HAL
open science

Digital, digitized, and numerical humanities

Camille Roth

To cite this version:

Camille Roth. Digital, digitized, and numerical humanities. Digital Scholarship in the Humanities, 2019, 34 (3), pp.616-632. 10.1093/llc/fqy057 . halshs-02281134

HAL Id: halshs-02281134

https://shs.hal.science/halshs-02281134v1

Submitted on 26 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Digital, Digitized, and Numerical Humanities

Camille Roth

Digital Humanities and Computational Social Science Team
Centre Marc Bloch, CNRS/HU Berlin
Friedrichstrasse 191, D-10117 Berlin
roth@cmb.hu-berlin.de

This is a pre-copyedited, author-produced version of an article accepted for publication in Digital Scholarship in the Humanities *following peer review. The version of record, Digital Scholarship in the Humanities, Volume 34, Issue 3, September 2019, Pages 616–632, is available online at: https://doi.org/10.1093/llc/fqy057.*

Abstract

The term "digital humanities" may be understood in three different ways: as "digitized humanities", by dealing essentially with the constitution, management and processing of digitized archives; as "numerical humanities", by putting the emphasis on mathematical abstraction and the development of numerical and formal models; and as "humanities of the digital", by focusing on the study of computer-mediated interactions and online communities. Discussing their methods and actors, we show how these three potential acceptations cover markedly distinct epistemological endeavors and, eventually, non-overlapping scientific communities.

7671 words

Introduction

The "digital humanities" label may appear as an ideal keyword to scholars who empirically study online social systems using methods borrowing to both social and computational sciences. Their work is indeed connected with three potential acceptations of the term: they do digital research on digital worlds using digital datasets. Put differently, they may construe their activity as "digitized humanities" (relying on digitized data), "numerical humanities" (using numerical and, more broadly, formal models) and "humanities of the digital" (studying computer-mediated interactions).[1]

[1] This ambiguity may be even stronger in the French-speaking realm where "digital humanities" are generally translated as "humanités numériques": "numériques" means both "digital" and "numeric" — besides, the translation of "digitized" is "numérisé", which is not far from "numeric" either, even though no numbers are usually involved.

In general, however, academics who identify themselves as "digital humanists" may be more or less sensitive to this polysemy, as it often covers distinct epistemological endeavors. The purpose of this contribution is to shed some light on this trichotomy and to make some remarks on the corresponding scientific differentiation. We also aim at discussing the conditions of a better articulation between these streams.

1 An overarching dichotomy

We first focus on the two most obvious acceptations of "digital humanities" which, as we shall later show, correspond to what appears to be in practice an overarching dichotomy between digitized and numerical humanities.

1.1 Digitized humanities

The perhaps most widespread acceptation of "digital humanities" relates to the creation, curation and use of digitized datasets in human sciences and, to a lesser extent, social sciences.[2] In broad terms, these approaches include the development and application of computer tools to, *inter alia*, digitize, store, process, gather, connect, manage, make available, mine and visualize text collections and corpuses, image banks or multimedia documents of various origins.

In practice, these tasks raise a series of fundamental infrastructure-related issues, in terms of dataset construction (involving the design of appropriate ontologies), storage (featuring questions such as long-term availability) and access (requiring ergonomic and remotely accessible interfaces). Yet, they also prompt the integration of diverse methods stemming from more formal disciplines, most importantly artificial intelligence (AI). This ranges from established and rather low-level tools such as optical character recognition (OCR) to higher-level data processing techniques, which are connected to open research questions in their origin fields such as named entity detection and classification [Nadeau and Sekine, 2007], optimal graph visualization [Tamassia, 2013] or the more recent techniques of crowdsourced archive curation [Carletti et al., 2013]. We discuss below the links between these AI techniques and the older field formerly denoted as "humanities computing".

This integration leads, in turn, to epistemological advances in the humanities by enabling novel questions, thanks to productivity gains as well as genuinely new ways of looking at traditional data. Examples abund in linguistics, as shown by the development of "corpus linguistics" [see e.g., Kennedy, 1998], Moretti [2000]'s notion of "distant reading" and further illustrated by the long dominant position of literature and linguistics in digital humanities journals [Huggett, 2012]. This prominence of linguistics should not come as a surprise. The field that became called as "humanities computing" has emerged long before the term of "digital humanities" has been coined. Humanities computing started in the late 1940s on the premises and promises of applying quantitative approaches on text corpuses [Hockey, 2004]. For instance, statistical methods have been developed in order to automatically analyze stylistic features, under a field denoted as "stylometry", using lexical,

[2] In all generality, we employ the term "dataset" without assuming any level of post-processing: data may consist of raw digitized corpora or measurements which may be completed or documented with comprehensive metadata. The term "archive" is in this respect less generic, as it often relates to digital material which is based on a unique source and where the conservation of the original material particularly matters.

syntactic, semantic markers and thus making it possible to provide, for instance, stastistical confidence in author attribution [Stamatatos, 2009] or in the chronological ordering of the works of some author [Holmes, 1998]. The field became further institutionalized from the late 60s to the mid-80s through the creation of various associations, such as the Association for Literary and Linguistic Computing (ALLC) or the Association for Computers and the Humanities (ACH), publication venues, such as "Computers and the Humanities" (1966) and "Literary and Linguistic Computing" (LLC, 1986; which recently took a broader name as "Digital Scholarship in the Humanities", DSH), or normalization efforts, such as the Text Encoding Initiative, towards the end of the 1980s.

Nonetheless, many other disciplines beyond humanities also have an established record of contributions which fit into this "digitized humanities" acceptation, by relying on archives which are not, at least primarily, in a purely textual format. Several social sciences have pioneered the use and computer-assisted exploration of digital archives, including first and foremost archaeology and its old interest in "digital archeology", or geography and its long-standing development of geographical information systems (GIS) — the intersection of these two digital humanities subfields, GIS & archeology, is already a vast domain in itself. The comprehensive textbook of Conolly and Lake [2006], for one, reviews many of the application domains of these endeavors, ranging from data acquisition, the computer-assisted analysis of regions of interest or various types of routes, to data management, for instance through elevation models. Political sciences have relied on the mining of public discourses in a variety of manners, including the study of nationalist reference frames in digitized newspaper corpuses [van den Bos and Giffard, 2016], a material that communication research increasingly exploited in a digital humanities perspective [Nicholson, 2013], while applications at the interface of history and sociology also emerged, such as the reconstruction of historical social networks by extracting names and inferring relations from the *Oxford Dictionary of National Biography* [Warren et al., 2016]. (I will come back to the specific expectations of social sciences with respect to archives and, more broadly, datasets, in Section 2.1.)

1.2 Numerical humanities

Another perspective on "digital humanities" puts the emphasis on mathematical abstraction and modeling. Undoubtedly, digitized corpuses are already, to some extent, abstractions of the original empirical material. In what I shall denote in the remainder of this chapter by the term "numerical humanities", however, researchers primarily appear to *develop* mathematical frameworks and computer science methods with the specific goal of formalizing and stylizing some systematic social processes. In other words, the objective is to capture the possibly *general* mechanisms at the root of the observed data. In this sense, this acceptation does not relate to the computer-assisted realization of tasks which are difficult or impossible to do by hand — as achieved by OCR, or the application of GIS tools, clustering or statistical methods on large datasets, for example. Rather, it aims at developing numerical models of human or social behavior per se. In this respect, in practice, it generally builds more often on social science research issues than humanities.

