AN ABSTRACT OF THE THESIS OF

Cole Crawford for the degree of Master of Arts in English presented on May 19, 2017.

Title: Respect the Gap: From Big to Boutique Data through *Laboring-Class Poets Online*

Abstract approved: _____

Megan Ward

This thesis consists of two major components. The first is *Laboring-Class Poets Online* (*LCPO*), a database-driven website that provides information about the more than 2,000 British laboring-class poets who published between 1700 and 1900 and their writing, lives, and literary relationships. I developed *LCPO* to demonstrate the importance of laboring-class writing to social and literary history by enhancing access to the underlying dataset, which numerous scholars have contributed to over the past thirty years. *LCPO* transforms this unstructured document into a relational database and provides new interfaces for users to query, visualize, and explore this rich source of biographical and bibliographic information. In the second written section, I argue that current big data understandings and implementations of databases fail to adequately meet the needs of many humanities scholars. I introduce the concept of boutique data as an alternative framework that respects the ambiguous gaps in humanities datasets. Using *LCPO* as an example of a boutique data project, I contend that a boutique approach to data better accommodates complex understandings of space and time as continuously unfolding events.

Respect the Gap: From Big to Boutique Data through *Laboring-Class Poets Online*

by
Cole Crawford

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Arts

Presented May 19, 2017
Commencement June 2017

Master of Arts thesis of Cole Crawford presented on May 19, 2017

APPROVED:

Major Professor, representing English

Director of the School of Writing, Literature, and Film

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Cole Crawford, Author

ACKNOWLEDGEMENTS

I must begin by thanking Megan Ward, my thesis advisor. Her patience, constructive criticism, and encouragement to experiment and push boundaries have enabled me to improve both my writing and digital skillset and complete this project, the first hybrid digital humanities thesis from the OSU School of Writing, Literature, and Film. I greatly appreciate the feedback and insight my committee members, Evan Gottlieb and Ray Malewitz, have provided throughout this process. I would also like to thank Eric Walkingshaw for serving as my Graduate Council Representative.

Outside of the department, thanks are due to Anne Bahde and Korey Jackson for their guidance and collaboration in digital project development during my internship with OSU Libraries and Press, and to Charles Robinson, for his quiet brilliance and enthusiasm for meaningful public scholarship. I would additionally like to acknowledge John Goodridge for his willingness to allow me to transform and carry on his life's work. Lastly, I must thank Bridget Keegan, my undergraduate thesis advisor and mentor, for her unwavering support of my academic career.

My final and most important thanks go to Mom, Dad, Brandon, and Julia. Thank you for continuously reminding me of the importance of family and my Midwest roots.

TABLE OF CONTENTS

LIST OF FIGURES

Respect the Gap: From Big to Boutique Data through *Laboring-Class Poets Online*

The emergence of the database as the predominant form of information storage and structure in the late twentieth century has inspired numerous scholars to respond to the database as both a data structure and a constructed cultural artifact. While there are numerous other terms that accurately describe different types of digital scholarship within the humanities, the database is the technology that undergirds all of them.[1] Databases provide a means of efficiently storing, querying, retrieving, and manipulating data, but they simultaneously imply that the world is quantifiable and computable[2]. From such a position, any real-world entity can be distilled into digital form if sufficient dimensions are captured and structured as discrete elements.

For Lev Manovich, writing in 2002, databases thus represent a fundamentally "new way to structure our experience" compared to reading a story or seeing a film (*New Media* 219). Textual and cinematic narratives emphasize certain elements through causation and narrative stress, whereas databases resist privileging any single record; meaning and interpretation suffuse narrative, while databases make no inherent claims as to the value or meaning of the information they present. These differences lead Manovich to provokingly frame these two media forms, database and narrative, as locked in existential conflict:

> As a cultural form, the database represents the world as a list of items, and it refuses to order this list. In contrast, a narrative creates a cause-and-effect trajectory of seemingly unordered items (events). Therefore, database and narrative are natural enemies. Competing for the same territory of human culture, each claims an exclusive right to make meaning out of the world. (*New Media* 225)

From this perspective, digital media have elevated the database from a possible worldview to a near-perfect form for structuring information, one which threatens to eradicate narrative from new media productions entirely.

Manovich's argument for the database as the ascendant form of modern media hinges on the opposition he finds between his two terms – narrative and database – which persists beyond his use of them. Ed Folsom extends Manovich's metaphor of database-narrative conflict in a special issue of *PMLA* devoted to remapping genre (2007). For Folsom, genre and canon constrain texts because through systematization and standardization in pursuit of consistency, they necessarily exclude what is random, messy, and tangential. Database functions largely metaphorically for Folsom, providing an opportunity to explore vast amounts of information by building on "a universality of particulars" rather than flattening detail (1574). Database is not only a new genre of narrative descended from the winding and comprehensive epic, but also a new cultural form which challenges narrative genres entirely. Responding to Folsom in 2007 and in expanded form in 2012, N. Katherine Hayles rejects both Manovich's framing of database and narrative as "natural enemies" and Folsom's techno-utopian extension of this conflict metaphor. Instead, she sees narrative and database as "natural symbionts" that engage in a "mutually beneficial" relationship (176). The database can organize vast amounts of information but still requires the interpretive power of "narrative to make its results meaningful," while narrative needs database to "enhance its cultural authority and test the generality of its insights" (Hayles 176). Through their inclusivity, databases provide scale to narratives, and through their selectivity, narratives create meaning from data.

Hayles' concept of symbiosis more accurately represents the relationship between narrative and database than Manovich's and Folsom's figurations of conflict; these cultural forms will coexist rather than database succeeding and extinguishing narrative, as Manovich argues. Yet each of these frameworks depends upon understanding the database primarily as

what Kenneth Price calls a "suggestive metaphor" for prolific and fluid information, rather than in a "strict" sense as a digital data structure (pars. 18-19). Metaphorizing databases encourages users not to see them as tools, but to instead adopt a worldview influenced by the internal logic of databases, a perspective in which all phenomena can be structured and quantified as records or objects. It also encourages associating a certain set of qualities and properties with the database form, such as large volume, reconfigurability, and comprehensiveness. Folsom makes the connection between database and volume particularly explicit, variously describing the database as a "huge," "endless," and "massive" form, a genre which "suggests … endless ordering and reordering, and wholeness" and strives for "completeness" by "grow[ing] exponentially" (1573, 75, 77). Such metaphors present databases as forms inseparable from their potentially large volume, falsely implying that the technology is equivalent to the types of information it stores.

The unspoken assumption of the narrative and database debate – that data and databases are necessarily high volume and comprehensive, and that these qualities are what establish the dominance of the database form – is flawed. Data do not need to be big to be useful – indeed, small humanities datasets suggest that data is valuable at any size.[3] Unlike narratives, databases are generally scalable and amenable to expansion, but the information that they house may or may not be.[4] Drawing on Anna Tsing's theory of nonscalability, Katie Rawson and Trevor Muñoz describe data as containing both scalable and nonscalable elements. Scalable elements can be explicitly articulated, captured through standard processes, and are generic components of systems, while nonscalable phenomena "are enmeshed in multiple relationships, outside or in tension with the nesting frame" of the hierarchical data model, and resist easy categorization (Rawson and Muñoz). Scaling data becomes difficult when trying to model extremely

heterogenous entities with many nonscalable elements, or when the data collection workflow includes many unique, time consuming, non-repetitive tasks.

