December 2014

Todd Presner, **"The Ethics of the Algorithm: Close and Distant Listening to the Shoah Foundation Visual History Archive,"** forthcoming in: *History Unlimited: Probing the Ethics of Holocaust Culture* (Cambridge: Harvard University Press, 2015), paired with Stephen Smith, "On the Ethics of Technology and Testimony: A response to Todd Presner."

Abstract:

   With more than 52,000 testimonies, 100,000 hours of video footage, and a database of some 6 million records, the Shoah Foundation's Visual History Archive is the largest archive of Holocaust testimony in the world. But more than an archive of eyewitness testimony, it is also an information management system, a patented digital library, and a generalizable database for indexing and cataloguing genocide. Presner examines how forms of computation – specifically databases, data structures, algorithms, and information visualizations – function as specific modes of historical emplotment that raise significant ethical questions. Through an investigation of the entirety of the Shoah Foundation's database, Presner shows how computational analysis can be "read against itself" in order to reveal certain assumptions and patterns in the data. In so doing, he argues for the development of an "ethics of the algorithm" based on insights from the Jewish ethical tradition. In response, Stephen Smith, Executive Director of the USC Shoah Foundation, argues that the ethical dimension of digital technologies is to be found in the way in which they preserve memory in perpetuity through "data integrity" and facilitate global access and, potentially, understanding. In so doing, Smith shows how the mission of the Shoah Foundation has broadened to include the preservation of first-hand testimony of survivors of other genocides and human atrocities, including the Armenian genocide, the Rwandan genocide, and the Nanjing Massacre.

Todd Presner is Professor of Germanic Languages and Comparative Literature at the University of California Los Angeles. He is the Sady and Ludwig Kahn Director of the UCLA Center for Jewish Studies and is also the Chair of the Digital Humanities Program. He is the author or co-author of four books: *Mobile Modernity: Germans, Jews, Trains* (Columbia University Press, 2007), *Muscular Judaism: The Jewish Body and the Politics of Regeneration* (Routledge, 2007), *Digital_Humanities* (MIT Press, 2011), with Anne Burdick, Johanna Drucker, Peter Lunenfeld, and Jeffrey Schnapp), and *HyperCities: Thick Mapping in the Digital Humanities* (Harvard University Press, 2014), with David Shepard and Yoh Kawano.

Stephen Smith is Executive Director of the USC Shoah Foundation Institute for Visual History and Education. He founded the UK Holocaust Centre in Nottinghamshire, England and cofounded the Aegis Trust for the prevention of crimes against humanity and genocide. He was also the inaugural Chairman of the Holocaust Memorial Day Trust in the United Kingdom. He was the project director responsible for the creation of the Kigali Memorial Centre in Rwanda and provided consultation for the Cape Town Holocaust Centre, where he still serves as a trustee. He has taught extensively in Lithuania and has been a member of the International Task Force for Cooperation on Holocaust Education, Remembrance and Research since its inception in 1998. He is the author of *Making Memory: Creating Britain's First Holocaust Centre*; *Forgotten Places: The Holocaust and the Remnants of Destruction*; and *The Holocaust and the Christian World*.

Todd Presner

The Ethics of the Algorithm:
Close and Distant Listening to the Shoah Foundation Visual History Archive[1]

I.

I begin with two sets of computer-generated visualizations: The first is the Digital Monument to

the Jewish Community in the Netherlands[2] (Fig. 1). It's a Holocaust memorial that has no physical or

built counterpart; it exists only on the web. It is a digital image consisting of about 831,432 colored

pixels. Each little box of pixels represents a single person, and they vary in size according to the age of

the victim. The monument is a raster graphic, or bitmap, which is comprised of a rectangular grid of

pixels viewable in a web-browser on a computer monitor. The graphic represents the 100,000 or so

Dutch Jews who were killed by the Nazis. Clicking on an individual color box brings a viewer to a

webpage containing information about the victims, including their names, dates of birth and death (if

known), place of birth, and family members, including information about whether they survived the war

or not. The graphic organization of the monument is based on the alphabetical order of the place of

residence of the victims when they were deported.

INSERT FIGURE 1 HERE

The second computer-generated visualizations are based on the general indexing categories

developed by the Shoah Foundation to organize the genocide-related concepts and experiences described

in the 49,000 Jewish survivor testimonies in the Visual History Archive (Figs 2-5). These categories

form the most general or highest level in the 50,000-word thesaurus created by the Foundation,

including: Captivity, Culture, Daily Life, Discrimination, Feelings and Thoughts, Movement,

Organizations, Places, People, Politics and Economics, Religion and Philosophy. Under each of these

broad categories are hierarchical vocabularies to facilitate searching at a more precise level. For example, Captivity includes Camp Experiences and under that category, among other things, are camp adaptation methods, which are further broken down into camp barter, camp begging, camp betrayals, camp bribery, camp smuggling, and camp stealing.[3] Each visualization shows 200 different testimonies along the y-axis; time is shown along the x-axis, divided into one-minute segments (as per the indexing guidelines of the foundation). A red box means "yes" (that a given category was mentioned at that moment in the testimony) and a white box means "no" (that it was not). The length of the testimonies varies from under an hour to over 15 hours in length, although the vast majority is around two hours. Any given segment can contain multiple keywords or indexing terms, thus a "red box" may appear at the same time marker across multiple categories.

INSERT FIGURES 2, 3, 4, and 5 HERE

A few things become apparent from these visualizations: Certain general categories (and, hence, their specific topics) crop up significantly more frequently in the course of the testimonies: Places, Organizations, and Activities are marked-up (and presumably described) significantly more often than feelings, emotions, and attitudes. Almost all testimonies begin with the mention of place, which makes sense as a starting point for a survivor's life story. We can also track some general structural trends in the narrative arc of the testimonies: Discrimination tends to cluster in the first third of the testimonies, often keyed to life before the War, and Still and Moving Images tend to cluster in the final third, often keyed to present-day life, pictures of family, and messages to the future. Part of the reason for this is that the goal of the interview was to produce a story-like narrative that followed the chronology of the survivor's life, beginning with experiences in the pre-War period before moving to the War and the Holocaust, and, lastly, the post-War period, which concludes with a segment with family members and a future message.