In part, this type of approach relies again strongly on artificial intelligence — for instance when pattern detection methods are used to confirm or refute the existence of some regularity in a given dataset, such as using cell phone call records to exhibit some mobility trends or showing the presence of a specific type of interaction preferences [Sobolevsky et al., 2013,

Miritello et al., 2013].

Yet, the distinctive element here is that a given digital dataset merely constitutes one among many possible empirical realizations. In other words, what matters is the theorizing of an underlying process which may account for some sort of empirical measurements on a whole class of *comparable* datasets. In this case, scholars usually aim at designing mathemetical frameworks where theory-based hypotheses and their empirical manifestations can be formalized. They then develop numerical or algebraic models aimed at reconstructing some key characteristics of the studied systems. These key features are often construed as *universal* classes of empirical configurations, independently of a specific archive. They include *systematic* spatial and/or temporal correlations between some types of variables, typical over- or under-representations and, more broadly, distributions of values rather than *specific* values. For instance, one may look for the systematic emergence of configurations where "few have much, many have little" (e.g., so-called Pareto laws, Zipf laws, power laws) predicted by some theory and generic process (e.g., so-called Matthew effect, or "rich-get-richer" reinforcement). By contrast, knowing who precisely is "rich" and possesses a lot in a given case study is of lesser importance.

Sometimes, the abstraction goes as far as modeling entire artificial societies and describing a typology of stylized facts emerging from their very simulation — an approach epitomized by the eponymous field of "social simulation" [Epstein and Axtell, 1996, Gilbert and Troitzsch, 1999], where models include a thorough description of social or behavioral processes (they are usually "agent-based" models [Bonabeau, 2002, Helbing and Balietti, 2012]) while the computer provides its own experimental data. More often, real-world empirical observations are used, yet again to validate an abstract class of processes that the chosen digital dataset is supposed to document. The data helps confirming an *a priori* model, or appraising the divergence between the model and reality. For instance, the above-mentioned mobility trends and interaction preferences are actually being described in regard to models which play the role of *laws* — that is, the description of these patterns makes sense within a certain modeling and theoretical framework. These laws are either (i) unknown and have to be determined from the data, or (ii) they are hypothesized and thereby provide a baseline or a benchmark from which the data may diverge (for instance, expected mobility based on a gravity model [Sobolevsky et al., 2013], expected interaction patterns influenced by a demographic bias [Miritello et al., 2013]). In both cases, case studies play no further role than helping to validate (or parameterize) a target (and rather generic) model.

The epistemological setting is thus slightly different, for datasets are not anymore exploited as singular recordings corresponding to given empirical case studies, but simply as exemplar instances of a much wider and, more importantly, interchangeable phenomenon. This approach is not dissimilar from the one usually ascribed to natural sciences, in that they seek nomothetic rather than idiographic regularities[3], i.e. general laws rather than local patterns. In this context, datasets are considered as equivalent, possibly reproducible empirical instances of an "identical" underlying process. One implicitly assumes that there exists an isomorphism between various data stemming from distinct yet *similar* empirical situations — for instance, any dataset based on cell phone records should, in principle, confirm the

[3]This argument builds upon the classical dichotomy proposed by Windelband and Rickert to distinguish nomothetic and idiographic descriptions. While the latter is usually associated to natural sciences and the former to human and social sciences, both natural and social sciences alternatively tend to look for both types of descriptions (see also Bouvier [2010]).

existence of similar phenomena.

This second perspective on digital humanities corresponds to an epistemological breakthrough illustrated by the recent surge of interest in so-called "computational social sciences" (CSS) [Lazer et al., 2009, Conte et al., 2012], building upon the older fields of mathematical sociology (mainly animated by social scientists) and social complex system modeling (leaning more towards applied mathematics, computer science and statistical physics). Over the last decade, numerous conferences, journals and even departments devoted to these approaches have emerged. Many have been initially connected to quantitative social science research programs, especially in mathematical sociology (featuring the Journal of Mathematical Sociology since 1971) and the Social Network Analysis (SNA) community (around an international association founded in the late 70s [INSNA, Intl. Network for SNA] and yearly international conferences since the early 80s). The field of "Social Simulation" [Gilbert and Troitzsch, 1999] has been relatively institutionalized since around the 1990s, with international conferences currently drawing hundreds of attendants, supported by core journals such as JASSS (Journal of Artificial Societies and Social Simulation) and several international associations, the oldest being ESSA (European Social Simulation Association). Complexity sciences have also a clear connection with numerical humanities, through the study of social complex systems, which most complex system journals and conferences devote a significant attention to, and which spreads to neighboring disciplines as well (for example, the German Physics Society [DPG – Deutsche Physikalische Gesellschaft] has a special chapter on Socio-Economic Systems which is very much aligned with this program).

Venues explicitly referring to CSS have emerged in the last few years, both at the national level (in Germany, let us mention the series of GESIS symposia on CSS since 2015 which attract an international crowd) and international level (including SocialCom and SocInfo, both since 2009, and since 2016 the freshly launched ICCSS, all international conferences respectively on "Social Computing", "Social Informatics" and CSS). Their success owes to the joint and commonplace development of unprecedented computational resources and, most importantly, large-scale databases — either digitized datasets stemming from archives (rather related to human sciences) or natively digital datasets, produced through the recording of human behavior by digital devices (not least from the internet, as will be evident in the latter sections of the present chapter), in the wake of the currently fashionable and prevailing "big data" movement.

2 An overlapping dichotomy

2.1 Digital divides and overlaps

While digitized and numerical humanities have been defined as distinct perspectives, they may well coexist within the same scientific community, for numerical approaches may inform digitized humanities, and digitized datasets may feed numerical humanists. In this light, what might nevertheless remain unexpected is the relative lack of explicit connections between the communities who respectively identify themselves as *digital humanists* or *computational social scientists*. For instance, explicit references to "digital humanities" within computational social science communities remain relatively rare; and vice versa. By contrast, it is not rare to meet social scientists who identify better the signification of keywords

such as “mathematical sociology”, “social computing” and “computational social sciences” than “digital humanities”.

General epistemological causes can and will be invoked, here, to explain this peculiar and perhaps temporary situation. Social sciences and humanities certainly had markedly distinct development paths with respect to digital and numerical methods, as regards both data collection and processing, and epistemological attitudes.

Experimental and nomological social sciences

On the one hand, the empirical material of social sciences is often made of either (i) textual records (essentially based on interviews or ethnographic observations) or (ii) tabulated data (be it, for instance, demographic, query-based, relational or geographical databases).
Each kind of data corresponds, in turn, to noticeably distinct approaches.