Information may also fail to scale because many humanities datasets have gaps, unknown components that cannot be filled in because of the vagaries of the historical record or the ineffable qualities of the subjects they address. These datasets are usually of manageable sizes (hundreds or thousands rather than millions or billions of records) because they are often created manually or focus on a domain of study that is also tractable for traditional humanities methodologies. Theorizing data and databases solely through a high volume or big data mindset denigrates smaller datasets as partial, incomplete, and insufficiently large to lead to substantial conclusions. To reconfigure our understanding of databases and the role of data within humanities research, we need to jettison the assumptions that dense, high volume data is inherently good and that small datasets with gaps and questions are undesirable. Instead, I would like to introduce an alternative understanding of data as boutique rather than big, a formulation more fitting for typical humanities research. For evidence of this concept, I will examine and critique *Laboring-Class Poets Online*, a database-driven website which addresses the more than 2,000 British laboring-class poets who published between 1700 and 1900 and their writing.[5]

Big and Boutique Data

Big data and databases are attracting attention not just from within new media studies, but also from diverse departments across the academy as well as the private and public sectors. Corporations and governments are pouring money into big data aggregation and analysis techniques to better track and target customers and citizens and to increase the efficiency of all

aspects of their operation. The rise of big data parallels the acceptance of digital methodologies within the humanities and the relabeling of humanities computing as the more mainstream digital humanities, an interdisciplinary field which maintains a conflicted relationship with big data. The allure and seeming omnipresence of big data shape how we think of data in general, prompting us to see information through Folsom's expansive, pervasive metaphors. When data and databases are consistently framed as big, we naturalize those structures and connotations and begin to accept data rather than questioning it like any other constructed, situated text. Uncritically adopted big data mindsets are problematic, but a boutique approach to data provides a potential solution because it more closely reflects the values of humanities research.[6]

Big data challenge traditional modes of information storage, retrieval, and analysis because of their increased size and complexity. The term also implies the use of new data analytics methods that turn large amounts of data into actionable information and provide insight into complex datasets. Big data is often described as having three key dimensions – volume, velocity, and variety ("Big Data"). Big data is large in volume because it often engages in full-scale tracking of phenomena rather than random sampling, resulting in petabytes of data that demand new massively distributed computing techniques compared to traditional processing methods. Big data is often high velocity, consisting of billions or trillions of datapoints generated constantly and automatically from sensors such as website clicks, video feeds, geospatial pings, or device interactions; it must be processed rapidly (often in real time) because it quickly becomes outdated and loses value. Finally, big data is of heterogeneous variety, drawing novel conclusions by connecting many different sources and file types, including unstructured text, images, and video, rather than relying on a single dataset.

The size and complexity of big data can create the illusion that such datasets can perfectly model an imperfect and unpredictable world, gaining authority simply by increasing in volume. The computational authority of big data is persuasive and seemingly presents an entirely new way of knowing reality – an epistemology of the database. This fascination with the potential power of big data is currently pervasive in the digital humanities. Big data encourages certain research techniques because larger datasets increase the potential return of complex, computationally intensive methodologies such as latent semantic analysis, machine learning, topic modeling, and data visualization. Franco Moretti uses what he calls distant reading to analyze large corpora of novels (*Maps*, *Graphs*, *Trees*; *Distant Reading*), while Matthew Jockers makes similar moves with the methodology he terms macroanalysis (Jockers). Stephen Ramsay wonders how to best use large databases of millions of books ("Hermeneutics"), pushing for an algorithmic criticism that employs computational text analysis to generate rather than validate theories (*Reading Machines*). Researchers at the Software Studies Initiative and Cultural Analytics Lab extend these textual investigations into new media by using "cultural analytics," a combination of data science, machine learning, and large-scale visualization, to study the properties of millions of Flickr images (Ushizima et al.). Big data digital humanities projects attract disproportionate amounts of funding compared to smaller endeavors, as evidenced by grant programs such as the T-AP Digging into Data Challenge, which is currently sponsored by eighteen international organizations.[7] Such projects aim to use databases to tell not only *a* story, but "*the* story," a single account that, through its enormity, encompasses and surpasses all individual narratives (Hayles 183).

While such research can provide tentative answers to large questions, big data projects within the humanities have also generated resistance. Some critiques of big data are methodological. Katie Trumpener, for instance, criticizes Moretti for relying too heavily on the invisible hand of the literary marketplace to explain generic shifts as reflected in datasets of novel titles and thus ignoring the complex historical interactions of writers, publishers, and audiences in favor of a "monocausal" force that better fits his distant reading approach (168). Other objections are infrastructural, rightly arguing that the extensive personnel, technical skills, and funding required for big data analysis exclude many potential audiences and institutions. Some scholars protest because of a fundamental mistrust of big data, or even resist digital scholarship and methodologies altogether. Of these broad critiques, the most productive explore the so-called "dark side" of digital humanities as a field by addressing issues of academic labor, funding, and collaboration (Chun et al.). Less generous assessments draw questionable links between the growth of the digital humanities and the emergence of the neoliberal academy, raising accusations of complicity or even causation (Allington et al.).

Criticizing an entire methodological approach and field of study in this way is unproductive. Instead of lambasting all digital scholarship, we should be pushing for more ethical approaches to digital research, especially how big data are collected, represented, and used. Just like user interfaces and algorithms, data are never neutral and do not speak for themselves, but are rather products of cultural, political, and economic situations that must be treated as such. Berendt et al. recognize this when they argue that "data analytics regulation is also a political, legal, and societal task" that cannot be reduced solely to the computational aspects of knowledge production (226). Alan Liu's well-reasoned 2012 plea for the digital

humanities to embrace cultural criticism ("Where Is?") has begun to be answered by efforts to decolonize the digital humanities (Risam) and by movements such as #transformDH ("About #transformDH"). However, many of the tools, datasets, and databases which drive digital scholarship have not yet fully implemented these insights which are central to humanistic inquiry. I do not find big data inherently problematic, but unquestioningly adhering to an uncritical and inflexible big data mindset is troubling, especially when such approaches fail to confront the ethical implications of big data collection, production, and analysis. Frédéric Kaplan recognizes this discord when he argues that the new "structuring tension" in digital humanities is between "Big Data Digital Humanities," which "focuses on large or dense cultural datasets which call for new processing and interpretation methods," and "Small Data Digital Humanities," which "do not use massive data processing methods and explore other interdisciplinary dimensions linking computer science and humanities research" (1-2).

Kaplan's key observation – that datasets of different scales are not equivalent – is rarely explicitly articulated. While big data continues to grow within the humanities, even "big" digital humanities projects remain significantly smaller than those in the sciences or private industry, a trend which has not changed since Lev Manovich recognized it in 2012 ("Trending" 2). If there is a typical size of dataset used within the humanities, it is not big data. Most humanities researchers oppose framing their objects of study (texts, films, paintings, historical documents, cultural artifacts, or even people) as data at all, preferring to treat them as unique, contextually situated, analog entities that resist digital reformation because of their inherently qualitative forms. Miriam Posner notes that for many humanities scholars, the term data has uncomfortable connotations. Describing an artifact as data implies "that it exists in discrete, fungible units; that

it is computationally tractable; that its meaningful qualities can be enumerated in a finite list; that someone else performing the same operations on the same data will come up with the same results" (Posner). Posner argues that calling a family album a dataset is not erroneous, but that such an approach "fundamentally doesn't understand why you value this artifact" and consider it worthy of study (Posner). Researchers are often less interested in the quantifiable, countable, or objective aspects of humanities data than in the intangible qualities that hide in database gaps or reside outside of the database entirely. If humanities scholars overcome their mistrust and do digitally represent these objects and store those digital representations in a database, they are still interested in artefactual minutiae and item-level complexity alongside and as representative of aggregate trends. Within big datasets, entities are subsumed within the whole and are rarely examined individually, whereas humanities scholars often recursively shift between distinct records and the collective dataset and between close and distant readings of these items.[8]

Understanding humanities data as boutique rather than big respects the continued importance of the individual text and the interpretive, hermeneutic methodologies at the core of humanistic inquiry. Ball et al. describe boutique data as "small," because of its limited size and "local context," and "dark," because it is often "inaccessible" due to never being published or publicly released (5).[9] Projects relying on boutique data "require the *extraction* of data from ongoing research … rather than its *preservation* and *containment*," often producing "data collaboratories" that join disparate datasets rather than repositories where data is merely deposited (Ball et al. 6, 9). A boutique approach to data differs greatly from a big data mindset by refusing to rank or value datasets according to their size, comprehensiveness, or ultimate

quantifiability. Small, partial, and qualitative datasets inhabit a valuable place within the humanities research process when viewed from this perspective.

A boutique approach to humanities data treats datasets as created entities with inherent biases and inbuilt assumptions about the value of different types of information and their audience. Databases serve as both a text (in the aggregate) and a collection of texts, each of which can be read and interpreted. Trevor Owens argues along these lines when he writes that we should understand humanities data as "constructed artifacts […] interpretable texts [… and] processable information […] holding evidentiary value" (Owens). While big data is frequently high velocity and automatically generated, a boutique approach to data respects the often-hidden labor of data creation, extraction, and processing. The construction of boutique data is rarely fully automated, and the traces of human contact with the database are embedded at every stage of production. Rather than concealing the createdness of data, a boutique approach to data acknowledges its sources and collection and cleaning methods.

