Thomas Padilla, “Conditions of Possibility”

Today I am going to speak about 3 conditions of possibility, that frame the way I look at the collections as data conversation. Agency. Empowerment. Ethics.

But first, a story.

Early last week, I exited the library at UC Santa Barbara, put on my sunglasses, and made my way across campus. It was a typical day, 72 degrees and sunny, no humidity. The quarter is just starting, so new students are out and about learning their way. It is a time of uncertainty. It is a time of possibility.

Tree of Life, Emmanuel Lubezki

I took a deep breath and tried to clear my mind. Lots of things compete for attention. To be honest it’s not always clear what matters, what *really matters*. It can be difficult to hold close to your heart the *reason of the things you do* and who you do them for. One foot in front of the other, I started to gain a measure of clarity. After near misses with bikes, longboards, and enterprising pokemon go players, I reached my destination. It was a building distinguished from those around it in that it rose higher than the rest – a tower.

I entered the elevator, red digits flicked to signal passing floors, the doors opened and I stepped toward a meeting with a director of a center on campus. I am sure there are folks in our respective communities walking to similar conversations right at this moment.

It was one of those first conversations where the pressure of possibility is everywhere.

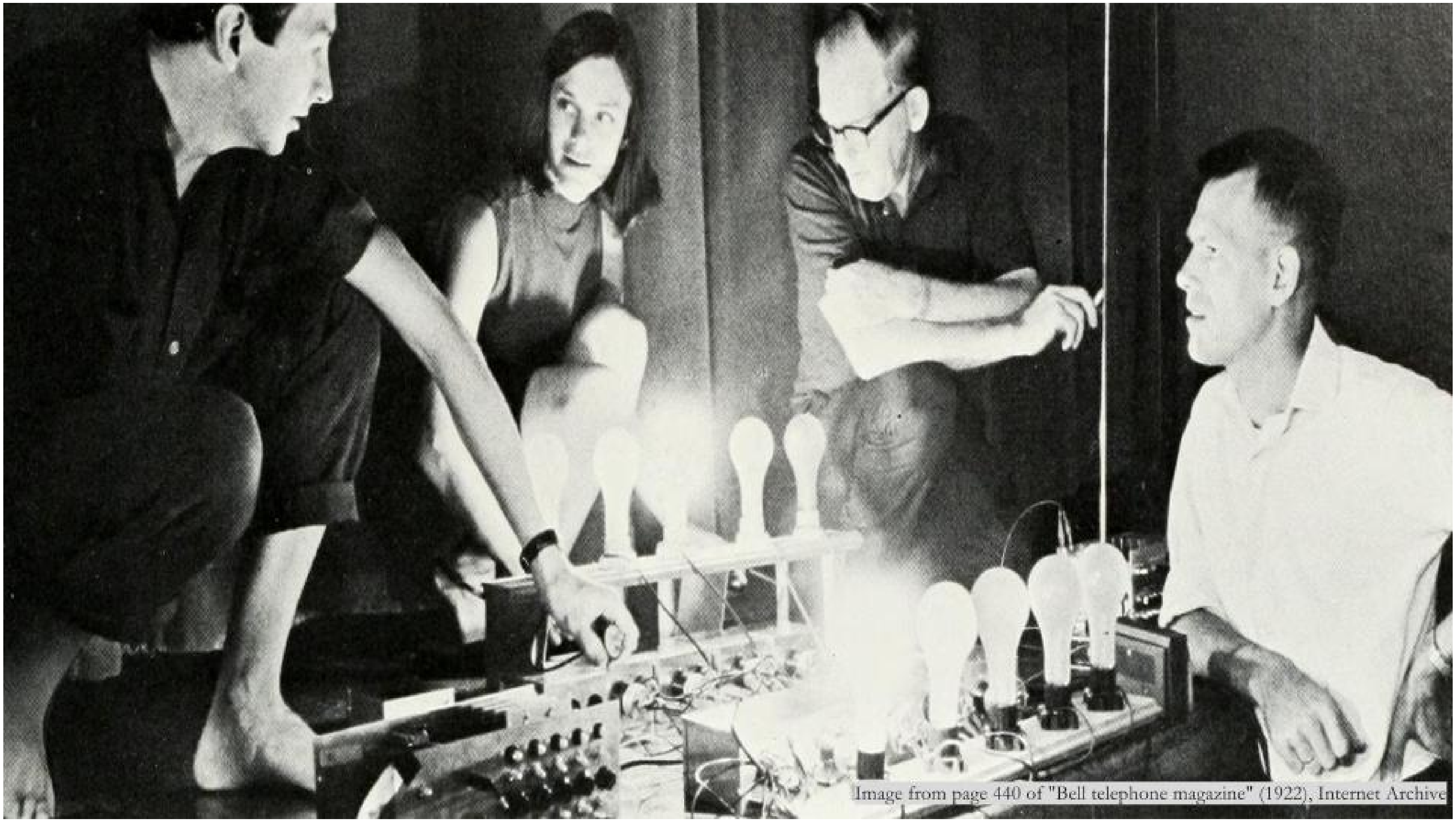


Image from page 440 of "Bell telephone magazine" (1922), Internet Archive

Our conversation ran a familiar course. Despite a bit of variation here and there the subtext buzzing below the surface is often the same. Who are you, what do you do, and why does it matter to me, *to my people*. In these conversations I often find myself in a bind of intelligibility and relevance – Humanities data, curator, librarian, digital humanities.
As the conversation progressed, there was a lot of academic hand gesturing around the table. We are good at that in university land.

After initial introductions I had transitioned to open the collections as data topic when the director interrupted me, took a measured breath, looked me squarely in the eyes and asked pointedly, “Yes, Thomas, but what are the stakes?”

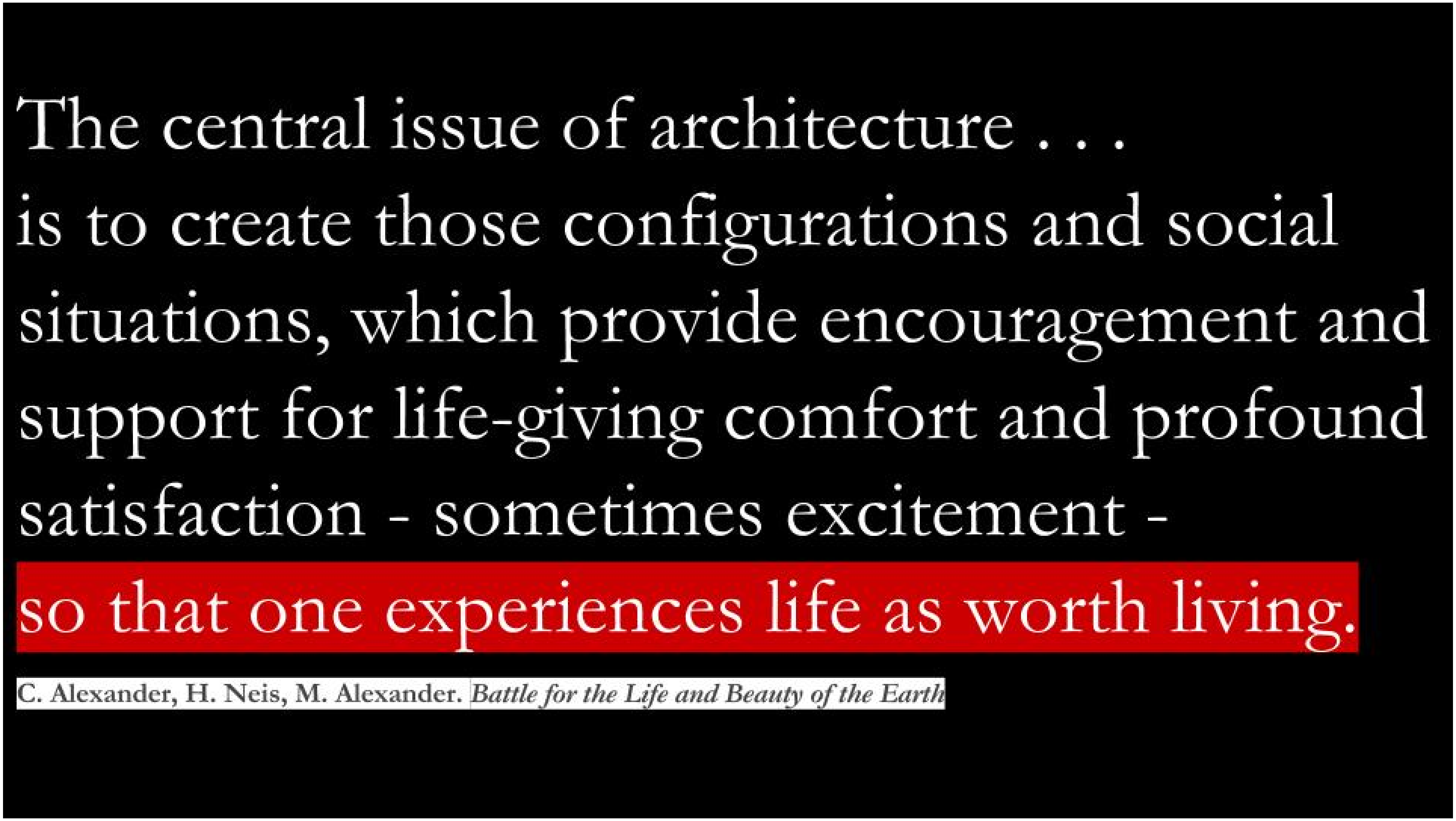
It’s a good question. Think on it. What are the stakes? Why are we here?

Many of us probably have zines our collections. We could make collection records for items like these available relatively easily the way the British Library has.

Many of us probably have works of art in our collections. Maybe not Van Gogh, but probably some form of artwork nonetheless. We could relatively easily make measurements of features of these images available, measurements that relate to brightness, hue, and saturation for example.

Some of us may have even have collections that can help us to make sense of the Paris attacks that occurred earlier in the year. Nick Ruest made available a Twitter dataset, which consists of more than ten million tweets captured during and after the event.
But back to the original question, what are the stakes?

I learned to identify the stakes, in architecture, which seems fitting. Like architects we are builders. We build infrastructure and collections. We support the formation of community through these efforts. The collections as data conversation is an extension of these commitments.

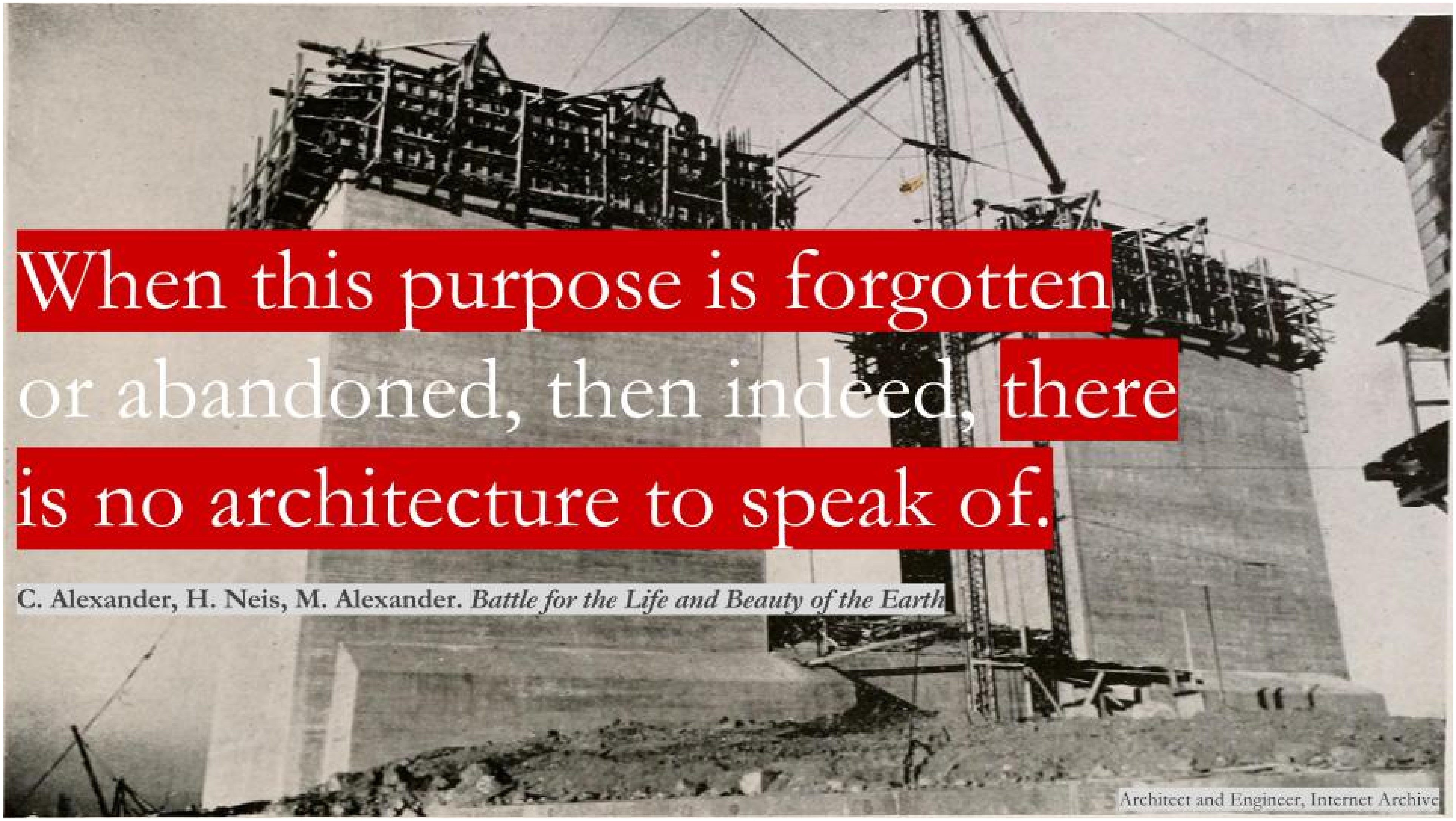


The central issue of architecture . . .
is to create those configurations and social situations, which provide encouragement and support for life-giving comfort and profound satisfaction - sometimes excitement -
so that one experiences life as worth living.

C. Alexander, H. Neis, M. Alexander. *Battle for the Life and Beauty of the Earth*

Though it is important, to emphasize in these efforts, that the conversation is not about data for data's sake, or computation for computation's sake – rather the work is intended to build upon our commitments to support nothing less than the ability to experience life as worth living.

This is not a new role for us. We have not been without imperfection in this role. But collections as data provides an opportunity to continue learning to do better.



When this purpose is forgotten or abandoned, then indeed, there is no architecture to speak of.

