Google: The Ultimate Cultural Prosumer

Jesse Fulton

June 2012

University of California, Santa Cruz

A thesis paper submitted in partial satisfaction of the requirements for the degree of Master of Fine Arts in Digital Arts and New Media

Thesis Committee:

Dee Hibbert-Jones, Warren Sack (chair). Noah Wardrip-Fruin

Dedication

To the future...

Acknowledgements

Thank you to my committee for their guidance throughout the development of this thesis.

Thank you to the DANM program for giving me access to so many invaluable resources and individuals.

Thank you to my friends for keeping me sane.

And thank you to my family for their unconditional support.

Contents

List of Figures

List of Tables

"Instead of the traditional information blackout, we face an information blizzard—a whiteout. This forces the individual to depend on an authority to help prioritize the information to be selected. This is the foundation for the information catastrophe, an endless recycling of sovereignty back to the state under the pretense of informational freedom."

Critical Art Ensemble, *The Electronic Disturbance*

"It is...the possibility of transmitting individual experience that makes possible the process of exteriorization. *And this is what we call culture.*"

Bernard Stiegler, *Leroi-Gourhan: L'Inorganique Organisé*(transl. Charlie Gere in *Art, Time, and Technology*)

"Library science scholars in particular concern themselves with the changing locus of access to information and knowledge (from public shelves and stacks to commercial servers). The 'Google effect'... may be studied in terms of the demise of the expert editor, and the rise of the back-end algorithm."

Richard Rogers, *The End of the Virtual: Digital Methods*

"What the work of art looks like isn't too important. It has to look like something if it has physical form. No matter what form it may finally have it must begin with an idea."

Sol Lewitt

"Science is what we understand well enough to explain to a computer. Art is everything else we do."

Donald Knuth, Foreword to $A{=}B$

Introduction

> "The issue no longer is how much of society and culture is online, but rather how to diagnose cultural change and societal conditions using the Internet. The conceptual point of departure for the research program is the recognition that the Internet is not only an object of study, but also a source."[1]

Google has defined and redefined what we[2] think of when we hear the word *Internet*. Google is responsible for handling the large majority of the world's Internet searches[3][4][5] and they also provide a number of other incredibly useful web-based tools such as email and mapping services, free of charge. They are a major gateway for all things digital. However, for all of the good that Google has provided, the company has come under scrutiny for a variety of business practices, largely focused on its PageRank algorithm, its privacy policies, and its data storage policies. This paper and the accompanying projects focus on one more aspect which I believe deserves more attention: "who (or what) is responsible for curating the content made available to me when I open my web browser?" The answer to that question is a complicated one, but for most users connected to the Internet, the largest common denominator is Google. Outside of government censorship, Google's software plays the primary role in enabling (or revealing and concealing) the possibilities of the Internet.

As technology becomes more accessible and more ubiquitous, websites and software applications are becoming a primary resource for research, entertainment, and personal development. I am interested in the societal impact of these large-scale information-based software systems, and specifically how they influence human knowledge (research & learning; information retrieval; individual and social memory) and culture (preserving culture; creating/influencing cultural trends; redefining the notion of a "cultural divide.")

For my MFA thesis, I have a developed a pair of projects looking at these issues: *Everybody's Google* is an exploration of the multiple "digital life-worlds"[6] presented by Google's search & personalization algorithms. *Ocular Character*

[1] Rogers (2009)

[2] Throughout this paper, I will be using the terms "we" and "us" to refer to the majority of the networked Western world

[3] `http://searchenginewatch.com/article/2174642/Yahoo-Search-Market-Share-Losing-Streak-Extends-to-8-Months`

[4] `http://gs.statcounter.com/#search_engine-ww-monthly-201104-201204`

[5] `http://www.statowl.com/search_engine_market_share.php`

[6] To avoid misconceptions, I am using this term to refer to the digital experience mediated through software. It is influenced by phenomenological thinking, but with an understanding of software studies. As Lev Manovich succinctly put it, "there is no such thing as 'digital media.' There is only software."Manovich (2011) Just as we perceive the physical world through our bodies and senses, we come to understand the rules and possibilities of the digital world only through software. While the data and bits on a hard drive do not change, our interactions with those digital objects are governed by multiple layers of software through which we engage it. This set of software-mediated experiences with the digital is what I refer to as a "digital life-world."

Recognition investigates how we interact with and archive cultural artifacts, specifically the digital representations of traditional print media.

Everybody's Google focuses on Google's search personalization algorithms. Google is profitable as a business because it makes money serving highly personalized advertisements[7]. Google tracks users as they navigate the web, constructing representative data models by analyzing their browsing behaviours (rather than relying on say, volunteered demographic information.)[8] The sites you visit, the things you search for, the date and geographic location of your searches - all of this information is stored by Google and assembled into a model in order to then serve you the "most relevant" advertisements possible[9]. This process of selecting "relevant" advertisements is executed by an algorithm ranking items based on the inferred user models. If some ads are ranked as relevant, then by definition others must be considered to be less relevant, or even irrelevant. In addition to fueling Google's AdSense platform, this is the same fundamental process underlying the personalization of Google search results on a user-by-user basis. The personalization process filters out results which Google's algorithms deem irrelevant for you based upon your user model; it also artificially inflates the scores for pages which it believes you might be interested in. However, in a privatized system run by a for-profit company, one must assume that information deemed "relevant" exists in a gray area between "important" and "profitable."

End users have little to no control over how Google is evaluating reults for search queries. The items which are selected and the order in which they appear are controlled by algorithms, invisible to the user. And because they are classified as trade secrets, Google has no responsibility or reason to disclose how they work; what information they are collecting and analyzing; or how certain behaviours affect Google (personalized) search rankings. These algorithms are executed in a black box[10], making it impossible for outsiders to gain a true understanding of the filtering and selection process. "Personalization renders search engines practically immune to systematic critical evaluation because it is becoming unclear whether the (dis)appearance of a source is a feature (personalization done right) or a bug (censorship or manipulation.)"[11]

While it may be primarily known for indexing web sites, Google also indexes videos[12], consumer goods[13], geographic locations[14], news[15], books and schol-

[7] In 2009, 98% of Google's revenues came from advertising. Rose (2009)

[8] "We are not Google's customers: we are its product. We—our fancies, fetishes, predilections, and preferences—are what Google sells to advertisers. When we use Google to find out things on the Web, Google uses our Web searches to find out things about us." Vaidhyanathan (2011)

[9] Pariser (2011)

[10] A black box is a system in which the input and output are transparent and visible, but its inner workings and internal processes are not.

[11] Stalder and Mayer (2009, 110)

[12] `http://www.youtube.com/`

[13] `http://www.google.com/shopping`

[14] `https://maps.google.com`

[15] `https://news.google.com/`

arly articles[16][17], social interactions and communications[18], images[19] and much more[20]. Google makes it possible to find the exact frame in a video when a particular phrase is uttered[21]. Google makes it possible to find the best gluten-free Pizza shop nearby and can then give us turn by turn directions optimized for our mode of transportation. Google makes it possible to get close enough to Starry Night to make out individual threads of the canvas[22]. Google can manage our stock portfolio; it can inform us of breaking news; and it can email us transcriptions of voice mails left on our home phones when we're half-way around the world. Google will even play the role of in-car tour guide if you drive a particular model of BMW[23]. In short, Google is indexing culture and then making it widely available through its various services. The algorithms running these services are not just filtering information or data, they're filtering culture and influencing its trajectory. In this sense, Google is engaged in a never-ending cycle of simultaneously consuming culture as well as producing it - it is the ultimate cultural prosumer.

The first step in making all of this possible is to index, store, and analyze cultural artifacts: music, films, books, museums, restaurants, opinions, events, etc. But at which point do the digital representations overtake or replace the originals? What are the cultural ramifications when the physical object is destroyed and all that remains is its digital representation? And what happens when the digital representation becomes corrupted, is censored, or is lost in a power outage or natural disaster? The more interesting question to me is, "how might this create a new method of interacting with, understanding, and appreciating (digital representations of) physical cultural artifacts?" How does digitization affect culture - (culture "proper", as well as digital culture)? For example, how have these technologies changed our notion of literature? Or what are the qualities of a book? Is a book simply words or images printed on bound paper? Can a series of screens and buttons on an iPad or Kindle still be considered a book? What if the words on the printed page are clearly digital symbols or artifacts? Is something's "book-ness" determined by its material, content, or some other set of qualities?