A boutique approach to data does not perceive the relatively small volume of boutique datasets as a weakness, but as an opportunity for deep analysis of individual entities alongside an interpretation of the entire dataset. A boutique approach to data reflects complex understandings of objects, texts, people, relationships, places, and times, and therefore the structure of boutique databases emerges out of this information (as well as each project's goals) rather than being imposed upon the data. Boutique datasets respect the cultural and historical contexts of data rather than unilaterally leveling these dimensions into more computationally tractable forms. Within a boutique dataset, people and texts retain their individuality rather than being flattened into objective data points. This is possible because boutique databases are often bespoke and are

designed to meet the requirements of specific populations and projects rather than a wide variety of generic users and use cases. Boutique databases can thus be described as situated software, which Clay Shirky defines as programs "designed in and for a particular situation and context." This design strategy relies on implicit domain knowledge and existing "social infrastructure" to overcome development problems, such as scaling. Situated software recognizes that "scaling problems aren't inherently fatal" (Shirky). Boutique data projects leverage their small size to create experimental databases and interfaces that may not scale, but fit the needs of local communities and purposes. Furthermore, members of these communities are often the designers of boutique databases.

Rather than being constructed by database administrators or software engineers, boutique databases are often built by end users, who rely on their expert content knowledge to outweigh any technical deficiencies. End users "face different motivations and work constraints than professional programmers" and rarely focus on formal requirements or specifications, reusability, testing, or quality control (Burnett 17). The development of boutique databases is related to the practice of end-user software engineering, which is "characterized by its unplanned, implicit, opportunistic nature, due primarily to the priorities and intents of the programmer" (Ko et al. 8). However, where Ko et al. argue that end-user programmers design primarily for themselves or small groups, most boutique databases are intended to be shared with a wide audience (4). The two approaches often overlap in their methodologies and typical activities, but differ slightly in their goals.

Finally, a boutique approach to data considers the inevitable incompleteness of databases to be a feature rather than a bug. From this perspective, database gaps function not as empty

voids but as evocative absences worth investigation and explanation. A boutique approach to data invites users to see both the database's vacuums and plenums as events – dynamic possibilities filtered through and altered by history, texts, and human labor, rather than static, unquestionable facts. To more fully explore the potential of boutique data for humanities research, the rest of this article turns to *Laboring-Class Poets Online* (*LCPO*) to present an alternative to big data and critique the limitations of a big data mindset.

Laboring-Class Poets Online is a database-driven website that aggregates biographical and bibliographical information about the more than 2,000 laboring-class poets who published between 1700 and 1900 and the texts they produced. It functions as a clearinghouse for data about poets from the lower classes who lived in the British Isles or in British colonies, and thereby helps demonstrate the importance of laboring-class writing to social and literary history. While *LCPO* is now a digital resource, it draws on a long history of textual and archival scholarship that began over thirty years ago. John Goodridge began collecting information about non-canonical poets in the late 1980s as a response to a *Dictionary of National Biography* request for missing names. Goodridge "was increasingly frustrated at the refusal of those who taught and researched canonical writers to even acknowledge that the poets they knew nothing about and hadn't read might affect their perceptions of literary history," and submitted a relatively small collection of 500 British poets to the DNB to bring attention to this underappreciated tradition (personal communication).

This list was later narrowed to focus on solely laboring-class figures, and in 2001 Goodridge released the first version of his "superlist" as a tie-in to the six-volume *Eighteenth-* and *Nineteenth-Century English Labouring-Class Poets* resource. The list featured 659 "named

poets" with additional "anonymous, pseudonymous, and group productions," and compiled "a concise paragraph on each individual including vital dates and a short description, key publications in short form, and secondary sources" (*Database* 2001). Sixteen years later, the superlist includes over 2,000 poet entries, many of which have been significantly expanded from their original terse forms. Goodridge continues to update this document monthly and incorporate new information supplied by the project's editorial board and other collaborators. While the superlist has been available for download as a Microsoft Word document since 2011, *LCPO* adds significant functionality by transforming this static list into a standalone website driven by a relational database. Compared to an unstructured text-centric dataset, *LCPO* natively includes multimedia assets such as images and audio; improves collaboration between project researchers; enhances the project's extensibility, maintainability, and visibility; provides direct references to internal and external resources and secondary sources; empowers users to perform complex queries across numerous controlled fields; and invites users to interact with content through numerous data visualizations including geographic maps, timelines, charts, and network graphs.

The information collected for *Laboring Class Poets Online* presents a classic case of boutique humanities data: a collaboratively and manually created and curated small dataset of several thousand entities extracted during ongoing research.[10] Accordingly, *LCPO* attempts to adhere to not only the characteristics intrinsic to boutique data (small volume, manual processes, and collaborative production) but also to the ethos of a boutique data perspective by interacting with time, places, and events as meaningful, contextually situated, interpreted entities. While human scholars often use context to interpret data points in historical documents, databases and computational methods lack this inherent capability. Uncertainty is embedded in historical data,

but databases often strip away this ambiguity to perform the computational functions that make their use worthwhile. By taking a boutique approach to historical and literary data, *LCPO* retains much of this ambiguity and offers insight into how humanities researchers can accommodate a complex understanding of space and time as continuously unfolding events.

Geography

One way in which historical, boutique humanities datasets conflict with computer-generated big data is through differing abstract representations of physical space. Geographical information systems (GIS) can represent any earth-based point as a set of *x*, *y*, *z* coordinates corresponding to latitude, longitude, and elevation. However, modern GIS databases and services are ill-equipped to address the spatial ambiguities of thickly layered historical humanities data. Modules such as Location, Addressfield, and Geofield allow Drupal developers to store locations as strings and turn them into geographic points or polygons through services such the Bing or Google Maps APIs. This process, called geocoding, attempts to match a location name with a pair or triplet of coordinates. These services are generally designed for GIS developers working with modern locations, and often fail or return inaccurate information when querying incomplete or historical places. While the Getty Thesaurus of Geographic Names provides some historical results, and GeoNames stores toponyms (alternate placenames),[11] I have been unable to find a comprehensive spatial service which can reliably geocode historical British locations.[12]

The difficulty of mapping historical places was a major problem when developing *LCPO*, as about 25% of the over 1,300 identified locations failed to geocode correctly through the major GIS APIs (Google, Bing, GeoNames). Some failures occurred because of variant spellings. The

Cosletts, a Welsh family of poets including William Coslett (Gwilyn Elian) and Coslett Coslett (Carnelian), lived at Nantyceisiaid farm near Machen in the historic county of Monmouthshire. Given the modern spelling of the Nant y Ceisiad stream, Google could geocode the location, but failed to recognize the slightly different historical spelling used by the *Dictionary of Welsh Biography* and its historical sources. Other locations, such as Llanbedr-ar-Fynydd (St. Peter's Church) are now in ruins; the chapel last held services in 1812, and only appears on historical survey maps. Sometimes less popular locations are incorrectly geocoded as their more well-known counterparts. Nether Bogside, the birthplace of Janet Little, the Scotch Milkmaid, geocodes as a farmhouse in Elgin, but in this case it is actually a region south of Ecclefechan. Houses, estates, and farms that carry names rather than road numbers also challenge modern GIS. Some of these locations, like Main of Nairn, the birthplace of pattern-drawer James Taylor, I could never distill into coordinates, and had to fall back to the next most specific location instead – in this case, the nearby town of Stanley. Because of such issues, I decided against saddling *LCPO* with an API service or module built for fully quantifiable information that would fail on ambiguous data. Instead, I geocoded what placenames I could automatically and added those locations and coordinates to *LCPO* as a taxonomy. I then manually tracked the remaining locations with a variety of resources, including British placename gazetteers, geographical message boards, tax rolls, and historical maps. Generic database services designed for contemporary big data and modern locations proved inadequate for boutique, historic humanities data, while traditional humanities research methodologies succeeded, albeit with a brute force time penalty.

As digital scholarship in the humanities matures, researchers have begun to develop and adapt tools for humanities data, though these remain somewhat uncommon. There are some digital services designed for historical locations, such as Pleiades, a gazetteer of ancient places which primarily covers Greek and Roman sites, and the Pleiades+ toponym extension which associates Pleiades entity URIs with GeoNames records. Because Pleiades was designed by classicists, its underlying data structure reflects the complexity of humanities data. Rather than representing geographic entities through a single object, Pleiades uses three related content types: places, which are human-constructed "conceptual entities" which do not have "spatial or temporal attributes of their own" because they may or may not be associated with a definitive, geocodable location; locations, which are associated with places and store geocoded coordinates and date ranges; and names, which provide additional toponyms for places at a specific period in time ("Pleiades Data Structure"). This tripartite spatial structure respects the intricacy and possible unknowability of places within boutique humanities data. Arcadia, for example, is a region of the Peloponnese peninsula, but within the European literary tradition (including British pastoral and utopic poetry) "Arcady" usually alludes to an unspoiled, lost place rather than the physical Greek location. The Pleiades geographic data structure stores such recondite, mythical places as place content types without locations rather than discarding them entirely because of their spatial unquantifiability. Such places are embedded within the literary imagination even if they are not physical entities, and a boutique data approach recognizes their importance. *LCPO* does not yet use the Pleiades data structure but will eventually incorporate some of these concepts to better represent the instability and historicity of space. The situatedness of space is already reflected in *LCPO* through the local places that permeate the structure of laboring-class

writing and literary communities. *LCPO* engages with complex space by capturing dialect usage and location-based literary publics.

