Scaled Entity Search: A Method for Media Historiography and Response to Critiques of Big Humanities Data Research

Eric Hoyt, Kit Hughes, Derek Long, Anthony Tran
Department of Communication Arts
University of Wisconsin-Madison
Madison, WI, USA
ehoyt@wisc.edu, kthughes2@wisc.edu
drlong@wisc.edu, adtran3@wisc.edu

Kevin Ponto
Department of Design Studies
University of Wisconsin-Madison
Madison, WI, USA
kbponto@wisc.edu

***Abstract*—Search has been unfairly maligned within digital humanities big data research. While many digital tools lack a wide audience due to the uncertainty of researchers regarding their operation and/or skepticism towards their utility, search offers functions already familiar and *potentially* transparent to a range of users. To adapt search to the scale of Big Data, we offer Scaled Entity Search (SES). Designed as an interpretive method to accompany an under-construction application that allows users to search hundreds or thousands of entities across a corpus simultaneously, SES balances critical reflection on the entities, corpus, and digital with an appreciation of how all of these factors interact to shape both our results and our future questions. Using examples from film and broadcasting history, we demonstrate the process and value of SES as performed over a corpus of 1.3 million pages of media industry documents.**

***Keywords*-big data critiques; film; historiography; radio; search**

I. Introduction

For the computational analysis of “Big Humanities Data” to gain broader acceptance, scholars pursuing such methods must address the concerns of critics. Exciting techniques, such as topic modeling and network visualizations, hold the promise of generating new knowledge by mining enormous collections of texts and data and transforming them into abstractions and visualizations. However, digital humanities scholars such as Timothy Hitchcock and Johanna Drucker have critiqued these techniques for lacking in transparency, failing to adequately answer research questions, and making it difficult to think critically about how and why the underlying data was collected and organized [1], [2]. These concerns must be addressed to widen the appeal and impact of “Big Humanities Data” research. The digital humanities needs to innovate methods that can harness the affordances of digital technology and, at the same time, facilitate the pursuit of research questions and the critical interrogation of texts, histories, and data structures.

In this paper, we introduce an analytical and interpretive method called Scaled Entity Search (SES) that we believe can accommodate these diverse demands. Unlike traditional keyword searches, SES allows users to submit hundreds or thousands of queries to their corpus simultaneously. In so doing, SES restores some of the context lost by keyword searches by helping the user to establish and analyze relationships between entities and across time. In addition to the technical method of SES, we propose an analytical framework for SES users. The analytical framework can be conceptualized as a triangle with three points: the entities, the corpus, and the digital. As we explain, users and researchers need to think critically about all three points on the triangle as well as the relationships between the points. After explaining SES’s technical and analytical methodologies, we share and discuss some of our results from applying SES to the Media History Digital Library’s 1.3 million page corpus of magazines and books about film, broadcasting, and recorded sound (http://mediahistoryproject.org). We queried the corpus using entity lists of radio stations and early film directors. This research is informing the ongoing development of Project Arclight, a web-based application for the study of 20th century American media that we are developing in collaboration with PI Charles Acland at Concordia University and with support from a Digging into Data grant sponsored by Canada’s Social Sciences and Humanities Research Council and the U.S.’s Institute of Museum and Library Services.

II. Existing Literature and Projects

As efforts to digitize text collections continue and humanities researchers watch their sources transform into data, scholars have proposed several new methodologies and frameworks to meet the scaled challenges of “big data” [3], [4]. One strain of such research has concerned itself with new methods for identifying, gathering, and sorting evidence across a corpus of texts (or multiple corpora). Using topic modeling, a researcher might locate the themes that best distinguish authors by gender and nationality, survey 20 years of women’s history scholarship to locate points of over-representation and persistent lacunae, or describe the political issues that animate legislative discussion [5], [6], [7]. Or, for scholars more interested in the relationships between records or between specific entities appearing across records, network analysis offers opportunities to track and even predict these connections [8], [9]. Yet another strategy for making large datasets meaningful is geographic modeling that "grounds" texts within representations of our physical world [10], [11]. Although space precludes a full accounting of the mass of methods and new questions inspired by digital corpora, these projects nevertheless give an account of how researchers continue to make these sources useful for locating new evidence and scholarly vantage points, creating productive classifications, and working towards making sense of the inhuman scale of big data.

One casualty, however, of this pursuit of new methods is a relatively senior strategy for gathering evidence from large-scale digital corpora: search. In discussing applying text analytical procedures to digital corpora, Stephen Ramsay refers in passing to "that most primitive of procedures: keyword search" [12]. Moreover, as data mining techniques pick up speed, "beyond search" threatens to become a watchword for large-scale digital humanities research.[1] According to Matthew Jockers, a leader in data mining methods for literature, "the sheer amount of data now available makes search ineffectual as a means of evidence gathering," and, further, "is not terribly practical" [4]. One of the authors of this paper, Eric Hoyt, similarly distinguished data mining methods from search in an earlier article [14]. The phrase "beyond search" works marvelously for rhetorical purposes. Because most humanities researchers are familiar with keyword search, "beyond search" becomes a shorthand way of calling for scholars to adopt less familiar digital processes.

