

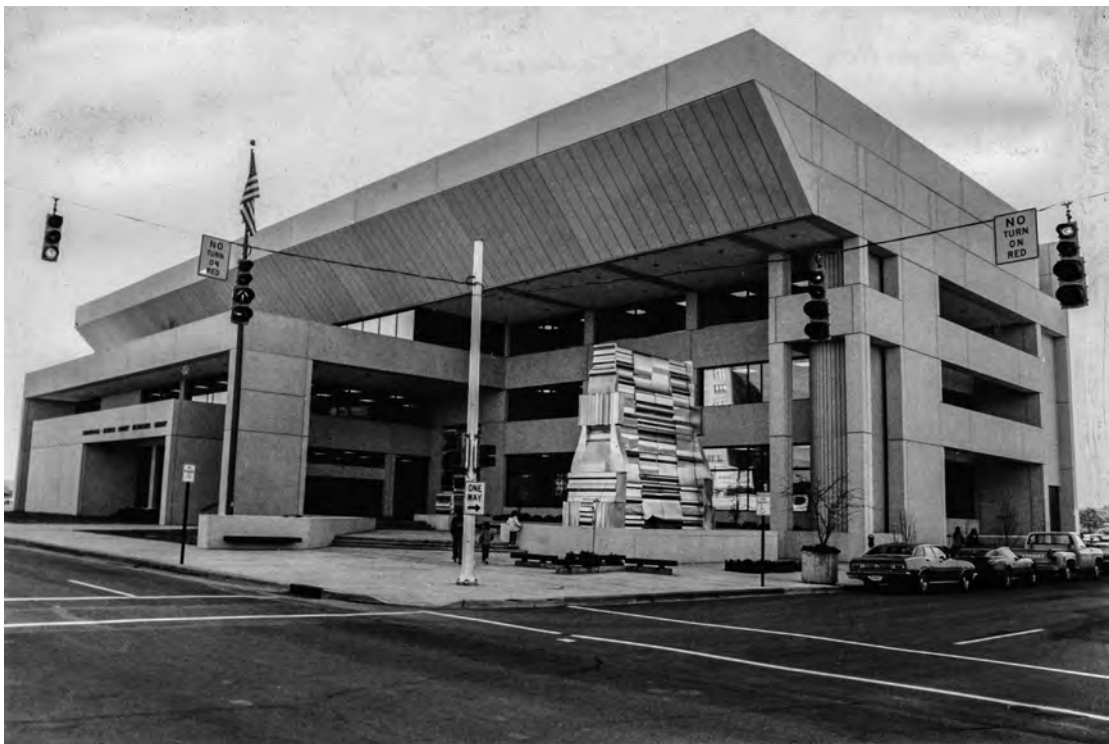
COLLECTING INFRASTRUCTURES

```
{ "DATAPROVIDER": "SMITHSONIAN INSTITUTION LIBRARIES", " "ADMIN":  
  { "VALIDATION_MESSAGE": "'RIGHTS' IS A REQUIRED PROPERTY, "  
"VALID_AFTER_ENRICH": FALSE}, "@ID": "HTTP://DP.LA/API/  
ITEMS/AF994D5BDBE589A9367E79980BFD7470, " "_REV":  
"8-138F4B0EE3C24865699469D2A900AAD9, " "OBJECT": "HTTP://LIBRARY.  
SI.EDU/SITES/DEFAULT/FILES/STYLES/BOOK_COVER_SMALL/PUBLIC/  
BOOKS/COVERS/CATALOGUEOFGAMES00MILT_COVER.JPG?ITOK=2IK8LJP8, "  
"AGGREGATEDCHO": "#SOURCERESOURCE, " "PROVIDER": { "@ID": "HTTP://  
DP.LA/API/CONTRIBUTOR/SMITHSONIAN, " "NAME": "SMITHSONIAN  
INSTITUTION"}, "INGESTDATE": "2014-07-31T11:14:53.630859, "  
"ID": "AF994D5BDBE589A9367E79980BFD7470, " "INGESTIONSEQUENCE":  
14, "ISSHOWNAT": "HTTP://COLLECTIONS.SI.EDU/SEARCH/RESULTS.  
HTM?Q=RECORD_ID%3ASIRIS_SIL_986353&REPO=DPLA, " "SOURCERESOURCE":  
{ "DESCRIPTION": ["TRADE LITERATURE, " "NEW YORK AGENCY, WILSON  
BROS. TOY CO., 119 CHAMBERS STREET., " "LITH. OF MILTON BRADLEY  
CO., SPRINGFIELD, MASS.\ "—P. [4] OF COVER, " "GAMES AND  
PUZZLES—SECTIONAL PICTURES AND MAPS—TOYS AND BLOCKS—NOVELTIES"],  
"LANGUAGE": [{ "NAME": "ENGLISH, " "ISO639_3": "ENG"}], "FORMAT":  
["55 P.: ILL., " "24 CM"], "@ID": "HTTP://DP.LA/API/ITEMS/AF  
994D5BDBE589A9367E79980BFD7470#SOURCERESOURCE, " "TEMPORAL":  
[{ "BEGIN": "1889, " "END": "1889, " "DISPLAYDATE": "1889"},  
{ "BEGIN": "1889, " "END": "1889, " "DISPLAYDATE": "1889–90 I.E.  
1889"}], "TITLE": "CATALOGUE OF GAMES, SECTIONAL PICTURES,  
TOYS, PUZZLES, BLOCKS AND NOVELTIES / MADE BY MILTON BRADLEY  
COMPANY, " "COLLECTION": [{ "@ID": "HTTP://DP.LA/API/  
COLLECTIONS/9463ABD671D67157E760344619BFBB9C, " "ID":  
"9463ABD671D67157E760344619BFBB9C, " ...
```

Source: Digital Public Library of America (excerpt of record)¹

A COMPARATIVE SETTING

In November of 2012, at the top of the Chattanooga Public Library—a concrete monument to an older era of collecting—I first caught a comparative view of data. I was in Chattanooga, Tennessee, for an “Appfest,” the first of its kind hosted in support of the Digital Public Library of America (DPLA), introduced at the beginning of this book.² The DPLA was initiated only two years earlier, on October 1, 2010, by a coalition of computer scientists, lawyers, librarians, and philanthropists from across the United States who



3.1

The brutalist era building where the first DPLA Appfest was held. Courtesy of the Chattanooga Public Library.

convened with the goal of gathering widespread collections data into an integrated index of cultural history “dedicated to the public good.”³ Since then, the DPLA has enlisted a broad array of institutional collections and collaborators: private and public universities, national institutions like the Smithsonian, and elite collections like the Getty. It has also developed tools for public access and succeeded in attracting a broad community.

I spent two days in Chattanooga, working with early participants in this community. We sorted through data compiled by the DPLA on books, newspapers, maps, photographs, and other objects of cultural import in a hanger-scaled space that was as grand as the nascent aspirations of the DPLA yet also windowless, with little connection to Chattanooga and what it might have meant to converge in this place. Our work was intended to help liberate data from their origins so as to make them accessible anywhere. Instead, the experience nudged me to question the goal of separating data from their originating institutions of collecting. My encounter with the DPLA—an experience that stretched over the next few years—helped me to articulate the third principle that structures this book: *data are assembled from heterogeneous sources, each with their own local conditions*.

The previous chapter focused on one data setting, the Arnold Arboretum. And even there, differences in data formats and uses were evident. But what happens when data from different institutions are brought together? How might we jointly hold or reconcile their incongruous place attachments? When data from multiple settings are juxtaposed, as in the DPLA, there is an inevitable clash between discordant originating data cultures.

The DPLA is an example of a data infrastructure: a meta collection, which agglomerates digital resources from distributed sites of production.⁴ Data infrastructures are often understood in terms that are seemingly technical. *Ingestion* grapples with how to collect and store heterogeneous data sets.⁵ *Interoperability* considers how to make those data speak to one another.⁶ *Enrichment* contends with how to add more information to ingested records.⁷ And finally, *interface* has to do with how to make those records accessible and actionable.⁸ But data infrastructures also awaken an ancient cultural ambition, as old as stories of Babel, to sever knowledge from its origins.

Using the DPLA as an illustrative case and harnessing a variety of techniques for local reading, I seek to uncover the culturally rooted place attachments that persist in data infrastructures. Reading DPLA data up close, with a focus on their classifications, schemata, constraints, errors, absences, and rituals, can reveal telling traces of their origins. Meanwhile, visualizing the encompassing data structures nested in the DPLA can offer us a glimpse of what is lost in practices of normalization: a process by which data are made to conform to an expected range of categories and values. For the DPLA, that format is its “MAP,” an internally defined metadata application profile.⁹

Despite the DPLA’s relatively low profile, the example demonstrates better than any other I know why data differ from one another. The data of the DPLA were created



3.2

The Library Observatory evolved to incorporate new contributing collections to the DPLA. Image by the author, Matthew Battles, and Jessica Yurkofsky. See <http://www.libraryobservatory.org>.

as part of entirely different knowledge systems, some of which have evolved over decades at long-lived institutions like universities, museums, and public archives. Others were invented in seclusion in order to serve the owners of small private collections. Acknowledging these conditions means learning to engage data infrastructures not as large, homogeneous sources of information but rather as sites of controversy where varied conceptions of data come into conflict.¹⁰ If we are to develop critical perspectives on data, I believe that the DPLA can help us learn to see both the forest and the trees.