C. Alexander, H. Neis, M. Alexander. *Battle for the Life and Beauty of the Earth*

Architect and Engineer, Internet Archive

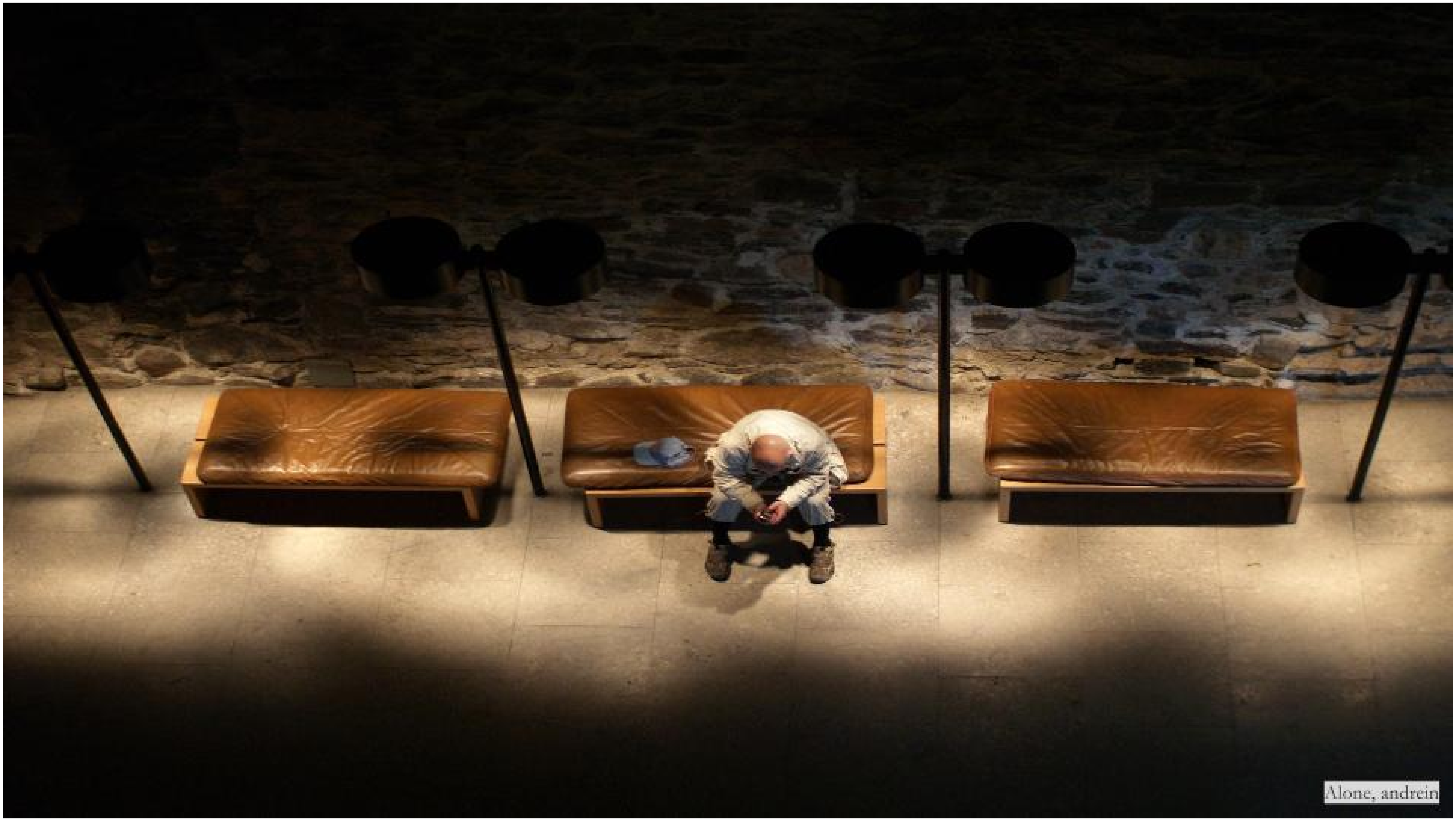
In my perspective, the first condition in the collections as data conversation centers on agency. To see collections as data is to step toward claiming agency – to extend individual capacity to act.

Agency

Empowerment

Ethics

I acknowledge the chance of initial alienation around this framing. Data after all, are not likely to be what the majority choose as a first conceptual frame for interacting with a book, an image, or a song.



In the face of the familiar, data may seem a cold path to discovery.

Yet, data are vibrant with possibility.

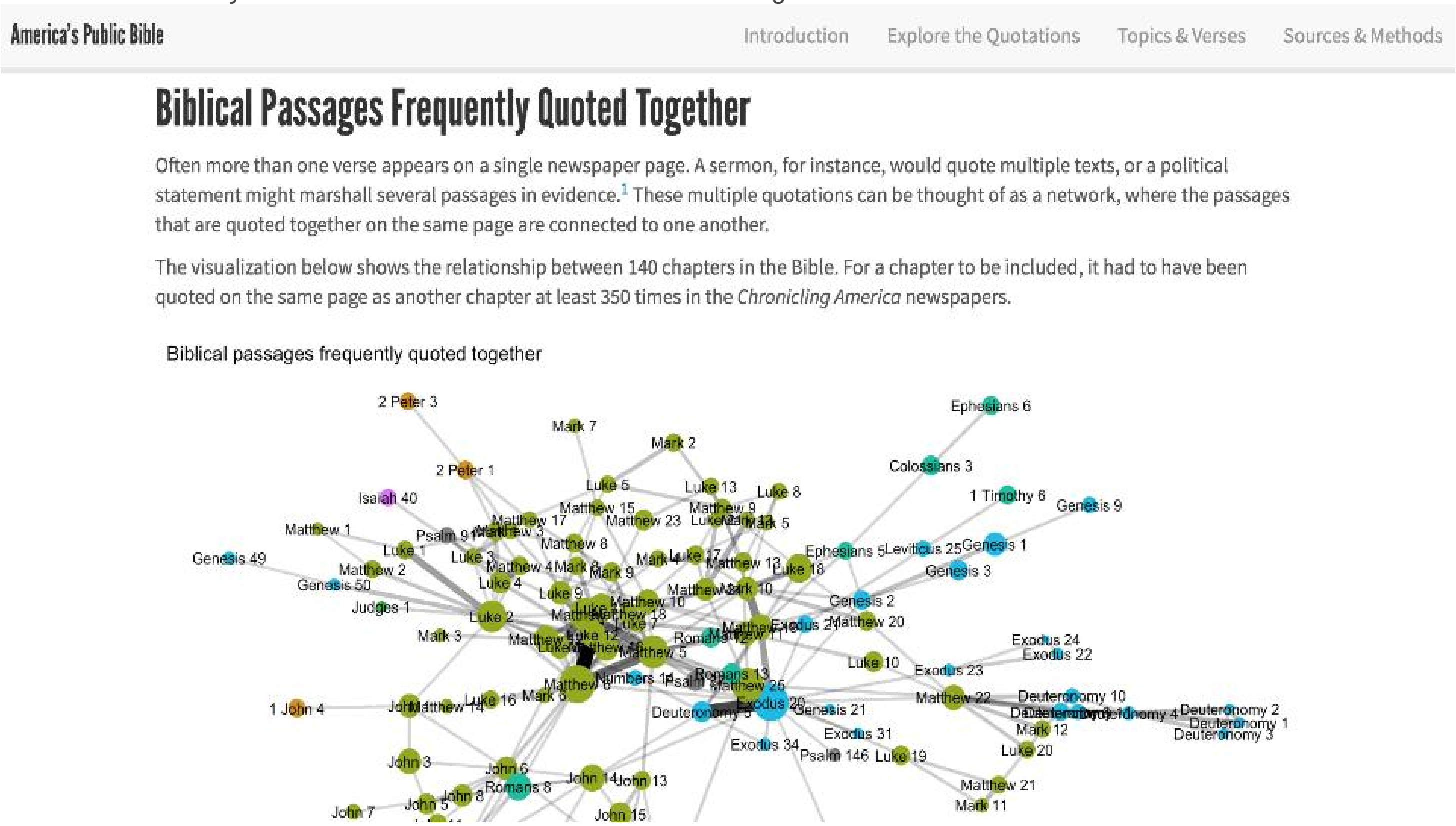
Hands, by vixenrose

They are the product of human design and world view. Reclamation of agency entails learning how to recognize, interpret, and act upon the facets of human intention in the data. By figuring collections as data we seek to make these facets known *and* usable.
There is more than one path to enabling the capacity to act with data, but I'll talk about one in the interest of time. This is a path that is generally functional in orientation, and depends on the notion of affordance. It's a pragmatic approach.

Basically I approach a seemingly familiar object and ask what functions it affords to help me resolve a question. I then transition to consider what additional functions it affords if I engage it as data – something comprised of vast numerical difference, often bound in codified systems that speak to each other across great distances of space and time. As this path is traversed, it is helpful to think on Martin Mueller's reminder that, "every surrogate has its own query potential, which for some purposes may exceed that of the original."

As a physical object, a newspaper readily affords certain types of functions. You can pick it up, fold it, unfold it, and generally orient yourself to a significant amount of heterogeneous content in a condensed space. Once digitized, the collection object has become data. In the Library of Congress project, Chronicling America, we see collections as data being leveraged to enhance the ability to act upon the data at micro and macroscopic levels. Shifting between these scales we witness a platform that extends familiar expectations for working with the source objects – the ability to browse, to search, to read. While the platform is committed to supporting these familiar functions, the designers of Chronicling America, in a nod toward the data centric possibilities, provisioned an API. The API has since supported a number of projects that explore the collection as data.

Spurred by a recent NEH funded contest, Lincoln Mullen leveraged machine learning to identify over 866,000 quotations and verbal allusions from the bible across 10.7 million pages of historical newspaper. Amidst nearly 56 billion words, Mullen not only identified the phenomena he was seeking, but further augmented the data by creating new linkages between it – indicating not only when and where bible verses occurred, but also discerning and subsequently explicitly structuring the data so that they indicate how often verses tended to occur together.



Biblical Passages Frequently Quoted Together

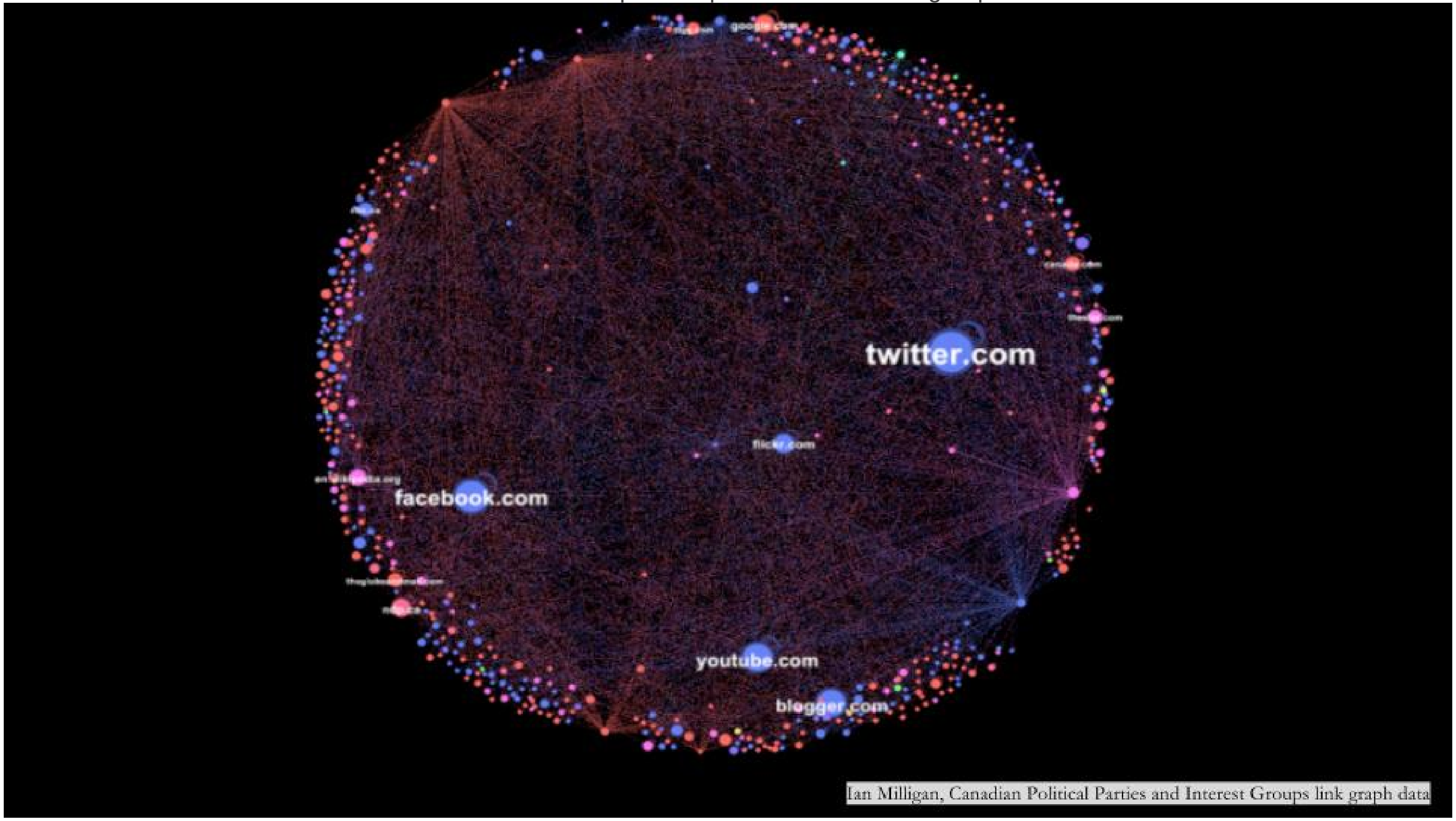
Often more than one verse appears on a single newspaper page. A sermon, for instance, would quote multiple texts, or a political statement might marshall several passages in evidence.[1] These multiple quotations can be thought of as a network, where the passages that are quoted together on the same page are connected to one another.

The visualization below shows the relationship between 140 chapters in the Bible. For a chapter to be included, it had to have been quoted on the same page as another chapter at least 350 times in the *Chronicling America* newspapers.

This act sheds a more complex view on use of the bible to support the formation of our varied imagined communities over time.

Of course, our collections are not limited to surrogates of physical objects. Increasingly, we collect so called “born digital” data. Web archive collections are significant efforts here at the Library of Congress, the Internet Archive, and throughout the world. A key component of the data orientation with respect to these collections is learning how to recognize what their nature affords for supporting exploration of questions. It will continue to be the case that we will have users that want to interact with these archives as a user may have interacted with them at the time of their creation through a web browser, a mobile device, and so forth.