In fairness, there are legitimate concerns that conventional keyword searches—in which users browse snippets of results for relevant documents—cannot possibly accommodate all relevant evidence within the massive scale of today's digital corpora. Furthermore, others have noted that the algorithms used by search engines may be shaped by the hidden desires of institutions or the profit-motive of corporations rather than more academic interests [15], [16]. Compounding the impact such algorithms bear on research is the design of search system interfaces that obscure the rules governing search and the limitations of returned results [15], [17]. Even leaving aside search algorithms potential lack of transparency, Ted Underwood argues that search is little more than "a Boolean fishing expedition" that strengthens confirmation bias and filters out oppositional evidence by "only show[ing] you what you already know to expect" [17].

Partially for these reasons, researchers place increasing emphasis on "unsupervised" methods of data discovery that use non-proprietary open-source tools adaptable to humanistic pursuits. Contrary to search, which presumes a user motivated by a hypothesis—no matter how ill-defined—unsupervised methods like topic modeling require no major hypothetical input from the user before generating results.[2] Referred to as "perhaps the greatest strength" of certain big data projects, "tabula rasa interpretation," aiming "to banish, or at least crucially delay, human ideation at the formative onset of interpretation" supposedly offers the novel opportunity to encounter defamiliarized texts free from pre-existing hypotheses, thus avoiding analyses that bend interpretation to the preconceived notions of the researcher [7], [18].[3] When keyword search does gain attention as a potential tool for large-scale corpora analysis—as it did most publicly with the release of Google's Ngram viewer—concerns arise over the limitations of temporally tracking complex concepts through a handful of single terms (the primary suggested use of the service), the inconstancy of word meaning over time, and the shape of the corpora [19], [20].

Such services, however, can be adapted to exhibit greater critical possibility. Similar to the Ngram viewer is Bookworm, a collaboration between Harvard University, the Encyclopedia Britannica, the American Heritage Dictionary, and Google that allows users to perform keyword searches over several corpora, returning line graphs that show how terms trend, over time, through each corpus [3]. By allowing users to easily return to documents within the corpus by clicking on trend lines and using facets to narrow the publications searched, Bookworm allows for toggling between several scales of research, moving between corpus, publication, and document levels. While Bookworm and SES both pursue this "middle ground," SES builds away from Bookworm through its emphasis on massive relational search, its specific technical process, and an interpretive framework based on transparency and reflexivity. As Jo-

[1] "Beyond Search" served as the title of a 2006-2009 workshop series—which eventually became the Stanford University Literary Lab in 2010—helmed by Matthew Jockers. "Beyond search" has also been taken up within online marketing research as a means of describing new methods of distributing sponsored content and targeted search results to users. See, for example, the "Beyond Search" workshop, conference and awards developed by Microsoft Research in 2008 and 2009, a reminder of the unlikely parallels between business and academe in the pursuit of "big data." The same phrase has also been used to describe the promise of improved user (affective) experience in the context of web search engines [13].

[2] A user does need to specify the number of topic models they want to generate and, in most cases, provide a "stop list" identifying words that should be ignored by the modeling program.

[3] Liu further suggests this goal of "tabula rasa interpretation" is more difficult to realize than most admit; "It is not clear epistemologically, cognitively, or socially how human beings can take a signal discovered by machine and develop an interpretation leading to a humanly understandable concept unless that signal (in order to be recognized as a signal at all) contains a coeval conceptual origin that is knowable in principle because, at a minimum, the human interpreter has known its form or position (the slot or approximate locus in the semantic system where its meaning, or at least its membership in the system, is expected to come clear)." Although Underwood's response to Liu highlights several ways in which "topic modelers" are well aware of how certain hypotheses shape topic modeling algorithms, e.g., the selection of the number of topics and the "blurriness" of topics, he does not respond to Liu's larger critique which seems more interested in the promise of "discovering" topic *content* free from hypotheses and assumptions.

hanna Drucker argues, visualizations such as trending charts reify data, forcing them to fit uncomfortably into standardized metrics, and seductively suggest certainty and self-evidential results [1]. The SES interpretive framework has been designed to respond to calls to maintain the strengths of the humanities—critical uncertainty, nuanced and careful interpretations, the development of subjective and situated knowledge—within the context of big data analysis [1], [2].

Information retrieval researchers, drawing from psychology, education, and information and library sciences, continue to emphasize the value of search and its ability to save time and serve a range of information needs. Indeed, no matter how sophisticated the end results, most research begins with online searching in some form or another. Furthermore, the actual practice of search is rarely the simple and uncritical entry of one or two key terms into a box that it is frequently made out to be. Operating from a descriptive rather than prescriptive intent, work on "exploratory search" reveals certain processes by which users gather and understand information to be complex and adaptive to different information environments. That exploratory search is already endemic to the way many people use digital tools to gather and consult data recommends the process.

