Information Services & Use 35 (2015) 23–29
DOI 10.3233/ISU-150764
IOS Press

The user of the future: Reimagining how we think about information

Kalev Hannes Leetaru
Senior Fellow, George Washington University, Center for Cyber and Homeland Security,
2000 Pennsylvania Ave, NW, Washington, DC 20052, USA
E-mail: Kalev.leetaru5@gmail.com; URL: http://kalevleetaru.com/

Abstract. We live in world today where the average citizen can zoom into the planet from Google Earth, walk down a virtual street in Google Street View, access live imagery and video on breaking events from remote regions via YouTube and Instagram, and search 100 petabyte indexes that learn and evolve to tailor themselves to hundreds of millions of individual users and even adjust the material returned based on our daily life patterns, calendar, and present location. Searches look across hundreds of millions of distinct archives and deliver their results via APIs and advanced analytics accessible through mobile devices while we walk down the street. In this brave new world of information access, what does the information platform of the future look like and how does the user of the future think about, access, consume, and interact with this reimagined world of information? This transcript of the NFAIS 2015 keynote opening address explores the user of the future and what we can learn from current trends to reimagine how we think about information and how it is delivered.

Keywords: User interface, information access, data mining

The world today is awash with data. Facebook alone claims to have over one billion members connected by over one trillion friendship links that have collectively uploaded more than 240 billion photographs to the service over its lifetime. Every day more than 2.5 billion new items are uploaded to Facebook totaling more than 500 terabytes, including 300 million photographs. The growth of social media far outpaces traditional media: there are as many words posted to Twitter every day as have appeared in the entire *New York Times* in the last half century, and every month there is as much video uploaded to YouTube as the combined output of all three major U.S. television networks in their entire histories. It is likely that by the close of 2015 there will have been as many words posted to Twitter as in all of the books published worldwide over the last half-millennium. The vast majority of all this data is either freely-accessible on the open web or owned by a handful of new media companies in Silicon Valley, rather than by traditional publishers. What does this mean for the role of professional information services in the Internet era as companies like Google seem to be on a march to index the entire planet? The answer has a lot to do with how organizations understand their users.

Turning back the clock a moment, this year marks a fascinating milestone that goes right to the heart of this conference. It was 50 years ago this year that the earliest origins of the Dialog system were built, following on the work of Doug Englebart at SRI International (formerly the Stanford Research Institute) two years prior and related work at the Massachusetts Institute of Technology (MIT), Harvard, and SDC all happening right around that period. It has been a little more than half a century since the development of what we would think of as the modern client-server database information retrieval system. What is most amazing to me is that apart from a fancier hardware interface and larger and faster datasets, the basic search experience has changed little in that half century since those very first search systems. Think about that for a moment. The majority of professional information platforms today in 2015 are little different than those pioneering systems of half a century ago – a user sits down at a computer terminal in her office and enters a carefully-constructed set of keywords using extensive domain knowledge to align as closely as possible to the exact wording of the original source material. For his/her trouble the user receives back a massive bulleted-list of search results that requires "next paging" for the rest of the afternoon to find the actual piece of information desired. To me this is truly amazing that half a century later with access to absolutely unprecedented volumes of material and unimaginable computing power we're still thinking in terms of text and using keywords and exact text-matching to search all of it.

Of course, even major web search engines like Google are based on keyword searches. However, that's where things get really interesting. You see, Google is not just a search engine that we pull up a few times a day to conduct a targeted query and then log out. Google has really become part of the fabric of our lives. It is frequently faster to look something up on Google than it is to try and remember it ourselves. Many of us use Google's Gmail service for our personal email accounts, Google Calendar for our appointments, rely on Google Maps to try and find everything from our next meeting to a new restaurant for a family night out, and access all of it through a Google Android-based cell phone. In essence, Google has the holy grail of all search services – it knows us inside and out across our professional and personal lives. Through Gmail, Google knows our plane, train, and hotel reservations, through our Android cell phone, it knows where we are at this moment as well as our daily travel routines, through Google Calendar it knows where we are supposed to be 10 minutes from now, and through Google Maps it knows that we're looking for last-minute dinner reservations in the vicinity of our meeting two hours from now. In short, Google has the most valuable kind of information to assist with our increasingly-mobile and personalized searches: detailed knowledge and understanding of who we are and what we want throughout our day. When we plan a trip, Google Maps even personalizes its map display to label our various transportation reservations so we know when we will be at each location and knows that if our flight leaves later this morning and we search for "SFO", it should show us the current status of our flight and a link to print our ticket, not return the homepage of San Francisco International Airport. Through its massive array of advertiser networks Google even knows a large fraction of the websites we visit, how we navigate through them, and the searches we used to get there.