While dialects encompass and are produced by numerous categorical factors, such as ethnicity, social class, and education, geography plays a dominant role in shaping linguistic patterns. Compared to canonical writers, far more laboring-class poets wrote dialect verse that reflected the speech patterns of their local peers rather than defaulting to Standard English. While many writers used dialect in dialogue or for comic effect, working-class poets deployed their native vernacular throughout their writing at the narrative level. Scottish writers were particularly likely to craft dialect verse, especially in the wake of Robert Burns. Scots declined as a literary language after the Acts of Union in 1707, when political power shifted to England. With this migration, Standard English in a Received Pronunciation gained prestige, while regional variations from this new norm became associated with provincialism. Many writers attempted to eliminate Scotticisms from their writing, and the use of Scots and other dialects became a marker of working-class status. This changed somewhat in the late eighteenth-century. Drawing on the dialect poetry of Allan Ramsay and Robert Fergusson, Robert Burns helped revive literary Scots, creating a market for Scots language writing that English dialect writers never acquired. The publication of his *Poems, Chiefly in the Scottish Dialect* (Kilmarnock: 1786) inspired dozens of imitations, many from working-class writers who consciously fashioned their writing after Burns' and advertised their dialect usage in the titles of their volumes. Other laboring-class poets wrote in Geordie (Tyneside), Cumbrian, Dorset, Buchan, and in dialects from Devonshire, Nidderdale, Northamptonshire, Lancashire, Leicester, Yorkshire, and Shetland. Many Welsh laboring-class poets wrote and published in Welsh, while smaller

numbers of poets wrote in both Scottish and Irish Gaelic, Manx, Jèrriais (Jersey), and Guernesais (Guernsey). *LCPO* captures dialect usage in both individual publications and by writers more generally through a continuously expanding taxonomy of dialects, thereby reflecting how location is threaded through writing.

Beyond their dialect usage, many writers identified primarily as members of a local geographic community. Large groups of working-class writers existed in Glasgow, Dundee, Paisley, Tyneside, Manchester (the "Sun Inn" group), Liverpool, Blackburn, Bradford, Sheffield, Nottingham, Leicester, Bristol, and Birmingham. The writing from each of these localities differs in dialect usage, but also in the topics that poets addressed, their general political allegiances, and the forms their writing assumed. In Sheffield, balladeers "sold single copies of their popular songs," while in Scotland there were more newspapers that solicited laboring-class poems, such as the Dundee-based *People's Journal* (*Database* 2017). Poets knew their local contemporaries and often addressed poems and letters to each other, forming literary networks that influenced their writing. Though contemporary literary criticism typically contextualizes this broad heterogeneous tradition of writing as a function of socio-economic class determined primarily by occupation, locality was at least as strong a component of writing identity for poets as their professions.[13] While many working-class writers identified themselves by profession, from well-known writers such as Stephen Duck ("The Thresher Poet") to minor poets like Patrick MacGill ("The Navvy Poet"), more than twice as many of the figures in *LCPO* who used such pseudonyms identified by location compared to occupation.[14]

When taking a boutique approach to data, places cannot be reduced to mere coordinates because they retain meaning beyond their physical locations and shape how humans act, write,

and collaborate. Places affect characteristics such as dialect and writing communities, but the opposite applies as well. Places are not static entities; they change over time as people and natural processes shape them or let them decay. This presents another problem for many geographic information systems, which can rarely show or capture the impact of chronology on places. Over time, boundaries of administrative units can be revised, locations can change names or meaning, and spaces may even cease to exist (Merry). The administrative geography of Britain has shifted numerous times as the boundaries and names of parishes, counties, and even countries have been altered to meet political demands. For instance, the area in northeast England consisting of Newcastle upon Tyne, Gateshead, and Durham was home to a major grouping of English laboring-class writers. Goodridge writes that these "Tyneside poets, documented in *Allan's Tyneside Song* and *Rhymes of the Northern Bards,* include printers and other city artisans, and a very strong contingent of mineworkers" (*Database* 2017). The tradition of songwriting started by Tyneside figures such as Tommy Armstrong, Ned Corvan, and Joseph Skipsey, "The Pitman Poet," continues today: "The Festival of Mining Literature and Poetry in North-East England" was held in June 2014 to celebrate not just written work, but also the enduring oral culture of the area (Whetstone).

Yet if those long-dead Tyneside poets had been able to return for the festivities, they would likely be confused by the intervening geographic upheaval (Figure 1). Newcastle upon Tyne used to be part of the historic county Northumberland, while Gateshead, just a few miles further south, was part of county Durham. Today, Newcastle upon Tyne and Gateshead have

largely merged to form the new ceremonial county Tyne and Wear, which carves out a small, dense space between Northumberland and Durham counties. Through such changes, place assumes a quotidian, governmental function rather acting as a unifying force. But altering spatial boundaries does not just modify political constituencies, governments, and trade systems; it profoundly influences communities of writers and citizens whose identities are bound to the land. This is why properly representing the messiness of spatial data matters. While a big data approach to spatial information might correctly geocode where people physically lived, a boutique data approach goes beyond mapping to engage geography as an organizing force that supersedes objective coordinates and is tied to cultural heritage and belonging.

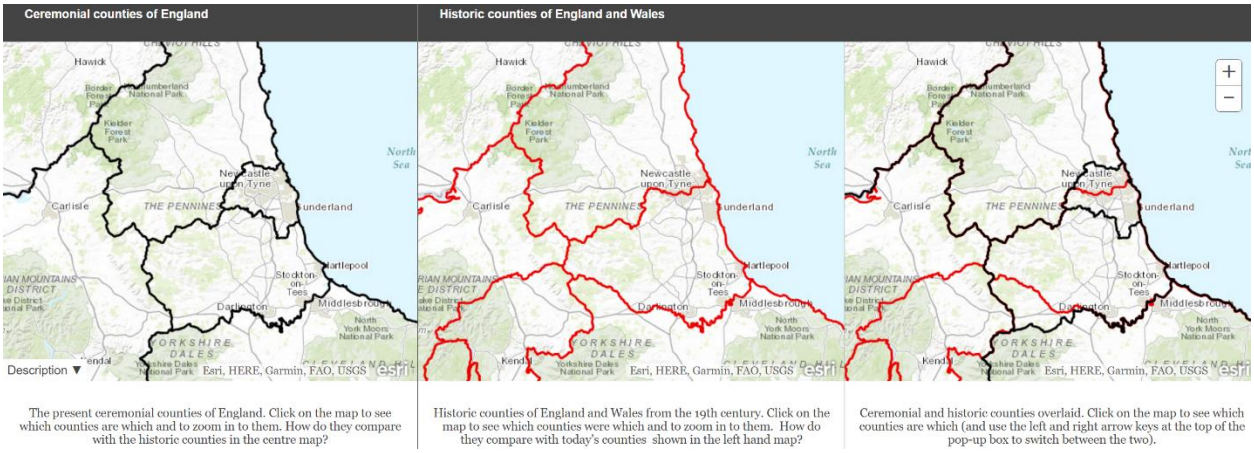

Figure 1: Change in English ceremonial and historic counties over time (Gillet)

The Association of British Counties is an organization which attempts to instantiate such an approach by influencing public policy. The ABC is dedicated to establishing a "fixed popular geography, one divorced from the ever-changing names and areas of local government but, instead, one rooted in history, public understanding and commonly held notions of community and identity" ("About the ABC"). This organization wishes to abolish ceremonial counties in favor of restoring the historic counties, which they argue are "bedrocks of the history, culture,

and geography of Britain … places where people live and 'come from,' where they 'belong' " ("We Promote"). Using Pleiades' data structure terms, because communities and cultures draw on the organizing structure of the historical county, these entities feel like places for the ABC, while the newer ceremonial counties function merely as arbitrarily demarcated locations. Geographic systems based on locations rather than places elide these complex, dynamic spatial representations, ignoring identity, heritage, and belonging in favor of a more easily chartable, ahistorical, and altogether flatter geography.

