From Facts to Judgments: Theorizing History for Information Science

by Ryan Shaw

There is increasing interest in representing the past as a database of historical facts. Drawing upon the increasing availability of digitized historical texts and advances in text mining and semi-structured databases, these projects seem set to fulfill Paul Otlet's dream of extracting the factual content from texts and making it available to answer queries about the past.

For example, the academic project *DBpedia* [1] aims to extract facts from *Wikipedia* infoboxes – the sections found on certain categories of *Wikipedia* articles that present basic facts using a standardized template. For example, the "Historical Event" infobox template includes fields for the event's preferred name, alternate names, date, location, participants and result, as well as a representative image. On a *Wikipedia* article such as "French Revolution," these values are presented as an HTML table. After parsing and processing by the *DBpedia* project's algorithms, these values are transformed into a standardized data model for representing subject-predicate-object expressions (for example, "French_Revolution : location : France"). *Freebase* is a commercial service that similarly parses *Wikipedia* infoboxes into structured data, but which has the more ambitious goal of integrating this data with all other available public domain data and furthermore providing interfaces for editing and adding to it.

Projects like *DBpedia* and *Freebase* mine historical facts primarily from the riches of *Wikipedia*, hoping that the collaborative effort there will trickle into their own projects. Other projects aim to mine the web at large for historical knowledge. Bruce Robertson calls the web "a historian's fantasy" and envisions sophisticated tools for organizing and querying not only encyclopedia articles but also digitized archives, scholarly editions, journal articles and blog posts [2]. Pursuing a similar vision, digital historian Dan Cohen and programmer Simon Kornblith have developed a system called *H-Bot* [3] that parses *Google* search results to answer historical questions. And *Google* itself has begun incorporating timeline results – historical facts mined from the web – into its search results, as a search for a historical personage or event will usually show. All of these systems might be considered descendants of IBM's "Professor RAMAC" computer program, which at the 1958 World's Fair impressed audiences with its ability to answer historical questions using its "stack of 50 fast spinning disks" on which were stored "the principal historical events of the world from the birth of Christ to the launching of Sputnik I" [4].

H-Bot claims to demonstrate the "automatic discovery of historical knowledge" by using *Google* search results to answer simple factual questions such as "When was George Washington born?" or "Who was Lao-Tse?" The program uses simple sentence parsing techniques to transform questions into keyword searches that are passed to *Google*. It then uses statistical techniques to extract frequently repeated information such as dates and names from the returned web pages. (In practice, however, *H-Bot* usually draws its answers from a handful of online reference sources such as *Wikipedia* or *Wordnet*.) With its focus on simple factoids, its creators admit that *H-Bot* "offers an impoverished view of the past." But they divert attention from this critique by implicitly equating such fact-finding with historical knowledge and by promising greater things to come (a favorite technique of AI researchers for decades). Most tellingly, however, they argue that the profusion of historical evidence surviving from the recent past poses a problem for historians, who must adopt the same techniques for managing information overload that scientists use.

Ryan Shaw is a Ph.D. candidate at the School of Information at the University of California, Berkeley. He can be reached at ryanshaw<at>ischool.berkeley.edu.

This rhetoric should be familiar to information scientists. From Paul Otlet to Vannevar Bush to the National Science Foundation (NSF)-funded projects of today, much research in information management and retrieval has focused on the needs of scientists and specifically on how to help scientists avoid reading. The problem as it has been framed by this research is that scientists must stay current with ever more scientific literature in a finite amount of time. This dilemma has led to a focus on text summarization, filtering of irrelevant information and extraction of key facts from explanatory narrative. With the advent of large-scale corpuses of digitized texts, these techniques are now being proffered for the humanities as well. The call for proposals from a recent *Digging into Data* program funded by the National Endowment for the Humanities (NEH), NSF and others to develop data-mining tools and techniques for humanist scholars exemplifies how the problem of too much text is being framed (emphasis added by this author): "Now that scholars have access to huge repositories of digitized data – *far more than they could read in a lifetime* – what does that mean for research?" Databases of historical facts are in part a response to this perceived problem: when there are too many histories to be read, boil them down to bare facts that can be subjected to powerful selection and retrieval mechanisms.

