Identifying and Normalizing Date Expressions

Introduction: This project focuses on annotating temporal phrases as a first step on placing dates on a calendar timeline. Therefore, the scope of the temporal expressions identified (referred to as labeling) and normalized are:

(i) those periods that can be placed on a calendar with the shortest measurement of a day,

(ii) lengths of time, referred to as durations, that are 24 hours or longer or refer to a measurement of a day or longer, and

(iii) periodic intervals of 24 hours or longer or which refer to a measurement of a day or longer.

Normalization follows the ISO8601 except for issues not addressed by the format.

For the normalization, it will be done on theL (i) sentence level, meaning that the date will be identified to the precision allowed by the information solely contained in the sentence plus the date the transcript was created and (ii) the document level, meaning that the date will be indenfied by the information in the document (by utilizing text before and after the sentence at issue) and utilzing the dates specified in the initiating document, if given. In addition, anchor dates will be identified along with the normalized date at the document level. Anchor dates are the dates implicitly pointed to by the speaker when using relative date terms such as “today” and “last week” and when referring to earlier points in the discussion by using pronouns and other terms.

Example 1:
Transcript Date: March 3, 2023
I went to a party on July the 4th 2013.
Labeling:
I went to a party on xxxJuly the 4th 2013xxx. Note: prepositions that indicate that something occurred on/during the given date span are not included.
Sentence-Level Normalization:
2013-07-04
Document-Level Normalization: (which includes the anchor date after a triple dash (or NA when no anchor date is involved.)
2017-07-04—NA

Example 2:
Transcript Date: May 5, 2000
Q: Did you go swimming on August 1, 1999?
A: Yes, I do go then.

Labeling:
Did you go swimming on xxxAugust 1, 1999xxx?

Yes, I do go xxxthenxxx.

Sentence Level Normalization
1999-08-01
1999-08-01

Document Level Normalization
1999-08-01—NA
1999-08-01—1999-08-01

Example 3
Transcript Date: 2025-03-05

Q: Let's discuss your activities last year.
Did you go to the store on Friday, March 1st?
A: No, it was the Monday after that.

Labeling:
Q: Let's discuss your activities xxxlast yearxxx.
Did you go to the store on xxxFriday, March 1stxxx?
A: No, it was xxxthe Monday after thatxxx.

Sentence-Level Date Normalization
2024-XX-XX
XXXX-03-01-05 (Note: follows YYYY-MM-DD-DOW(2 dight)
XXXX-XX-XX-01

Document-level Date Normalization
2024-XX-XX—NA
2024-03-01-05—2024-XX-XX
2024-03-04-01—2024-03-01

An anchor date can also be the creation date of the document being referenced. Below is an example.

Example 4
Q: I will read you an email sent on April 5, 1999.
My personal belief is that Enron stock is an incredible bargain at current prices, and we will look back a couple of years from now and see the great opportunity that we currently have.
And that's the latter part of the quote where it mentions that you had been buying stock in the last couple of months.
Did you do that?

A: I bought stock in February and March of 1995.

Labeling:

Q: I will read you an email sent on xxxApril 5, 1999xxx.
My personal belief is that Enron stock is an incredible bargain at current prices, and we will look back xxxa couple of years from nowxxx and see the great opportunity that we currently have.
And that's the latter part of the quote where it mentions that you had been buying stock in xxxthe last couple of monthsxxx.
Did you do that?

A: I bought stock in xxxFebruary and March of 1995xxx. (Note that two dates are included in one span because they are nested.)

Sentence Level Date Normalization:
1999-04-05
XXXX-XX-XX/+PXY (Note: since it is unclear how many months an "X" is used and the "/" plus "+PXY" specifies months after the date specified before the slash. A further explanation of this is below.)
XXXX-XX-XX/-PXM
1995-03,,,1995-02

Document Level Date Normalization:
1999-04-05—NA
1999-04-05/+PXY Note: since it is unclear how many months an "X" is used and the "/" plus "+PXY" specifies months after the date specified before the slash. A further explanation of this is below.
1999-04-05-PXM—1994-04-05
1995-03—NA,,,1995-02—NA

Normalization Simplification
For the present project, date modifiers such as before, after, or since will not be considered in normalizing the date, but are included in labeling.