TAKING DATA INFRASTRUCTURES APART

The Appfest, as its name suggests, was a minimarathon for designing software applications that would extend the DPLA, including rapid project pitches, informal tutorials, and plenty of head-down coding time. Such gatherings, more often called “hackathons,” are meant to both kick-start new computing projects and foment community growth around a set of technical problems. They predate contemporary data infrastructures. In fact, the term *hack* comes from an earlier culture of computing, which Sherry Turkle defines in terms of its creative and sometimes subversive use of bricolage—an approach to making software that she compares to a conversation.¹¹ Today, hackathons still espouse this bottom-up, conversational approach to computing, but they are often focused on data and how to wrangle them for productive means. They are meant to redress, without overtly acknowledging, a pervasive problem: data infrastructures are frequently brought together with little concrete sense of how they should be used.

In Chattanooga, the Appfest began, as many do, with an introduction to the data. At the time, the DPLA had a much smaller trove, drawn from only three contributing institutions: the Digital Library of Georgia, Minnesota Digital Library, and South Carolina Digital Library. We were encouraged to produce projects that might enrich those initial offerings for a broad public audience. In considering the task from my burgeoning local perspective, I was struck by the fact that it was so hard to see the collection in terms of its parts. Could a panoramic view, of the sort introduced in chapter 2, enrich our understanding of the DPLA’s current holdings? I was curious to learn who had contributed what and when. Since this view wasn’t possible at the time, I went about trying to construct it, together with several other Appfest participants.¹²

Our project, later dubbed the “Library Observatory” by one of my collaborators, the writer and polymath Matthew Battles, is structured as a tree map: a hierarchical form of visualization composed of nested boxes for data that conceptually mimics the branching structure of a tree. Each box contains or represents all the entries at one level of the collection.

The largest boxes in the visualization, representing the contributing institutions of the DPLA, form the metaphoric base of the tree. In our initial design, the box containing entries from the Georgia library takes up half the overall image because it contributed that portion of the total DPLA holdings at the time. Minnesota and South Carolina are

about equal. Inside each of these contributor boxes are smaller ones, labeled according to the next level of organization for each collection.

This is where their structures start to diverge noticeably. Georgia's subcompartments include "GA Obituaries," "1730-1842," "Civil Rights Digital Library," "Georgia Government Publications," and a large subcollection confusingly labeled the same way as the higher-level container, "Digital Library of Georgia." The South Carolina and Minnesota collections have their own subcollections, many of which also seem bemusing or simply unhelpful from an outsider's perspective; they are often named for local cities, institutions, or people that have little significance for a national or international audience.

Our crude map of these conditions was a mess, reflecting the discordant structure of the data contained therein. Nevertheless, it was a useful first step toward the goal of taking apart the DPLA in order to reveal its invisible localities. Such a map can provide a useful illustration of how a data infrastructure is composed. But how might a tree map be made into a useful navigation tool for visitors? When a search query is formed by a visitor, would it help them to know where, among the contributing collections, the results are returning from?

The Library Observatory project, which called attention to more problems than it solved, did not win any prizes at the Appfest.¹³ Although antithetical to my agenda to reveal the heterogeneity of the DPLA, the winner that weekend, aptly named "Dedupe," is also helpful for understanding data infrastructures and the challenges they face. Dedupe treats the entire DPLA as if it were a one-dimensional data array, plagued by redundancies that need to be nulled.

Ironically, nowhere are the differences within a data infrastructure more apparent than in efforts to normalize them. The term *dedupe* is shorthand for an automated process that will rid the DPLA of duplicate entries, deemed repetitive by an algorithm. Yet the process of identifying seemingly identical digital versions of books, newspapers, and other collections objects can also reveal key differences that have the potential to illuminate what each object means in its originating context. Duplicates are a key to learning about the heterogeneity of data infrastructures.

Instead of ridding the DPLA of redundancies, why not learn from them? When seen in this way, the initiative presents an opportunity to use data infrastructures to study the production of data—raising important questions about the local histories of collecting across the country: Who makes data and why? What does *data* mean in different cultures of collecting? What kinds of values and assumptions can data hold?

Independently of the DPLA, a number of recent projects have shed light on the importance of duplicates in library collections. For example, a project called "Book Traces" at the University of Virginia implicitly demonstrates the relevance of duplicates by systematically documenting the unique attributes of individual, physical books. The authors of "Book Traces" invite us to explore how "old library books bear fascinating traces of the past."¹⁴

" Thus, seamed with many scars,
 Bursting these prison bars,
 Up to its native stars
 My soul ascended !
 There from the flowing bowl
 Deep drinks the warrior's soul,
 Skoal ! to the Northland ! Skoal ! " *
 — Thus the tale ended.

* In Scandania this is the customary salutation when drinking a health. I have slightly changed the orthography of the word, in order to preserve the correct pronunciation.

Then you looked at your watch & said —

" Now, shall we go & make that visit, for at 5 o'clock I have to go to Washington " & we meant, you & I, & we had a happy walk —

Our last walk together in this world — Never to see each other more — Never, oh, never !

It was after this, I called you — " Noseman " the name we always used to me and, in our letters. Do you remember ?
THE WRECK OF THE HESPERUS.
You added to it " your Noseman " and you " devoted Noseman " —

It was the schooner Hesperus,
 That sailed the wintry sea ;
 And the skipper had taken his little daughter
 To bear him company.

Blue were her eyes as the fairy-flax,
 Her cheeks like the dawn of day,
 And her bosom white as the hawthorn buds,
 That open in the month of May.

The skipper he stood beside the helm,
 With his pipe in his mouth,
 And watched how the veering flaw did blow
 The smoke now West, now South.

3.3

An 1891 book of poems and ballads by Henry Wadsworth Longfellow, annotated by Jane Chapman Slaughter. Image from the "Book Traces" project.

Over the course of the “Book Traces” project, which is ongoing, the authors have enhanced existing collections data to describe what they call “interventions” in thousands of books. “Readers wrote in their books, and left pictures, letters, flowers, locks of hair, and other things between their pages.”¹⁵ In one example (figure 3.3), a book of poems by Henry Wadsworth Longfellow is annotated with bittersweet memories of reading it with a lost friend.¹⁶ The creators of “Book Traces” have also developed generalized methods for physically surveying a large number of volumes for such interventions. Following on their project, others have been inspired to explore interventions as evidence of the complex lives of physical books and thus grist for the mill of academic research. Devoney Looser, an English professor at Arizona State University, writes that “the items found in the books ... provide important information about circulation and authorship, and are of interest to critics, historians and biographers.”¹⁷

But beyond identifying unique conditions in redundant copies of books for the purposes of narrowly defined scholarship, “Book Traces” demonstrates the educational merit of examining physical collections. Rather than trying to streamline the digitization and ingestion process, as the DPLA has sought to do, what can be learned by seeing that process as an opportunity for critical reflection and hence a point of entry for student engagement? The challenges of digitization can, when seen in the right way, become lessons on the working of knowledge systems like library collections. What the “Book Traces” project misses, and I am seeking to unravel, are the ways in which data can also act as cultural markers of past collection practices, and how they differ from one era or institution to the next. The agglomerated data of the DPLA provides a number of opportunities to understand what it means to look within data infrastructures for the local conditions in all data.

IDENTIFYING THE LOCAL

The DPLA became a self-supporting nonprofit a year after the Appfest. On its first anniversary in 2014, the organization reportedly contained “over 7 million digitized cultural heritage items from 1,200 contributing institutions across the United States.” Today, in 2019, the DPLA is a thriving nonprofit organization, guided by a board of directors that includes academics, librarians, publishers, and businesspeople. The primary interface to the DPLA—a standard search bar with the heading “A Wealth of Knowledge: explore 11,578,169 items from libraries, archives, and museums”—promises equal access to each individual repository (figure 3.4). The initiative describes its mission of cultural collecting as all encompassing: “It strives to contain the full breadth of human expression, from the written word, to works of art and culture, to records of America’s heritage, to the efforts and data of science.”¹⁸

But a basic search of the DPLA’s unified collections (figure 3.5) conceals the striking heterogeneity and unevenness of the underlying data. Below, I demonstrate local readings of collections data using examples drawn from the DPLA. These efforts variously

reveal local classifications, schemata, constraints, errors, absences, and rituals, all of which are rooted in the contingencies of the places in which DPLA data are made.

Questioning Classifications

At the conceptual level, data are shaped by local classifications. Yet audiences typically don't take notice of them, unless their origins are unfamiliar. Jeffrey Licht, a technologist working for the DPLA, calls attention to a record that would appear unfamiliar outside South Carolina.¹⁹ The DPLA contains a group portrait, contributed by Clemson University, with a field labeled "coverage" containing a single string, "upstate," presumably referring to a place in South Carolina. The field (coverage) and string (upstate), however, have little meaning to either Licht or me. Neither of us are from the region. But such language shouldn't be presented as anomalous: a mere obstacle to accurate geocoding, the process of turning places into coordinates on a map. Instead, the example should compel us to think about the local nature of all place-names. Local classifications are a product of geographies as well as other social boundaries, such as those that separate disciplines.

Seeing Schemata

In the collections data of the New York Public Library, a major civic institution and early contributor to the DPLA, one can find at least 1,719 unique date schemata: ways of recording the moment that a book, image, or other library artifact came into the world. This detail was already introduced at the beginning of the book, but below I offer a deeper look at a sample of abstracted date schemata. These are not actual dates. Rather, each represents one way of documenting a date of publication.

