

Lessons Learned in Implementing the Extended Date/Time Format in a Large Digital Library

Hannah Tarver
University of North Texas
Libraries, USA
hannah.tarver@unt.edu

Mark Phillips
University of North Texas
Libraries, USA
mark.phillips@unt.edu

Abstract

In 2012, the University of North Texas (UNT) Libraries implemented the Library of Congress Extended Date/Time Format (EDTF) into the metadata guidelines for their digital holdings which now contain more than 460,000 records. This paper discusses the evaluation process to identify the number of previously-existing dates that meet EDTF standards and those that need to be edited for conformance. It also outlines practical steps used for implementing the standard, such as date validation for metadata creators and changes to date displays for public users. Finally, it presents some of the challenges encountered during the implementation process and considerations for other institutions that may want to use the EDTF.

Keywords: metadata; date formats; standardization; Extended Date/Time Format; standards implementations; digital libraries

1. Introduction

The University of North Texas (UNT) Libraries have made digital library holdings a priority by supporting the creation of digital library infrastructure and the growth of digital collections that are primarily open access. The UNT digital collections comprise three public-facing repositories: The Portal to Texas History (<http://texashistory.unt.edu>), the UNT Digital Library (<http://digital.library.unt.edu>), and The Gateway to Oklahoma History (<http://gateway.okhistory.org>). In total these collections contain more than 460,000 digitized and born-digital objects, including text, images, and audio/visual items. All of the item records follow the same metadata schema, which is based on Dublin Core.

Before formally switching to the Extended Date/Time Format (EDTF) as the primary date standard, the UNT Libraries used the International Standards Organization (ISO) 8601 standard for formatting dates: Data elements and interchange formats – Information interchange – Representation of dates and times. Although the ISO standard meets usage needs for general dates, it does not address many of the complex kinds of dates represented in library and archival collections, such as approximate or partially-known dates, which are common for historical objects.

Due to the large range of date types, the UNT Libraries decided to move to the EDTF because it contained standardized ways of representing many of the dates that the ISO standard does not address. Although this shift has primarily been positive, any change involves some adjustments and challenges. This paper will outline steps that the UNT Libraries have taken to implement the EDTF, including the evaluation of existing dates for compliance with the EDTF, the use of date validation for metadata creators, and the normalization of date displays for users. Additionally, it will summarize some of the challenges that the UNT Libraries have encountered and considerations for other institutions.

2. UNT Libraries Metadata

The UNT digital collections use locally-qualified Dublin Core metadata to represent a range of information about the original physical or digital object. Metadata is available for harvesting from these collections in several formats, including basic Dublin Core records. Metadata labeled “UNTL” refers to the original metadata created in the UNT system that contains all local fields and qualifiers without any conversions or simplifications and matches the records displayed to users. The UNT digital collections have an in-depth set of formatting guidelines (UNT Libraries, *Input Guidelines*, 2013) to support metadata creation; when appropriate, the guidelines also reference external standards, such as date standard information from the EDTF (UNT Libraries, *Date*, 2013).

2.1. Date Fields

There are two fields in UNTL metadata that contain date information. The Date field represents important dates in the lifecycle of the physical or digital object, including the creation date of the original item, submission and acceptance dates for patents, harvest dates for web archiving, and embargo dates for scholarly papers and items that have restricted access. There is a qualifier for each of these date types: Creation Date, Submission Date, Acceptance Date, Harvested Date, and Embargoed Until Date.

Creation dates take a variety of forms that are often related to the types of items. In the library and museum fields, it is not unusual to have “circa” dates for items or a general range of time when an item may have been created; however, a specific creation date is sometimes known and may be at the year, month, or day precision. In the case of born-digital photographs, there is often an exact date-time stamp on the items but metadata creators do not always have access to the information in order to include that level of specificity. Serial texts almost always have a publication date clearly noted, but over the course of the publication the frequency often changes so that dates for one title may vary between years, months, seasons, and sometimes even issuance days. To address the broad range of associated dates, creation dates in the digital collections use nearly all of the date types specified in the EDTF.

The Coverage field is the second element containing dates and includes date, time period, and geographic location information about the content of the item. Coverage date information is labeled with a “Coverage Date” qualifier or “Start Date” and “End Date” qualifiers for representing date ranges. Unlike creation dates, coverage dates tend to be a single date or date range and rarely contain more specific or complex dates since they reflect the content, which is usually stated or apparent in the item.