(i) The analysis of the former, typical of the ethnographic approach and particularly prevalent in sociology and anthropology, generally relies on human interpretation and manual induction of typologies of discourses or behaviors from a collection of interview transcripts. There is nonetheless a small yet long tradition of quantitative text analysis [Popping, 2000, especially in chapter 8] related, in part, to the so-called “grounded theory” which notably advocates an inductive approach from (field) data [Glaser and Strauss, 1967] to discover social categories. While this branch concurrently led to significant software development [Klein, 2001], it remains infrequently applied in the field, and manual text processing is still widespread.

    Regardless, there is nowadays a widespread culture of digitization, as many researchers arrange their ethnographical records into computer files. Yet, there are few efforts aiming at normalizing the format of archives, securing their conservation or publicizing them, essentially for privacy reasons — transcripts are certainly not made available to a wider audience, outside of the small circle of scholars who primarily designed and carried out the empirical protocol.

    As a result, even though they do share a common epistemological and technological background in terms of discourse and narrative analysis with digital humanists working on text corpuses, such social scientists appear to be rarely concerned by the issues commonly addressed in the “digitized humanities”, especially its subfields related to text processing.

(ii) The latter type of data admittedly lends itself readily to numerical analysis. But it also resonates with a long-standing stream of research devoted to formalization rather than digitization. While tabulated and structured databases indeed enable social scientists to apply systematic procedures to detect or assert the existence of relevant patterns and phenomena — an objective which is aligned with the epistemological promises ascribed in the previous section to digitized humanities — they have also played a key role, from early on, in fostering the development of a *nomological* stance. Durkheim, for one, in his study of pan-European suicide at the end of the 19th century, sought the statistical signature of anomie as a general social process, rather than discussing the specific national or seasonal figures per se.

In this nomological context, empirical data has an instrumental and even, one may say, almost accessory status. It is not uncommon to witness the predominance of a normative attitude over a descriptive attitude: some approaches even assume the existence or relevance of some abstract structures or formal designs *before* observing them in real datasets (let alone archives). In concrete terms, for instance, the quantitative study of sociological communities relied first on conceptual constructs such as cliques [Luce and Perry, 1949] or structurally equivalent sets [Lorrain and White, 1971]. Many of the seminal works in this area dealt with a variety of issues without resorting to empirical data, spanning otherwise concrete topics such as information propagation in social systems [Rapoport, 1953], social field configurations [Cartwright and Harary, 1956], urban segregation dynamics [Schelling, 1971], conditions of the emergence of riots [Granovetter, 1978], cultural epidemiology in cognitive anthropology [Claidière and Sperber, 2007], or even a significant portion of economic theory [Samuelson, 1947]. Studies diversely affiliated with the paradigms of agent-based and multi-agent modeling [Axelrod, 1997], artificial societies and social simulation [Gilbert and Troitzsch, 1999] and social self-organization [Helbing, 2012], are practically dataset-free or primarily rely on artificial datasets, what is admittedly diametrically opposed to digitized humanities. Here, real-world datasets are admittedly useful to empirically validate the prediction of the existence of some stylized or so-called "emergent" phenomena (i.e. macro-scale phenomena which are not obvious from micro-level behavior and seem to be only describable through a macro-level observation framework) — but they are far from being a requisite.

The long-lasting presence of this approach in many social sciences [in sociology, see Edling, 2002] plausibly facilitated its affiliation to modern-day computational social science (CSS). It is also no wonder that CSS gathers fields traditionally associated with natural sciences, such as statistics (for obvious reasons), discrete mathematics (including classification, graph, game theory), system science (to configure social processes as dynamic iterative systems), and statistical physics (to study large interactional systems). By contrast, this stance appears to have much less appeal in fields traditionally associated with human sciences.

Archive-based humanities

On the other hand, humanities plausibly pay a marked attention to the specificities of singular datasets e.g., by focusing on a given author or a given historical case study, where social sciences tend to assume, as said before, the interchangeability of datasets within a similar empirical context. This may explain why textual and even visual archives are more ubiquitous in human science and, as a result, why "digitized humanities" have a stronger urgency than in fields more related to social science: sophisticated archiving issues regarding, for instance, indexation, data processing and availability. On the contrary, social scientists may feel less concerned with the preservation of specific datasets since (i) their increased focus on reproducible and/or interchangeable observations of a social system makes them less dependent on unique archives, and (ii) they often have the option of being in control of the empirical protocol and, thereby, of the conditions of data construction. Their interest in archives is further complicated when datasets cannot be released because of privacy concerns (typically, small-scale field research such as interviews) or commercial interests

(typically, when a dataset provided by a private company remains a current business asset).

This relative lack of enthusiasm (and, perhaps, of good practices) regarding archives, on the part of social science, is further reinforced by a certain preference for the present. Social science studies appear to focus more often on contemporary issues[4] or, at least to some extent, atemporal research questions (as regards the most nomological and naturalized part, such as the above-mentioned works on cultural transmission, kinship structure, or urban segregation). This warrants the collection of recent and even new data to fulfill the needs of a research study: in practice, empirical validation consists in proposing a protocol and then creating the conditions to build a(ny) new dataset, or hunt for an(y) existing database fulfilling some criteria essential to the protocol, without necessarily being peculiar to a specific spatio-temporal empirical situation. In other words, studying the current sociological impact of the introduction of IT in German, French, European or worldwide university departments is certainly less idiographic and dependent on a unique set of observations than writing the history of the development of digital humanities specifically in Germany, and articulating its key actors, dates and institutions.

Furthermore, the fact that numerical humanists often rely on a genuinely *experimental* approach (i.e. doing experiments) is not to be underestimated in their relative disaffection with archives. In effect, the assessment of the future value and relevance of a raw dataset collected for a given case study is intrinsically linked to the question of the experimental protocol, on at least two levels:

- The first one relates to the collection protocol. A given database may indeed become more or less obsolete for further research: either because it has revealed most of its mysteries at a certain level of granularity (resolution limit), or because new questions require a richer data collection protocol (ontology limit). For example, in order to advance the above-mentioned works of Miritello et al. [2013] and Sobolevsky et al. [2013], one can expect that more detailed datasets would be needed, and therefore built.

- The second one touches the complexity of the pre-processing protocol. Oftentimes, very big datasets need to be pre-processed before being manageable — think, here, of the readily-available digital dumps of the complete history of the Wikipedia, documenting all revision and interaction histories. Most scholars would only be interested in a subset of the data: interaction networks between pages [Zlatic et al., 2006] or contributors [Capocci et al., 2006], dynamics of reference lists on pages [Chen and Roth, 2012], edition diversity of contributors [Adamic et al., 2010], votes for administrators [Leskovec et al., 2010], to cite a few.

For both these reasons, there will most likely be eventually as many archives and associated tools as there are classes of research questions, complicating further a convergence of interests around a given huge original database, as could be the case for digitized humanities.

2.2 Two epistemic communities?

To summarize, we first observe a disjunction between a contemporary (or non-historical) focus, encouraging the creation of relevant data from scratch, and a historical focus, impos-

[4] Clearly, few sociologists would be interested in decade-old interviews, except for historical comparative purposes (thereby leaning again towards humanities).

ing a heavy dependence on archives. Second, we may differentiate a nomological stance (and a relative interchangeability of datasets) from a more idiographic stance (and relative singularity of datasets). We contend that this double dichotomy creates a combination of factors which partially help explain the relative disconnection between the two scientific communities dealing respectively with digitized and numerical humanities.

