LITERARY GEOGRAPHIES

Visualising the Uncertain in Heritage Collections: Understanding, Exploring and Representing Uncertainty in the First World War British Unit War Diaries

Johannes Liem

Danube University

Aidan Slingsby

City, University of London

Eirini Goudarouli

The National Archives, UK

Mark Bell

The National Archives, UK

Cagatay Turkay

University of Warwick

Charles Perin

University of Victoria

Jo Wood

City, University of London

Abstract:
This paper argues that cultural heritage data is inherently ambiguous and may involve different types and levels of uncertainty. Using a variety of examples based on The National Archives (UK)'s Unit War Diaries collection unveiling stories of the British Army and its units on the Western Front in the First World War, we discuss the ways in which visualisation can help us approach heritage collections as data, enabling their visual representation in a constructive and informed way. It also aims to open up the discussion about the theoretical and methodological challenges that uncertainty, which is often hidden, can bring to the understanding of ambiguous heritage data.

In brief, we discuss ways in which uncertainty appears in cultural heritage collections, either as something innate in the collections or resulting from the data extraction and narrative construction process. We identify three main types of uncertainty: *inaccuracy*, *incompleteness* and *ambiguity*, with the latter then subdivided into *inconsistency*, *imprecision* and *non-specificity*. Distinguishing, considering and quantifying these different types of uncertainty can help understand the level of confidence that we can have in narratives, source data and the extraction process. This can then enhance the discoverability of cultural heritage collections that involve high levels of uncertainty.

In this way, we suggest that cultural heritage organisations should strategically focus on improving the understandability and discoverability of their digital collections by exposing and embracing uncertainty in cultural heritage collections and by innovating in its visual presentation to researchers and the public.

Keywords: Uncertainty; representation; visualisation; cultural heritage; collections; data.

Author contact: Johannes.Liem@city.ac.uk; A.Slingsby@city.ac.uk; Eirini.Goudarouli@nationalarchives.gov.uk; Mark.Bell@nationalarchives.gov.uk; Cagatay.Turkay@warwick.ac.uk; cperin@uvic.ca J.D.Wood@city.ac.uk

Heritage collections, held by recordkeeping and cultural heritage organisations, provide access to information that gives insights into our past. They are comprised of quantitative and qualitative artefacts in analogue, digitised and born-digital formats, such as administrative records, newspapers, maps, websites, social media activities, videos, photographs and audio recordings. To uncover new insights and create narratives, researchers study and analyse these artefacts, synthesising contextual knowledge from different archival resources, exploring new types of knowledge and providing new interpretations. Depending on the scope of the research and target audiences, outputs vary – including those that are more descriptive (e.g. essays, textbooks), narrative-based (e.g. documentaries) or quantitative (e.g. timelines of key events, family trees, infographics, maps) in nature.

Different types and degrees of uncertainty are inherent in all such compiled outputs, resulting from error and ambiguity in the original source material (e.g. misspelt names and names that do not unequivocally describe a single place or person) and subjective inference (e.g. relating common places or people between sources; making informed decisions about the order of events; or estimating timings based on other facts). The degree to which this uncertainty is exposed, and the methods for doing so, will depend on the research goals, target audiences, and output type. Outputs for a general audience may reveal little of this uncertainty for the sake of clarity or simplicity. This is particularly the case for some of the more systematic quantitative outputs such as maps, timelines, and social networks. For example, a narrative presented as a map of key events would need a specific location associated with each event, and so if a key event is missing a location, it would either need to be omitted or a plausible location inferred to complete the narrative. In the latter, expressing uncertainty in location may be considered an unnecessary complication, particularly if it has little effect on the narrative. However, if the researcher extracts such inferred facts without knowing either that they are inferred or the degree of uncertainty associated with them, this may be problematic when these facts are used in other contexts.

Visualisation is an expressive means to present outputs. Infographics have been popularised by news outlets for explaining how events have unfolded or for highlighting links between events, people and places. The arrangement of visual elements laid out in space can convey relationships between elements, including geographical (e.g. maps), similarity, temporal (e.g. timelines), hierarchy (e.g. tables) and membership (e.g. Venn diagrams). Lines connecting elements indicate relationships between them (e.g. flowcharts and node-link diagrams). Visual variables (Bertin 1983) such as size, colour, transparency, texture, motion and symbols can represent quantitative data about these elements or relationships, such as quantity, average and category. Visualisation is also an expressive means to represent uncertainty, via, for instance, colour lightness, transparency, fuzziness/sketchiness and dotted/dashed lines (MacEachren et al. 2012). Techniques based on statistical graphics (e.g. error bars) are also often considered intuitive means to indicate ranges of possible values or times. Labels, too, are effective ways of augmenting these with other contextual information.

This paper is the joint work of an interdisciplinary group of authors consisting of Cultural Heritage professionals, with an expertise in digital research in Cultural Heritage, and of Computer Science scholars specialising in Information Analysis and Visualisation. In it, we focus on visualisation outputs for presenting knowledge extracted from heritage collections. We argue that the quantitative and systematic nature of visualisation provides a means of explicitly expressing gaps and uncertainties in knowledge and making them transparent within the inferences made from heritage collections. We discuss this in general terms and through a case study. The case study is based on The National Archives' Unit War Diaries collection and the crowdsourcing project, Operation War Diary (OWD). The OWD project was launched in 2014 and was a collaboration between The National Archives, the Imperial War Museum and University of Oxford's Zooniverse, the world's largest platform for citizen research. During this 5-year project, thousands of volunteers engaged with an online crowd-sourcing platform to crowdsource transcriptions of these fascinating documents, unveiling stories of the British Army and its units on the Western Front in the First World War.[1]

While this paper focusses on our work with the theoretical and methodological challenges of uncertainty and ambiguity in heritage data, there is a potentially useful implied link to current literary geographies work on literary maps, mappings and cartographies. Scholarship in these areas also often has to consider issues of uncertainty, even if the discussion is not always couched in those terms. For example, the term 'slipperiness' (Bushell 2012) is regularly used by Bushell et al. when referring to literary space's 'tenuous relationship with real geographies' (Bushell et al. 2021). Fictional space is described as 'fragmentary with uncertain, vague boundaries, and difficult to localise' (Reuschel and Hurni 2011). GIS software removes vagueness by encouraging the reduction of areas to points (Dai Prà and Gabellieri 2021), but this creates a tension between 'generalisation and the "infinite variety of detail"' (Taylor et al. 2018). Techniques used to counter this generalisation include, fuzzy shapes, animating indeterminate locations, inferring paths between places, density smoothing, fuzzy boundaries, and deep mapping (Reuschel and Hurni 2011; Murrieta-Flores et al. 2017; Taylor et al. 2018). Others have used approaches such as network graphs, either as topological models of space (Bushell et al. 2021) or as representations of qualitative spatial relationships (Stell 2019). In our work, we engage with related critical interpretations of uncertainty in our consideration of spatio-temporal relationships and the mapping and visualisation of place and space.[2]

In particular, we wish to reflect on the importance of understanding, exploring and representing uncertainty by distinguishing the different types that may exist in cultural heritage collections, and by discussing how they arise. For example, we aim to consider the implications of the uncertainty introduced into heritage collections when extracting entities (e.g. places, people and events) and creating links between them (e.g. associations and dependencies). We also reflect on how visualisation techniques can help us approach heritage collections as data that inherently are ambiguous by unlocking new ways of presenting uncertainty in cultural heritage data to researchers and to the public.

Identifying sources of uncertainty

The intention of this paper is to demonstrate how uncertainty permeates archival collections not only in the documents themselves but through the processes we use in presenting their contents to users of the archive. In this section, we begin by introducing a typology of uncertainty drawing on Smithson's 'Typography of Ignorance' (Smithson 1989) and Klir's taxonomy of uncertainty (Klir and Wierman 1999). This helps us demonstrate three layers of process between historical events and presentation of documentary evidence as data, each introducing their own uncertainties, through the Unit War Diaries case study. At the top level of Figure 1, uncertainty is separated into three types: *Inaccuracy*, *Incompleteness* and *Ambiguity*. The latter category is further broken down by *Inconsistency*, *Imprecision*, and *Non-specificity*. The data visualised in our case study are the result