Unlike topic modeling and other tools that have been criticized as "black boxes" for their lack of transparency, search—especially through modifications like visible facets and a simple yes/no registration of search terms on each page to determine relevance rankings—has the potential to be relatively well understood by a range of users [20]. Furthermore, considering digital tools' low rate of adoption—due in large part to "traditional" humanities researchers' confusion or skepticism over their value and operation—search may be a productive site for negotiating between the promise of big data and the expectations and desires of the majority of academic researchers [21]. While search may not allow for the same level of "non-subjective" discovery as unsupervised computational analytics, the precision of search, powered by users' domain expertise, offers researchers valuable mechanisms for locating and gathering data. Moreover, due to SESs emphasis on self-reflexive analysis at each step of the data collection and interpretation process, the method responds to scholars' calls to recognize search as itself a special mode of text mining—one well overdue for theorization. SES thus opens up opportunities to reevaluate the possibilities and pitfalls that attend an entrenched scholarly practice. Given this continuing value of search, the question becomes: how might we best leverage users' comfort and expertise with search to create new digital tools and methods scaled to "big data"?

III. SES: Technical Method

The SES method utilizes an Apache Solr search index as its algorithmic backbone.[4] In addition to being open source, Solr possesses five qualities that make it well suited for SES. First, Solr accommodates scale; libraries and companies routinely store and search millions of documents in Solr indexes.[5] Second, Solr is optimized for speed, enabling it to return results in seconds compared to databases that take several minutes. Third, Solr offers faceted search capabilities, which enable the organization and counting of search results by core metadata fields (e.g., year of publication or magazine title). Fourth, Solr offers flexible query parameters, which can be used to target certain fields and print only their facet counts. Fifth, and not to be overlooked, Solr is already used by a large number of university libraries and digital collections. Researchers may find pre-existing Solr indexes that they can use for SES analysis. If not, they may be able to find experts on campus who are familiar with Solr and can help them create an index.

In our case, we began testing and improving SES on the pre-existing Solr index of Lantern, which is the Media History Digitals Library's (MHDL) search platform [22]. The MHDL's dataset and Lantern's index consist of a collection of roughly 1.3 million discrete XML documents representing individual pages from thousands of out-of-copyright trade papers, magazines, and books related to film, broadcasting, and recorded sound. The core of the collection spans the years 1905 to 1964. The high resolution image files are stored at the Internet Archive (which serves as the MHDL's scanning vendor and preservation repository). However, the page-level XML was created through XSLT transformation and Python scripting. Each transformed MHDL XML document collates metadata for the publication with the OCR body text of each individual page. One major strength of the MHDL dataset (apart from its simultaneous claims to scope and focus) is the relatively high quality of OCR text, which was generated from high-quality print originals.

Either before or after building the Solr index, researchers need to generate entity lists—from existing databases or other sources—relevant to the indexed corpus. In our case, we generated entity lists related to the histories of early cinema and radio (the MHDL has digitized numerous trade papers and fan magazines that covered these industries, and we have domain expertise and ongoing research projects focused on these areas). To generate the entity list for cinema, we used an existing dataset containing credits information for all known films produced between 1908 to 1920 (35,686 films total).[6] Using Perl and XSLT, we output the names of

[4] For more on Solr's documentation and download instructions, see http://lucene.apache.org/solr/.

[5] Librarians and developers designed Blacklight to provide an interface for the Solr indexes that many libraries are using. See http://projectblacklight.org/.

[6] This is the same credits data detailed in [23] and [24].

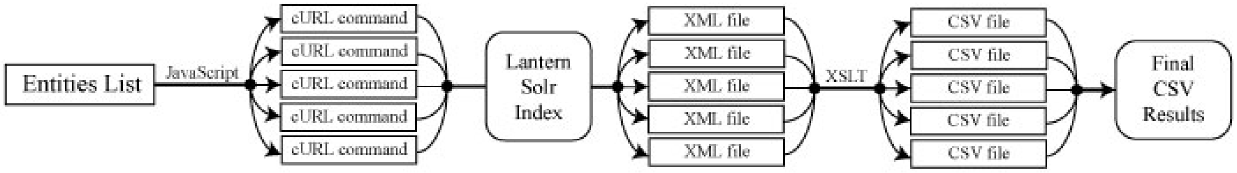

Figure 1. The SES technical process uses an entity list to query Solr, then transforms the results into a single CSV file.

the 1,548 directors listed in the dataset. We could not locate any comparable structured dataset for radio, so we generated one ourselves by going through the 1948 *Radio Annual* and entering the call letters, established date, location, power, frequency, owners/operators, airtime, and market size for every U.S. and Canadian radio station noted in the book (2,002 stations total) [25]. 1948 holds special significance to our field as the first year of the FCC Freeze, which is widely understood as the period that cemented network control over the first few decades of television [26].