Chronology

Boutique data also confound the structuring impulses of modern databases through the difficulties presented by historic temporal information within the humanities. Humans understand time (or temporally situated events) and time duration subjectively, and as I show below, our historical systems for measuring time often reflect that arbitrariness. Computers, however, measure and represent time as an inexorably progressing, unbroken continuum of milliseconds and enforce this ontological perception of time through their structure. Big databases standardize and atomize time to facilitate computation, but this regularity comes at the cost of rebuffing non-linear understandings of chronology and reducing time to a stopwatch. A boutique approach to temporal humanities data attempts to represent this information on its own terms and thereby respect complex time.

Databases cannot represent fluid, individual perceptions of time; they require standard structures. When computers or sensors generate chronological information, these dates and times validate because they follow a strict internal logic. Time is measured as discrete units elapsed

since a reference date or epoch, which serves as the origin of time within that chronologic system. The most common method for databases and many programming languages to store and reference dates is as a Unix timestamp, or an integer that measures the total number of seconds elapsed since the instant chosen as the Unix Epoch, 00:00:00 Coordinated Universal Time, 1 January 1970. This approach has its own issues due to impending hardware constraints, but generally works well for storing dates generated from activities in the range 1970 to 2038.[15] Extending this method backwards before 1970 quickly becomes messy not only because it requires negative numbers, but also because historical chronological information resists being parsed down to a definite integer.

Insisting on an unambiguous representation of time presents a challenge for humanities data because is often generated from such historical sources. Wai Chee Dimock writes that time is not "identical to the properties of number … standardization is not everywhere the rule. In many parts of the non-Western world, a very different ontology of time prevails" (*Continents* 2). Even Western Europe used multiple calendars in the recent past. When the Gregorian calendar was introduced by Pope Gregory XIII in 1582 to correct the Julian calendar's leap year problem and halt the seasonal drift of Easter, it skipped 10 days to align the new calendar with the Earth's seasonal cycle. Catholic countries quickly changed over, but Britain did not switch until 1752, creating a gap for nearly two hundred years between British dates recorded in the Julian calendar and continental European dates recorded in the Gregorian calendar; Greece and Turkey didn't adopt the Gregorian calendar until after World War I. This creates oddities such as Shakespeare and Cervantes sharing the death date April 23, 1616, because Spain was using the Gregorian calendar while Britain remained on the Julian calendar for another century; Cervantes actually

died ten days prior to Shakespeare (Armstrong). Without a clear consensus on how to accurately record historical dates digitally, each database management system (DBMS) represents them differently, though they all struggle with historical data due to ambiguity and a lack of standardization between database vendors.

This digital consternation is compounded when partial, approximated, mixed-format, or ancient dates are considered. When working with *LCPO* and historical documents, I have occasionally encountered partial dates, such as "May 1808" in a letter from Robert Tannahill to James Clark (Tannahill). Other documents may have even more ambiguous or estimated information that covers a wide range of time, such as "spring 1734" or "late sixteenth century." MySQL (and by extension Drupal, the content management system used by *LCPO*) cannot natively represent such "fuzzy" or approximated dates. Databases also fail to adequately handle mixed-format date fields with varying granularities. A field using the SQL DATE datatype requires information to be formatted as "YYYY-MM-DD" and cannot simultaneously handle dates formatted as "1808," "May 1808," "5 May 1808," and "5:00 am, 5 May 1808." The field will arbitrarily add day and month data to the first two examples and strip the time data from the third, all in the pursuit of regularity. Some developers have attempted to alter databases to better interface with historic time, such as the creators of Drupal's Partial Date module,[16] but like geographic systems designed for humanities concepts of place, such approaches are rare. While all *LCPO* records are from the early modern or modern eras, digital archaeological projects may grapple with what Dimock terms "deep time," a "thickened" and "irregular" chronology that extends back beyond any human attempt to regulate time and encompasses celestial events that span centuries or are tens of thousands of years old (*Continents* 4). Deep time stands in

opposition to an ontological understanding of time as discrete, "quantifiable" and "unidirectional" (*Continents* 123). Deep time events may clearly hold intense narrative meaning, but databases fail to fully encapsulate that value by truncating them and locking them into a single figure devoid of context.

Laboring-Class Poets Online brushes up against irregular chronologies and the messiness of time in several situations, such as recording vital dates. The level of information coverage within *LCPO* is extremely variable. For well-known or canonical figures such as Robert Burns, Robert Bloomfield, John Clare, Mary Collier, Stephen Duck, and Ann Yearsley, *LCPO* has significant amounts of information, including verified birth and death dates. On the other end of the spectrum there are poets such as Elizbeth M. Sinclair, a millworker from Ettrick Braes in New Lanark, and a figure for whom *LCPO* researchers have been unable to recover any chronological information at all. Most poets fall somewhere between these two extremes and *LCPO* implements numerous date fields besides birth and death to accurately represent these levels of temporal ambiguity.[17] Many birth and death dates are approximated, so there is an additional Boolean field to indicate whether a date is a "circa" date. Sometimes poets were entered in a parish baptismal registry, but we are unable to recover an actual birth date, so there is another field for baptism dates. We also use known publications to help approximate poet chronologies when no verifiable vital dates exist. While many poets only published single works, giving us a single data point, others published numerous works over longer timespans. For instance, Thomas MacQueen, a journeyman mason from Bakip, published at least three poetry collections: *Poems and Songs* (Glasgow: 1826), *My Gloaming Amusements, a variety of poems* (Beith: 1831), and *The Exile, a Poem in seven books* (Glasgow: 1836). This gives a "flourished

range" from 1826-1836. Another secondary source mentions that a "Thomas M'Queen" emigrated to Canada and published three volumes between 1836 and 1850, dying in 1861; while I have not been able to verify that these two poets are the same person, even a decade range of active publishing activity is far more valuable than no dates at all.

The temporal information included in *LCPO* is clearly boutique data. *LCPO* attempts to understand these messy, uncertain, and variable chronologies as dynamic, deep time events rather than static entities. These gaps in information density and certainty should not be construed solely as a lack of information. While there are missing pieces in *LCPO* which will be added in the future as additional research is completed and new collaborators join the project, there are also holes which will never be filled. These silent nulls do not corrupt the database, but accurately represent the ruthlessness of the historical record and the passage of time. Many laboring-class poets left few lasting traces on the world – a name, a few chapbooks or pamphlets, perhaps a few dates, and maybe a birthplace or an occupation, often inferred from a poem title rather than an explicit biographical source. Their publications have not survived for various reasons: small print runs, cheap paper and ink which deteriorated rapidly, and stereotypes of laboring-class writing as not valuable or literary and thus not worthy of inclusion in libraries or preservation in archives all contributed to their disappearance. Publishing venues also play a role. Poets often chose to publish in periodicals and magazines such as the Chartist outlet *The Northern Star* or the *People's Journal* because they reached a wide audience and provided a lower-stakes forum than full length collections. However, many smaller periodicals have huge gaps in their archives or no surviving copies at all, leaving only the thin memory of a publication derived from a reference or letter. While a big data mindset might understand these lacunae as

information failures, annoying breakdowns hindering the quest to perfectly model the world, a boutique data worldview seizes on those gaps as instances where class, the literary marketplace, and time intersect, and tries to make those fractures in the historical record visible.