Despite this, it is important to recognize the shared interests of these overlapping epistemic communities, at least from an *instrumental viewpoint*. The two perspectives are indeed intertwined. Numerical models of humanities data require digitally-available corpuses in the first place, while the evaluation and detection of specific patterns from such datasets fuels the design of hypotheses which, in turn, are used as the basis for the modeling of artificial societies and eventually the confrontation of simulations to empirical data.

Some interdisciplinary wide-ranging endeavors have already successfully pursued this kind of bicephalous research under a single umbrella — the most prominent and perhaps oldest such research program being the field of "bio-informatics", which features a virtuous mix of data visualization, pattern analysis, and modeling of artificial evolutionary processes. Why the junction between computational social sciences and digital humanities has not yet fully occurred, in terms of scientific communities, remains to be discussed. While there is no obvious *de jure* epistemic dichotomy between fields rather associated with human sciences which would be dedicated to "digitized humanities", and fields leaning towards social sciences and more dedicated to "numerical humanities", my experience seems to indicate that there is a *de facto* social split. Empirical evidence will support this intuition in section 4.

An obvious remedy would consist in encouraging the development of areas at the interface between the two stances. The recent agenda of several humanities scholars definitely reflects the preoccupation of appraising existing corpuses with a fundamentally nomological angle. In linguistics, we may evoke the work of Ghanbarnejad et al. [2014] which aims at discovering language evolution laws based on the adoption of spelling innovations, including orthographic variations, at the scale of an entire population, by relying on an extremely large corpus covering millions of digitized books from the two last centuries [described by Michel et al., 2011]. Focusing on so-called adoption "s-curves", they distinguish innovations related to exogenous causes (e.g., imposed reforms) from those due to endogenous causes (e.g., progressive regularization of verbs, or spelling simplifications) by positing the existence of distinct dynamical processes, evidenced in turn by different categories of time series. In an archeological context, Knappett et al. [2008] focus on maritime interactions in the Aegean sea during the Bronze Age. They assume a process of cost minimization empirically based on geographical distances between known settlement sites. Their model, which owes very much to Hamiltonian mechanics, enables them to propose a probable interaction network whose shape fruitfully questions existing field knowledge [as presented, in particular, by Broodbank, 2000]. In anthropology, Roth et al. [2013] assess the contribution of chance to the morphological properties of kinship and marriage alliance networks. They hypothesize and analytically study a random baseline of alliance formation to empirically scrutinize the matrimonial role of social groups as asserted by native or ethnological theory. To this end, they use around 20 digital corpuses of genealogical networks from varied temporal and spatial origins [most of them available on *kinsources.net*, see Bringé et al., 2014] and represent them in the uniform framework of weighted networks. Using historical archives, Boulet et al. [2008] discuss the relevance of a variety of significantly formal notions

on graph structure for the purpose of defining communities in interaction graphs stemming from agrar contracts made during the 14th century in a French seigniory. Even though they eventually illustrate their discussion using a specific digital dataset in a specific region at a specific time, it clearly appears that not only any similar database would have been a valid candidate as well, but they also seem to suggest that the conclusions drawn from this very case study are in essence extendable to any other similar dataset — a stance which is typical for the normal science of social network analysis [Hummon and Carley, 1993]. Similarly, White et al. [2007] statistically study a large historical database describing city sizes over the period of a thousand years [Chandler, 1987]. They aim at validating a diverse set of assumptions coupling demographic and socio-political dynamics. Here again, the observed stylized features are discussed in terms of universal, almost natural phenomena.

This very partial selection of examples showcases instances of strong interlock between digitized and numerical humanities, as we presented them so far. They highlight the large-scale, cross-temporal empirical validation made possible by numerical digitized humanities. They also illustrate how often digital archives may be considered as an interchangeable instrument — the idea of preparing archives for a third-party use is indeed not so frequent among numerical humanists. As said before, beyond privacy and commercial concerns, a certain appetence for renewed experiments might help explain why numerical humanists rarely devote additional time to systematic and concerted efforts to define common formats, ensure long-term data availability and convene on normalized analysis procedures and tools (possibly based on such common formats and meta-databases). In this respect, the objectives of, say, national libraries archiving web content on a massive basis (and possibly with an undocumented selection bias) might not be entirely aligned with those of numerical humanists inclined to use this kind of data, especially for nomological purposes.

This last example brings us to another area of fertile interface between digitized and numerical humanities. It relates to the study of the online world, a third and entirely distinct perspective on "digital humanities", perhaps even orthogonal to our dichotomy so far. This domain, where digital archives are *native*, attracted many scholars both from digital humanities and computational social science; it will be the focus of the next section.

3 An overlapping trichotomy

3.1 Humanities of the digital

A third understanding of "digital humanities" pertains to what may be called the humanities of the digital. This acceptation focuses on computer-mediated interactions and societies, such as the Internet and other online communities. It principally gathers sociologists, ethnologists and, to a lesser extent, linguists, political scientists and geographers, among others.

There has been a slow yet steady recognition of online communities as a legitimate investigation field *per se* — "a sense of the Internet as simply another context where social life is lived, where research methods are applied, and where contemporary social issues are addressed", as Hine [2004] nicely puts it. This may be originally attributed to two non-exclusive movements. First, the progressive use of electronic devices to "digitize" the classical toolbox of conventional field research. Murthy [2008] explains how scholars started using digital

cameras, web-based questionnaires, emails and, more recently, social networking sites, to carry out interviews or make surveys. He further underscores the role of online platforms and communities as a means of recruiting and observing participants (especially, in the earlier times, to help connect with normally hard-to-reach populations and underground groups). Second, the construal of computer networks as something essentially social. Flichy [2000] recounts how networked technologies were historically meant to foster the emergence of social interactions. Inter-networks were originally developed as a social tool (and even a socio-cognitive tool) rather than a technical tool, following the vision of Licklider and Taylor [1968] where computers should, more than anything else, be human communication devices facilitating distributed cognition within groups of common interest.

Scholars have essentially looked at the integration of online communication in everyday life, and at the everyday life in online communities. Regarding the former, sociologists in particular have been questioning the nature of electronic interactions and their intertwinement with offline communication channels, studying for instance the socio-demographics of the Internet and its usage (in terms of civic engagement [Norris, 2001] or access to information [Kayahara and Wellman, 2007]), or the share of computer networks in overall communication [Wellman, 2001], emphasizing the increased social reality of these virtual interactions — "the more virtual the more real" [Woolgar, 2002]. With respect to the latter, Rheingold [1993]'s essay on "The Well", a San Francisco-based virtual community founded by two former hippies, contributed greatly to the acknowledgment of the existence of self-organized online societies of a novel nature, originally related to a certain digital utopia [Turner, 2006]. This opened the way to a surge in ethnographic research on online communities [Wilson and Peterson, 2002]. Hine [2000] describes the specificity of this research, where digital devices constrain the ethnographer's behavior in a manner not experienced before in the offline world, especially with respect to tracking conversations flowing in a very different manner than a face-to-face encounter) — as well as its commonality, in the sense that ethnographers must, as usual, adapt themselves to the culture they wish to study, understand their codes and their rules.

