

## DATE NORMALIZATION UTILITY (DNU) DOCUMENTATION

Document owners: David Loy, Bill Landis  
Last updated: 2005-08-24

This document references a number of decisions made during several planning-stage meetings for date normalization that were held in the context on the OAI Harvest Capacity Development Project. Notes from those meetings are available internally at CDL at:

- [https://diva.cdlib.org/projects/american\\_west/harvest/harvest\\_core/planning\\_docs/datenorm\\_notes\\_20050309.doc](https://diva.cdlib.org/projects/american_west/harvest/harvest_core/planning_docs/datenorm_notes_20050309.doc)

### *Coding this DNU as a Component of the CDL Common Framework (CF)*

David Loy, beginning in March 2005, coded this DNU in Java as a component of CDL's CF technical infrastructure, specifically as a utility available as part of the CF's *Ingest* service.

This DNU utilizes a Document Object Model approach and is integrated at the top level with the CDL CF for the purposes of logging. This would need to be decoupled in order to use this DNU at an institution not using a CDL CF repository, and David has some good ideas about how to accomplish this.

### *Metadata Formats Normalized*

This DNU currently normalizes dates extracted from selected elements in the following metadata schemas:

- Simple Dublin Core (DC), OAI-PMH schema (oai\_dc)
  - <date> element
  - <title> element
  - <description> element

In the OAI ingest system in use at CDL, MODS OAI records are mapped into DC as part of the ingest process, and the DNU is run after this process is completed, so in the CDL environment this utility also normalizes dates in MODS records. This process of mapping to DC during ingest will facilitate this DNU handling other OAI metadata schemas as well.

### *Quality Assurance Testing of the Finished Product*

At this point the performance of this DNU has not been tested beyond that done by David Loy during programming. The OAI Harvest Core group will develop and implement a more formal QA testing plan for this utility late in 2005. Our success metrics should be to have some significant percent of the dates in the American West Project's prototype collection (harvested in July 2005) normalized appropriately (e.g., the normalized date(s) reflect the date(s) of creation of the original object) -- we are not aiming even remotely for 100% normalization.

### *Extraction and Normalization of Dates*

## Background decisions informing development of this DNU:

1. Focus is on extracting and normalizing dates relating to the creation of the content of a digital object or resource.
  - a. The attempt will be made, through the algorithm, to remove administrative dates relating to the digitization of non-digital resources, which appear in a significant number of the OAI-PMH records that the American West Project is harvesting and are not helpful to end users seeking dates relating to the creation of the original content.
2. Normalize only to 4-digit years (YYYY), both for single dates (e.g., 1956) and date ranges (e.g., 1868-1879).
  - a. Normalization of months and days is not needed for indexing for searches in the American West Project portal, which is driving the initial development of this particular DNU.
3. Collapse lists of contiguous dates that appear in a single date element into a date range (e.g., 1936, 1937, 1938, 1939, 1940, 1941 will be normalized as 1936-1941)
4. Dates in actual date elements will be given preference, meaning we will not look in other elements for dates if there is a date in a date element in a record.
  - a. This DNU looks for the following indicators of the likelihood of a date in addition to the 4-digit year pattern noted above:
    - i. Standard AACR2 date representations (see 1.4F in AACR2).
    - ii. Standard variations of month, day, year, with months in English, either spelled out or abbreviated.
    - iii. Between YYYY and YYYY (normalized as a date range)
  - b. In cases where no date or an indication of an unknown date are found in a date field, this DNU will attempt to find and extract dates, based on patterns indicated in 4.a, from the title and description fields. No other DC fields will be explored for the purpose of "guessing" dates.
  - c. Process for all records:
    - i. Look in oai\_dc record for date element(s); if found, extract and normalize date(s).
    - ii. If no date element(s) found (or if indication of unknown is found in a date element), look in title element for year-like strings of numbers; if found, extract and normalize date(s).
    - iii. If no title element(s) found, look in description element for year-like strings of numbers; if found, extract and normalize date(s).
    - iv. If no description element(s) found, STOP.
    - v. Proceed with normalization using step 9.
5. Dealing with multiple date elements and likely date(s) of digitization.
  - a. Many OAI-PMH records contain a digitization date in a generic date element, usually, though not always, in addition to at least one other date element. As a heuristic for this DNU, we decided to use 1995 as the date prior to which we think most dates would not reflect date of digitization of the item.
  - b. Process for records containing 1 or more date elements:

- i. If a record contains a single date element, extract and normalize date(s).
    - ii. If a record contains two or more date elements, extract dates; if all are earlier than 1995, normalize them all; if one or more are 1995 or later, do not normalize the most recent single date (the date we feel most likely represents the digitization date for the item).
    - iii. If a record contains no date element, or if it contains only a single date element with some version of an expression of an unknown date, normalize all dates extracted according to the process documented in 4.c.ii-v.
- 6. Dealing with date fields containing an indicator of an unknown date.
  - a. If a date field exists in a record, but contains an indicator of an unknown date, follow the process outlined in 4.c.ii-v.
  - b. This DNU recognizes and normalizes the following indicators of unknown dates:
    - i. unknown, date unknown, unkn
    - ii. unavailable, unavail, unav
    - iii. not determined
    - iv. n.d., nd
    - v. s.d., sd
    - vi. date not indicated
    - vii. no date
  - c. If no normalizable date(s) (see 4.a.i-iii) is found, stop and treat the date as "Unknown" with a TEMPER indication of ":unav". See *Appendix A* for an indication of how this would look in the AIP <dmdSec><sup>1</sup>.
- 7. Dealing with textual expressions of dates and date ranges.
  - a. This DNU recognizes and normalizes the following textual expressions of dates and date ranges:
    - i. 17th century/cent. = 1600-1699 (same pattern for later centuries)
    - ii. Specifically does not recognize various shadings of centuries, such as "early," "mid," or "late". In these cases this DNU will extract and normalize only the textual expression of the century.
    - iii. Also does not recognize centuries spelled out (e.g. "seventeenth" instead of "17th").
- 8. Dealing with indications of uncertainty in dates expressed with AACR2-like symbols.
  - a. All indications of uncertainty in dates expressed using AACR2-like symbols ? and - will be normalized as follows:
    - i. Ignore square brackets in uncertain date expressions (e.g., treat 192[?] as 192?).

---

<sup>1</sup> Regarding the AIP <dmdSec> reference, in the CDL CF repository in use at CDL normalized dates are stored in a CF SQL table (internal CDL architecture). The actual output of this DNU itself is an array of key-value (see the section *DNU Output* for a list of possible keys). A programmer can redirect results post process to whatever format is required. In the CDL CF repository, AIP is dynamic and can be provided as a variety of output formats as needed.

- ii. Uncertainty about decades (e.g., 192-, 192?, 192-?) will be expressed as the equivalent decade date range (e.g., 1920-1929).
  - iii. Uncertainty about centuries (e.g., 18--, 18??, 18--?) will be expressed as the equivalent century date range (e.g., 1800-1899).
9. Dealing with indications of uncertainty in dates expressed with date qualifiers.
- a. All indications of uncertainty in dates expressed with date qualifiers extracted from OAI-PMH records will be normalized to +/- 5 years (e.g., "circa 1942" is normalized to the date range 1937-1947).
    - i. The one exception to the above is when "circa" or "ca." appears before a date that includes a month and/or day. In this case the date will be normalized to just the year indicated, in keeping with the narrower scope of the indication of uncertainty.
  - b. Ignore square brackets in uncertain date expressions (e.g., treat [ca. 1922] as circa 1922).
  - c. When a question mark appears after a date (e.g., 1875?), it is also normalized to +/- 5 years.
  - d. In date ranges, indications of uncertainty at either end of the range are treated as an indication of the uncertainty of the entire range (e.g., "ca. 1890-1902" is normalized as 1885-1907)
  - e. This DNU recognizes and normalizes the following indicators of uncertainty associated with dates and date ranges:
    - i. circa, ca. (c. is specifically treated in its AACR2 sense, as an indication of a copyright date, and not normalized by this DNU).
    - ii. approximately, approx.
10. Recording normalized date when no date field or year-like strings are found in a record.
- a. If, using the process outlined in 4.c.i-v, no date fields or year-like strings in a title or description field are found, treat the date as "Unknown" with a TEMPER indication of ":none". See *Appendix A* for an indication of how this would look in the AIP <dmdSec>.