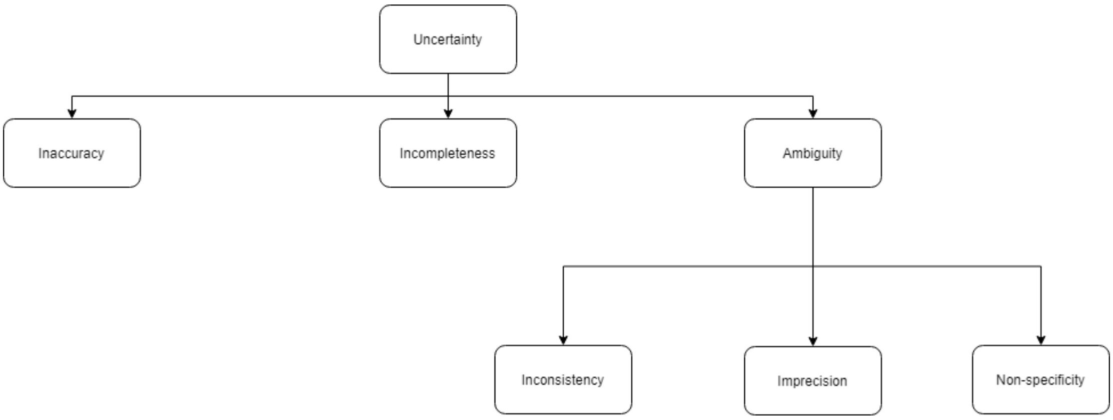


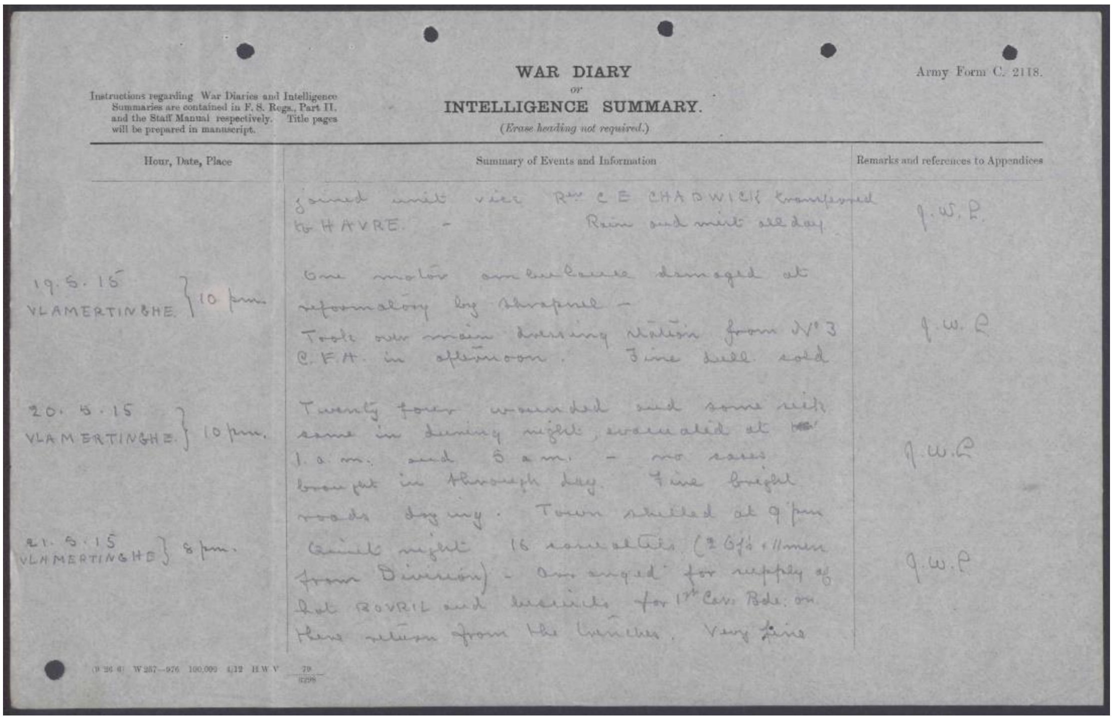
Figure 1. Types of uncertainty that may exist in heritage collections, drawing on Smithson's "Typography of Ignorance" (Smithson 1989) and Klir's taxonomy of uncertainty (Klir and Wierman 1999).

of a three-stage process, each introducing new sources of uncertainty, which we will term *layers of uncertainty*.

The first layer of uncertainty pertains to the recording of historical events and the creation of the record. For example, focusing on the Unit War Diaries, the journey of a war diary begins in the trenches when someone in an army unit writes an abbreviated account of the unit's activities for a period of time. This could be anything from a specific hour to a week, and there can be gaps when no activity is recorded. Figure 2 shows an example war diary page.[3] It consists of three columns of information, the first being the date of the entry and location, the second a narrative description of the day's activities, and the third (optional) additional remarks. This particular page mentions shrapnel damage to an ambulance, shelling, 16 casualties, and the arrangement of '…a supply of hot BOVRIL and biscuits on return to the trenches. Very fine'. It is unclear whether that last comment referred to the Bovril or the weather.

The entries are necessarily brief, except when regarding something particularly important such as an account of a battle, with heavy use of abbreviations and acronyms. People named are generally officers although lower ranks may get special mention in the case of heroic actions or death. This brevity and the existence of temporal gaps means they are *incomplete* accounts.

Places are mentioned frequently, either in discussing accounts of past events, current location, future intentions, or locations of unit members who are positioned elsewhere (possibly back in the UK, or in hospital). What we must remember is that these were soldiers in a foreign land, often in rural areas without clear signage or landmarks. The named places could therefore be based on imperfect knowledge, thus giving rise to toponymic *inaccuracy*. There was also a tendency to spell place names phonetically if they didn't know the correct spelling. This can be looked upon as *inaccuracy* but also as a source of *inconsistency* between spellings of the same place.



Instructions regarding War Diaries and Intelligence Summaries are contained in F. S. Regs., Part II. and the Staff Manual respectively. Title pages will be prepared in manuscript.

WAR DIARY
or
INTELLIGENCE SUMMARY.
(*Erase heading not required.*)

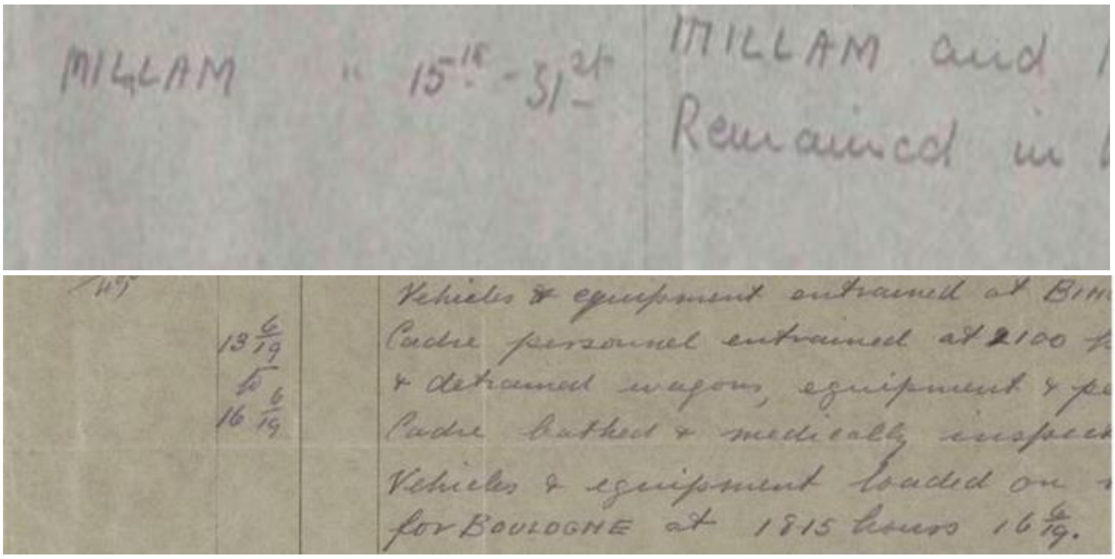
Army Form C. 2118.

Hour, Date, Place	Summary of Events and Information	Remarks and references to Appendices
	joined unit vice Rev C E CHADWICK transferred to HAVRE. - Rain and mist all day	J.W.P.
19.5.15 VLAMERTINGHE } 10 pm	One motor ambulance damaged at reformatory by shrapnel - Took over main dressing station from No 3 C.F.A. in afternoon. Fine dull cold	J.W.P
20.5.15 VLAMERTINGHE } 10 pm.	Twenty four wounded and some sick came in during night, evacuated at 1 a.m. and 5 a.m. - no cases brought in through day. Fine bright roads drying. Town shelled at 9 pm	J.W.P
21.5.15 VLAMERTINGHE } 8 pm.	Quiet night 16 casualties (2 Offs + 11 men from Division) - Arranged for supply of hot BOVRIL and biscuits for 1st Cav. Bde. on their return from the trenches. Very fine	J.W.P

(9 26 6) W 257—976 100,000 4/12 H.W.V 79/8298

Figure 2. War Diary image from 1st Cavalry Field Ambulance division; reference WO95/1104/3.

Each diary entry was dated with varying degrees of precision. At one end of the scale are exact times (e.g. 1815 hours in Figure 3, bottom) while at the other are more *imprecise* date ranges (e.g. 15th-31st also Figure 3). The dates themselves are usually *incomplete* with years often excluded and sometimes months too. The use of ditto (") marks is prevalent; these can occasionally refer to something on the previous page. From our experience, there is no reason to doubt the accuracy of the dates.