3. About the Extended Date/Time Format

The Extended Date/Time Format (EDTF) is a draft date-time standard initiated by the Library of Congress with the intention of creating more explicit date formatting and addressing date types that are not currently regulated by ISO 8601 (see the Appendix for examples). Current suggestions for additions are being noted and discussed within the EDTF community with the intention of formalizing the EDTF as an ISO 8601 amendment or as an extension to other Web-based date standards (Library of Congress, 2012).

3.1. Structure of the EDTF

There are three levels of support in the EDTF allowing an organization to implement only the most basic level (0), the first two levels (0 and 1), or the full complement of options (levels 0-2). Level 0 includes features supported by ISO 8601 while levels 1 and 2 include extensions to the features in ISO 8601 to allow for additional date types. Each level contains all of the functionality for the previous levels.

Features of Level 0 include basic dates (a year, a year and month, or a year, month, and day), dates that have timestamps, and intervals (a range of time between two dates that may be at varying levels of precision). Although all of these date types are supported by ISO 8601, the EDTF prescribes usage by requiring punctuation that is optional for compliance with the ISO standard (International Standards Organization, 2004).

Level 1 introduces uncertain and approximate dates, unspecified or unknown digits within specific parameters, intervals containing only one known date or some levels of approximation and uncertainty, years containing more than four digits (e.g., a date during the Cretaceous period, such as -70000000), and seasons.

Level 2 allows for the representation of partial uncertainty or imprecision by marking a part of the date that is uncertain, approximate, or unspecified in a single date or date interval. It also includes inclusive date lists that are not consecutive and date lists that represent “one of a set” (e.g., one of the years or dates in a non-consecutive list, or a date that is before or after a known date). Finally, Level 2 provides some ability for further clarification by adding a qualifier to seasons, exponential forms of years that exceed four digits, and masked precision for years.

3.2. Adoption of EDTF

Although EDTF is still in a draft state, it has been established and formalized for usage. The Library of Congress has integrated the EDTF into other standards managed by the organization, such as the Metadata Authority Description Standard (MADS), Metadata Encoding & Transmission Standard (METS), Preservation Metadata standard (PREMIS), and Metadata Object Description Standard (MODS). Additionally, other institutions – including the Digital Public Library of America (DPLA) – are considering the usage of EDTF as an alternative or addition to other date standards (Digital Public Library of America, 2013). The EDTF has potential benefits for organizations that have specific date/time needs both because it already outlines standards to handle various date uncertainties and because the EDTF community is still discussing amendments for the future to incorporate additional kinds of dates.

4. Establishing a Baseline in UNT Digital Collections

Although the UNT Libraries began implementation of EDTF-compliant dates in 2012, nearly 300,000 records existed in the system prior to the shift. Not only do some records contain “legacy” dates, but some metadata creators may have continued to enter non-EDTF-valid dates in the system. One of the initial steps toward EDTF compliance involves an evaluation of the dates in the system metadata records to determine how many dates meet EDTF standards and how many dates will require editing to bring them in line with the formatting guidelines.

4.1. Collecting Data

The authors conducted an investigation to better understand the range of values present in a large digital library installation by obtaining date values from the UNT digital collections. At the time of the analysis, there were 379,392 unique digital items present in the system. This paper contains an analysis of all dates from the Date field – including all qualifier types – but does not include dates from the Coverage field. Further research may include coverage date values, as they make greater use of range information.

The three repositories used in this analysis make their metadata publicly available via the Open Archives Initiative’s Protocol for Metadata Harvesting (OAI-PMH) in a variety of different record formats. The authors harvested UNTL-format metadata records using a Python OAI-PMH harvester written by the authors. The results were stored as three large eXtensible Markup Language (XML) files, one for the entire metadata holdings of each repository (see Table 1).

TABLE 1: Records harvested from UNT digital collections.

Repository	URL	Number of records
The Portal to Texas History	http://texashistory.unt.edu/oai/	258,455
The UNT Digital Library	http://digital.library.unt.edu/oai/	72,937
The Gateway to Oklahoma History	http://gateway.okhistory.org/oai/	48,000

The authors extracted dates from the harvested metadata with a general metadata analysis tool used by the UNT Libraries to convert OAI-PMH data into formats easily consumed by standard command-line tools (Phillips, *Metadata Analysis*, 2013). The date values for each repository were concatenated into a single file containing all values from the Date field across the three repositories (Phillips, *EDTF Datasets*, 2013). The single date file contains 390,751 total date instances, one per row in the file, and 55,212 unique date values (see Table 2).

