doi: http://dx.doi.org/10.23925/1980-7651.2018v21;p1-12

New proposals for organization of knowledge and their role in the development of databases for history of science

Ana M. Alfonso-Goldfarb; Silvia Waisse; Márcia H.M. Ferraz[·]

Abstract

As is known, the digital era brought new possibilities for creation, organization and work with large databases. However, some problems make such large databases difficult to manage, as e.g., lack of definite standards, or perhaps even full impossibility to develop any standard at all due to the complexity of the data. This certainly is the case of databases in the humanities, especially in regard to fields in which the aspect of temporality is constitutive. In this paper we discuss models for organization of knowledge, with particular attention to the modern fate of the traditional 'trees of knowledge', the emergence of decentered network models and new possibilities that emerged together with the digital humanities. We conclude with some considerations on very recent initiatives to solve the problem posed by temporality in large textual corpora.

Keywords

Databases; History of science; Digital humanities; Trees of knowledge; Networked knowledge; Digital environment

· Center Simão Mathias of Studies in History of Science (CESIMA), Pontifical Catholic University of São Paulo, Brazil. ✉ aagold@dialdata.com.br. The present paper derives from a presentation at symposium Doing History of Science in a Digital, Global, Networked Community: Tools and Services Linking Scholars and Scholarship, 25th International Congress of History of Science and Technology, Rio de Janeiro 22-29 July 2017.

Introduction

In this paper we would like to share some of our reflections on the state of the art in the theoretical-methodological discussions on organization of knowledge and its assimilation into the digital humanities. These reflections have, as a fact, a very long story. Our research center, CESIMA, was created more than 20 years ago, thus well before digital databases became the commonplace they are today. One of the reasons underlying the creation of CESIMA was precisely to develop a digital database to provide researchers in Brazil, and Latin America as a whole, access to documents for studies in history of science in any time and place. In the course of many years, tens of thousands of documents were acquired and processed. Finally the database content was ready. But unexpectedly, we faced a serious and seemingly unsolvable problem at the time of indexing and classifying the documents: none of the classification systems in use today seemed to apply to concepts which meaning varied substantially along more than 2,500 years.

While organization of knowledge is a subject we have paid continuous attention all along our work as historians of science, it gained particular momentum starting 2010, with the help of Brazilian research promotion agencies and national and international partners, among which we would like to mention in particular Research Group in Technology Applied to Education, Federal University of Rio Grande do Sul, Brazil (Grupo de Pesquisa em Tecnologia Aplicada à Educação - GTech.Edu/UFRGS) which developed the Sobek concept mining tool; the Committee of Bibliography and Documentation/International Union of History and Philosophy of Science/ Division of History of Science and Technology (CBD/IUHPS/DHST); and IsisCB board.

Our activities included active interaction with specialists in several fields, including bibliography, library science, information science and technology, among others. We performed an in-depth review of the historical, theoretical and methodological foundations of the main classification systems – Dewey Decimal (DDC), Library of Congress (LCC), Universal Decimal (UDC) and so forth. In time, S.R. Ranganathan's Colon Classification significantly awakened our interest, as its flexibility seemed to meet our needs. However, not even the elaborated theory developed by this distinguished Indian mathematician and documentalist succeeded in solving the problem of ***temporality*** –of paramount importance when dealing with the concepts found along work with documents for history of science.

Here we summarize the main results of our inquiry. First we address the modern fate of the traditional hierarchical approach to the classification of the sciences, i.e., the so-called 'trees of knowledge'. In the second part we focus on the possibilities afforded by the novel field of digital humanities. We conclude with some considerations on very recent initiatives to solve the problem of temporality in large text corpora.

Recreating the tree of knowledge

As is known, ever since antiquity the organization of knowledge was often represented through resource to the ***tree metaphor***, consisting of one single trunk, which in

orderly and hierarchical manner splits into several 'branches' of knowledge. Growing in complexity and increasingly more diagrammatic, this figure was meant to emphasize the allegedly unitary nature of knowledge, a belief held until late into the nineteenth century - despite the rise of pie charts in the 1800s, first within economics to make room for the increasing application of statistics.[1]

However, the exponential growth and diversification of science and technology in the twentieth century were attended by demands for less centrality, hierarchy and symmetry, and more interconnectivity and interdependence among the branches of the older trees. As a result, while the node-link model survived, the tree-like diagrams were thoroughly reshaped, especially from the end on the century onward. So for instance, in the 1990s, Ben Shneiderman observed that successive decompositions of a given field into multiple smaller elements evidences many different hierarchical levels previously hidden by the massive weight of the full structure. One result of this approach was the rectangular treemap, which in turn influenced the much-used circular treemaps, and more recently the beautiful Voronoi treemaps. In turn, the pie charts gave rise to the radial treemaps.[2]

