DOI: http://dx.doi.org/10.14236/ewic/EVA2017.13

Understanding Textual Uncertainty in Dates Using Interactive Timelines

Sam Cottrell
Royal College of Art
London SW7 2EU, UK
sam.cottrell@network.rca.ac.uk

This paper explores what is understood by common uncertainty terms (e.g., circa, approximately, around) and how the associated date information affects the perceived range of a given date. This is accomplished through a study incorporating an interactive timeline. Participants are presented with a date and associated uncertainty term, and instructed to mark the possible and probable extents. The results from the study have been used to produce guidelines on the capture of dates, in order to include textual uncertainty and represent them visually.

Uncertainty. Crowdsourcing. Timelines. Collections.

1. INTRODUCTION

Uncertainty is a property of data that refers to imperfect or unknown information. Epistemic temporal uncertainty is present in varying degrees in all practical time based data and adds a layer of complexity to visualisation. While this type of uncertainty cannot be controlled, it can be described (Bonneau et. al. 2014). Digital processes often do not accommodate the presence of uncertainty and as such, it is frequently ignored when collections data are captured, processed or represented graphically (Kräutli & Boyd Davis 2013), disregarding valuable information and allowing potential insights to go unrecognised, or incorrect inferences to be made.

Cultural institutions are capturing increasingly large amounts of metadata regarding their collections, enriching and converting existing records, often making them available to the public through online catalogues and APIs. The records in these collections frequently include uncertainty terms connected with the dates.

To visually represent dates that include associated uncertainty terms visually, it must be determined what is most commonly intended when they are used, and what is understood when they are encountered.

2. INFLUENCES ON PERCEPTION

It is expected that the following factors influence the perceived range of the date described using uncertainty terms:

- The resolution (and the perceived resolution) of the date;
- The temporal distance of the provided date from the present;
- The term used to describe the uncertainty.

The resolution of the date connected with the uncertainty term will inform much of the perceived uncertainty as it will be related to the provided level of information, e.g., when presented with “circa. 1951”, it is assumed that the source is unsure of the year in which year it lies.

The resolution of information can, however, be ambiguous, or misinterpreted, e.g., “circa 1900” whilst normally associated with the year, may be interpreted as the decade or even the century beginning at that year. This is further complicated by digital processes as the constraints of date formatting may mean that a false resolution of information has been recorded, e.g., when a date entry field must include day and month level information then they are often input as the 1st of January.

Due to the improvement of data recording technologies over time, and the nature of human and institutionalised memory (Basu & Waymire

2006), the further from the present date, the less accurate dates appear, and so greater perceived uncertainty is expected.

By changing and recording the resolution, temporal distance and term variables as part of the study it should be possible to determine to what extent they affect the perceived possible range of the date.

Since different terms associated with dates convey varying amounts of certainty, to fully transfer textual date information into a visual medium the nature of the terms used must be understood. Uncertainty terms are by their nature and intent ambiguous. They are also subject to the interpretation of the parties involved (both audience and the author).

The terms to be investigated have been compiled through investigation and experience. Kräutli (2016) identified 12 ambiguous terms used in date descriptions in the Victoria & Albert museum collections data. The most common terms, and the ones that will be used in the study are: *circa*, *about*, *approximately*, *around*, *perhaps*, *possibly* and *probably*.

Table 1: *Online collection records containing "Circa"*

Institute	Total Records	Containing "Circa"	%
State Hermitage Museum	*4,424*	*2,998*	*67.8*
Los Angeles County Museum of Art	*55,614*	*25,430*	*45.7*
Fitzwilliam Museum	*205,069*	*36,828*	*18.0*
Courtauld Institute of Art	*43,228*	*2,925*	*6.8*
National Trust Collections	*898,083*	*50,873*	*5.7*
Bristol Museums, Galleries and Archives	*198,374*	*6,531*	*3.3*
Royal Museums Greenwich	*190,503*	*6,037*	*3.2*
Royal Collection Trust	*257,083*	*6,807*	*2.6*
Horniman Museum & Gardens	*37,668*	*954*	*2.5*
London Transport Museum	*5,980*	*52*	*0.9*
Wellcome Library	*1,086,216*	*6,716*	*0.6*

The term *circa*, a date specific term, is by far the most common, Table 1 indicates its presence in an assortment of online collections.

