0000-00-00	0000-00	0000/0000	0000	0000-00-00/0000-00-00
Kentucky Digital Library	0000				
Minnesota Digital Library	0000	0000-00-00	0000?	0000-00	0000-00-00?
Missouri Hub	0000-00-00	0000	0000-00	0000/0000	0000?
Mountain West Digital Library	0000-00-00	0000	0000-00	0000?	0000-00-00a00:00:00a
North Carolina Digital Heritage Center	0000-00-00	0000	0000-00	0000/0000	0000?
South Carolina Digital Library	0000-00-00	0000	0000-00	0000?	
The Portal to Texas History	0000-00-00	0000	0000-00	[0000-00-00..0000-00-00]	0000~

1st most-used overall | 2nd most used | 3rd most used | 4th most used | 5th most used

Content hubs vs. service hubs

Similarities:	Differences:
•Overall level of use of **Date** metadata : **83.4%** for **content hubs** & **80.9%** for **service hubs**.	•Much more common for **service hubs** to include EDTF-valid date strings.
•Single hub in each group with EDTF-**Level2**-valid date strings; in similarly small percentage of records: • GPO •Portal to TX History	• More complex EDTF-**Level0**-valid date formats (**yyyy-mm-dd**) used predominantly in **service hubs**' metadata. •Most common date patterns simpler (**yyyy**) in **content hubs**' metadata •EDTF-**Level1**-valid date strings appear in metadata of a higher proportion of **content hubs**.

10 most-used non-EDTF Date patterns

Date Pattern	Number	Example Date String
0000-0000	1,117,718	2005-2006
00/00/0000	486,485	03/04/2005
[0000]	196,968	[2006]
[aaaa aaaaaaaaaaa]	183,825	[Date Unavailable]
00 aaa 0000	143,423	22 Aug 2006
0000 – 0000	134,408	2000 – 2005
0000-aaa-00	116,026	2003-Dec-23
0 aaa 0000	62,950	3 Jan 2012
0000]	58,459	1933]
aaa 0000	43,676	Oct 2015

Simple solutions available (next slide)

Possible solutions

Simple lossless transformations:

- **0000-0000** and **0000 -- 0000** to **0000/0000** [change over **1 M Date** values to EDTF- valid format].
- **00/00/0000** to **0000-00-00** [if the original date structure is known: 00/00 for mm/dd (US) or dd/mm (non-US)]
- **[0000]** to **0000**

Will allow to:

- convert most of non-EDTF dates in DPLA into EDTF-valid
- enrich the DPLA metadata with ~**2 M** additional date strings -- to 86% of all **Dates** in DPLA
- improve metadata interoperability

Summary

- **83%** DPLA records contain **Date** metadata -- somewhat lower than observed in previous studies of digital repository metadata
- **49%** date strings in DPLA **Date** fields are non-EDTF-valid
  - most follow patterns that can easily be brought into conformance.
- **51%** date strings are EDTF-valid:
  - 99.1% of these conform to Level 0 standards (incidental?)
  - basic **Date** string formats also conform to ISO or W3C and have likely been used in native metadata without intention of EDTF conformance: e.g., yyyy, yyyy-mm, and yyyy-mm-dd
- DPLA hubs that might want to convert **Date** metadata to a machine-readable format may already have high proportion of metadata in conformance with EDTF where date format matches the EDTF Level 0 date format specifications.

Limitations

- Exploratory nature of the study
  - single research method – exploratory content analysis – to address its research questions.

- Focus on only one temporal property in metadata:
  - the dates that represent events in the lifecycle of information objects such as the date of creation.

Future Research

- Combine content analysis of metadata records with:
  - content analysis of DPLA content hubs' and service hubs' local policies for **Date** metadata guidelines
  - survey of DPLA partners about **Date** metadata creation decisions and factors affecting these decisions.

- Focus on dates/times that represent **intellectual content** of an information resource
  - **Temporal Coverage** metadata element of Dublin Core metadata scheme.

Cited Works

•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.

•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.

•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.

•Loy, D., & Landis, B. (2005). *Date normalization utility (DNU) documentation*. Retrieved August 1, 2015 from http://www.cdlib.org/services/access_publishing/dsc/projects/docs/datenorm_documentation.pdf

•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 August 1, 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 August 1, 2015 from http://digital.library.unt.edu/ark:/67531/metadc174739/

•Tennant, R. (2004). *Specifications for Metadata Processing Tools*. Retrieved August 1, 2015 from http://roytennant.com/metadata_tools.pdf

•Ward, J. (2003). A quantitative analysis if unqualified Dublin Core metadata element set usage within data providers registered with the Open Archives Initiative. *Proceedings of the 2003 Joint Conference on Digital Libraries*, pp. 315-317.

•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.

Questions?
Comments? Ideas?

Please contact us:

Oksana L. Zavalina: Oksana.Zavaina@unt.edu