Sometimes an academic field is defined by a “great project”—a laudable and generous goal, shared by all or most members of the field, that determines the aim and scope of its work for years and sometimes decades. For social and cultural anthropology, from the postwar years to around the 1980s, that project was to represent, through the method of participant observation and the genre of ethnography, the planet’s great diversity of peoples, languages, and cultures, which were rapidly being transformed or destroyed by the expansion of the world system. The product of that great collective labor is a vast ethnographic record, comprising essays and monographs focused on specific communities and linguistically or culturally uniform regions. Although sometimes these ethnographies were focused on a specific aspect of culture, such as language or ritual or economics, the goal was always to create a confederated and inclusive atlas of world cultures, even as efforts to formally centralize these efforts, such as Yale’s Human Relations Area File, were not widely embraced by the field. Today, anthropology has moved on from this goal. One reason is that, since the 1980s, it has not been possible to frame research in terms of the retrieval and authentic representation of local societies, if it ever was valid to do so. Aside from the rise of critical and postcolonialist perspectives that led to an inward and more literary turn in the field, the situation in anthropology was produced by a change in the subject of anthropology itself. Although at one time it seemed possible to filter out the influence of Christian missionaries on the beliefs of, say, a community of headhunters, it became impossible to ignore the effects of chainsaws felling their trees.

In the digital humanities, we too have been involved in a great project. In the early days of the field, back when it was called humanities computing, that project was the retrieval and remediation of the vast collection of primary sources that had accumulated in our libraries and museums and, in particular, those textual sources that form the foundation of two fields that define, along with philosophy, the core
of the humanities: literature and history. The signature offering of this project was the digital collection, exemplified by such projects as Edward Ayers' *Valley of the Shadow*, which would evolve into what Unsworth and Palmer called the “thematic research collection” and what others would label, with some degree of inaccuracy, the “archive.” Almost everything that characterized the field prior to its rebranding as digital humanities can be related to this project: the work of text encoding; the concern for textual models and formal grammars (a side effect of and motivation for encoding in SGML and XML); a parallel but less intense focus on image digitization; the desire to develop effective digital critical editions; the inclusion of librarians and academic faculty under the same umbrella; the eventual development of tools like Zotero, Omeka, and Neatline; the interest in digital forensics (the need for which became apparent to those actually building these archives); and so forth. Even speculative computing, one of the most innovative branches of humanities computing, led by Drucker and Nowviskie, developed on top of this fundamental concern for the textual archive.

Just as anthropology’s project was undone by the overwhelming forces of globalization, so too has that of the digital humanities, though perhaps without the horror. In our case, the pervasive technological changes associated with Web 2.0 and big data—two marketing labels that nonetheless index very real manifestations of globalization within the datasphere—have altered our great project by shifting the foundations on which it had long been built. In place of the vertical archive has emerged the horizontal networked database, both as a symbolic form that thwarts the will to narrative and unseats the prominence of ontology, and as a platform of participation that decenters the local project. Partly as a result of this shift, the older concern for well-curated collections founded on philosophically satisfying models has been displaced by a wider range of concerns, from an embrace of data science and distant reading to the exploration of maker labs and the Internet of Things to an engagement with the public humanities on a series of political and even eco-theological fronts.

Among these concerns, perhaps the most profound has been the engagement with data science. Anyone who has attended the annual DH conference over the years will have noticed the change. Statistical criticism has always been a feature of these conferences and the field, but the number of presentations describing the use of text analytics methods (such as topic modeling) has increased dramatically, to the point where large portions of the program guide could be mistaken for an IEEE conference. The change became visible around 2009, when the NEH’s Office of Digital Humanities announced its Digging into Data Challenge, which asked digital humanists to answer the question, posed by Crane in 2006, “What do you do with a million books?” A year later, Google’s N-Gram Viewer was revealed to the world, which provided its own answer to the challenge, although it was developed independently of the NEH initiative. By 2011, the noted historian Anthony Grafton, then president of the American Historical Association, would write about
an “astonishing” encounter with the creators of the viewer and its associated theoretical framework, “culturomics.”

Although the impact of data science on the digital humanities can be measured by the sheer volume of attention that has shifted toward the newer methods, the greatest effect has been a reorientation of the field's most fundamental practice: the production and use of digitized primary sources, usually in the form of text. In place of the digital scriptorium, in which scholars painstakingly mark up texts according to well-conceived schema (such as those of the Text Encoding Initiative; TEI) to surface structure and semantics, there has emerged the data-mining operation in which, for example, texts are converted into “bags of words” and vectorized for use in a variety of black-box procedures developed by systems engineers and computer scientists. In place of the concern for the establishment and criticism of individual texts, the text itself has been “unbundled” and replaced by other containers—the paragraph or text segment, the year of publication—which then become the units of interpretation. At no point is the difference between these practices more clear than when a text miner, making use of a legacy archive of marked-up documents, first strips its texts of all markup—often representing years of labor—regarding it as noise.