Such approaches are better suited to the sciences, given that scientists are assumed to be engaged in a cumulative research effort in which later researchers build upon the work of earlier researchers. This ideal of cumulative research effort requires that the complexity of earlier work be distilled down to reusable conclusions or facts. Bruno Latour in his *Science in Action* [5, 1-17] famously describes this process of fact production as "black-boxing." A black box is a metaphor for a mechanical or computational component that is used to fulfill a functional requirement without knowledge of its internal implementation. While it may be possible to know how the black box works, all that is relevant to its users is that it produces expected outputs from given inputs. While scientists can in principle go back and recreate the experiments of their predecessors, efficient cumulative research requires that most of the time they simply trust that the facts they inherit work as advertised.

Yet as Louis Mink [6] points out, work in history does not produce "detachable" conclusions of this kind. It is rare for historians to simply accept an earlier account of some historical subject. Re-examination of primary evidence is the rule. Mink argues that this practice is not common because historiography is less developed than the sciences, but instead is due to a difference in the nature of the conclusions that historians produce. Rather than simply producing facts about the past, the historian aims to produce what Mink calls "synoptic judgments" of some complex of actions and events in the past.

A synoptic judgment is an interpretive act in which one moves from seeing that a series of things happened (the facts about the past) to seeing those happenings as a synthetic whole. Once she has reached such a judgment herself after immersion in the historical evidence, the historian's task is to lead others through the interpretive process via the medium of a written text. Through the techniques of narrative representation, the author of a historical text aims to show past actions and events as a coherent whole when seen from a certain perspective. The exhibition of this whole as represented by the thick description of the historian's narrative is the conclusion and as such cannot be detached from that narrative. In other words, the historian's conclusions inhere in the structure and organization of her narrative. Even when the historian summarizes her narrative in separate statements, Mink argues, these statements are not detached conclusions but simply reminders to the reader of how the historian has ordered and organized her true conclusion, the narrative itself.

Databases of historical facts purport to help us answer questions about the past, and in a narrow sense they do that. But few of these systems take us further than the initial "Hey, neat!" reaction inspired by Professor RAMAC. The problem is not that the facts are wrong – Cohen and Rosenzweig [3] show that, on the contrary, they can be quite accurate by most standards – or that they are incomplete (though they certainly are). Nor is insufficiently advanced technology to blame – even if we were able to perfectly extract historical facts from texts, disambiguating every name and indexing each fact in the absolute grid of time and space, we would still face this problem. The problem is that systems like this are grounded in an impoverished conception of how we represent the historical past, a conception that focuses on atomic facts rather than synoptic judgments.

The problem is an old one. It can seem obvious that what we need to understand the past are facts about the past, and that a perfect history is thus one that identifies and enumerates "everything that happened" in terms of such facts. Yet upon careful consideration these notions are quite problematic. Philosophers have often hypothesized the idea of a complete historical database in order to demonstrate what the problems are. Arthur Danto imagines an "Ideal Chronicle" with descriptions of "absolutely everything that happened," in the order it happened, thus providing the "whole map of the Past" [7, 148-181]. Danto argues that even such a complete database of the past would not obviate the need for historiography, since the role of historians is not simply to recount factual data about the past, but to represent the significance of episodes in the past from the perspectives of the present. These perspectives (and thus our criteria for significance) are constantly changing. It is this kind of change, not simply the discovery or refutation of historical evidence (addition or deletion of facts from the database) that results in new historiographical conclusions. Or, as Mink puts it, even if we could "sit before a screen and directly review the past in its minutest details," we would still need some imaginative representation of the past to help us make sense of it all in light of our current historical situation, and it is the role of history to develop such imaginative representations and not simply gather the detailed facts.