More broadly, each online community configures a potentially unique socio-technical system, in that each specific socio-technical setting can define a distinct society, with its own artificial rules. Consider again Wikipedia: its success made it a lively and hierarchized community of thousands of regular contributors evolving within an environment subject to its own (self-appointed) rules, enforced by a technical interface whose functioning is decided by the community itself. As a result and as said before, the collaborative encyclopedia has been the focus of many studies at the interface of various disciplines: sometimes leaning more towards natural sciences, by considering Wikipedia's website(s) as an almost physical system [Zlatic et al., 2006, Capocci et al., 2006], sometimes leaning more towards sociological issues, by exploring Wikipedia as a radically new community where the interface plays a strong role in disciplining both socialization and content creation (for instance by emphasizing the existence of either entirely novel interaction rules [Adamic et al., 2010, Chen and Roth, 2012] or rather traditional social processes, such as apprenticeship [Bryant et al., 2005] and internal democracy [Leskovec et al., 2010]). Blogs define a different ecosystem [Karpf, 2008], being both a public space [Etling et al., 2009] and a conversation arena [Herring et al., 2005] animated by peculiar temporal patterns [Goetz et al., 2009]; micro-blogs (such as twitter) borrow partly to the blogsphere but define yet again a distinct socio-cognitive environment [Kwak et al., 2010], with its own information diffusion processes [Weng et al., 2012]

and reputation dynamics [Marwick and d. boyd, 2011]. It is far beyond the scope of this paper –let alone this section– to provide a comprehensive overview on the wealth of themes currently structuring the humanities of the digital, even in a very succinct manner. Let us nonetheless close this partial enumeration with the evocation of gaming, whose practice also belongs in full to the scope of this stream [Mäyrä, 2008], for it offers a perhaps extreme illustration of the particularity of online socio-technical systems where entire universes, often driven by unrealistic rules are being populated by agents who spontaneously dialogue, form groups (guilds), solve problems (quests) together [Williams et al., 2011].

3.2 A bridge to digitized and numerical humanities

On the whole, this perspective on digital humanities seems to be apparently orthogonal to the two previous ones: although interdisciplinary, this study field has a very specific and bounded object focusing on human-computer and computer-mediated interactions. Yet, it actually touches to the respective cores of digitized and numerical humanities:

1. First, because internet corpuses, i.e. corpuses produced by the web itself, are readily available and, almost by design, already digitized: virtually all online actions and interactions are recordable in a native format, theoretically at any level of granularity.

   The impact of online data is for instance significantly visible in linguistics, as emphasized by Kilgarriff and Grefenstette [2003] in their introduction to the special issue of *Computational Linguistics* dedicated to the web as a corpus (for an overview, see also Hund et al. [2007]). It is also an established means, as Rogers [2011] expresses it, to "study culture and society with the Internet". Ruppert et al. [2013] give a comprehensive account of its contribution to the evolution of social science methods as a whole, including the possibility of observing actions and transactions (instead of documenting and interpreting stories), on well-delimited populations (featuring easier uniform recruitment procedures), with possibly as many timepoints as needed.

   In practice, online communities also appear to constitute the bulk of the data exploited by numerical humanities [Lazer et al., 2009, Lazer and Radford, 2017]. Here, however, just as for the other types of datasets that they are using, computational social scientists are again rarely aware of the good practices developed in digitized humanities regarding the questions of archive formatting or long-term availability. This data is not less interchangeable, it is a priori reproducible, but, what is more, it can often be easily re-collected or re-exported from a given platform or set of websites as long as they still exist — the web sometimes appears, rather illusorily, to archive itself automatically: why would one worry about securing a specific dataset?

2. Second, because the understanding of computer-mediated interactions is profitable for both stances. As regards digitized humanities, it sheds light on the prospects and limitations of the usuability and viability of interfaces and online research communities in dealing with digitized archives. Of prime interest, here, are the social and technical aspects of virtual collaboration groups dedicated to the collaborative constitution or curation of archives. It is indeed key to have a good knowledge of the sociological phenomena occuring in communities driven by voluntary participation [Mockus et al., 2002, Marlow et al., 2006, Gallant et al., 2007], as well as being aware of

the advances in human computing and "games with a purpose" [Ahn, 2006] and participative archives [Huvila, 2008] or crowdsourced curation [Oomen and Aroyo, 2011]. As more and more digital humanists consider that it is also essential to be a connected scientist e.g., by managing a research blog or Twitter account, the relevance of studies on information dynamics and social hierarchization in blogspace also increases.

With respect to numerical humanities, the analysis of the dynamics of online communities is often considered as a proxy for the *in vivo* observation of societies on an unprecedented scale. It enables, in principle, the formulation of testable and sometimes replicable hypotheses on social processes and human behavior at large. Many of the above-studies on Wikipedia, blogs, Twitter or gaming platforms belong indeed to a computational social science endeavor: looking at the evolution of the structure of a self-organized system (of Wikipedia webpages), the behavioral regularity of information producers (in the case of bloggers), content epidemiology across a network (of Twitter users), and typical interaction graph topology (of online gamers).

While online life offers a fertile research playground to both of the above-described perspectives, one may thus wonder which place occupies this strand in venues commonly affiliated with either "digital humanities" or "computational social science". Let us now examine this situation in concrete terms and, more precisely, the respective share of each stance in the various venues.

4 Digital overlaps in practice

DH and the digital, digitized, and numerical

Admittedly, the three above-described understandings induce distinct epistemological underpinnings and, possibly, distinct scientific objectives and audiences. I would like to briefly sketch their concrete presence within fields which are explicitly related to the DH label. To this end, I shall first look at two distinct types of venues for "digital humanities". The choice of the empirical material could be debated, as it relies on relatively arbitrary decisions. Even if I intend to make no claim of strong statistical representativity (to be fair, it should be the aim of a rigorous scientometrics paper devoted to this issue), I contend that the following rough assessment already provides a convincing glimpse on the coexistence of the three acceptations of DH, and their instantiation in different scientific communities.

To start with, I chose the journal "Digital Humanities Quarterly" (DHQ) which has one of the oldest history in the field as a review specifically referring to the DH label. As mentioned in the introduction, "Digital Scholarship in the Humanities" (DSH) also has a long history, it operated however under the name "Literary and Linguistic Computing" (LLC) until the end of 2013. I also harvested recent data for this journal. It is published by the Association for Computers and the Humanities (ACH), which organizes the yearly international DH conference. For this reason, my third venue is the 2015 edition of this conference, held in Sydney, Australia, later denoted by DH2015.

I then examined the representativity of the three stances by carrying out a manual categorization of papers published in recent issues over five years in DHQ and DSH/LLC (covering both 2012-16), to the exclusion of editorial / introductory papers. I also considered abstracts of talks presented at DH2015.