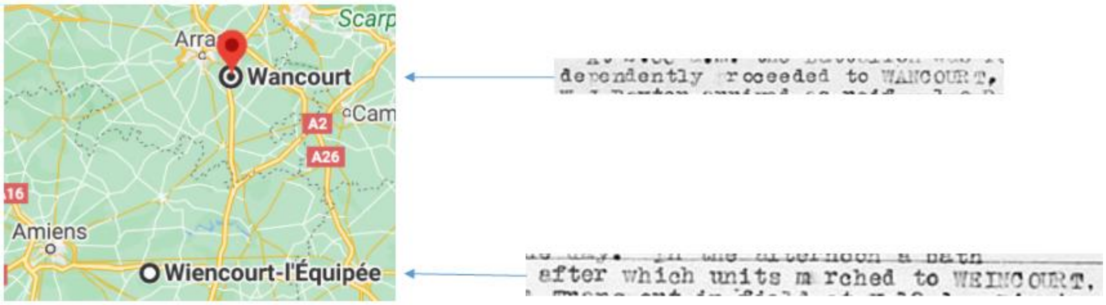


MILLAM " 15th - 31st | MILLAM and
Remained in

13 6/19 to 16 6/19 | Vehicles & equipment entrained at B
Cadre personnel entrained at 2100
& detrained wagons, equipment & p
Cadre bathed & medically inspec
Vehicles & equipment loaded on
for BOULOGNE at 1815 hours 16 6/19.

Figure 3. Images depicting date ranges in diaries.

The second layer of uncertainty occurs when information is extracted from the documents. For example, 100 years after the beginning of the war, The National Archives, in collaboration with the Imperial War Museum and Zooniverse, commenced the OWD project to transcribe and annotate over one million battlefield notes from this fascinating historical collection. By examining the transcription process, we see how further sources of uncertainty are introduced. Through the Zooniverse platform, volunteers were asked to highlight entities (people, places) and events (dates, activities, movement, weather) on a randomly selected diary page presented to them through a browser-based interface, label them by type, and transcribe the highlighted text. In addition to this, they were asked to geo-locate places on a map. The volunteers in this instance were in a similar situation to the soldiers, locating places in unfamiliar territories. An incorrect spelling either in the diary or introduced by transcription may result in geo-locating the wrong place, an *inaccuracy*, or not being able to geo-locate at all, an instance of *incompleteness*. Some places like "No Man's Land" cannot be located but are nevertheless tagged as locations in the data.



dependently roceeded to WANCOURT,

after which units m rched to WEINCOURT.

Figure 4. The towns of Wancourt and Wiencourt-l'Equipee.

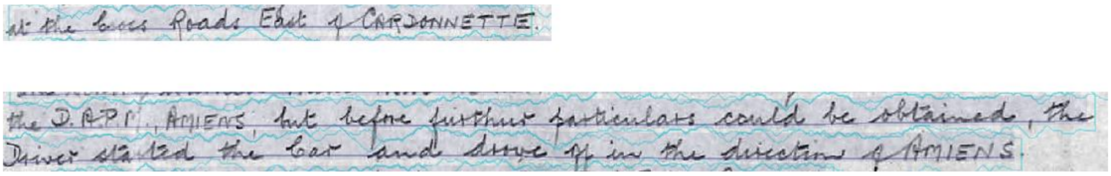
Another common scenario is that of *ambiguous* place names, an example of *non-specificity*. The geolocation mapping interface on the platform was reported, by some users, as being too small to identify the best of multiple options on the map. Consider the example (Figure 4) of the village of Wiencourt-l-Equipee whose name was shortened to Weincourt (note the switching of the 'e' and 'i').[4] The same battalion also travelled to Wancourt some 70km to the north. The misspelling of Wiencourt meant that volunteers sometimes geo-located it as Wancourt. What's more, disambiguation relies on local or contextual knowledge. Volunteers with a good historical knowledge of key WW1 campaigns could leverage that knowledge to identify locations, rather than having to rely on guesswork. Elsewhere, the random order of diaries meant that volunteers missed the opportunity to follow a narrative and use context to fill in the gaps. The experience of two professional transcribers working on a selection of war diaries was that they often went back to correct errors as they progressed through the story, learning personalities and places, and becoming increasingly proficient in deciphering handwriting.[5] They also highlighted how difficult it was to interpret the handwriting of multiple authors and noted that the writing of a single person would change when under duress. The luxury of going back was not afforded to the crowdsourcing volunteers but, if they transcribed enough pages, they could still develop similar knowledge of people and places. There is evidence of transcribers filling in data which was not on the pages, such as the year of the diary entry, or the rank of a named officer. Since it is not clear whether this was desired behaviour, or what the source of their knowledge about the missing information was, it is therefore unverifiable.

The third layer of uncertainty may be introduced during the interpretation of the data. When data is ambiguous, the usual approach is to resolve the ambiguity (e.g. by simple majority) to create a single crisp outcome. Traditional transcription projects, whether undertaken by professionals, academics, or volunteers, usually use a double keying approach in which each document is transcribed by two people. If the two transcribers agree they are accepted, but if there is any disagreement an expert arbiter views the document and transcriptions and makes the final decision.

However, crowdsourcing works without a final decision maker. Instead, up to five opinions on each page are gathered and the resulting data needs to be reconciled before it can be interpreted or presented to users of the archive. This reconciliation process is complex and contains frequent disagreement (*inconsistency*) between transcribers, so a simple majority vote system does not always work. In the case of an even number of transcribers, the vote may be split, but we could also consider whether transcriber opinions should be weighted by experience with the project and material. There are ethical questions around weighting volunteer contributions, and this is not an avenue we explored in the case study. Then there is the case of whether data from a page transcribed by a single volunteer should be treated equally to one transcribed by five. In this instance, an insufficient number of transcriptions may undermine confidence in the data and thus result in a different form of *incompleteness*. Aside from inconsistency in transcriptions, there was also inconsistency in highlighting the position of entities on the page, particularly dates. The data from the crowdsourcing platform included coordinates of a single point representing the position of the tagged entity on the page. These coordinates were given as a rounded percentage relative to the height and width of the page (if the page was 20cm x 40cm, then an item at position (10.5, 30) would have coordinates (52, 75)). This coordinate system was *imprecise* but also led to *inconsistency* since transcribers might variously tag a date at the top, middle, or bottom of the entry it related to. So, in the case of long entries this resulted in large discrepancies.

Further *ambiguity* ensued when tags at the bottom of one entry were close to tags at the top of another. Date ranges could be transcribed in different ways as evidenced in the date transcriptions in Figure 3. Of the five volunteers transcribing the top image (15th - 31st), two transcribed only the 15th, one only the 31st, one both, and the one every day from 15th to 31st as separate dates. In this latter case each day would be given a different coordinate to cover the length of the entry, which in the data could be interpreted as events occurring on specific dates rather than at indeterminate points during a period; a false sense of *precision*.

In the lower image of Figure 3 (13th to 16th June 1919), all three interpreters transcribed the first date (13th) while only two captured the 16th. Of the two volunteers who transcribed both start and end date, one incorrectly set the year to 1916. This pattern was consistent for all date ranges on this page. One further source of uncertainty comes from the loss of context resulting from extracting entities rather than fully transcribing the pages. When a location appears in the data, we do not know whether it was being discussed in terms of the past, present, or future, or whether it was related to the location of the unit or some other entity (HQ, hospital, enemy troops). Figure 5 shows a report from a checkpoint set up "East of Cardonnette", North East of Amiens, encountering a car which "drove off in the direction of AMIENS". The words 'East', 'direction of', and indeed 'driver' are lost by the transcription process. This provides a further example of *incompleteness*.


at the Cross Roads East of CARDONNETTE.

the D.A.P.M., AMIENS, but before further particulars could be obtained, the Driver started the car and drove off in the direction of AMIENS.

Figure 5. An incident at a checkpoint.

The task of aggregating this crowdsourced data to establish some form of true representation of the documents is highly complex and may not be possible with such incomplete data. The power of visualisation is that it can embrace such ambiguity in the data and allow for the interpretation of wider patterns within it. This paper will now address specific types of uncertainty, that is *ambiguity* (particularly *inconsistency*) and *incompleteness,* through a novel visualisation approach.

Extracting and Presenting Narratives with Visualisation

Information visualisation can be used to help extract narrative by laying out visual elements in ways that indicate relationships and using visual variables — i.e. size, colour, transparency, texture, motion and symbols — to convey additional information about the data (Bertin 1983). Transcripts, entities, attributes and relationships between entities can also be summarised quantitatively in various overviews. For example, simply plotting the number of mentions of a person or place along a timeline or on a map provides a temporal or spatial summary that supports quality checking by identifying where there may be gaps and validations. Interactive visualisation techniques that facilitate zooming, filtering and details-on-demand enable summaries to be further explored through disaggregation (Shneiderman 1996).

Bar graphs, maps and histograms showing counts of entities and their distributions can indicate where there appear to be gaps in the data, giving the user or researcher the opportunity to check the original material. It can also be used to signpost potential problems with entity matching and links inferred between entities. In the case study described below, time histograms were used to identify the battalions that were more complete (Figure 6b – the figure can be found in the next section) and could therefore form the basis for rich narratives.