My project *Ocular Character Recognition* examines some of these questions by appropriating and manipulating one of the key innovations which has made the Google Books Library Project[24] feasible - the CAPTCHA. CAPTCHAs are the images of distorted text you must translate to do things like sign up for an

[16] `http://books.google.com/`

[17] `http://scholar.google.com/`

[18] `https://plus.google.com/`

[19] `http://images.google.com/`

[20] Including most recently, the concept of *things* (`http://www.google.com/insidesearch/features/search/knowledge.html`)

[21] Youtube indexes closed-captioning data which includes a textual transcript of all dialogue in the video mapped to timestamps in the feed

[22] `http://www.googleartproject.com/collection/moma-the-museum-of-modern-art/artwork/the-starry-night-vincent-van-gogh/320268/`

[23] `http://www.youtube.com/watch?v=xLSTfGZAJDE`

[24] `http://www.google.com/googlebooks/library.html`

account on a website or post a comment in an Internet forum. They're used to prevent spam by determining if the user of a site is a human, or a computer program (often referred to as a "bot.")[25]

A familiar CAPTCHA test

As part of its digitization process, the Google Books Library Project is an attempt to create digital copies of every book in existence. When Google digitizes physical books, OCR[26] software is used to "read" and convert the scanned pages from images to machine-encoded text, creating a machine-searchable and indexable copy of the physical book in the process. However, the OCR algorithms occasionally fail due to scan irregularities, low print quality, or similar problems. When this happens, the only way to decipher these portions of the digital scan is with a human eye. Thus, Google's goal of digitizing every book in existence requires an enormous amount of human labour. reCAPTCHA, the most popular CAPTCHA service on the Internet, distributes this labour as micro-tasks to Web users across the Internet, using their eyes to decipher the pieces of scanned text which the OCR software could not read. In exchange for access to enhanced website functionality (such as the ability to post a comment on a blog or forum), website visitors solve CAPTCHA tests, providing Google with the (free) labour required to fill in the gaps in its OCR translations[27].

While the Internet is largely viewed as embracing democratic ideals and guaranteeing equal access for all, to all, the fact that our Internet experience is mediated almost exclusively through privatized companies, is inherently undemocratic. The UN has dubbed the Internet "one of the most powerful instru-

[25] CAPTCHA is an acronym for "Completely Automated Public Turing Test To Tell Computers and Humans Apart"

[26] Optical Character Recognition

[27] A CAPTCHA served from the reCAPTCHA service consists of two text-based images: a word which could not be deciphered by OCR software (the "unknown" word); and a second word which the OCR software could decipher (the "known" word.) The user is then prompted to type in both words, with the assumption being that if the known word is correctly identified, then the identification of the unknown word is also likely to be correct. This human translation for the unknown word is then added to the original OCR translation.

ments of the 21st century for... building democratic societies."[28] However, the majority of the most heavily-trafficked web sites are developed and maintained by private companies running closed-source software systems. The monopolization and privatization of the Internet experience counteracts many of the goals and values of the idealized democratic Internet, specifically "the freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."[29] This freedom is fundamental in ensuring the diversity of global cultures.

Additionally, without the knowledge of how these closed-source software systems work, it will be impossible to truly understand the digital world of the Internet. This has the potential to continue to reinforce the electronic colonialism currently underway by large Western corporations. And if this is not the direction we want the Internet to head in, we must first gain "the ability to understand the world [in order] to change it."[30]

I view my thesis projects as experiments rather than as answers to any particular questions. They are intended to raise awareness of complex and important issues in an approachable and easily digestible manner. For something that largely remains unseen, it is difficult to create a physical form, but the projects detailed below represent a few of the possible manifestations of the concepts and ideas presented within each project.

For the MFA show, a number of works were shown - one for *Everybody's Google*; and three for *Ocular Character Recognition*. The approach I have taken with my thesis work has been that the ideas themselves constitute each project (the two projects being *Everybody's Google* and *Ocular Character Recognition*.) The software I have written is the manifestation of those ideas. Interacting with the corresponding websites (in a non-gallery setting) is the experience of those ideas. And the pieces shown at the MFA show were a condensation of those ideas.

[28] The United Nations Human Rights Council (2011)
[29] ?? (UN-)
[30] Oscar H. Gandy (1988, 109)

Everybody's Google

> "Access to the Web is preconfigured in subtle but politically important ways, resulting in exclusion of significant voices... The politics of search engines thus represents the broader struggle to sustain the democratic potential of traditional media, the Internet, and the World Wide Web in particular."[31]

In his short story *The Library of Babel*, Jorge Luis Borges clearly illustrates how physical access to an abundance of information does not bestow upon one ultimate knowledge. A blizzard of information creates a white-out. An overwhelmingly massive library of information is unusable without being able to filter out the useful from the useless, the genius from the insane, the insightful from the nonsense. To make use of the information, one must possess the ability to find the desired information as well as the ability to discover new information.

Much like Borges' library, the World Wide Web is an enormous collection of independent documents on topics containing nearly every type of information imaginable - some of it incredibly useful, but most not so much. While the actual size of the Internet is near-impossible to calculate, a *partial* copy of the Web hosted by the Library of Congress's web archive, weighs in at over 285 petabytes[32][33]. Humans do not have the mental resources to sift through so much content unaided. In a society suffering from a severe case of information overload, we must closely manage how we focus our attention. With such an overabundance of data on the Web, we need a guide - a search engine.

In their simplest form, search engines store a copy of every page on the Internet and return a ranked list of the most relevant pages from its database based upon text-based queries. "If Herbert Simon was right in 1971 when he declared attention a scarce resource consumed by an overabundance of information, we have to recognize that ranking is not only very useful but also inevitable... Any system of ranking will favor certain sites over others; the question is which ones."[34] Google has often been criticized for its PageRank[35] system, being accused for unfairly favoring certain websites (or itself[36]) over others. But the Google interface is very simple and because Google search results are structured as linear text, they *must* have an order or a ranking. The original PageRank algorithm was a form of citation analysis[37], ranking web pages according to number of incoming links or citations.

However, search engines have become much more complex since their inception. In order to remain competitive and profitable, they are constantly refining their ranking algorithms. One integral innovation to Google's algorithms has been search personalization. Google tracks user behaviours on its

[31] Introna and Nissenbaum (2000)
[32] 1 petabyte = 1,000,00 gigabytes
[33] `http://www.loc.gov/webarchiving/faq.html#faqs_05`
[34] Rieder (2009, 139)
[35] PageRank is Google Search's algorithm for ranking web pages
[36] Zapler (Zapler)
[37] Brin and Page (Brin and Page)

own sites, as well as on external websites through the use of various embedded widgets. They use this information to generate "user interest" models. These inferred data models are then used to personalize Google Search, prioritizing results which its personalization algorithms deem relevant to each individual user. In this sense, the Google algorithms are the arbiters of taste in the digital realm. They present you with the information you should be most interested in. Personalization creates potential for a "loss of autonomy... related to the fact that we are presented with a picture of the world (at least how it appears in search results) made up of what someone else, based on proprietary knowledge, determines to be suitable to one's individual subjectivity."[38]

Everybody's Google explores these ideas and systems by programmatically controlling a set of Google user accounts. In an effort to hyper-personalize their Google Search results, these accounts have been scripted to navigate the web biased towards specific cultural interests or political orientations. These scripted users browse the internet and perform searches, while feeding the data from these processes to a real-time visualization displaying each account's top 10 results for particular Google Search queries, side-by-side.

With *Everybody's Google*, I'm not showing a new way to view the digital world, but rather *revealing* the way it actually is (or the multiple ways in which it's presented.) This project demonstrates that there is a black box, and it is not empty - there is a very complex process taking place within it. While we can see the outcome, the process is hidden. Whenever information goes into this black box, "personalized" data comes out. Each time Google applies a personalization algorithm (not just in search, but also in ads, news feeds, and YouTube suggestions), they're making assumptions about individuals and using these judgements to direct their digital world view.