While computer-generated data visualizations may illuminate certain commonalities, patterns, or structures through quantitative analyses, ethical questions immediately come to the foreground. These pixels signify Holocaust victims (in the first case) and the testimony of Holocaust survivors (in the second case). Even if we don't object to the "digitization," there is certainly some kind of "aestheticization" in the digital image: After all, the gridded organization of the Dutch digital monument conjures up a Mondrian painting. To turn Holocaust victims into quantifiable entries in a database and to visualize their lives as data points using colored pixels on a bitmap is, on the face of it, problematic: It presents victims as numbers and digital colors; it abstracts and reduces the human complexity of the victims' lives to quantized units and structured data. In a word, it appears to be de-humanizing and, even worse, might even partake in the same rationalized logic of modernity that Zygmunt Bauman identified in his seminal work, *Modernity and the Holocaust*, as the condition of possibility for genocide, namely the impulse to quantify, modularize, distantiate, technify, and bureaucratize the subjective individuality of human experience.[4] And, as we know from the work of Edwin Black, computational processing in the form of IBM's Hollerith punch card and card sorting technologies automated the process of identifying Jews from census data, registration forms, and other governmental records, and these computational technologies were deployed in Germany and occupied countries throughout the Reich to manage, accelerate, and automate the annihilation of the Jews.[5]

Might, then, the realm of the "digital" and the "computational"—precisely because it is, by definition, dependent on algorithmic calculations, information processing, and discrete representations of data in digitized formats (such as numbers, letters, icons, and pixels)—present some kind of *limit* when it comes to responsible and ethical representations of the Holocaust? In other words, are the "digital" and the "computational" at loggerheads with the ethical, and, if not, what might "ethical" modes of computation look like in terms of digital interfaces, databases, and data visualizations? To answer

these questions, I will be taking both a close and distant view of the Shoah Foundation Visual History Archive: The archive currently contains a total of 53,583 video testimonies (primarily of Jewish survivors of the Holocaust), in 39 languages, from 61 countries, amounting to over 100,000 hours of testimony.[6] I have watched only a tiny fraction of the video testimonies, but have spent considerable time examining the significance of the meta-data scaffolding and data management system, including the numerous patents for its information architecture, that allow users to find and watch testimonies. This chapter is not another contribution to the debates over the relevance or reliability of survivor testimony for historical writing or the possibilities of navigating the blurred lines between history and memory in eyewitness accounts of the Holocaust; instead, it is an analysis of the *computational genre* itself in historical representation. This includes databases, structured data and queries, and all the algorithmic means in which data is mined, analyzed, and visualized. It seeks to locate the ethical in digital and computational modalities of representation – hence the title: "the ethics of the algorithm."

II.

     Let's begin by considering what the question is to which these visualizations may be an answer. For the Dutch Digital Monument the question might be: How could one create a visual representation of the entire Dutch Jewish community destroyed by the Nazis? This is a problem of scale, scope, and complexity, and the interface provides one answer: An interactive bitmap of hundreds of thousands of pixels connected to a database documenting every person. Without the visual interface, the database is still searchable by way of the tables containing structured data (name, place of birth, date of birth, date of death, family members, and so forth); however, the totality cannot be seen without an interface that visualizes the scope and scale of the database. In fact, given the infinitely extensible nature of the digital, the physical limitations of built memorials (construction materials, available land, the legibility of

inscriptions, among other things) are no longer an issue. One need only recall that one of the reasons the first winning proposal for the Berlin Holocaust memorial by Christine Jackob-Marks was scrapped in 1995 was because there simply wasn't enough physical space to legibly inscribe the names of some 4.5 million identifiable victims.[7]

With the Shoah Foundation Visual History Archive, we are, of course, speaking about survivors, but the question of scale is equally daunting: With about 110,000 hours of survivor testimony, it would take a viewer 24 years to watch every testimony, assuming one watched 12 hours a day, 365 days of the year (and could understand 39 different languages). The scope of the archive—that is, its sheer scale measured in terms of hours of testimony—is not readily comprehensible to the human facilities of listening. Thus the database exists to organize, categorize, and search the content of the testimony based on a series of parameters. The visualizations above attempt to extrapolate the "whole" of the data in the database (which is—and this is critically important—a very different thing than the "whole" of the event called "the Holocaust").

But such problems of scale and scope are not new or unique to these digital archives. In fact, many of the same problems of representation and human comprehension came to the foreground in attempts to create a mode of history-writing and visual representation to capture "modernist events" of the 20th century, which, according to Hayden White, seemed to be different from events that historians from Herodotus to Arthur Schlesinger typically wrote about.[8] Against the backdrop of the new experiences of mass death in World War I, Walter Benjamin wrote his famous essay on Nikolai Leskov, "The Storyteller," about the social and historical conditions of *impossibility* of certain modes of representation. He argues, we have lost "the ability to exchange experiences" or tell stories precisely because the scale, scope, and depth of modernist events doesn't reflect or cannot be captured by the structures of storytelling in a realistic mode of narration.[9] The experiences of the war event and mass

death could no longer be observed, described, and communicated using the structures and meaning-making strategies reserved for historical realism, which was part and parcel of the tradition of storytelling with clear agents, a coherent plot, and narrative strategies characterized by the unities of time, place, and action that gave rise to the logic of a story. In other words, in modernism we see a breakdown of the homology between real events (*Geschichte*) and the narrative strategies (*Historie*) used to represent, capture, communicate, and render these events meaningful.

With the Holocaust and other catastrophic modernist events, we are faced with several challenges for historical representation: The first concerns the scale, scope, and complexity of the events themselves; the second concerns the lack of homology between the reality of "what happened" and the modalities of representation, whether through narrative, visual, or computational techniques; and the third is the problem of limited human faculties to observe, comprehend, read, listen to, and finally adjudicate the vastness of the different accounts of the events in question. This, I would suggest, is the "data sublime" that both the Digital Dutch Holocaust Monument and the Shoah Foundation Visual History Archive are confronting through computational modes of representation. And yet, as I will show here, the data sublime of the 100,000 hours of testimony provided by more than 50,000 survivors arranged in some 6 million tables of keywords in the database is structured by an information management system that is remarkably literalist: It accounts for the reality of "what happened" without attending to the heterogeneity of testimony as a representational form about a modernist event. Even as the Shoah Foundation's database assures factuality and facilitates access and preservation, it has the side effect of flattening differences between the testimonies and rendering listening one directional. As I will argue here, computation—as a genre of historical representation that includes data, databases, algorithmic processing, and information visualization—can be used against itself, so to speak, to not only deconstruct assumptions of objectivity and mathematical certainty but also give rise to a renewed

attention to the ethical. As such, far from simply replicating the structures of automation and information processing used in the planning and execution of the Holocaust, I will argue that computation also contains the possibility of an ethics of the algorithm.

III.

We need to begin with the specific genre of Holocaust video testimony because it is here that we can appreciate the conventional ethical imperatives structuring the creation, encounter with, and dissemination of survivor testimony. Much has been written on the history, significance, and media-specificity of audiovisual testimony, and I can give only the briefest overview of that history here, focusing primarily on how the recording, archiving, and dissemination of Holocaust video testimony have been defined through a Jewish ethics of individualized listening and personal obligation.