This was compiled by simply searching for the term "circa" in each online catalogue, as such it does not include the common abbreviation *"c.",* as many searches would disregard the punctuation and return any record containing the letter *"c".*

3. STUDY

In addition to the core of the study, further information about the participant including gender, age and occupation is captured. Once the user has completed this information as part of their profile and read the directions, they are presented with a date, associated description of the uncertainty and instructed to input first the "Most Probable" and then the "Maximum Possible" extents for the given date and term (Figure 2).

The terms "Most Probable" and "Maximum Possible" introduce additional ambiguous terms. However, omitting this and having a single instruction per task of (i.e., "Input the date range for ***x*** y" where ***x*** is the uncertainty term and y is the provided date) removes assumptions made by the participant that would be present with a lack of direction, and by capturing what the user considers to be both the maximum and probable extents more complete conclusions may be drawn.

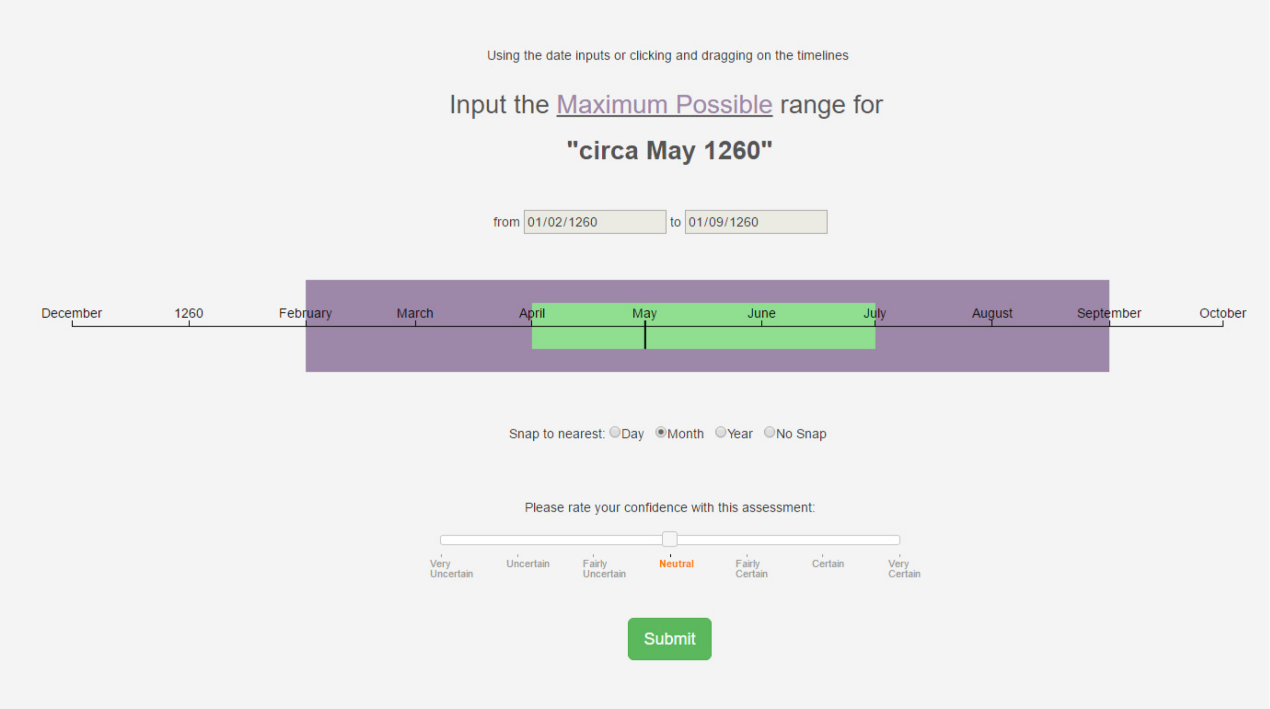
The variables identified in the previous section are randomly selected on page load. A random date between 900AD and 2200AD is rounded to the generated date resolution and presented to the user along with the uncertainty term in an accessible format relevant to the resolution:

- Day – DD^{nth} MMM YYYY
- Month – MMM YYYY
- Year – YYYY
- Decade – YYYYs
- Century – CC^{nth} century

The scale at which the timeline is presented is dependent on the resolution of date randomly selected. The user may then manually adjust the zoom to suit their preference.