Yet, if we peel back the layer, to engage the underlying, less visible structures organizing the representation we see on the screen it becomes possible to ask more questions of the collection as data. Consider Ian Milligan’s work on Canadian Political Parties and Interest Groups. In this network visualization, Milligan has drawn connections between all links that criss-cross Canadian political parties and interest group websites.



Ian Milligan, Canadian Political Parties and Interest Groups link graph data

From this vantage point, Milligan gains a sense of topology in the data, a multidimensional view of connectivity that may have not been readily apparent and/or intelligible without an awareness of the possibility latent in the data sitting below the surface of the familiar.

This push toward reaching past our familiar orientation to digital environments is what the collections as data conversation is all about. The deeper we go down this path, the better we serve our communities ability to peel back the layers covering seemingly mundane interactions in everyday life in order to recover an ability to act that they may have not known they have.

In order to do the work that lies ahead, we will need to be empowered to think differently. When I say we, I am referring to the builders at cultural heritage institutions large and small. When I say that we must be empowered, I refer to the concept in its truest sense as a move toward self-actualization.

I'll speak now from the perspective of libraries in particular, since that is the community I have the most familiarity with. There is a base claim to universality often touted in libraries. Universal collections with impartial aims. Services with universal scope.

Black box 3, by skeletalmess

In this manner, libraries are figured like transaction driven benevolent blackboxes.

But seriously, down with the blackbox.

Libraries are people.

Hands, by leonie_x

Individual hands, hopes, and dreams build the collections we use.

Let’s acknowledge that, celebrate it, and give credit where credit is due.

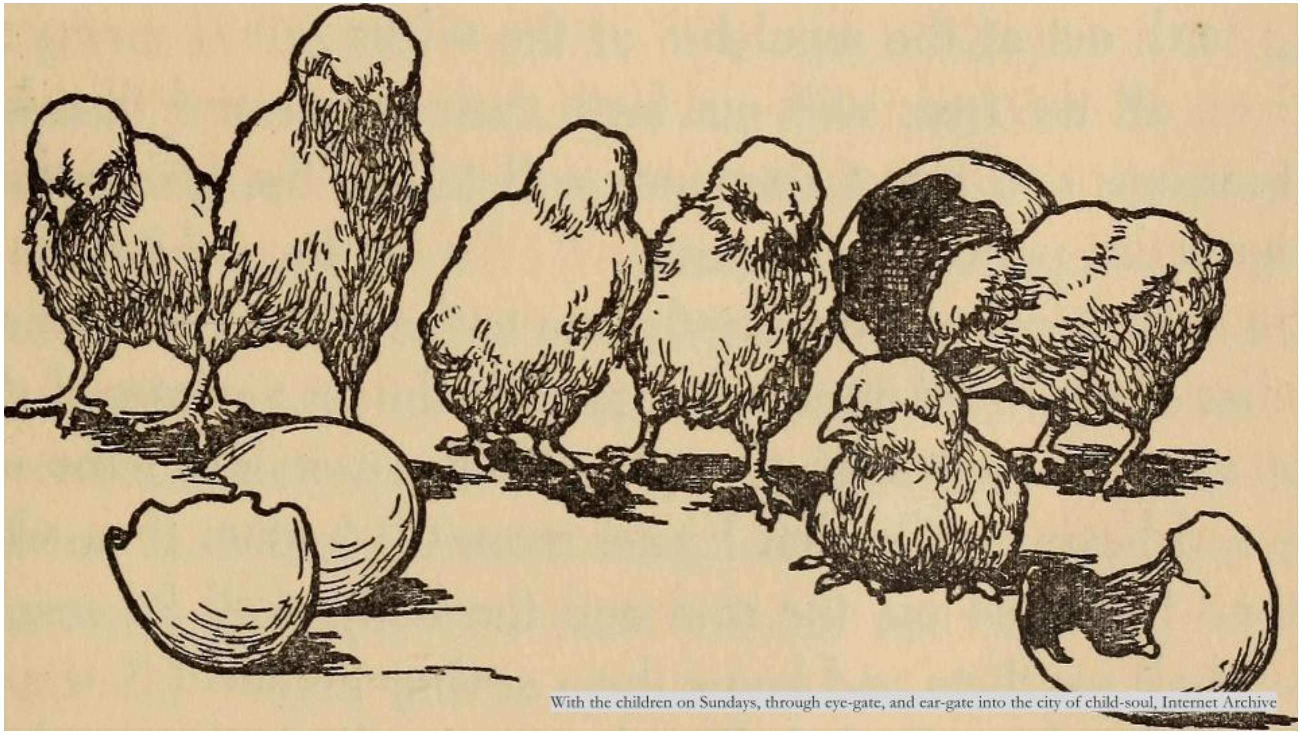
Collections as data provide an opportunity.

Who designed the schema? Who did the transcription? Who decided why the acquisition should be made and for whom? Why did a certain normalization occur? This is about recognizing the labor and the intellectual value of librarian contributions to crafting the materials with which a wide array of communities gain a sense of meaning in the world.

It is also about surfacing the design decisions in order to give our collections their fullest integrity, so that they can actually be used to substantiate claims based on them.

There are many directions the collections as data project could go. In almost any direction, we will bump into preexisting norms for going about this work. There will undoubtedly be expectations couched in terms of scalability.

To be honest I am kinda tired of the scalability framing. It’s like a chicken or the egg debate where someone keeps insisting that chickens sprang from an ur-chicken forehead and won’t consider the possibility of the egg. That both could actually happen in parallel, that they might actually need to happen that way in order to push the conversation forward.



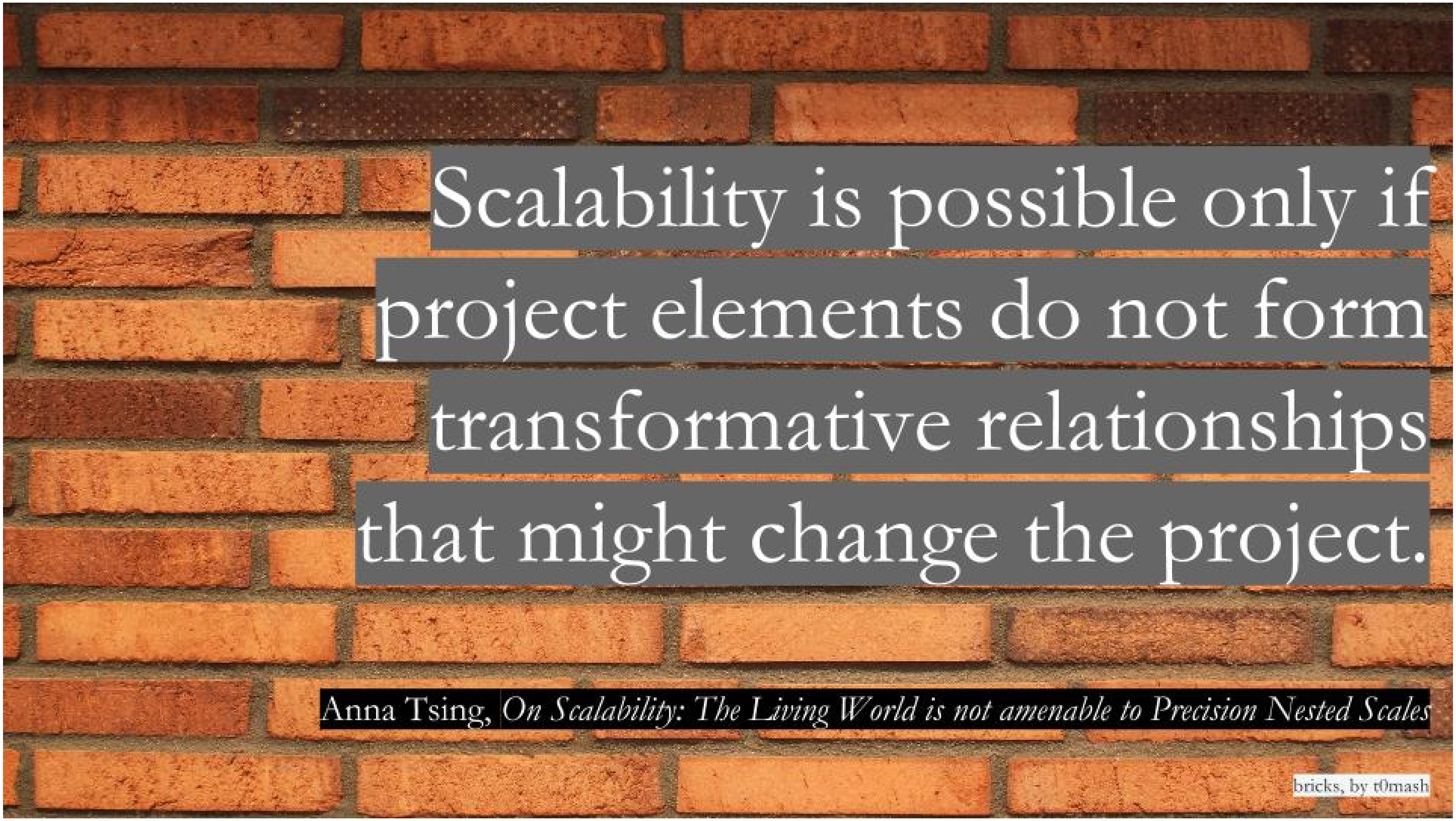
With the children on Sundays, through eye-gate, and ear-gate into the city of child-soul, Internet Archive

A quick sidebar – I want to thank Katie Rawson and Trevor Munoz for introducing me to Anna Tsing's, *On Scalability: The Living World is not amenable to Precision Nested Scales* in their excellent piece *Against Cleaning*. I highly recommend it. In what follows I will quote Tsing because she really gets at the notion of scalability I am focusing on.

When small projects become big without changing the nature of the project, we call that design feature 'scalability'.

Anna Tsing, *On Scalability: The Living World is not amenable to Precision Nested Scales*

Grotto in an iceberg, National Library NZ



Scalability is possible only if project elements do not form transformative relationships that might change the project.

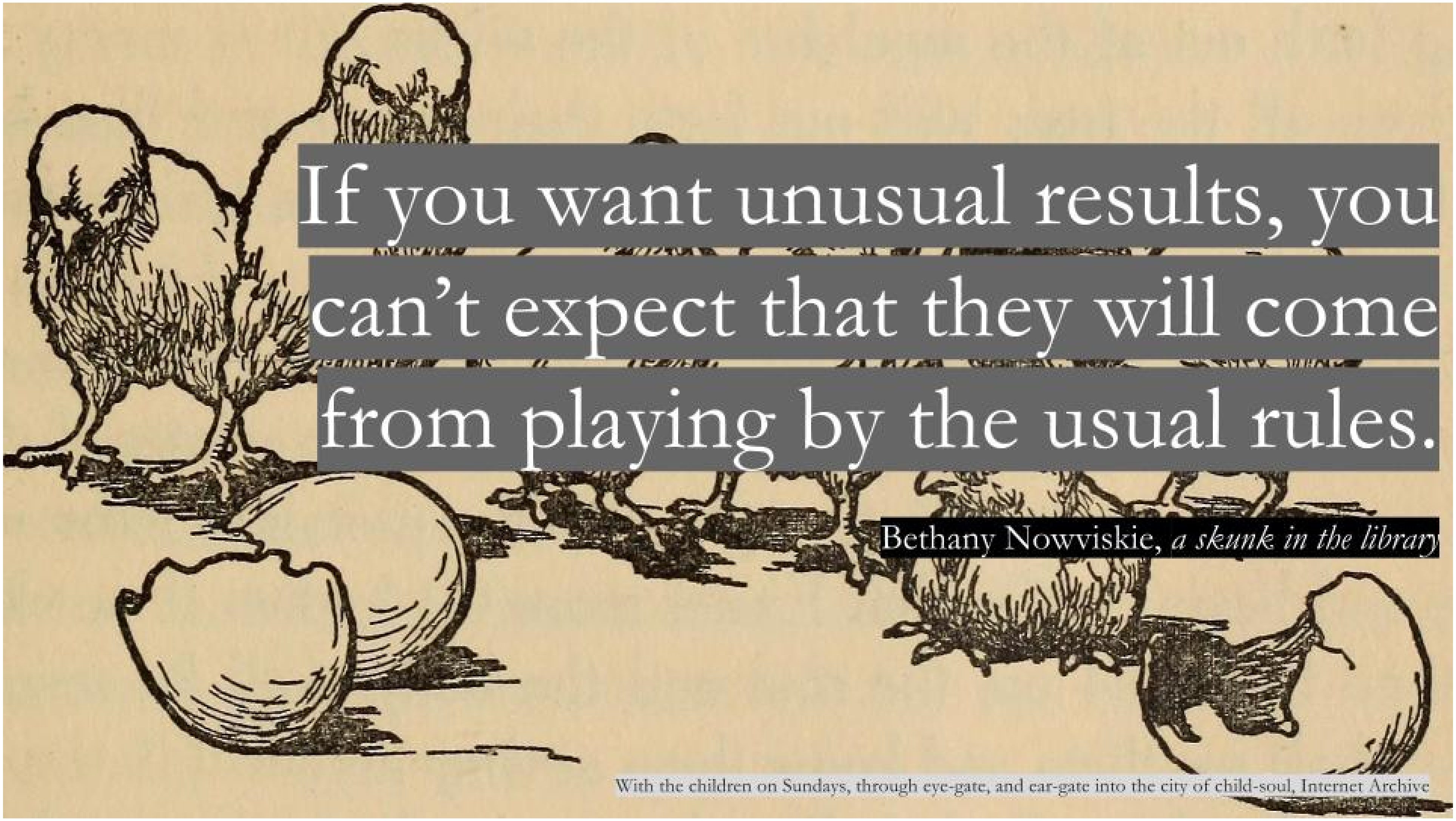
Anna Tsing, *On Scalability: The Living World is not amenable to Precision Nested Scales*

bricks, by t0mash

But transformative relationships are the medium for the emergency of diversity.

Anna Tsing, *On Scalability: The Living World is not amenable to Precision Nested Scales*

Grotto in an iceberg, National Library NZ



If you want unusual results, you can't expect that they will come from playing by the usual rules.

Bethany Nowviskie, *a skunk in the library*

With the children on Sundays, through eye-gate, and ear-gate into the city of child-soul, Internet Archive

A couple of issues typically arise as librarians aim to extend their commitments in the collections as data space. First, it is often expected that new ground is broken yet the potential of the work is mitigated from the outset by attempting to predicate effort on a scalability paradigm.