Acts of synoptic judgment produce imaginative representations that are articulated as historical narratives. Once historians have developed narratives that relate sets of facts under some synthesizing ideas, they usually label these narratives with phrases like "The Renaissance" or "The French Revolution." The philosopher of history W. H. Walsh calls this process *colligation*, appropriating the philosopher of science William Whewell's term for "the binding together or connection of a number of isolated facts by a suitable general conception or hypothesis" [8, 59-64]. Walsh supports a hermeneutic conception of historical method in which the goal of historians is to imaginatively and empathetically reconstruct the experiences and thoughts of people in the past. Accordingly, Walsh's notion of colligation initially depended upon happenings being intrinsically connected by virtue of being intentions or consequences of some past actor's plans. A "suitable general conception" for binding together facts was one that illuminated those facts as being part of a (conscious or unconscious) policy guiding the behavior of people in the past. Later Walsh expanded his notion of colligation to include any case in which some set of events as is interpreted as a connected process or development, whether or not such a policy could be discerned [9].

Even though concepts are of primary interest in library and information studies, colligatory concepts have been mostly overlooked. Even the most sophisticated theoretical discussions of concepts in the literature tend to equate concepts with classes or categories. For example in his recent survey of concept theories Hjørland [10, p. 1522] asserts that "[c]oncepts are dynamically constructed and collectively negotiated meanings that *classify* the world according to interests and theories" (emphasis added). This preoccupation with classification is perhaps understandable in light of the aforementioned focus on scientific domains. The sciences seek to abstract away from unique individuals to generalized classes that can be related by laws. While historians do generalize, they also – arguably primarily – seek to assemble descriptions of unique past events into connected and coherent but no less unique representations. Concepts like "The Renaissance" colligate rather than classify.

The most fully developed theory of colligation to date has been developed by Frank Ankersmit, who in his *Narrative Logic: A Semantic Analysis of Historian's Language* [11] seeks to explain how colligatory concepts – which he calls "narrative substances" – are constructed from sets of statements expressing facts. A narrative substance is a point of view from which to regard the past, articulated by means of a specific historical narrative. Ankersmit contends that each individual historiographical narrative constructs a narrative substance so that, for example, there are as many "Renaissances" as there are narratives on the subject, since each narrative articulates a specific point of view. So when we speak generally about "The Renaissance," we are really talking about a whole family or type of narrative substances that have been given the same name.

Furthermore, Ankersmit claims that when we define such types, we do so *extensionally* rather than *intensionally*. An intensional definition of a type is one that defines some necessary and sufficient conditions for belonging to the type. For example, one might define a *mug* intensionally as "a type of cup made of glass or ceramic and having a handle large enough to accommodate a whole hand." An extensional definition of a type, on the other hand, enumerates the members of a set of individuals considered to be instances of that type. An extensional definition of *mug* would collect all the world's individual coffee mugs and beer steins and so on and thereby declare "these are mugs."

Ankersmit argues that one can define types of narrative substances extensionally by clustering narrative substances that contain overlapping sets of statements. He proposes a thought experiment in which a giant matrix is constructed. Along one axis of the matrix are aligned all the declarative statements made about the past that have actually appeared in some text or another. Along the other axis are aligned all the narrative substances constructed by means of those statements. Each cell in the matrix is then filled with a "0" or a "1" indicating whether or not the corresponding statement was used to help construct the corresponding narrative substance. Given such a matrix, we could then try to identify types of narrative substances by grouping together narrative substances with similar patterns of 0s and 1s, in much the same way that we might identify types of drinking vessels by looking for similar shapes or handles or materials. Ankersmit posits that we will observe that "certain classificatory patterns automatically appear." These clusters in "narrative space" reflect the fact that historians write in response to other historians and construct their narrative substances by distinguishing them from those that came before (which implies a significant degree of overlap).

As Ankersmit points out, such an extensional procedure for identifying types can never be precise. Depending on how we interpret similar patterns there will be many possible groupings into types. Moreover, for any given interpretation of similarity, there will always be boundary cases that could belong to more than one cluster. At best, extensional typification can identify regularities in how we have chosen to conceptualize reality, but it cannot tell us anything about reality itself. In other words, looking at written history this way tells us something not about the reality of the past, but about the contours of the concepts developed by historians over time. "The Renaissance," "The Cold War," "The French Revolution" and "9/11" are not objectively existing entities in the past, but are names of types of stories we tell to understand the past.