One of the earliest efforts to videotape Holocaust survivors began in 1979 when Dori Laub and Laurel Vlock created the Yale Video Archive for Holocaust Testimonies. It was later named the Fortunoff archive, and today has more than 4,400 testimonies and consists of some 10,000 hours of video. But oral history recording projects of survivors and other documentary efforts to capture eyewitness testimony began in the immediate aftermath of the War, of which one of the earliest and most extensive was David Boder's wire recorded audio narratives in Displaced Persons Camps in 1946.[10] Many of the early testimonies in Yad Vashem's collection were recorded before it was established in 1953 and, today, Yad Vashem has an archive of 36,000 testimonies, of which 11,000 are video testimonies (the remainder being oral and written testimonies). Started in 1994 and funded by the Spielberg Foundation, the Shoah Foundation Visual History Archive, with more than 53,000 video testimonies and over 100,000 hours of testimony, is the largest such archive in the world.

It is not coincidental, as Annette Wieviorka points out, that the impulse to record audiovisual testimonies in the late 70s and early 80s was spurred by televisual realities, ones that go back to the immediacy of first-person accounts by survivors at the Eichmann trial and go forward to the public impact of the television miniseries *Holocaust* in 1979 and, later, the global success of *Schindler's List* in the early 1990s.[11] But the visual power of televisual and cinematic modes of presenting and representing history is, as we know, not only alluring and captivating but also demands an interrogation of the reality effect produced by such ways of seeing and experiencing.

Geoffrey Hartman, one of the founders and project directors of the Yale Archive, has written extensively on the ethical dimensions of video testimony and distills the essential meaning of video testimony to be about the "duty to listen and to restore a dialogue."[12] For Hartman, video testimony offers what he calls an "optic" for viewers to immediately experience their non-experience of the Holocaust: That is to say, it mediates the geographic, temporal, experiential, and psychological remove that most of us have with the events of the Holocaust. This happens first through the relationship between the interviewers and the survivors and then through the generations of viewers who contribute to the creation of an "affective community" of witnesses to the witnesses.[13]

Testimony is a kind of performative embodiment of Martin Buber's "Ich-du" relationship, although one which is not symmetrical, in which the listener and the survivor, in Laub's words, enter into a "contract" through listening, bearing witness, and being heard.[14] Every survivor, writes Laub, has a need to be heard, to tell his or her story to a listener who is actively present for the other, listening to both silence and speech, trauma and survivorship.[15] "The unlistened-to story," as in Primo Levi's recurring nightmare in *Survival in Auschwitz*, is a trauma akin to re-experiencing the event itself.[16] In essence, video testimony—in so far as it instantiates a relationship of intersubjective relationality through the ich-du pact between the survivor and the listener—becomes a practice of ethics as a relation

of obligation and responsibility to the other. Bearing witness, then, is as much a testimony of the self as

it is a testimony for the other, and Hartmann will explicitly situate it within a framework derived from

Emmanuel Levinas. For Hartmann, testimony implies a "covenant" between the self and the other, one

which is in the face of an "infinite demand." "Ethical testimony," is for him, about being present: "Here

I am" – I am ready to listen, I am attentive, I am all ears, I am standing open, ready to be summoning to

this infinite demand, to this injunction to "hear" (the central prayer of Judaism, Shema Israel).[17]

It is the philosophy of Levinas, perhaps more than any other, that has informed much post-War

thinking about ethics as obligation and responsibility to the other. In survivor testimony, it is the

physical face of the other – the traumatized, wounded face of the survivor – which calls forth in its

alterity and infinity. The face of the survivor is a face of difference and rupture, but one which is

brought into a relationship of proximity, vulnerability, and closeness with the listener's own face. For

Levinas, ethics—defined as the relation to the other—is "the first philosophy," prior to any ontological

structure, origin, or attempt to ground being. It is not coincidental that Hartman will use the term

"optics" to highlight the media-specificity of video testimony, since Levinas will use the same term,

"optics," to define ethics as a relation of seeing and being for the other.[18] Indeed, Levinas' greatest

works, *Totality and Infinity* and *Otherwise than Being*, posit a philosophy of ethics as a relationship to

the other, such that the other is never reduced to the same, which he considers to be the violently

universalizing or totalizing impulse of ontology. Ontology is "a philosophy of power," violence, and

injustice because it subordinates and even negates the relationship of the subject to the other.[19] Ethics is

a relationship of vulnerability marked by responsibility to and difference from the other, perhaps most

notably in the fragile relationship between survivor and listener.

But what place, if any, does Levinas have in the realm of the computational, where relationships

are characterized by data placed within tables and fields in a database to be queried, displayed, and

visualized? And simultaneously, we may ask, what place, if any, does the computational have in the realm of listening to survivor testimonies? What would it mean for a computer to "watch," "hear," and "listen to" testimonies? What might be seen or heard beyond the faculties of human cognition and the optics of human perception? These are the questions to which we will now turn as we delve into the Visual History Archive.

IV.

While the media-specificity of the first generation of Holocaust testimony has been discussed at great length—ranging from Boder's wire recordings to cassette tape and audiovisual documentation— there is virtually no literature on the digitization of the Holocaust archive and its transformation into an information management system. With regard to the Shoah Foundation's VHA, this is particularly noteworthy because the very condition of possibility for watching any testimony is the information architecture standing behind the testimonies themselves. This information architecture consists of several components: First, there is the interface itself, which runs in a web-browser, allowing a user to type in keywords, names, and other search terms in order to listen to segments of testimony (Fig. 6); behind that, is a relational and structured query language database (SQL database, for short) in which content is organized into tables, records, and fields (Fig. 7); all of this data was inputted after the videos themselves were indexed with keywords and other associated information was manually entered (such as the information on the pre-interview questionnaires that each survivor had to fill out before the interview took place). But before this indexing could happen, standards and protocols—which were derived from the National Information Standards Organization's Z39.19 standard for the construction, format, and management of monolingual controlled vocabularies—provided the guidelines for what and how to index the content of the videos.[20] The standard governed the creation of a unique thesaurus to

achieve consistency in the description of the content through a controlled vocabulary and thereby facilitate its search and retrieval. A special piece of software called a Video Indexing Application or a Cataloguing Facility was developed to do this.[21] Beyond this, we have the hardware, such as the archive servers and storage servers, where the videos are stored in digital formats for streaming in a video player.

INSERT FIGURES 6 and 7 HERE

As such, for every survivor, we have two texts: first, the video testimony itself and, second, the data in the database about the testimony. With regard to the latter, every survivor is assigned a TestimonyID, and his or her testimony is broken into segments, which are generally one-minute in length. Each segment is assigned a segmentID that is correlated with a keywordID, which, in turn, corresponds further to a type label in the index hierarchy. The effect is to turn the narrative into data amenable to computational processing. Significantly, this process is exactly the opposite of what historians usually do, namely to create narratives from data by emploting source material, evidence, and established facts into a narrative.