Furthermore, visualisation can be used for presenting narratives. As discussed, uncertainty is inherent in narratives constructed from information held in heritage collections. When narratives are presented, some expose this uncertainty by giving caveats and alternative explanations, whereas others gloss over it to present a clear and unambiguous narrative. Where visualisation is the medium to present narratives, it has the expressive power to indicate the presence and degree of different types and sources of uncertainty using visual variables and symbolism (Thomson et al.. 2005; Potter, Rosen and Johnson 2012; Padilla, Kay and Hullman 2021). Colour lightness, transparency and fuzziness/sketchiness are commonly used to indicate uncertainty (Wood et al. 2012). Examples include how much transcriber agreement there was for an allocated category, or a measure of confidence in the value returned by a statistical algorithm. Dotted/dashed lines that link entities (Boukhelifa et al. 2012) are intuitive means to indicate uncertainty in the relationships. Geographical positional uncertainty can be indicated with visually fuzzy boundaries, standard ellipses (Yuill 1971), or partially transparent extents. Techniques based on statistical graphics (e.g. error bars, Olston and Mackinlay 2002) are also intuitive means to indicate ranges of possible values or times. Finally, labels are also effective ways of augmenting these with caveats and other contextual information.

When representing narratives, *incompleteness* can be problematic. For example, if a narrative is presented as a map of key events in which the whereabouts of some of these events are unknown, a decision needs to be made about whether to omit these events or to infer a plausible location to give the narrative more clarity. In the latter case, a further decision needs to be made as to whether the *ambiguity* in some of the inferred locations should be identified or not. Whichever decision is taken, visualisation can express this *incompleteness*, either qualitatively (through a descriptive annotation) or more quantitatively (with the amount of missing information indicated by bar length, colour lightness or transparency). As we will discuss in detail in the next section, we use transparency to indicate where we are less sure about the geographical location.

Case Study: *The Unit War Diaries*

Time and place play an important role in cultural heritage data. Both data dimensions are well suited to presentation by narrative visualisation since time and space are integral to storytelling (Mayr and Windhager 2018). For example, animation can display movement through narrative time, thus providing a continuous representation of the depicted data. In the case of the Unit War Diaries, animation allows us to communicate the movements of military units over time. In this section, we discuss in detail an interdisciplinary collaboration formed with the aim to study, explore, analyse and visually represent data derived from the OWD project.

In 2018, The National Archives, in collaboration with the giCentre based at the Department of Computer Science at City, University of London, organised two interdisciplinary workshops to explore new ways of accessing and visually representing collections at scale. The group used data from the Unit War Diaries collections, consisting of hand-written diaries from the First World War that was crowdsourced during the OWD project. The data provided fascinating content for the group to explore the challenges that occur at different stages of the data visualisation process and particularly in relation to uncertainty. At both workshops, an interdisciplinary group consisting of data visualisation experts, digital humanists, computer scientists, archivists and historians came together to share different ideas of how to visually represent these spatio-temporal data and to create different narratives designs.

Through this collaboration, we extracted data from the Unit War Diaries, documenting the story of the British Army and its units on the Western Front, and used visualisation output to unlock their stories. We focussed on geographical and temporal aspects of the advance and retreat of the battalions, as well the types of activities they undertook over time. The visualisation was built around the two following *aims*:

- To graphically communicate the movement of troops over the course of the war, and the uncertainties and ambiguities associated with the OWD data so that, instead of giving the illusion that the data is complete and clean, we can leverage uncertainty to produce a more organic view of a unit's movement over time.
- To visualise the life behind the trenches (see Grayson 2016). The intention here is not to only communicate the battles, victories or defeats of WWI but, instead, to visualise a sense of the day-to-day life in the war zone. What's more, by including fighting and non-fighting activities as documented by the diaries, the aim was to shed light on aspects of war that are often not conveyed in history books, but remain buried in documents such as these diaries.

Data Extraction

The OWD project was a collaboration of The National Archives and Zooniverse during which a crowd-working platform was created to allow volunteers to contribute to digitising The National Archives' Unit War Diaries collection. Volunteers were provided images of individual pages of the Unit War Diaries and were asked to extract times, places, casualties, indications of unit strength, weather, everyday army life (or domestic activities), military activities, soldier names and ranks, location names and dates. This resulted in a rich spatiotemporal dataset (Figure 6a). The domestic and the military activities can be spatially distinguished based on their characteristics (Grayson 2016). While many military activities took place at the front (e.g. combat activities, digging trenches, repairing positions or patrolling), other activities were located behind the lines (e.g. being in reserve or support, moving from place to place, training, casualty treatment or re-supplying), as well as domestic activities (e.g. sports and leisure, practicing of religion or hygiene). This distinction was made to attain our narrative intent to communicate the life behind the trenches.

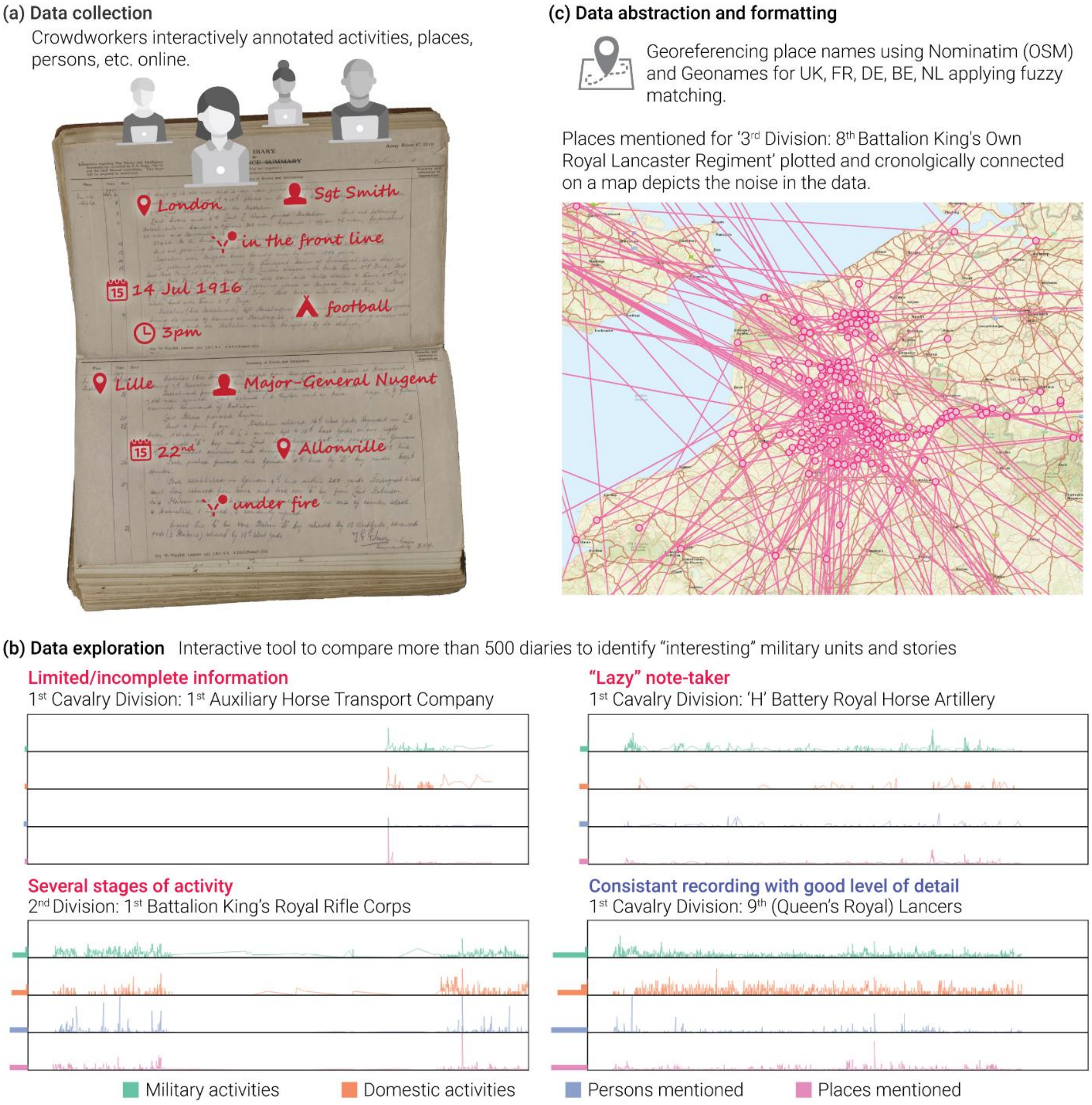
Exploratory visualisation of the quantitative summaries was an important tool to help validate the extracted data, gain an overview and, therefore, greater understanding. To that end, we implemented an interface that allowed us to open and parse datasets, break the data down according to military units, and visualise summary statistics for each unit. These summary statistics consist of time on the *x*-axis, from the beginning to the end of WWI, and, on four *y*-axes, mentions of military activities, domestic activities, persons and places per day in the unit's diary. It also made it possible to quickly identify firstly, the units whose data were relatively complete, and therefore, considered as good candidates for the outputs, and secondly, the units with less data which therefore bring different types and levels of uncertainty. Figure 6b shows the summaries of three examples of less complete diaries and one example with a consistent recording and a good level of detail.

Another important use of exploratory visualisation was to help validate the geographical locations identified during the georeferencing process (Figure 6c). Geographical coordinates were allocated to the place names by using the geocoder services Photon (based on OpenStreetMap) and Geonames.[6] Plotting the results and connecting them with lines in chronological order made it clear that some locations were unlikely to be correct when considered in their temporal context. This literal spatiotemporal mapping of locations has limitations for the visual communication of troop movements, but it did enable us to select the most appropriate locations based on where the battalion was before and after using the approach introduced below. It also allowed us to realise the volume and magnitude of errors in the data that may occur when automating the extraction.