Printed by Thomas; Badger, Jun (1)

pref _____] (1)

__ March, _____ (3)

probably before _____ (7)

[c_____]/ _____ (130)

_____ - _____, _____ - _____ (209)

_____ - _____, reissued through _____ (240)

_____ - _____ - _____ / _____ - _____ - _____ (438)

ca. _____'s (640)

The “_” in each schema is a variable standing in for a variety of possible integers. The number in parenthesis indicates the total times that the format appears in the New York Public Library catalog. Thus the common schema ca. _ _ _'s, used 640 times, might be encountered as ca. 1950s. The less common formats are at the extreme ends of uncertainty: either highly ambiguous or strangely specific. In one case cited here, the format includes the name of the printer. It appears only once. Although we can't

BROWSE BY TOPIC | BROWSE BY PARTNER | EXHIBITIONS | PRIMARY SOURCE SETS | ABOUT DPLA | NEWS | DPLA PRO





D P L A DIGITAL PUBLIC LIBRARY OF AMERICA [Donate](#)

Discover 21,455,806 images, texts, videos, and sounds from across the United States

Search the collection [Search](#)






[Browse by Topic](#) [New? Start Here](#)

Online Exhibitions [Browse all Exhibitions >](#)

Two Hundred Years on the Erie Canal American Empire Battle on the Ballot: Political Outsiders in US Presidential Elections Race to the Moon

Primary Source Sets [Browse all Sets >](#)

Cotton Gin and the Expansion of Slavery Elie Wiesel's *Night* and the Holocaust Victorian Era Immigration through Angel Island Dutch New Netherland

3.4

The main page of the DPLA website.

HOME
BROWSE BY TOPIC
BROWSE BY PARTNER
EXHIBITIONS
PRIMARY SOURCE SETS
ABOUT DPLA
NEWS
DPLA PRO

DPLA
DIGITAL PUBLIC LIBRARY OF AMERICA
America
Search

2,469,217 results for America
Sort by: Relevance
Items per page: 20

Refine your search

Type

- image: 1,787,318
- text: 578,775
- moving image: 16,717
- sound: 6,887
- physical object: 406
- collection: 22

Subject

- Plantae: 982,653
- Dicotyledonae: 697,880
- Asterales: 278,567
- Asteraceae: 267,451
- Anthropology: 202,238
- Rosales: 156,877
- Fabaceae: 152,203
- Archaeology: 138,626
- Pteridophyte: 138,611
- Monocotyledonae: 129,181
- Cyperales: 110,897
- Cyperaceae: 95,657

Date

Between Year

and Year

Update

Location +
 Language +
 Contributing Institution +
 Partner +

In America
[1906] - Gorky, Maksim, 1868-1936
At head of title: Massimo Gorki.
[View Full Item](#) in University of Illinois

This is America
[2002?]
Shipping list no.: 2002-0134-P: Special English, Cover title.
[View Full Text](#) in Purdue University

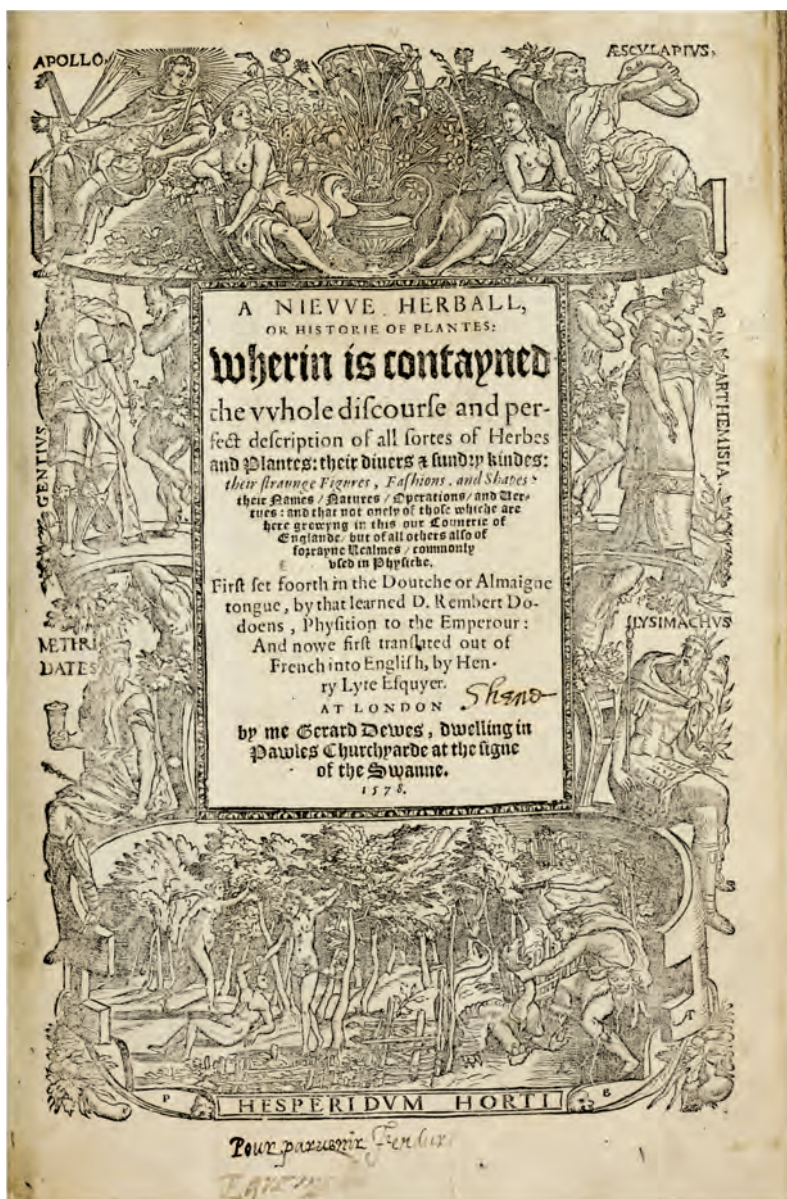
America
1803
Anonymous (British 19th Century).
[View Full Image](#) in The Miriam and Ira D. Wallach Division of Art, Prints and Photographs: Print Collection. The New York Public Library

America
1801
Anonymous (British 19th Century).
[View Full Image](#) in The Miriam and Ira D. Wallach Division of Art, Prints and Photographs: Print Collection. The New York Public Library

America
1705 - Harris, John (fl. 1680-1740)
Lawrence H. Slaughter Collection. National Endowment for the Humanities Grant for Access to Early Maps of the Middle Atlantic Seaboard. Prime meridian: Ferro. Relief shown pictorially. Shows Californi...
[View Full Image](#) in Lionel Pincus and Princess Firyal Map Division. The New York Public Library

3.5

An example search on the DPLA website.



3.6

An example of a historical book that causes numerous production artifacts from the DPLA collections.

understand the most obscure date schemata without further inquiry, they stimulate curiosity about their own invisible local histories.

Attending to Constraints

Ways of inscribing data are always constrained by local conditions. One well-known example of the technical limits on data comes from the turn of the millennium. Leading up to the year 2000, digitized date codes had to switch from two to four digits. Otherwise “00,01,02” might be mistaken for 1900, 1901, 1902 versus 2000, 2001, 2002. Across contributing collections to the DPLA—and indeed the world—databases had to be updated at great cost. And still, fears persisted that some unseen conflict in formats might cause whole systems to fail. Two-digit date formats are local markers of a bygone era. In previous eras, when storage space was much more costly, programmers used two-digit date codes to save space. We should read such legacy constraints as evidence of the way that data are located in a technological moment. But the lesson is more general: without the software and hardware of their era as well as operating knowledge thereof, data would not be accessible at all.

Decoding Errors

Every contributing collection to the DPLA contains errors. One doesn’t need much instruction to notice misspelled words or misplaced punctuation. But it takes a degree of local knowledge to see that such errors are not random. In fact, we can unpack them as evidence of localized cataloging practices.²⁰ They stem from situated processes of data production.

Badly scanned text, blurred photos, and moiré effects—all common in DPLA records—are a result of specific imaging technologies and ways of making use of them within a local setting. Typographic errors sometimes originate in the use of type from a particular historical moment (figure 3.6). They are brought on by the optical character misrecognition of unusual typefaces, ligatures, or unexpected characters. For instance, the standing *s* in early modern English typography is routinely mistaken for an *f* by character recognition systems. Understanding this has value beyond the amusement of contemporary readers. Alternatively, errors can be brought about when content is mistaken for code. Brackets, dollar signs, and semicolons can be interpreted as instructions to be carried out by a computer program.

We often hear about the arduous but imperative need to rid data sets of such flagrant errors through acts of cleaning or filtering. As I first explained in chapter 2, however, such instances of data dirt are simply out of place. In other words, errors in collections data might be better understood as signifiers taken out of their original interpretative contexts. We should learn to read data dirt as important traces of their own local production.

Revealing Absences

More subtly than in the previous illustrations, data are defined by what they leave out. Former Smithsonian historian Marya McQuirter recounts having searched her institution's catalog, one of the largest contributors to the DPLA, for the terms *black* and *white* (figure 3.7).²¹ The first brings up lots of examples of African American artists; the museum diligently documents work created by or about people who identify racially as black. But the search term *white* brings up little about race, other than the occasional piece linked to white supremacy. White is not a racial identity that the Smithsonian typically tracks. Instead, the category exists as an absence that reveals a bias.²² Whiteness is not critically examined by the institution. Yet it is the racial identity of the vast majority of artists whose work is shown at the Smithsonian. As this example demonstrates, absences are part of deeply rooted systems of representation—in this case, white supremacy—reified in data.

Observing Rituals

Finally, and less overtly visible, are the local rituals that shape data. I use the term *ritual* here to identify cultural practices with data that have their own significance as symbolic expressions or community-making activities.²³ Thomas Ma, a cataloger at the Harvard Library—another of the heaviest contributors to the DPLA—reflects on the way that cataloging has changed over the course of his career.