Next, we created a simple for loop that extracts the entities and enters them as variables into a Solr query. We designed the Solr query to target the fields and facets that interested us. Here is an example of the query automatically generated and run for WCCO, a Minneapolis radio station that we became interested in based on its unexpected prominence in our results:

http://solrindex.commarts.wisc.edu:8080/solr/select?q=%7B!dismax%20qf=body%7DWCCO&rows=0&facet=true&facet.limit=110&facet.range=year&f.year.facet.range.start=1890&f.year.facet.range.end=2001&f.year.facet.range.gap=1&f.year.facet.missing=true&f.year.facet.mincount=-1&facet.field=title&stats=true&stats.field=year

This query returns: A) the number of matching pages for each entity for every year between 1890 to 2000 (even years with zero hits are returned to allow for a standardized set of columns and easy comparisons within Excel); B) the titles of books and magazines mentioning the entity and the number of matching pages for any given title; C) Solr's StatsComponent, which provides a count of the total number of pages that each entity appears in. To run queries on an entity described by multiple words (like a movie title or person's name), the query parameters operate almost identically, except they include escaped quotation marks around the entity and escaped spaces between words.

After running each query, the for loop saves the results locally as an XML file named after the query (for instance, "WCCO.xml"). Using an XSLT script, the XML files are converted into CSV files (for instance, "WCCO.csv"). Finally, the CSV files are merged into a single CSV file that can be opened and analyzed in R or Excel. Our aggregated CSV file from running SES on the radio station list generated 2,002 rows (each one representing a different station) and 113 columns that track the number of matching pages per year, the total number of matching pages, and associated metadata that we had collected for the entities (for the radio stations, this included all database information mentioned in the above list).

Tracking only whether or not an entity appears on a page (yes/no) on a year-by-year basis may strike some as an overly blunt method for comparing how entities trend over time. Why not count an entity mentioned ten times on a page more highly than an entity mentioned only once? We recognize this potential objection, but we believe the yes/no tracking on a page-by-page basis makes sense for several reasons. First, when applied at scale to 1.3 million pages of text, the distinctions between the amount of attention entities get on a certain page become less important; the outliers and exceptional entities still rise to the top. Second, if a single entity is named multiple times in a single page, then the redundancy helps mitigate the problem of SES missing instances of the entity due to imperfect OCR. Third, the page-level logic of SES makes the process much easier to conceptualize for users who aren't experts in how search algorithms or logarithmic smoothing work. Grounding our process in Boolean logic and simple mathematical addition and division helps to keep the method transparent and less like the proverbial "black box."

When running SES on lists of names, we recommend pre-processing the entity list to reduce the likelihood of returning false positives or false negatives. Our early cinema dataset, for example, tracked 1,548 names who were credited as having directed at least one film between 1908 and 1920. An inspection of the list, however, revealed that the same individuals were credited slightly differently, resulting in duplicate entries. To generate better results, we wrote queries using the Boolean OR operator to collect all instances of the individual. For instance, we combined the four entities "Al Christie," "Al E. Christie," "Albert E. Christie," "Al. E. Christie" into one query.

Similarly, pre-processing entities can help provide a level of disambiguation that reduces the number of false positives. One of the most important film companies of the 1910s and 1920s was named Universal (which is still an active studio, though its ownership has changed several times). To

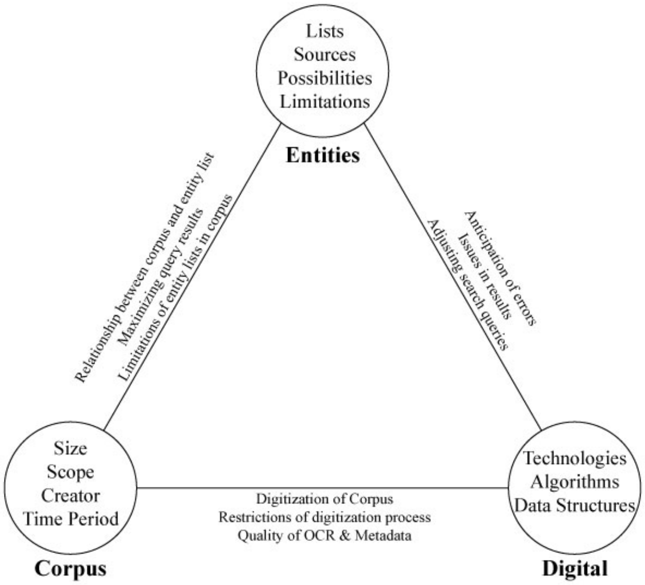

Figure 2. The SES triangle method of interpretation.

cut down on the number of instances in which "universal" appears as an adjective, one could design a Boolean query that searches for

"Universal Film" OR "Universal Manufacturing" OR "Universal Picture" OR (Universal AND Laemmle)

As we describe in the next section, this transparency and flexibility is a strength of SES.

IV. SES: Interpretive Method

Of prime importance to the full realization of SES as a humanistic method of big data analysis is a triangulated interpretive framework that balances critical understandings of the entities, the corpus, and the digital, with particular care given to the relationships between each of these elements (see Figure 2). This framework aims to keep SES transparent and self-reflexive. We address each point of the triangle and the relationships between them below.

