Managing Uncertainty in the Humanities: Digital and Analogue Approaches

Jennifer Edmond*
Trinity Long Room Hub
Trinity College Dublin
College Green
Dublin 2, Ireland
edmondj@tcd.ie

ABSTRACT

This paper takes a high-level view of both the sources and status of uncertainty in humanities research and the attributes a digital system would ideally have. It draws upon both the experience of a number of digital projects and research into the many-faceted concept of uncertainty in data. Its intention is to support a dialogue between the humanities and computer science, able to realise the promise of digital humanities without a reversion to a new positivism in disciplines such as history and literary studies.

CCS CONCEPTS

• Information systems → Uncertainty • Information systems → Data analytics • Applied computing → Arts and humanities

KEYWORDS

Arts and Humanities Research; Digital Humanities; Research Processes; Uncertainty

1 Introduction

In the natural sciences, we are used to being able to speak of certain 'laws:' the law of gravity, the laws of thermodynamics, even some biological phenomena adhere to what seem to be predictable laws. As with the laws of human judicial systems, breaking these 'laws' extracts a 'penalty,' such the enormous amount of fuel required to lift an aircraft from the ground, or the death of a human cell in an inhospitable environment. What gives these 'laws' their status as such, therefore, is the matter of their predictability, reproducibility and stability over time.

The social and human sciences, and in particular the humanities, have no such verifiable, durable laws. These disciplines may be based upon 'facts,' such as, for example, a letter having indeed been delivered to a certain person on a certain day, an event for which there may be incontravertible evidence and multiple corroborating witness statements. What the humanities lack, however, is 'truths,' transcendent statements of confidence that might allow the meaning of such facts to be reliably and repeatedly interpreted in the exact same way. What these disciplines have instead are the ability to build well-grounded interpretations, although even these may well expose the unconscious biases of their authors, in particular after a time step of a number of decades. Influential in the recognition of this was Francois Lyotard, who, in his seminal work *The Postmodern Condition* [9], argued that scientific knowledge is *"*a kind of discourse.*"* Humanistic scholarship, at its best, gives us the durable philosophical reflections of Plato or Aristotle, but sometimes our insights are also all too evidently a product of the social and cultural norms of the place and time in which we produce them, and it is one of the most central skills of the humanist to recognise their own biases and those of their place and time, and seek to transcend them. For this reason, to return to the metaphor of law, humanistic scholarship tends to work on the basis of a 'preponderance of evidence' more than ever really being 'beyond doubt.'

Because of these contingencies, the place and nature of uncertainty in humanities research is also different from in the natural sciences. This paper will therefore look at how this differentiation between humanistic 'fact' and interpretation shapes the nature of humanistic research questions and attitudes toward sources in the face of such uncertainty. It will look at this phenomenon first in its analogue manifestations, then in the context of digital tools and data-driven research, thereby exposing some of the challenges inherent in designing digital systems to reduce uncertainty in research in the humanities. The frame of reference will be largely drawn from the disciplines of history and literary studies (for the humanities themselves are highly heterogenous), though the conclusions will be applicable across many such qualitative approaches to understanding the artefacts of human culture and creativity.

2 The Nature of Humanities Research Questions and Processes

*Jennifer Edmond is Associate Professor of Digital Humanities

TEEM'18, 24-26 October, 2018, Salamanca SPAIN

Humanities research questions can be quite complex, with multiple parts, and are mostly unable to be answered adequately though a single source of information. As a part of developing its portal and Virtual Research Environment, the CENDARI Project produced a very instructive set of use cases, featuring research questions from medieval and modern history [3]. Some of the questions explored in that report are as follows:

> *"My project examines how the rural-urban divide shaped Habsburg Austrian society's experience of the war from about 1915"*

> *"I want to investigate the relationship between the Bec-Hellouin Abbey, in Normandy, and other monasteries, priory, archbishopric and the kingdom of England, from its foundation to the XV century."*

> *"I wish to examine the ways in which the Ottoman empire and Islam were perceived by the political (liberal and Catholic) elites in the Slovenian lands of the Habsburg monarchy in the decade before the outbreak of the First World War."*

Eef Masson makes the useful observation that humanities scholars "do not seek to establish unassailable, objective truths" and "instead […] approach their objects of study from interpretive and critical perspectives, acting in the assumption that in doing so they necessarily also preconstitute them," [10] a statement which is borne out by these examples. None of these questions has a straightforward, factual answer: indeed, each of them proposes a multifaceted investigation of an issue, perception or relationship that, even at the time of its happening, would have been complex to explain. These questions, in other words, are not only rife with uncertainty, but suffused with and dependent upon it. One can imagine factual layers that could be recruited to support these investigations, such as correspondence flows in to and out of the Abbey of Bec-Hellouin or records of trade relationships between the agricultural heartlands of the Austro-Hungarian Empire and its great cities, such as Vienna and Budapest. But even from and within these, selections and interpretations will need to be constantly made.