Uncertainty in the Unit War Diaries

The visualisation design and development process engaged fully with the inherent ambiguity by identifying, understanding, quantifying and reducing (where possible) the different types of uncertainty that existed in the extracted data. The different types of uncertainties included possible errors in the original source material (e.g. misspellings); missing entries (*incompleteness*); text that was difficult to read; *imprecise* terms (e.g. France, instead of a specific location where an event took place); *non-specificity* (e.g. names that may refer to multiple different entities or names out of use) and *inconsistency* (e.g. inconsistency in source material, in extracted material, or disagreement amongst the transcribers).

There were also some persistent issues. As transcribers were self-selecting volunteers, they often lacked specific local and historical knowledge that might have helped them to make more informed judgements. They were also allocated random pages, so they did not have the benefit of temporal context that would have helped them in making judgements. For example, some of the data identified as locations of events were not locations at which the unit resided at the time, but actually a soldier's hometown. The temporal data typically do not capture the exact chronological order of locations on a single day.



(a) Data collection

Crowdworkers interactively annotated activities, places, persons, etc. online.

(c) Data abstraction and formatting

Georeferencing place names using Nominatim (OSM) and Geonames for UK, FR, DE, BE, NL applying fuzzy matching.

Places mentioned for '3rd Division: 8th Battalion King's Own Royal Lancaster Regiment' plotted and cronolgically connected on a map depicts the noise in the data.

(b) Data exploration Interactive tool to compare more than 500 diaries to identify "interesting" military units and stories

Limited/incomplete information
1st Cavalry Division: 1st Auxiliary Horse Transport Company

"Lazy" note-taker
1st Cavalry Division: 'H' Battery Royal Horse Artillery

Several stages of activity
2nd Division: 1st Battalion King's Royal Rifle Corps

Consistant recording with good level of detail
1st Cavalry Division: 9th (Queen's Royal) Lancers

Figure 6. Data extraction workflow. (a) Volunteers annotated pages of more than 500 diaries using a prescribed tagging catalogue; (b) Exploratory small multiple visualisations provided a quantitative and qualitative summary for the digitised diaries; (c) Georeferencing of mentioned place names and connected chronologically on a map.

For the narrative information visualisation, we chose to only consider positional uncertainty of the units through time. This locational uncertainty arises from a combination of the vague and ambiguous examples mentioned above as well as the fuzzy matching of place names during the georeferencing process. We designed a distance-based measure within a temporal window in which the greater the distance of a location from those within a short temporal window, the more uncertain it is (Figure 7). This allows unlikely locations to be treated differently within the visualisation design.

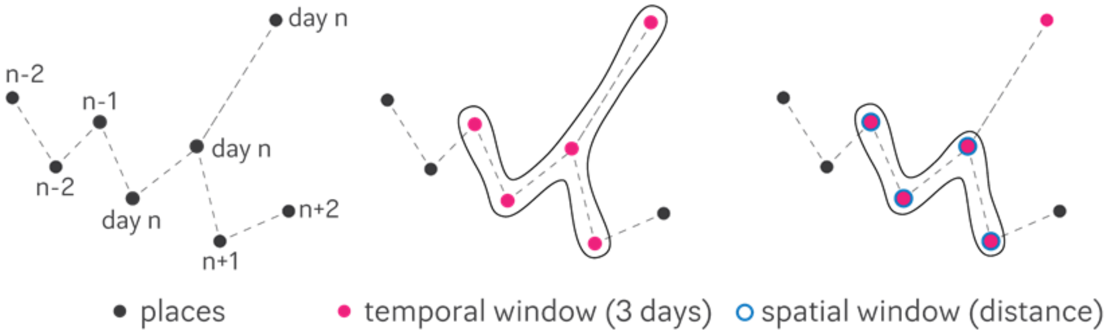


Figure 7. A temporal and a spatial window were used to select locations and aggregate them. Uncertain locations can be identified and, if necessary, excluded or treated separately within the visualisation.

Representing Uncertainty in the Unit War Diaries with Visualisation

In order to design visual representations of military units that incorporate and acknowledge uncertainty in the data, we explored possible geometries to visualize a unit's movement during the war. We found the visual certainty conveyed by crisp lines or point geometries (e.g. flow lines with arrowheads, animated glyphs) to be ill-suited to the characteristics of the OWD data. We instead concluded that, with animated areal geometries, we were able to achieve our first aim. Based on the selection method introduced above (Figure 7), we define GeoBlobs as an abstract representation of spatiotemporal data dedicated to conveying uncertain positions and uncertain temporal information of entities that move over time. Instead of showing an entity at a given point in time, GeoBlobs convey an unordered estimation of the possible locations over a temporal window using enclosed shapes. While line or dot/glyph visualisations become cluttered when the number of data points is increasing, GeoBlobs suffer less from such scalability issues. As described above, a start and end date define a window for temporal aggregation of the included locations, and sliding the window along the temporal axis animates the GeoBlob over time.

Many different parameters contribute to the design of GeoBlobs and can shape the aspects of conveying uncertainty or visual storytelling. We defined and explored the following design variations in the context of the OWD data, and several options were implemented in online prototypes. The design space and prototypes focused on geographical positional uncertainty, as outlined in the previous section. Different visual styles can be used to differentiate between more or less certain locations. Colour, transparency, focus (blur effect), pattern, or gradient can be used to vary the stroke or fill style of a GeoBlob (Figure 8) and therefore support an intended message. For example, using a blurry or sketchy (Wood *et al.,* 2012; Boukhelifa *et al.*, 2012) style can help to visually convey the uncertainty of the data.

Shape and form parameters can influence the saliency on the screen. We distinguish convex-hull-like, skeleton-like, and graph-like shapes (Figure 9, top). While skeleton-like geometries connect locations along the shortest distance, graph-like GeoBlobs consider the temporal order of locations. Buffers for nodes and edges can influence the appearance

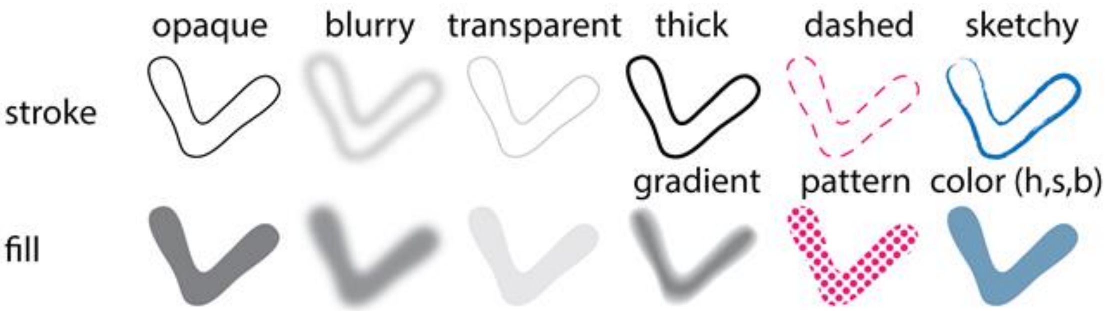


Figure 8. Design variations for stroke and fill style of GeoBlobs.

of a GeoBlob; we distinguish between wide, narrow and mixed buffers (Figure 9, bottom). The uncertainty of locations can be reflected in the buffer type. While unlikely locations are enclosed with a narrow buffer, more certain places have a wider buffer and a stronger saliency on the screen. The temporal and the spatial windows define the size and shape of the GeoBlob. Keeping both windows constant over time, the size and shape of the geometry convey information about the speed of a unit. While narrow and long shapes imply a greater distance covered, small circular forms indicate local or no movements.

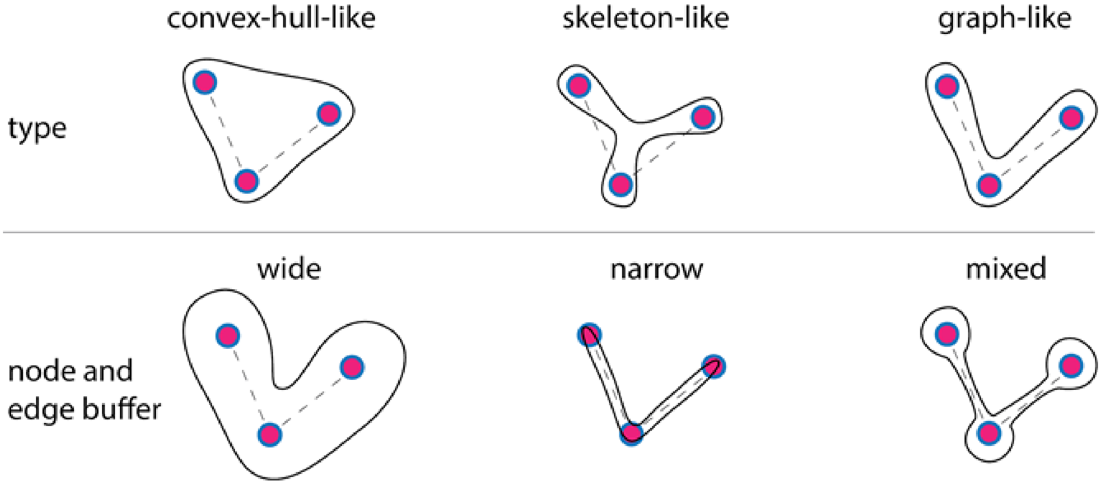


Figure 9. Design variations of the shape and form of GeoBlobs.