1. For DHQ, I specifically considered seventeen issues 6.2 to 10.4, i.e. from the second issue of 2012 up to the last issue (included) of 2016. This makes a total of 113 articles, of which 16 were providing reflexive viewpoints on the history, epistemology or carreers of DH as a field; they were thus excluded from the categorization.

   The final selection of 97 articles consists of:

   - 84 articles (86.6%) attributed to "digitized humanities", with topics such as: archive curation, management, access, formatting, status of digital artifacts, digital art; sixteen of them also deal with the application of existing AI techniques on archives for the purpose of a specific case study (PCA, OCR, SNA, and lexicometry);
   - 4 articles (4.1%) attributed to "numerical humanities": here, namely, one paper focused on agent-based modeling, one on information contagion models, two on the possibility of a statistical characterization of style or automatic assignment of gender;
   - and 9 articles (9.3%) for the "humanities of the digital", featuring three discussions on the Web 2.0, one on game studies, two on coding practices, one on the use of hashtags in Twitter, one on using mobile apps in research, and one paper on online reading groups.

2. For DSH/LLC, I considered the first issues of each volume published between 2012 and 2016, i.e. a total of five issues. DSH/LLC issues contain more articles than DHQ, so this made a total of 47 articles, or 46 after excluding one paper discussing the history of DH.

   - 34 articles (73.9%) to "digitized humanities". A large proportion of them deals however with the development of maps or linguistic classifiers (e.g., detection of verbs or some morphemes) often through the application of existing clustering or statistical methods on digitized corpora. Although relying more on numerical methods than in the case of DHQ, the main focus still lies on a specific corpus or collection of corpuses, rather than the abstraction of new models or the development of new quantitative methods per se. In contrast to DHQ, only few in this category focus principally on a specific archive, dataset, or visualization software or approach.
   - 10 articles (21.7%) to "numerical humanities". This category principally contains papers which introduce a stylometric model, rather than just applying it — often by proposing a model of language generation, hence involving a behavioral mechanism. Only one paper focuses on the modeling of cycles in a given society, while two papers discuss laws behind shot length in films.
   - and 2 articles (4.3%) to the "humanities of the digital". One addresses the use of mobile apps to do language research, the other proposes a case study on the use of a text mining software.

3. For DH2015, I subsequently considered 48 (long) abstracts, randomly chosen within the list of talks found in the conference program. I then attributed:

   - 39 articles (81.2%) to "digitized humanities", including 10, this time, featuring the use of quantitative or AI tools;

- 7 papers (14.6%) to "numerical humanities", which were almost systematically involving narrative models: even though they often deal with classification tasks (e.g. authorship attribution), these papers make the assumption that text production may be described with the help of generic models);
- and 2 papers (4.2%) to the "humanities of the digital": while one paper deals with Twitter, the other one analyzes crowdsourcing communities for a task of database curation.

To summarize, an overwhelming majority of papers had to deal with the "digitized" understanding of DH. Very few papers dealt with internet life and corpuses, and not many more had to do with the "numerical" stance — when they did, they essentially had to do with stylometry (8 out of 14). In this respect, if DSH/LLC and, to a lesser extent, DH2015 featured more numerical humanities than DHQ, the corresponding papers were almost all framed within a literary research endeavor. This seems to support further the notion that, in spite of occasional epistemological similarities, there are two distinct communities of practice (humanists and social scientists).

Computational social science and DH

By comparison, I also looked at a large and plausibly representative computational social science venue. To this end, I considered the most recent international conference explicitly referring to this subject: the International Conference on CSS, which a bit more than 400 participants attended in Helsinki during the summer of 2015. I focused on talk titles in the conference program for which the corresponding papers (i.e. with an identical title) were readily available as an open-access article: in order to eventually categorize 48 talks, I had to randomly pick 62 talks (i.e. more than three quarters of the original random selection of titles had been self-archived prior to the conference).

The empirical results yield the contours of a CSS landscape which stands in stark contrast with that of DH. A majority of the presentations (29, i.e. 60.4%) had to do with the "humanities of the digital", including 24 talks (exactly half of the whole sample) which featured a markedly numerical approach, from multi-agent models to behavioral modeling, through the development of original and flexible machine learning models. Of the remaining 19 papers not related to online communities, 17 (35.4%) could be ascribed to "numerical humanities" proper. While no talk focused on archiving issues per se, two consisted in applying existing quantitative tools on archives — digitized humanities thus accounted for a small share of 4.2% of all talks.

On the whole, a total of 5 + 2 talks did not qualify as a truly numerical approach (as defined earlier). They had nonetheless a distinguishable numerical flavor, by contrast with the similarly-categorized DHQ-DSH/LLC-DH2015 papers. Here, these seven papers often consisted in applying a series of existing pattern recognition, regression or classification tools (especially when internet data was involved). The used toolbox was virtually always discussed, i.e. a significant part of the paper was devoted to presenting and adapting the quantitative method to the case study (e.g. working on a binary classification algorithm and discussing its performance in machine learning terms), while many coauthors had a primary background in computer science or statistical physics.

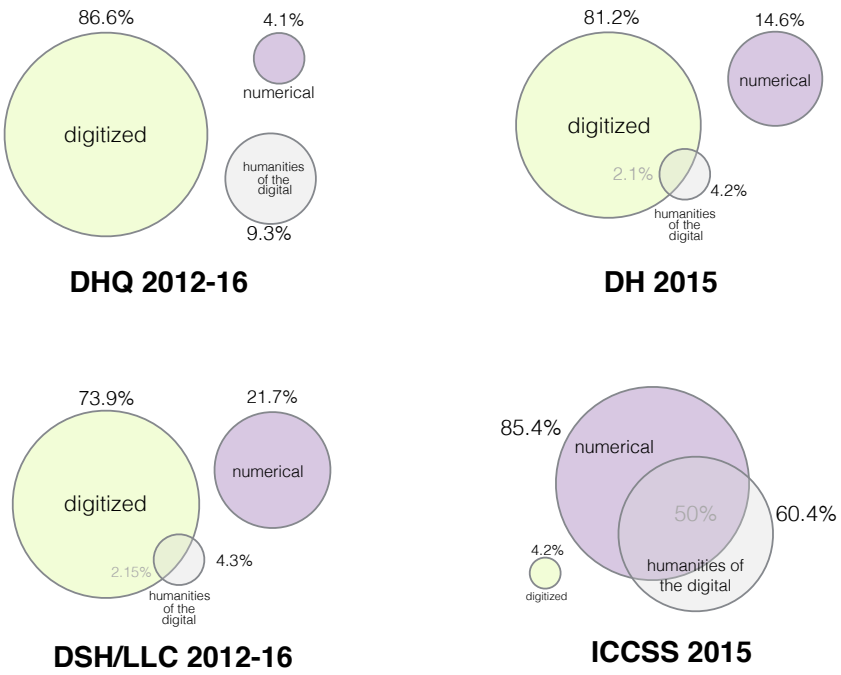

Figure 1: Intersection diagram representing the share of papers dealing with humanities of the digital and relying on an approach typical of either digitized or numerical humanities, for each of the three considered venues referring respectively to the label "digital humanities" (DHQ and DSH/LLC over 2012-16, and DH 2015 in Sydney) or "computational social science" (ICCSS 2015 in Helsinki).