Use of such source materials will often be opportunistic, as there is no one pathway toward the resolution of these questions, and none that is not almost assuredly partial, unreliable or biased. Knowing how to judge the provenance and authority of your sources is an essential part of the formation of the historian, a necessary and complementary skill to the kinds of questioning the above examples demonstrate.

3 Managing Uncertainty in the Humanities

The breadth of such humanistic research questions leaves a great latitude for encountering, incorporating and managing uncertainty. Not all uncertainty is equally present in the sources, however, or has an equivalent impact on the ability to produce research. As Petersen writes, *"*Uncertainty takes many forms, whether epistemic, statistical, methodological, or sociocultural. *"* [14] Outside of the humanities, many models and tools have been developed, either for capturing uncertainty in data (such as Petersen's *"*Uncertainty Matrix*"* [14]) or for capturing aspects of processing that could introduce uncertainty (such as the NASA EODIS [12]). These would, however, be cumbersome to adapt to and apply in the humanities.

Epistemically closer to home, Kouw, Van Den Heuvel, and Scharnhorst [7] cite Brugnach et al's [2] idea of a relational concept of uncertainty, in which three possible sources of uncertainty exist: unpredictable systems, incomplete knowledge of a system, or incompatible frames of reference for the system. [7] As a theoretical basis for exploring data and uncertainty in the humanities, this is a powerful model, but somewhat removed from the wide variety of observable practices in humanistic disciplines. From a more example-driven perspective, therefore, sources of uncertainty encountered by the humanist could be characterised instead to include the following:

- **Inherent uncertainty :** Somebody did something, or something happened - why did they do this? Why did it happen? Is this mark in the manuscript a doodle or a representation of a face? What was the inspiration behind this author's use of this word, this image?
- **Partial, missing, perspective-limited or conflicting Information**. I have a letter, I know who received it, but who wrote it? I know the age of an object, where did it come from? I know a document is from 1944 (or, more commonly, '5th Century,' 'medieval era,' 'around 1650' etc.) but what specific date? One account claims 20 people were killed in the skirmish, another account claims 200, which is accurate? Was this story written before or after the author heard about a particular event?
- **Errors, especially in cataloguing, but also in interpretation.** I don't know where this document came from, is it in the right place/box? The date or origin given for this object in the finding aid or catalogue record does not make sense to me as an expert, what is the source of this gap? By revisiting evidence, I can see that a medical diagnosis made decades earlier was probably wrong, or that an earlier interpretation was based on biased or incomplete records.
- **Bias**. I can verify a statement was made and by whom, but how can I verify the veracity or intention of the speaker? I am working with a collection that supports a particular conclusion, but is there material excluded here (intentionally or unintentionally) that would contradict it? This woman's writing (which we no longer have) is described as inferior by a male contemporary, but was that an aesthetic judgement or a gender-based one?
- **Sense of something wrong, without knowing why** (an opposing force to the 'happy accident' of discovery that is serendipity). The number of items found in a digital search seems to low, but I can't be sure what the source of the problem is. I know a certain object in in this collection, why can't I find it in the catalogue? This visualisation (eg. in a GIS) doesn't match my tacit understanding of a phenomenon. The author writes that his

intention was to portray a character in a certain way, but that does not match my own interpretation.

Humanistic data streams (defined here as the sources and other inputs that are used to inspire questions and build interpretations) are comprised of with these kinds of ambiguous, contradictory, 'messy' attributes. The need to verify a single individual fact is a subordinate task for the historian, whose method embraces a much broader uncertainty. Uncertain data (defined as data whose meaning is unresolved or unresolvable) would in many fields be epistemically 'off limits,' an unusable and unstable ground for any conclusions to be made. This, however, is the norm and not the exception in the humanities, and any humanist should be well able to isolate and build around, either via direct proxy sources or other corroborative material, those aspects of a useful source that is also somehow flawed.