A number of artists & technologists have been working with Google as a medium and also investigating personalization, but I'm not aware of any who have gone quite as in-depth as I have with *Everybody's Google*. Ben West & Felix Heyes' recent *Google* is a 1,240 page book of images containing the first Google Image Search result for every word in the English dictionary and is a wonderful example of using Google and the Internet as a medium. Two projects recently came out of Rhizome's "Seven on Seven" event earlier this year, both addressing issues of personalization and the multiplicity of digital life-worlds. *Peep* by Xavier Cha & Anthony Volodkin allows you to recreate any other user's Twitter feed and view the information they subscribe to. *Cultural Differences* by Taryn Simon & Aaron Swartz allows website visitors to see the different results returned by Google Image Search for various queries. Both of these are interesting projects, but neither actually allows you to truly experience the personalized web as another, authenticated user. *Everybody's Google* is not recreating or simulating anything, it is able to show the actual, real results and experience.

[38] Stalder and Mayer (2009)

Formal Description

Everybody's Google is a piece of software which programmatically controls a set of Google user accounts and navigates them through the World Wide Web. Four new Google accounts were created, each representing a particular set of stereotypes: a young man living in London who is interested in technology and gambling; a teenage girl from Atlanta interested in fashion, celebrities, and pop music; a middle-aged conservative Christian businessman from rural Virginia; and a liberal, Californian mom-to-be, interested in women's rights[39].

However, unlike traditional characters in a story, these personas only exist in code as a set of URLs. "Character development" occurs as the software guides them through the web, visiting blogs and news sites. Each account has a series of "seed" URLs on which they begin. The software then selects links on each page (for example, all "featured" links) to follow and continue reading. With each site they visit, information is sent to Google[40], allowing the company to hone and refine its idea of these users' interests. Google builds models of this information, using it to send targeted advertisements, present relevant news information, and personalize search results. For example, one of the accounts is constantly reading MTV hip hop blogs - Google is likely to assume that this person is very interested in hip hop music and news and will filter the results accordingly.

The specific character interests (defined by their subscribed URLs) were intentionally chosen to get a diverse set of search results for a wide variety of searches. The websites chosen generally have specific political leanings or serve various ethnic or cultural demographics. In an effort to hyper-personalize their Google Search results, these accounts have been scripted to navigate the web biased towards specific cultural interests or political orientations. In a sense, their "character development" (from Google's perspective) is generated through their scripted browsing behaviours. The more they browse the web and view sites with Google's tracking cookies, the more Google "knows" about them (in a certain area), and the "better" they can hone in on personalized search.

In its current form, the project is engaged with through a web interface. Rather than exploring raw data in a traditional tabular or list format, visitors are presented the data as an animated series of rotating cubes showing each account's top 10 results for a revolving set of Google Search queries. Each cube in the visualization represents one item in the top 10 search results returned by Google; each face of each cube represents one of the four personas. As the cubes rotate and a set of cube faces becomes perpendicular with our field of vision, we are presented screenshots of the 10 URLs returned to that Google account for the current search term being visualized.

[39] For more detailed information about the personas, and the sites they visit see Tables 1-4 in the Appendix

[40] all of the "seed" URLs were chosen because they use one or both of Google Analytics or Google AdSense. Simply having these tools present on a web page allows Google to collect information about users.

Everybody's Google WebGL visualization

Visitors to the website may engage with the data by entering a search query which will then be queued and executed by each of the 4 personas. Once completed, their search results are added to the main visualization. The web site homepage uses the same design treatment as the main Google search page. The colors have been manipulated, but the layout and fonts are the same. The visualization also makes use of a modified version of the Catulli font, the font face used for the Google logo and the familiar red/yellow/green/blue color scheme. At the MFA show, a monitor was running the visualization and visitors could add items to the queue/visualization by visiting the `everybodysgoogle.com` on their mobile devices.

Technical description

Everybody's Google began as a website and an accompanying plugin for the Chrome web browser which would allow anybody to transparently upload and compare their personal Google search results to those of all plugin users. The project was then refocused to create custom personas which could be controlled through code in order to maximize the potential variability in search results[41].

[41] This original version of the project is available at `https://github.com/jessefulton/everybodys-google`. The second code repository was named "google-views" and is at `https://`

The second approach was taken when initial tests showed promising results after Google searching for "romney" from two test accounts. One of these test account had been configured to visit all of the pages the politics section of `npr.org`; the second, to do the same, but beginning from `foxnews.com`. After they had browsed each respective site, both user accounts executed a Google search for "romney." The results were similar barring one important difference: the results for the NPR contained `spreadingromney.com`, a "Google Bomb" meant to embarrass presidential hopeful Mitt Romney; while the results for the Fox News reader included a pro-Romney news blog (`committedtoromney.com`) instead[42]. Given the great initial results, I created four new Google accounts and developed a persona for each, as described in the previous section.

I created a set of scripts using CasperJS[43] which let me programmatically navigate the web, authenticated as a Google user. Three main scripts were used to perform actions on behalf of these users: "authenticate," "browse," and "search." Before every action, the "authenticate" script was run to log in to a particular Google account, after which we could perform the "browse" or "search" action. The "browse" script would open an initial webpage at a "seed" URL. It would then select specific links on that page (i.e., only links in the "featured stories" section of a news site) and visit (or "read") each link. The "browse" script was executed once every ten minutes for roughly one month before & after the MFA show. Each time the script ran, it would sign in as one of the four users, and click on particular links of predefined webpages which correlated to that personality. The "search" script would execute a Google search query as the logged in user, and then save the search results to a database. Whenever these processes completed, they'd fire off a series of events which would begin the image manipulation and generation required for the data visualization.

The first step to building the visualization was to capture screenshots of each URL in the search results set. When the "search" script completed, a second script would search the database for the newly added search result page URLs and create "screen grabs" or generate images of those webpages on the server. Next, each screen grab was converted into a format and size compatible with WebGL[44] textures[45] using ImageMagick software. Finally, after all of the screenshots had been generated for each user who had performed the query[46], a socket.io[47] message was sent to any running WebGL visualizations, which would then update themselves with the new content and textures in real-time[48]. The

`github.com/jessefulton/google-views`. There are plans to eventually merge the two projects back together.

[42] See Figure 1

[43] CasperJS allows scripting of web browser behaviors and navigation

[44] WebGL is a web browser port of OpenGL, the de facto 3D graphics implementation. WebGL enables real-time 3D graphics processing inside of a web browser.

[45] In order to texture map a 3D object in WebGL, the dimensions of the image used must be a power of 2

[46] There was usually some overlap, so it averaged out to about 13 images per query.

[47] socket.io is a transport which allows for near-real-time communication between a web-server and a client browser

[48] So, if I were viewing the visualization on my own computer, while at the same time another visitor added an item to the search queue, I would see that action in my own web

visualization itself was built using Three.js, a library for scripting WebGL and building 3D environments.

The website itself (and all of these other processes - search/browse; screenshot service; image manipulation) were created using NodeJS[49] running on a server at UCSC. However, Google's personalization algorithms also rely heavily on a user's geographic location. In order to prevent Google from thinking all of my characters lived in Santa Cruz, CA (and shared the same computer/IP address), I set up globally distributed web proxies in specific geographic areas running on Amazon EC2 [50] instances.

Shortly after the initial "soft-launch" opening of the MFA show, a number of "inappropriate" search queries began entering the queue for *Everybody's Google*. I assumed it was a few students having fun with the site. However, the list grew and the submissions overloaded the server to the point that the site crashed a number of times. I finally found the source of the traffic[51], and the night before the show reopened had to implement throttling, limiting IP addresses to two submissions maximum.[52]

Goals and expected outcomes

At the MFA show, visitors could watch the visualization of these searches happening in real time. There were users queuing search queries who were not at the show (as mentioned above, regarding the KnowYourMeme discussion), but people at the show could also interact with the piece provided they had a smart phone. Using the smart phone, they could visit `http://everybodysgoogle.com` and add items to the search queue and see their results within a few minutes (after all of the queries had been processed.) It was important for me to not introduce a keyboard and mouse interface for this piece. With that setup I believe it becomes very difficult for more than one person to engage with the piece at a time. In this situation, nobody is in a position to "assume control" of the piece. People are in charge of their own experience - whether that consists of walking past, stopping and viewing the piece for a bit, or interacting with it via cell phone. I believe this kept the focus of the piece on the content, as opposed to the interaction (or lack thereof.) By presenting the process as relatively straightforward data visualization, there wasn't much "digital magic" to figure out. I also feel that the minimalist implementation and design, while

browser in near real-time.