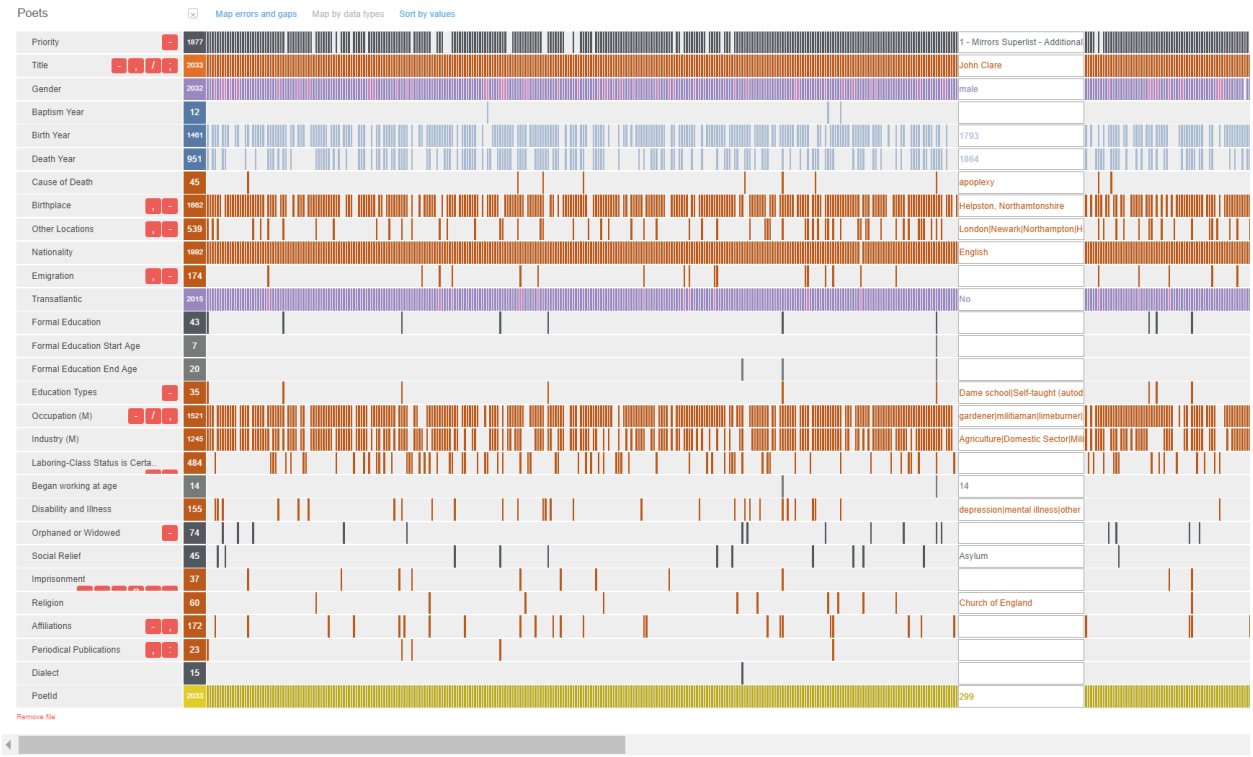

Figure 2: All poet records and major fields exported from *LCPO* and visualized through Breve

Breve, a web application designed for qualitative analysis of "incomplete and messy … historical data," can help accentuate these breaks by visualizing data density ("Breve"). Breve is typically used as a meta-scholarly tool by database designers or data curators during the data entry and aggregation processes. Because boutique data is generated manually over time, boutique databases often change dramatically between iterations. Breve helps database administrators see these changes by visualizing an entire dataset in a single snapshot.[18] Figure 2 shows the first 350 *LCPO* poet records as visualized by Breve. Like an inverted spreadsheet, each column represents a single poet and each row a single field; filled fields are colored, with

potentially mismatched datatypes highlighted in a different color, while grey cells represent null fields in the database. Breve allows users to readily distinguish the poet records with the highest data density. John Clare is highlighted in this figure, but to his left are Thomas Chatterton and Robert Burns; the numerous non-null education fields, which few poets currently have filled, make them easily identifiable. As more information is added to *LCPO*, these spaces will slowly fill up for other poets in the database, but eliminating the gaps entirely is neither possible nor desirable. Instead, researchers should be using tools such as Breve not only during backend database development but as visualizations that can help to draw end users to boutique datasets and to explore both their gaps and abundance.

Events

Events emerge through the overlap of these two primal forces, geography and chronology. Events occur when human or nonhuman actions are stitched into this folding of time and space; we constantly experience events in our everyday lives, but we only know of events outside of our immediate experience because they were registered in the historical record through evidence such as texts, artifacts, buildings, or memories. Humans often perceive events as things that occurred in a fragment of linear time (past, present, or even future) and in a specific, physical place, and are circumscribed by these boundaries. Jacques Derrida introduces the concept of *arrivance* which resists such a simple understanding of events:

> … an event that remains an event is an arrival, an absolute arrival [*arrivance*]: it surprises and resists analysis after the fact. At the birth of a child, the primal figure of the absolute *arrivant*, you can analyze the causalities, the genealogical, genetic, or symbolic premises, and all the wedding preparations you like. Supposing this analysis could ever be exhausted, you will never get rid of the element of chance [*l' aléa*], this place of the taking-place … (20)

Events come into being because of human actions, but if we treat them as static things with inert properties, we ignore the random chance that produced them. Looking back at historical events, we are no longer are surprised by their shape because "analysis always tends to diminish surprise" (Derrida 20). Their order and causality appear predetermined. If we instead see events through the lens of *arrivance*, they instead withdraw from any definitive interpretation or final, complete human knowledge. As continuously developing assemblages of time, place, and action, such events are irreducible and cannot be or fully explained or exhausted. Given the difficulty of properly representing geographic and chronologic humanities information from a big data perspective, creating eventful databases provides a possible solution for boutique data.

Currently, *LCPO* does not fully engage with *arrivance*-based eventfulness. The Poet content type has numerous fields that capture biographical attributes and thereby help describe the historical figure which the record represents. These properties include the aforementioned vital dates; a series of different fields to capture names, including married and maiden names, bardic names (for Welsh poets), and pseudonyms; cause of death; birthplace, other locations, and emigration history; nationality; education types, levels, and starting and ending ages; occupations; whether their status as a laboring-class writer is certain or not; disabilities and illnesses; whether the poet was orphaned or widowed; interactions with the social relief and criminal justice systems; religion; affiliations with different literary, political, or geographic groups; and dialect usage. Poet entities are closely linked to the publication[19] content type, which captures texts poets authored, edited, or translated, and the relationship content type, which connects two poets who interacted and explains the strength and nature of their association. Together, these direct and related attributes form poet records as represented in *LCPO*.

Most of this biographical information is oddly frozen, unstitched from time and space. While each publication has publication years and locations for each print run (where available), each poet record only has one somewhat eventful datapoint, which represents the event of his or her birth, and even this event is split over multiple fields. Taken together, these poet records collect useful but static biographic data, decontextualized from chronology and geography. Rather than representing the actions and being of each poet as a series of overlapping, conflicting, progressing and regressing events strung across space and time, *LCPO* crystallizes them as dormant figures that simply *were* rather than people that *lived*. *LCPO* poet records more closely resemble collections of fixed properties than sites of *arrivance*. By speculatively re-imagining *LCPO* poet biographical as eventful rather than static, we can begin to recognize their continual eventfulness in ways the database currently does not accommodate.

An "eventful" *LCPO* would have far fewer fields for each content type – only intrinsic features which do not tend to change over time. Most fields would be disassociated from content types and re-instantiated on a new Event content type. Events would point to Poets or Non-Poet Figures which were involved in the Event, and would also be able to store a date or date range, a location or array of locations, a description, and an event type and subtype. The event type field would store the information previously captured directly by content fields such as industry, religion, or immigration, while the event subtype field, if applicable, would store the field value. This would allow *LCPO* to model John Clare's complex early work history as a series of occupation events, with each job having a start and end date; between 1807 and 1813 Clare apprenticed as a cobbler, a stonemason, and a garden-boy, and worked as a ploughboy, inn boy, general laborer, gardener, lime-burner, lawyer's clerk, and soldier. A similar approach could be

used to show how poets changed political or literary affiliations over time, married, had children, or moved about Britain and the world.

Decoupling biographical data from content types and imbuing it with space and time in this way would reintroduce eventfulness as a guiding paradigm for *LCPO*. Users would be able to view timelines of the events in individual poet's lives and see via time-maps how they moved between locations – or, as in most cases, remained local. The benefits would not be limited to single poets; it would be possible to pull events from the lives of many poets, filter them by any number of criteria (such as imprisonment events that affected female poets over the age of 40, in the time range 1700-1800), and recombine them in a new timeline or map. This alters the function of time; rather than serving as a unidirectional force that emphasizes temporal distance, time folds events together into new groups. The time between events collapses as they are combined and structured through categories, presenting a new order which can then again be recursively rearranged.

Several practical limitations to an eventful approach come to mind. A series of complicated joins would be required to reassemble all this information on a single poet landing page in a consumable fashion. The same data sparsity that is already clearly visible in the database (see figure 2) could limit the usefulness of temporal and spatial visualization of events because not all events would have chronologic and geographic dimensions. But perhaps this data sparsity would prompt developers to design different interfaces that would help users understand how such informational gaps retain meaning. I have begun to implement such approaches with basic charts on the *LCPO* website. Figure 3 shows how users can choose whether they include or exclude unknown dimensions within a visualization; users are also able to choose whether to plot

poet birth years, death years, or total lifespan, and can deform the data by altering how long an estimated lifespan is for poets with either a birth or death date but not both.