The Entities: SES users reflect on how they select their entity list(s). Questions to ask: How and why did you select this grouping to compare? If you did not generate the entity list yourself, where did it come from? What sources were used to generate the data? How does this list open up new possibilities for research? How does it limit or close down other possibilities?

The Corpus: SES users reflect on the corpus that is being queried. Questions to ask: What is the size and scope of the corpus? Who created it and why? What are its strengths and weaknesses in terms of the time periods covered and diversity of publications?

The Digital: SES users reflect on the digital technologies, algorithms, and data structures that comprise the process. Questions to ask: What schema, fields and facets were used in creating the search index? What historical materials, processes, and experiences do not easily lend themselves to digitization and what effect does their omission have on results? How does making materials machine-readable change the research process?

The Entities-Corpus Relationship: What is the relationship between the list of entities you are querying and the corpus? How could you design an entity list that plays to the strengths of the corpus? At the same time, if we only design research questions and entity lists on the basis of what is likely to generate interesting results in the corpus, how does this limit scholarship?

The Corpus-Digital Relationship: How did the digitization process change the nature of the corpus? What is the quality of the OCR text? How did intellectual property restrictions and other factors influence what material was digitized and what was left out? How granular is the metadata that describes the corpus and is it consistent? Is the underlying corpus data openly accessible, viewable, and reusable? We contend that it should be to keep the process transparent and repeatable.

The Entities-Digital Relationship: What issues of disambiguation, false positives, and false negatives can you anticipate before querying the entities? What issues do you recognize in examining the queried results? How do you adjust the search queries to try to mitigate these problems? Do you make these adjustments consistently or selectively?

We think of the SES analytical triangle much like an algorithm—an iterative process that researchers can return to again and again as they work. However, we do not believe that SES researchers need to limit their analysis to this triangle model. As noted earlier, many digital humanities data mining techniques could benefit from more critical interrogation and self-reflexivity. But this does not mean that the end goal of the SES research process should be only to generate meta-commentaries and critiques of all Big Data analysis and visualization. The triangle model helps researchers interpret their results and qualify historical claims, but, as we demonstrate below, it can also answer research questions, spark new inquiries, and generate knowledge.

V. Results and Discussion

We tested the SES method on the MHDL corpus by querying two large entity lists: 2,002 radio stations and the names of 1,548 individuals who are credited as having directed at least one film between 1908 to 1920.

Before drawing conclusions from the radio station results, we reflected on the relationship between the entities and the digital, following one of the edges of the triangle interpretive model. Out of the 2,002 entities, 63 stations, or 3.1%, returned extremely high page counts due to call letters that doubled as words and, therefore, flagged numerous false positives in the OCR (examples include Peoria's WEEK and Urbana's WILL). We removed these 63 entries from the remainder of our SES analysis, but noted that an interesting future research question might be how stations attempted to brand themselves by asking the FCC for call letters with a semantic meaning, rather than a four letter ID that lacked pre-established meaning or memorability. After setting aside the obvious false positives, we returned to analyzing the results.

Prior to running SES on the station list, we hypothesized that the stations in the largest U.S. markets would have received the most attention in the industry press and, therefore, have the highest page counts. In our results, we found this largely to be true. For the 20 stations mentioned most in industry publications between the years of 1920 to 1964, see Table I.

Table I
RADIO STATION RESULTS.

Rank	Station ID	Market
1	WGN	Chicago
2	WJZ	New York
3	KDKA	Pittsburgh
4	WMCA	New York
5	KYW	Philadelphia
6	WLS	Chicago
7	WBBM	Chicago
8	WBZ-WBZA	Boston & Springfield
9	WCAU	Philadelphia
10	WHN	New York
11	KHJ	Los Angeles
12	WSB	Atlanta
13	KNX	Los Angeles
14	KFI	Los Angeles
15	WGY	Schenectady
16	WWJ	Detroit
17	WCCO	Minneapolis
18	WIP	Philadelphia
19	KGO	San Franscisco
20	WFAA	Dallas

In evaluating this table, we can see that 8 of the 10 most discussed stations were from the three most populous American cities between 1920 and 1950: New York, Chicago, and Philadelphia. The station that ranked third, Pittsburgh's KDKA, holds historical significance that helps to explain its prominence. As historians of American radio have noted, KDKA's broadcasts of the 1920 presidential election results and a 1921 boxing match proved highly influential on the public's perception of what radio could offer as a medium [27], [28].