I remember when I first started at the law school, I was told by the person in charge of technical services that “you weren’t worth anything if you didn’t have a backlog.” And nowadays it’s like if you have a backlog, you have the cooties. So the backlog [used to be] evidence of a certain kind of care and quality and attention in the catalog processing. And now the backlog is a distinct liability.²⁴

In this instance, practices with data are closely tied to professional identity and status. Moreover, the change in cataloging practices has significant practical implications. For when backlogs pile up—sometimes consisting of tens of thousands of accumulated books that need to be processed—cataloging is outsourced. “We used a company in Arkansas. I guess they find cheap labor. ... They just grab people off the street and say, here, slap a record together and move it on.”²⁵ Ma’s words are laden with an implicit argument for preserving the social milieu in which he was trained. His data rituals are evidence of a local social order. Indeed, without the proper rituals, argues Ma, the quality of library data is in danger. Ma forecasts a dark future for collections: they contain more entries than ever before, but their data—the maps to those collections—are increasingly thin. From this perspective, the movement toward data infrastructures can paradoxically make individual records less accessible.

As the examples above illustrate, there is no such thing as universal data. Data are situated within the means of their production, the infrastructures required to maintain them, their systems of representation, and the social order that they reproduce. This perspective extends social studies of information that focus on the specificity of data at the level of the institution, such as the museum or laboratory.²⁶ Such studies often assume that standards are the primary forces that shape data. Lisa Gitelman writes that “every discipline and disciplinary institution has its own norms and standards for the imagination of data.”²⁷ As the above readings of DPLA data suggest, though, local variations can also be subject to a number of historical, technological, and cultural contingencies, which transcend disciplinary boundaries. Too often, these differences are passed off as anomalies, to be resolved by normalizing data. But we can learn to see differences in data as markers of otherwise-invisible local conditions that must be understood for meaningful analysis.

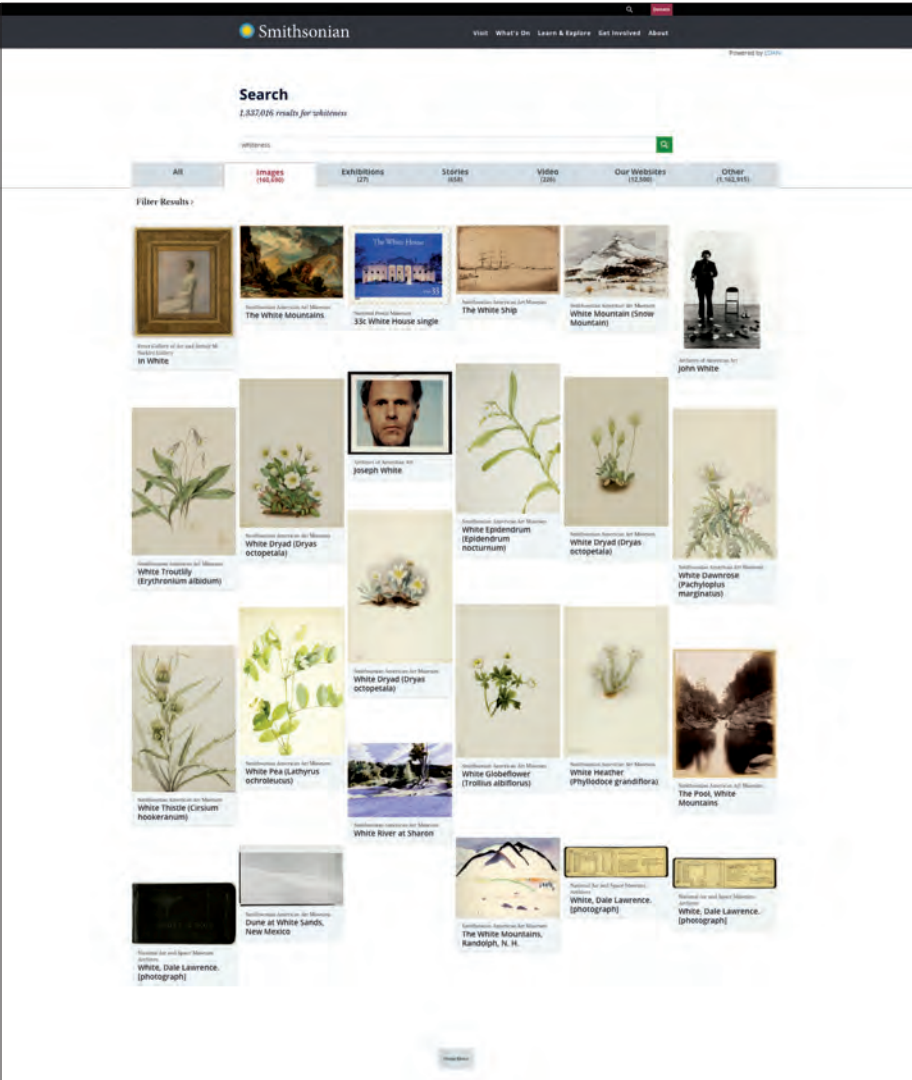
I do not mean for the six features examined above—classifications, schemata, constraints, errors, absences, and rituals—to be taken as a fixed typology for local readings of data. Rather, they convey the contingent character of data through examples that cover a range of possible scales and local ties. What appears to be local in data depends on emergent differences among data aggregated from many places. As first suggested in the introduction yet more fully illuminated here, the local is only intelligible when seen through a comparative lens. When data are drawn together from disparate origins, conflicting practices of data production are suddenly apparent.

Local markers are especially relevant when exploring big data.²⁸ For although the term indicates a departure from the local, the rise of the big data phenomenon has ironically made the local qualities of data more significant. Under big data, distributed records with discordant local ties are increasingly estranged from their creators and presented to audiences other than those first intended.

The DPLA might not conform to the strict definitions of big data, mentioned earlier in the book, emerging from critical data studies: high in volume, variety, and velocity. Yet it and other data infrastructures explored in this book are important references for grappling with the encompassing cultural phenomenon of big data, which is best understood as a desire for data sets that are intended to be comprehensive and autonomous, capable of yielding insights without contextual information.²⁹ This definition speaks to the universalizing ambitions of many data infrastructure projects, and prompts us to think about when and why they fall short.

VISUALIZING THE LOCAL

Reading the data of the DPLA one entry at a time, and looking for local instances of schemata, errors, constraints, classifications, absences, and rituals, can be revelatory but also time consuming. Furthermore, some of those conditions are instances of patterns



3.7
Example searches conducted by the author for *whiteness* and *blackness* on the Smithsonian website.

634.834 results for blackmen

Business

[Images](#)

Exhibitions

Stories

Video

Our Websites

Other

Filter Results: 0



that appear at larger scales. Revealing the larger pattern and its causes requires different methods of data presentation. Data visualization, first utilized in the last chapter as a means of seeing the forest, all the data at once, can also illuminate the trees, markers of difference within. Used in this way, visualization can become a mechanism for “infrastructural inversion,” through which foundational though invisible patterns are made visible.³⁰

When Geoffrey Bowker and Susan Leigh Star introduced the concept of infrastructural inversion in *Sorting Things Out*, they proposed that the most straightforward way to make infrastructure visible is to break it. Indeed, errors in DPLA data, such as the misrecognition of the long s by optical character recognition, can call our attention to the infrastructural processes by which those data are produced. Yet there are other strategies for seeing infrastructures that we might invoke. Here, as in previous examples, I use a comparative approach. Comparative visualization can illuminate how locally distinct collections are assembled and assimilated differently in data infrastructures like the DPLA. I developed two distinct visualization programs to demonstrate this at different scales.

The first of the two, entitled “A Comparative Visualization of Temporalities in DPLA Data,” or “Temporalities” for short, revisits the startlingly distinctive date formats to be found in the ingested collections data of the DPLA. These are data that have already been normalized: changed in order to conform to a single format. But wisely, the DPLA chose to preserve the original records. This visualization unearths those records and organizes them from most legible to least, but by not human legibility. Their machine legibility is determined by the prevalence of date signifiers like months, days, and years, which a computer can easily be made to identify.

This way of putting the date formats on the page gives the reader a sense of the range of original schemata drawn together in the DPLA as well as the normalization work necessary to make them conform to a single format. The temporalities that do not conform are subject to procrustean measures: stretched out or cut short as needed. This is the evidence of what must be obscured or added in order to normalize just one data field.

A Comparative Visualization of Temporalities in DPLA Data

Temporalities (figure 3.8) is coded in JavaScript to circumvent the DPLA’s default web interface (www.dp.la/) and communicate directly with its application programming interface (API).³¹ This backdoor interface is the only means of accessing the original records from contributing institutions like the New York Public Library, Smithsonian, or Digital Library of Georgia.

Here is a brief explanation of how the program works. It begins with a search request to the API. I have used the search term *America* in the trial run above as a gesture to the full scope of resources that the DPLA aims to encompass. The results of this

search, returned in JavaScript object notation, are at first stored in memory. For each individual record returned by the API, the code identifies a field titled “sourceResource.date.displayDate.” This is the yet-to-be-normalized date format provided to the DPLA by contributing institutions. Other fields found in each DPLA record, date.begin and date.end, contain normalized date values: representations of the date that have been created to facilitate consistent searches from the main DPLA interface.