Example 5:
Transcript Date: January 1, 2024
Q: When did you have the big turkey dinner made by Mary?

A: It was some time before Christmas.

Labeling:
It was xxxsome time before Christmasxxx. Note: that some time was included in the date expression phrase because it indicates estimation.

Sentence Level Normalization
XXXX-12-25

Document Level Normalization
XXXX-12-25—NA Note: It is unclear to what year the speaker is referring to given the text available.

ISO 8601 Formatting for Date Normalization

I. Calendar Dates

A. Basic Points in Time

The formats are as follows. Exactly the components shown here must be present, with exactly this punctuation.

Year:

YYYY (eg 1997)
Year and month:
YYYY-MM (eg 1997-07)
Complete date:
YYYY-MM-DD (eg 1997-07-16)
where:
YYYY = four-digit year
MM = two-digit month (01=January, etc.)
DD = two-digit day of month (01 through 31)

Days of the Week: (only list the day of the week if it is referenced in the sentence)

Monday	01 (eg 1997-07-16-01)
Tuesday	02
Wednesday	03
Thursday	04
Friday	05
Saturday	06
Sunday	07

Weeks of the Year
Week 01, meaning the first week of the year, is the first week that has the majority of its days in the new year. Week 01 might also contain days from the previous year and the week before week 01 of a year is the last week (52 or 53) of the previous year even if it contains days from the new year. Weeks start with Monday (day 1) and end on Sunday (day 7).
Example:

The first week of 2025:

2025-W01

The week notation can include the day of the week by addin at the end the number for the day. For example, the day 2024-12-31, which is the Tuesday (day 2) of the first week of 2024, can also be written as

2024-W01-2

B. Points of Time Utilizing Durations

Points of Time utilizing durations are time periods that refer to a starting or ending time plus a block of time, also referred to as duration.

Pattern:

Use the point of time pattern specified above plus the below

/

\+ time period after the point in time

\- time period prior to the point in time

Example:

I swam every day for two months starting on the first of July 2024.

Labeling:

I swam xxxevery dayxxx for yyytwo months starting on the first of July 2024yyy. Note: that more than one clause is found, then the

clauses are labeled consecutively starting with x and then going through the alphabet as follows: x, y, z, a, b . . .x

Sentence-Level Date Normalization

R1D,,,2024-07-01/+P2

Document-Level Date Normalization

R1D—NA,,,2024-07-01/+P2M—NA

Fiscal Years should be written as “FY” plus the four-digit year.

Example: FY2009

Quarters/Seasons

Values for quarters and seasons are below.

21-24 = Spring, Summer, Autumn, Winter

33-36 = Quarter 1, Quarter 2, Quarter 3, Quarter 4 (3 months each)

Example:

Transcript Date: January 1, 2025

I will buy Easter candy in the spring and Christmas chocolates this winter.

Labeling: I will buy Easter candy in xxxthe springxxx and Christmas chocolates xxxthis winterxxx.

Sentence-Level Date Normalization:

2025-21,,,2025-25

Document Level Date Normalization:

2025-21—2025-01-01,,, 2025-25-2025-0101

Decades

Decades are denoted by putting an X for the last digit in the year.

Example:

He worked at the university in the 60s.

Labeling:

He worked at the university in xxxthe 1960sxxx.

Sentence-Level Normalization:

196X

II. Durations and Periodic Time Periods

A. **Date Durations**. Durations are lengths of time that are 24 hours or greater or use as a measurement a day or greater.

Examples: I had the measles for xxxtwo weeksxxx.

We got out of school for xxxa half a dayxxx.

They are represented by starting with a “P” and then adding the number and then a measurement.

Example:	One year	P1Y
	One month	P1M
	Two Days	P2D

The duration part (e.g., P1Y2M10D, meaning the period of one year, two months, and 10 days) follows this structure:

- **P:** Indicates the start of a duration component.
- **Y:** Represents years.
- **M:** Represents months.

- **D:** Represents days.

Ages

A person's age is a duration.

Example:

She was seven year's old.

Labeling:

She was xxxseven year's oldxxx.