[49]NodeJS is a web server technology written in JavaScript which has a large and active developer community, providing numerous open-source libraries and modules with a wide range of applications.

[50]Amazon Elastic Cloud Computing is a service which provides dynamic "cloud" servers which can be geographically targeted to certain regions of the world

[51]A user on KnowYourMeme.com (a popular website for generating, documenting, and discussing internet memes) had visited the show at UCSC the previous weekend and posted a message about *Everybody's Google* in their forums. That online community then began overloading the DANM web server. See original forum thread at `http://knowyourmeme.com/forums/general/topics/15561-everybodys-google`

[52]Interestingly, looking through the logs, a group of people had turned the search query queue into a sort of chat room or message board.

simply being consistent with my overall aesthetic, also helped direct focus to the content of the piece as opposed to formal qualities of the installation itself (which were not necessarily "bad" but I would not consider the installation to be a part of the project as it could really live on any monitor attached to any computer with an Internet connection.)

Most web users are not aware that their Internet experience is personalized and different from that of other users. Even those who are aware of this do not realize to what extent it occurs. The primary goal of this project is to raise awareness of web personalization by clearly and concisely showing it in action. To get people to question their "digital life worlds" and begin thinking about how Google models the world. Is it correct? Can the concept of "correct-ness" even apply here? How can anybody or anything organize the massive amounts of data available on the Internet in a way that's consistent with everybody's way of thinking? How can it efficiently and accurately sort through it all? Surely there must be compromises in this process.

Ocular Character Recognition

> "Google is designed to absorb and respond to culture as much as it influences culture."[53]

Up until this point, I have largely been focused on Google's web search which allows users to find web pages relevant to search queries. But as discussed in the introduction, Google is indexing much more than web sites. We can search geographic locations and restaurant menus; we can walk through virtual reproductions of museums; and we can search through digital representations of physical books. Google is making the physical world searchable through digitization.

What does it mean to be digitized? How does digitization change our understanding of the physical? While it is important to consider who is determining which objects are worthy of digitization, first one must ask what *can* be digitized? Digitization entails more than simply capturing data or models of certain objects. It is also the ability to create a representation of that object using display technologies. The ability to be digitized depends upon a combination of the state of technology and the qualities of the object itself.

The ability to transfer visual input to physical media has been around in some form for centuries, with sound recording coming more recently. Understandably, photos, texts, and music already have digital counterparts, whereas technology for recreating stimuli for tactile or olfactory senses is rare. But there are difficulties in all types of digitization processes - even those which lend themselves well to the current state of digital media (books, image scans, etc.) OCR algorithms have difficulties with font changes, superscripts/subscripts, mathematical formulas, handwritten notes, or even full handwritten pages (journals, diaries, legal documents/letters.) Issues also arise during the scanning process where digital artifacts can appear, or pages may be lost or corrupted due to aberrations in the scanning process[54]. As we move towards text-based search as our primary mode of retrieving this type of information, what are the consequences when it doesn't exist (or exists in a corrupted format) in our searchable database? How can we find the mis-translated words if they don't show up when we look for them? Does the content exist if we're not aware of its existence or cannot find it?

I'm very interested in how we treat physical objects differently from their digital representations. And as the physical and the virtual continue to merge together, it is interesting to materialize digital content and see the reverse process. Do we treat digital content differently if it's encountered in a physical format? Perhaps by doing this, we can gain any insight into the processes and changes which occur when transplanting artifacts from the physical realm to the virtual, or vice versa. This is the driving force behind my project *Ocular Character Recognition*.

[53] Vaidhyanathan (2011)

[54] `http://theartofgooglebooks.tumblr.com/` has some great examples of these issues

In 2009, German artist Aram Bartholl performed a series of urban interventions titled *Are you Human?* in which he created physical cutouts of CAPTCHA codes and installed them on public walls amongst graffiti tags. CAPTCHAs demonstrate the human cultural knowledge required to decipher signs of written language. By placing physical CAPTCHAs on walls heavily covered by graffiti tags, Barthol is making a connection from the process of deciphering and interpretation of digital signs to the signs used in graffiti culture. A graffiti tag is a sign for graffiti tagger as well a a representation of the tagging process itself. Similarly, CAPTCHAs signify not so much the underlying meaning of the word depicted, but rather the processes of scanning, distorting, and deciphering. While Bartholl's work more clearly speaks to language & signs, as well as the relatively new idea of relying upon computers to verify qualities of one's "human-ness," we are both interested in placing internet artifacts in physical or "unnatural" surroundings.

Bicycle Built for Two Thousand by Aaron Koblin & Daniel Massey is another great example of appropriating digital artifacts. "Daisy Bell (Bicycle Built for Two)" was the first song ever sung by a computer. Koblin & Massey took the computer generated audio and split it into hundreds of individual split-second sound clips. These isolated clips were then distributed as micro-tasks to users of Amazon's Mechanical Turk, who then used their computer microphone to reproduce the sounds they heard. These samples were then combined together to reproduce the original track using human voices as opposed to computer synthesized sounds creating a human-ized version of a digital rendition of a song.

This topic of digitization is a very large one, and I decided to focus on perhaps the most well-known, and most invasive/criticized aspect of this process: the Google Books Library Project. Unbeknownst to them, most web users have not only encountered the project, but they have likely contributed to it as well through the usage of CAPTCHAs. As mentioned in the introduction to this paper, Google uses CAPTCHA tests to supplement the automated translations being performed as part of the Google Books Library Project's digitization processes[55]. By solving a CAPTCHA, you're proving to a computer that you are a human. It is a somewhat ironic situtation: we are granting computers total access to create digital representations of our cultural artifacts. And the process by which computers exploit human labour to digitize these physical objects is the same process used to grant humans access to those very same digital representations.

I presented three projects under the umbrella of the *Ocular Character Recognition* project at the MFA show. These were largely experiments with materiality and exploring how we engage with the digital and the process of taking a digital artifact and presenting it in a physical form. *Ocular Character Recognition* consists of three separate, but related projects: *Webpages for Humans* -

[55]The project has also undergone heavy criticism for things like copyright infringement as well as for promoting a type electronic colonialism (initially, the project only digitized western language books and magazines (predominately English) from a small set of US & European libraries.) However, since the program began, it has become much more inclusive of non-Western cultures.

a series of prints of CAPTCHA-fied[56] webpages; *CAPTCHA-fy!* - handmade booklets expressing the underlying goals and ideas under the OCR umbrella (a manifesto); and the *Serendipity Engine* - a modified printer used to serendipitously dispense/produce CAPTCHA-fied pages from eBooks digitized through Google's book/library project.

The title *Ocular Character Recognition* is a play on the term Optical Character Recognition (defined in the introduction.) However, as opposed to an optical process - one which deals primarily with light and its properties - an ocular process implies an emphasis on the human eye. *Ocular Character Recognition* tries to put the focus of digitization back on the human experience, and preserving human culture by reversing the process of digitization, defamiliarizing the familiar, and reclaiming the physical/claiming the digital.