4 Humanities, Uncertainty and the Digital

The powerful training a humanist receives for managing uncertainty in their source material does not necessarily translate well to digital or data-driven environments. In particular, exchanging a set of diverse and varied sources for a homogenous corpus of 'data' that cannot be surveyed and 'seen' in the same way, is a challenging shift. Kouw, Van Den Heuvel, and Scharnhorst in particular acknowledge this "highly ambiguous meaning of data in the humanities" [7], a position that Christine Borgman advances in her conjecture that "…[b]ecause almost anything can be used as evidence of human activity, it is extremely difficult to set boundaries on what are and are not potential sources of data for humanities scholarship." [1]

But it is not just words that are being shifted as the humanities move from sources to data, it is methods and values as well. As Masson describes it: "with the introduction of digital research tools, and tools for data research specifically, humanistic scholarship seems to get increasingly indebted to positivist traditions. For one, this is because those tools, more often than not, are borrowed from disciplines centred on the analysis of empirical, usually quantitative data. Inevitably, then, they incorporate the epistemic traditions they derive from." [10]

As we have seen from the examples given above, positivism is not an approach currently favoured in historical or literary scholarship and indeed is rather discredited in these disciplines. But the push toward a sort of 'new positivism,' arising not from a research or epistemic cultural imperative so much as an opportunistic one based on tool availability, cannot be ignored. Christine Borgman describes the challenge of the humanist using a digital tool not necessarily developed for her accustomed mode of questioning as follows: "they are caught in the quandary of adapting their methods to the tools versus adapting the tools to their methods. New tools lead to new representations and interpretations." [1] Digital humanities should, by all means, open up the way to new interpretations (which must be informed, of course, by an understanding of the function of tools) but it should also be able to resolve this quandary state by moving the tool to the user, drawing strength from the different perspective the humanist brings to the use of quantitative approaches, rather than resisting them.

5 Productive and Unproductive Management of Uncertainty in Humanities Research

By and large, humanistic researchers are not looking for tools that change what the study or how they undertake their investigations, so much as an enhancement of and supplement to their already heterogenous sources and adaptable methods. But this gap can be quite broad, with humanists viewing such widely accepted practices of data-driven research as 'data cleaning' or 'data scrubbing' with great suspicion, viewing these activities in the much more negative light of 'data manipulation' [6]

They have good reason to be suspicious. In his analysis of the process by which the metadata was created for the massive oral history collection of the Shoah Visual History Archive, Todd Presner speaks of the potential algorithmic approaches seem to have to be 'ethical,' that is neither losing the suffering of the individual in the masses, nor focussing only on particular well-known stories, such as Anne Frank's or Elie Wiesel's. And yet, as Presner shows, this apparent ethical viewpoint is deeply flawed when it comes to representing uncertainty, leading to the tagging of some material as 'indeterminate data' or 'non-indexable content.' Presner's account of this is worth quoting in full, as it not only raises a number of concerning ethical issues, but also highlights the real difficulties facing researchers when it comes to encoding particularly difficult or rich materials into a dataset:

"'Indeterminate data' such as 'non-indexable content,' must be given either a null value or not represented at all. How would emotion, for example, need to be represented to allow database queries? While certain feelings, such as helplessness, fear, abandonment, and attitudes, are tagged in the database, it would be challenging to mark-up emotion into a set of tables and parse it according to inheritance structures (sadness, happiness, fear, and so forth, all of which are different kinds of emotions), associative relationships (such as happiness linked to liberation, or tears to sadness and loss), and quantifiable degrees of intensity and expressiveness: weeping gently (1), crying (2), sobbing (3), bawling (4), inconsolable (5). While we can quickly unpack the absurdity (not to mention the insensitivity) of such a pursuit, there are precedents for quantified approaches to cataloguing trauma [...] Needless to say, databases can only accommodate unambiguous enumeration, clear attributes, and definitive data values; *everything else is not in the database.* The point here is not to build a bigger, better, more totalizing database but that database as a genre always reaches its limits precisely at the limits of the data collected (or extracted, or indexed, or variously marked up) and the relationships that govern these data. We need narrative to interpret, understand, and make sense of data." [15]

Presner gestures towards a complete rethinking of the database as a genre: specifically regarding representations of 'the indeterminate.' "Such a notion of the archive specifically disavows the finality of interpretation, relishes in ambiguity, and constantly situates and resituates knowledge through varying perspectives, indeterminacy, and differential ontologies." [15]

Perhaps, if we are to realise the potential of the digital humanities, this is the direction in which we must look?

Knowledge organisation frameworks and their associated tools, like metadata standards, taxonomies, controlled vocabularies and ontologies, have all provided powerful frameworks to increase our ability to connect and find information. They also, however, reduce the complexity of the information around a particular object and its digital surrogate, stripping it of its original context and its provenance. Ironically, therefore, the way in which cultural data, in particular historical data, is prepared may well increase its findability, but reduce its usability. I would therefore propose that systems looking to reduce uncertainty for the humanist might focus on the following measures:

- provide access to context and provenance.