TABLE 2: Breakdown of date instances in each repository.

Repository	Total Records	Records Without Dates	Records With Dates	Total Date Instances	Unique Date Instances
The Portal to Texas History	258,455	24,074	234,381	262,930	52,066
The UNT Digital Library	72,937	2,662	70,275	79,821	10,562
The Gateway to Oklahoma History	48,000	0	48,000	48,000	11,510

4.2. Analyzing the Data

Once the dataset was established, the next step was to determine which of the dates already met EDTF standards. To understand how the three repositories make use of the EDTF, the authors wrote a classifier that takes a date instance as input and returns information regarding whether the date meets valid EDTF standards, and, if so, the level to which the date conforms. All of the dates in the concatenated file were fed into the program and classified (see Table 3).

TABLE 3: Breakdown of validation for all date instances.

	Not Valid EDTF	Valid EDTF	Level 0	Level 1	Level 2	Total Dates Analyzed
All Instances	11,069	379,682	377,059	2,609	14	390,751
Unique Instances	2,369	52,843	52,361	471	11	55,212

As evidenced in Table 3, a large percentage (97%) of the dataset conformed to the EDTF specification. This is likely due to the fact that the UNT Libraries have always recommended the use of ISO 8601 for date formatting whenever possible; since Level 0 compliance overlaps with ISO 8601, many of the values are also compliant with EDTF. Another factor may be related to the content in the three repositories. Newspapers represent a large percentage of items – 166,322 records or 64% of items in The Portal to Texas History and 100% of the 48,000 records in The Gateway to Oklahoma History – and typically contain unambiguous publication dates, making it relatively easy to include a fully-formed date value for more than half of the total digital holdings (roughly 56%).

For dates that met EDTF standards, each compliance level was broken down by specific features. The 377,059 (52,361 unique) Level 0 dates can be classified by measures of precision and dates containing intervals (see Table 4).

TABLE 4: Classifications of Level 0 dates in UNT repositories.

Classification	Number of Date Instances	Unique Date Instances
Year only	58,232	361
Year and Month	25,892	1,620
Year, Month, and Day	290,670	49,735
Year, Month, Day, and Time	110	107
Interval	2,154	537

The Level 1 instances encompass 2,609 date values (471 unique) in the dataset. These instances can be further classified by the features available in the Level 1 feature set (see Table 5).

TABLE 5: Classifications of Level 1 dates in UNT repositories.

Classification	Number of Date Instances	Unique Date Instances
Uncertain/Approximate	2,283	343
Unspecified	94	16
Extended Interval (L1)	207	74
Year Exceeding Four Digits	0	0
Season	232	112

Finally, 14 dates conform to Level 2 values and fall into several categories of Level 2 features (see Table 6).

TABLE 6: Classifications of Level 2 dates in UNT repositories.

Classification	Number of Date Instances	Unique Date Instances
Partial Uncertain/Approximate	3	2
Partial Unspecified	7	7
One of a Set	4	2
Multiple Dates	0	0
Masked Precision	0	0
Extended Interval (L2)	0	0
Season - Qualified	0	0

4.3. Invalid EDTF Dates

There are 11,069 (2,369 unique) date values in the dataset that do not conform to the EDTF standards (see Table 3). These values are of particular interest to the authors as they may represent date instances entered by metadata creators before the EDTF standard was adopted, malformed EDTF instances, or date types that are difficult to format in EDTF. Non-EDTF-valid dates can be categorized in several broad groups including common non-EDTF punctuation and text strings (see Table 7).

TABLE 7: Types of non-EDTF-valid dates in UNT repositories.

Incorrect Punctuation	Number of Date Instances	Unique Date Instances
Enclosed in square brackets	3,623	996
Trailing periods	166	74
Space hyphen space	2,142	511
c., circa, ca., etc.	4,397	654
Tilde with incorrect date formations	157	80
Question marks (possibly incorrect)	1,124	223

The authors noticed that many of the non-EDTF-valid dates contain more than one of the previously-mentioned indicators. For example, the value [189?] which the authors interpret as “a date supplied by the metadata technician that is sometime during the 1890s” contains both square brackets and a question mark in place of a digit; this date can be represented by the EDTF value of 1890~/1899~.