In combination with the above options, multiple, stacked layers of GeoBlobs can be used to convey location uncertainty (Figure 10, left). Locations within a description may refer to the current location or places in other contexts, for example, hospitals or nearby towns. Decreasing transparency and increasing distance indicates the probability of soldiers residing at the indicated location. The prototypes, for example, used a two layered blob symbology: locations, which are more certain, are enclosed with a thickly-outlined blob; and those which are uncertain appear as a transparent blob with blurred edges (Figure 10, right).

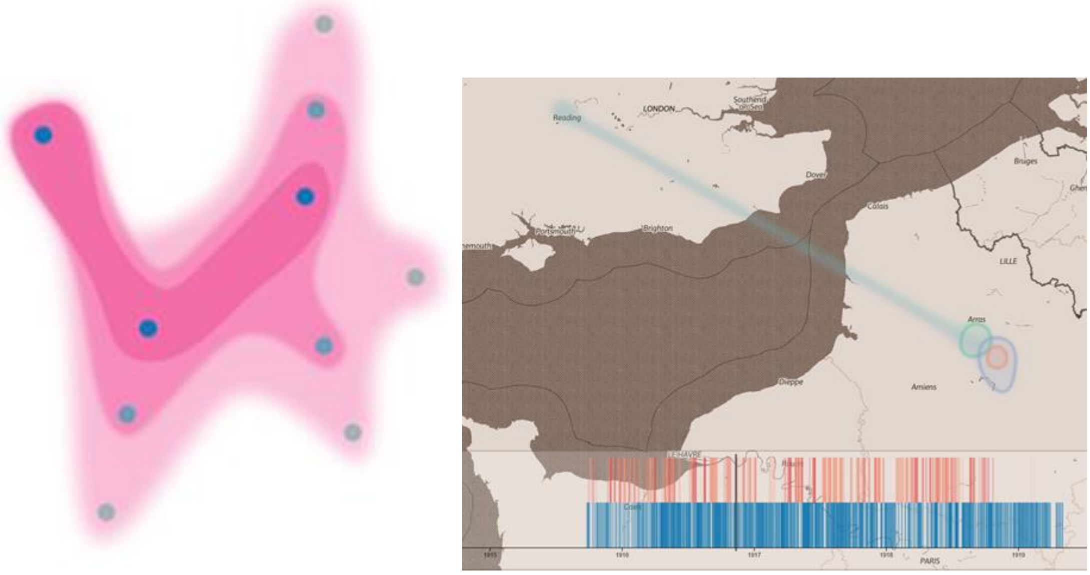


Figure 10. Left: Generating multiple layers for a single GeoBlob taking the different location probabilities into account and using transparency allows us to convey the uncertainty of the locations. Right: Prototype with a two layered GeoBlob. For example, while the troops resided in northern France (thick outline symbology), Reading was mentioned most likely in a different context (transparent fill and blurry edges).

Besides the visual aspects addressing uncertainty, several other design variations aim to support narrative aspects of the visualisation. Mapping multiple entities (e.g. several battalions or even enemies) allows the comparison of their individual movements (Figure 11). For example, the visualisation can demonstrate how two units were located at the same front, but then split after a battle.

In narrative visualisation, contextualisation provides a richer picture of events. Besides spatio-temporal aspects, the implemented prototypes conveyed information about the activities behind and at the front. In this case, contextualisation is possible through motion and animation or overlays.

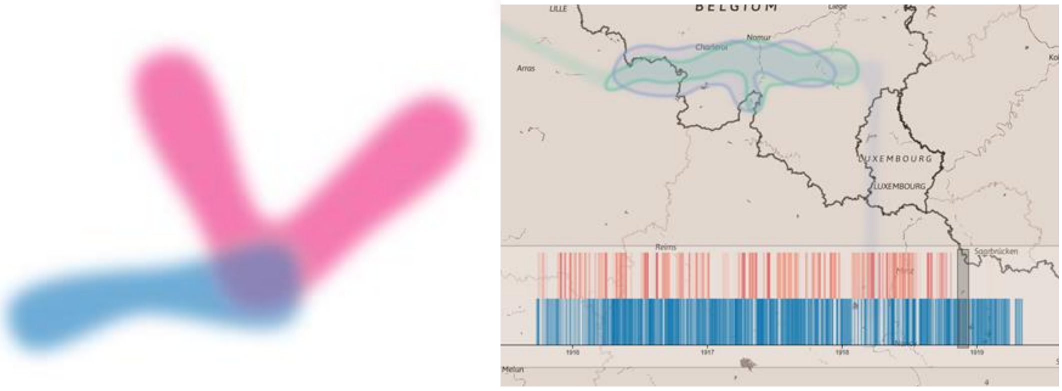


Figure 11. Left: Multiple entities to allow comparison. Right**:** Prototype showing two battalions moving side by side towards Germany at the end of the war.

When animating a GeoBlob, its motion encodes the overall direction of the displayed army unit. Here the temporal and spatial window might introduce a certain generalization level (e.g. movements back and forth within a larger temporal window are not obvious) which also reflects the uncertainty in the data. We also investigated how motion can be used to convey additional information such as, for instance, a shaking GeoBlob to convey combat activities.

Besides motion, outline, and fill styles, overlays can provide context by visually integrating unit activities such as fighting, re-supplying the front, and resting behind the lines or events like famous battles. Overlays can range from abstract dot geometries to more figurative depictions (Figure 12).

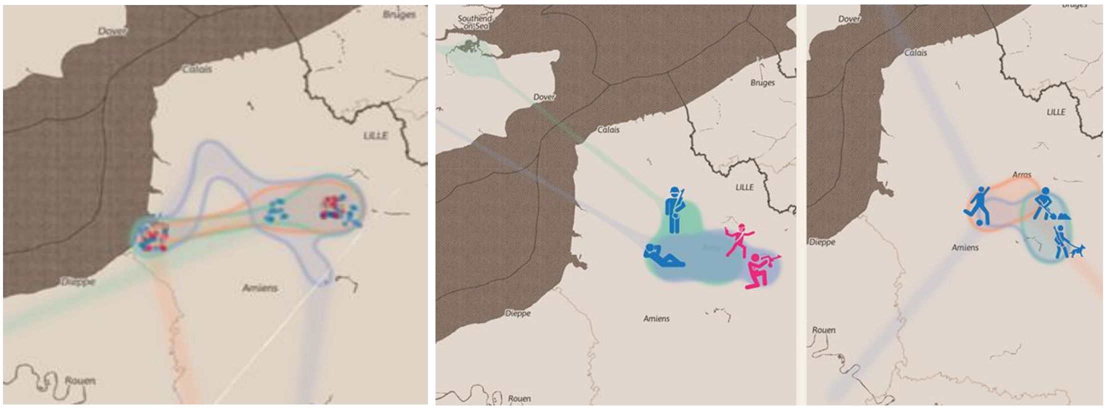


Figure 12. Contextualization through overlays depicting different types of activities.

In an additional prototype (Figure 13), we used text, colour, and screen space to communicate the different types of activities. While the blue area indicates the proportion of activities behind the front, red areas indicate the proportion of activities at the front (non-combat action in light red and combat action in red). The text is scaled regarding the number of the activities mentioned within the selected timespan. The timeline, which helps to navigate through the narrative sequence, includes a density plot providing a visual summary of the two types of activities (behind the lines and at the front). Adding such text-based and visual context to the map display, and depicting the scanned diary pages, helped to support the message about how much time soldiers spent behind the lines.

Additional narrativisation in the context of the GeoBlob design space is possible. For instance, additional information from the OWD data or other archival data sources about WWI could be incorporated, thus adding more aspects describing the life behind the trenches. This could be accomplished with additional map layers (e.g. showing weather information for the given time), as well as media elements such as photographs of activities. Whilst the background of the visualisation gives the appearance of being a historical map; it is not. Using a map from the period showing borders and other war-related information (e.g., trench maps) could help users to put themselves in the situation.