Formal Description

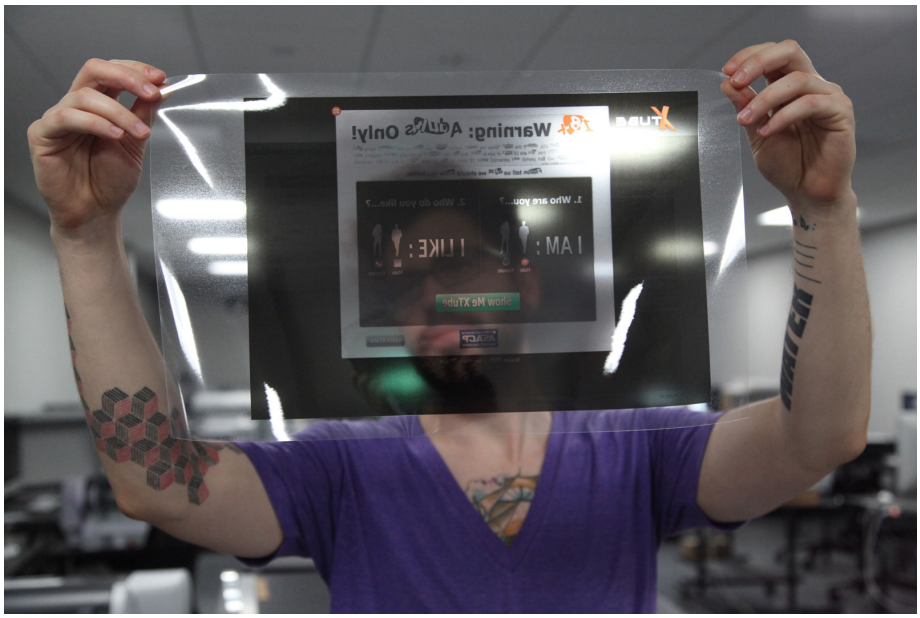
Webpages for Humans is meant to be encountered on the web. It is a web application which generates the "Human Web" - a series of navigable CAPTCHA-fied webpages[57]. When users first visit the site, they can enter a URL into a form which brings them to the CAPTCHA-fied HTML page for that URL. From this point on, any links clicked within the page will continue to be CAPTCHA-fied by the web server. However, as mentioned earlier in this paper, I have a strong aversion to showing digital works on a computer in a gallery setting. For this reason, I used the software & website developed for *Webpages for Humans* to create a series of digital prints using screenshots from the "Human Web." The installation for *Webpages for Humans* consisted of a series of 5 digital prints suspended from the ceiling using monofilament. These prints were generated by replacing the text on webpages with CAPTCHAs, transforming a digital document specifically designed for machine readability into a physical document which can no longer be "read" by a computer.

Each image was printed on a sheet of transparent digital mylar[58], and mounted between a sheet of clear acrylic and a sheet of translucent white acrylic. The familiar images had something slightly askew (the CAPTCHA-fied text) but their shape and size resembled that of a laptop screen, without the rest of the laptop. While it was a struggle to properly light the pieces due to the location (in a long hallway, the opposite wall being floor-to-ceiling windows), at night, with controlled lighting, the white acrylic backing would catch and diffuse the light, causing the floating screens to emanate a soft glow. The luminescence of the images under proper lighting conditions reinforced the connection to the screen.

[56] CAPTCHA-fy is the term I use to distort text to the point where it become unreadable by Optical Character Recognition. This technique is the only reason why CAPTCHAs work.

[57] See Figure 2

[58] transparent mylar prepared with a digital ground for use in digital printing



Webpages for Humans mylar prints

It has been brought to my attention that web pages are primarily created for human consumption. However, the web pages we view and read are simply visual representations of the underlying code. The CAPTCHA-fication process is not simply replacing text with images on a web page, it fundamentally changes the underlying markup of the HTML pages in a way such that not only is the code itself unreadable, but its visual representation (or rendering) is as well. And it is done in such a way that it can no longer be deciphered by OCR processes.

CAPTCHA-fy! served as the manifesto for the *Ocular Character Recognition* project, linking the concepts of *Webpages for Humans* and the *Serendipity Engine* to the ideas within the project as a whole. Each manifesto was hand-made and hand-bound. The covers to the booklets were identical - each with the computer-generated "CAPTCHA-fy!" text block printed onto a piece of thick, dark drawing paper. The block prints were created by etching a digital image into a linoleum block using a laser cutter. This created very crisp, pristine, and machine-like lines which contrasted well with the hand-made, DIY aesthetic. The inner pages were low-quality xerox copies, further reinforcing this conflicting man-made/machined aesthetic. As part of the installation, there was also a reading area consisting of two chairs and a footstool, where the pamphlets were displayed. There was originally a holder for the pamphlets to signal that they were takeaways, but it was not properly installed and shattered after falling shortly before the show opening. The pamphlets were then arranged along the

window behind the reading area.

CAPTCHA-fy! booklet covers

The *Serendipity Engine* consisted of a modified printer installed in the ceiling above the reading area. The printer would dispense individual CAPTCHA-fied pages from "The Mad Tea Party," chapter seven of Lewis Carroll's book *Alice's Adventures in Wonderland*. The pages were taken from Google's free eBook version of *Alice*, and after manipulation, still contain the "digitized by Google" watermark. Each page was printed on high quality cotton paper and cut to the size of a standard book page. They are clearly recognizable as pages from a book, yet the CAPTCHA-fied text, a digital artifact, immediately defamiliarizes the object. In addition to being a writer, Lewis Carroll was also a logician and mathematician. He was disillusioned with the radical changes occurring in the fields of Mathematics around this time, and The Mad Tea Party was a critique of William Rowan Hamilton's concept of the Quaternion, or more directly it was a critique of what Carroll felt was Hamilton's misuse or appropriation of time in order to solve an unrelated problem. Similarly, I have appropriated this piece of writing to provide a critique of another unrelated system and set of processes - Google's consumption and production of digital culture.

The printer mechanism for the *Serendipity Engine* was configured to dispense sheets of paper at set time intervals throughout the duration of the exhibition rather than making the piece reactive or responsive to some external stimulus.

The title of the piece[59] implies an unexpected discovery. If this piece were interactive, there is an expectation for a result. This is something I wanted to avoid. It is also a way to reduce the amount of "tech magic" in the piece. Placing something which is completely non-interactive at a digital art/new media show allows the object to retain its power. It cannot be taken away once the technology driving the piece has been "figured out."

Serendipity Engine and reading area

Technical description

The core functionality for these projects is a shared JavaScript library I created which generates images of CAPTCHA-fied text[60]. The library uses the HTML Canvas element to draw various shapes and manipulate text in a style emulating that of CAPTCHAs. This data was then saved as a DataURIs[61] to allow it to be directly embedded into HTML documents or saved as separate image files. This library was then implemented in a web site[62]. Originally, this functionality was

[59] "Serendipity engine" is a term Facebook and Google have both used with regards to the future of their services. Rather than having users actively search for information, these companies are aiming to reach a point where (relevant and interesting) information would find users before they even think about searching for it.

[60] `https://github.com/jessefulton/node-captchafy`

[61] `http://en.wikipedia.org/wiki/Data_URI_scheme`

[62] `http://nodecaptcha.herokuapp.com/`

going to be written in Java and served via a separate web service[63], but hosting issues made this unrealistic. By using Javascript however, it was incredibly easy to embed this functionality in web browsers, which greatly simplified the HTML manipulation process for the *Webpages for Humans*[64] project. The server was able to inject the JavaScript classes into any webpage, using CasperJS scripts. Writing the library in JavaScript also made it trivial to create a bookmarklet[65] version of the *Webpages for Humans* project.

The JavaScript library was used to generate the cover image and some of the internal pages of the *CAPTCHA-fy!* booklets. A high-resolution version of the CAPTCHA-fied cover text was etched into a block of linoleum using a laser cutter. The images for the inner pages where laid out using Adobe InDesign and then copied at a low-quality setting on a standard Xerox machine.

The code to run the printer for the *Serendipity Engine* is very straightforward, but is not available online. A printer was dismantled until just the document tray, rollers, and motor were still intact. An arduino was connected to the printer's DC motor and was used to run the motor (and feed out sheets of paper) at standard time intervals, The prints were generated by another set of scripts using the node-captchafy library[66]. Initially this code was written in C++ using OpenFrameworks[67], but this proved to be more overhead than necessary and the project was rewritten to use simple bash scripts. The scripts take a PDF file as input and run OCR algorithms on the document using the Tesseract library. Once Tesseract has identified the text in the PDF, the translation is saved in hOCR format, which is essentially an HTML document. Finally, this HTML document is processed by a custom PhantomJS script which positions and styles all of the elements accordingly and injects the node-captcha library in order to CAPTCHA-fy the whole page, replacing roughly ever other word on the page with a CAPTCHA-fied version of itself.