The code then reformats each sourceResource.date.displayDate: integers (frequently used to represent days, months, and years) are converted to underscores (\d to _), and months are converted to double underscores (Jan|Feb|Mar| ...) to __). Integers and months are the most common features in these dates. Seeing their arrangements provides a useful comparison of underlying commonalities. Finally, each format is saved to a list, used to keep a tally of how many instances of each common string (or schemata) are found. The final visualization shows this list: all the date formats saved for that one search.

In order to help the reader make sense of this list, it is sorted in terms of its machine legibility—defined here by the ratio of underscores to other characters. This ratio indicates the amount of machine-readable information in the format. When examples have the same ratio, those with more underscores are listed first. Date formats that are text heavy and thereby less machine readable will be pushed to the bottom. These dates are more difficult to normalize. The hardest dates to normalize, the strangest or most unexpected ones, require manual translation. For example, Roman numerals cannot easily be read by a machine. A program is likely to interpret all Vs, such as the V in “Version,” as the Roman numeral equivalent of 5. Data cleaning has its limits, or at least requires a lot of locally sensitive rules. The numbers beside each format indicate how many times each format appeared in records returned by the API search.

Temporalities inverts the infrastructural process of ingestion. It does so by sectioning the DPLA along a single field and exposing traces of the deep intricacy of contributed collections. In Temporalities, that intricacy is exhibited in the range of localized date formats to which data are subject across cultures of collecting. In the context of the DPLA, meant to be an inclusive repository of US cultural history, such variations in data are worthy of attention and even beautiful in their own right. The second program, entitled “The Shape of DPLA Data,” or “Data Shapes,” also grapples with this heterogeneity and how to present it, but with a focus on the entire data structure of each DPLA entry.

3.8

The Temporalities application visualizes date formats from the DPLA in order of their machine legibility. Image by the author and Peter Polack.

31 -----
1 -----
16110 ----
18 -----
2 -----
827 -----
404 -----/
1 -----
1 -----,
2301 -----
23 -----
12 -----/
11 -----
3 -----
2 -----
12 -----/
11 -----
2 -----
1 -----[]
3597 -----
16 --/ --/
13 -----]
5 -----?
5 -----,
4 c -----
4 -----
3 --" --
3 -----, c
2 -----/
2 -----
2 -----]
1 -----
1 -----/
580 -----]
347 c -----
168 -----
29 -----?
12 -- --
11 - ----
8 -----~

3 _____,
 3 © _____
 2 _____s
 1 d _____
 64 _/_/_/_____
 9 ____u-_____
 4 _/_/_/_____
 2 _____-____u
 2 _____-____
 1 _____
 1 _____-____
 1 _____-____
 12 _____-____/_/____-____-____
 4 _____-____-____-____-____-____
 1 _____-_____, c _____-_____
 3 _____, c _____-_____
 8 _____-____]
 6 _/_/_/_____
 2 _____]-____
 2 _____. _____
 1 c _____-____
 1 _____-____
 1 _____-____
 1 _/_/_/_/_____
 1 _____-____~
 1 _____-____'
 1 _____-____?
 1 ____-_____
 8 ____u
 2 l_____
 1 ____-____
 6 _____-____-____T____:____:____
 466 _____-_____
 108 _____, c _____
 20 [_____-_____
 11 _____-[_____
 8 [_____,_____
 7 _____?/_/____?
 7 _____, _____
 6 _____[c_____
 4 _____ [_____
 2 c_____, _____
 2 _____, ©_____
 2 _____, ____
 2 [_____-_____
 1 c_____-_____
 1 _____~/_/____~
 1 _____?-_____
 1 _____-____?]
 1 _____[-_____
 1 _____?-_____
 210 _____-____-____T____:____:____-____:____
 3 _____-____, _____
 13 _/_/_/_/_____
 1 ____-____-____
 1 _____s (_____-_____
 6 _____-____-____T____:____:____Z
 5 _____, _____
 1 _d _____
 3 _____-____, _____
 1 {_____-____-____._____-____-____}
 2 _____-____ [c_____-____]
 3 _____ [c_____-____]
 1 _____-____ [c_____
 110 [_____._____
 57 _____ [c_____
 17 c. _____-____
 5 [_____, _____]
 3 _____ to _____
 3 [c_____-_____
 2 _____, [_____
 2 _____ [©_____
 2 _____, [c_____
 2 _____ or _____
 2 [_____, c_____
 2 [_____-____?]-_____
 1 _____, c._____
 1 [_____-____]-c_____
 1 c[_____-_____
 1 ____y. _____

- 1 ____-[c____]
 1 [____?~____]
 1 _____, c____-
 1 __th____
 1 c____-c____]
 1 [____], ____
 1 ____-[____?]
 1 [____-____?]
 24 __u u-____
 10 ____-__u u
 9 [____-__]
 5 ____-[__]
 3 __u-__u
 2 [____]-__
 1 c____-__]
 1 _____, '____
 1 [__]-____
 1 ____-____.]
 1156 [____]
 220 ____?]
 42 c____]
 12 ____-].
 8 c____-
 5 ____?~
 2 c.____
 2 ____-~
 1 -____]
 1 c____"
 1 _/_/_/
 1 -c____
 1 _____,
 1 [____-
 1 c____?
 1 ____?>
 1 _-~____
 1 ____ (
 1 ____-____ [v.____]
 1 ____ [c____]-____ [c____]
 1 _d __, ____
 1 ____ [____ or ____]
- 20 ca. ____-____
 12 [____, c____]
 6 _____, [c____]
 2 [____]-[____]
 2 ____ or ____]
 2 ____-ca. ____
 1 ____ [c.____]
 1 c____-[c____]
 1 [c____-c____]
 1 [____?~____?]
 1 [____?], ____
 1 ____ and ____
 1 [____? c____]
 1 [____ [c____]
 1 [____?, c____]
 1 [c____, ____]
 16 [c____]-__
 10 [c____-__]
 1 c____-c____]
 1 [____?~__]
 3 ____-]
 1 ____-?
 3 [____ or ____]
 2 ____ i.e. ____
 1 [____?, ____?]
 1 [____, '____-__]
 1 _____, [c____-__]
 1 ____? to ____?
 587 [____?]
 452 [c____]
 37 ca ____
 9 c. ____
 8 [©____]
 4 [____]-
 2 [____s]
 2 ____?~]
 1 c[____]
 1 ____ AD
 1 [____!]
 1 ____-]

1 [____.]
 1 __[____]-
 1 c ____-
 1 wc____]
 1 ____?]
 1 -[____]
 1 [c____-__] v.__, ____[c____]
 1 ____-__, t.p. ____
 1 ____ [i.e.____]-__
 3 __u-__uu
 2 _uuu-____
 2 __uu-__u
 3 ____-__ [v. __, ____]
 3 [c____-c__]
 1 [©____-c__]
 1 [__-?]-____
 8 circa ____-____
 5 ____-circa ____
 1 ____, t.p. ____
 1 ____,repr. ____
 1 ____, i.e. ____
 1 ____- [v.__, ____]
 6 circa ____ - ____
 1 ____ [i. e. ____-__]
 11 ____ [i.e. ____]
 1 ____ [cop. ____]
 1 [etc.] ____-____
 111 ca. ____
 3 ca ____s
 3 ____- __
 3 ____, ca
 2 [c.____]
 2 [____-]
 2 [____?]-
 1 [c____.]
 1 [____?]-
 1 [c ____]
 1 [c____]-
 1 cop.____
 1 [c____-]

1 c. ____s
 1 [©____-]
 1 [c____?]
 1 [__ -__]
 12 ____-?]
 9 [____-]
 2 [____?]
 1 [l____]
 9 __uu
 2 __--
 1 Vol. __- (____-____)-
 1 ____-__, [v. l, ____]
 3 ca. ____-ca. ____
 1 modeled ____-____
 1 ____, [i.e. ____]
 1 ____ [cover ____]
 1 Circa __ ____, ____
 1 Vol. __ (____, ____ to ____ __, ____)-v. __ (____ __ to ____ __, ____)
 1 ____-between ____ and ____
 1 after __ ____
 1 Circa __ ____
 2 __. _st ____ [i.e. ____]
 1 Vol. __ (____ __, ____ to ____ __, ____)-v. __ (____ __, ____ to ____ __, ____)
 1 ____-____, [reprinted ____]
 1 ____, i.e., [____]
 5 __uu-__uu
 2 ca. ____?
 1 _uuu-__u
 1 c____. --
 1 c. ____s?
 1 cop. ____
 1 ca. ____s
 1 ca. ____]
 1 __r. _th ____ [i.e. ____]
 1 after __, ____
 26 [____-?]
 1 [____-]-
 1 __. ____-v. __, issue __ (____. ____)
 4 ____, ____ printing
 1 __th ed. (____/____)-__th ed. (____/____)

1 ____- [v. __, pt. __, ____]
 1 Began with ____-____
 1 ____; reprinted ____
 1 __th ed. (____/____)
 17 circa ____
 10 [ca. ____]
 7 after ____
 6 [cop. ____]
 3 Circa ____
 1 ____ circa
 1 pref. ____
 2 __--?
 2 __--]
 1 __nd ed. (____/____)-__th ed. (____/____)
 1 __th ed., ____-____-
 1 Vol. __ (____-____)-vol. __, no. __ (____/____)
 1 __th ed. (____/____)-__th ed. (____/____)
 1 Vol. __, no. __ (____/____)
 3 between ____ and ____
 1 ____ [reprinted ____]
 1 __th ed. (____/____)-
 1 __th ed. (____/____)-__th ed. (____/____)
 1 ca. ____-
 1 between ____ and ____]
 2 before ____
 2 pref. ____]
 1 [p.d. ____]
 1 [cop. ____]
 1 _d print. ____
 2 ____, probably __
 1 __th ed. (____/____)-
 47 [between ____ and ____]
 1 Egyptian_____SIU.tif
 1 Vol. __, no. __ (____, ____)-Vol. __, no. __ (____, ____)
 1 Began in: ____-____
 1 Began with __-____
 7 [pref. ____]
 1 [cop. ____-]
 2 _uuu-__uu
 1 __-uuuu