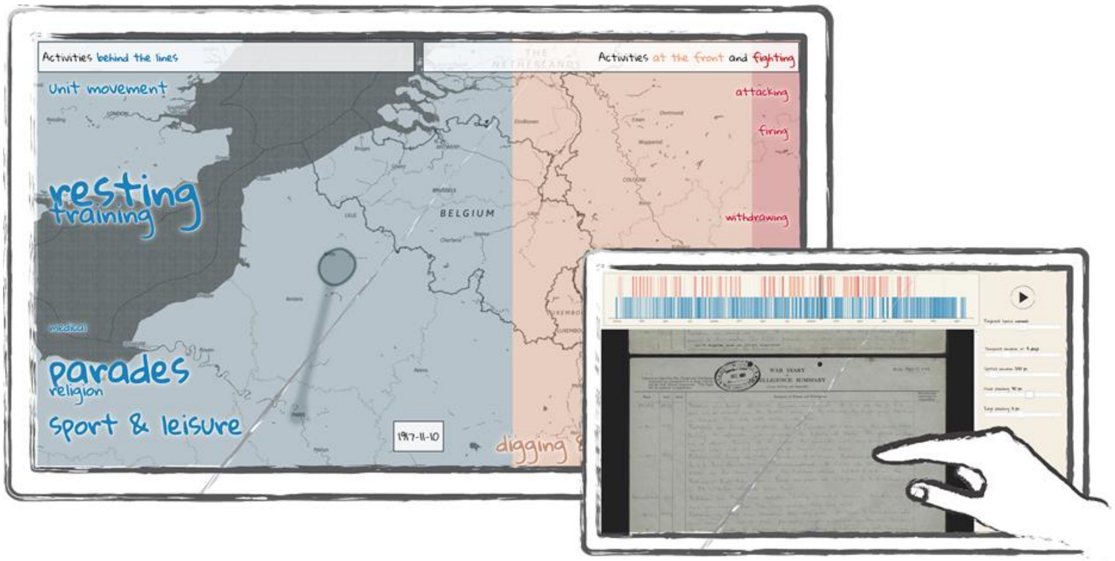


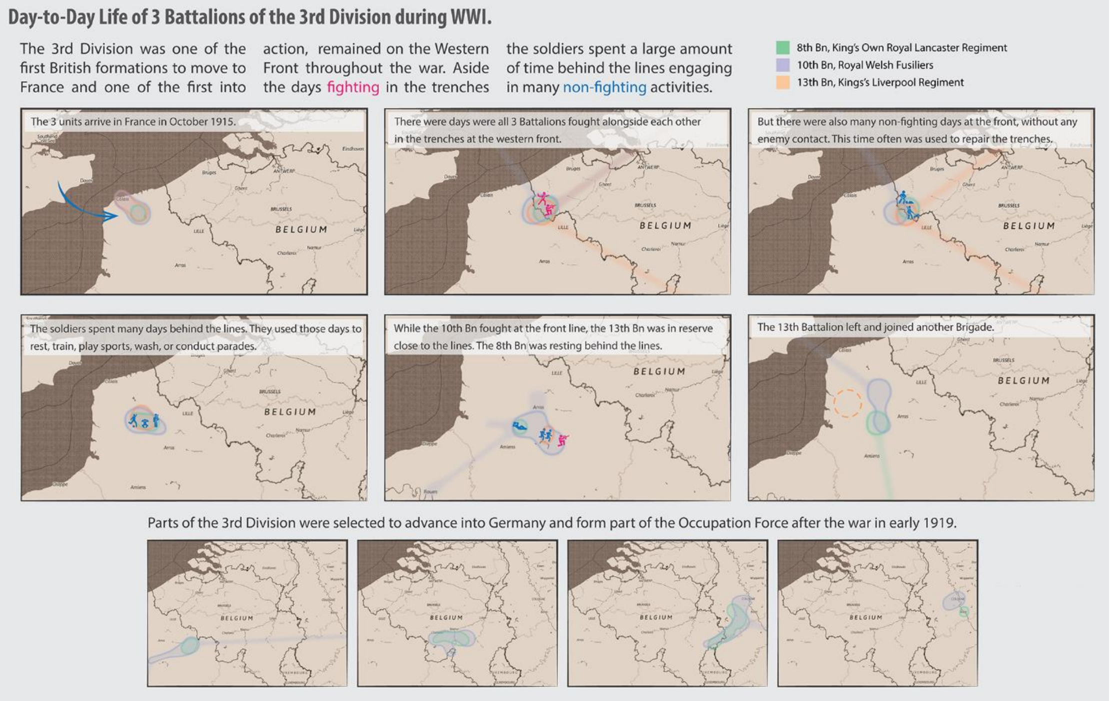
Figure 13. Double screen prototype with (left) the main map display and (right) the interaction display including facsimiles of the war diary pages.

The browser-based prototypes were implemented using the mapping framework leaflet with custom D3 SVG overlays to visualize the OWD data.[7] To calculate the geometry for the GeoBlobs, we make use of the Bubble Sets algorithm (Collins, Penn and Carpendale 2009). In the prototypes, a set of sliders allows the user to explore and vary the design space. It includes interactive adjustment of the temporal window (how many days are visualized by a unit GeoBlob); the spatial window (the distance from a blob's centroid to outlier locations forming the GeoBlob), the blob properties (like size or style); the playback speed; or the duration and fade out time of overlaid events.

GeoBlobs are well suited to any dataset with uncertain locations. An obvious comparison would be with hurricane prediction maps, since they too have spatial and temporal uncertainty. One problem they have in common with GeoBlobs is that more uncertain data can lead to much bigger footprints on the map, and users may be confused by the difference between large magnitude events and highly uncertain events. Moreover, the technique can be transferred to domains where the movement data is not uncertain, but has a high density, or where the area of coverage is of interest. For example, GeoBlobs are also a promising way of visualising the movement, spatial coverage, and the pressure in team sports, as well as visually aggregating the spatio-temporal trajectories of players. All of these are important topics in sports visualisation (Perin et al. 2018). Due to their configurable visual appearance, we also think GeoBlobs could be used in Data Comics (Bach et al. 2017). Figure 14, for example, shows a comic-like, short visual narrative about three battalions of the 3rd Division during WWI.

Reflections for Record-keeping Organisations

In the world of archives, the arrival of new types of records, from digitised to born-digital, is fundamentally changing the landscape and role of archivists and cultural heritage



Day-to-Day Life of 3 Battalions of the 3rd Division during WWI.

The 3rd Division was one of the first British formations to move to France and one of the first into action, remained on the Western Front throughout the war. Aside the days fighting in the trenches the soldiers spent a large amount of time behind the lines engaging in many non-fighting activities.

🟩 8th Bn, King's Own Royal Lancaster Regiment
🟪 10th Bn, Royal Welsh Fusiliers
🟧 13th Bn, Kings's Liverpool Regiment

The 3 units arrive in France in October 1915.

There were days were all 3 Battalions fought alongside each other in the trenches at the western front.

But there were also many non-fighting days at the front, without any enemy contact. This time often was used to repair the trenches.

The soldiers spent many days behind the lines. They used those days to rest, train, play sports, wash, or conduct parades.

While the 10th Bn fought at the front line, the 13th Bn was in reserve close to the lines. The 8th Bn was resting behind the lines.

The 13th Battalion left and joined another Brigade.

Parts of the 3rd Division were selected to advance into Germany and form part of the Occupation Force after the war in early 1919.

Figure 14. The day-to-day life of three battalions of the 3rd Division during WWI.

organisations. The emergence and application of new-generation technologies is also a factor that is rapidly leading to the appearance of a variety of complexities and challenges to the archival frameworks, requiring new capabilities and approaches on how best to contextualise and present the ambiguous nature of this new type of records. It also brings new opportunities, for example, in unlocking large-scale collections (either physical, digitised or born-digital) for research and experimentation, by enabling the extraction of their content as data and the exploration of new ways of providing discoverability and understandability of the collections (Goudarouli, Sexton and Sheridan 2019).

We have shown that uncertainty in cultural heritage data may be present in the source material and may be subsequently introduced through further data extraction and inference steps. Understanding, distinguishing and communicating the sources, types and degrees of uncertainty is key. In this paper, we have explored the forms these may take, noting how they are so often hidden from view, and have discussed the theoretical and methodological challenges that arise due to uncertainty, as well as how it can be embraced, through understanding, representation and exploration. We have also reflected on the implications as well as the benefits of cultural heritage organisations not only providing their users with access to the source data, but also providing them with the option to explore extracted entities, inferences and the resulting uncertainty that may derive from the source data. This is, potentially, an important shift in the role of such organisations, from traditionally being considered to provide "definitive" source data to being able to provide users with the space and tools to study, explore and understand the enriched data and all the associated uncertainty. In order to maintain the trust of users in this type of data, appropriate, detailed, and intuitive means of communicating this uncertainty is essential.

Therefore, this paper suggests that cultural heritage organisations should strategically focus on improving the understandability of their digital collections, by embracing uncertainty in their collections and by innovating the ways such uncertainty is displayed to academic and public audiences. Advanced computational methods use available data to summarise, link, make inferences and enable large-scale analysis. However, as discussed, data are often incomplete and ambiguous. Through research and exploration (such as close reading and comparing various resources), researchers might be able to disambiguate or resolve disagreements in archival resources. Going beyond exploration at record level and moving towards the comprehension of collections at scale, visualisation can be used as an expressive means to present data that involve gaps and uncertainties and to derive and infer more information to enrich collections.