Based on a long-standing practice in the library cataloging and metadata world, many of the invalid dates contain square brackets to represent values not taken from the physical piece, but rather from a source that is not the item in hand. However, the use of square brackets conflicts directly with a feature in Level 2 of the EDTF specification used to represent “one of a set” for a date list. The presence of trailing periods is another holdover from cataloging practice, though it is not generally used for date values. It is possible that the trailing periods were copied or carried over from another catalog or database into UNTL metadata records.

Before the UNT Libraries transitioned to EDTF, they employed a number of practices to help metadata creators represent dates, including the use of question marks to represent unknown digits in a date string; since the adoption of the EDTF specification, the use of a question mark has been superseded by the use of a “u,” which stands for “unknown.” Similarly, previous metadata practice allowed for a date range to be separated by the pattern “space-hyphen-space” which has been replaced by the use of a slash (/) to represent an interval.

A majority of the non-EDTF-valid dates contain text strings to represent approximate dates – “circa” or “ca.” – or a copyright date. Although these notations are common for archival objects, they are not supported by the EDTF standard. Instead, the use of a tilde (~) represents approximate dates, and copyright dates are no longer formally noted in the date field, although they may be represented elsewhere in a metadata record depending on local practice.

Finally, a number of the non-EDTF-valid dates do not conform to either the ISO standard or the EDTF although they may be common representations outside of standards-regulated dates. For example, some dates use a slash instead of a hyphen (e.g., 2001/02/20) or contain fewer than four digits.

The authors tried applying a series of transforms to the invalid date strings to determine the ease of converting some values directly into valid EDTF strings. This work was experimental and further testing is needed before implementing the process in a data cleanup workflow for live collections. Table 8 contains a list of heuristics applied to the date strings.

TABLE 8: Examples of rules for converting non-EDTF-valid dates into valid EDTF dates.

Problem or Incorrect Punctuation	Rule/Change	Example of Conversion
Dates in square brackets	Remove brackets	[2000] >> 2000
Dates with a trailing period	Remove period	1922. >> 1922
Dates with ca., c., or circa	Remove strings, add “~”	c. 1920 >> 1920~
Dates with “ – ”	Convert “ – ” to “/”	1920 – 1922 >> 1920/1922
Dates with internal “?”	Convert “?” to “u”	19?? >> 19uu
Dates with dddd-dddd	Convert “-” to “/”	1922-1926 >> 1922/1926

The resulting dataset was fed into the date validator to determine the remaining number of non-EDTF-valid dates (see Table 9).

TABLE 9: Breakdown of validation for all date instances after applying initial date conversions.

	Not Valid EDTF	Valid EDTF	Level 0	Level 1	Level 2	Total Dates Analyzed
All Instances	1,397	9,672	8,581	1,065	26	11,069
Unique Instances	683	1,178	992	161	25	1,861

The authors feel that the automated cleaning of dates provided insight into the level of complexity that is required to convert non-EDTF-valid dates, however, doing so in an automated fashion without verification of the resulting dates may introduce unexpected errors. Furthermore, some dates require manual evaluation and adjustment in order to verify that the intended date is represented correctly in the EDTF format.

5. Implementation

Implementing the EDTF in UNTL metadata requires an ongoing process to ensure that dates in new records meet the standards and to correct dates in the older records when necessary. A successful implementation includes tools to help metadata creators format dates appropriately and measures to make dates understandable by public users viewing the records.

5.1. EDTF Validation and Metadata Entry

The UNT Libraries developed an EDTF validation Web service (UNT Libraries, *Validation Service*, 2013) in order to provide feedback to metadata creators when they are inputting data into the Web-based metadata entry forms used by the UNT Libraries digital library system.

This service takes a candidate EDTF date and returns a JavaScript Object Notation (JSON) of “True” if the submitted value conforms to any level of the full EDTF specification or “False” if it does not validate against the EDTF. Additionally, the Web service allows arguments that provide the ability to designate the level of specification a date should verify against and an optional callback function for using the service within a Javascript environment.

Once developed, the validation service was embedded in the UNT Libraries metadata editing application to provide real-time validation of dates added by metadata creators and editors. The goal of this validation step is to improve the quality of dates as metadata creators become more familiar with the standards.

As someone enters text into a date field, the date validation service checks the string. Valid date fields are highlighted in green, while invalid date fields are highlighted in red. An invalid date also cues a reminder notice for metadata enterers to use EDTF formats (see Figure 1).