Dozens of new tree-like diagrams are currently available, the recent development of which will doubtlessly be a source of pleasure for historians of science and technology. However, some scholars criticize the very figure of the tree, as they consider that a true leap in the organization of knowledge involves shifting from the hierarchical tree to the non-hierarchical network model. To add to the debate, other scholars believe that the network model is not really a substitute, but a variant of the tree-like diagrams.[3]

Yet, one should not lose from sight that the subject here is ***graphs***, i.e., structured diagrams which through hard mathematizing are very far from implying a rhizome-like and indeterministic (dis)organization of knowledge of the kind suggested by Gilles Deleuze and Felix Guattari.[4] While the trees gained in hierarchical flexibility, and networks increasingly provide more and more possibilities of coupling and extension, both remain as topological objects and thus as quantitatively, spatially and morphologically conceived.[5]

---

[1] See e.g., Ian Spence. "No Humble Pie: The Origins and Usage of a Statistical Chart," *Journal of Educational and Behavioral Statistics* 30, no. 4 (2005): 353–68; Michael Friendly, "The Golden Age of Statistical Graphics," *Statistical Science* 23, no. 4 (2008): 502–35.

[2] See e.g., Ben Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," in *IEEE Symposium on Visual Languages* (Los Alamitos [CA]: IEEE Computer Society Press, 1996), 336-43; Ben Shneiderman, Cody Dunne, Puneet Sharma, & Ping Wang, "Innovation Trajectories for Information Visualizations: Comparing Treemaps, Cone Trees, and Hyperbolic Trees," *Information Visualization* 11, no. 2 (2011): 87–105.

[3] On these various perspectives see e.g., Cathleen McGrath, David Krackhardt, & Jim Blythe "Visualizing Complexity in Networks: Seeing Both the Forest and the Trees," *Connections* 25, no. 1 (2003): 37-47; Lyn Robinson, & Mike Maguire, "The Rhizome and the Tree: Changing Metaphors for Information Organization," *Journal of Documentation* 66 (2010): 604-13.

[4] Gilles Deleuze, & Felix Guattari, *A Thousand Plateaus* (Minneapolis [MN]: University of Minnesota Press, 1987).

[5] On this comparison see Fulvio Mazzocchi, "Images of Thought and Their Relation to Classification: The Tree and the Net," *Knowledge Organization* 40, no. 6 (2013): 366-74, esp. 368-9. On new theoretical-methodological and nomenclature perspectives, see Ágota Fóris, "Network Theory and Terminology," *Knowledge Organization* 40, no. 6 (2013): 422-9. For further detail on the statistical work and differences between these two instances see e.g., Russell Lyons, & Yuval Peres, *Probability on Trees and Networks* (Cambridge [MA]: Cambridge University Press,

That the network model is occupying increasingly more space in the organization of knowledge is an undeniable fact. With their potential for both detailed and broad-scoped visualization and growing ability to detect several spatiotemporal nodes within one and the same mobile diagram, networks provide one of the best perspectives for analysis of the transformation of concepts over time, which it goes without saying is particularly relevant for history of science. Once again Shneiderman and colleagues played a substantial, if not pioneering role here, with the development of approaches to assimilate the aspect of temporality into the network model – even though they explicitly assert their work is just incipient and that much still remains to be done to achieve the results many among us would like to see.[6] The problems these authors mention are neither few nor new, and are related to semantic substrates required for identification of temporality. According to these authors, semantic substrates are composed of several high-complexity layers, which in the case of history – the so-called 'historiographs' – range from linguistic and philological to social, cultural and economic changes. Moreover, they imply domain properties which are not free from user interference, and thus are infinitely more complex than the well-known structural properties assumed in digital networks from their inception. Incidentally, the former are considered to be independent from the latter, and thus from the very structure of a given network. The outcomes of this type of studies are still sparsely published, and depend considerably on the collaboration of users, as is also our case.[7]

It is thus not difficult to understand why network-based approaches made fast advance and were quickly assimilated into the digital world for fields in which history seldom has a central part. Such is, e.g., the case of the social sciences, in which network social analysis (NSA) was long under way and fit like a glove when the global Internet revolution exploded. Similarly, it is neither difficult to understand why an early paper by Linton Freeman – one of the first advocates of digital NSA – became mandatory reading for anyone involved in the field initially known as 'humanities computing.'[8]

Problems and shortcomings notwithstanding, substantial work in fields in which history does have the central place – like Franco Moretti's study on literary history – has already contributed to the shift of the former 'humanities computing' into something more properly called 'digital humanities.' Namely, a new field (although already with a considerable history) in which the humanities ***interact with*** and ***contribute to*** the digital

2016), ch. 2-5 are specially relevant for the case of history of science.