The outliers on this list are Atlanta's WSB, Minneapolis's WCCO, Dallas's WFAA, and Schenectady's WGY. These cities had populations that were less than a tenth the size of the New York market.[7] What explains their prominence? In the case of the list's smallest market station, WGY Schenectady was one of the five owned-and-operated stations of NBC used as a launch pad for new technological innovations [27]. Additional research would be required to more fully explain why WSB, WCCO, and WFAA received such attention within the industry's press. However, SES analysis allows us to quickly recognize that these stations were outliers; a historian using WCCO as a case study would know that this Minneapolis station was especially noteworthy. Conversely, if a historian wanted to profile a station that was not an outlier, one that was utterly representative of the industry-wide attention most stations received for a given year (let's say 1929), then Sioux City's KFWB or Shreveport's KOIN would be great examples, occupying the median for that year. As these examples show, SES allows humanities researchers to apply big data techniques to better understand individual entities (e.g., radio stations) from a historical and comparative perspective.

By using the triangle model and reflecting on the relationship between the corpus and the entities, we also came away with a better understanding of the magazines and the strengths and weaknesses of the corpus itself. The two broadcasting-oriented trade papers most represented in the MHDL corpus are *Broadcasting* and *Sponsor* (representing 110,138 and 60,726 pages, respectively). Even after accounting for *Broadcasting's* nearly double page count, we found that nearly all the stations were discussed more frequently in *Broadcasting* than *Sponsor*. This trend applied to both large market and small market stations, suggesting that *Broadcasting's* coverage of the industry gave far more attention to individual stations than *Sponsor* and that the managers and representatives of stations were a more important audience for *Broadcasting*. This insight, obtained from distant reading, provides a contextual frame for the close reading and interpretation of articles from *Sponsor* or *Broadcasting* discussing particular stations.

From reflecting on the corpus, the entities, and the digital, we also realized that the corpus has a weakness in its broadcasting coverage during the early-1930s and mid-1940s. We recognized this gap after applying Excel's conditional formatting color filter over the SES results and noticing that all stations seemed to be trending downward during these periods. This downward trend cannot be reflective of U.S. broadcasting, which grew at a rapid pace during the early 1930s and mid-1940s. Instead, the trend reflects that the Media History Digital Library has scanned far fewer broadcasting-oriented publications from those years than the periods of the 1920s and the periods of 1936 to 1944 and 1948 to 1963. We were able to arrive at this insight because, unlike Google Ngram search, SES enables and encourages

[7] City populations determined from historical U.S. census data, https://www.census.gov/population/www/documentation/twps0027/tab15.txt.

researchers to ask questions about the corpus and investigate the underlying texts. This insight may also lead toward improvements to the corpus; an author of this paper, Eric Hoyt, is also co-director of the Media History Digital Library and now recognizes that early-1930s radio publications need to be a priority for scanning.

Fortunately for research into early American film history, the MHDL corpus has no comparable gaps. For the years that match our film director credits list (1908 to 1920), the MHDL collection includes extensive runs of five film trade papers (*Moving Picture World*, *Motion Picture News*, *Motography*, *Film Daily*, and *Exhibitors Herald*), two theatrical papers that also covered film (*Variety* and *New York Clipper*), and four movie fan magazines (*Motion Picture Magazine*, *Picture-Play*, *Photoplay*, and *Film Fun*). Certain years have more magazines and pages indexed than other years, but by dividing an entity's number of page hits for a given year by the total number pages indexed for that year, we can normalize the results.

One question that interested us was when "the director" became a category of film worker who received significant attention in the motion picture press. To begin to answer this question, we plotted the number of page hits between 1905 to 1920 for the directors represented in the early cinema credits dataset using Kernel Density Estimation (KDE) as shown in Figure 3. Formally, KDE can be expressed as:

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K(x - x_i) \qquad (1)$$

where K is the kernel function, h is a smoothing parameter called bandwidth, n is the number of datapoints, and x is an independently and identically distributed sample drawn from some unknown density. This method has been as an effective visualization method for large amounts of samples [29], [30], [31]. We chose to set the kernel function as the inverse of the squared distance. The color gradient was selected from the cubehelix color ramp in order to give a better perceptual understanding for the viewer [32].

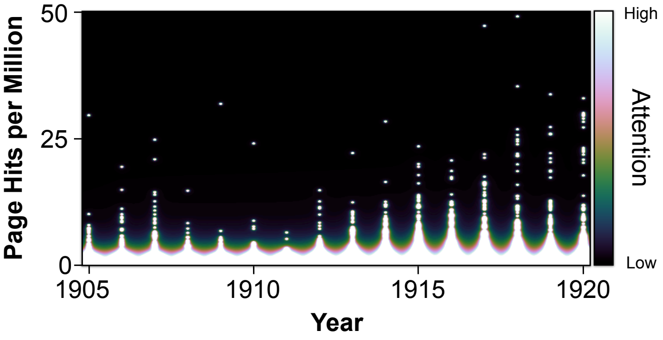

Figure 3. Scatter plot of number of page hits between 1905 and 1920 for the directors represented in the early cinema credits dataset.