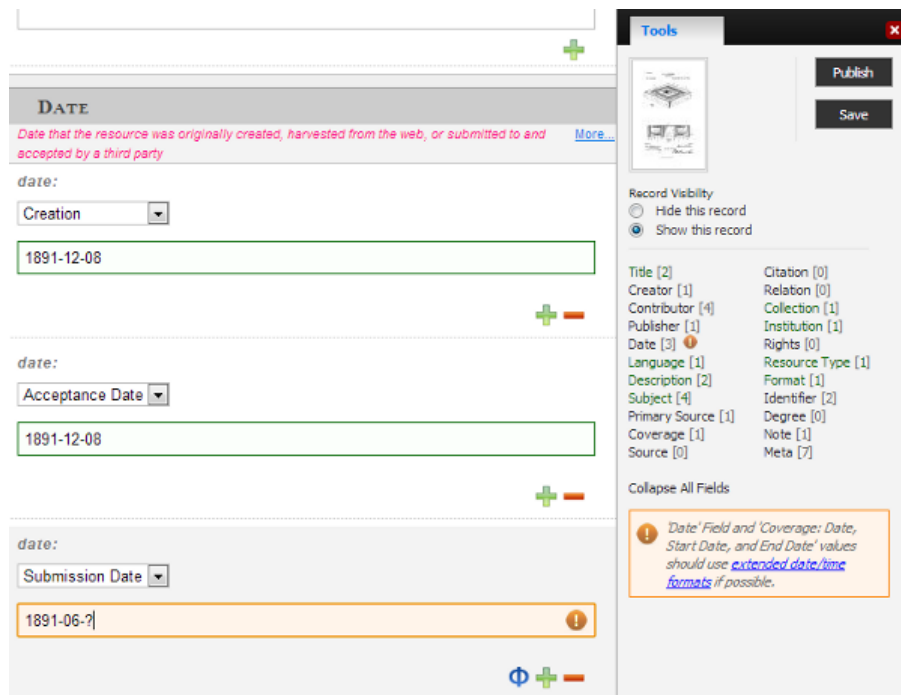


FIG. 1. Metadata record containing correct dates (top), a non-valid date (bottom), and the reminder notice (right).

5.2. User Displays

Since the EDTF is a machine-readable format, the notation used may not be immediately understood by many users. Even the basic form, year-month-day, could be misinterpreted since 2013-03-07 could represent “March 7, 2013” or “July 3, 2013” depending on the user’s social context and familiarity with date-time formats. Additionally, some users might misinterpret partial dates (e.g., 1897-04) as incomplete, even though the partial date might represent all of the date information that is available or relevant to a particular item. Seasons, approximate dates, unknown digits, and date ranges could be even more confusing (see Table 10).

TABLE 10: Examples of common EDTF notations that could be confusing to users.

Kind of Date	EDTF Notation	Example
Season	21, 22, 23, or 24 in place of a month	1968-24
Approximate (“circa”)	~ at the end of the date	1895~
Unknown digits	“u” to replace unknown digits	1923-uu-30
Interval (date range)	/ between the start and end dates	1906-08/1910-12

To mitigate this problem, the UNT Libraries have initiated some changes to the user display that translate EDTF date strings into fully-written dates. For instance, 1937-08-23 displays as August 23, 1937 in the user interface (see Figure 2).