Goals and expected outcomes

A CAPTCHA is something most Internet users have encountered, yet few know why they appear, and even fewer know how the process is connected to the digitization of physical books. However, based upon my interactions with students who visited the show, a fairly high proportion of them did know quite a bit more about the process. The goal for this project is not necessarily to educate or raise awareness about reCAPTCHA, but to get people thinking on a larger scale about how we digitize physical objects. What is the process of taking a scanned copy of Alice's Adventures in Wonderland and then being brought to a particular phrase in that digital representation after Google searching for a quote from the book? Why that book? Why books instead of other cultural

[63] `https://github.com/jessefulton/captcha-servlet`

[64] `https://github.com/jessefulton/webpages-for-humans`

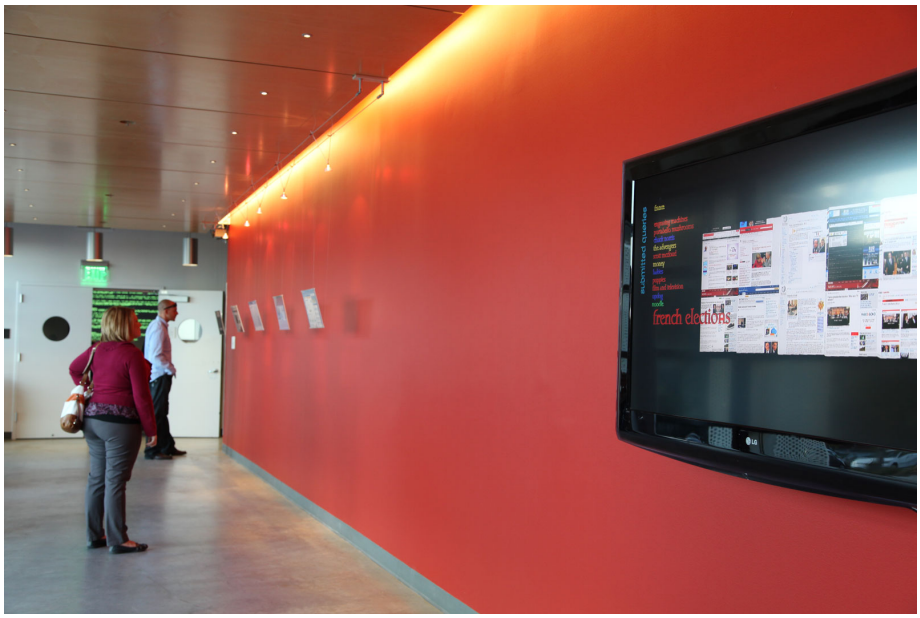
[65] a bookmarklet is a piece of JavaScript stored as a URL on a webpage which may be saved as a "bookmark" in a user's web browser. This allows the user to later execute that JavaScript functionality at any point in time simply by clicking on the link in the saved bookmarks.

[66] `https://github.com/jessefulton/phantomjs-hocr`

[67] `http://www.openframeworks.cc/`

objects, visual or non-visual? By making a "Human Web" the goal was to have visitors totally immersed in this "alternate" Internet, and question the possibilities of how we interact with the digital, and what it means to view a digital document. Is a digital document simply a combination of zeros and ones? Or does it have structure from which information can be gathered and analyzed (by both humans and computer algorithms)?

MFA Show Response



MFA Installation in the Digital Arts Research Center

People were very receptive to and genuinely interested in the concepts behind the two projects. Either they "got it" immediately, or after I explained it, they became very intrigued. There seemed to be a certain level of technical proficiency associated with the depth at which people connected with the pieces. There were a number of people who were incredibly excited by the ideas who seemed to have a very solid understanding of how software in general works, despite not necessarily being programmers (most simply worked with software and the Internet on a daily basis.) In this sense, I believe that my expectations of the primary audience being "normal" Internet users will be met and am looking forward to feedback received after publicizing the pieces online.

Interaction with the pieces at the show was less than I had hoped. A few people interacted with *Everybody's Google*, but far less than I had expected. Part of this may be due to the fact that cell phone reception was not very good in the building. I decided not to set up a separate wifi network for the project despite it being inaccessible from cruznet (the standard university wifi) due to the website running on a "nonstandard," blocked port. And the wall text (for both projects) did not overly emphasize the fact that these were software projects living on the web. Even with the addition of the web page URLs on the wall text, few people seemed to realize that these were both products or

manifestations of software/websites.

CAPTCHA-fy! and the *Serendipity Engine* received mixed responses. Because of the limited available space in the printer mechanism and the long duration of the show, the *Serendipity Engine* only dispensed paper roughly every eight minutes. I did not witness anybody remaining in the hallway area for more than eight minutes at a time, so it was in fact extremely serendipitous when a sheet fell from the ceiling and *was noticed* by a visitor. The "reading area" over which it was installed was very simple and visually unassuming. I believe this was a large factor in why most people did not even seem to notice it. The pages from the *Serendipity Engine* did accumulate on the floor, but people tended to either ignore them, or thought that they had been knocked over, and stacked them together into a neat pile on top of the *CAPTCHA-fy!* booklets. This created two problems: first, the next time a page fell unnoticed to the ground, the next visitor would pick it up and place it on the now-stacked pile of previous pages; second, the pile was placed on the footstool, on top of the *CAPTCHA-fy!* booklets, essentially rendering them invisible. There were a series of booklets along the window sill, but without the context of a series of 4 or 5 clearly marked as "DANM Copies", they looked more like decorative pieces and visitors were reluctant to touch them, let alone take one home[68].

While documenting, we timed the printer to dispense paper as people were walking by using a remote control and got great reactions. But again, it quickly devolved into people waving their hands at the ceiling, expecting a reaction from the magical printer in the ceiling. From this, I think it would've been important to have the piece be much more impactful rather than stressing its subtlety. People rarely noticed it - it may have been better to have it run for 10 or 15 seconds straight over the course of an hour or so. Another option could have been to use a scroll as opposed to individual sheets to stress the intentionality of the piece.

A Tangent: The Cubicle and The White Cube

> "Things become art in a space where powerful ideas about art focus on them."[69]

My primary medium as an artist is software. However, I often use the Internet as a medium as well, not simply in the sense that much of my work is encountered online, but it also re-uses and repurposes content found on the Internet, often hacking apart services and mashing them back together for unintended uses. Much of my work focuses conceptually on various aspects of Internet behaviour, communication, and surveillance. Questions arise when considering exhibition of such pieces, especially within the context of a physical exhibition: most often they are along the lines of "Why don't you just show the software itself at the

[68] The pages from the *Serendipity Engine* and the *CAPTCHA-fy!* booklets were intended to be takeaways. From my estimates, slightly more than half of the booklets were taken, and slightly less than half of the *Serendipity Engine* pages were taken (or possibly thrown away.)

[69] O'Doherty (1999, 14)

show?" Or if the piece in question exists on the Web, "Why don't you just set up a computer and let visitors click around on the website?"[70] I have two large concerns with these approaches.

A major problem with digital art is that it is often viewed merely as a gimmick or a technical demo. Rather than appreciate the work as a piece of art, the encounter devolves into a game of "how does it work?" And once that puzzle has been solved, it's on to the next one[71]. However, this seems to be the case only when such pieces are presented in the context of a museum or gallery. When encountered serendipitously or as a part of one's own explorations, most people engage with these types of projects much differently[72].

The second issue is one which I believe often applies to Internet-based art or net.art. The software I had created for the MFA show (and the majority of the software that I create) is not intended to be shown in The White Cube, but rather the Cubicle. It was designed to be encountered with an "internet mindset" - having just performed a google search or filled out a CAPTCHA. A key component to an artwork's aura is its "presence in time and space, its unique existence at the place where it happens to be."[73] However, the near-real-time data-transmission made possible by the Internet collapses our ideas of space and time. "Speed destroys space, it erases temporal distance. In both cases, the mechanism of physiological perception is altered."[74] With Internet-based works, the notion of space and time exists largely in a virtual form, constructed within ones own mind. In this sense, an altered state of mind changes the space in which the piece is encountered. Eliminating the experience of being **on** the internet ultimately decontextualizes the piece.

I have ideas which are compelling to me, and questions I try to grapple with. Software is a difficult thing to present in an (traditionally) aesthetically pleasing and approachable way. Perhaps code can have an aesthetic form, however, I'm not so much concerned with the code itself as much as I am with the processes and ideas the code represents. Although I have primarily produced software for these projects, code is not the only form in which I see these ideas manifesting. I'm more concerned with processes and systems than I am with objects, but social constructs inform us to focus on the isolated object of contemplation when inside a gallery. This is not the intention with the software I have written, but rather than fight those constructs, I've experimented with them and created physical objects and images for the MFA show. The goal was to condense the ideas into forms which could later expand and develop.

[70] The mere fact that a piece exists on the web does not mean it is a "web site" per se nor that one can "click around" on it.