Sentence Level Normalization:

P7Y

B. Periodic Time Periods.

Periodic time periods of 24 hours or greater or periodic time periods using a day or greater should be measured. The format periodic time period such as R1Y (meaning annually) follows this structure:

R: Indicates the start of a periodic time period

Y: Represents years.

M: Represents months.

D: Represents days.

C. **Time Periods with start and end dates**

The start and end dates should be connected by a slash.

Example:

Q: When in 2024 did you move the boxes?

A: We moved boxes starting in July and ending in August.

Labeling:

We moved boxes xxxstarting in July and ending in Augustxxx.

Sentence Level Normalization

2024

XXXX-07/XXXX-08

Document Level Normalization:

2024-07—2024/2024-8—2024 Note: 2024 is included because this is the anchor date.

III. Date Span Labeling Rules

1. Words of Estimation, Uncertainty and Approximation should be included with the span it modifies.

Example: It was about June of last year that it happened.

Labeling:

It was xxxabout June of last yearxxx that it happened.

Example: I think it was in January.

Labeling:

I xxxthink it was in Januaryxxx.

Example: It was probably June 2021.

Labeling:

It was xxxprobably June 2021xxx.

2. When more than one date is given as an approximation, they should be included in one phrase.

Example:

It was either on July 12th or July 14th.

Labeling:

It was either on xxxJuly12th or July 14thxxx.

3. When the speaker corrects themselves, the incorrect and correct date should be included in one phrase.

Example:

It was November of 1994, no I mean, July of that year.

Labeling:

It was xxxNovember of 1994 no I mean July of that yearxxx.

4. Time periods with beginning and ending dates should all be included in one span.

Example:

I walked my dog from July 1 to August 2nd.

Labeling:

I walked my dog xxxfrom July 1 to August 2ndxxx.

5. Nested dates should be included in one phrase.

Example:

He had parties in the first quarter of FY2025 and FY2025.

Labeling:

He had parties in xxxthe first quarter of FY2025 and FY2026xxx.

6. Superfluous langage should be included when it is sandwiched between words needed to form a full date expression.

Example: Thus, between July the 13th and–excuse me, June the 13th and July the 3rd, I was on vacation.

Labeling:

Thus, xxxbetween July the 13th and–excuse me, June the 13th and July the 3rdxxx, I was on vacation.

Sentence-Level Normalization:

XXXX-07-13/XXXX-07-03

XXXX-06-13/XXXX-07-03 (Note: two date ranges are given because two were communicated within the span.)

Working with the Spreadsheet

1. Sentence Column.

A. Date expressions should be labeled if the Data_Split column has text in it. (The text will either be “TST” or “TR”). The labeling of the sentences should be done in the “Sentence” column.

B. If no date expressions are found in the Sentence, go to the next sentence where the Data_Split column has “TST” or “TR”.

B. The first date expression should be marked with xxx immediately before the start of the word which begins the expression and xxx immediately after the last word in the phrase. Each additional date expression found should be marked by three of the next consecutive letter at the beginning and end of the phrase. The order then being: x, y, z, a, . . .

C. Include within a phrase modifiers relating to uncertainty, estimation, and approximation.

D. Include within a phrase date and duration modifiers such as “more than” and“before”.

E. Include articles such as “the” and “a” if they modify words within the date expression.

2. Sentence Date Normalization Column. For the date expressions labeled in the Sentence column, list any dates within these expressions that can be put in the ISO format in the order they are found in the sentence using the information in the sentence and, if necessary, the transcript date, which is in the Transcript_Date column. Put three comas with no spaces between each iso-formatted date provided.

3. Document_Date_Normalization_Anchor_Date Column. List all ISO Dates in the order they are found in the sentence, but use all information in the transcript (and in any other information provided such as if the original pleading is included) to determine the date as best one can. After the ISO Date, add three dashes and then put the Anchor Date (in the ISO Date format), if no Anchor Date is applicable put NA after the three dashes. Separate each ISO Date Anchor Date pair by three commas. The Anchor Date is a date that the sentence is implicitly pointing to such as the transcript creation date, a date earlier referenced in the transcript, or a date commonly spoken of such as the date of the alleged crime, which the participants are expected to know.