[6] Jae-wook Ahn, Catherine Plaisant, & Ben Shneiderman. "A Task Taxonomy for Network Evolution Analysis," *IEEE Transactions on Visualization and Computer Graphics* 20, no. 3 (2014): 365-76.

[7] On issues proper to the semantic approach, including the need for user collaboration, see Ben Shneiderman, & Aleks Aris "Network Visualization by Semantic Substrates," *IEEE Transactions on Visualization and Computer Graphics* 12, no. 5 (2006): 733-40; a similar discussion is provided by Fóris, 423-5.

[8] Linton C. Freeman, "Visualizing Social Networks," *Journal of Social Structure* 1 (2000) art. no. 1. For a more thorough view of his work, including antecedents and consequences, see Freeman, "The Development of Social Network Analysis – with an Emphasis on Recent Events," in *The SAGE Handbook of Social Network Analysis*, ed. John Scott, & Peter J. Carrington (Thousand Oaks [CA]: SAGE Pub., 2014), ch. 3. The vast literature on the subject includes not only dozens of studies by enthusiastic followers of Freeman, but also rather critical views, as e.g., Mark C. Taylor, *The Moment of Complexity: Emerging Network Culture* (Chicago: The University of Chicago Press, 2001).

world.[9]

So, enter the digital humanities

Few people would dispute that digital technologies are fundamentally changing the way investigators engage in research. According to David Berry,[10] research is being increasingly mediated by digital technologies to the point that such mediation is beginning to change the very notion of research thus affecting the epistemologies and ontologies that underlie research programs. In the case of the digital humanities (DH), while its predecessor, the so-called humanities computing, merely applied computing techniques to subjects proper to the humanities, most notably, major edition projects the change in name involved a thorough conceptual shift that resulted in a new and autonomous field.[11]

According to Matthias Kirschenbaum, DH is better understood as a common methodological outlook than as specific sets of texts or technologies.[12] In a similar vein, Todd Presner considers that DH is "an umbrella term for a wide array of practices for creating, applying, interpreting, interrogating and hacking both new and old information technologies."[13] So much, that Kathleen Fitzpatrick was led to ask: "Digital Humanities – singular or plural?"[14]

Some scholars describe two waves of DH work. The first took place in the late 1990s and early 2000s and tended to focus on large-scale digitization projects and the establishment of technological infrastructure.[15] In the specific case of history of science, we might mention the development, starting in the 1990s, of databases on the life and work of individual actors,

---

[9] Franco Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History* (Brooklyn/London: Verso, 2005) which provides a careful computer-based analysis of literary history, which without the considerable scholarship of the author would not have succeeded in overcoming several of the problems stated above. Some later studies by Moretti's associates give hints of the methods applied; see e.g. Ryan Heuser, & Long Le-Khac, "Learning to Read Data: Bringing out the Humanistic in the Digital Humanities," *Victorian Studies* 54, no. 1 (2011): 79-86.

[10] David M. Berry, "Introduction: Understanding the Digital Humanities," in *Understanding Digital Humanities,* ed. David M. Berry (New York: Palgrave Macmillan, 2012) [eBook].

[11] Ibid.; Benjamin Caraco, "Les digital humanities et les bibliothèques," *Bulletin des bibliothèques de France* 2 (2012) available at: http://bbf.enssib.fr/consulter/bbf-2012-02-0069-002 (accessed 8 Feb 2016); Katherine Hayles, *How We Think: Digital Media and Contemporary Technogenesis* Chicago: The University of Chicago Press, 2010); Patrik Svensson, "Humanities Computing as Digital Humanities," *Digital Humanities Quarterly* 3, no. 3 (2009), available at: http://digitalhumanities.org/dhq/vol/3/3/000065/000065.html (accessed 8 Feb 2016); Patrick Svenson, "The Landscape of Digital Humanities," *Digital Humanities Quarterly* 4, no. 1 (2009), available at http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html (accessed 8 Feb 2016); Willard McCarty, *Humanities Computing* (New York: Palgrave, 2005).