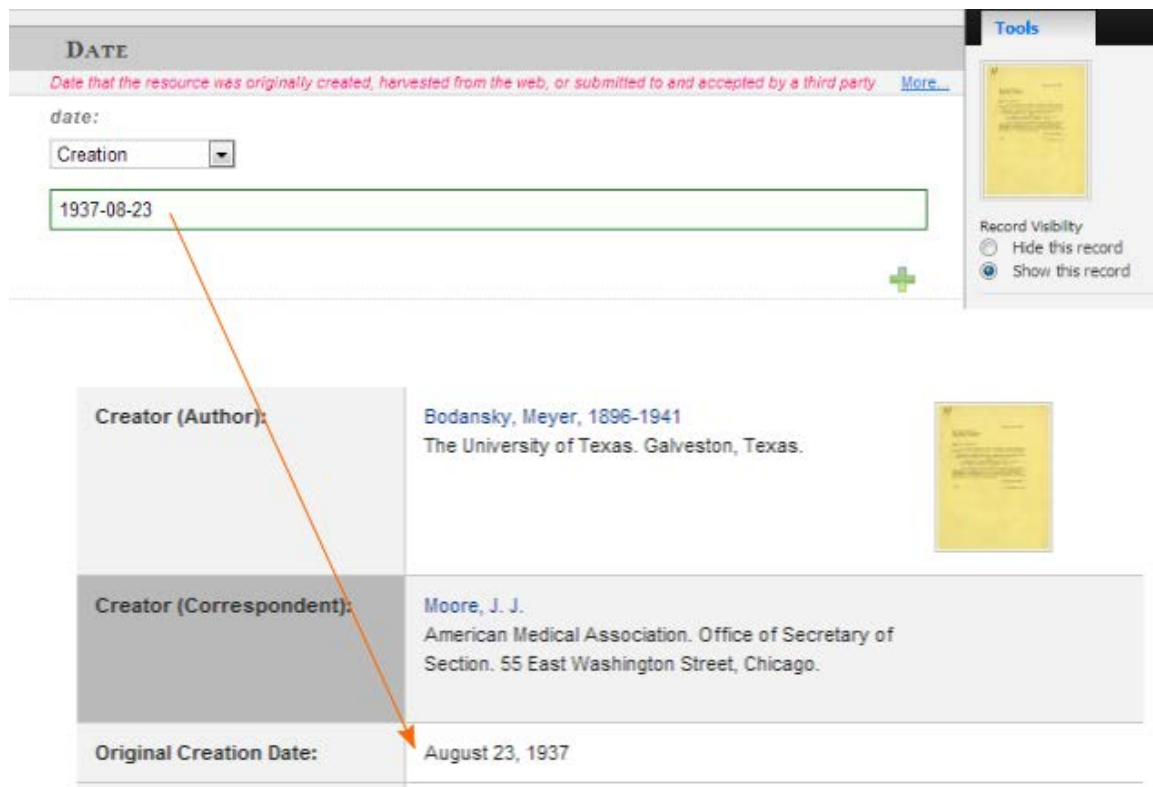


FIG. 2. Date entered according to EDTF (top) displayed as written words in the user interface (bottom).

Level 0 dates and some Level 1 dates are relatively easy to represent for public users (see Table 11), although more complex dates pose some challenges.

TABLE 11: Examples of date notations with user display equivalencies.

Kind of Date	EDTF Example	Display
Year and month	1854-02	February 1854
Year, month, and day	2013-03-07	March 7, 2013
Season	1968-24	Winter 1968
Approximate ("circa")	1895~	Approximately 1895
Interval (date range)	1906-08/1910-12	August 1906 through December 1910
Decade	193u	1930s

6. Challenges and Considerations

As with any new standard, using the EDTF in UNTL metadata includes some challenges in implementation. For example, many of the partners who enter metadata in the digital collections copy information from traditional library catalogs or institutional databases, which can result in dates that have square brackets, periods, or some of the other problems noted previously. Institutions implementing the EDTF may find similar discrepancies between common practice in other fields of study and the notation used to function in machine-readable ways.

There are also some challenges in representing complex dates for public users. In some cases, determining an appropriate user display equivalence is complicated enough to create problems. For example, some letters and documents have only partially-visible dates due to age or damage. If the year and day are known, but not the month, it is easy to represent in EDTF (e.g., 1892-uu-07), however, there is a hole at the start of the date when attempting to write it out in the same format as other user displays (e.g., ____ 7, 1892). Although the missing information can be represented in various ways (see Table 12), finding a reasonable alternative that works in multiple scenarios and makes sense to users has proven difficult.

TABLE 12: Examples of date notations without obvious user display equivalencies.

Kind of Date	EDTF Example	Some Possible Display Options
Partial date missing the month	1892-uu-07	____ 7, 1892 ? 7, 1892 [month unknown] 7, 1892
Uncertain date	1890-10-05?	October 5, 1890? Possibly October 5, 1890 October 5, 1890 (not verified) October 5, 1890 (uncertain)
Date range with an uncertain start or end date	1867-02-04/unknown	February 4, 1867 through Date Unknown February 4, 1867 through ? February 4, 1867 through [unknown]
Partial date missing a year	197u-12-08	December 8, 197? December 8, 1970s

To date, the UNT Libraries have not finalized a solution for displaying many of these dates. However, when dates do not have an equivalent user display, an icon next to the date links users to a glossary of date terms to help them interpret the date that they are looking at (see Figure 3).