Finally, Ankersmit argues that an extensional procedure like this is the only way to identify types of narrative substances. The alternative would be to intensionally identify types in terms of logical definitions based on attributes of the things being classified, the way we define the type *mammal* as "warm-blooded," "vertebrate" and "having hair or fur." But this intensional identification is precisely what we cannot do for narrative substances. There is no logical definition, no core set of properties both necessary and sufficient for making a particular narrative a narrative about "The French Revolution." While it's easy to identify statements that would not appear in any narrative of the French Revolution – for example that the storming of the Bastille occurred in 1967 in Tokyo, Japan – we cannot identify statements that *must* appear in such stories or by virtue of which we *must* consider a given narrative to be a narrative about the French Revolution. Given all the narratives that have ever been written about the French Revolution, we may not be able to identify a single statement that appears in every one. Thus we cannot and do not identify types of narrative substances intensionally. Decisions about what "The Cold War" is can only be justified pragmatically, not logically.

With the large-scale digitization of books it may become possible to investigate Ankersmit's theory by analyzing the full texts of historical narratives. Before undertaking such a project we must address some methodological problems. First is the issue of how to identify statements about the past. Ankersmit's matrix involves statements (also known as *propositions*) about the past, not the sentences that express these statements. (Ankersmit contends (p. 19) that, "states of affairs in the past can be unambiguously described by means of constative statements.") We cannot conflate statements with the sentences that appear in historical texts, as it is unlikely that any two texts will contain precisely the same sentence. So we must find a way to move from the sentences that appear in texts to the propositions those sentences express, a problem that has occupied linguists and philosophers of language since Bertrand Russell. Fortunately this is an area where information extraction technology might show its worth, as such technology is precisely concerned with transforming sentences into logical propositions. Despite well-known problems with propositional theories of language [12, 70-74] and the fact that information extraction technologies are plagued by errors, their output still might be usable for exploring Ankersmit's theory.

The second problem, as Ankersmit points out, is that a given historical text constructs multiple narrative substances, and it is difficult to determine exactly which statements are being used to construct which narrative substances. Indeed, Ankersmit argues (p.104) that in order to identify narrative substances reliably we must, "compare historiographical topics studied and discussed by generations of historians." Fortunately the catalogers who maintain the *Library of Congress Subject Headings* (*LCSH*) [13] have done this for us, creating subject headings for historiographical topics and assigning them to historical texts. Using the LCSH is not ideal for investigating narrative substances, however. Catalogers usually do not identify more than a couple of the narrative substances constructed in a given text. Since they are concerned with characterizing whole books, they will not identify historical narratives in books that are not primarily works of historiography. And since the goal is to collocate texts, there is no effort to distinguish differences among the narrative substances constructed by different texts. In essence, what catalogers have done is group various narrative substances into types a priori.

Notwithstanding these problems, it is plausible that one could use a historiographical subject heading to obtain a set of texts within which authors have constructed comparable narrative substances. Ankersmit suggests that we can refer to narrative substances with terminology such as "Renaissance$_1$," "Renaissance$_2$," "Renaissance$_3$," etc. where the subscript *n* indicates that we are referring to the specific representation of the Renaissance constructed in the historical text *n*. Likewise, we could posit that a set of *N* texts found under the *LCSH* for the "Renaissance" constitute a set of narrative substances "Renaissance$_1$," "Renaissance$_2$," "Renaissance$_3$"... "Renaissance$_n$." We could then compare the propositions made in these texts to each other to build a model of the extension of the type "Renaissance." This model could provide a basis for highlighting differences among the individual narratives constructed by the different texts. Though such a model would enable us to examine the structure of types already identified by the *LCSH*, it could not reveal new types not yet identified there. But we have to start somewhere.