[12] Kirschenbaum, Matthias G. "What is Digital Humanities and What's Doing in English Departments?" *ADE Bulletin* 150 (2010): 1-7.

[13] Todd Presner, "Digital Humanities 2.0: A Report on Knowledge," *OpenStax CNX*. 8/6/2010, available at: http://cnx.org/contents/2742bb37-7c47-4bee-bb34-0f35bda760f3@6 (accessed 8 Feb 2016).

[14] Kathleen Fitzpatrick, The Humanities, Done Digitally," *The Chronicle of Higher Education,* May 8 2011, available at: http://chronicle.com/article/The-Humanities-Done-Digitally/127382/ (accessed 8 Feb 2016).

[15] Presner, "Digital Humanities".

as e.g., Isaac Newton,[16] André M. Ampère,[17] Albert Einstein,[18] Charles Darwin,[19] and Henri Poincaré,[20] among several others,[21]or of institutions, such as the Marine Biological Laboratory, Woods Hole,[22] and the Max Planck Society,[23]an approach pursued to this day. Organization and cataloguing of these databases gave rise to a parallel wealth of relational metadata, which allowed detecting networks of interaction among actors, related production and most significant institutions. Other relevant examples in this regard are projects *Six Degrees of Francis Bacon*,[24] and *Régistres de l'Académie*,[25] which datasets are close to the ones obtained for the social sciences.

The second is a currently evolving wave, which is rather qualitative, interpretive, experiential, emotive and generative in character, and deals with 'born digital' knowledge. However, Berry notices that neither wave truly problematized the essential core of the humanities, and thus suggests looking at the digital component of DH "in the light of its medium specificity as a way of thinking about how medial changes produce epistemic changes."[26]

Against the initial expectations, disciplinary-based specificities did not take long to come to the fore. So, for instance, while digital technologies were quickly and easily assimilated by the social sciences, the same was not the case of the various fields of history research.[27] The reasons, according to the sociologist Nina Baur are several: 1) historians have a long-standing awareness that different data can be read in very different ways as a function of the interpreters' perspective (i.e., historiographical approach) and the specificities proper to different times and places; contrariwise, interpretation becomes normative and biased; 2) the focus of sociologists is on research-elicited data and the secondary analysis of them, just as in the natural sciences; in turn, historians focus on process-generated data, to wit, data generated through the very process of living, working, interacting in society; as such they are much more complex compared to what the usual sociological classifications of data suggest;

---

[16] John A. Walsh, & Wallace E. Hooper, "The Liberty of Invention: Alchemical Discourse and Information Technology Standardization," *Literary and Linguistic Computing* 27 (2012): 55-79; Robert Iliffe, "Digitizing Isaac: The Newton Project and an Electronic Edition of Newton's Papers," in *Newton and Newtonianism*: *New Studies*, ed. J.E. Force, & S. Hutton (Dordrecht: Springer, 2004), 23-38; Cesare Pastorino, Tamara Lopez, & John A. Walsh, "The Digital Index Chemicus: Toward a Digital Tool for Studying Isaac Newton's Index Chemicus," *Body, Space & Technology* 7, no. 2 (2008), available at: http://people.brunel.ac.uk/bst/vol0702/cesarepastorino/home.html (accessed 8 Feb 2016).

[17] *Ampère and the History of Electricity*, http://www.ampere.cnrs.fr/.

[18] *Einstein's Archives Online*, http://www.alberteinstein.info/.

[19] *Darwin Correspondence Project*, https://www.darwinproject.ac.uk/; *Darwin Manuscripts Project*, https://www.amnh.org/our-research/darwin-manuscripts-project.

[20] *Henri Poincaré Papers*, http://henripoincarepapers.univ-nantes.fr/.

[21] A regularly updated list is available at Digital History and Philosophy of Science: http://digitalhps.org/projects.

[22] *History of the Marine Biological Laboratory*, http://history.archives.mbl.edu/.

[23] *Forschungsprogramm Geschichte der Max-Planck Gesellschaft*, http://gmpg.mpiwg-berlin.mpg.de/de/.

[24] http://www.sixdegreesoffrancisbacon.com/?ids=10000473&min_confidence=60&type=network

[25] https://akademieregistres.bbaw.de/

[26] Berry, "Introduction."

[27] Stephen P. Weldon, "Historians and Their Data," in *Crossing Oceans: Exchange of Products, Instruments and Procedures in the History of Chemistry and Related Sciences*, ed. A,M. Alfonso-Goldfarb et al. (Campinas [SP]: CLE/UNICAMP, 2015), 299-322.