This exploration of recent research self-labeled as CSS clearly shows that the humanities of the digital have become, in this field, a somewhat essential category, in contrast to the situation observed in DH-labeled venues. This is all the more visible by drawing Venn diagrams depicting the intersections of papers relying on internet corpuses and borrowing to either digitized or numerical humanities approaches — see Fig. 1.

Concluding remarks

This last empirical examination of a bit more than 250 research studies diversely affiliated to DH or CSS confirms the existence of two non-overlapping epistemic communities with respect to digital methods. In basic terms, this partition may be characterized by a double dichotomy based, on the side of objects, on the *attention* given to the humanities of the digital and, on the side of methods, on the *status* both of models and of archives.

It is admittedly a conceivable situation. The existence of two distinct epistemological settings and objectives may indeed suffice to explain the *de facto* historical development of two disconnected streams, sometimes reinforced by a *de jure* delimitation between two understandings where either qualitative or quantitative approaches are put forward [Porsdam, 2013]. Yet, it is also a non-desirable situation. CSS could certainly improve their curiosity for the variety of objects addressed by digital humanists (about 15% of the ICCSS'15 communications were singularly focused on Twitter) as well as profit from their field expertise and, perhaps, finer knowledge of the conditions of construction of the datasets. DH would

not be last to benefit from the nomological attitude of computational social scientists and their flexible modeling skills. In this regard, DH might also be currently at the periphery of a revolution which is related *in a nomological way* to big data (especially big internet data) and which enjoys a notable momentum in social sciences. It is also one of the teachings of the share of the "humanities of the digital" in CSS topics.

By making this state of affairs visible, this contribution hopes to give some hints about the causes of this relative yet definitely perceivable divide, by articulating the connections between the various understandings. Beyond the rather natural idea of encouraging the emergence of more joint venues and networks appealing to both scientific communities, this chapter finally intends to insist on the possible broker role of the scholarship *on* online communities — that is, humanities of the digital bridging the gap between digital humanities and numerical humanities.

References

Lada A. Adamic, Xiao Wei, Jiang Yang, Sean Gerrish, Kevin K. Nam, and Gavin S. Clarkson. Individual focus and knowledge contribution. *First Monday*, 15(3):1, 2010.

Luis von Ahn. Games with a purpose. *IEEE Computer*, 39(6):92–94, 2006.

Robert Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press, Princeton, New Jersey, 1997.

Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *PNAS*, 99(3):7280–7287, 2002.

R. Boulet, B. Jouve, F. Rossi, and N. Villa. Batch kernel SOM and related Laplacian methods for social network analysis. *Neurocomputing*, 71(7-9):1257–1273, 2008. doi: doi: 10.1016/j.neucom.2007.12.026.

Alban Bouvier. Connaissance de l'individuel et sciences du général. une comparaison entre sciences de l'homme en société et sciences de la nature. In Thierry Martin, editor, *La scientificité des sciences humaines*, pages 35–52. Vuibert, Paris, 2010.

Arnaud Bringé, Marie-Hélène Cazes, Pascal Cristofoli, Isabelle Daillant, Laurent Gabail, Anne Garcia-Fernandez, Michael Gasperoni, Cyril Grange, Klaus Hamberger, Vincent Hirtzel, Michael Houseman, Olivier Kyburz, and Ismaël Moya. Kinsources and puck: Open data and free tools for analyzing kinship networks. Project description poster presented at the 1st European Conference on Social Networks, U. Barcelona, Jul 3, 2014, 2014.

C. Broodbank. *An island archaeology of the early Cyclades*. Cambridge University Press, Cambridge, 2000.

Susan L. Bryant, Andrea Forte, and Amy Bruckman. Becoming wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Group'05, Sanibel Island, FL, USA*, Nov 6-9 2005.

A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical Review E*, 74 (3):036116, 2006.

Laura Carletti, Gabriella Giannachi, Dominic Price, Derek McAuley, and S Benford. Digital humanities and crowdsourcing: An exploration. In *Proc. MW2013: Museums and the Web 2013, Portland, Ore.*, Apr 2013.

D. Cartwright and F. Harary. Structural balance: A generalization of Heider's theory. *Psychological Review*, 63:277–292, 1956.

T. Chandler. *Four thousand years of urban growth: an historical census*. Edwin Mellon Press, Lewiston, NY, 1987.

Chih-Chun Chen and Camille Roth. {{citation needed}}: The dynamics of referencing in wikipedia. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, pages 8:1–8:4, New York, NY, USA, 2012. ACM.

Nicolas Claidière and Dan Sperber. Commentary: The role of attraction in cultural evolution. *Journal of Cognition and Culture*, 7: 89–111, 2007.

J. Conolly and M. Lake. *Geographical Information Systems in Archaeology*. Cambridge Manuals in Archaeology. Cambridge University Press, 2006.

R. Conte, N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J.-P. Nadal, A. Sanchez, A. Nowak, A. Flache, M. San Miguel, and D. Helbing. Manifesto of computational social science. *European Physical Journal Special Topics*, 214(1):325–346, 2012.

Christofer R. Edling. Mathematics in sociology. *Annual Review of Sociology*, 28:197–220, 2002.

Joshua M. Epstein and Robert Axtell. *Growing artificial societies: social science from the bottom up*. The Brookings Institution, Washington, DC, USA, 1996. ISBN 0-262-55025-3.

Bruce Etling, John Kelly, Rob Faris, and John Palfrey. Mapping the Arabic blogosphere: Politics, culture, and dissent. Technical Report 2009-06, Berkman Center Research Publication, 2009.

Patrice Flichy. Internet or the ideal scientific community. *Réseaux – The French Journal of Communication*, 7(2):155–182, 2000.

Linda M. Gallant, Gloria M. Boone, and Austin Heap. Five heuristics for designing and evaluating web-based communities. *First Monday*, 12(3), 2007.

Fakhteh Ghanbarnejad, Martin Gerlach, José M. Miotto, and Eduardo G. Altmann. Extracting information from s-curves of language change. *Journal of The Royal Society Interface*, 11(101), 2014. ISSN 1742-5689.

Nigel Gilbert and Klaus G. Troitzsch. *Simulation for the Social Scientist*. Open University Press, Buckingham, 1999.

Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory*. Aldine, Chicago, 1967.

Michaela Goetz, Jure Leskovec, Mary McGlohon, and Christos Faloutsos. Modeling blog dynamics. In *ICWSM 2009 Proc. 3rd International AAAI Conference on Weblogs and Social Media*, 2009.

Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

Dirk Helbing, editor. *Social Self-Organization*. Springer, 2012.

Dirk Helbing and Stefano Balietti. Agent-based modeling. In Dirk Helbing, editor, *Social Self-Organization*, pages 25–70. Springer, 2012.