As Kouw, Van Den Heuvel, and Scharnhorst state : "Metadata provide context, but the question of whose context is particularly contentious in the humanities." [7] Systems that capture and make it possible to explore the provenance of data streams, the variety and richness of its contexts, who has contributed to them, and what they may have been produced in proximity to could greatly enhance a researcher's ability to overcome inherent weaknesses in a source.

- don't focus on an unrealistic 'single source' model. Researchers not trained in the humanities may assume that deeply interrogating a single source is a norm for the humanities, as it may be in other disciplines. While this does occur, knowledge of that single source must always be supported by corroborating evidence from elsewhere. While there are some emerging examples of powerful single source corpora for humanistic research (such as historic newspapers, social media, or parliamentary records), the humanists will still long maintain an inconvenient tendency to: "[draw] on all imaginable sources of evidence," [1] a fact that should not just be accepted, but celebrated.

- focus on interoperability and 'comparative legibility' in corroborating sources. This is a corollary to the item above: if a single source will never be enough, then finding new ways to move fluidly between sources would be the far greater gain. This does not mean pulling all data into a universal federated information bank, a process that would inevitable lose context and flatten complexity. Rather, the goal would be to enable sources that are siloed to be combined and compared more easily than they are now, that is, more easily than as a linearly accessed succession of searches operating in different environments with different affordances and norms of interaction in each. One might think of the Orange tool chain platform [13] as an inspiration for this kind of linked, rather than siloed or merged, experience.

- provide a 'fuzzy search' that can reduce false negatives, such as is incorporated in the excellent interface of the Transkribus handwriting recognition tool. [17] While such a capacity will not solve all of the problems uncertainty about data might instil, it will at least promote an interrogability that may increase confidence.

- Interrogability of processing must also become more the norm. In a universe where a majority of humanistic sources were textual, a methodological source (such as a work of critical theory) could be read, evaluated and then used or discarded. Only the foolhardy scholar would attempt to use a source s/he had not read. And yet, in the digital age, such equivalent tools for framing arguments and approaches, from topic modelling to stylometric tools, do not always explicitly expose their lines of argument, their thought processes. Instead, they often seem to run the risk of promoting the maxim 'garbage in, gospel out,' leaving the user who may not be aware of the limitations of the tool to accept the authoritative voice of its output. DARPA's research into "Explainable AI" [4] points in a direction that could provide models for this, in spite of the additional cost and limitations such an approach may place on the technology deployed.

- enable trust. Digital tools will speed up some aspects of a humanities researcher's process, but others aspects will almost certainly defy interrogation by digital methods. According to Tenopir et al.'s substantive report on trust and authority in scholarly communications, the top criteria scholars used to judge their sources were: "criteria … associated with personal perusal and knowledge, the credibility of the data and the logic of the argument and content." [16] All of these processes are ones that the digital presentation of source material has the potential to impede, by reducing perusability, removing context, restructuring an existing logic framework, or indeed presenting data stripped of the interpretive narrative meant to accompany it. To disrupt these elements is to disrupt the internalised, tacit verification system of the humanist, without which, sources and tools are of no use at all.

- explore embodied practices. Part of the strength of the humanistic research process, and its adaptation to the heterogenous and uncertain sources it relies upon, comes from its multimodality. The embodied elements of humanistic research practices are highly complex, enabling a very subtle management of time and space, of kinds of knowledge and of complementary sources, which is antithetical to work on a single platform or device. A better match with these strategies would create a far more fluid relationship between their needs and digital tools and environments.

- finally, and most importantly, **don't try to remove uncertainty, but signal where it is.** Humanists will never have certainty, because the sources, and the humans who created them, are flawed. Because of this, honing a human instrument able to draw conclusions under these circumstances is a value and a process humanists hold dear. There are many things a researcher has to learn to deal with just by 'slogging through' them, this is a part of the discovery and learning process. But, properly deployed, the digital can contribute a lot to what a humanist does with the uncertainty they have, and how they move toward a greater and better-grounded confidence in their interpretations.