3) historians pay close and careful attention to the authenticity of the data, while sociologists rather focus on sampling and interpretation/analysis of the data.[28]

As Stephen Weldon pointed out, the sociological style of analysis, different from the historical one, attempts to cut through the individual and contingent factors that produced unusual data in order to see the bigger generalized picture. For that reason, he says, "the features that make historical work so powerful are not well suited to the current digital environment."[29] For this reason, the more historians rely on digital tools, a tension arises between the historical framework and the information environment. The reason is that in DH, the generation of datasets is not previous to or independent from actual research work. In the words of Susan Schreibman and colleagues, in DH, "critical inquiry involves the application of algorithmically facilitated search, retrieval, and critical process that […] originat[es] in humanities-based work."[30] Thus Joris van Zundert observes that the computational tools "should also warrant that existing heuristics and hermeneutics are appropriately translated into their equivalent digital counterparts, especially in a field where heterogeneity of data and multifaceted approaches are not regarded as reducible noise but as essential properties of the research domain."[31] It goes without saying that history is one such field per definition.

Among historians, the potential conflict between historical thinking and computational data management was already discussed about 40 years ago. According to James Levitt & Claude LaBarre the traditional historical methodology differs significantly from the computational methodology because historians generally think in terms of specific instances and individual cases, and not in terms of how cases can be understood as a group in a statistical way.[32] Those authors thus stressed that historians needed to develop new statistical techniques and methods, while ensuring that the information needed remains part of the dataset. In other words, to import statistical practices from other disciples does not work if it means that data critical for good historical analysis is lost.

Weldon further stresses that the digital environment is not a neutral medium for the flow of information, but it was intentionally built for specific purposes, which in turn shape how it is used.[33] And naturally, there are limits to what the information can do, as a function of its very nature, namely, a binary encoding of everything which makes it intrinsically reductionistic, being able to respond precisely as it was coded to do. As such, it imposes limits to history work and its required sources, since the historical methodology, by its very nature, tends to oppose the mechanistic digital environment. The stable resources studied by

---

[28] Nina Baur, "Problems of Linking Theory and Data in Historical Sociology and Longitudinal Research," *Historical Social Research* 34, no. 1 (2009): 7-21.
[29] Weldon, 300.
[30] Susan Schreibman, Ray Siemens, & John Unsworth, "Frontmatter," in *Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, & John Unsworth (Oxford: Blackwell, 2004), i-xxviii, on xxv.
[31] Joris van Zundert, "If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities," *Historical Social Research* 37, no. 3 (2012): 165-86, on 173-4.
[32] James H. Levitt, & Claude E. LaBarre, "Building a Data File from Historical Archives," *Computers and the Humanities* 9, no. 2 (1975): 77-82.
Levitt & LaBarre (1975, p. 77)
[33] Weldon, 301.

historians are idiosyncratic data, one-of-a-kind phenomena and unstructured information, which are impossibly difficult to put into a formal structure.

In simple terms van Zundert asks: "If only everybody would wear size 9 shoes, wouldn't that be a blessing for the shoemaking industry?"[34] Standardization poses more problems than offers solutions in the case of heterogeneous data and specific requirements. One immediate example is overtagging: when standards (like precise thesauri or well-defined concepts) are not available, retrieval of any item in datasets demands a too large number of tags, making the technology go back to the earlier humanities computing, where it originated.[35] An even more difficult problem is posed by the requirement to analyze any text in its specific time, space and sociocultural context, which is one of the most peculiar features of the historical method. As Weldon showed, it is extremely difficult to encode time, and even more difficult to encode context. In the best of cases, coding standards (like EAC-CPF, *Encoded Archival Context-Corporate Bodies, Persons and Families*, Society of American Archivists, 2011) allow for the establishment of relationships among actors, their work (as e.g., publications) and loci of activity (*Régistres de l'Académie Project*) from which context might be inferred.[36] Schemas like EAC-CPF, says Weldon, provide an excellent example of how historical methodology can be incorporated into a medium that was not initially designed to manipulate and work with objects of such complexity. Indeed, EAC-CPF was developed by archivists as a finding aid.

What this shows, essentially, is that tool building is not a mere research-independent act to enable data processing. Rather, it is the act of modeling humanities data and heuristics as an intrinsic aspect of research. Tool and software development thus represent in part the capture and expression of interpretations about structure and properties of data, as well as interactions with that data.