3 __--?]
 3 [__--]
 2 [__--?]
 1 Vol. __, issue _ (____-v. __, issue _ (____)
 1 Began with: _____
 1 Print began and ceased with: No. __-____ (____, ____-____, ____)
 1 patented _____
 1 _____ printing
 1 anno XI--____
 1 [pref. _____.]
 1 Egyptian_____Rencher.tif
 1 Egyptian_____Antoine.tif
 1 Vol. __, issue _ (____-v. __, issue _ (winter ____)
 1 reprint _____, updated _____, original _____]
 1 Began in _____?
 1 _____ and later
 1 Began in: _____
 4 [__--?]
 1 [__--]-
 1 __--?]-
 1 Began with _st ed., _____. ____, _____
 1 Began with __th ed., _____. ____, _____
 1 ceased with __th ed., _____. ____, _____
 1 Colonial period, ca. ____-____
 1 Began with _st ed., ____, _____
 5 Began with _____
 1 MDCCLXXI [____]
 1 [foreword _____]
 2 Began with: _____
 1 [__--?]-
 2 _uuu
 1 Circa early _____s
 1 possibly ca. _____
 1 [not before _____]
 1 __th-__th century
 1 MDCCLXXXIII [____]
 2 Began in the _____s?
 2 _____, printed later
 1 probably after _____
 1 Print began in _____

1 [etc., etc., ____?]
 1 Yale University Press ____-__
 1 Publication began with v. __, no. _ (____, ____)
 1 Yale University Press ____-__
 1 Print began in ____?
 1 Vol. __-v. __
 1 Ceased with v. __, issue _ (winter ____)
 4 Printed in the year ____
 2 __th century
 1 Unpublished (before ____)
 1 Print began with: _st ed. (____)
 1 Began and ceased with: ____
 2 late __th-early __th century
 1 Print began with: Winter ____
 1 Ceased, per publisher, in ____
 2 late __th century
 1 Print began with: _st ed., published in ____
 1 Printed by Thomas Badger, __
 1 -
 1 n.d
 7 uuuu
 5 null
 1 n.d.]
 14 unknown
 5 Unknown
 3 undated
 1 Undated
 1 C. Tilt,
 1 M.DCCC.I
 1 [undated]
 1 unpublished
 1 C.T. Dillingham,
 23 [Date Unavailable]
 1 National capital press,
 1 Cambridge Botanical Supply,
 1 Printed by Gillespie Bros.,
 1 The Womens social & Political union,
 1 [Press of National printing co., etc.]

The Shape of DPLA Data

The second program, Data Shapes (figure 3.9), also begins with an API query to the DPLA. It returns a visualization for each individual data structure in the resulting DPLA entries. Each entry is represented as a network graph, which displays all the individual nested fields included therein. Related fields are connected by green threads and grouped at different levels in the data hierarchy within translucent convex shapes. Fields from the original contributor record are juxtaposed with fields created by the DPLA.

Like Temporalities, Data Shapes is an expressive project that examines what visualization can show us about the heterogeneity of the data infrastructures. Whereas most data visualizations focus on global patterns across a data set, Data Shapes show patterns of difference. It visualizes data fields and their relationships as opposed to data values, which is more typical of conventional visualizations.

Data Shapes interrogates the hierarchical structure of each DPLA entry as follows. The code starts at the root of the JavaScript object notation returned by the API. It then recurses through the entire structure of nested fields using a depth-first search. This means that for each (parent) field, the program visits all of its (children) subfields, then its (children's children) sub-subfields, before moving to the next (parent) field. For every individual (child) field on the tree, a new node is visualized and linked to the (parent) field one level up.

Data Shapes reveals the otherwise-invisible structures inside each record, including the original data from the contributing collection, and the Dublin Core data created for the DPLA.³² The pink areas show different subcollections of data, and the green lines show how they are related. The bold numbers are indexes for fields used in the records and listed in a legend on the far left of the screen. The light numbers show the frequency of the field's use in the current image. This legend can help an informed reader decipher the original data from their DPLA-generated counterparts, formed during the process of ingestion. Taken together, Temporalities and Data Shapes suggest new ways in which visualization can help us see data rather than seeing through them. In collections of cultural history for our digital age, even data deserve to be engaged on cultural terms.

In addition to seeking out the cultural histories of data, we should acknowledge the histories behind data infrastructures. For such infrastructures are no more neutral than the data they contain. Since their earliest imaginings, information processing infrastructures were meant to draw data together, regardless of their origins, and mine them for insights.