Yet, instead of displaying all of this to us in a science fiction-like dashboard with thousands of technical indicators, Google wields this vast knowledge quietly and transparently in such a way that we are not even aware of how much it is customizing our search experience. Take the example of someone who has just flown to DC from San Francisco and is walking down M Street in Georgetown and pulls out her phone to search for "pizza". Google knows that he/she likely wants to see a list of nearby pizza restaurants and their menus and reviews, not Pizza Hut's latest earnings reports. If the person tends to search for certain kinds of pizza back home in San Francisco, such as deep dish pizza, Google might even prioritize deep dish restaurants in their search. Conversely, the Wall Street Analyst at her desk in New York City who is deep in the midst of a sequence of searches for earnings reports of various restaurant chains who types "Pizza Hut" is more likely to want to see a link to its latest financial report, not a menu of the afternoon's specials at the Pizza Hut store 50 miles away. Most incredibly, despite serving a customer base of billions of users, Google individually learns the preferences of every single one of its users and customizes its results for each of them in real-time, constantly refining, filtering, and reranking the results on a person-by-person and search-by-search basis. Google has become the go-to search service because it has created a "frictionless" interface for searching the world. It intuitively and transparently tries to deeply understand our information needs and behavior in order to act as the ultimate intelligent search agent.

Perhaps most powerfully, Google and its Silicon Valley kin drive relentless and continuous change, focusing all of their resources on the ongoing reinvention of themselves: A/B testing and data-driven design informs nearly every decision. These companies do not hire a graphic design firm, hold a series of focus groups, issue a request for proposals, and finally roll out a new technology platform three years later. They implement change as a perpetual sequence of micro-improvements, rolled out tens or sometimes hundreds a day, and tested in the real world on the scale of their entire customer base. Their success also derives from thinking of the entire lifecycle of how people access information, rather than focusing purely on user interface design. For example, a search service that rolls out a new "responsive" web template for mobile may have a fantastic-looking version of their search interface that works great on every mobile device currently used, but overlooks the fact that a user on a tiny screen trying to search while walking to their next meeting probably does not want to see fifty advanced search buttons or huge scrolling pages of interactive responsive taxonomic trees. BMW learned this the hard way when they first rolled out their iDrive system that worked fantastically in the lab setting, but they did not appreciate the fact that someone barreling down the interstate at 65 MPH might need to just quickly turn on the defrost and not want to take their eyes off the road to wade through five levels of interactive animated clickable menus to do so. It is pretty amazing when you see the results from A/B testing compared side-by-side with the results of a traditional design-build-deploy approach. Things like the most subtle difference in what shade of white a sidebar is can have a huge impact on usability, while the flow of users through a system can completely hinge on the placement of a single label.

Things become even more interesting when we think about the difference between Facebook and LinkedIn. Google has gotten to where it is by becoming a master of understanding our personal information needs as we go about our daily lives. Similarly, Facebook has become the social network of our personal lives, while LinkedIn connects our professional lives. Think about the user interfaces, the workflows, the features of Facebook versus LinkedIn: each is highly-customized to the very different needs of our personal and professional lives. Historically, Facebook has emphasized spur-of-the moment immediacy and the sharing of rich multimedia content, while LinkedIn has emphasized contemplation and connections. Most professional information service platforms, on the other hand, enforce a "one size fits all" approach to interface, offering a "basic" and "advanced" search interface to which every user, regardless of field or search need, must conform. To remain relevant in the emerging era of information access, platforms must offer "frictionless" interfaces and recognize that, for good or bad, Google, Facebook, LinkedIn and their kin have set the interface standards to which all search products are held. Even more importantly, however, is that the way people use information products has fundamentally changed. Internet search engines have accustomed users to seamless meta-search – a single search box to query every available piece of information, rather than a myriad of walled gardens that have to be individually searched as is the norm in the professional information services world. Users demand intuitive and transparent search interfaces that do not require them to sit down and work with a reference librarian for half a day to utilize their advanced search features. Think for a moment about the "saved search" feature that many professional information services offer. Users can bookmark frequent searches and run them again as needed. Google does this too, but it is transparent – Google realizes as a user starts typing that she wants to rerun one of her past queries, fills the rest of the query in for her, and even recommends the links that she found most useful last time.