In recent years, DH received a boost with the emergence of big data programs. These are programs which in terms of technology maximize computation power and algorithmic accuracy and compare large data sets. In terms of analysis, they draw on large data sets to identify patterns in order to make, in the case of this project, scientific and educational claims.[37]

In the specific field of science and technology, big data deals with innovations for hypothesis testing and knowledge discovery whose applications are turned to high-throughput instrument based data collection, fine grained multiple-modality and large-scale records.[38] But, what is complementary to data? Analysis. This is why, in connection with big data, analytics is classified in five critical technical areas: big data analytics, text analytics,

---

[34] Van Zundert, 173.
[35] Elise Hanrahan, & Markus Schnöpf, "Scholarly Digital Editions: Connecting Archives and Libraries". *Conference: New Directions in Digital History of Science* (Berlin: Max Planck Institute of History of Science/Committee of Documentation and Bibliography, IUHPS, 2013) [forthcoming].
[36] Weldon, 313-4.
[37] Dana Boyd, & Kate Crawford. "Critical Questions for Big Data," *Information, Communication & Society* 15, no. 5 (2012): 662-79.
[38] Hsinchun Chen, Roger H.L. Chiang, & Veda C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* 36, no. 4 (2012): 1165-88.

network analytics, web analytics and mobile analytics.[39] Another connected field to big data is visualization for necessary interpretation, which must accompany the processing and analytics of big data. According to Cesar Hidalgo & Ali Almossawi the ability to understand and see a large volume of data is entering a stage of evolution similar to that which was brought by Galileo in astronomy in the seventeenth century.[40]

Accordingly, we performed a broad-scoped survey of approaches for DH developed in the past years which we summarize next.

Topic modeling

It consists of a body of mathematical and statistical methods for inferring 'topics' (recurring themes discussed in a textual corpus) from large collections of digitized texts. These methods were developed within the field of machine learning, with the primary objective of optimizing online search tools. The first such method was Latent Semantic Analysis (LSA) which extracts and represents the contextual-usage meaning of words by statistical computations applied to a large corpus of text.[41] LSA tries to explore the latent or implicit semantics in the text, given by global relations among the terms, rather than by the meanings of isolated words. LSA assumes that words that are close in meaning will occur in similar pieced of text. It produces measures of word-word, word-passage and passage-passage relations that are well correlated with several human cognitive phenomena involving association or semantic similarity.

LSA served as starting point for the development of a family of methods collectively known as ***topic modeling***. That is a way of providing a set of algorithms to discover the hidden thematic structure of a vast collection of texts. The results of topic modeling algorithms can be used to summarize, visualize, explore and theorize about a corpus.[42] Of particular interest for history research, the family includes ***Dynamic Topic Modeling***, which targets the transformation of topics over time.[43] Topic modeling has seen widespread adoption across DH, usually for exploring the thematic content of very large historical or literary corpora.[44] It should be noticed that tools like LSA lend themselves to network representations because they provide a formal and quantitative way of expressing relationships among both texts and latent concepts.

---

[39] Ibid.

[40] Cesar A. Hidalgo, & Ali Almossawi, "The Data-visualization Revolution," *Scientific American*, published March 17, 2014, available at:
http://www.scientificamerican.com/article/the-data-visualizationrevolution (accessed 15 September 2015).

[41] Thomas K. Landauer, & Susan T. Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge," *Psychological Review* 104, no. 2 (1997): 211-40.

[42] David M. Blei, "Topic Modeling and Digital Humanities," *Journal of Digital Humanities* 2, no. 1 (2012).

[43] David M. Blei, & John D. Lafferty, "Dynamic Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning – ICML 2006* (Pittsburgh: Carnegie Mellon University, June 25-29, 2006), available at:
http://repository.cmu.edu/cgi/viewcontent.cgi?article=2036&context=compsci (accessed 20 Feb 2016).

[44] Walsh & Hopper, "Liberty;" Jaimie Murdock, Colin Allen, & Simon DeDeo, "Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks," *Cornell University Library*, 10/12/2015 arXiv:1509.07175v2 [cs.CL] (accessed 8 Feb 2016).