Figure 3 shows the concatenation of 24,000 datapoints from over 1,500 directors. This visualization demonstrates a generally upward trend over the years between 1905 and 1920. We found that directors received significantly more attention in the industry press from 1914 to 1920 than they did from 1908 to 1913. What explains this shift? The category of the director had existed since at least 1907, so it cannot be because directors themselves were new arrivals on film sets [33]. One part of the explanation has to do with the rise of the feature film, which occurred in the mid-1910s. Feature films required greater product differentiation than the program shorts that preceded them and continued to be distributed in the mid-1910s [34], [35]. In the feature era, trade papers placed greater emphasis on film reviews. Similarly, advertisements became oriented toward particular films (rather than simply the "manufacturer," to use the term of the day). Directors were noted more in ads and reviews than they had been prior to the feature era.

The rise of the feature film has been well documented by film historians [34], [35], [36], [37]. However, by applying SES and digging deeper into the relationship between the entities and corpus, we found another explanation for the increasing prominence of directors as a class of workers: directors themselves were seeking and obtaining more publicity. Consider, for instance, one director from the credits list: Harry Millarde. When we explored the matching pages associated with Millarde, we noticed that numerous pages were personal advertisements that he had taken out, promoting himself as a director-for-hire.[8] Elsewhere, Millarde's name popped up in short news items, which Millarde or a paid publicist likely placed. This finding suggests something new and interesting about the film industry's labor marketplace in the mid-to-late-1910s. Directors sought to differentiate themselves from one another within the industry—a parallel development to how companies, stars, and film titles were used to differentiate feature films to exhibitors and the public. Directors used the trade press to position themselves within the industry, and the trade press used directors to increase their advertising revenue. By thinking about the entities in relation to the corpus, we improve and broaden our understanding of film history and the media industry ecosystem.

VI. Limitations

The ability of SES to address massive numbers of entities simultaneously is both its key strength and a potential limitation. As the size of named entity lists grow to the tens and hundreds of thousands, researchers' results will increasingly need additional tools, such as visualizations, to make them human-readable; in turn, this incorporation of additional computer-aided techniques requires further theorization. Increasing the scale of analysis also places

[8] See, for example, [38].

greater strain on those tasked with the production of high-quality, high-volume named entities lists; as we note above, this is an area where collaboration is key. Likewise, we designed these methods with historical analysis in mind, it is up to further research to determine how we might apply SES to questions of textuality and aesthetics. Last, although SES can broadly sketch correspondences between terms insofar as they trend over time, it cannot measure terms' co-location in specific texts. For those interested in such questions, which are best served with topic modeling and other methods, we hope that SES offers an additional lens of analysis that can deepen these existing approaches.

VII. Conclusion

Scaled Entity Search offers humanities researchers a method that accommodates scale, applies a familiar technique (search) in a novel way, and invites critical interrogation of the entities, corpus, and digital. As our examples of the radio station and film director lists demonstrate, the process makes a valuable contribution to our home discipline of film and media history. However, we also believe SES makes a worthwhile intervention in big data humanities research by building off the important work of both scholars who have innovated data mining techniques for humanities datasets and scholars who have critiqued those techniques. We hope that scholars in the digital humanities use SES and develop more methods that can meet the needs and demands of a large base of humanities users. Our own immediate challenge with Project Arclight is to deploy a web-based version of SES, with a graphical user interface, that does not sacrifice the methods emphasis on asking critical questions about the entities, corpus, and digital. Our success or failure in achieving this objective will have to wait for another research paper—and the feedback of digital humanities scholars and critics.

Acknowledgment

The authors wish to thank the collaborators and sponsors who have made this research possible. Project Arclight has brought together a large network of interdisciplinary scholars, and we are grateful for their ongoing collaboration. We would especially like to thank Charles Acland for his contributions and feedback on this paper. Project Arclight received a Round 3 Digging into Data grant with funding from the Social Sciences and Humanities Research Council in Canada and the Institute of Museum and Library Services in the U.S. Additionally, our work has received support from Concordia University and the University of Wisconsin-Madison's Graduate School. Finally, we are grateful to the Media History Digital Library and Paul Spehr for making their datasets available for analysis.

References

[1] J. Drucker, "Humanistic theory and digital scholarship," in *Debates in the Digital Humanities*, M. K. Gold, Ed. Minneapolis: University of Minnesota Press, 2012, pp. 85–95.

[2] T. Hitchcock. (2014, July) Big data for dead people: Digital readings and the conundrums of positivism. [Online]. Available: historyonics.blogspot.co.uk

[3] J. M. et al., "Quantitative analysis of culture using millions of digitized books," *Science*, vol. 331, pp. 176–182, 2011.

[4] M. L. Jockers, *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

[5] M. L. Jockers and D. Mimno, "Significant themes in 19th century literature," *Faculty Publications – Department of English, University of Nebraska, Lincoln, Paper 105*, pp. 1–23, 2012.

[6] S. Block and D. Newman, "What where, when, and sometimes why: Data mining two decades of women's history abstracts," *Journal of Women's History*, vol. 23, no. 1, 2011.

[7] K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev, "How to analyze political attention with minimal assumptions and cost," *American Journal of Political Science*, vol. 54, no. 1, pp. 209–228, 2010.