Another common trend of Silicon Valley companies is that they recognize that they cannot themselves be the source of all innovation and think up every great idea for their products. Instead, they build platforms for others to use to build those great ideas for them for free. The latter part of that statement is particularly novel: by building software developer platforms, these companies enable millions of developers to build millions of tools for them at no cost to their companies. Social media companies today view themselves nearly identically to professional information services companies: exclusive holders of vast archives of incredibly unique and highly-valuable data with enormous commercial potential. What is so unique about them is that they let their user communities themselves build their own ultimate tailored interfaces. In essence, while professional information services companies pay for all of their own development work and try to prioritize customizations for each vertical market, social media companies outsource all of that work to their users, getting them to do a huge amount of the development work in building a relentless source of new interfaces for their platforms.

Just one year after launching its developer's platform in 2010, Facebook's 800 employees were bolstered by over 400,000 third party developers that had created more than 33,000 applications for Facebook. That's right – Facebook had nearly half a million developers working for it, but only had to pay the salaries of 800 and got more than 33,000 customized interfaces built targeting every vertical and use case imaginable. Last year Twitter launched its Data Grants program to seed high-risk highly-innovative applications of its data, while LinkedIn recently launched its Economic Graph Challenge. NetFlix famously used its NetFlix Prize to get over 20,000 teams from 150 countries to sign up to tackle trying to make its recommender service better. In just nine months 2,000 teams had submitted over 13,000 solutions, besting NetFlix's own algorithms by 10%. The first winner reported over 2,000 hours of development work creating a final combination of 107 algorithms working together. In the case of NetFlix, it offered a million dollar prize, but in each of these other cases there was no money being spent for all of that development effort.

How might this model work for professional information services companies? Think about Google Cloud for a moment – it only provides access to hardware, not data. It is possible to rent petabytes of storage and hundreds of thousands of processors, but you can not rent access to Google's datasets – it is a strictly "bring your own data" model. Now imagine information companies. They have vast archives of incredibly unique one-of-a-kind data, but the real bottleneck is access. A common refrain from their users is the desire for direct API access to their underlying data stores to run data mining algorithms or other sophisticated analyses. What if companies created their own miniature clouds that combined a custom-tailored computing platform built for the needs of their customers' industries, with the incredible datasets they hold? For an additional fee, their customers can essentially rent a set of machines running physically on their premises in their private cloud that have access to specific datasets and run data mining algorithms on that data or build custom interfaces to that data. None of the data ever leaves the premises, but users are able to run any analysis imaginable. Developers could even build their own customized interfaces to those data platforms, creating custom-tailored ecosystems.

What might this look like in real life? In collaboration with the Internet Archive, we developed what we call the "Virtual Reading Room" model, which is essentially a framework for allowing precisely this kind of cloud-based computing where the raw data never leaves the publisher's premises. In this case, I had an interest in exploring the geography of American television news. According to a Pew report, three quarters of Americans still get their news from television. What does American television news tell us about the world? What events make it and what events do not? Identifying and disambiguating geographic mentions in half-a-million hours of television closed-captioning required applying enormously complex algorithms that required raw access to the closed-captioning streams. However, those streams could not leave the Archive's physical premises. So, all of the geographic processing was run on a set of virtual machines physically located on the premises of the Archive. Instead of the traditional model of a web-based file server where a user downloads data locally to work on it, here the analysis is uploaded to run alongside the data. Only the final computed metadata – a list of locations – left the Archive and was used to create an interactive mapping interface. The resulting map generated a tremendous amount of buzz about IA's television archive and got a lot of people in the television industry talking about mapping interfaces to their archives.

Many professional search platforms are text-only, but for those that have rich material that includes visual content like digitized page scans, think for a moment about all of the images in that material. In fall 2013 I approached the Internet Archive with an idea. What if we inverted the concept of the book? Instead of thinking of books as collections of words, what if we thought of them as containers of images? If you think about it, if you took all of the world's books and pulled out every image from every page of every book, you would have one of the greatest art collections in the history of the world. I certainly was not the first to think about extracting the images of books and there have been lots of previous projects, perhaps most famously the British Library's collection. However, all of these past efforts were relatively small-scale. What I was interested in is what happens when you take this to full-scale. The Internet Archive's collection spans 600 million pages from 500 years of books scanned from over 1,000 libraries worldwide. The other key piece lay in thinking about how users would interact with all of these images. All of the previous image projects I know of were substantively driven by technical folks who approached it as a technical challenge. To me, it was about reimagining the book – focusing on the content first and technology second. Most past projects simply created massive piles of images and either posted them as giant ZIP files or shoveled them into browse-only digital libraries. Manually browsing a massive thumbnail gallery of tens of millions of images is not very useful. What makes Google Images so powerful? It is that you can actually search the collection using keywords. Thus, I focused on finding ways of extracting the text surrounding each image so that they could be made keyword searchable.