It was suggested that the zoom level may have influenced the dates input by the participant. The same starting scale regardless of resolution of data was considered, however this was found to irritate users, as they would quite often have to zoom for some time to present the timeline at an appropriate scale. This could have been improved upon by slightly randomising the starting scale, whilst still having it remain appropriate for the resolution provided.

The timeline is centred upon the earliest time for the generated date, as this is how dates converted to digital are most often captured and displayed. It was considered that perhaps it should be placed at the midpoint, but this could bias the results.



Using the date inputs or clicking and dragging on the timelines

Input the Maximum Possible range for

"circa May 1260"

from 01/02/1260 to 01/09/1260

Snap to nearest: ○Day ●Month ○Year ○No Snap

Please rate your confidence with this assessment:

Figure 1: *Screenshot of the application after the "Most Probable" range (indicated by the green box) and the "Maximum Possible" range has been selected and submitted (as indicated by the purple box)*

The participant enters the date ranges for the given task using the standard HTML date entry boxes, or by interacting with the timeline using their mouse. Entering a valid date via the date entry boxes dynamically updates the timeline and vice versa. The number of times the date is updated and the method of changing it is captured.

By using the "snap to" radio buttons the participant can make the values input constrained to the nearest day, month, or year. It also modifies any already input dates not yet submitted to fit the constraint.

Users also have the option to input their confidence with their assessment using a 7 scale slider to indicate.

The date range selection task was estimated to take around a minute to complete and participants were encouraged to complete it multiple times.

The study was made available through the web (http://study.samctrl.com) and circulated to the public through websites such as the "sample size" subreddit to and professionals through mailing groups and internal emails at the National Archives.

4. RESULTS

By compiling and analysing the results we can gain an understanding of how different factors affect the perceived range. The study gauges the effect on participants' responses with respect to the term used to describe the uncertainty, the temporal distance of the provided date from the present and the resolution of the date.

150 responses were recorded from 35 participants across Europe and the United States and from a variety of backgrounds. Including students, data professionals and archivists.

Both means of entering dates were used roughly equally, with some participants choosing to use a mixture of text input and clicking the timeline during a single-entry task.

Figure 2 shows the data captured in the study. The date ranges entered by the participants are plotted on the y axis and have been normalised with resepct to the date resolution. As expected the "maximum possible" values sit largely outside of the "most probable".

The provided values were highly symmetrical indicating little bias in perception towards either side of the date presented.

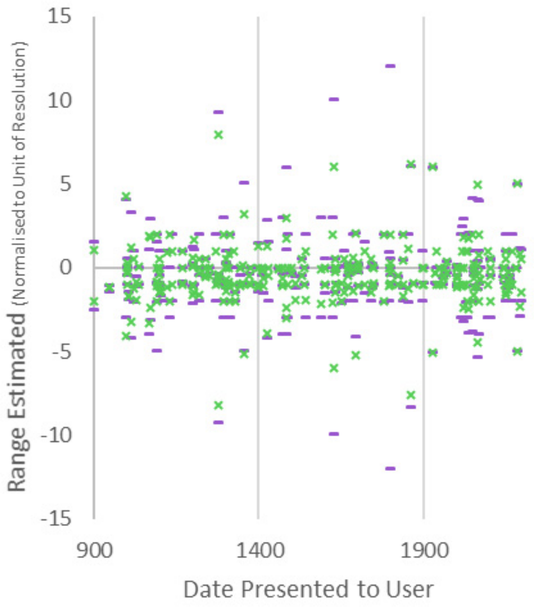


Figure 2: *Date ranges normalised to their unit of resolution captured during the survey ("Most Probable" extents indicated by green crosses, "Maximum Possible" extents indicated by purple dashes)*

The data also showed little to no trend that would indicate that that dates closer to the present are percieved as more accurate. This would imply that when numerically literate individuals are provided with a date, without additional context about the source, the distance from present is not a determining factor for the perceived range.

However, dates from real world sources that are further from the present are more likely to be described in less specific terms (i.e. with a larger resolution).

Further interrogation revealed that the perceived ranges were mostly determined by terms with which they are described. It was found that the data fell into two groups with similar average ranges.

The first group had similarly low average ranges, and comprised the terms "About", "Around", "Approximately" and "Circa". The usage of these implies a continuous distribution. The average ranges of these terms are displayed in Figure 3.

The second group of similar terms included "Perhaps" "Possibly" and "Probably"; statements that do not imply a continuous, but instead suggest separate discrete likelihoods. As such the method of selecting ranges for these terms may not have been optimal. This is supported by lower participants' confidence assessments associated with these terms. Nevertheless, the analysis produces a useful indication of the perceived possible ranges of dates for this term.