The global architecture of the Shoah Foundation's Digital Library System was developed by Samuel Gustman, the Chief Technology Officer, and consists of the following elements: Data capture (starting with the transfer of the video tape to digital format and cataloguing) to the storage of data (both the videos themselves and the indexing server that knows where all the catalogue metadata is) and, finally, the interface to play, search for, and distribute data and its related content. In what follows, I will be focusing on that realm of information architecture between the user interface and the server storage – in other words, the metadata, the data structures, and the database. It is precisely here that we see a fundamental dissociation of the presentation of the content (that is, the testimonies and the interface to watch them) from the information architecture, database, and metadata scaffolding that lies behind the content. Such a dissociation is not unique to the VHA but bespeaks a common practice in

digital library systems and computation more generally, stretching back to Claude Shannon's theory of information as content neutral.[22] In the words of media theorist Alan Liu applying the principles of Friedrich Kittler, what we are witnessing is emblematic of "the discourse network 2000"[23]: A mode of organizing information characterized by the "separation of content from material instantiation ... [such that] the content management at the source and consumption management at the terminus [are] double-blind to each other."[24] In other words, the content of the testimonies knows nothing of the information architecture, and the information architecture knows nothing of the testimonies. In this sense, the database is simply an empty, neutral bucket to put content in, and the goal of the information system is to transmit this content as noiselessly as possible to a receiver or listener.

Between 1996 and 2002, ten separate patents were filed by inventor Samuel Gustman and the Survivors of the Shoah Visual History Foundation, the assignee, for the VHA information architecture. The inventions include the following: a Method and Apparatus for Cataloguing Multimedia Data; several patents for a Method and Apparatus for Management of Multimedia Assets; a Digital Library System; and, finally, a Method and Apparatus for Cataloguing Multimedia Data. Some of the patents— such as the "Digital Library System" and "Methods and Apparatus for Management of Multimedia Assets"—have been referenced by more than 70 other patents from companies such as Xerox (for developing a browser-based image storage and processing system) and Microsoft (for semi-automatic annotation of multimedia objects). In 2011, the Shoah Foundation granted an exclusive right to all ten of its patents to a company called Preservation Technologies, a company with a specialty in audio-visual preservation, media transfer, digital archiving, and media streaming.[25]

The first patent, "A Method and Apparatus for Cataloguing Multimedia Data," was filed in 1996 and established the method for indexing the testimonies and creating a search and retrieval system for their playback. I quote the summary of the invention: "The invention catalogues data such as

multimedia data. A catalogue is a collection of one or more catalogue elements. An index is used to access a catalogue. An element of a catalogue has one or more attributes. An attribute provides information that can be used to search for, answer questions about, and navigate through a catalogue ... Attribute elements and attributes are used to build an index that can be used to facilitate catalogue access."[26] This summary can be elucidated using a diagram from the patent itself (Fig.8): At the top are video segments, generally chunked into one-minute units; they contain narrative elements (sentences and phases) said by the survivor; these phrases have a number of different attributes: they mention particular people (and the particular information about the person is stored in the database); they contain particular keywords (which may already exist in the thesaurus, or may need to be added, hence, "proposed keywords"); and, most importantly, the keywords have a certain hierarchy in that they can be contained in more general "types." Altogether, the keywords and types form a catalogue consisting of an index of attributes connected to phrases uttered during segments of video. This is the metadata scaffolding or "metatext" that resides behind the video themselves. In the words of Johanna Drucker on the significance of such metadata structures: "Arguably, few other textual forms will have greater impact on the way we read, receive, search, access, use, and engage with the primary materials of humanities studies than the metadata structures that organize and present that knowledge in digital form."[27] This is certainly true of the VHA, whose knowledge model, as we will see, is fundamentally aimed at the transformation and dis-ambiguation of narrative into data that makes it amenable to computational processing and structured logic. It is a process that can be called "de-figuration" – precisely because it evacuates all traces of the figurative in its literalism.

INSERT FIGURE 8 HERE

      Within the index, there are three different kinds of relationships that can exist between any two (or more) indexing elements, and these relationships form the "pillar" of the index, according to

Gustman: Inheritance, whole/part, and associative relationships.[28] Inheritance relationships are characterized by "is_a" (for example, in the patent, he cites a "Ford Bronco" is a "car," where the specific keyword is "Ford Bronco" and the type is a "car").[29] The second relationship is whole/part (for example, cars and tires); and the third relationship is associative (such as "car" and "driver" – where neither "is" the other and they are not in a whole/part relationship).   As I mentioned earlier, these principles derive from the application of a specific standard (Z39.19) to consistently and unambiguously describe "content objects" (the survivor testimonies) in order to produce a monolingual controlled vocabulary (the thesaurus) to facilitate their search and retrieval.[30] The goal of the standard, as explained in its documentation, is to provide "guidelines for the selection, formulation, organization, and display of terms that together make up a controlled vocabulary" for the purposes of "knowledge management" and "knowledge organization."[31] The indexing terms are generally nouns and form subject headings, underneath of which one finds keywords in various relationships (inheritance or hierarchical, whole/part, and associative).

It is important to underscore that none of the testimonies in the Shoah Foundation VHA was automatically tagged with keywords; instead, every component of the cataloguing system—from the development of the indexing terms and the thesaurus to the database itself—was created by the staff working at the Foundation who listened to all the testimonies and indexed them according to the guidelines developed by the Foundation. This is because there are no transcripts of the testimonies themselves. In fact, the keyword indexing system—which consists of a thesaurus term (or terms) linked to a particular segment of video—is the *only* way to search the content of the testimonies. On average, testimonies have about 120 indexed terms associated with one minute segments (although many have more and some less), yielding about 6.2 million tables of data.