Computer Corpus Linguistics (CCL)/Computational linguistics (CL)

CCL uses statistical machine learning to study the structure and evolution of language, and thus allows detecting and comparing distinct thought structures and patterns of language associated with localized communities.[45] Thus many scholars have recognized a natural fit between the methods of computational linguistics and the long-standing research goals of concept-based history research, as the study by Stephen Pumphrey & Paul Ashcroft shows.[46] Work by David Hall and colleagues has explicitly linked formal theories in computational linguistics to the concept of scientific paradigms. [47] One important application of computational linguistics to history of science is the use of named entity recognition and relation extraction to develop network models of historical actors and concepts in large digitized text collections.[48]

Social network analysis (SNA)

SNA is a computationally oriented area of sociology is increasingly used by historians of science and science and technology scholars, among others, to address questions about scientific communities. SNA provides a battery of quantitative tools for interpreting static networks based on interpersonal relationships. In history of science and science and technology studies, social network analysis has been used to study the transmission of knowledge.[49] Several authors used SNA to document sociotechnical processes at several key moments in the history of science and technology.[50] These kinds of studies generally focus on single, unimodal networks, that is, involving a single class of actors or concepts, and arrive at historical inferences by interpreting static parameters and distributions. Social network methods have been extended by the defense research community for large-scale analyses (e.g. screening for unusual actors) that include a temporal component; these techniques have also been introduced to the history of science in recent years.[51]

---

[45] Lars Engwall, Enno Aljets, Tina Hedmo, & Raphaël Ramuz, "Computer Linguistics: An Innovation in the Humanities," in *Organizational Transformation and Scientific Change: The Impact of Institutional Restructuring on Universities and Intellectual Innovation*, ed. Richard Whitley, & Jochen Gläser (Bingley: Emerald Group Publishing Limited, 2014), 331-65; Tony B. Sardinha, "Linguística de Corpus: Histórico e Problemática," *D.E.L.T.A.* 16, no. 2 (2000): 323-67.

[46] Stephen Pumphrey, & Paul Aschcroft, "Alchemy, Chemistry or Chymistry: An Analysis of Actors' Categories in Early Modern England," in *16th SHAC Postgraduate Workshop. Programme and Edited Abstracts.* Oxford, 30 October 2015, 5-6.

[47] David Hall, Daniel Jurafsky, & Christopher D. Manning, "Studying the History of Ideas Using Topic Models," *Proceedings of the Conference on Empirical Methods in NATURAL Language Processing, EMNLP/2008*, Edinburgh, 25-27 October 2008, available at: http://web.stanford.edu/~jurafsky/hallemnlp08.pdf (accessed 8 Feb 2016).

[48] John Kizito, Ismail Fahmi, Erik T.K. Sang, John Nerbonne, & Gosse Bouma, "Computational Linguistics and History of Science,"in: *Storia della scienza e linguistica computazionale: Sconfinamenti possibile,* ed. Liborio Dibattista (Milano: FrancoAngeli, 2009), 55-73.

[49] René Sigrist, & Eric D. Widmer, "Training Links and Transmission of Knowledge in 18th Century Botany: A Social Network Analysis," *REDES* 21, no. 7 (2011): 319-59.

[50] Francis C. Moon, ed. *Social Networks in the History of Innovation and Invention* (Dordrecht: Springer, 2014).

[51] Catherine A. Bliss, Morgan R. Frank, Christopher M. Danforth, & Peter S. Dodds, "An Evolutionary Algorithm Approach to Link Prediction in Dynamic Social Networks," *Journal of Computational Science* 5, no. 5 (2014): 750-64.

Text analytics

A new information technology (IT) discipline that emerged within the context of business intelligence to considerably expand its scope of applications, formulated as an answer to the 'unstructured data' problem, i.e., 80% of business information originates and is locked in 'unstructured' form.[52] It focuses on the extraction, categorization and classification of text-extracted data originally based on linguistics and data mining. More recently it expanded first via extension to data mining workbenches, and later on in the form of term-extraction and analysis interface, thus providing ability to discern features in text and extract them to databases. Text analytics first emerged in the 1990s as simple 'text (data) mining' to evolve through the use of linguistics to handle variant words and multi-word terms and looks for hidden relationships and other complex patterns within datasets. Techniques include classification, clustering, link analysis and decision trees, among others, as well as predictive modeling, all of which can be applied to data derived from textual sources. Within this context, it is worth to call the attention to *Sobek*, a text-mining tool developed by a group from Federal University of Rio Grande do Sul chaired by Eliseo Reategui (sobek.ufrgs.br) which we are currently applying to our library (CESIMA Digital).

Digitization and parallel adoption of digital and computational research methods led to a proliferation of new forms of scholarly production. These include new forms of communicating history and philosophy of science to the public, for example, the so-called 'maps of science.'[53] These new forms are not only relevant for scholars, but also for educators and the public at large.