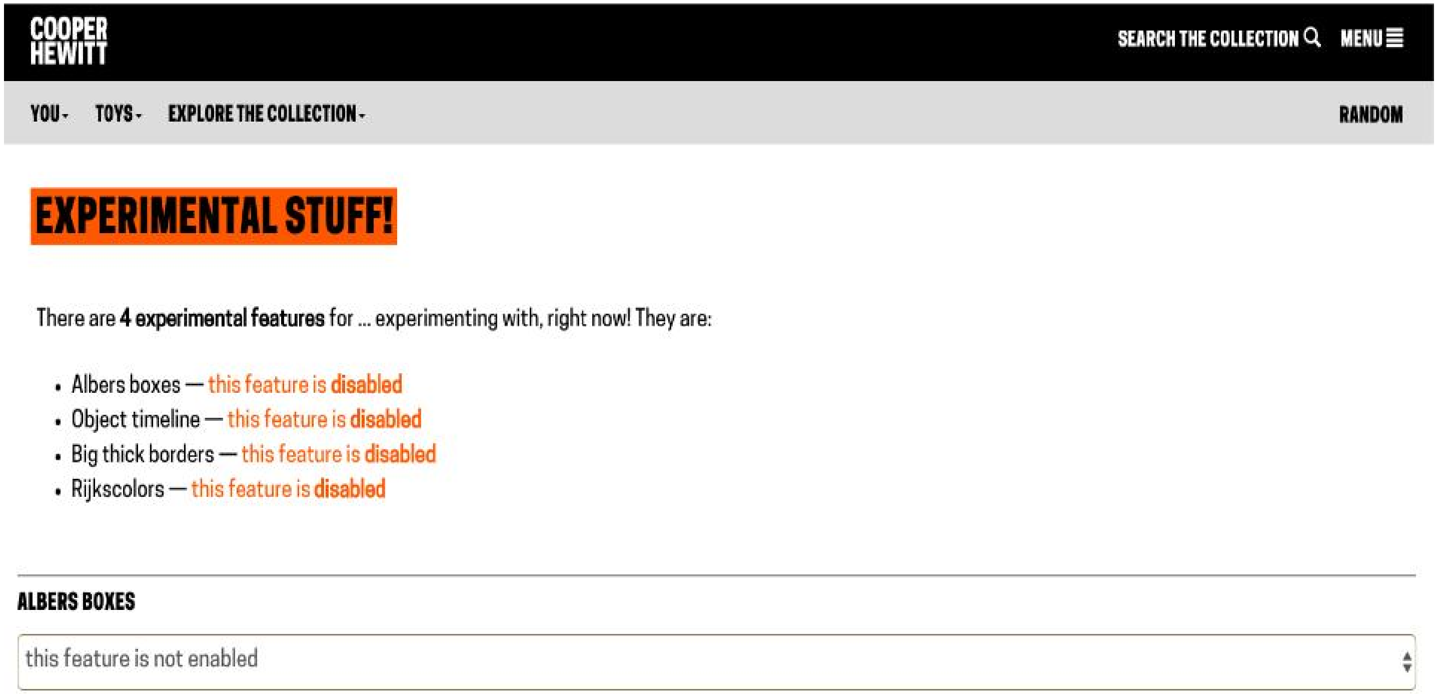
Scalability cannot be the sole precondition of possibility in the collections as data space. In order for diverse solutions to occur we must learn to embrace experimentation that accommodates and even embraces the value of failure as equally as success.

Some may say that the ability to engage this span of outcomes constitutes a luxury, a privileged position. I would say to look to liberal arts colleges, often times dwarfed by the R1 research universities throughout the country. They are doing some of the most interesting work in this space and with vastly limited resources comparatively.

The second issue, impeding forward progress relates to time. As previously mentioned it is often the case that there is an expectation that new ground is broken but little is done administratively to free up individual time to contribute to new projects. I find more often than not that personal interest or excitement, what you might call the preconditions of empowerment, are not the barrier to doing collections as data work – rather its simply a question of time whose lack of resolution can often be traced to administrative mismatch between goals and reality.

Simply put, lets free people up to empower themselves in their work, whether thats collections as data or some other new initiative.

I’ve long admired the work that the Cooper Hewitt has done with collections as data. They are enthusiastic and very public about their experiments. They actually have a menu on their website called toys, with an experiment section. There are currently four experiments listed, all disabled. Does this mean they failed? Could we consider the possibility that they may have been meant to fail? That that may actually be a good thing?



EXPERIMENTAL STUFF!

There are **4 experimental features** for ... experimenting with, right now! They are:

- Albers boxes — this feature is **disabled**
- Object timeline — this feature is **disabled**
- Big thick borders — this feature is **disabled**
- Rijkscolors — this feature is **disabled**

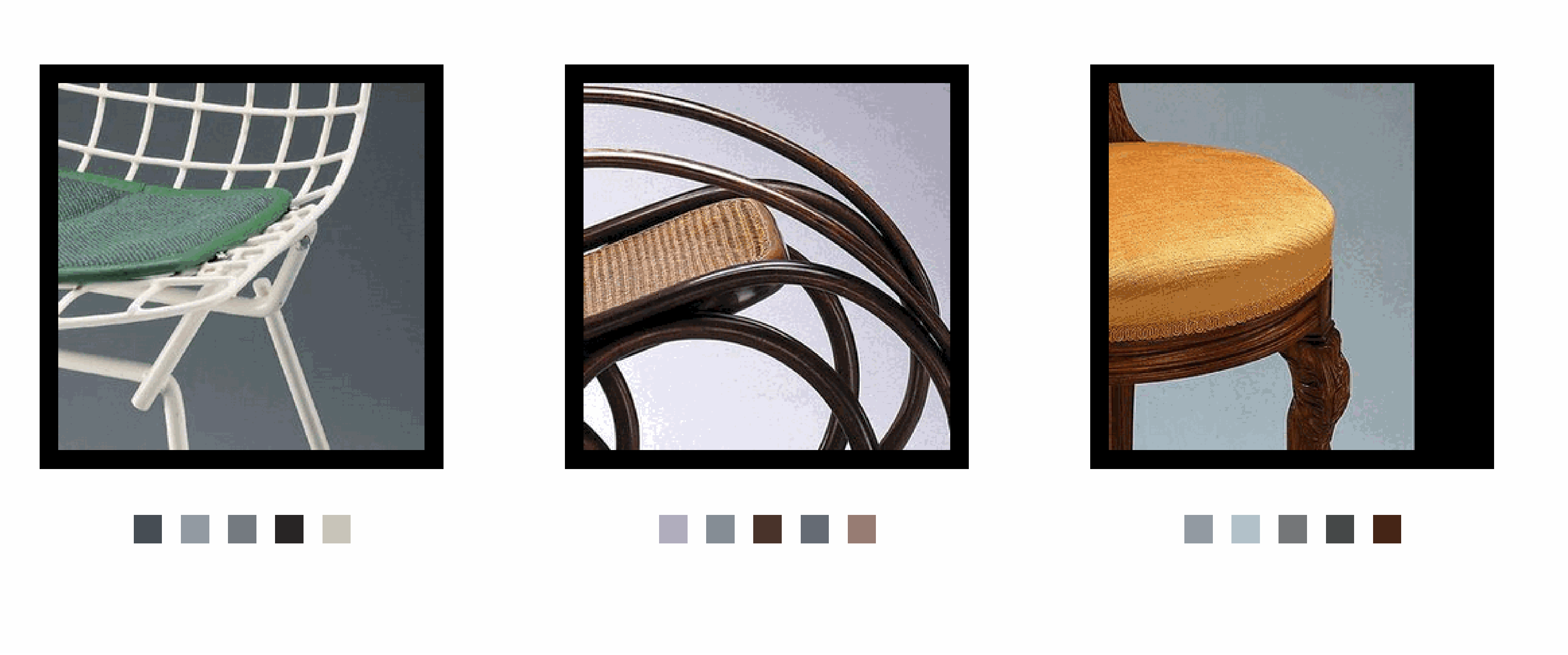
ALBERS BOXES

this feature is not enabled

As we move out of the experimental section of the toys menu to the exploration menu we are presented with a dizzying array of ways to engage the collection. Search by color, you got it. Search by concordance, why not.

I don’t know with 100% certainty but I’d bet that more than one of these features started in the mind of someone who felt empowered to explore – that exploration was run as an experiment, disabled, and then walked into the primary collection interface.

Here we witness one of these experiments in action. What we are seeing is a novel solution to conveying information visually about objects of widely varying sizes. When presented with a photo of a chair at a distance of 10 feet and a pencil at a distance of 1 foot, how do you go about presenting images that present the user with an equivalent amount of information visually to decide whether the object interests them? In this case, you write a python script that calculates the Shannon Entropy value for every image, which in turn enables cropping images at the optimal place to generate thumbnails that convey the most amount of information to the user.



The British Library Labs has been doing fantastic work in this space, for years now. Folks like Nora McGregor, Mia Ridge, and formerly James Baker. They are an enterprising group, engaged in teaching digital scholarship within their institution, and working toward releasing a whole slew of different types of collections as data. You may have heard of their pioneering work, automatically extracting a million images from historical text, and making them available via Flickr under a cc-0 license.

Subsequently, they have encouraged a number of competitions for working with the collections that they release as data. These efforts are not purely about publicity. They provide opportunities to refine the way that we prepare and provide access to collections, and can lead to concrete reciprocal benefits from outside our institutions.

For example, Mario Klingemann's experimentation with machine learning and semi-automated image classification to generate additional metadata for the Flickr image release, as well as enabling further subsetting of the data – for example leading to the creation of smaller datasets comprised of portrait images.

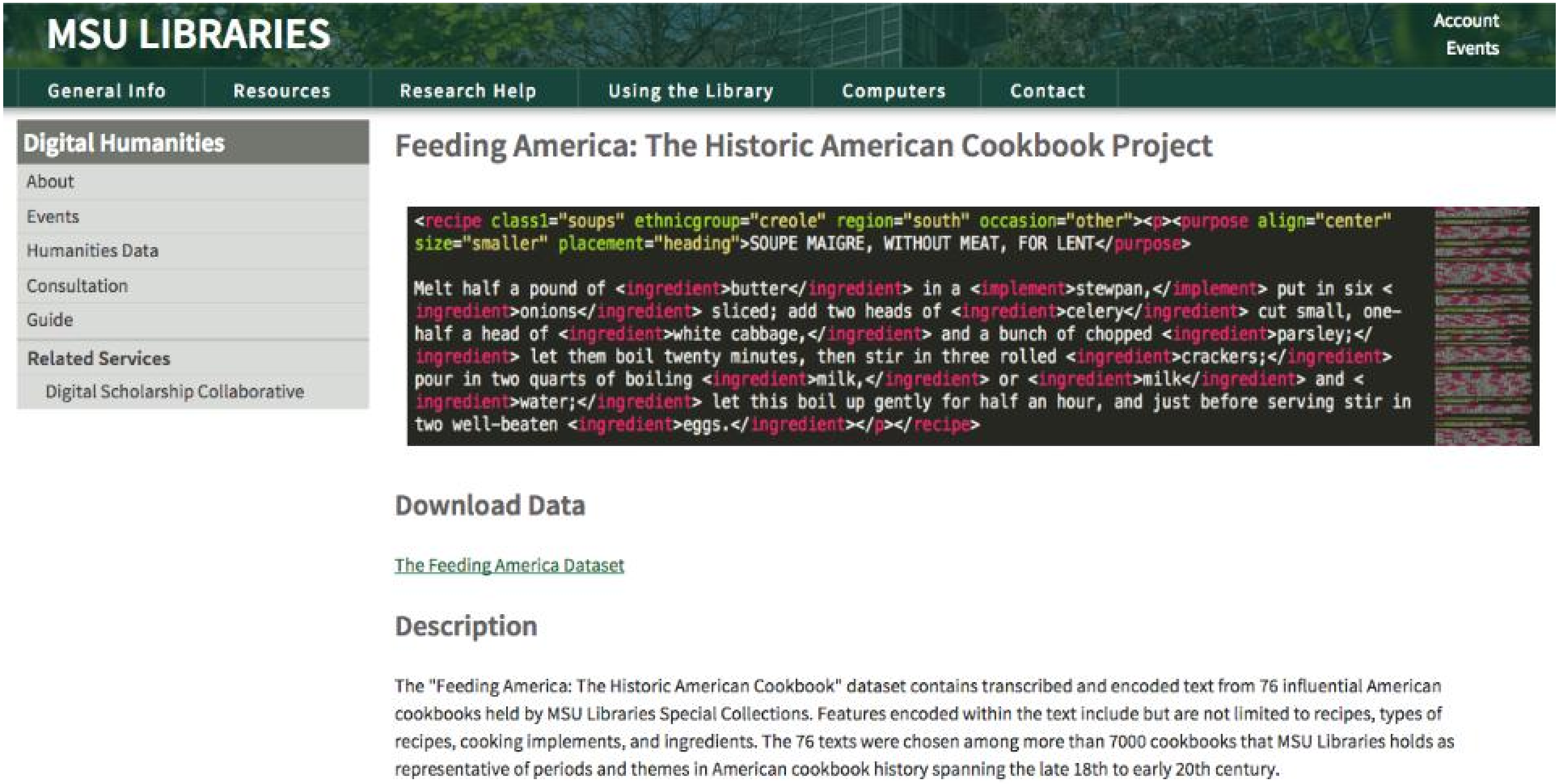
16 x 16 Colourful Faces from the British Library Collection, by quasimondo

There have been a number of academic libraries in North America working in this space as well. I am formerly of Michigan State University, the green helmet there on the top left. While collections differed in scope among these institutions, we shared in common a desire to release our collections as data in order to encourage additional types of use.

Penn
Libraries
UNC
UNIVERSITY LIBRARIES

The process in my case was fairly straightforward – identify existing digital special collections. Evaluate the possibilities latent in the data. Are they unstructured or structured? What is the quality of OCR? Is it transcribed? How representative is the data?

The next step was to create derivatives that served anticipated use. So out with the JPG thumbnails, in with the PDFs, tei-encoded files, and plain text derivatives. Wrap all of the above in new documentation that supports computational use.