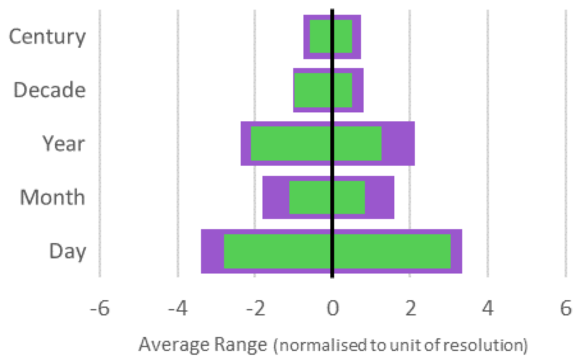


Figure 3: *Normalised average date ranges of the "continuous" terms for each of the tested resolutions. The inner green box indicates the "Most Probable" values, and the outer purple indicates the "Maximum Possible"*

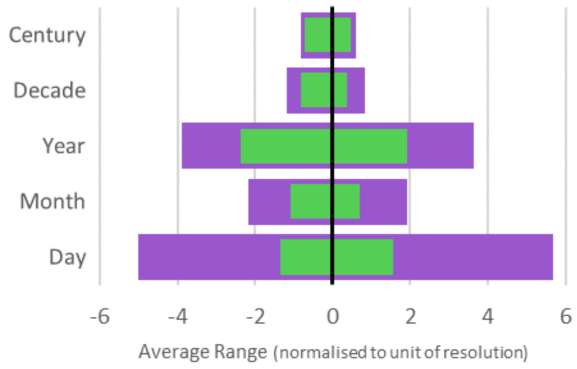


Figure 4: *Normalised average date ranges of the "discrete" terms for each of the tested resolutions*

Dates described using discrete terms have a greater perceived maximum possible range than those of the continuous group, but have a similar and in some cases smaller probable range. This may be because the discrete terms themselves describe the most probable value of the date, but not the extent of the possibility.

In both groups an increase in relative range from the date presented is seen with growing resolution, with the relative range for centuries being much smaller than that of days. This could be due to the periods of time being much larger and having less perceived likelihood of falling far from the given value.

Interestingly, the relative average of the probable range for months is around 50% smaller than years. This could be caused by textual presentation of the date in the instruction of the survey. Due to the cultural understanding of textual descriptors such as the names for months or seasons, these may be perceived to be more bounded than purely numerical date values.

Another possibility is that if a date lies within or across several months it is more likely to be described in terms of subdivisions, such as early, late or seasonal descriptors. This also implies that if a date is given with a month resolution that the error is perceived as less likely to cross into another boundary such as a different season.

5. OUTCOMES

The results of this experiment could be used in any visualisation that includes entries on a time scale. The average range values for the provided resolutions as shown in Figures 3 and 4 could be added to any entry on a visualisation described with one of the tested uncertainty terms.

Kräutli and Boyd Davis (2015) propose the use of a probability density function to calculate likelihood. The values provided in this paper can be used to inform the standard deviation to be used, and the probabilities may then be presented in a manner such as those they suggest.

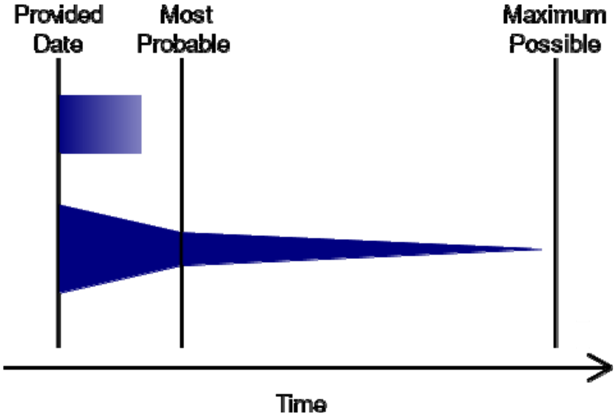


Figure 5: *Simple example of how transparency or width may be used to depict the probable date range of an event on a timeline*

The values could also be integrated into date parsing and visualisation tools such as TimeLineCurator (Fulda, Brehmer & Munzner 2015) which employ Natural Language Processing (NLP) but does not currently interpret uncertainty values terms.