To develop the metadata, the Shoah Foundation employed about 50 indexers who worked for several years watching each and every video using a specially developed application (also patented) that allowed the human indexer to assign a keyword to a video segment. Keywords were assigned to the narrative content of the video from the thesaurus and, at the same time, new keywords could be proposed to describe experiences not already in the thesaurus.[32] For the first 5,000 testimonies, the segments were variable in length and could be determined by the indexer; however, this was quickly replaced by another system (used for the remaining 46,000+ testimonies), in which The Video Indexing Application would automatically "chunk up" the testimony into discrete, one-minute segments and prompt the indexer to assign a keyword. The "chunking" of the video was automated but the assignment of the keyword was determined by a human listener. Not every minute segment, however, has a keyword, something that may indicate the continuation of the previous keyword but may also mean, according to the Shoah Foundation staff, "the lack of indexable content."[33] Lack of indexable content can mean many things, ranging from an interviewer asking a question to a survivor repeating him or herself, a pause in the conversation to reflect or search for the right words, an emotional moment, noise, silence, or even content that the indexer doesn't want to draw attention to (such as racist sentiments against Hispanics, for example, in one testimony). In other words, indexable content is manifest content, in a declarative or imperative mode – in general, what is literally and objectively said. Altogether, the indexing system produces a kind of "normative story" (purged of certain contingencies and unwanted elements) in which – on the level of the data in the database – many of the testimonies, but certainly not all, become quite like each other.[34]

The result is a massive data ontology that has expelled the latent content, the performative, the figural, the subjunctive, the tone of questioning and doubt, the expressiveness of the face and the very acts of telling (and failing to tell) that mark the contingency of all communication. And while its aim is

objectivity, it is important to underscore that a human listener decided what to index and what not to index; a human listener decided what indexing term to use and what indexing term not to use; and a human listener decided if a given narrative segment could be described by a keyword or not. This is a fundamentally interpretative process. The result is the removal of the potentialities of the narrative in the application of the data ontology. In the end, it has the effect of turning the narrative into data. In this regard, it is exactly the opposite of the problem that Berel Lang bemoaned about the use of figurative language and aestheticization "adding to" the factual reality of the events[35]; here, we are speaking about "subtracting from" or "abstracting of" the narrative as told by the survivors. In other words, what goes missing in the "pursued objectivity"[36] of the database is narrativity itself: from the dialogical emplotment of the events in sentences, phrases, and words in response to the interviewer's questions, to the tone, rhythm, and cadence of the voice, to the physical gestures, emotive qualities, and even the face itself.

Of course, this is because databases are not narratives or people telling stories; instead, they are formed from data (such as keywords) arranged in relational tables which can be queried, sorted, and viewed in relation to tables of other data. The relationships are foremost paradigmatic or associative relations, to use Ferdinand de Saussure's terms, since they involve rules that govern the selection or substitutability of terms, rather than the syntagmatic, or combinatory elements, that give rise to narrative.[37] Database queries are, by definition, algorithms to select data according to a set of parameters. Whenever I enter a search string in the Shoah Foundation interface, I am performing a SQL query based on parameters that can be searched on.

"Indeterminate data," such as "non-indexable content," must be given either a null value or not represented at all. How would emotion, for example, need to be represented to allow database queries? While certain feelings, such as helplessness, fear, abandonment, and attitudes, are tagged in the database,

18

it would be challenging to mark-up emotion into a set of tables and parse it according to inheritance structures (sadness, happiness, fear, and so forth, all of which are different kinds of emotions), associative relationships (such as happiness linked to liberation, or tears to sadness and loss), and quantifiable degrees of intensity and expressiveness: weeping gently (1), crying (2), sobbing (3), bawling (4), inconsolable (5). While we can quickly unpack the absurdity (not to mention the insensitivity) of such a pursuit, there are precedents for quantified approaches to cataloguing trauma, including a method developed by David Boder following his analyses of the interviews he conducted with survivors in DP camps.[38] Needless to say, databases can only accommodate unambiguous enumeration, clear attributes, and definitive data values; *everything else is not in the database*. The point here is not to build a bigger, better, more totalizing database but that database as a genre always reaches its limits precisely at the limits of the data collected (or extracted, or indexed, or variously marked up) and the relationships that govern these data. We need narrative to interpret, understand, and make sense of data.

So that leaves us with a critical question: What do we need databases for? With regard to the Shoah foundation VHA, the database exists to provide meaningful access to the testimonies on a scale that is both tailored and comprehensible to a human viewer whose faculties of attention and knowledge (most likely) preclude 24 years of viewing and listening. In other words, a database and, hence, the very genre of computational representation exists, first of all, to manage scale. Secondly, as I show below, a relational database, by definition, functions by virtue of the relations or cross-connections between the data in the database. As such, the database can gave rise to infinitely many search queries and thereby allow innumerable combinations that identify larger thematic issues, reveal patterns and structures, and create new associations between experiences that may not otherwise be considered together. And, finally, computational analysis can provide insights and ethical perspectives that human listening cannot

precisely by the way in which it allows a kind of "distant listening" based on the whole of the archive rather than a selection of representative or even canonical testimonies.

The visualization below (Fig. 9) and the detail (Fig. 10) are examples of network relations based on just 100 testimonies, in which names are connected to keywords mentioned in the testimonies. The large circles (nodes) are survivors and all the lines (edges) that extend out from them are keywords used not only in their testimony but in the testimony of other survivors. The thicker the line, the higher the frequency of use; the larger the circle, the more keywords are associated with the person. Keywords at the center are more common (and this also moves the person to the center); keywords describing less common experiences gravitate toward the periphery. In this particular example, one survivor, Arie Leopold Haas, appears on the periphery with comparatively fewer lines connecting the keywords in his testimony to those of other survivors. Perhaps this is because the experiences he describes in his testimony—being an Italian Jew who was hidden, who converted to Christianity, who attended church— are ones that are less typical, at least when compared to the experiences of others in the archive.

INSERT FIGURES 9 and 10 HERE

The graphic was generated by a data visualization program called Gephi, which algorithmically determines "communities" based on topics mentioned. From the 100 testimonies, it detected 16 different communities. In some cases these "communities" appear to be based on nationality (Russian, Ukrainian in blue), but in other cases they seem to be based on places mentioned or shared experiences. Visualizations like these might provide new starting points for delving into the more than 6 million records in the database and seeing connections that a human eye could not possibly detect or track. In this particular case, we might be able to identify "outlier" experiences or non-canonical stories that help us reassess certain assumptions or provide a more differentiated set of perspectives.

20

The querying of the database, particularly through facetted searching that allows a user to apply multiple filters, can reveal sites of overlap and linkages between experiences. I would contend that the possibility of infinite "query-ability" and visualization of the relations in a database is, in fact, a critical part of its ethical dimension. Consider, for a moment, the alternative: 52,000 atomized testimonies searchable by unique identifiers such as name or record ID but without the ability to traverse orthogonally through the tables. The more "thick" the possible relations and intersections are between tables, the more possibilities of interconnection, the more ethical the database. In other words, the potential to facilitate an ever deeper relationality among the data in a database is one of the conditions of possibility for an ethics of the algorithm.