We live in a world today where the average citizen can zoom into the planet from Google Earth, walk down a virtual street in Google Street View, access live imagery and video on breaking events from remote regions via YouTube and Instagram, and search 100 petabyte indexes that learn and evolve to tailor themselves to hundreds of millions of individual users and even adjust the material returned based on our daily life patterns, calendar, and present location. Searches look across hundreds of millions of distinct archives and deliver their results via APIs and advanced analytics accessible through mobile devices while we walk down the street. In this brave new world of information access, what does the information platform of the future look like and how does the user of the future think about, access, consume, and interact with this reimagined world of information?

The information world of today is profoundly different from the iconic image of the solitary scholar sitting at a desk for years among piles of books and assisted by human librarians intimately familiar with every book in their collection. Discovery today must reach across petabytes spanning disciplines, geography, and modalities and is increasingly machine-assisted as search engines literally "learn" the interests, access preferences and discovery styles of each of their hundreds of millions of individual users, constantly evolving with each search to deliver ever-more personally tailored results. From Google Now to Siri to Cortana, automated virtual assistants increasingly integrate contextual knowledge about their user's daily lives into the search process, directing a restaurant search to a live map and phone number for the user walking nearby, a menu and reviews to a user at her desk, and financial statements to the Wall Street trader. Information is increasingly consumed on-the-go via mobile devices with unique interface demands and in novel use scenarios. Yet, this must be balanced with the need for precision, reproducibility, and exactness demanded by the professional searcher. Unified search has become the norm, with users growing accustomed to services like Google Scholar that transparently unify hundreds of thousands of sources into a single meta-search interface offering one-click authentication via institutional subscription.

Information is itself increasingly a platform – Google, Facebook, Twitter all provide APIs and allow third party developers to build new tools and services on top of their archives, leveraging the creative energy of millions of outside developers to create entrepreneurial ecosystems around their brands. The "consumerification" of data mining means that users are no longer satisfied with the classic search result of a bulleted list of millions of articles mentioning the keyword "drone" – they want to make a timeline of coverage (temporal analysis), see the topics mentioned alongside of it (semantic/contextual analysis), create a network diagram of who's talking about it (network analysis), visualize a map of where drones are being talked about (geographic analysis), and sort by emotional context to separate articles lauding drones from those criticizing them (emotional/sentiment analysis). At the same time, the printed word is no longer the central access point of information: users accustomed to the post-textual world of live imagery and video from Instagram and YouTube are demanding rich multimedia access. More advanced users demand direct access to entire archives totaling petabytes to run sophisticated data mining algorithms. As research becomes increasingly collaborative and interdisciplinary, scholars demand the ability to share information with a single mouse click and to access and bring together information from across sources at the atomic level of individual passages, tables, images, and concepts. In short, "information" is no longer defined by keyword searches of walled gardens returning bulleted lists of documents – it is about helping users make sense of all of that information and delivering users the pinpoint specific pieces of knowledge they need to make decisions.

In our post-textual, always-connected, mobile on-the-go, real-time, unified, individually-tailored, analytic world, what does the concept of "information" look like in the future, how will users interact and consume it, and how do we evolve and adapt the model of today's professional information service into the coming decade? One trend is for certain: Silicon Valley is largely defining what this world will look like and has a lot to teach us in how to get there and when we do get there, it will look a lot more like Hal9000 than it will the mainframe keyword prompts that have been the mainstay of the industry for more than half a century.

About the author

One of Foreign Policy Magazine's Top 100 Global Thinkers of 2013, Kalev H. Leetaru is a Senior Fellow at the George Washington University Center for Cyber & Homeland Security and a member of its Counterterrorism and Intelligence Task Force, as well as a Council Member of the World Economic Forum's Global Agenda Council on the Future of Government. From 2013–2014 he was the Yahoo! Fellow in Residence of International Values, Communications Technology & the Global Internet at Georgetown University's Edmund A. Walsh School of Foreign Service, where he was also adjunct faculty. Kalev has been an invited speaker throughout the world, from the United Nations to the Library of Congress, Harvard to Stanford, Sydney to Singapore, while his work has appeared in the presses of more than 100 nations and from *Nature* to *The New York Times*. In 2011 *The Economist* selected his Culturomics 2.0 study as one of just five science discoveries deemed the most significant developments of 2011, while the following year HPCWire awarded him the Editor's Choice Award for Edge HPC (High Performance Computing) "representing the highest level of honor and recognition given to the thought leaders in the

HPC community" and in 2013 noted "his research helped usher in the era of petascale humanities", while later that year his work on Twitter was recognized by Harvard's Neiman Lab as the top social media study of 2013.