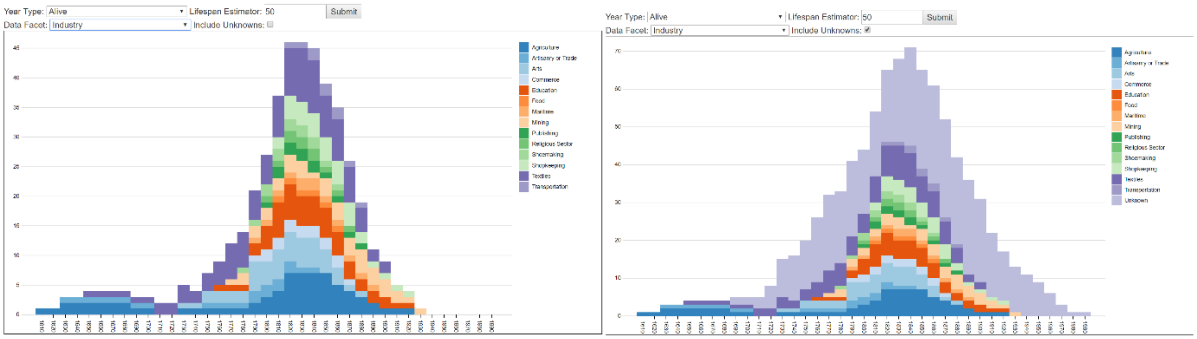

Figure 3: Stacked bar charts showing industries Welsh poets worked in, without and with unknowns included

Maps or timelines that allow intentional data deformation could further push the concept of eventfulness. For instance, for a known event without a specific date and location, users could enter a very generous range of times and places where it may have occurred, such as between 1800 and 1815 somewhere in Northamptonshire, and this fuzzy information would then be used to randomly map or plot the event. These fuzzy dimensions would change upon each viewing of the visualization; in once instance, Clare may have worked in Kettering in 1807, while in another he would be shown as working in Northampton in 1810. The event could be clearly marked as fuzzy or uncertain, but this conscious deformation of geography and chronology would draw attention to the ambiguity and inexhaustibility of events as much as any visual marker. Through such approaches, poet records would begin to evolve from a collection of static properties into an assemblage of insuppressible events composed of deep time and space.

An eventful approach to poet biographical data pushes back against the big data fantasy of information that is high-volume, trustworthy, comprehensive, and well-structured by weaving complex understandings of time and space back into *Laboring-Class Poets Online*. Moving

forward, boutique databases present an opportunity to advance digital scholarship at any scale by minding the gaps, small size, incompleteness, and messy fuzziness of historical and humanities data. In an increasingly digitally inflected world, a boutique approach to data retains the core strengths of humanistic inquiry and interpretation by striking a critically informed balance between rejecting computational methodologies outright and enthusiastically embracing a big data mindset. Building boutique databases and understanding humanities data from this perspective allows us to ask new questions rather than circumscribe possibilities. It is in the gaps which big data worldviews reject that boutique approaches to data hold the greatest potential.

---

[1] Price (2009) critiques some of the terms currently used to describe and classify digital textual studies scholarship, including "project," "archive," "edition," "database," and "digital thematic research collection." Each of these terms has distinct connotations which impact how digital scholarship is positioned and perceived.

[2] See Ramsay (2004) for a humanities-centered introduction of databases and related concepts such as database and schema design, queries, and database management systems.

[3] Brian Croxall identifies three "red herrings" about big data: 1. Data can "save" us because they are enough on their own; this is false because data always requires interpretation. 2. "Digital humanities methods only work on massive datasets"; this is untrue because digital methods can scale to datasets of many different sizes. 3. Data-driven digital approaches fit every situation; this isn't true, as smaller datasets provide diminishing returns for computationally- or time-intensive methods, and not all questions can be answered through digital methods. (Croxall)

[4] Relational databases such as MySQL, MariaDB, and Oracle can be expanded or altered relatively easy. They store the attributes for individual records (rows) as fields in tables organized in columns, which can be joined through relations via keys. Other database models are less flexible. Relational databases were preceded by hierarchical databases such as IBM's Information Management System, which organizes data in a hierarchical tree structure with child and parent segments; like a file system, each child can only have one parent. Because such structures use pointers to store the hierarchical path, retrieval of information along the path is very fast, but updating information can be problematic because changes must occur in multiple locations. NoSQL databases such as MongoDB, Cassandra, and Neo4J are increasingly popular. They eschew the schemas and tables of relational databases and instead store data as key-value pairs, documents, column families, or graphs. Compared to relational databases, NoSQL databases are more flexible, handle "blobs" of unstructured data better, and scale horizontally better, which is useful for cluster computing. However, they are not as consistent as relational databases and may lack support for complex queries.

[5] I have been involved in the development of *Laboring-Class Poets Online* (http://lcpoets.org) as a digital resource since 2013, when I helped create a preliminary data model, began manual data entry, and built a prototype Omeka website. As the digital component of this MA thesis, I revived the project and developed a fully functional Drupal-based website with a more complex and comprehensive data model. While I am the primary developer of *Laboring-Class Poets Online*, the project's underlying data is the product of decades of collaborative research collected in *A Database of British and Irish Labouring-Class Poets and Poetry, 1700-1900*. John Goodridge is the general editor

and principal writer; Bridget Keegan is the eighteenth-century editor; Mary-Ann Constantine is the Welsh poetry editor; and over two dozen other scholars have contributed information or served on the editorial board.

[6] I borrow the term "boutique data" from Cheryl Ball, Tarez Samra Graban, and Michelle Sidley, who use it to describe small and context-specific "currently inaccessible sets of qualitative and quantitative data" (5). My argument for boutique data emphasizes its size and contextuality rather than its hiddenness or combinative potential.

[7] Formerly the Digging into Data Challenge sponsored by the NEH Office of Digital Humanities and several other international funders, this program was renamed in 2016 to the T-AP Digging into Data Challenge to reflect co-sponsorship by the Trans-Atlantic Platform for the Social Sciences and Humanities, bringing together a total of eighteen different funding organizations. (https://diggingintodata.org/about)

[8] Hayles quotes Matthew Kirschenbaum, who in a 2009 personal interview with Hayles calls this shifting practice "rapid shuttling" (31). See also: Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*.

[9] Humanities data management practices vary widely in terms of accessibility, but generally trail the hard sciences, which have rapidly accepted Open Data principles. Since February 2013, scientific research funded by federal grants must make publications resulting from that research freely available within a year, and researchers must also "account for and manage the digital data resulting from federally funded scientific research with the goal of making these data publicly accessible" ("Public Access Mandates"). Commercially collected big data is proprietary and is rarely shared. I do not focus on open access in this article, but strongly believe that humanities researchers should follow the natural sciences in minimizing inaccessibility by embracing open data principles to minimize dark data and increase access regardless of dataset size. A boutique data approach respects the privacy of cultural heritage collections and vulnerable populations, but in most instances should follow Open Data principles.

[10] As of April 2017: 2,037 poets, ~3000 publications, ~200 non-LC poet figures with ties to poets, ~400 relationships between figures, and numerous controlled vocabularies with several thousand terms.

[11] GeoNames includes an "alternate names" attribute for variant names, including local dialects, and a "isFormerName" attribute and associated chronologic field in the database that could be used to track name changes in locations over time. This attribute has not yet been exposed or populated with information (http://forum.geonames.org/gforum/posts/list/1242.page).

[12] Services such as CHALICE (Connecting Historical Authorities with Links, Contexts and Entities), a historical UK placename gazetteer, its successor DEEP (Digitisation and Exposure of English Place-names), and their Unlock Places API have failed to materialize fully. The DEEP API has been removed as of 2016, and while *The Historical Gazetteer of England's Place Names* (http://placenames.org.uk/index.php/search) provides the ability to geocode individual locations, without support for Welsh, Scottish, and Irish locations and an API for mass geocoding, it is an incomplete solution at best. The Association of British Counties (discussed below) also provides a single-search geocoder and Gazetteer with spatial data for purchase.

[13] There is significant overlap in the distribution of occupations and locations because some industries tended to cluster, such as the numerous Paisley weaver poets, Tyneside miner poets, or Nottingham textile worker poets.

[14] As of April 2017, of the 270 poets in LCPO who used pseudonyms, 102 identified places and 51 identified professions other than "poet," "poetess," and "bard."

[15] When Unix time reaches 19 January 2038, new timestamped dates added to 32-bit databases will fail as integer overflow will occur because a 32-bit field can only hold integers up to 2,147,483,647.