This shift in our orientation toward text has not been total—text markup continues to be a core practice within the digital humanities—but it has produced something of a shake-up in the field that remains surprisingly unnoted. At the very moment when digital humanities is on the tip of everyone’s tongue, a tagline to every IT-related grant and initiative within the liberal arts, its identity is at risk. For the practice of text encoding, limited as it may seem in light of the field’s new developments, remains the ancestral practice of digital humanities, and its displacement by methods whose mathematical and computational requirements are far beyond the scope of most humanists’ training must be regarded as a kind of crisis. What now distinguishes the digital humanist from the data scientist interested in working on historical or literary texts, especially when the former actually knows what is in the black box? What specific expertise does the digital humanist bring to the interdisciplinary table? Recall that the inventors of culturomics are biologists, not historians. To be sure, the data scientist may retain the humanist as a subject matter or “domain” expert—but that fits no one’s definition of digital humanities.

Adeline Koh’s notorious “DH Will Not Save You” post touched on this issue, although from the opposite angle. She chastised digital humanists for privileging computation over culture, a move that can only push the humanities into further irrelevance by becoming a “handmaiden to STEM.” But Koh replaces one form of redundancy with another. Instead of being overshadowed by engineers, scientists, and mathematicians, digital humanists are asked, in effect, to be more like scholars in media studies, Science, Technology, and Society Studies (STS), or some variant of cultural studies. This vision of digital humanities misrecognizes the central eros of the field: the ludic and critical engagement with computational and cultural forms,
the “situation” created by engaging in the collaborative and iterative work of interpretation by means of digital media (Alvarado, “Digital Humanities Situation”). Such work is neither merely technophilic nor purely critical; it reflects the authentic and perhaps naïve desire of humanists to work with digital technology on their own terms. The result has been both critical and practical, a liminal mixture that will not satisfy the purist and remains easy to mischaracterize by outsiders.

Among the concerns to emerge in the space opened up by the digital humanist’s engagement with data science is one that both continues in the spirit of the earlier focus on digital collections and text encoding and that promises to lay the groundwork for an inclusive and fruitful research agenda commensurate with the historical and literary alignments of digital humanities. This is the work of operationalization, highlighted by Moretti in the 2013 essay “Operationalizing,” a term of art from data science that refers to a specific way of representing knowledge for machine use. Although the word has its origins in the natural sciences, referring to the practice of defining the observable and measurable indices of a phenomenon (or its “proxies”) so that it may be studied experimentally and quantitatively, Moretti generalizes the idea to include the practice of translating a received concept or theory (typically about society, culture, or the mind), into machine-operable form. As an example he describes the contrasting ways that Woloch’s concept of character-space—which defines the amount of attention a character receives in a novel in terms of the number of words used to represent it—can be translated into specific metrics suitable for machine processing. Other examples include Finlayson’s conversion of Propp’s theory of the folktale, which defines stories as sequences of elementary narrative “functions,” into a machine-learning algorithm operable on a collection of (heavily) annotated texts; the use of Lakoff and Johnson’s theory of metaphor, which emphasizes the importance of the body in the creation of metaphors, to classify motion-captured human movements (Wiesner et al.); and the effort by Mohr et al. to translate Kenneth Burke’s “grammar of motives,” which provides a “dramatistic” framework for describing the attribution of motives, into an automated process for text encoding and analysis. Such projects, diverse as they are, share the trait of appropriating an existing theory more or less as is and translating it into computational form to meet a research project’s requirement to achieve some level of coherency with its material. The theory serves as a resource for the development of an ontology—in the narrow, computational sense of a “formalization of a conceptualization”—that may be used for a variety of practical purposes, such as the definition of a database schema or the writing of class libraries to process a project’s data.

Among operationalization’s useful and interesting consequences, Moretti emphasizes the critical opportunities that arise from the work of translating a discursively constituted idea into machine-readable code. To demonstrate, he translates Hegel’s theory of tragic opposition, which describes the process by which equally valid human values come into conflict, and notes that the work of operationalization itself can actually cause us to rethink the original theory in refreshing
ways, even if in retrospect we may imagine having arrived at the new perspective by other means. Here Moretti echoes Unsworth’s earlier observation, experienced by many, that “there’s definitely something to be learned from the discipline of expressing oneself within the limitations of computability” (Unsworth, paragraph 2), as well as the larger point made by Drucker and Nowviskie that “[d]igital humanities projects are not simply mechanistic applications of technical knowledge, but occasions for critical self-consciousness” (432). For these thinkers, operationalization produces a rationalization effect, a disruption of tacit knowledge caused by the computer’s representational demand for explicit, discrete, and often reductive categories, which frequently requires one to refine held ideas into clear and distinct form. Along the way, lively philosophical questions, long hidden in the foundations of an idea, are reopened for debate, since the coded representation of the original idea is never the only one possible, but inevitably demands choosing among alternatives.