There are exciting opportunities for using data extraction techniques and visualisation to help enrich heritage data. Interactive visualisation can help cultural heritage organisations to engage users in new and exciting ways and enable user feedback to help to improve the understanding of collections, including those with high levels of uncertainties. It can also provide the information needed to help the users interpret this information and to weave it into new narratives. The National Archives' ongoing research project *Engaging Crowds: Citizen research and heritage data at scale* promotes public participation in heritage research, and is an instance of how collaboration in virtual space engenders the creation and sharing of knowledge.[8] The project draws experience from OWD, which was the first crowdsourcing cultural heritage project for The National Archives and Zooniverse. The Engaging Crowds project works with sample data to develop post-processing workflows and focuses on iterative design of tasks to help avoid types of uncertainty that may occur during the design of the project and the interface.

Acknowledgments

We would like to thank Andrea Kocsis, Research Fellow (Advanced Digital Methods) funded by the Friends of The National Archives and Steven Hirschorn, Technical Architect at The National Archives UK, for providing insight and expertise that greatly improved parts of this paper.

Notes

[1] The Unit War Diaries (record series WO 95) represent one of the most popular collections held by The National Archives: https://www.nationalarchives.gov.uk/first-world-war/centenary-unit-war-diaries/. The National Archives digitised around 1.5 million pages of war diaries from the Unit War Diaries collection, which allowed volunteers around the world to access the diaries and embark on the hugely exciting crowdsourcing project, OWD. The OWD is now archived on the Internet Archive Wayback Machine.

[2] For example, recent work on literary mapping which engages with representation, time-space and place-space relationships, and uncertainty within literary studies includes: Cooper, Donaldson, and Murrieta-Flores (2016); Hones (2022); Thacker (2005); Bushell at al. (2022); Bushell et al. (2021); Stell (2019); Taylor et al. (2018); Reuschel and Hurni (2011).

[3] https://discovery.nationalarchives.gov.uk/details/r/C7351411

[4] We would like to thank Andrea Kocsis, Research Fellow (Advanced Digital Methods) at The National Archives, for providing this example that hugely benefit this section of the paper. In 2020 – 2021, Andrea extensively researched the data produced during the OWD project. You can find out more about her work at her most recent blogpost published at The National Archives' blog in April 2022: https://blog.nationalarchives.gov.uk/the-challenges-of-working-on-the-operation-war-diary-records/. Andrea's fellowship was funded by the Friends of The National Archives: https://www.nationalarchives.gov.uk/about/get-involved/friends-of-the-national-archives/.

[5] We would like to thank Celine Fernandez and Lisa Balderson for their valuable contribution to the transcriptions of this fascinating historical collection. Celine and Lise used *Transkribus* to transcribe a selection of war diaries from the Unit War Diaries collections. For more on *Transkribus*, please visit: https://readcoop.eu/transkribus/.

[6] https://www.geonames.org/.

[7] https://leafletjs.com/ and https://github.com/christiankaiser/Leaflet.D3SvgOverlay.

[8] The project received funded in 2020 by Arts and Humanities Research Council as part of the Towards a National Collection research programme. It is a collaboration between The National Archives, University of Oxford, Royal Botanic Gardens, Edinburgh and National Maritime Museum. https://tanc-ahrc.github.io/EngagingCrowds/.

Works Cited

Bach, B., Riche, N.H., Carpendale, S. and Pfister, H. (2017) 'The Emerging Genre of Data Comics.' *IEEE Computer Graphics and Applications*, 37(3), pp. 6-13.

Bertin, J. (1983) *Semiology of Graphics.* University of Wisconsin Press.

Boukhelifa, N., Bezerianos, A., Isenberg, T. and Fekete, J.-D. (2012) 'Evaluating Sketchiness as a Visual Variable for the Depiction of Qualitative Uncertainty.' *IEEE Transactions on Visualization and Computer Graphics*, 18 (12), pp. 2769-2778.

Bushell, S. (2012) 'The Slipperiness of Literary Maps: Critical Cartography and Literary Cartography.' *Cartographica: The International Journal for Geographic Information and Geovisualization*, 47, pp 149-160.

Bushell, S., Butler, J.O., Hay, D. and Hutcheon, R. (2022) 'Digital Literary Mapping: I. Visualizing and Reading Graph Topologies as Maps for Literature.' *Cartographica: The International Journal for Geographic Information and Geovisualization*, 57(1), pp. 11-36.

Bushell, S., Butler, J., Hay, D., Hutcheon, R. and Butterworth, A. (2021) 'Chronotopic Cartography: Mapping Literary Time-Space.' *Journal of Victorian Culture*, 26(2), pp. 310-325.

Collins, C., Penn, G. and Carpendale, S. (2009) 'Bubble Sets: Revealing Set Relations with Isocontours over Existing Visualizations.' *IEEE Transactions on Visualization and Computer Graphics*, 15(6), pp. 1009-1016.

Cooper, D., Donaldson, C. and Murrieta-Flores, P. (eds) (2016) *Literary Mapping in the Digital Age*. London: Routledge.

Dai Prà, E. and Gabellieri, N. (2021) 'Mapping the Grand Tour Travel Writings: a GIS-Based Inventorying and Spatial Analysis for Digital Humanities in Trentino-Alto Adige, Italy (XVI-XIX c.).' *Literary Geographies*, 7(2), pp. 251-274.

Goudarouli, E., Sexton, A. and Sheridan, J. (2019) 'The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK.' *Philosophy & Technology*, 32(1), pp. 173-183.

Grayson, R.S. (2016) 'A Life in the Trenches? The Use of Operation War Diary and Crowdsourcing Methods to Provide an Understanding of the British Army's Day-to-Day Life on the Western Front.' *British Journal for Military History*, 2(2), pp 160-185.

Hones, S. (2022) *Literary Geography*. Routledge.

Klir, G. and Wierman, M. (1999) *Uncertainty-Based Information: Elements of Generalized Information Theory. Studies in Fuzziness and Soft Computing.* 2nd edition. Physica-Verlag Heidelberg.

MacEachren, A.M., Roth, R.E., O'Brien, J., Li, B., Swingley, D. and Gahegan, M. (2012) 'Visual Semiotics Uncertainty Visualization: An Empirical Study.' *IEEE Transactions on Visualization and Computer Graphics*, 18(12), pp. 2496-2505.

Mayr, E. and Windhager, F. (2018). 'Once upon a Spacetime: Visual Storytelling in Cognitive and Geotemporal Information Spaces.' *ISPRS International Journal of Geo-Information*, 7(3), 96.

Murrieta-Flores, P., Donaldson, C. and Gregory, I. (2017) 'GIS and literary history: Advancing digital humanities research through the spatial analysis of historical travel writing and topographical literature.' *Digital Humanities Quarterly*, 11(1).

Olston, C. and Mackinlay, J.D. (2002) 'Visualizing data with bounded uncertainty.' In *IEEE Symposium on Information Visualization, 2002*. INFOVIS 2002, October 2002. pp. 37-40.

Padilla, L., Kay, M. and Hullman, J. (2021) 'Uncertainty Visualization.' In *Wiley StatsRef: Statistics Reference Online. American Cancer Society*. pp. 1-18.

Perin, C., Vuillemot, R., Stolper, C.D., Stasko, J.T., Wood, J. and Carpendale, S. (2018) 'State of the Art of Sports Data Visualization.' *Computer Graphics Forum*, 37(3), pp. 663-686.

Potter, K., Rosen, P. and Johnson, C.R. (2012) 'From Quantification to Visualization: A Taxonomy of Uncertainty Visualization Approaches.' In Dienstfrey, A. M. and Boisvert, R. F. (eds) *Uncertainty Quantification in Scientific Computing*. Berlin, Heidelberg: Springer. pp. 226-249.

Reuschel, A.K. and Hurni, L. (2011) 'Mapping Literature: Visualisation of Spatial Uncertainty in Fiction.' *The Cartographic Journal*, 48(4), pp 293-308.

Shneiderman, B. (1996) 'The eyes have it: a task by data type taxonomy for information visualizations.' In *Proceedings 1996 IEEE Symposium on Visual Languages*. September 1996, pp. 336-343.

Smithson, M. (1989) *Ignorance and Uncertainty: Emerging Paradigms*. New York: Springer-Verlag.

Stell, J.G. (2019) 'Qualitative Spatial Representation for the Humanities.' *International Journal of Humanities and Arts Computing*, 13(1-2), pp. 2-27.

Taylor, J.E., Donaldson, C.E., and Gregory, I. N., and Butler, J. O. (2018) 'Mapping Digitally, Mapping Deep: Exploring Digital Literary Geographies.' *Literary Geographies*, 4(1), pp. 10-19.

Thacker, A. (2005) 'The idea of a critical literary geography.' *New Formations*, 57, pp. 56-73.

Thomson, J., Hetzler, E., MacEachren, A., Gahegan, M. and Pavel, M. (2005) 'A typology for visualizing uncertainty.' *Visualization and Data Analysis 2005. International Society for Optics and Photonics*, 5669, pp. 146-157.

Wood, J., Isenberg, P., Isenberg, T., Dykes, J., Boukhelifa, N. and Slingsby, A. (2012) 'Sketchy Rendering for Information Visualization.' *IEEE Transactions on Visualization and Computer Graphics*, 18(12), pp. 2749-2758.

Yuill, R.S. (1971) 'The Standard Deviational Ellipse; An Updated Tool for Spatial Description.' *Geografiska Annaler: Series B, Human Geography*, 53(1), pp. 28-39.