[16] The Partial Date module (https://www.drupal.org/project/partial_date) uses multiple fields to bypass PHP/SQL date formatting limitations, allowing blank components, fuzzy dates, text labels for ranges such as "eighteenth century," and a plaintext description field. Partial Date faces some substantial compatibility issues with other modules.

[17] As of April 2017, there are 1461 poet records with birth years; 951 with death years; 883 with both birth and death years; and 11 with baptismal years. 98 birth years and 41 death years are circa dates. 427 poet records use flourished dates, generally ascertained from publications.

[18] Breve also supports assigning labels to show how the data in different fields was produced, using values such as generated automatically, authored by a researcher, edited by a researcher, and untouched source data. Breve does not currently support tracking a database over numerous versions, which somewhat limits its usefulness in seeing how a database changes during the data entry process.

[19] Publications are most often collections of poetry and songs, but some poets also published non-poetic works such as biographies, travelogues, sermon collections, almanacs, and serialized romance novels.

Bibliography

"About the ABC." *Association of British Counties*. 26 Aug. 2012. Web. 17 Mar. 2017.

"About #transformDH." *#TransformDH*. 2 June 2015. Web. 20 Apr. 2017.

Allington, Daniel, Sarah Brouillette, and David Golumbia. "Neoliberal Tools (and Archives): A Political History of Digital Humanities." *Los Angeles Review of Books*. Web. 20 Apr. 2017.

Armstrong, Richard. "Time Out of Joint." Interview by John Lienhard. *The Engines of Our Ingenuity.* Houston Public Media. 2008. Transcript of radio broadcast.

Ball, Cheryl, Tarez Samra Graban, and Michelle Sidler. "The Boutique Is Open: Data for Writing Studies." *Networked Humanities: Within and Without the University*. Ed. Jeff Rice and Brian McNely. Parlor Press, Forthcoming. Print.

Berendt, Bettina, Marco Büchler, and Geoffrey Rockwell. "Is It Research or Is It Spying? Thinking-Through Ethics in Big Data AI and Other Knowledge Sciences." *KI - Künstliche Intelligenz* 29.2 (2015): 223–232. Web.

Burnett, Margaret. "What Is End-User Software Engineering and Why Does It Matter?" *End-User Development*. Springer, Berlin, Heidelberg, 2009. 15–28. Web. 7 June 2017.

"Breve: See your data." Humanities + Design. Stanford University. 2015. Accessed 6 April 2017.

Chun, Wendy Hui Kyong, Richard Grusin, Patrick Jagoda, and Rita Raley. "The Dark Side of the Digital Humanities." *Debates in the Digital Humanities.* Minneapolis, MN: University of Minnesota Press, 2016. Web. 20 Apr. 2017. Debates in the Digital Humanities (2016).

Croxall, Brian. “The Red Herring of Big Data.” Fresno State University, Fresno, CA. Aug. 2013. Invited Lecture. *www.briancroxall.net*. Web. 7 Apr. 2017.

Derrida, Jacques. “Artifactualities,” trans. Jennifer Bajorek. *Echographies of Television: Filmed Interviews,* Ed. Jacques Derrida & Bernard Stiegler. Cambridge: Polity Press, 2002. 1-28.

Dimock, Wai Chee. *Through Other Continents: American Literature Across Deep Time.* Princeton University Press, 2006. Print.

Folsom, Ed. “Database as Genre: The Epic Transformation of Archives.” *PMLA* 122.5 (2007): 1571–1579.

Gillet, Dave. “Counties of England (current and historic).” ArcGis map; British Ordnance Survey boundryline data and Association of British Counties historic county data. 2016. https://www.arcgis.com/home/item.html?id=7b0e661ef66b4a7aacb5a9acf55108ac

Goodridge, John, ed. *A Database of Self-Taught and Labouring-Class English Poets, 1700-1900*. 2001. Web.

---. Ed. *A Database of British and Irish Labouring-Class Poets and Poetry, 1700-1900.* March 2017. Web.

---. Personal interview. March 2017.

Hayles, N. Katherine. *How We Think: Digital Media and Contemporary Technogenesis*. Chicago: The University of Chicago Press, 2012. Print.

---. “Narrative and Database: Natural Symbionts.” *PMLA* 122.5 (2007): 1603–1607.

Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History*. Champaign, IL: University of Illinois Press, 2013. Print. Topics in the Digital Humanities.

Kaplan, Frédéric. “A Map for Big Data Research in Digital Humanities.” *Frontiers in Digital Humanities* 2 (2015): n. pag. *Frontiers*. Web. 1 Mar. 2017.

Kirschenbaum, Matthew. Interview with N. Katherine Hayles. College Park, MD and Hillsborough, NC. 2009.

Ko, Andrew J., Robin Abraham, Laura Beckwith, Alan Blackwell, Margaret Burnett, Martin Erwig, Joseph Lawrance, Henry Lieberman, Brad Myers, Mary Beth Rosson, Gregg Rothermel, Chris Scaffidi, Mary Shaw, and Susan Wiedenbeck. “The State of the Art in End-User Software Engineering.” *ACM Computing Surveys.* 43.3 (2011): 21:1–21:44. *ACM Digital Library*. Web.

Liu, Alan. “Where Is Cultural Critique in the Digital Humanities?” *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press, 2012. Print. Debates in the Digital Humanities.

Manovich, Lev. *The Language of New Media*. Cambridge, Mass.: MIT Press, 2002. Print.

---. “Trending: The Promises and Challenges of Big Social Data.” *Debates in the Digital Humanities*. Ed. Matthew Gold. University of Minnesota Press, 2012. Print.

Merry, Mark. *Designing Databases for Historical Research*. School of Advanced Study: University of London, 2011. Web.

Moretti, Franco. *Distant Reading*. Brooklyn, NY: Verso, 2013. Print.

---. *Graphs, Maps, Trees: Abstract Models for Literary History*. Brooklyn, NY: Verso, 2007. Print.

Owens, Trevor. “Defining Data for Humanists: Text, Artifact, Information or Evidence?” *Journal of Digital Humanities* 1.1 (2011). Web.

"Pleiades Data Structure." *Pleiades.* 14 Nov. 2015. Web.

Posner, Miriam. "Humanities Data: A Necessary Contradiction." Harvard Purdue Data Management Symposium, Cambridge, MA. 17 June 2015. *Miriamposner.com*. Web. 22 Mar. 2017.

Price, Kenneth M. "Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?" *Digital Humanities Quarterly* 3.3 (2009): n. pag. *Digital Humanities Quarterly*. Web. 11 Jan. 2017.

"Public Access Mandates for Federally Funded Research." *Colombia*. N.p., n.d. Web. 7 Mar. 2017.

Ramsay, Stephen. "Databases." *A Companion to Digital Humanities*. Ed. Susan Schreibman, Ray Siemens, and John Unsworth. Blackwell, 2004. Web. 11 Jan. 2017.

---. *Reading Machines: Toward an Algorithmic Criticism*. Champaign, IL: University of Illinois Press, 2011. Print. Topics in the Digital Humanities.

---. "The Hermeneutics of Screwing Around; or What You Do with a Million Books." *Pastplay: Teaching and Learning History with Technology*. Ed. Kevin Kee. University of Michigan Press, 2010. 111–121. Print. Digital Humanities.

Rawson, Katie, and Trevor Muñoz. "Against Cleaning." *Curating Menus.* 6 July 2016. Web. 1 May 2017.

Risam, Roopika. "Decolonizing Digital Humanities in Theory and Practice." *The Routledge Companion to Media Studies and Digital Humanities*. Ed. Jentery Sayers. New York: Routledge, 2017. Print.

Shirky, Clay. “Situated Software.” *Clay Shirky’s Writings About the Internet: Economics & Culture, Media & Community, Open Source*. 30 Mar. 2004. Web. 7 June 2017.

Tannahill, Robert. “To Robert Clark.” May 1808. *Transforming Robert Tannahill*. Ed. Cole Crawford.

Trumpener, Katie. “Paratext and Genre System: A Response to Franco Moretti.” *Critical Inquiry* 36.1 (2009): 159–171. *JSTOR*. Web.

Ushizima, Daniela, Lev Manovich, Todd Margolis, and Jeremy Douglass. “Cultural Analytics of Large Datasets from Flickr.” *The Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (2012): 30–34. Print.

“We Promote the Historic Counties.” *Association of British Counties*. 27 June 2013. Web. 17 Mar. 2017.

“What Is Big Data?” *Villanova University*. N.p., n.d. Web. 7 Mar. 2017.

Whetstone, David. “Culture of the Coalfield Celebrated in Newcastle Mining Institute.” *The Journal* 17 June 2014. Web.