The philosophical boon yielded by operationalization is enough to establish it as a core practice in the digital humanities. But operationalization promises more than an occasion to reflect on foundations: it may alter fundamentally the aims and increase the scope of DH research projects. By shifting focus from the remediation of content to the remediation of ideas (which have been developed to interpret that content), digital humanists may reconnect with the production of theory, an area where the humanities and interpretive social sciences have developed expertise (its excesses notwithstanding). Although digital humanists will always be invested in the building of digital collections, both individual thematic research collections and large digital libraries, operationalization allows us to build on the scale of these collections to pursue questions about how these materials will be interpreted to answer questions with merit beyond the formalism inherent in collection building. Imagine a project involving historians and literature scholars applying an operationalized version of Benedict Anderson’s thesis in Imagined Communities, regarding the effects of print capitalism on the formation of national consciousness. Such a project would take advantage of the substantial digital collections of novels and newspapers from the long eighteenth century and would employ data science methods such as topic modeling to represent and visualize nationalism in ways that might support or falsify Anderson’s claims. Instead of only building collections based on shared authorship, genre, provenance, or period, we might focus on how such categorized collections can be connected (via linked data protocols) and aggregated to pursue deep research questions that cut across these boundaries.

But perhaps most important, there is a critical opportunity opened up by an operational turn that should appeal to our desire for a more public humanities. By virtue of our collective familiarity with the body of social and cultural theory, digital humanists are in a position to evaluate the work of nonhumanistic data scientists who routinely grab the low-hanging fruit of social science to complete the narrative arc of the arguments they make without understanding the historical and theoretical contexts in which such ideas must be evaluated. In doing so, we can guard
against the great danger of operationalization: the selection and amplification of ideas whose main qualification for inclusion in an argument is their ease of being represented by digital means.

Kavita Philip’s account of the Indian national census database illustrates the point. In 2011, after eschewing the equivalent of what in the United States is called postracialism, the people and government of India decided to reintroduce the category of caste into the national census for the first time since 1931. In creating the database to capture this information—in the field, through form-driven interviews, and in the schema of a relational database—the developers drew, through imitation of the 1931 census, from the works of British ethnographer and colonial administrator, Herbert Hope Risley, including his 1908 study, *The People of India*. Apparently, the thinking among the software developers was that by the 1930s English anthropologists had reached a sufficiently advanced understanding of culture, caste, and race that Risley’s ideas would provide a sound foundation for the data model. After all, by this time, many anthropologists had moved beyond the more egregious theories of race that had characterized the discipline in the late nineteenth and early twentieth centuries. However, Risley was no Boas, the American cultural anthropologist who was an early critic of the race concept and dispelled many attempts to link physiological traits to behavior. In contrast, instead of viewing human anatomical variation of features such as head shape as results of environmental conditioning, Risley was strongly committed to notions of genetic determinism and the efficacy of anthropometry, and he saw caste as a reflection of these dimensions. Moreover, he believed endogamy to be more consistently practiced than we know to be the case. Because of this, the census database encoded not only a set of received categories about caste but also a particular understanding about the nature of categorization itself. In effect, the 2011 census operationalized and thereby naturalized an antiquated and dangerous understanding of the caste system, encoding in its data model a theory of caste to which no current stakeholder would subscribe, at least openly. But since one rarely questions the data model of a database, because there is no practice or discourse with which to have such a discussion in the public sphere, the silence of the model in effect establishes its transcendence.¹

At issue, then, in the digital humanist’s engagement with operationalization is the transmission of knowledge to succeeding generations. Which ideas and ontologies will be taught, and which will be forgotten? For operationalization is, whether practiced by the digital humanist or data scientist, especially at this juncture, a selective transducer of concepts and theories, an evolutionary conduit through which some ideas will survive and others will not. Many of these ideas will have social consequences, as the census database example tells us. As humanists, we should not accept the glib premise that the most easily operationalized ideas are the best ideas, but should instead engage in an overt and critical review of operationalization as a form of argument, even as we employ this form to test and explore a grand theory.
The digital humanities has before it the opportunity to engage in a new great project, the embracing of operationalization as a form of deep remediation. This project has several virtues, including being inclusive of the big tent, synthetic of theoretical traditions and new research agendas, critical of emerging forms of digital culture, and—perhaps above all—being both backwardly compatible with our great work in the building of thematic research collections and forwardly comparable with our engagement with data science and our generous vision of a public humanities.

NOTE

1. For a complete account of Philip’s talk, see Alvarado, “Purity and Data.”

BIBLIOGRAPHY


Crane, Gregory. “What Do You Do with a Million Books?” D-Lib Magazine 12, no. 3 (March 2006).