[71] Obviously this is not always the case, but it occurs frequently enough that I do not feel like it's a gross miscategorization.

[72] Pieces encountered online (day-to-day browsing) or in a public space seem to have this effect

[73] Benjamin (1968)

[74] Huyssen (1994)

Conclusion

> "The Internet...is the first modern communications medium that expands its reach by decentralizing the capital structure of production and distribution of information, culture, and knowledge... This basic change in the material conditions of information and cultural production and distribution have substantial effects on how we come to know the world we occupy and the alternative courses of action open to us as individuals and as social actors. Through these effects, the emerging networked environment structures how we perceive and pursue core values in modern liberal societies."[75]

Google is much more than a search engine. It is a company which specializes in indexing, harvesting, and disseminating information. Google controls, creates, and distributes information in a multitude of forms, and it does this all incredibly well. But Google also has near-monopoly status on search, and is extremely competitive with other web-based services such as email and video. Google *is* the Internet to many people and it's one of the primary gateways to all things digital. Its laws and rules influence how we can interact with the Internet – they directly shape our understanding of the space of the Internet, what we can do there, and what the Internet can be. But all of these rules and laws are software algorithms. "There can only be one interpretation of every piece of code. Unlike people, computers are not able to guess or interpret a meaning if it's not stated exactly."[76] There is no ambiguity in code, and there is no such thing as nature in software - everything is explicitly designed and man-made. All rules and regulations are put in place intentionally - they are nothing other than political. To quote Lawrence Lessig regarding software design, "its architecture is its politics."[77]

As dependence on digital interfaces and software grows, these issues will become ever-more prevalent. The hardware interfaces and software algorithms of today will shape the the cultures and networked societies of tomorrow. In accordance with the *exception culturelle* clause, it is important to "extend the principle of plurality from the level of opinion to that cultural expression" in order to protect and promote cultural goods and practices[78]. Respecting and acknowledging the various cultural perspectives from an algorithmic viewpoint is just as important as respecting them in law. The focus on speed, efficiency and profitability needs to shift towards a more human-centric viewpoint. Enhanced functionality such as personalization or digitization technologies must be viewed just as critically for the information they present as they are viewed for the information which they omit. With a trend towards information consumption through digital networks, software algorithms will have the growing ability to impact the development of individual autonomy on a cultural and societal level.

[75] Benkler (2006, 30)
[76] Reas et al. (2010, 15)
[77] Lessig (2006)
[78] Rieder (2009, 144)

Nobody owns the web, yet is inherently a privatized public space. Public in the sense that the majority of sites do not have barriers to entry (such as subscription-based sites) but private in the sense that at some point in the process of visiting a website, a private corporation will become a method of data transport - whether it is the site and server itself, the Internet Service Provider, or the companies maintaining the hardware infrastructure underlying it all. As long as the web remains private, there is a lack of accountability, and private interests have powerful control over the flow of information under the guise of making it free and widely available.

At the risk of sounding too pessimistic, I look forward to the future. Shortly, when realtime 3D scanning and affordable 3D printing become widely available, the underlying systems of representing and modeling the world, along with the mechanisms for interacting with those systems, will radically change the way we interact with and think about the digital. And, in the past year alone, there have been great improvements by companies such as Google at addressing many of the concerns addressed in this paper and being more transparent about where their information comes from[79][80] (but there have also been a few missteps[81]). The great thing about working in software and using the Internet as a medium is that it is constantly changing, sometimes from a day-to-day basis. What I've done in the past year may be irrelevant next month, at which point, I will adapt and learn to think about and address a whole new set of issues and ideas. Another case in point of Google not just consuming data, but producing it and influencing ideas of individuals and how we think about and understand the world. This is only the beginning.

[79] http://googleblog.blogspot.com/2012/01/search-plus-your-world.html

[80] http://www.google.com/ads/preferences/html/about.html

[81] http://www.washingtonpost.com/blogs/compost/post/googles-no-opt-out-privacy-changes-and-the-end-of-the-anony
2012/01/25/gIQAtZuUQQ_blog.html

Appendix

Everybody's Google Personas

Age	22
Gender	Female
Location	Atlanta, GA
Seed URLs	http://rapfix.mtv.com/ http://www.bet.com/news/national.html http://www.bet.com/news/politics.html http://hightimes.com/ http://www.tumblr.com/tagged/hip+hop http://streetpeeper.com/news http://perezhilton.com/ http://www.rottentomatoes.com/news/

Table 1: *Everybody's Google* Persona 1 Information

Age	33
Gender	Male
Location	London, UK
Seed URLs	http://www.bbc.co.uk/news/uk/ http://landofllostcontent.blogspot.com/ http://www.guardian.co.uk/world/us-politics http://www.bbc.co.uk/sport/0/football/ http://uk.pokernews.com/news/ http://gizmodo.com http://techcrunch.com/ http://www.joystiq.com/tag/@breaking http://news.google.co.uk/news

Table 2: *Everybody's Google* Persona 2 Information

Age	62
Gender	Male
Location	Lynchburg, VA
Seed URLs	http://www.foxnews.com/politics/index.html http://www.foxnews.com/us/index.html http://www.foxnews.com/entertainment/music/index.html http://www.cmt.com/news/ http://www.cbn.com/cbnnews/ http://news.google.com/news/section?topic=b http://www.economist.com/ http://www.forbes.com/ http://www.drudgereportarchives.com/dsp/links_recap.htm

Table 3: *Everybody's Google* Persona 3 Information

Age	36
Gender	Female
Location	Santa Cruz, CA
Seed URLs	http://www.npr.org/sections/politics/ http://www.npr.org/sections/us/ http://www.npr.org/music/ http://rhizome.org/editorial/ http://sfmoma.tumblr.com/ http://front.moveon.org/ http://front.moveon.org/category/war-on-women/ https://www.google.com/search?q=veganism http://www.themarthablog.com/ http://family.go.com/ http://www.momcentral.com/blogs

Table 4: *Everybody's Google* Persona 4 Information

Github Projects

Everybody's Google

The code for the `everybodysgoogle.com` is available at `https://github.com/jessefulton/everybodys-google`. The code for the original version of the project (consisting of a website and Chrome browser plugin) may be viewed at `https://github.com/jessefulton/google-views`.

Ocular Character Recognition

Node Captchafy (`https://github.com/jessefulton/node-captchafy`) is the JavaScript library used to generate CAPTCHA styled images. The source code for `webpagesforhumans.com` can be found at `https://github.com/jessefulton/webpages-for-humans`. Code used to CAPTCHA-fy PDF files can be downloaded from `https://github.com/jessefulton/phantomjs-hocr`. And the updated code for the web interface hosted at `nodecaptcha.herokuapp.com` is on github at `https://github.com/jessefulton/node-captcha-site`.

Miscellaneous

Two other versions of the PDF CAPTCHA-fication software were created. One of the abandoned projects is at `https://github.com/jessefulton/tesseract-pdf`. The second one is not online, but was a result of work contributed to OpenFrameworks at `https://github.com/jessefulton/ofxTesseract`. Finally, an abandoned Java version of the CAPTCHA software can be found at `https://github.com/jessefulton/captcha-servlet`.

Thesis Paper

A copy of this paper in its raw and completely incoherent forms can be found at `https://github.com/jessefulton/thesis-paper`.