Whether or not initial attempts to automatically discern and model them prove successful, it is time information science paid closer attention to colligatory concepts. In the digital environment, traditional components of the scholarly apparatus such as term lists, classification and categorization schemes, and thesauri are evolving into generalized semantic tools for enumerating and disambiguating concepts and mapping the relations among them [14]. Though in the past such tools have mainly been used for indexing and retrieval, in an era of full-text search I believe we will see other applications move to the fore. Specifically, semantic tools that map a given conceptual domain can be integrated into reading and writing environments to help users contextualize some fragment of interest. Given such applications, "fuzzy" concepts reflecting particular interpretive stances are at least as important as traditional categorical concepts. Ankersmit and his predecessors' theory of colligation provides grounds for investigation of such concepts. Done properly, such an investigation may lead to new tools (or new ways of using existing tools) for research that avoids reducing conceptions of the past to bare facts. ■

ACKNOWLEDGEMENTS – Support from the Institute of Museum and Library Services through award LG-06-06-0037-06 "Bringing Lives to Light: Biography in Context" and from the Advancing Knowledge grant PK-50027-07 "Context and Relationships: Ireland and Irish Studies," jointly funded by the National Endowment for the Humanities and the Institute of Museum and Library Services, is gratefully acknowledged.

RESOURCES on next page

Resources Mentioned in the Article

Website links

DBpedia: http://dbpedia.org
Digging into Data: www.diggingintodata.org/
Freebase: http://www.freebase.com
Google: www.google.com
H-Bot: http://chnm.gmu.edu/tools/h-bot
Wikipedia: www.wikipedia.org
Wordnet: http://wordnet.princeton.edu

Other Resources

[1] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009, September). DBpedia – A crystallization point for the Web of data. *Journal of Web Semantics 7*(3), 154-165. Retrieved October 21, 2009, from http://dx.doi.org/10.1016/j.websem.2009.07.002. Preprint available at www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/publications/Bizer-etal-DBpedia-CrystallizationPoint-JWS-Preprint.pdf.

[2] Robertson, B. (2009). Exploring historical RDF with Heml. *Digital Humanities Quarterly 3*(1). Retrieved October 21, 2009, from www.digitalhumanities.org/dhq/vol/3/1/000026.html.

[3] Cohen, D. J., & Rosenzweig, R. (2005, December 5). Web of lies? Historical knowledge on the Internet. *First Monday 10*(12). Retrieved October 21, 2009, from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1299.

[4] Morris, M. E. (1981). Professor RAMAC's tenure. *Datamation 27*(4), 195–198.

[5] Latour, B. (1987). *Science in action.* Cambridge, MA: Harvard University Press.

[6] Mink, L. O. (1966). The autonomy of historical understanding. *History and Theory 5*(1), 24–47. Retrieved October 21, 2009, from www.jstor.org/pss/2504434.

[7] Danto, A. C. (1985). *Narration and knowledge.* New York: Columbia University Press.

[8] Walsh, W. H. (1951). *An introduction to philosophy of history.* London: Hutchinson's University Library.

[9] Walsh, W.H. (1974). Colligatory concepts in history. In Gardiner, P. (Ed.), *The philosophy of history* (pp.127–144). Oxford: Oxford University Press.

[10] Hjørland, B. (2009). Concept theory. *Journal of the American Society for Information Science 60*(8), 1519-1536. Retrieved October 21, 2009, from http://dx.doi.org/10.1002/asi.21082.

[11] Ankersmit, F. R. (1983). *Narrative logic: A semantic analysis of the historian's language.* The Hague: Martinus Nijhoff.

[12] Lycan, W. G. (2008). *Philosophy of language: A contemporary introduction* (2nd ed.). New York: Routledge. 70–74.

[13] Library of Congress. *Library of Congress Subject Headings.* Information about editions and formats available at www.loc.gov/cds/lcsh.html. Online search of LC subject headings available at http://authorities.loc.gov/.

[14] Hjørland, B. (2007). Semantics and knowledge organization. *Annual Review of Information Science and Technology 41*, 367-405. Retrieved October 21, 2009, from http://dx.doi.org/10.1002/aris.2007.1440410115. Preprint available at http://dlist.sir.arizona.edu/2312/.