As Lev Manovich asks in *The Language of New Media*: "How can our new abilities to store vast amounts of data, to automatically classify, index, link, search, and instantly retrieve it, lead to new kinds of narratives?"[39] As an explicit uptake of Manovich's question of how classification, indexing, search, interlinking, and retrieval can lead to new narratives, the Shoah Foundation VHA allows users to create their own project narratives from the search results, essentially, building remixed and hybridized narratives from any number of constitutive narrative segments. In this regard, we see a symbiosis between narrative and database, such that the paradigmatic structure of the database contributes to the syntagmatic possibilities of combination at the heart of narrative.[40] And I would point out that this is not fundamentally different from what historians already do: make selections from the trove of archival sources in order to combine elements together to form a narrative. The database performs this selection and combinatory process in every query and, hence, literalizes an instance of historical emplotment. The metadata database of the Shoah Foundation VHA thus represents a kind of "paratext" insofar as it can be reordered, disassembled, and reassembled according to the constraints and possibilities of computational logic.[41] The visualizations of the Shoah Foundation VHA are representations of the

21

paratext, the metadata scaffolding that runs behind the testimonies and, with every query to the database, represents an algorithmically transformed text.

In a computational mode of representation, it is common to "toggle" between the singular and the global, the individual experiences of particular eyewitnesses and all the experiences as recounted by the survivors, which in this case is the summation of all the data in the VHA. The latter does not represent the reality of "the Holocaust" (as a complete or total event) but rather the totality of the archive, and therefore, can only present structures, patterns, and globally-oriented visualizations of data. But, again, this is not very different from what historians do, insofar as they emplot events at various levels of "zoom" in order to convey different kinds of meaning. In other words, we "toggle" back-and-forth between macro-level accounts of the totality of the event (zoomed out) and micro-level accounts of individual experiences (zoomed in), which are, by their very nature, defined by specific experiences, perspectives, spectatorship, language, and so forth. Saul Friedländer's "globally oriented inquiry" into the history of the Holocaust not only examines the encompassing "ideological-cultural factors" and mythologies of the Nazi regime while recounting the totality of events, actions, and numbers to convey the overwhelming efficiency and scope of the destruction, but he also calls upon the individual voices and personal chronicles of diary and letter writers "to illuminate parts of the landscape ... like lightning flashes," and thereby "pierce the (mostly involuntary) smugness of scholarly detachment and 'objectivity.'"[42] In essence, there are certain parallels between the compositional practices of historians and those of computation.

But the computational mode also allows another kind of reading and listening practice, which is quite different from what individual readers and listeners tend to do with memoirs and video testimony. The computational allows us to perform what literary scholar Franco Moretti has termed "distant reading" – a practice that moves away from the close, hermeneutical reading of texts in favor of an

algorithmic approach that presents over-arching structures and patterns.[43] For Moretti, distance is "a condition of knowledge" because it allows a scholar to "focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems."[44] In other words, the perspective of distance allows us to see different things than the perspective of closeness (characterized by close, attentive, detailed reading). By confronting scale, "distant reading" – or, in our case, "distant listening" – reveals structures, patterns, and trends that are not discernable when the focus remains on just a handful of close readings of individual texts. And the stakes are much higher than just revealing structures: Distant listening facilitates whole corpus analysis and, potentially, the democratization of knowledge. Instead of privileging "human listening" (in which we necessarily have to limit ourselves to a tiny canon of works, probably a few hundred), distant listening is performed by a computer and can easily "listen to" thousands, if not millions, of works.[45]

So what might this kind of large-scale, full corpus, "distant listening" mean for the Shoah Foundation VHA? For one thing, it brings into stark relief the tiny fraction of memoirs and testimonies of survivors that are actually read, listened to, and taught. We tend to privilege a very small canon of witnesses, whose stories stand in—rightfully or not—for the stories of almost everyone else. We know Elie Wiesel, Anne Frank, and Primo Levi, but what about Anna Neuman-Goldman, Daniel Geslewitz, and Arie Leopold Haas? Distance listening can facilitate a democratization of witnessing, since it has a leveling effect in that all testimonies are granted equal importance and weight, such that no one testimony takes priority or assumes canonicity. To extrapolate structures, trends, patterns, frequencies, and correlations from the entire database will produce claims that are grounded in the experiences of exponentially more people than tend to enter into conventional historical accounts grounded in a significantly smaller sample size. I would posit that "distant reading" – or, in our case, "distant listening" – is ethical precisely because it takes into account the metadata (specifically, the keywords

linked to testimony segments) of *every* survivor who had his or her story recorded in the VHA. I am not arguing that the computer should replace the human listener and the intersubjective experience at the heart of testimony, but I am saying that computational or algorithmic analysis can be ethical precisely because it takes into account the fullness of the archive insofar as all the indexed data related to the narrative of every survivor is part of the analysis.

V.

Let me now conclude with some speculative questions with the aim of re-imagining the database of the Visual History Archive in a modernist register, considering data as figuration, and implementing a practice of humanistic computing characterized by an ethics of the algorithm. We might begin by asking: How would a Levinasian database operate? What would it mean to bring the realm of the ethical as defined by Levinas as "a first philosophy" to the backend information architecture (the database, the data structures, and the metadata standards)? In other words, I want to imagine an information architecture that is fundamentally connected to the content, and not just any content, but the specific narratives of Holocaust survivors and the listener's responsibility to that testimony through an ethics of obligation. This means the database, like all of the information architecture, is not a neutral container to store or put content into, and the goal of the information system is not simply to noiselessly and seamlessly transmit messages to receivers. Instead, the database must be conceived through the same ethical optic as watching the testimonies and, therefore, fundamentally connect testimony to the information architecture, the data ontologies, the data structures, the indexing systems, and the viewers who are engaged in a participatory mode of listening.

For Levinas, ontology is the problem because it is a philosophy rooted in being and the attempt to ground meaning though identity, objectivity, and even certain kinds of linguistic structures, namely

what he calls the literalism of "the said". This is essentially the same literalism of the data in the VHA database: dis-ambiguated, manifest content, objectively said. Instead of ontology, Levinas poses a philosophy of relationality, in which the self is connected to the other through bonds of responsibility, vulnerability, proximity, and even rupture. Here, the linguistic operation is the act of "saying" or, more radically, the possibility of "unsaying the said." For Levinas, the challenge is to undo the paradigms of wholeness and totality, which are implicated in philosophies that are grounded in ontology and identity, in favor of an intersubjective philosophy of relationality, alterity, fragility, and uncertainty.

I wonder how we might rethink the very genre of the database as a representational form vis-à-vis the specific experiences of bearing witness, testifying, surviving, and narrating. How might the database reflect the fragility of life, the uncertainty, ambiguity, and figuration of narrative? How might it preserve (rather than undo) the "hauntedness" that informs so much of the testimony? In other words, how might a database be open to the haunt of the past, the trace of the unknown, the spectral quality of the indeterminate, and, simultaneously, be oriented to the uncertainty of the future, the possibility of the unknown, what Jacques Derrida calls "the spectral messianicity" at the heart of the archive? Such a notion of the archive specifically disavows the finality of interpretation, relishes in ambiguity, and constantly situates and resituates knowledge through varying perspectives, indeterminacy, and differential ontologies.