Software and Frameworks Used

Tessseract - http://code.google.com/p/tesseract-ocr/
- NodeJS - http://nodejs.org/
- socket.io - http://socket.io/
- WebGL - http://get.webgl.org/
- Three.js - http://mrdoob.github.com/three.js/
- HTML5 Canvas - https://developer.mozilla.org/en/HTML/Canvas
- PhantomJS - http://phantomjs.org/
- CasperJS - http://casperjs.org/
- ImageMagick - http://www.imagemagick.org/
- Google Chrome - https://www.google.com/intl/en/chrome/browser/
- Amazon EC2 - http://aws.amazon.com/ec2/
- MongoDB - http://www.mongodb.org/
- Redis - http://redis.io/
- Git/Github - https://github.com/

Figures

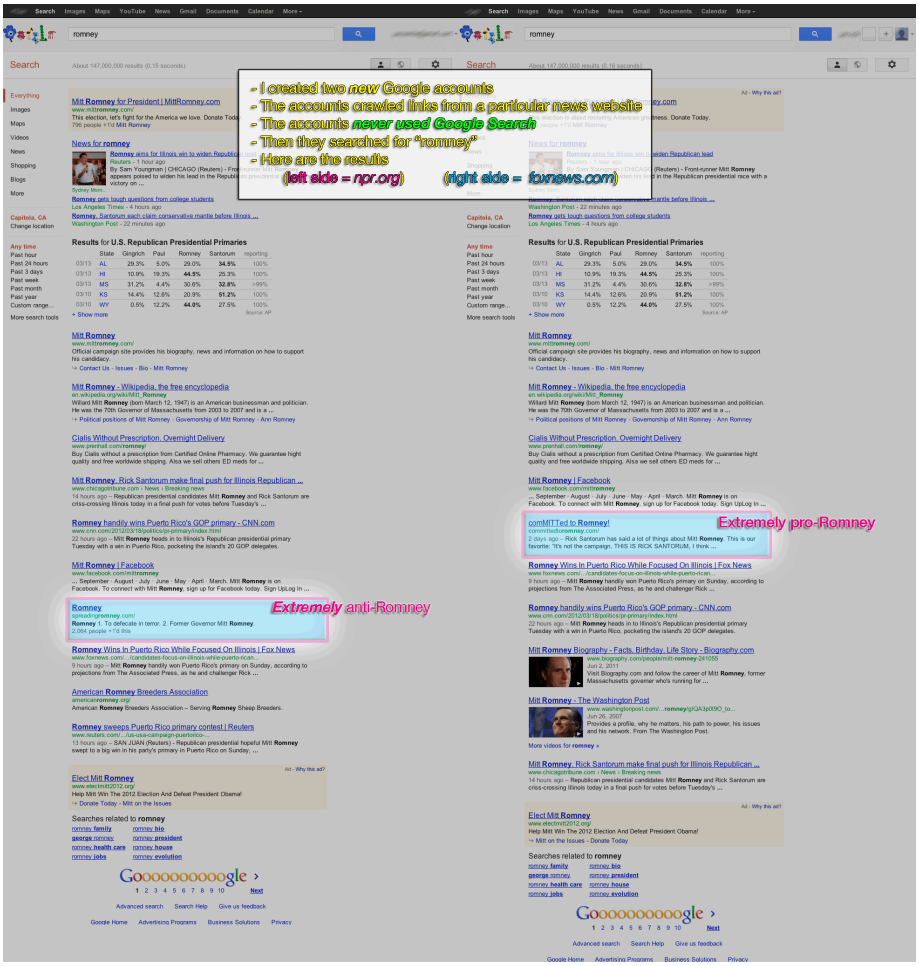


Figure 1: Initial testing results from *Everybody's Google*

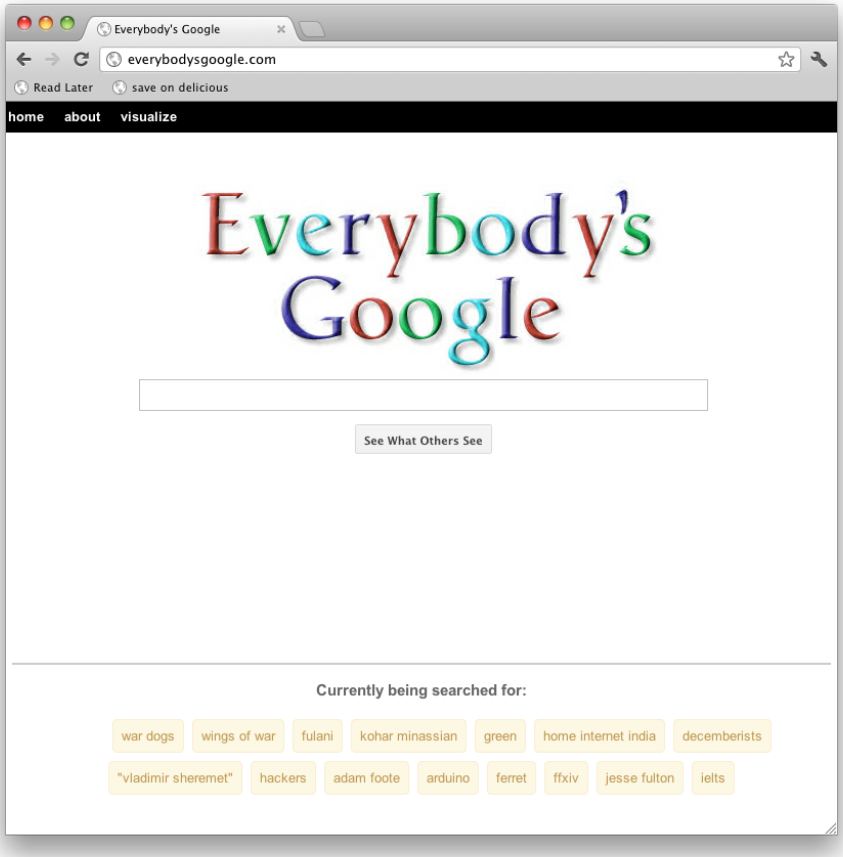


Figure 2: *Everybody's Google* home page

Figure 3: *Webpages for Humans* screenshot

References

Un universal declaration on human rights.

Benjamin, W. (1968). The work of art in the age of mechanical reproduction. In H. Arendt (Ed.), *Illuminations*. Shocken.

Benkler, Y. (2006). *The Wealth of Networks*. New Haven: Yale University Press.

Borges, J. L. (2000). The library of babel. David R. Godine.

Brin, S. and L. Page. The anatomy of a large-scale hypertextual web search engine.

Bush, V. (1945). As we may think. *The Atlantic Monthly*.

Curtis, T. (1988). The information society: A computer-generated caste system? In V. Mosco and J. Wasko (Eds.), *The Political Economy of Information*. The University of Wisconsin Press.

Ensemble, C. A. (1993). *The Electronic Disturbance*. Autonomedia.

Gere, C. (2006). *Art, Time, and Technology*. Berg.

Huyssen, A. (1994). *Twilight Memories: Marking Time in a Culture of Amnesia*. Routledge.

Introna, L. and H. Nissenbaum (2000). Shaping the web: Why the politics of search engines matters.

Lessig, L. (2006). *Code 2.0*. Basic Books.

Lewitt, S. (1967, June). Paragraphs on conceptual art". *Artforum*.

Manovich, L. (2011). There is only software.

Marko Petkovsek, H. S. W. and D. Zeilberger (1997). *A=B*. Addison Wesley.

O'Doherty, B. (1999). *Inside the White Cube: The Ideology of the Gallery Space*. University of California Press.

Oscar H. Gandy, J. (1988). The political economy of communications competence. In V. Mosco and J. Wasko (Eds.), *The Political Economy of Information*. The University of Wisconsin Press.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*.

Pasquinelli, M. (2009). Google's pagerank. In K. Becker and F. Stalder (Eds.), *Deep Search: The Politics of Search beyond Google*, pp. 152–162. Vienna: World-Information Institute.

Rawls, J. (2001). *Justice as Fairness: A Restatement*. Harvard University Press.

Reas, C. and B. Fry (2007). *Processing: A Programming Handbook for Visual Designers and Artists*. The MIT Press.

Reas, C., C. McWilliams, and LUST (2010). *Form + Code in Design, Art, and Architecture*. Princeton Architectural Press.

Rieder, B. (2009). Democratizing search? In K. Becker and F. Stalder (Eds.), *Deep Search: The Politics of Search beyond Google*, pp. 133–151. Vienna: World-Information Institute.

Rogers, R. (2009). *The End of the Virtual: Digital Methods*. Amsterdam University Press.

Rose, C. (2009). A conversation with eric schmidt, ceo of google.

Stalder, F. and C. Mayer (2009). The second index: Search engines, personalization and surveillance. In K. Becker and F. Stalder (Eds.), *Deep Search: The Politics of Search beyond Google*, pp. 98–115. Vienna: World-Information Institute.

Stiegler, B. (1998). Leroi-gourhan: L'inorganique organisé. *Cahiers de Médiologie 6*(2).

The United Nations Human Rights Council (2011). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. The United Nations Human Rights Council.

Vaidhyanathan, S. (2011). *The Googlization of Everything*. University of California Press.

Williams, R. (1976). *Keywords*. Oxford University Press.

Zapler, M. Mike lee calls for closer look at google.