Final remarks

Computational and big data driven approaches are increasingly applied to many areas of the arts, humanities and social sciences, including history and philosophy of science (HPS) and science and technology studies (STS). Methods deployed in these areas range from various forms of network analysis to statistical and semantic analysis of large and small text corpora. These developments are beginning to yield novel insights, however, some hints strongly indicate that successful computational and big data approaches require specific adjustment to the heuristics and hermeneutics proper to the humanities.

Therefore we believe it relevant to conclude this paper with some considerations on initiatives which since the beginning of the 2010s have sought to provide solutions to scholars like us, who mandatorily need to detect conceptual shifts, and thus consider the aspect of temporality in large datasets. One example is, for instance, ***culturonomics***, which consists in quantitative analysis of culture using large corpora of digitized texts. The idea was launched

---

[52] Ashok N. Srivastava, & Mehran Sahami, ed, *Text Mining: Classification, Clustering, and Applications* (Boca Raton [FL]: Taylor and Francis, 2009); Seth Grimes, "A Brief History of Text Analytics," *BeyeNetwork* October 30 2007, available at: http://www.b-eye-network.com/view/6311 (accessed 5 March 2016); Adam Schencker, Graph-theoretic Techniques for Web Content Mining," PhD Dissertation, Tampa: University of South Florida, 2003.

[53] Ismael Rafols, Alan L. Porter, & Loet Leydesdorff, "Science Overlay Maps: A New Tool for Research Policy and Library Management," *Journal of the American Society for Information Science and Technology* 61, no. 9 (2010): 1871-87.

by Jean-B. Michel and colleagues, who constructed a corpus of digitized texts containing about 4% of all books ever printed to investigate, quickly and accurately, historical and contextual trends in fields as diverse as lexicography and epidemiology. The authors realized that the possibilities were such, that they granted the creation of a new evidence-based field in the humanities, as grounded on "fossils" of ancient creatures as paleontology.[54]

The repercussion of this and following initiatives notwithstanding, the fact that the major gains brought by this approach were mainly statistical soon became evident. Temporality still eludes the attempts to capture it within large databases, and consequently also the conceptual changes hidden in them. While no efforts were spared to complement culturonomic studies, its process still needs considerable refinement, including more through and qualified data extraction to reach the deeper layers of corpora, and attempts to come increasingly closer to their original sources. Once again, as expectable, this approach demands active and integrated participation of users specialized in cultural and humanities studies.[55]

On the other hand, the fact that such specialized users tend to apply theoretical frameworks of their own to data has been long recognized. Yet, data are just signals extracted from texts, and thus when they captured using preset conceptual frames, the ones that do not seem to fit risk being discarded, or are deprived of their more subtle and complex nuances.[56] Therefore it seems that in spite of all the advance made in the last decade, we still must cope with the dilemmas formulated long ago by Ben Shneiderman. Perhaps the key to find the exit out of this maze is to increasingly refine training in digital humanities, so that the 'humanities' and 'digital' sides might become fully integrated.

---

[54] Jean-Baptiste Michel et al, "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science* 331, no. 6014 (2011): 176-82. This widely known article, a virtual manifesto, was signed by authors from many different areas, including Google staff, which contributed to the vast data extraction. Held as a milestone, it was followed by many complementary initiatives and publications, as e.g., Yoav Goldberg, & Jon Orwant, "A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books," in *Second Joint Conference on Lexical and Computational Semantics, Volume 1: Proceedings of the Main Conference and the Shared Task*, Atlanta [GA], 2013: 241–7.

[55] An example of such demands is provided by Nina Tahmasebi et al, "Visions and Open Challenges for a Knowledge-based Culturomics," *International Journal of Digital Library* 15, no. 2/4 (2015): 169–87.

[56] Heuser & Le-Khac, 81; as mentioned above, these authors are associates of F. Moretti, one of the founders of the Stanford Literary Lab and the one who coined the concept of 'distant reading;' see Franco Moretti, "Conjectures on World Literature," *New Left Review* 1 (2000): 54-68. The notion of distant reading involves a wider and quantitative approach to the global literature allowing to detect previously unnoticed data and thus provide more thorough and accurate analyses. It was further developed in Moretti's later work with the help of new digital technology, achieving much repercussion to the point it came to be considered as the matrix for approaches such as the ones applied in studies on culturonomics; see e.g., Lars Borin et al, "Mining Semantics for Culturomics: Towards a Knowledge-based Approach," in *UnstructureNLP '13 Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing*, New York, 2013: 3-10; doi: 10.1145/2513549.2513551. However, by the time the manifesto was launched, Moretti and associates at Stanford Literary Lab had already developed a more refined and complex view of the relations between the digital world and specialized users.