6 Conclusions

I have long been inspired on the work of historians on the phenomenon known as epistemicide, the systematic marginalisation to the point of extinction of certain ways of

creating knowledge, which was particularly pronounced in at the height of long 16th Century, with it's many examples of colonial expansion activity. [5] In our digital, quantitative age, with its keyword searches, artificial intelligences and statistical profiling algorithms, I worry we may be facing into another great wave of this same phenomenon. Bruno Latour seems to harbour the same fear, proposing that we need to: "...recalibrate, or realign, knowledge with uncertainty, and thereby remain open to a productive disruptive aspect of uncertainty." [18] As Kouw, Van Den Heuvel, and Scharnhorst point out, "uncertainty is often explained as a lack of knowledge, or as an aspect of knowledge that implies a degree of unknowability. Such interpretations can result in commitments to acquire more information about a particular situation, system, or phenomenon, with the hope of avoiding further surprises." [7] But for the humanist, the joy of discovery, and of reaching across a gulf of time and text to connect with others, is a surprise one could never wish away. The authors of this passage continue to say that we need to understand and appreciate: "how uncertainty can be a source of knowledge that can disrupt categories that provide epistemological bearing." [7] If our attempts to assist researchers to manage uncertainty with digital tools are to succeed, we must be ever mindful of this.

Historian Christina Larner sounded a warning bell as early as 1984 that: "Inadequate data [does] not become scientific information simply by virtue of being processed through a computer." [8] This does not mean, however, that uncertain sources cannot be made into knowledge with the assistance of a computer, however, and it is toward this goal we must strive.

ACKNOWLEDGMENTS

This research in a partial result of work in the PROVIDEDH project, funded within the CHIST-ERA programme under the national grant agreement: of the Irish Research Council. It owes a significant debt to some of the unpublished work of Dr Georgina Nugent Folan produced in the Knowledge Complexity Project (KPLEX), funded by the European Commission under grant agreement 732340.

REFERENCES

[1] Borgman, C. 2017. Big Data, Little Data, No Data. MIT Press

[2] Brugnach M. et al 2008. Toward a Relational Concept of Uncertainty : about Knowing Too Little, Knowing Too Differently, and Accepting Not to Know. Ecology & Society 13.2.30.

[3] CENDARI Project team, 2016. Domain Use Cases. Accessible at : http://www.cendari.eu/sites/default/files/CENDARI_D4.2%20Domain%20Use%20Cases%20final.pdf

[4] Gunning, D. Explainable Artificial Intelligence (XAI). Accessible at : https://www.darpa.mil/program/explainable-artificial-intelligence

[5] Hall, B. 2015. Beyond Epistemicide : Knowledge Democracy and Higher Education. Accessible at : http://hdl.handle.net/1828/6692.

[6] Kouw, M.,Van Den Heuvel, C. and Scharnhorst, A.. 2013. Exploring Uncertainty in Knowledge Representations: Classifications, Simulatons, and Models of the World in Wouters et al., eds., Virtual Knowledge: Experimenting in the Humanities and the Social Sciences. MIT Press

[7] Larner, C. ed., 1984. Witchcraft and Religion. The Politics of Popular Belief. Oxford.

[8] Lyotard, F. 1979. The Postmodern Condition : A Report on Knowledge. University of Minnesota Press.

[9] Masson, E. 2017. Humanistic Data Research An Encounter between Epistemic Traditions in Mirko Tobias Schäfer and Karin van Es, The Datafied Society. Studying Culture through Data,

[10] NASA's Earth Observing System Data Information System (EOS DIS). n.d., accessible at: https://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products.

[11] Orange, https://orange.biolab.si/

[12] Petersen, A. 2006. Simulating Nature: A Philosophical Study of Computer-Simulation Uncertainties and their Role in Climate Science and Policy Advice, Uitgeverji Maklu (2006), reprinted in Wouters et al., *Virtual Knowledge*, MIT Press.

[13] Presner, T. 2015. Probing the Ethics of Holocaust Culture, in Fogu, Kansteiner, and Presner, History Unlimited. Harvard University Press. Accessible at : http://www.hup.harvard.edu/catalog.php?isbn=9780674970519.

[14] Tenopir, C. et. al. 2013. Trust and Authority in Scholarly Communications in the Light of the Digital Transition, Final Report. University of Tennessee & CIBER Research Ltd.

[15] Transkribus, https://transkribus.eu/Transkribus/

[16] Wouters P. et al., eds., *Virtual Knowledge: Experimenting in the Humanities and the Social Sciences* (MIT Press, 2013), 95, http://www.jstor.org/stable/j.ctt5vjrxn.

[17] Huber, E, Lehmann, J and Stodulka, T. 2018.Report on Data, Knowledge Organisation and Epistemics. Accessible at: https://kplexproject.files.wordpress.com/2018/06/k-plex_wp4_report-data-knowledge-organisation-epistemics.pdf