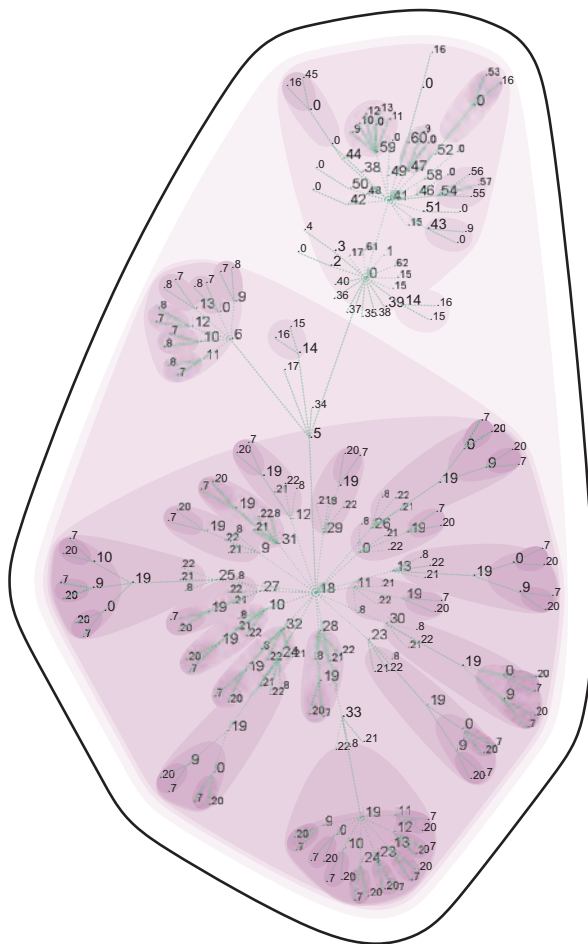
3.9

The Data Shapes program visualizes the structure of individual JavaScript object notation records from the DPLA. Image by the author and Peter Polack.

```

0 24 0
1 1 context
2 1 dataProvider
3 1 admin
4 1 object_status
5 1 originalRecord
6 1 controlfield
7 37 #text
8 23 tag
9 12 1
10 5 2
11 4 3
12 4 4
13 4 5
14 2 provider
15 5 id
16 5 name
17 2 _id
18 1 datafield
19 17 subfield
20 31 code
21 17 ind1
22 17 ind2
23 2 6
24 2 7
25 1 8
26 1 9
27 1 10
28 1 11
29 1 12
30 1 13
31 1 14
32 1 15
33 1 16
34 1 leader
35 1 object
36 1 aggregatedCHO
37 1 ingestDate
38 2 type
39 1 ingestionSequence
40 1 isShownAt
41 1 sourceResource
42 1 publisher
43 1 description
44 1 language
45 1 iso639_3
46 1 title
47 1 format
48 1 rights
49 1 contributor
50 1 creator
51 1 extent
52 1 spatial
53 1 country
54 1 date
55 1 begin
56 1 end
57 1 displayDate
58 1 specType
59 1 identifier
60 1 subject
61 1 ingestType
62 1 score
63 1 hasType
64 1 collection
65 1 relation
66 1 stateLocatedIn
67 1 physicalDescription
68 1 tmp_high_res_link
69 1 tmp_image_id
70 1 usage
71 1 namePart
72 1 valueURI
73 1 authority
74 1 displayLabel
75 1 titleInfo
76 1 supplied
77 1 tmp_rights_statement
78 1 relatedItem
79 1 note
80 17
81 18
82 1 tmp_item_link
83 1 originInfo
84 1 dateIssued
85 1 place
86 1 placeTerm
87 1 location
88 1 shelfLocator
89 1 physicalLocation
90 1 version
91 1 genre
92 1 rightsStatementURI
93 1 schemaLocation
94 1 typeOfResource
95 1 topic
96 1 geographic
97 1 stringValue
98 1 keyDate
99 1 encoding

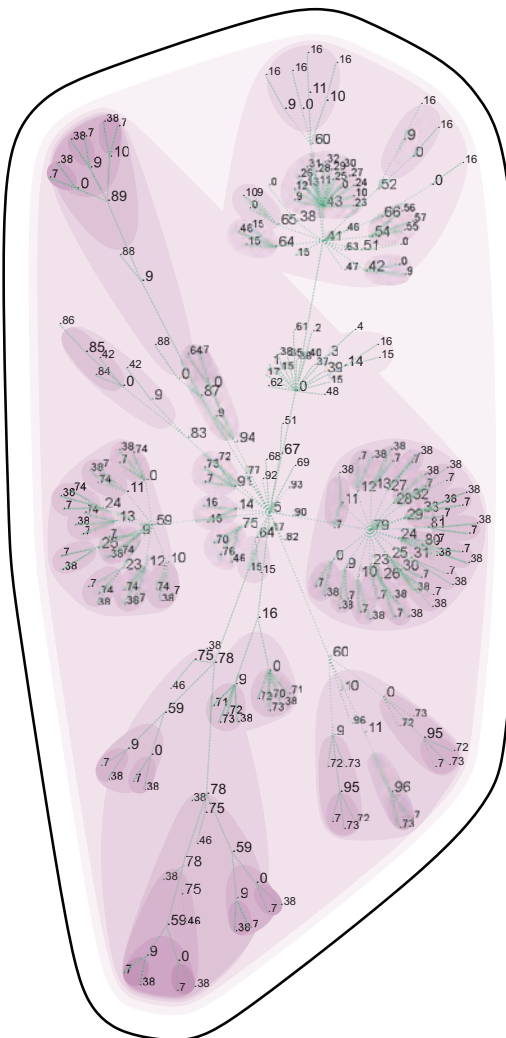
```




```

0 20 0
1 1 context
2 1 dataProvider
3 1 admin
4 1 object_status
5 1 originalRecord
6 controlfield
7 42 #text
8 tag
9 16 1
10 7 2
11 5 3
12 3 4
13 3 5
14 2 provider
15 9 id
16 10 name
17 2 _id
18 datafield
19 subfield
20 code
21 ind1
22 ind2
23 3 6
24 3 7
25 3 8
26 2 9
27 2 10
28 2 11
29 2 12
30 2 13
31 2 14
32 2 15
33 1 16
34 leader
35 1 object
36 1 aggregatedCHO
37 1 ingestDate
38 44 type
39 1 ingestionSequence
40 1 isShownAt
41 1 sourceResource
42 3 publisher
43 1 description
44 language
45 iso639_3
46 6 title
47 1 format
48 1 rights
49 contributor
50 creator
51 2 extent
52 1 spatial
53 country
54 1 date
55 1 begin
56 1 end
57 1 displayDate
58 specType
59 4 identifier
60 2 subject
61 1 ingestType
62 1 score
63 1 hasType
64 3 collection
65 1 relation
66 1 stateLocatedIn
67 1 physicalDescription
68 1 tmp_high_res_link
69 1 tmp_image_id
70 2 usage
71 2 namePart
72 7 valueURI
73 8 authority
74 8 displayLabel
75 4 titleInfo
76 1 supplied
77 1 tmp_rights_statement
78 3 relatedItem
79 1 note
80 1 17
81 1 18
82 1 tmp_item_link
83 1 originInfo
84 1 dateIssued
85 1 place
86 1 placeTerm
87 1 location
88 2 shelfLocator
89 1 physicalLocation
90 1 version
91 1 genre
92 1 rightsStatementURI
93 1 schemaLocation
94 1 typeOfResource
95 2 topic
96 2 geographic
97 stringValue
98 keyDate
99 encoding

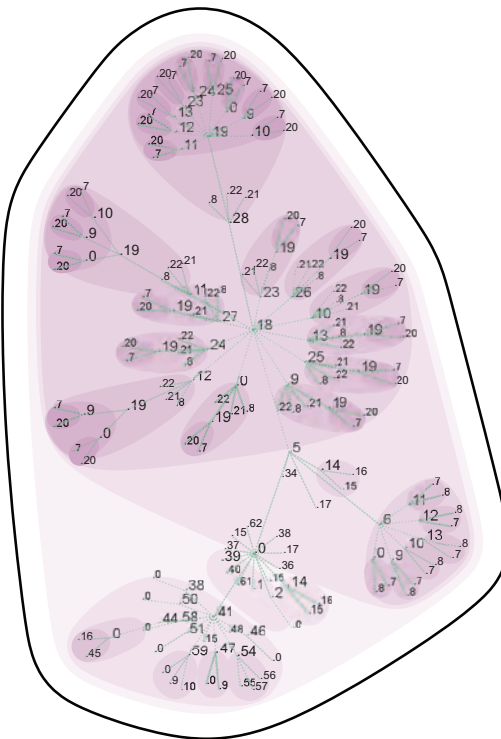
```



```

0 15 0
1 1 context
2 1 dataProvider
3 1 admin
4 1 object_status
5 1 originalRecord
6 1 controlfield
7 29 #text
8 18 tag
9 7 1
10 5 2
11 3 3
12 3 4
13 3 5
14 2 provider
15 5 id
16 3 name
17 2 _id
18 1 datafield
19 12 subfield
20 23 code
21 12 ind1
22 12 ind2
23 2 6
24 2 7
25 2 8
26 1 9
27 1 10
28 1 11
29 12
30 13
31 14
32 15
33 16
34 1 leader
35 1 object
36 1 aggregatedCHO
37 1 ingestDate
38 2 type
39 1 ingestionSequence
40 1 isShownAt
41 1 sourceResource
42 1 publisher
43 1 description
44 1 language
45 1 iso639_3
46 1 title
47 1 format
48 1 rights
49 1 contributor
50 1 creator
51 1 extent
52 1 spatial
53 1 country
54 1 date
55 1 begin
56 1 end
57 1 displayDate
58 1 specType
59 1 identifier
60 1 subject
61 1 ingestType
62 1 score
63 1 hasType
64 1 collection
65 1 relation
66 1 stateLocatedIn
67 1 physicalDescription
68 1 tmp_high_res_link
69 1 tmp_image_id
70 1 usage
71 1 namePart
72 1 valueURI
73 1 authority
74 1 displayLabel
75 1 titleInfo
76 1 supplied
77 1 tmp_rights_statement
78 1 relatedItem
79 1 note
80 17
81 18
82 1 tmp_item_link
83 1 originInfo
84 1 dateIssued
85 1 place
86 1 placeTerm
87 1 location
88 1 shelfLocator
89 1 physicalLocation
90 1 version
91 1 genre
92 1 rightsStatementURI
93 1 schemaLocation
94 1 typeOfResource
95 1 topic
96 1 geographic
97 1 stringValue
98 1 keyDate
99 1 encoding

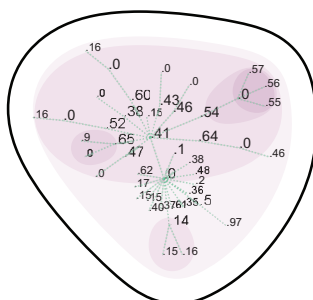
```

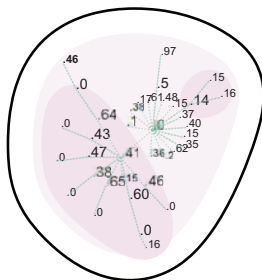


```

0 10 0
1 1 context
2 1 dataProvider
3 admin
4 object_status
5 1 originalRecord
6 controlfield
7 #text
8 tag
9 1 1
10 2
11 3
12 4
13 5
14 1 provider
15 4 id
16 3 name
17 1 _id
18 datafield
19 subfield
20 code
21 ind1
22 ind2
23 6
24 7
25 8
26 9
27 10
28 11
29 12
30 13
31 14
32 15
33 16
34 leader
35 1 object
36 1 aggregatedCHO
37 1 ingestDate
38 2 type
39 ingestionSequence
40 1 isShownAt
41 1 sourceResource
42 publisher
43 1 description
44 language
45 iso639_3
46 2 title
47 1 format
48 1 rights
49 contributor
50 creator
51 extent
52 1 spatial
53 country
54 1 date
55 1 begin
56 1 end
57 1 displayDate
58 specType
59 identifier
60 1 subject
61 1 ingestType
62 1 score
63 hasType
64 1 collection
65 1 relation
66 stateLocatedIn
67 physicalDescription
68 tmp_high_res_link
69 tmp_image_id
70 usage
71 namePart
72 valueURI
73 authority
74 displayLabel
75 titleInfo
76 supplied
77 tmp_rights_statement
78 relatedItem
79 note
80 17
81 18
82 tmp_item_link
83 originInfo
84 dateIssued
85 place
86 placeTerm
87 location
88 shelfLocator
89 physicalLocation
90 version
91 genre
92 rightsStatementURI
93 schemaLocation
94 typeOfResource
95 topic
96 geographic
97 1 stringValue
98 keyDate
99 encoding