Feeding America: The Historic American Cookbook Project

```
<recipe class1="soups" ethnicgroup="creole" region="south" occasion="other"><p><purpose align="center"
size="smaller" placement="heading">SOUPE MAIGRE, WITHOUT MEAT, FOR LENT</purpose>

Melt half a pound of <ingredient>butter</ingredient> in a <implement>stewpan,</implement> put in six <
ingredient>onions</ingredient> sliced; add two heads of <ingredient>celery</ingredient> cut small, one-
half a head of <ingredient>white cabbage,</ingredient> and a bunch of chopped <ingredient>parsley;</
ingredient> let them boil twenty minutes, then stir in three rolled <ingredient>crackers;</ingredient>
pour in two quarts of boiling <ingredient>milk,</ingredient> or <ingredient>milk</ingredient> and <
ingredient>water;</ingredient> let this boil up gently for half an hour, and just before serving stir in
two well-beaten <ingredient>eggs.</ingredient></p></recipe>
```

Download Data

The Feeding America Dataset

Description

The "Feeding America: The Historic American Cookbook" dataset contains transcribed and encoded text from 76 influential American cookbooks held by MSU Libraries Special Collections. Features encoded within the text include but are not limited to recipes, types of recipes, cooking implements, and ingredients. The 76 texts were chosen among more than 7000 cookbooks that MSU Libraries holds as representative of periods and themes in American cookbook history spanning the late 18th to early 20th century.

Subsequently this data was used to support digital humanities pedagogy, became a test dataset in a text analysis program, and was sought after to support a computational analysis of ingredient usage in cooking throughout American History. Had we not treated this collection as data and promoted it as such it’s not likely that it would have seen these types of use.

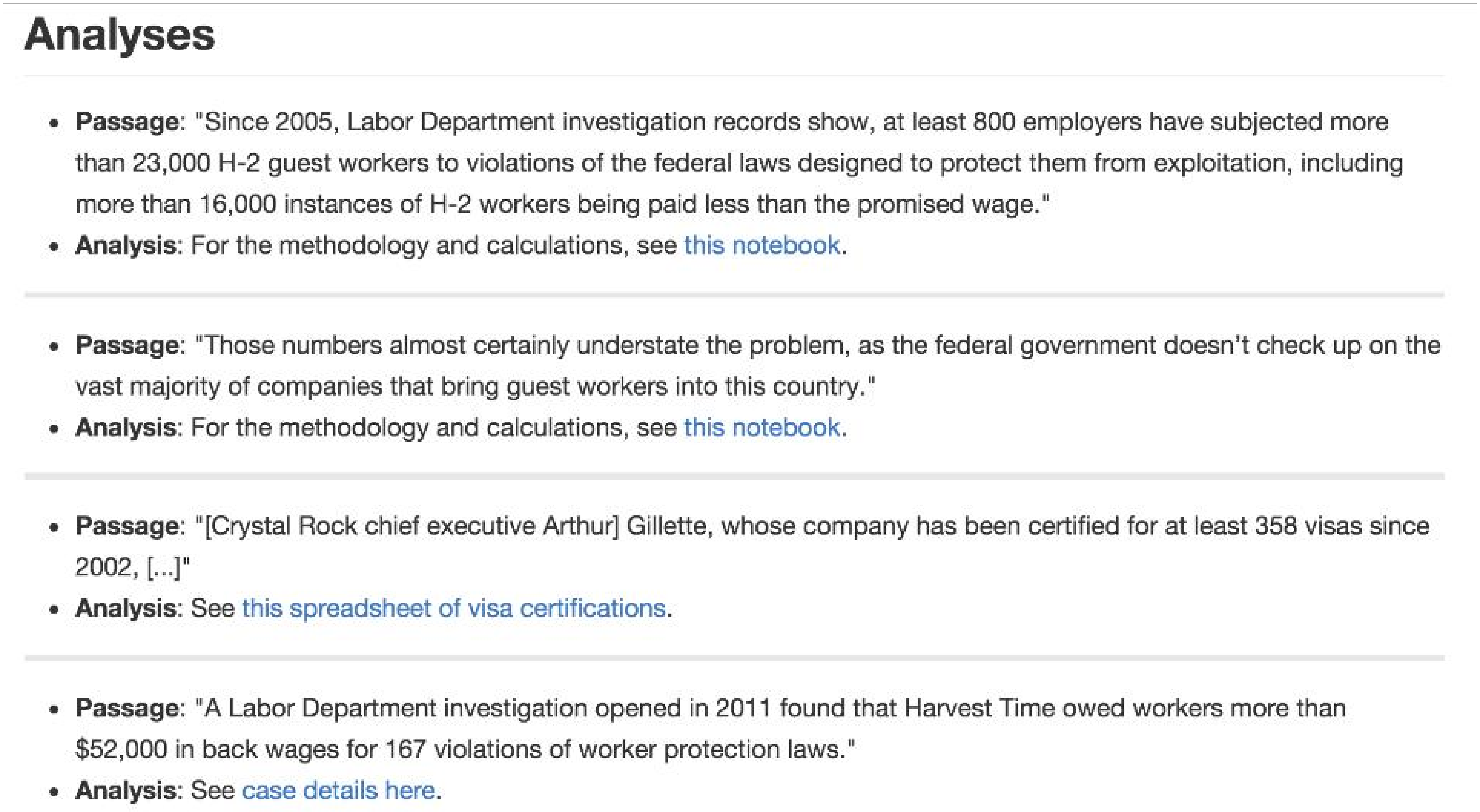
Work of this kind only becomes possible when individual people feel empowered to pursue their passions, to self-actualize through them. The cost of that self-actualization is time and the confidence to pursue projects, whether they succeed or fail. Really, failure should be a goal.

Hand study, by nicdalic

That said, embracing failure doesn't mean doing our work without attention to ethics. We must commit to transparency, inclusivity, and respect in the work that lies ahead. A number of ethical considerations are raised. Some of these are new. Others are familiar, yet possibly more complex in light of the medium.

The first issue relates to transparency in the process of collection building. How were the data processed? What was left in what was left out? Why? How representative is the data? What bias does our organization of the data reflect? Can we admit a bias?

I think for a number of reasons that data driven journalism has some qualities that we can learn from as we consider these questions. Recently, journalists at Buzzfeed supplemented an article, *New American Slavery*, with a blog post that links to data, methods, analysis, and code used to produce the work.



Analyses

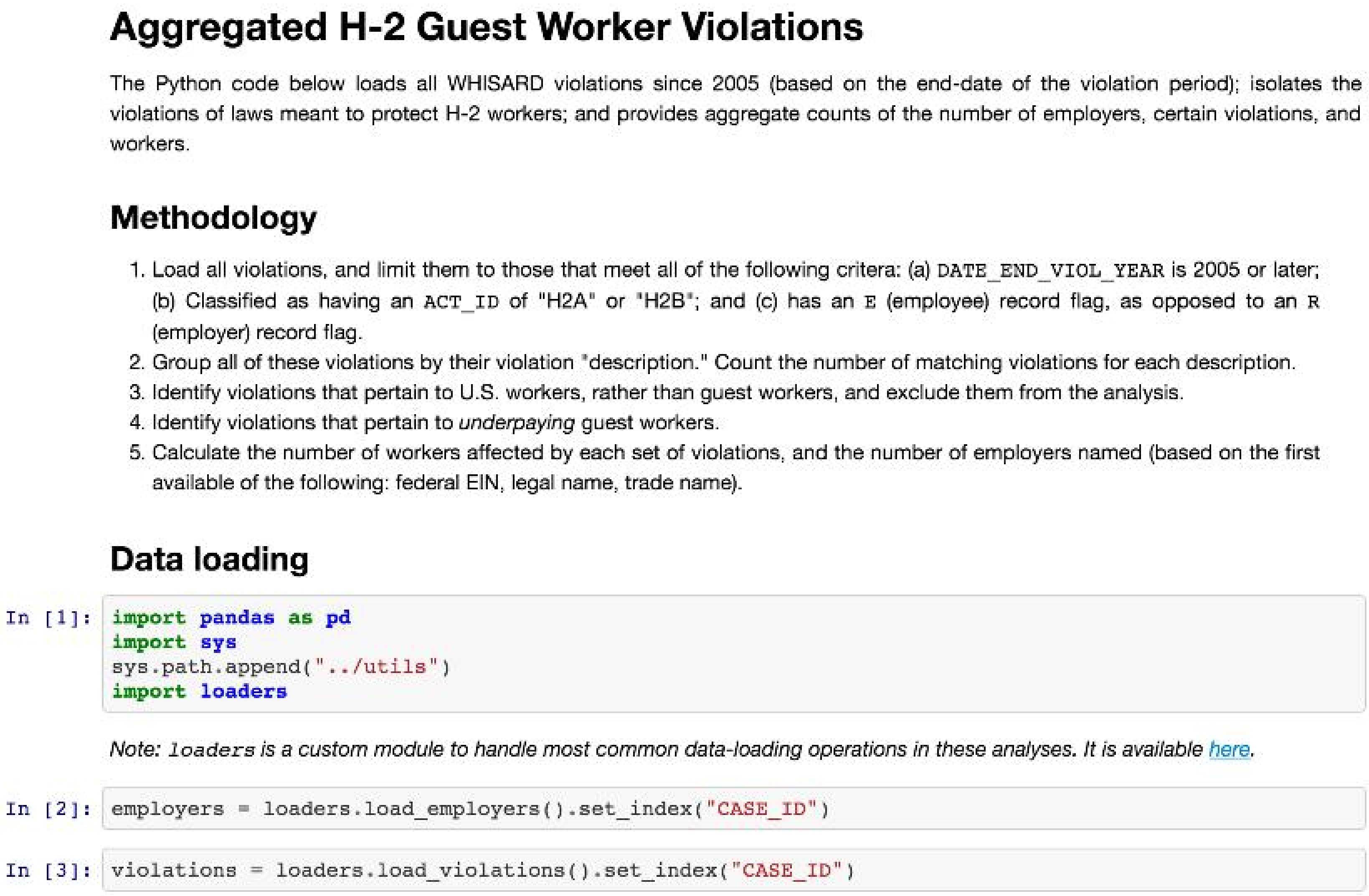
- **Passage**: "Since 2005, Labor Department investigation records show, at least 800 employers have subjected more than 23,000 H-2 guest workers to violations of the federal laws designed to protect them from exploitation, including more than 16,000 instances of H-2 workers being paid less than the promised wage."
- **Analysis**: For the methodology and calculations, see this notebook.

- **Passage**: "Those numbers almost certainly understate the problem, as the federal government doesn't check up on the vast majority of companies that bring guest workers into this country."
- **Analysis**: For the methodology and calculations, see this notebook.

- **Passage**: "[Crystal Rock chief executive Arthur] Gillette, whose company has been certified for at least 358 visas since 2002, [...]"
- **Analysis**: See this spreadsheet of visa certifications.

- **Passage**: "A Labor Department investigation opened in 2011 found that Harvest Time owed workers more than $52,000 in back wages for 167 violations of worker protection laws."
- **Analysis**: See case details here.

We see that the level of transparency is quite granular. The writers go down to the level of individual claims, with links to the ipython notebook that contains the code that generated the data to support the claim.



Aggregated H-2 Guest Worker Violations

The Python code below loads all WHISARD violations since 2005 (based on the end-date of the violation period); isolates the violations of laws meant to protect H-2 workers; and provides aggregate counts of the number of employers, certain violations, and workers.

Methodology

1. Load all violations, and limit them to those that meet all of the following critera: (a) `DATE_END_VIOL_YEAR` is 2005 or later; (b) Classified as having an `ACT_ID` of "H2A" or "H2B"; and (c) has an `E` (employee) record flag, as opposed to an `R` (employer) record flag.
2. Group all of these violations by their violation "description." Count the number of matching violations for each description.
3. Identify violations that pertain to U.S. workers, rather than guest workers, and exclude them from the analysis.
4. Identify violations that pertain to *underpaying* guest workers.
5. Calculate the number of workers affected by each set of violations, and the number of employers named (based on the first available of the following: federal EIN, legal name, trade name).

Data loading

```
In [1]: import pandas as pd
        import sys
        sys.path.append("../utils")
        import loaders
```

Note: `loaders` is a custom module to handle most common data-loading operations in these analyses. It is available here.

```
In [2]: employers = loaders.load_employers().set_index("CASE_ID")
```

```
In [3]: violations = loaders.load_violations().set_index("CASE_ID")
```

As I cast my eye over this approach, I wonder what it can teach us in libraries as we work to provide collections as data.

Will our users expect similar levels of transparency in the documentation of computational processes and analytic decisions that were made to generate the collections that we provide access to? They may. Transparency in this vein may be a precondition of viable use.

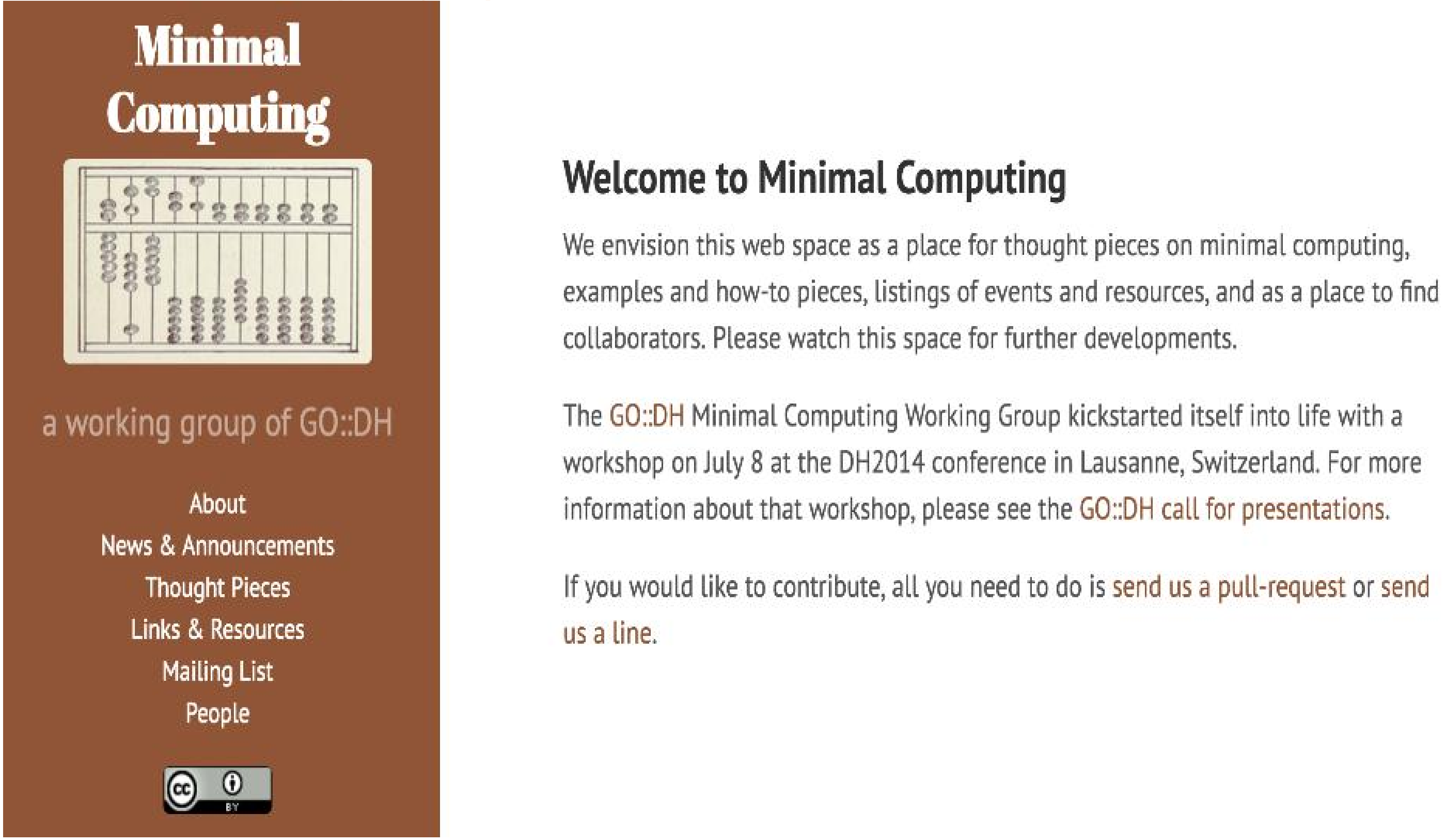
As collections as data are developed we cannot lose sight of interest beyond North American borders. We must have an inclusive vision. Global Outlook::Digital Humanities, affectionately referred to as GODH reminds us of this pointedly.