### *DNU Output*

This DNU outputs up to 6 different dates in an array of key-value pairs:

<b>key</b>	<b>Description of key-value pair</b>
<b>date.found</b> --OR-- <b>date.guess</b>  These fields are mutually exclusive and should never appear together in the same record.	The date(s) selected for normalization by the DNU from the as-harvested metadata. If the date(s) selected for normalization are from a DC date field (or its equivalent in another OAI schema), the date(s) are output with the key <b>date.found</b> ; if they are from a title or description field, they are output with the key <b>date.guess</b> .

<b>date.temper</b>	The DNU first creates a TEMPER date and outputs it using the key <b>date.temper</b> . This version records a normalized date that is easily sortable, and standardizes the expression of uncertainty. See <i>Appendix B</i> for a description of how TEMPER is being used by the DNU.
<b>date.normalize</b>	This normalizes the TEMPER version of the date with any uncertainties (~ in TEMPER) expanded to +/- 5 years.
<b>date.decade</b>	A token for the decade(s) in which the <b>date.normalize</b> value falls. Specifically designed for use with the American West project at CDL. See <i>Appendix C</i> for more information.
<b>date.era</b>	A token (1-4) mapped to 4 specific eras of use to the CDL American West project. See <i>Appendix C</i> for more information.
<b>date.tokens</b>	A token for indexing representing each year indicated in the <b>date.normalize</b> value.

## Appendix A: Indicating Unknown Dates in DNU Output

In cases where there was no discernable date to normalize *anywhere* in the original OAI-harvested record, the DNU output would look like this (there would be no date.found or date.guess):

```
date.temper=(:none) Unknown
date.normalize=(:none) Unknown
date.era=(:none) Unknown
date.decade=(:none) Unknown
```

In cases where there was a discernable indicator of an unknown date in the <date> field, but the normalization algorithm was unsuccessful in finding an alternate year-like string in a title or description field to normalize, the DNU output would look like this:

```
date.found=(:unav) No date
date.temper=(:unav) Unknown
date.normalize=(:unav) Unknown
date.era=(:unav) Unknown
date.decade=(:unav) Unknown
```

The indicators (:none) and (:unav) are used at the recommendation of John Kunze and are drawn from the "Kernal Glossary of Elements and Values" of the Draft Kernal Metadata Specification (August 2004)<sup>2</sup>. (:unav) indicates that we are basing our status of an "unknown" date on the best date field metadata available to us from the original source. (:none) indicates that we are basing our status of an "unknown" date on the fact that the original record had no date field value and never will.

---

<sup>2</sup> Available at <<http://www.jiscmail.ac.uk/files/DC-KERNEL/Aspec.html>>.

[https://diva.cdlib.org/projects/american\\_west/harvest/harvest\\_core/planning\\_docs/datenorm\\_documentation.doc](https://diva.cdlib.org/projects/american_west/harvest/harvest_core/planning_docs/datenorm_documentation.doc)

## Appendix B: Use of TEMPER<sup>3</sup> by the CDL DNU

1942 = single date  
1942-1956 = date range  
1942, 1952, 1964 = multiple single dates  
1942-1954, 1962 = multiple single dates and date ranges  
1942~ = uncertain single date  
1942~-1961 *OR* 1942-1961~ = one date in a range uncertain  
1942~-1958~ = both dates in a range uncertain  
-1965 = open-ended beginning of a date range  
1942- = open-ended end of a date range

## Appendix C: American West Era and Decade Categories

The four AmWest eras and their corresponding decades are:

-1800	1751-1760, 1761-1770, 1771-1780, 1781-1790, 1791-1800 + earlier decades if needed
1801-1890	1801-1810, 1811-1820, 1821-1830, 1831-1840, 1841-1850, 1851-1860, 1861-1870, 1871-1880, 1881-1890
1891-1940	1891-1900, 1901-1910, 1911-1920, 1921-1930, 1931-1940
1941-	1941-1950, 1951-1960, 1961-1970, 1971-1980, 1981-1990, 1991-2000, 2001-

---

<sup>3</sup> More information about the draft TEMPER (Temporal Enumerated Ranges) specification is available at <<http://www.ietf.org/internet-drafts/draft-kunze-temper-00.txt>>.