There are some avenues that could be developed for future investigation. This study was only available in English, as such it may not be suitable for uncertainty terms used in different languages and cultures.

Additional uncertainty terms not used as part of this study that can be categorised as either “discrete” or “continuous” could be a further exercise. However, it may be that identification as part of either group may already provide the values to be utilised from this study.

A modification could be made to the existing study to allow for multiple selections for terms that imply discrete distributions (e.g., “perhaps”, “possibly”).

As previously identified there may be a difference in perception depending on whether the dates are described textually or numerically. This could be investigated by describing months using numbers, or including the day of the week on dates described at the day resolution.

While there were some ambiguous dates included in the dataset, they did not appear as outliers. The means of producing dates in the study would need to be weighted to produce more ambiguous dates.
Seasons and other terms that describe a subdivision of time (such as early, middle or late) also include a level of ambiguity that could be investigated.

6. SUMMARY

This study shows that despite real world errors increasing with distance from the present, given no additional information about the source, the uncertainty range attributed to a presented date by participants is not affected.

Study of resolution indicates the assumption of uncertainty to be generally symmetrical and increasing with resolution. It also highlights the effect of textually presenting a resolution rather than numerically.

Uncertainty terms fall into two groups and this together with the resolution appear to give consistent results that can be used for standard deviation when visually representing temporal data entries.

The introduction of comparative questions, multiple selections for discrete terms, or the addition of context or opinion questions could provide additional information on users’ perception of uncertainty as opposed to the consistent numerical evaluations that they perform.

7. ACKNOWLEDGMENTS

This research is undertaken as part of an AHRC funded Collaborative Doctoral Partnership between Royal College of Art and The National Archives (grant number AH/M004775/1).

8. REFERENCES

Basu, S. and Waymire, G. B. (2006) Recordkeeping and human evolution. *Accounting Horizons,* 20(3), pp. 201–229.

Bonneau, G. P., Hege, H. C., Johnson, C. R., Oliveira, M. M., Potter, K., Rheingans, P., and Schultz, T. (2014) Overview and state-of-the-art of uncertainty visualization. *Scientific Visualization,* pp. 3–27. Springer, London.

Fulda, J., Brehmel, M., and Munzner, T. (2016) TimeLineCurator: Interactive authoring of visual timelines from unstructured text. *IEEE Transactions on Visualization and Computer Graphics*, 22(1), pp. 300–309.

Kräutli, F. and Boyd Davis, S. (2013) Known unknowns: Representing uncertainty in historical time. In K. Ng, J. P. Bowen & S. McDaid (eds.), *EVA London 2013 Electronic Visualisation and the Arts*, London, 29–31 July. Electronic Workshops in Computing, BCS.

Kräutli, F. (2016) *Visualising Cultural Data: Exploring Digital Collections Through Timeline Visualisations*. PhD Thesis. Royal College of Art.

Internet Resources

Bristol City Council: Museum Collections, http://museums.bristol.gov.uk/ (retrieved 20 March 2017).

Collections – National Maritime Museum, http://collections.rmg.co.uk/ (retrieved 19 March 2017).

Explore our collections – Collections – Horniman Museum and Gardens, http://www.horniman.ac.uk/collections (retrieved 19 March 2017).

Explore the Royal Collection online – Royal Collection Trust, https://www.royalcollection.org.uk/collection/search (retrieved 20 March 2017).

Fitzwilliam Museum Collections Explorer – University of Cambridge, http://webapps.fitzmuseum.cam.ac.uk/explorer/ (retrieved 20 March 2017).

Hermitage Museum – Collection Online, https://www.hermitagemuseum.org/wps/portal/hermitage/explore/artworks (retrieved 20 March 2017).

LACMA Collections, http://collections.lacma.org/ (retrieved 20 March 2017).

National Trust Collections: Home, http://www.nationaltrustcollections.org.uk/ (retrieved 19 March 2017).

Poster and Artwork collection online from the London Transport Museum, http://www.ltmcollection.org/posters/ (retrieved 19 March 2017).

Search The Collection – The Courtauld Institute of Art, http://courtauld.ac.uk/gallery/search-the-collection (retrieved 19 March 2017).

TimeLineCurator, http://www.cs.ubc.ca/group/infovis/software/TimeLineCurator/ (retrieved 22 March 2017).

Wellcome Library | Search the catalogues, https://wellcomelibrary.org/search-the-catalogues/ (retrieved 20 March 2017).