There is a wider world to engage with.

We should think deeply about how we can best interact.

Part of that consideration is linguistic.

Part of that consideration speaks to questioning assumptions of normative access to computational infrastructure.



Welcome to Minimal Computing

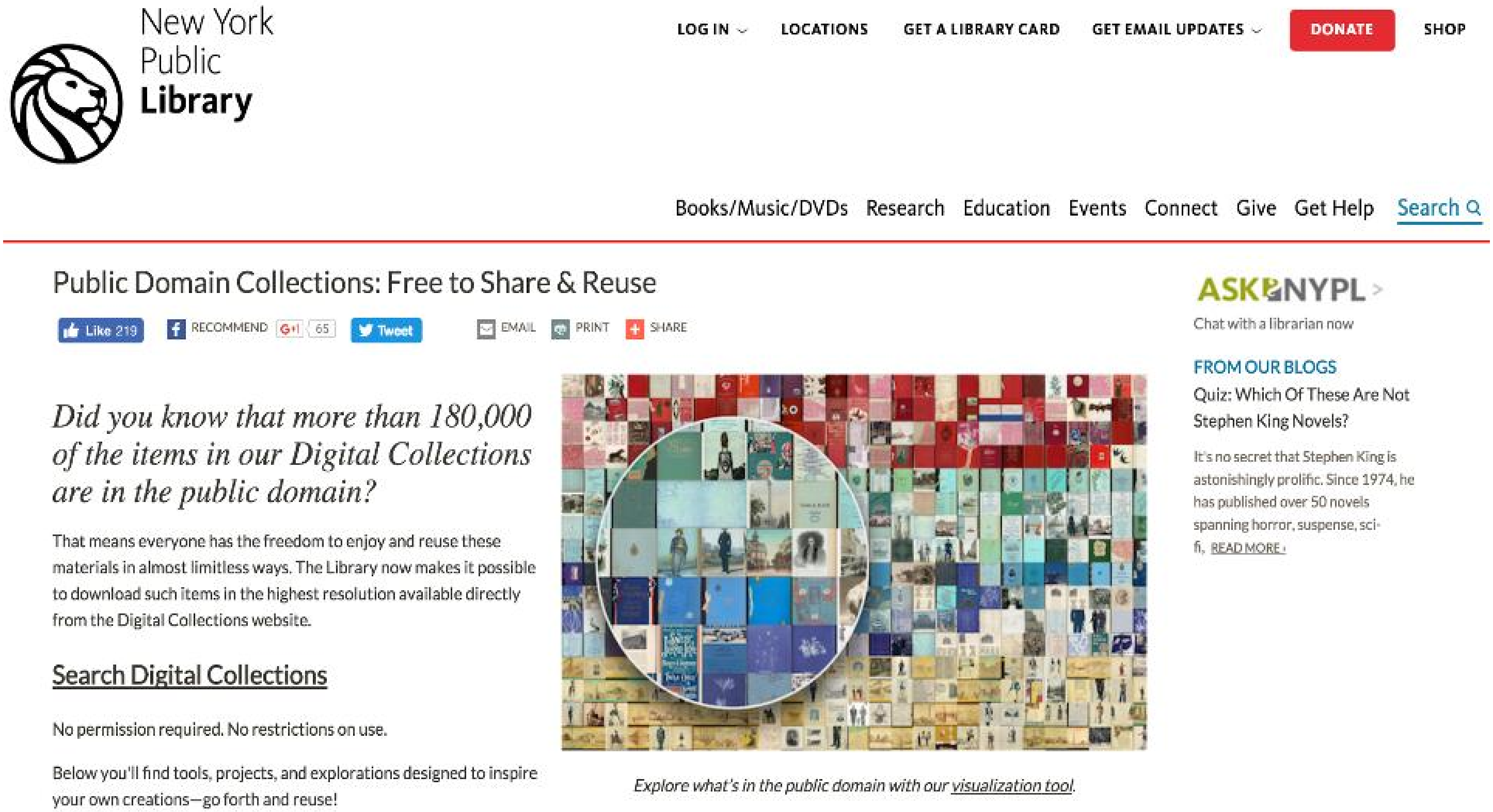
We envision this web space as a place for thought pieces on minimal computing, examples and how-to pieces, listings of events and resources, and as a place to find collaborators. Please watch this space for further developments.

The GO::DH Minimal Computing Working Group kickstarted itself into life with a workshop on July 8 at the DH2014 conference in Lausanne, Switzerland. For more information about that workshop, please see the GO::DH call for presentations.

If you would like to contribute, all you need to do is send us a pull-request or send us a line.

Through GODH's experiments with minimal computing we can discern a model for encouraging participation in this collections as data conversation that is more expansive than it would be otherwise.

Across the board, we require acts of bravery large and small to make our collections as open as possible. Few things make the light of curiosity die more quickly in someone's eyes when they shift to the words, "rights assessment is your responsibility". This is something I've witnessed personally, from primary school teachers seeking to use historical photos in their classrooms to university researchers interested in conducting large scale analysis of historical collections.



Public Domain Collections: Free to Share & Reuse

Did you know that more than 180,000 of the items in our Digital Collections are in the public domain?

That means everyone has the freedom to enjoy and reuse these materials in almost limitless ways. The Library now makes it possible to download such items in the highest resolution available directly from the Digital Collections website.

Search Digital Collections

No permission required. No restrictions on use.

Below you'll find tools, projects, and explorations designed to inspire your own creations—go forth and reuse!

Explore what's in the public domain with our visualization tool.

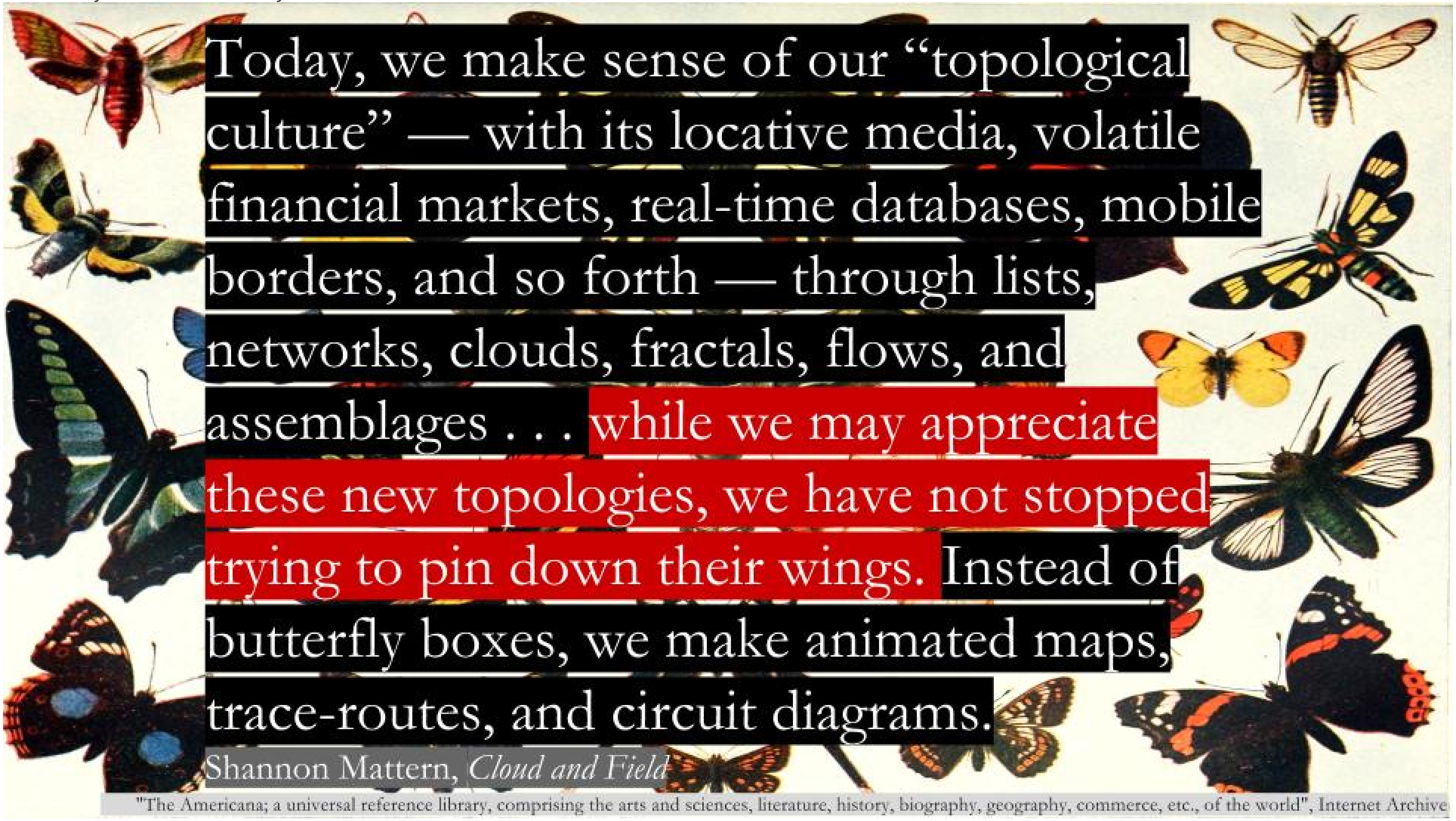
FROM OUR BLOGS

Quiz: Which Of These Are Not Stephen King Novels?

It's no secret that Stephen King is astonishingly prolific. Since 1974, he has published over 50 novels spanning horror, suspense, sci-fi, READ MORE ›

NYPL is a shining light in the open collections space that I am very grateful for. More open collections equal more access, more access equals a more inclusive range of use.
Now, this is very important.

Collections as data represent lived experience and we should respect that. Traces of human activity are not butterflies to be collected, cataloged, and pinned in 21st century cabinets of curiosity.

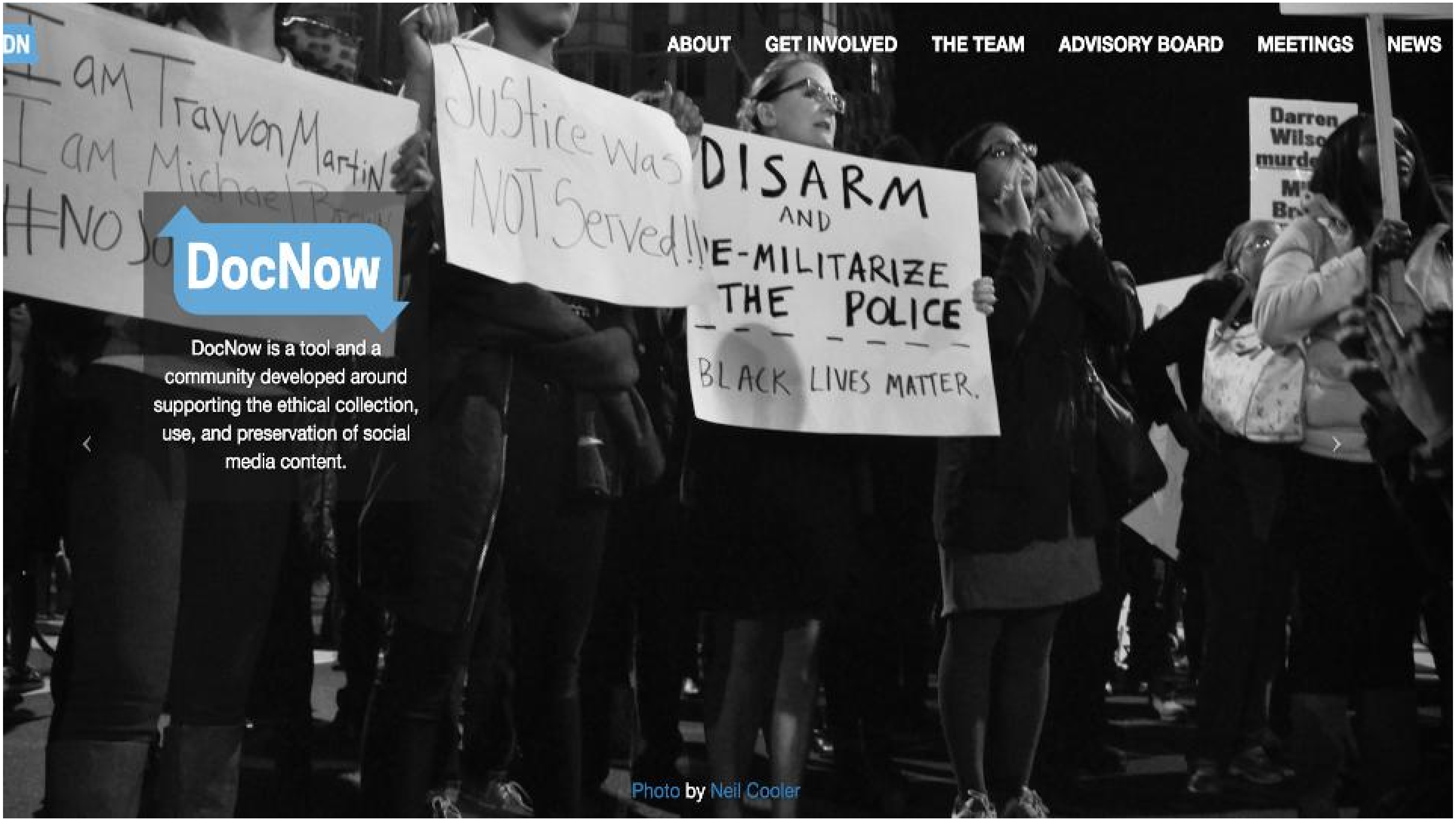


Today, we make sense of our “topological culture” — with its locative media, volatile financial markets, real-time databases, mobile borders, and so forth — through lists, networks, clouds, fractals, flows, and assemblages . . . while we may appreciate these new topologies, we have not stopped trying to pin down their wings. Instead of butterfly boxes, we make animated maps, trace-routes, and circuit diagrams.