As such, we might imagine how a fluid data ontology might work, by allowing, multiple thesauruses that recognize a range of knowledge, standards, and listening practices. For example, what if verbs that connected action and agent, experience and context were given more weight than hierarchies of nouns primarily in associative relationships? What if a more participatory architecture allowed for other listeners to create tags that could unsay the said, or in other words, undo – or, at least, supplement – the definitive indexing categories and keywords associated with the segmented

testimonies? Or more radically, what if the user interface was generated by network graphs or visualizations, such that the listener did not merely type terms into an empty search box but rather could browse the entirety of the archive in a dynamic way based on, perhaps, communities of experience, narrative structure, or even silences, gaps, and so-called non-indexical content?[46]

Such structures of saying and unsaying the database would constantly re-interpret and re-inscribe the survivors' stories in ways that not only place the listener into an active relationship of responsibility but unleash a potentiality of meaning in every act of "saying" and "browsing." Narratives would be heard in their polyphony, with some listeners hearing some things and others hearing quite different things. Through these acts of saying and unsaying, which are, according to Levinas, marked by an "allegiance" and "exposedness" to the other, the responsibility to the other might become part of the ethics of the information architecture itself. We might call it: Otherwise than the Database, or Beyond Essence.[47] In a sense, we would never be done listening, watching, and processing the testimonies because there is always more – a surplus of meaning – that is never finally captured in data or databases. And this is what the information architecture would facilitate**:** A hermeneutic of uncertainty, a modernist – or perhaps, Talmudic – writing and rewriting of the metadata through an ethics of obligation and ever thicker relationships between data and narrative, as a kind of Jewish ethics of responsibility, telling and retelling, interpreting and reinterpreting, listening and being present.

There is no reason, then, why the realm of information architecture, data structures, and databases should be considered apart from the realm of ethics and the subjective, contingent, meaning-making, interpretative practices at the heart of the humanities. What's at stake when the ethical philosophies of the humanistic tradition do not fundamentally inform the digitization of the archive, when data and data management "conform to a model of mathesis that assumes objective, totalizing, mechanistic, instrumental capability"?[48] This is the risk of completely separating content from

26

information architecture, of privileging dis-ambiguated data ontologies over probabilistic knowledge, potentialities of figuration, and interpretative heterogeneity. But computational representation does not have to be this way if it is guided by an ethics of the algorithm.

The challenge resides in imagining a kind of humanistic computing that not only deconstructs the assumptions of mathesis operating behind and imposed on top of the cultural record but that also propels an approach to information, the database, and the digital archive in general that does not seek to overcome or suppress the ambiguous, the unfinished, the differential, the multiple, and the spectral. Through ever thicker relationships between data and narrative, saying and unsaying, visualizing and listening, it is possible for computation to facilitate an ethics of listening that moves between the whole of the database and the individual testimony, transforming both in a never-ending, dynamic process of listening that gives rise to new narratives. As such, the ethics of the algorithm might begin by performing close and distant listening to the more than 52,000 testimonies in the Shoah Foundation archive: listening to them one-by one *and* by listening to them all-at-once.

---

[1] The research presented here owes a significant debt of gratitude to the USC Shoah Foundation, particularly Stephen Smith and Samuel Gustman, for providing our UCLA digital humanities research team with a copy of the entire database of metadata related to the testimonies in the Visual History Archive. All of the analyses presented in this paper were performed at UCLA under the direction of Todd Presner. This paper represents his views and the team's research and does not necessarily represent the views of the USC Shoah Foundation. Members of the UCLA digital humanities were critical for carrying out the research presented here, and I would like to thank: Rachel Deblinger, David Shepard, Monit Tyagi, Yoh Kawano, and Zoe Borovsky. In addition, I would like to thank Hayden White, David Myers, Alan Liu, Eric Rentschler, Judith Ryan, Kristine Stiles, Bill Seaman, and other colleagues at UCLA, Harvard, and Duke University for their valuable feedback on earlier versions of this paper.

[2] http://www.joodsmonument.nl/?lang=en (accessed June, 1, 2014).

[3] USC Shoah Foundation Institute Thesaurus (June 2010).

[4] Zygmunt Bauman, *Modernity and the Holocaust* (Ithaca: Cornell UP, 2000).

[5] Edwin Black, *IBM and the Holocaust* (New York: Dialog Press, 2001).

[6] In 2014, the Shoah Foundation added nearly 2,000 testimonies from San Francisco's Jewish Family and Children's Services (JFCS). Prior to that, there were 51,696 testimonies recorded and indexed by the Foundation. The analysis presented here does not include the recent addition of 2,000 testimonies from JFCS, which have not (as of this writing) been indexed.

[7] See the discussion by James Young, *At Memory's Edge: After-Images of the Holocaust in Contemporary Art and Architecture* (New Haven: Yale University Press, 2000).

[8] Hayden White, "The Modernist Event," in *Figural Realism: Studies in the Mimesis Effect* (Baltimore: Johns Hopkins University Press, 1999), 66-86.

[9] Walter Benjamin, "The Storyteller: Reflections of the Works of Nikolai Leskov," in *Illuminations,* ed. Hannah Arendt, trans. Harry Zohn (New York: Schocken Books, 1968), 83.

[10] Boder interviewed 130 survivors, in nine languages, in DP camps during the summer of 1946. The archive is available online at: http://voices.iit.edu/david_boder.   For an account of Boder's work, see: Alan Rosen, *The Wonder of Their Voices: The 1946 Holocaust Interviews of David Boder* (Oxford: Oxford UP, 2010). Rosen points out that Boder didn't use the term "testimony" but rather called his DP interviews "narratives, reports, personal histories and documents, stories, and even 'tales.'" (12). Moreover, Boder, reacting to the flood of newsreel footage of the liberation of concentration camps, considered the visual to be silent, in need of narrative and voices. In this regard, his wire recordings were aimed to provide "first-hand auditory material ... that sought to augment—and, perhaps, to challenge—the camera's work" (130).

[11] Annette Wieviorka, *The Era of the Witness*, trans. Jared Stark (Ithaca: Cornell UP, 2006).

[12] Geoffrey Hartman, *The Longest Shadow: In the Aftermath of the Holocaust* (Bloomington: Indiana UP, 1996), 133.

[13] Geoffrey Hartman, "The Ethics of Witness: An Interview with Geoffrey Hartman," by Ian Balfour and Rebecca Comay, in: *Lost in the Archives*, ed. Rebecca Comay (Toronto: Alphabet City Media, 2002), 492-509. Here, 495 and 501.