[8] V. Szabo, "Transforming art history research with database analytics: Visualizing art markets," *Art Documentation: Journal of the Art Libraries Society of North America*, vol. 31, pp. 158–175, 2012.

[9] J. Chang and D. M. Blei, "Hierarchical relational models for document networks," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 124–150, 2010.

[10] T. Brown, J. Baldridge, M. Esteva, and W. Xu, "The substatial words are in the ground and sea: Computationally linking text and geography," *Texas Studies in Literature and Language*, vol. 54, no. 3, pp. 324–339, Fall 2012.

[11] T. Elliott and S. Gilles, "Digital geography and classics," *Digital Humanities Quarterly*, vol. 3, no. 1, 2009.

[12] S. Ramsay, *Reading Machines: Toward and Algorithmic Criticism*. Urbana: University of Illinois Press, 2011.

[13] E. M. Bryant, R. Harper, and P. Gosset, "Beyond search: A technology probe investigation," *Library and Information Science*, vol. 4, pp. 227–250, 2012.

[14] E. Hoyt, "Lenses for lantern: Data mining, visualization, and excavating film history's neglected sources," *Film History*, vol. 26, no. 2, pp. 146–168, 2014.

[15] L. Gitelman, "Searching and thinking about searching jstor," *Representations*, vol. 127, no. 1, pp. 73–82, Summer 2014.

[16] F. Kaplan, "Linguistic capitalism and algorithmic mediation," *Representations*, vol. 127, no. 1, pp. 57–63, Summer 2014.

[17] T. Underwood, "Theorizing research practices we forgot to theorize twenty years ago," *Representations*, vol. 127, no. 1, pp. 64–72, Summer 2014.

[18] A. Liu, "What is the meaning of the digital humanities to the humanities?" *PMLA*, vol. 128, no. 2, pp. 409–423, 2013.

[19] M. Liberman. (2013, August) The culturomic psychology of urbanization. [Online]. Available: languagelog.ldc.upenn.edu

[20] B. M. Schmidt, "Words alone: Dismantling topic models in the humanities," *Journal of Digital Humanities*, vol. 2, no. 1, Winter 2012.

[21] F. Gibbs and T. Owens, "Building better digital humanities tools: Toward broader audiences and user-centered designs," *Digital Humanities Quarterly*, vol. 6, no. 2, 2012.

[22] Lantern. [Online]. Available: http://lantern.mediahist.org/

[23] E. Lauritzen and G. Lundquist, *American Film-Index 1908-1915*. Stockholm: University of Stockholm Akademiebokhandeln, 1976.

[24] ——, *American Film-Index 1916-1920*. Stockholm: Huddinge/Tonnheims, 1984.

[25] J. Alicoate, Ed., *The 1948 Radio Annual*. Radio Daily: New York, 1948.

[26] W. Boddy, *Fifties Television: The Industry and its Critics*. Urbana: University of Illinois Press, 1990.

[27] D. Gomery, *A History of Broadcasting in the United States*. Malden, MA: Blackwell Publishing, 2008.

[28] M. Hilmes, *Radio Voices: American Broadcasting, 1922-1952*. University of Minnesota Press, 1997.

[29] A. Huynh, K. Ponto, A. Y.-M. Lin, and F. Kuester, "Visual analytics of inherently noisy crowdsourced data on ultra high resolution displays," in *Aerospace Conference, 2013 IEEE*, March 2013, pp. 1–8.

[30] O. D. Lampe and H. Hauser, "Interactive visualization of streaming data with kernel density estimation," in *Pacific Visualization Symposium (PacificVis), 2011 IEEE*. IEEE, 2011, pp. 171–178.

[31] M. Florek and H. Hauser, "Quantitative data visualization with interactive kde surfaces," in *Proceedings of the 26th Spring Conference on Computer Graphics*. ACM, 2010, pp. 33–42.

[32] D. Green, "A colour scheme for the display of astronomical intensity images," *arXiv preprint arXiv:1108.5083*, 2011.

[33] D. Bordwell, J. Staiger, and K. Thompson, *The Classical Hollywood Cinema: Film Style and Mode of Production to 1960*. New York: Columbia University Press, 1985.

[34] M. Quinn, "Distribution, the transient audience, and the transition to the feature film," *Cinema Journal*, vol. 40, no. 2, pp. 35–56, 2001.

[35] ——, "Paramount and early feature distribution: 1914-1921," *Film History*, vol. 11, no. 1, pp. 98–113, 1999.

[36] B. Brewster, "Periodization of early cinema," in *American Cinema's Transitional Era: Audiences, Institutions, Practices*, C. Keil and S. Stamp, Eds. Berkeley: University of California Press, 2004, pp. 66–75.

[37] B. Singer, "Feature films, variety programs, and the crisis of the small exhibitor," in *American Cinema's Transitional Era: Audiences, Institutions, Practices*. Berkeley: University of California Press, 2004, pp. 76–100.

[38] N. Author, "Harry millard, director (advertisement)," *The Film Daily*, May 11 1919.