Shannon Mattern, *Cloud and Field*

"The Americana; a universal reference library, comprising the arts and sciences, literature, history, biography, geography, commerce, etc., of the world", Internet Archive

Documenting the Now focuses on capturing a liminal form of data – our Tweets.

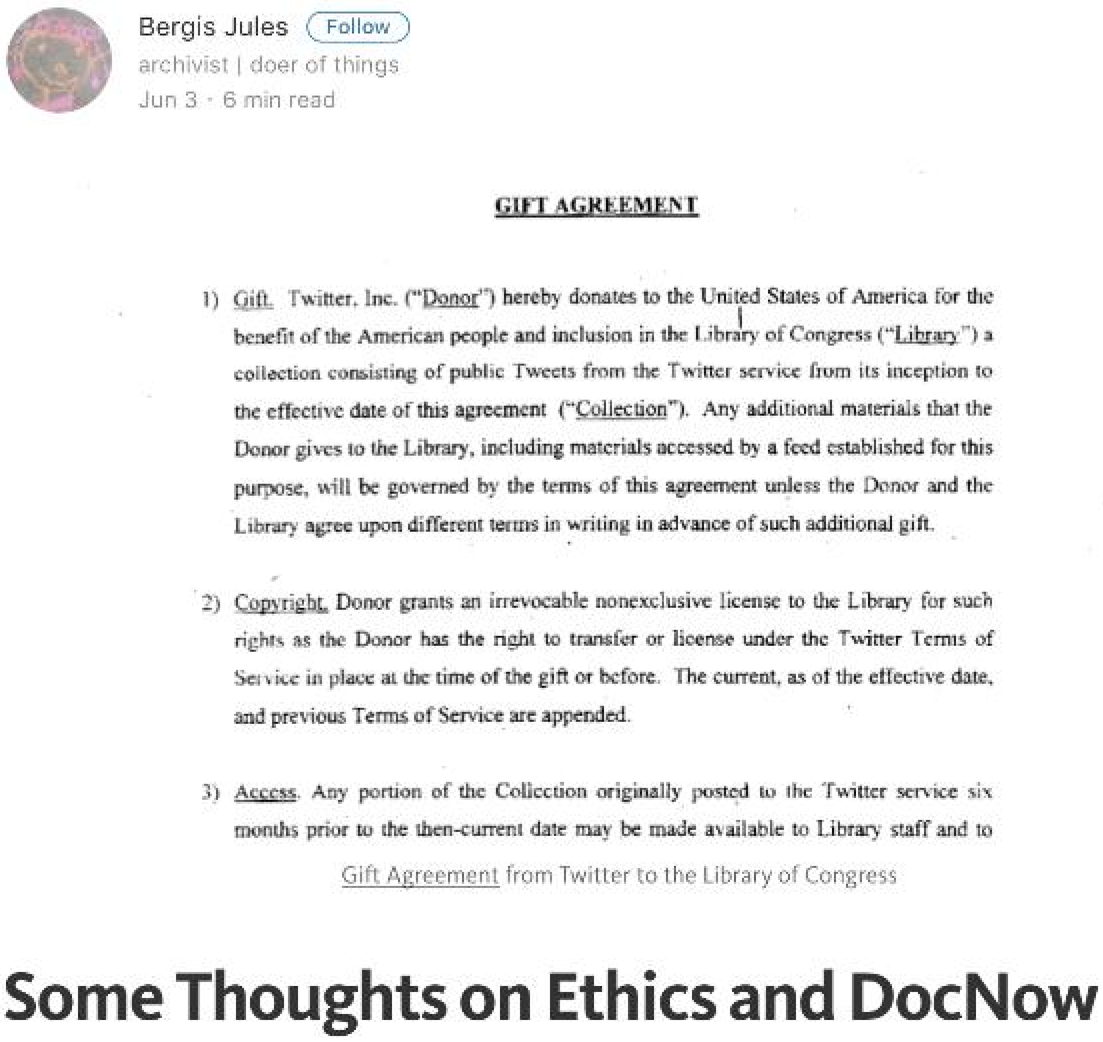


DocNow is a tool and a community developed around supporting the ethical collection, use, and preservation of social media content.

Photo by Neil Cooler

While the technical development is impressive, it also presents an approach that models how we might engage in collections as data activity that is expressly predicated on sustained attempts to ground the act of collection in community need.

Documenting the Now approaches this angle of work in a few different ways, but I'm particularly intrigued by Bergis Jules attempt to map precedent in archival practice, codified in the deed of gift, to gain a measure of purchase over the challenges that lay ahead – particularly as they pertain to structuring agreements between data producers and data collectors.



Bergis Jules Follow

archivist | doer of things

Jun 3 · 6 min read

GIFT AGREEMENT

1) Gift. Twitter, Inc. ("Donor") hereby donates to the United States of America for the benefit of the American people and inclusion in the Library of Congress ("Library") a collection consisting of public Tweets from the Twitter service from its inception to the effective date of this agreement ("Collection"). Any additional materials that the Donor gives to the Library, including materials accessed by a feed established for this purpose, will be governed by the terms of this agreement unless the Donor and the Library agree upon different terms in writing in advance of such additional gift.

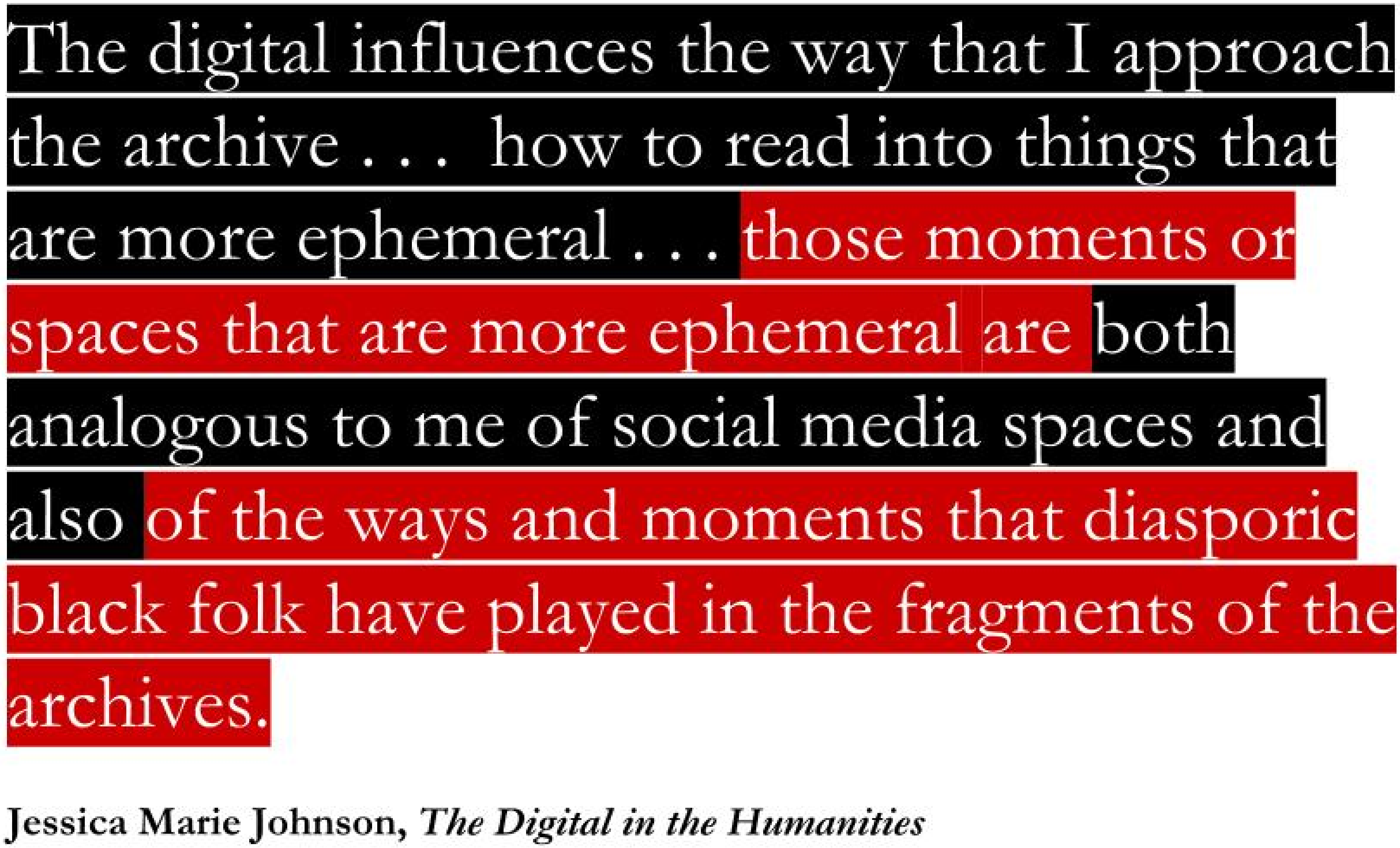
2) Copyright. Donor grants an irrevocable nonexclusive license to the Library for such rights as the Donor has the right to transfer or license under the Twitter Terms of Service in place at the time of the gift or before. The current, as of the effective date, and previous Terms of Service are appended.

3) Access. Any portion of the Collection originally posted to the Twitter service six months prior to the then-current date may be made available to Library staff and to

Gift Agreement from Twitter to the Library of Congress

Some Thoughts on Ethics and DocNow

I hope we see more efforts like this in this space. A contract after all is about agreeing on the terms of engagement. It’s about mutual respect.

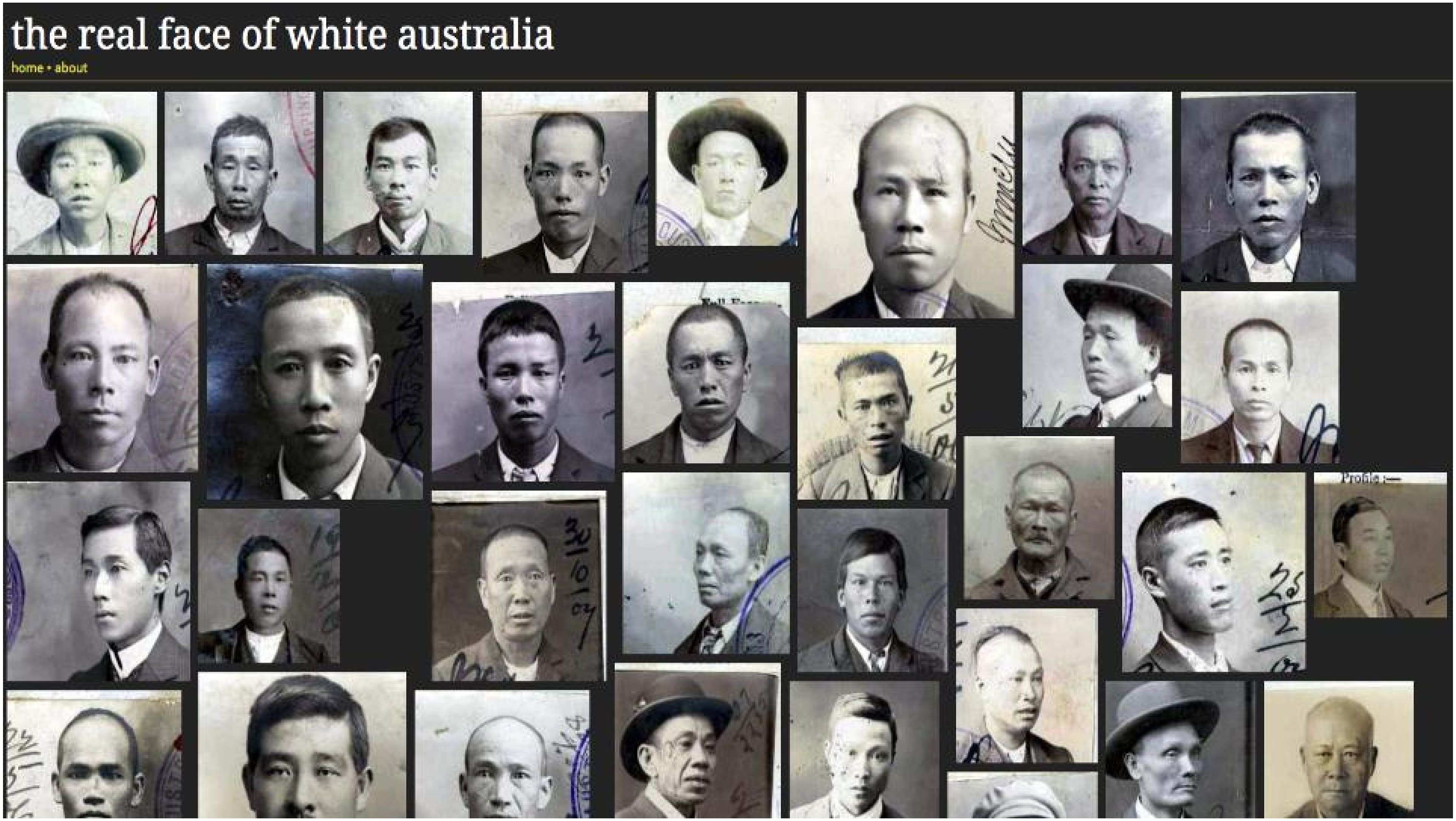


> The digital influences the way that I approach the archive . . . how to read into things that are more ephemeral . . . those moments or spaces that are more ephemeral are both analogous to me of social media spaces and also of the ways and moments that diasporic black folk have played in the fragments of the archives.

Jessica Marie Johnson, *The Digital in the Humanities*

In her interview with the LA Review of Books, Jessica Marie Johnson hits upon one of the key failings of our collections – their inclusivity. She like many others have by necessity sought meaning in the fragments and absences in archival holdings that testify to practices of systematic bias whose net effect is nothing less than historical erasure.