FIG. 3. Example of the icon displayed next to dates in the user interface.

Metadata creators unfamiliar with the EDTF and similar standards may also struggle initially with interpreting some of the guidelines. For example, the EDTF documentation does not explicitly state that “approximate” dates cannot also have a level of uncertainty (e.g., 189u~ for “circa 1890s”). However, the coded components in the EDTF that are used for validation can clarify some of the logical rules for applying various date types.

7. Conclusions

Despite the inherent challenges in modifying existing guidelines to new standards, the shift to EDTF has been primarily positive for the UNT Libraries. The EDTF contains standard formats for handling essentially all of the date requirements for items in the UNT digital collections; similarly, other cultural heritage institutions may find that the flexibility of the EDTF meets date needs that other date standards do not address. As the EDTF community grows, additional features may be added in the future and provide options for even more specialized needs.

Bringing non-compliant dates up to standards will be an ongoing process for the UNT Libraries; however, a majority of the dates have consistent formats that can be changed mechanically. The analysis of dates in the UNT digital collections suggest that dates in other institutions that have used ISO 8601 may also be mostly-compliant with EDTF if they are considering a move to the EDTF as a primary date standard.

Standardizing complex dates may help to promote metadata interoperability and provide more complete and consistent information to users. Based on the experiences of the UNT Libraries and the continuing efforts to provide useful metadata to digital library users, the EDTF appears to be a useful step for representing date information.

References

- Digital Public Library of America. (February 8, 2013). Metadata Application Profile, Version 3. Retrieved April 3, 2013 from <http://dp.la/about/map/>.
- International Standards Organization. (December 1, 2004). ISO 8601: Data elements and interchange formats – Information interchange – Representation of dates and times. Third edition.
- Library of Congress. (January 13, 2012). Extended Date/Time Format (EDTF) 1.0. Retrieved from <http://www.loc.gov/standards/datetime/>.
- Phillips, Mark Edward. (February 28, 2013). Extended Date/Time Format (EDTF) Dates Research Datasets. Retrieved from <http://digital.library.unt.edu/ark:/67531/metadc172363>.
- Phillips, Mark Edward. (2013). Metadata Analysis at the Command-Line. *Code4Lib Journal*. Retrieved from <http://journal.code4lib.org/articles/7818>.
- Tarver, Hannah. (2013). [Handout to Accompany Poster on Implementation of a New Date/Time Standard]. Retrieved from <http://digital.library.unt.edu/ark:/67531/metadc159536>.
- University of North Texas Libraries. (2013). Date. *Input Guidelines for Descriptive Metadata*. <http://www.library.unt.edu/digital-projects-unit/date>.
- University of North Texas Libraries. (2013). Extended Date Time Format Levels 0, 1 and 2 Validation Service. <http://digital2.library.unt.edu/edtf/>.
- University of North Texas Libraries. (2013). Input Guidelines for Descriptive Metadata. <http://www.library.unt.edu/digital-projects-unit/input-guidelines-descriptive-metadata>.

Appendix

Examples of EDTF formatting. Taken from *EDTF Cheatsheet* (Tarver, 2013).

Commonly Used Dates

	Example Item	Format	Example Date
Year	Book with publication year	YYYY	1902
Month	Monthly journal issue	YYYY-MM	1893-05
Day	Letter	YYYY-MM-DD	1924-03-03
Time	Born-digital photo	YYYY-MM-DDTHH:MM:SS	2003-12-27T11:09:08
Interval	Compiled court documents	YYYY/YYYY*	1887/1889
Season	Seasonal magazine issue	YYYY-SS	1957-23
Decade	WWII poster	YYYu	194u
Approximate	Map "circa 1886"	YYYY~*	1886~

* "Year" is used as an example, but each date could have month or day precision as well.

Some Complex Dates

Example Item	Kind of Date	Format	Example Date
Photo taken at some point during an event August 6-9, 1992	One of a Set	[YYYY..YYYY]*	[1992-08-06..1992-08-09]
Hand-carved object, "circa 1870s"	Extended Interval (L1)	YYYY~/YYYY~*	1870~/1879~
Envelope with a partially-legible postmark	Unspecified	"u" in place of digit(s)	18uu-08-1u
Map possibly created in 1607 or 1630	One of a Set, Uncertain	[YYYY, YYYY]*	[1670?, 1630?]

* "Year" is used as an example, but each date could have month or day precision as well.