[14] Shoshana Felman and Dori Laub, *Testimony: Crisis in Witnessing in Literature, Psychoanalysis, and History* (New York: Routledge, 1992), 72 and 85.

[15] *Ibid*, 58.

[16] *Ibid*, 67. Primo Levi, *Survival in Auschwitz*, trans. Stuart Wolf (New York: Collier Books Macmillan, 1993), 60.

[17] Hartmann, "The Ethics of Witness," 492. This call of humble readiness ("Here I am") is uttered not just by Abraham to God (Genesis 22:1), but also by God, as a warning, in Isaiah ("I was ready to be sought by those who did not ask for me... I said Here I am, Here I am, ...but when I called, no one answered, when I spoke they did not listen." (Isaiah 65:1-2, 66:4). "Here I am" is also said by Jacob (Genesis 46:2), Moses (Exodus 3:5), and Samuel (I Samuel 3:4).

[18] Emmanuel Levinas, *Totality and Infinity*, trans. Alphonso Lingis (Pittsburgh: Duquesne UP, 1969), 29.

[19] Ibid, 46.

[20] Cf. "Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies" (ANSI/NISO Z39.19-2005) (Bethesda, Maryland: National Information Standards Organization, 2005).

[21] Patent 5,832,495.

[22] Claude Shannon's foundational ideas of information theory were articulated in his article "A Mathematical Theory of Communication" (1948), in which the goal of a communication system was to transmit a message over a channel to a receiver with the minimal amount of noise possible. Available online: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

[23] The reference is to Kittler's *Discourse Networks 1800/1900*.

[24] Alan Liu, *Local Transcendence: Essays on Postmodern Historicism and the Database* (Chicago: U of Chicago Press, 2008), 216.

[25] http://themediapreserve.com/index.html. Preservation Technologies is owned by another company called Digitude Innovations, a company that acquires patent technologies and licenses them to other companies. Interestingly, since December of 2011, Digitude's subsidiary, Preservation Technologies, has instigated a litigation campaign against more than a dozen companies involved in digital media streaming based on patent-infringement. Preservation Technologies has filed lawsuits against companies including Hulu, Netflix, Vimeo, ESPN, CBS, Sony, Fox, Dish, and the New York Times Digital. In each case, the patents of Gustman, for which it has an exclusive license, are cited as the basis of patent-infringement.

[26] Patent 5,832,495, "Summary of Invention," 3.

[27] Johanna Drucker, *SpecLab* (Chicago: U of Chicago Press, 2010), 9.

[28] Conversation with author, February 6, 2012.

[29] Patent 5,832,495, 10.

[30] While the testimonies are in thirty-nine different languages, the database is only in English. Because all content is tagged with English metadata, the database also represents a profound translation effect that loses the linguistic and cultural specificity of the terms used by the survivors.

[31] "Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies," 2.

[32] The thesaurus and database structure were designed to be modular in that they can be applied to testimonies about other genocides. The foundation has already begun the indexing of testimonies from the Armenian genocide and the Rwandan genocide, using a revised version of the thesaurus to include different place names and experiences (such as "alcoholism," a key term in Rwanda) as well as remove certain terms that do not apply to these genocides. Interestingly, with the new testimonies from San Francisco's Jewish Family and Children's Services, there are also possibilities of using indexing terms from other genocides (such as "depression" and "reconciliation," both terms that appear frequently in the Rwanda testimonies) to tag Holocaust testimonies. As my work on this project continues, I hope to explore questions of comparative genocide studies through these comparative information architectures.

[33] Krispin Brooks, conversation with the author (February 6, 2012).

[34] Our digital humanities team at UCLA is analyzing all of the data in the database in order to develop "normativity indices" for each testimonyID based on patterns in the data. The "normativity index" is based strictly on the data in the database and is meant to provide a baseline to determine the effectiveness of the tagging process and also identify which testimonies (from the standpoint of the data) are "outliers." To do this, we looked at both narrative form and content in order to measure the likelihood that certain indexing categories would be assigned at a given percentage point in a testimony (form) as well as the likelihood that certain experiences or themes would be discussed (content). These indices are intended to help make the database better by providing flags for the VHA to identify testimonies that have not been indexed very well as well as identify the widest possible range of experiences described and narrative strategies used by survivors.

[35] Berel Lang, *Act and Idea in the Nazi Genocide* (Ithaca: Cornell UP, 1992).

[36] "Indexing Guidelines," Shoah Foundation, Visual History Archive, 5.

[37] Ferdinand de Saussure, *Course in General Linguistics*. Also, discussion by Lev Manovich, *The Language of New Media* (Cambridge: MIT Press, 2001), 230-31; and N. Katherine Hayles, "Narrative and Database: Natural Symbionts," in: *PMLA* 122.5 (October 2007): 1603-08, here, 1606.

[38] Boder developed an extensive method to quantify trauma based on a "traumatic inventory" of sufferings recounted by survivors: "The Impact of Catastrophe: I. Assessment and Evaluation," *The Journal of Psychology* 38 (1954): 3-50.

[39] Lev Manovich, *The Language of New Media*, 237.

[40] Also, see Hayles, "Narrative and Database: Natural Symbionts."

[41] See the discussion by Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism* (Urbana: U of Illinois Press, 2011), 85.

[42] Saul Friedländer, *The Years of Extermination: Nazi Germany and the Jews, 1939-1945* (New York: HarperCollins, 2007), xvi, xxv-xxvi.

[43] Moretti first explored this notion of "distant reading" in an article called "Conjectures on World Literature," in: *New Left Review* (Jan.-Feb. 2000): 54-68. A more thorough and wide-ranging analysis of the practice is found in his: *Distant Reading* (London: Verso Press, 2013).

[44] Ibid., 57.

[45] Analogously, Moretti argues that the vast majority of books are simply never read by human beings, amounting to what he calls the great "slaughterhouse of literature." In his estimation, 99.5% of the novels published in 19th century Britain are never read or taught; instead, literary scholars are fixated on a tiny canon of works that are radically un-representative of the massive number of books, authors, publishing houses, and markets of that period, and that this fixation on the canon greatly skews our understanding of its cultural texture and social history. Moretti, "The Slaughterhouse of Literature," in: *Modern Language Quarterly* 61:1 (March 2000): 207-227.

[46] Our UCLA digital humanities team is currently developing a series of alternative front-end interfaces based on network graphs that would allow users of the VHA to "browse" the entirety of the database from a multiplicity of perspectives and entry points.

[47] Levinas, *Otherwise than Being, or Beyond Essence*, trans. Alphonso Lingis (Pittsburgh: Duquesne UP, 2006), 49-50.

[48] Drucker, *SpecLab*, 17.