Sometimes it is the case that the histories are buried beneath a homogenized approach to representing our collections. Seeing collections as data offers an opportunity to surface difference in the historical record, by emphasizing people rather than bureaucracy. In the case of *The Real Face of White Australia*, Tim Sherratt, leveraged a facial detection script to automatically extract faces from a wide range of historical documents, and place those front and center. In doing so, Sherratt utilized a data oriented approach to humanize that data and surface a story that might have not been otherwise told.



the real face of white australia

home • about

As we transition our historical sources to meet the collections as data mark, we must be conscious of the extent to which the bulk of the digitized historical record privileges white, English language, predominately western history. This is a legacy born of early twentieth century microfilm collection development policies.

Canon
Microfilm Scanner 800
UC Roosevelt parade to protest US action
UC concerts to continue
Council will decide Robie zoning

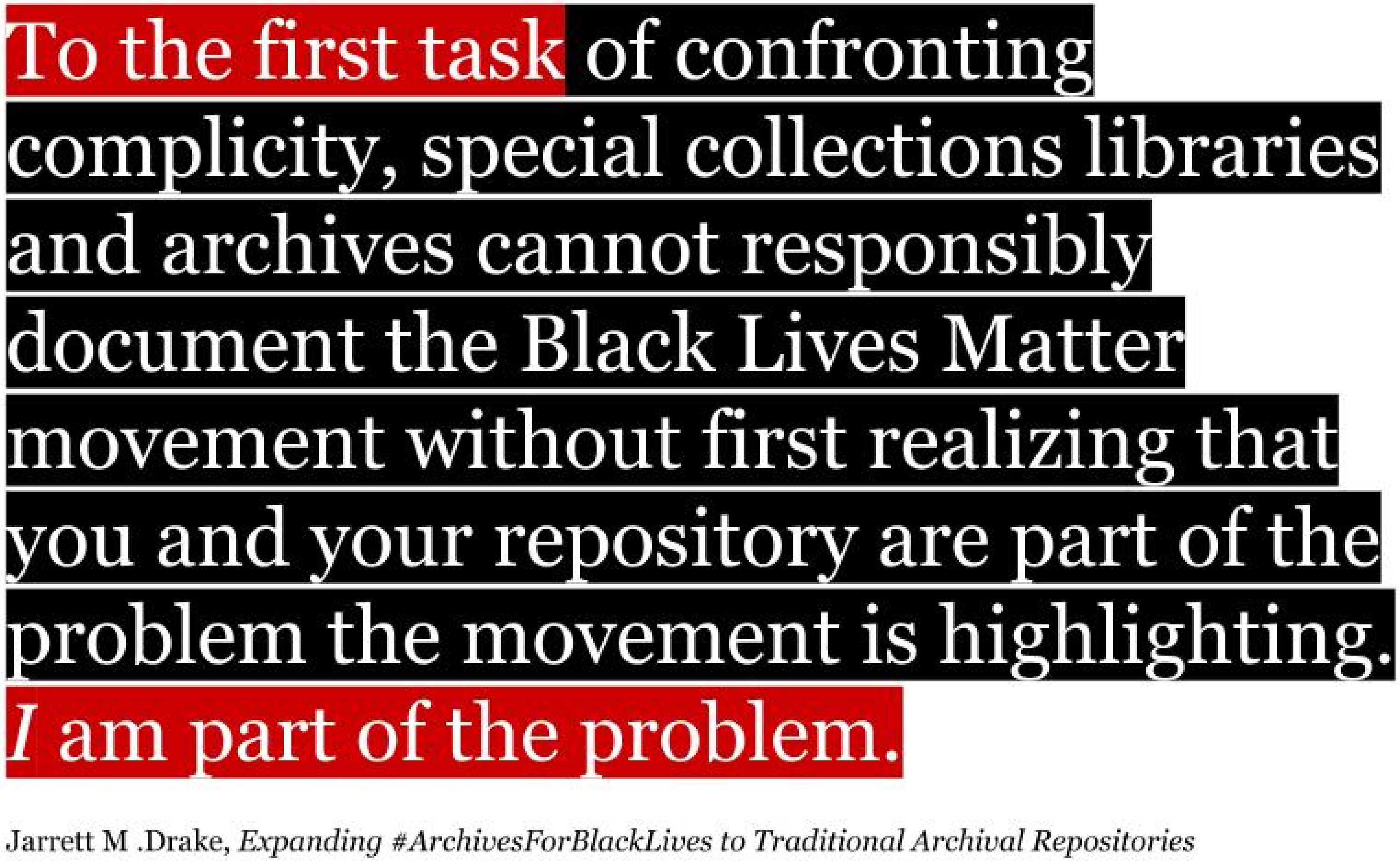
How I spend free time, by urbanoasis

Subsequently many digital projects have worked from these data sources, which biases said projects toward reinforcing a less pluralistic view of the world. Presently, we have an opportunity to create collections as data that better support the ability to craft narratives that reflect a greater diversity of lived realities.

The theory and practice of landscape painting in water-colours., by Internet Archive

Jarrett Drake reminds us that we are complicit in the creation of collections that reinforce inequity. In order to improve we must face that reality, look ourselves in the face, and aim to do better.

As I have mentioned throughout this talk, collections as data provides an opportunity to do better.

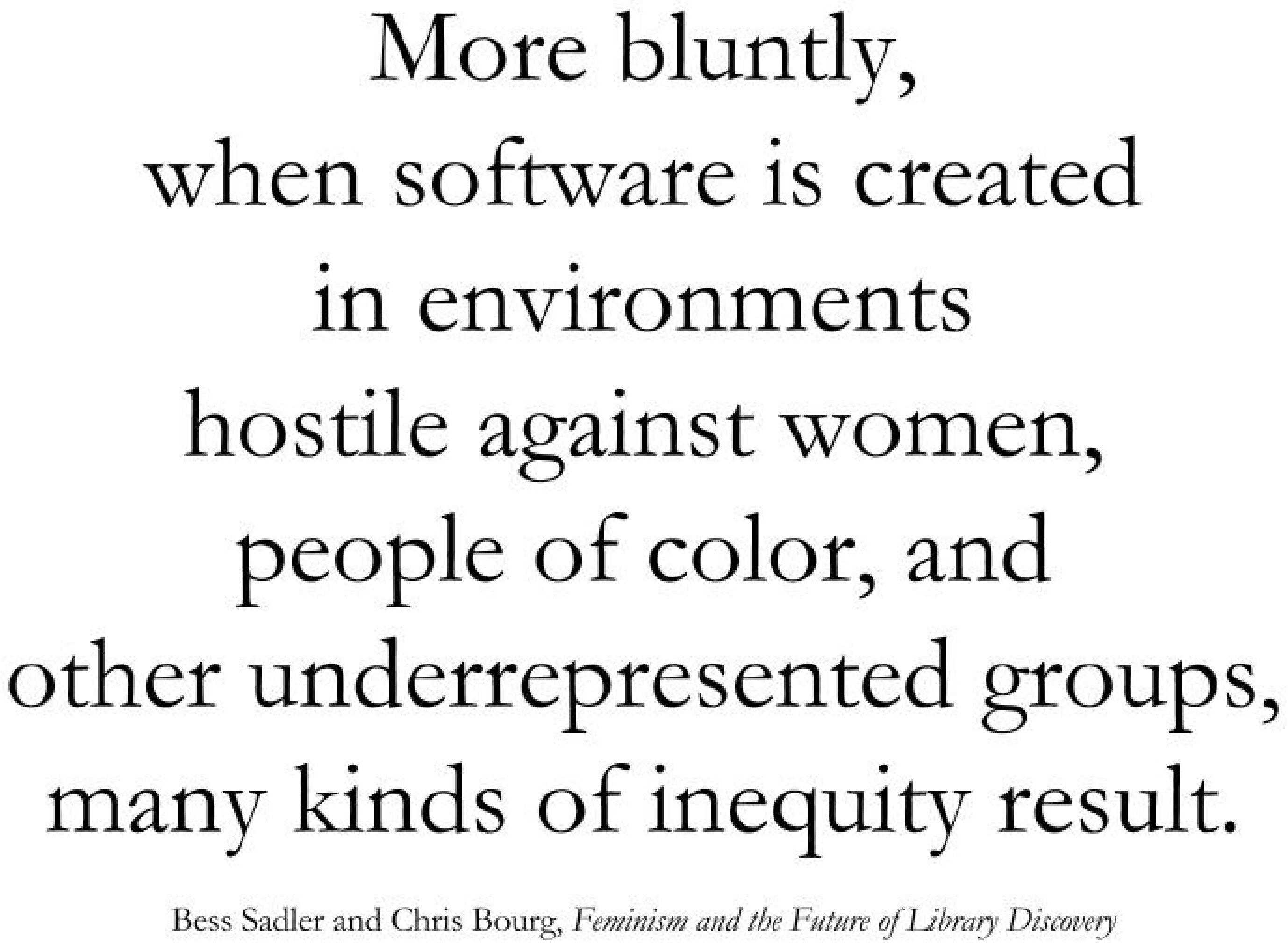


To the first task of confronting complicity, special collections libraries and archives cannot responsibly document the Black Lives Matter movement without first realizing that you and your repository are part of the problem the movement is highlighting. *I* am part of the problem.

Jarrett M .Drake, *Expanding #ArchivesForBlackLives to Traditional Archival Repositories*

There is some initial concerted effort seeking to improve our work in this space. The Digital Library Federation Cultural Assessment Group is one promising start. Its early but they are expressly geared toward identifying and questioning the underlying biases that are driving digital library collection development.
You may have noticed throughout this talk that I didn't talk much about infrastructure with any degree of specificity. It wasn't my goal to dig deep into technical considerations. That said, I wish it went without saying, but I feel I need to say it anyways –

. . . any infrastructure development in this space must take place in an environment that recognizes and works against historic and contemporary hostility toward women, people of color, and other underrepresented groups.

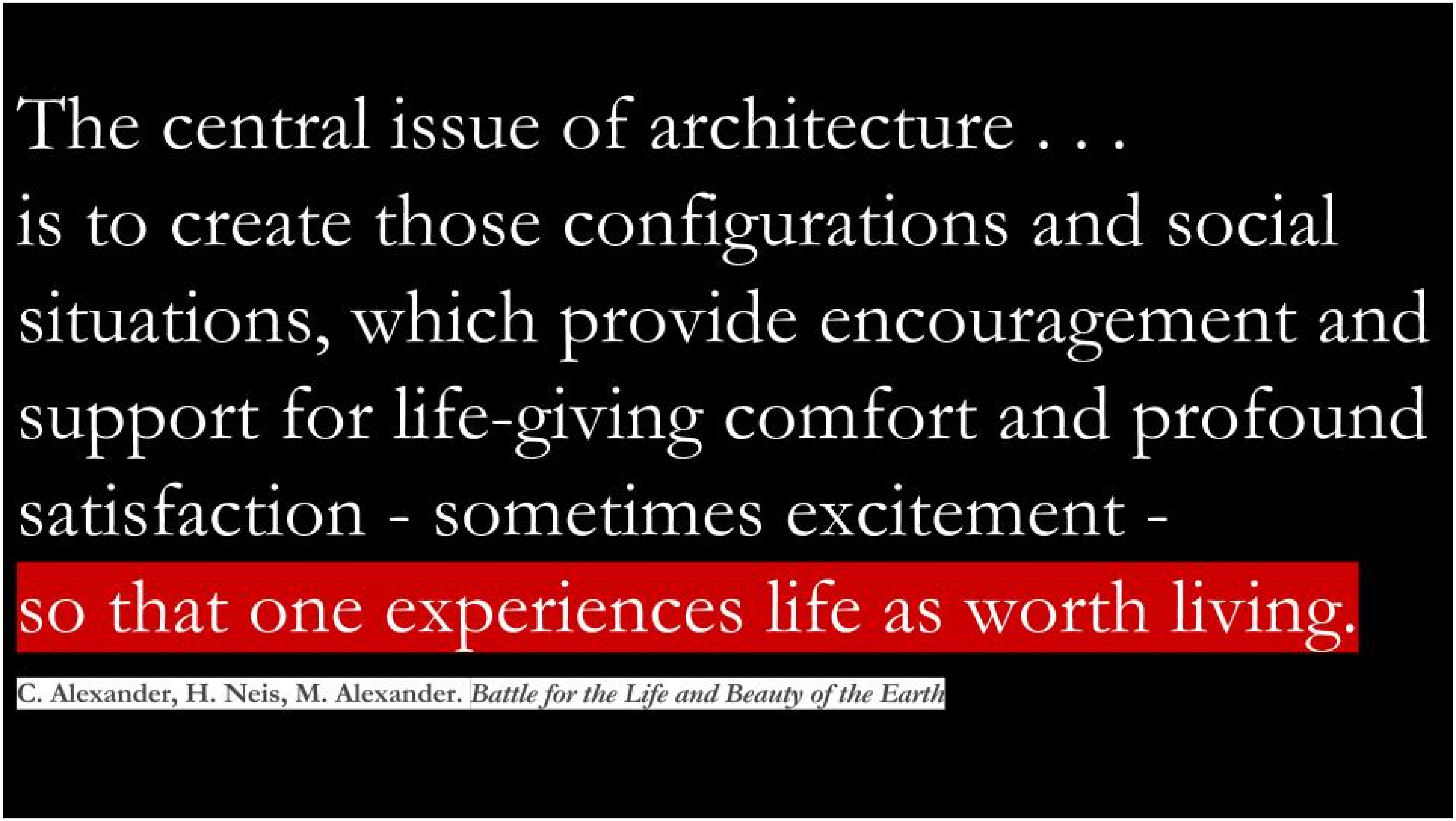


More bluntly,
when software is created
in environments
hostile against women,
people of color, and
other underrepresented groups,
many kinds of inequity result.

Bess Sadler and Chris Bourg, *Feminism and the Future of Library Discovery*

And we come full circle.

There is an incredible amount of opportunity that lies ahead. If we stay true to supporting the agency of our communities as they seek to make meaning from the data, if we empower ourselves to be inspired by our own experiments with data, and if we are guided by an ethics that focuses on transparency, inclusivity, and respect I truly believe that we are heading in a promising direction, a direction whose path charts a course to nothing less than supporting life as worth living.



The central issue of architecture . . .
is to create those configurations and social situations, which provide encouragement and support for life-giving comfort and profound satisfaction - sometimes excitement -
so that one experiences life as worth living.

C. Alexander, H. Neis, M. Alexander. *Battle for the Life and Beauty of the Earth*

Thank you.