Susan C. Herring, Inna Kouper, John C. Paolillo, Lois Ann Scheidt, Michael Tyworth, Peter Welsch, Elijah Wright, and Ning Yu. Conversations in the blogosphere: An analysis "from the bottom up". In *Proceedings of the Thirty-Eighth Hawai'i International Conference on System Sciences (HICSS-38)*, 2005.

Christine Hine. *Virtual Ethnography*. Sage, London, 2000.

Christine Hine. Social research methods and the internet: A thematic review. *Sociological Research Online*, 9(2), 2004.

Susan Hockey. The history of humanities computing. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A companion to Digital Humanities*, chapter 1, pages 3–19. Blackwell Publishing, 2004.

David I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

Jeremy Huggett. Core or periphery? digital humanities from an archaeological perspective. *Historical Social Research*, 37(3): 86–105, 2012.

Norman P. Hummon and Kathleen Carley. Social networks as normal science. *Social Networks*, 15(1):71–106, 1993.

Marianne Hund, Nadja Nesselhauf, and Carolin Biewer. *Corpus linguistics and the web*. Rodopi, Amsterdam-New York, 2007.

Isto Huvila. Participatory archive: towards decentralised curation, radical user orientation, and broader contextualisation of records management. *Archival Science*, 8.1:15–36, 2008.

David Karpf. Understanding blogspace. *Journal of Information Technology & Politics*, 5 (4):369–385, 2008.

Jennifer Kayahara and Barry Wellman. Searching for culture—high and low. *Journal of Computer-Mediated Communication*, 12(3):824–845, 2007.

Graeme Kennedy. *An introduction to Corpus Linguistics*. Longman, 1998.

Adam Kilgarriff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29 (3):333–347, 2003.

Harald Klein. Overview of text analysis software. *Bulletin de Méthodologie Sociologique*, 70:53–66, 2001.

Carl Knappett, Tim Evans, and Ray Rivers. Modelling maritime interaction in the aegean bronze age. *Antiquity*, 82(318): 1009–1024, 2008.

H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. 19th Intl. Conf. on World Wide Web WWW'10*, pages 591–600, Chicago, April 2010. ACM.

David Lazer and Jason Radford. Data ex machina: Introduction to big data. *Annual Review of Sociology*, 43:19–39, 2017.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915): 721–723, 2009.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Governance in social media: A case study of the wikipedia promotion process. In *Proc. AAAI ICWSM Intl. Conf. Weblogs and Social Media 2010*, pages 98–105, 2010.

J.C.R. Licklider and R.W. Taylor. The computer as a communication device. *Science and technology*, 1968.

F. Lorrain and Harrison C. White. Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1(49–80), 1971.

R. Duncan Luce and Albert Perry. A method of matrix analysis of group structure. *Psychometrika*, 14:95–116, 1949.

Cameron Marlow, Mor Naaman, danah boyd, and Marc Davis. HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. In *Proc. of Hypertext 2006*, New York, 2006. ACM Press.

A. E. Marwick and d. boyd. I tweet honestly, i tweet passionately: Twitter user, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133, 2011.

Frans Mäyrä. *Introduction to Game Studies: Games in Culture*. Sage Publications, 2008.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific Reports*, 3 (1950):1–7, 2013.

Audris Mockus, Roy T. Fielding, and James D. Herbsleb. Two case studies of open source software development: Apache and Mozilla. *ACM Transactions on Software Engineering and Methodology*, 11(3):309–346, 2002.

Franco Moretti. Conjectures on world literature. *New Left Review*, 1:54–68, 2000.

Dhiraj Murthy. Digital ethnography: An examination of the use of new technologies for social research. *Sociology*, 42(5):837–855, 2008.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26, 2007.

Bob Nicholson. Exploring the methodological possibilities of digital newspaper archives. *Media History*, 19(1):59–73, 2013.

Pippa Norris. *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide*. Cambridge University Press, 2001.

Johan Oomen and Laura Aroyo. Crowdsourcing in the cultural heritage domain: opportunities and challenges. In *Proc. 5th Intl. Conf. on Communities and Technologies*. ACM, June 2011.

Roel Popping. *Computer-Assisted Text Analysis*. Sage Publications, 2000.

Helle Porsdam. Digital humanities: On finding the proper balance between qualitative and quantitative ways of doing research in the humanities. *Digital Humanities Quarterly*, 7(3), 2013.

Anatol Rapoport. Spread of information through a population with socio-structural bias: I. Assumption of transitivity. *Bulletin of Mathematical Biology*, 15(4):523–533, 1953.

H. Rheingold. *Virtual Communities: Homesteading on the Electronic Frontier*. MIT Press, http://www.rheingold.com/vc/book, 1993.

R. Rogers. Das ende des virtuellen – digitale methoden. *Zeitschrift für Medienwissenschaft*, 5:61–77, 2011.

Camille Roth, Floriana Gargiulo, Arnaud Bringé, and Klaus Hamberger. Random alliance networks. *Social Networks*, 35(3): 394–405, 2013.

E. Ruppert, J. Law, and M. Savage. Reassembling social science methods: The challenge of digital devices. *Theory, Culture & Society*, 30(4):22–46, 2013.

Paul A. Samuelson. *Foundations of Economic Analysis*, volume 80 of *Harvard Economic Studies*. Harvard University Press, Cambridge, MA, 1947.

Thomas C. Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1:143–186, 1971.

Stanislav Sobolevsky, Michael Szell, Riccardo Campari, Thomas Couronné, Zbigniew Smoreda, and Carlo Ratti. Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLoS ONE*, 8(12):e81707, 2013.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

Roberto Tamassia. *Handbook of Graph Drawing and Visualization*. Discrete Mathematics and Its Applications. CRC Press, 2013.

F. Turner. *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. The University of Chicago Press, 2006.

Maarten van den Bos and Hermione Giffard. Mining public discourse for emerging dutch nationalism. *Digital Humanities Quarterly*, 2016.

Christopher N. Warren, Daniel Shore, Jessica Otis, Lawrence Wang, Mike Finegold, and Cosma Shalizi. Six degrees of francis bacon: A statistical method for reconstructing large historical social networks. *Digital Humanities Quarterly*, 10(3), 2016.

B. Wellman. Computer networks as social networks. *Science*, 293:2031–2034, 2001.

Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Filippo Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2:335, 2012.

Douglas R. White, Laurent Tambayong, and Natasa Kejzar. Oscillatory dynamics of city-size distributions in world historical systems. In George Modelski, Tessaleno Devezas, and William R. Thompson, editors, *Globalization as Evolutionary Process: Modeling Global Change*, pages 190–225. Routledge, 2007.

Dmitri Williams, Noshir Contractor, Marshall Scott Poole, Jaideep Srivastava, and Dora Cai. The virtual worlds exploratorium: Using large-scale data and computational techniques for communication research. *Communication Methods and Measures*, 5(2):163–180, 2011.

S. M. Wilson and L. C. Peterson. The anthropology of online communities. *Annual Review of Anthropology*, 31:449–467, 2002.

S. Woolgar. *Virtual society*, chapter Five rules of virtuality, pages 1–22. Oxford University Press, 2002.

V. Zlatic, M. Bozicevic, H. Stefancic, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74(1):016115, 2006.