```

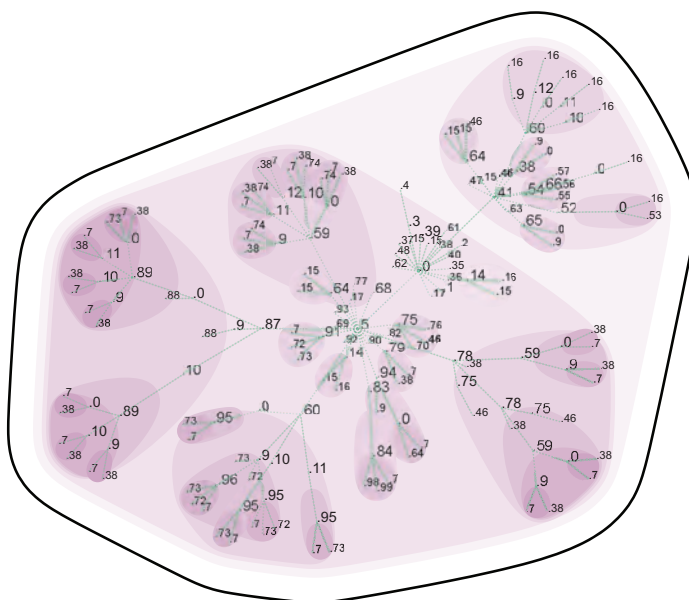




```

0 14 0
1 1 context
2 1 dataProvider
3 1 admin
4 1 object_status
5 1 originalRecord
6 controlfield
7 25 #text
8 tag
9 11 1
10 6 2
11 4 3
12 2 4
13 5
14 2 provider
15 9 id
16 9 name
17 2 _id
18 datafield
19 subfield
20 code
21 ind1
22 ind2
23 6
24 7
25 8
26 9
27 10
28 11
29 12
30 13
31 14
32 15
33 16
34 leader
35 1 object
36 1 aggregatedCHO
37 1 ingestDate
38 21 type
39 1 ingestionSequence
40 1 isShownAt
41 1 sourceResource
42 publisher
43 description
44 language
45 iso639_3
46 5 title
47 1 format
48 1 rights
49 contributor
50 creator
51 extent
52 1 spatial
53 1 country
54 1 date
55 1 begin
56 1 end
57 1 displayDate
58 specType
59 3 identifier
60 2 subject
61 1 ingestType
62 1 score
63 1 hasType
64 3 collection
65 1 relation
66 1 stateLocatedIn
67 physicalDescription
68 1 tmp_high_res_link
69 1 tmp_image_id
70 1 usage
71 namePart
72 4 valueURI
73 8 authority
74 4 displayLabel
75 3 titleInfo
76 1 supplied
77 1 tmp_rights_statement
78 2 relatedItem
79 1 note
80 17
81 18
82 1 tmp_item_link
83 1 originInfo
84 1 dateIssued
85 place
86 placeTerm
87 1 location
88 2 shelfLocator
89 2 physicalLocation
90 1 version
91 1 genre
92 1 rightsStatementURI
93 1 schemaLocation
94 1 typeOfResource
95 4 topic
96 1 geographic
97 stringValue
98 1 keyDate
99 1 encoding

```



PLACING INFRASTRUCTURES

There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less to remember, as they appear.

This passage offers a familiar depiction of a contemporary problem: an individual, working alone, is overwhelmed by sources beyond the scope of their expertise, collected by other people, in unknown times and places. This scene was described by the founder of the US National Science Foundation, Vannevar Bush, in 1945, thirty years before the invention of the personal computer. It is excerpted from his popular essay “As We May Think,” published in the *Atlantic* magazine, which has since become a landmark in the history of computing.³³

Bush’s proposed solution to the “mountain of research,” the memex (short for “memory expansion”), captures an early desire for a space in which a single investigator might navigate and connect data from disparate origins. He depicted the memex as an ordinary World War II era desk, outfitted with a pair of projection screens, a keyboard, an array of buttons and levers, and a memory store of unprecedented size for the time—all neatly concealed under the desk’s working surface. Harnessing a combination of dry photography and microfilm, the latest technologies of the postwar era, Bush believed that the machine might be capable of storing millions of independently addressed records: books, newspapers, photographs, and correspondence. The memex, writes Bush, would “instantly bring files and material on any subject to the operator’s fingertips.”³⁴

The idealized setting for data heralded by Bush’s machine—individualized, isolated, and unsurprisingly placeless—still speaks to the creators of today’s data infrastructures. Though never built, the memex articulated an emergent desire to liberate data from their attachments to institutionalized settings, which Bush argues are only obstructions to knowledge discovery:

Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can be in only one place, unless duplicates are used; one has to have rules as to which path will locate it, and the rules are cumbersome. Having found one item, moreover, one has to emerge from the system and re-enter on a new path.³⁵

Bush imagined that the memex could seamlessly extend human thinking—a function evoked in the title of his essay. He wanted a machine that, as Orit Halpern explains,

would “break the taxonomic and stable structure of the archive and would work ‘as we may think,’ by creating rhizomatic linkages and nonlinear associations between different pieces of information.”³⁶ To enable this nascent desire, Bush introduced the notion of *associative trails*, a precursor to what we now call *hyperlinks*, specifically meant to automate out librarians and catalogers.³⁷ He saw little merit in the work of “clerks,” offering his associative trails as a more “natural” alternative.³⁸ Bush wanted the investigator to be free to encounter data in a frictionless realm representative of and yet set apart from the world.

A similar ambition guides the development of contemporary data infrastructures, designed to collect, maintain, and distribute data across network technologies.³⁹ Indeed, contemporary data infrastructures act as Bush envisioned, but only in a superficial sense. They help scientists—but also educators, professionals, and an increasingly broad public—manage streams of data that would otherwise overwhelm an individual. Yet Bush did not predict some of the most important social changes in practices with data—changes that would make data simultaneously smaller and bigger than he could have imagined.

First, knowledge practices have rapidly diverged, leading to a variety of data cultures, each of which manages data in its own way. Second, data are widely distributed. The notion of a personal database has been replaced by a fascination with the potential of the web as a platform for access to data from almost anywhere. Third, data have become big business. As in the example of Zillow explored in chapter 5, the potential surplus value of data has stimulated aggregation, thereby, severing crucial ties between data and the local institutions in which they are made.

Learning to look for the local in data can help us see data infrastructures as composites; they juxtapose classifications, schemata, constraints, errors, absences, and rituals from diverse data sources. Data infrastructures would do well to acknowledge the history of collecting practices, for some day those same infrastructures will be historical relics too, and their choices will constrain future generations.

Seeing that all data are local means dismantling the image of data work manifest in the memex, a machine envisaged for a single investigator, a scientist (and man), sitting at a command-and-control center with access to the vast store of human knowledge. Instead, we must acknowledge that working with data is social; it requires continual communication and care.

I see this as an ethical shift from the model of the master sitting at his desk of power. We might yet move to a condition in which more intimate relationships with data and their subjects are widely fostered as an ethical responsibility. This is not some radical speculative future. It is demonstrated everyday by librarians, archivists, arboretum workers, and even some realtors.

One final caution about data infrastructures: simply using them can reinforce the social significance of the large, well-supported organizations that create and contribute

to them. Thus we also need counterdata or even antidata: tactical representations that challenge the dominant uses of data to secure cultural hegemony, reinforce state power, or simply increase profit. Counterdata might encode alternative perspectives to the cultural histories of the DPLA. Although counterdata have their own limits, they are necessary to counteract dominant representations of the past and ways of imagining the future; an alternative to the memex is long overdue.⁴⁰

CONCLUSION

In the *New York Review of Books*, Robert Darton sought to characterize the early efforts of the DPLA as an initiative that is at once universal and thoroughly American: “A library without walls that will extend everywhere and contain nearly everything available in the walled-in repositories of human culture. ... E pluribus unum! Jefferson would have loved it.”⁴¹

The tension expressed in this quote, between the all-encompassing ambitions of present-day data infrastructures and their local origins, attachments, and values—in this case, expressly US values—is a common refrain throughout this book. Darton portrays the DPLA as a “meta-mega-macro library,” a collection that could, in principle, contain everything. It is an old ambition that has only recently begun to seem like a possibility. But the name of the initiative (the Digital Public Library of *America*) alongside Darton’s evocation of the slave-owning “Sage of Monticello” should raise questions about the actual scope of the DPLA. Whose America does it represent? Moreover, cultural repositories are not simply “walled-in.” Walls are cultural constructions in their own right. They are a means of making a social order durable and hence important for understanding everything contained therein. Similarly, the unedited collections data contributed to the DPLA are the structural elements of its cultural histories, registering the norms that unite and separate diverse cultures of collecting within the United States.

Today there are a proliferation of initiatives that assemble data infrastructures from local, distributed sources, as is illustrated by the DPLA. Each data infrastructure promises to reveal new patterns across previously independent data sets. Yet these initiatives are not all the same in their motivations and goals. The DPLA is an example from the nonprofit world. Created with a mission “to educate, inform, and empower everyone in current and future generations” (www.dp.la), the DPLA is supported by a combination of private foundations, individual philanthropists, and US federal research agencies. Meanwhile, academia and industry have their own models. The next two cases in the book are also data infrastructures, but with different ends and means. NewsScape, an academic model supported entirely by public funds, is used for research in disciplines ranging from communication to computer science. Zillow is a model from the world of industry: a for-profit platform motivated by the potential for surplus value gained through the aggregation of data.

Despite their differences, the DPLA, NewsScape, and Zillow are unified by a consistent motivation: to assemble collections that are seemingly complete. Their creators aspire to build comprehensive perspectives on the world, offering vistas across all the books, all the news, or all the real estate opportunities. We might think of data infrastructures as an instance of what David Nye calls the “technological sublime,” for they test the limits of human perception and imagination.⁴² But as these cases reveal, data infrastructures only appear to be comprehensive. As I demonstrate, they have noteworthy limitations. There is an emergent need for alternatives to the universalizing discourses that surround data infrastructures, infusing them with a sense of truth and objectivity.⁴³

I propose that working effectively and ethically with data infrastructures means seeing them through a comparative lens—by acknowledging the whole as well as its local, heterogeneous parts. Local readings of data infrastructures, informed by interviews with those who make and use data, can reveal the variety of local ties that they harbor in classifications, schemata, constraints, errors, absences, and rituals. Seeing data infrastructures as assembled from local conditions opens up new opportunities and obligations for scholarship as well as pedagogy and practice. Yet we shouldn’t romanticize local ties; as some examples illustrate, they can be lacking in sophistication or even be discriminatory.

In the next two chapters, I deal with external though salient dimensions of data infrastructures: the algorithms that activate them, and the interfaces that recontextualize them. In chapter 4, I use NewsScape to uncover the deep historical and material entanglements between data and algorithms. In chapter 5, I explain, via Zillow, how we can develop critical perspectives on the mechanisms—visual, discursive, and algorithmic—by which interfaces make data actionable.

4