



Search...

Go →

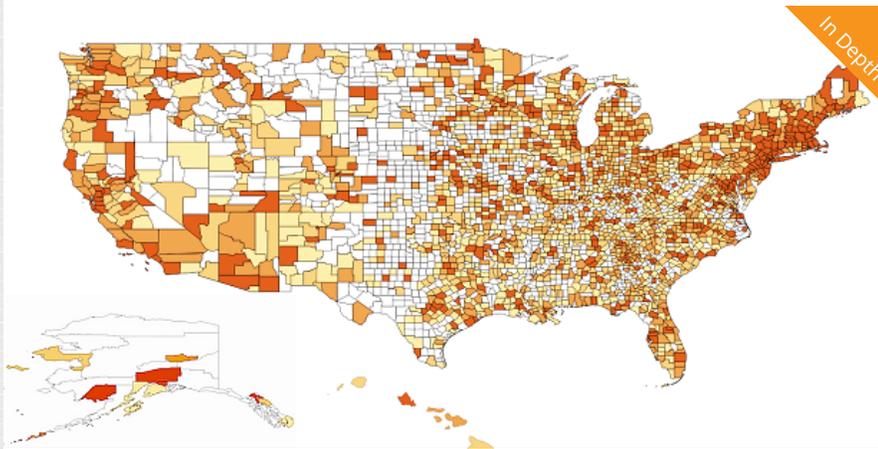
About Us

Publications

Blog

Events

Resources



Published on September 10th, 2014 | by Travis Korte

1

# Wikipedia Edits Reveal America's "Data Deserts"

For many, the concept of the "data revolution" conjures up images of online transactions, genomic medical data, and ubiquitous sensor networks. But user-generated data sets, such as social media updates, Flickr photos, and blog posts, are a major part of the data revolution as well. Internet users post **100 hours of video to YouTube every minute**, and tweet around 600 million times a day. These data sets reflect the lived experiences of millions of individuals, and collectively provide valuable information about many different communities.

User-generated data sets have been used for many purposes, from **using Twitter data to detect earthquakes** to **using Flickr and YouTube data to forecast political and economic attitudes**. For example, a number of recent initiatives have used user-generated data for public health purposes. The United Nations' Global Pulse initiative has **mined web search data** to detect non-communicable diseases such as cancer and diabetes, health officials in New York City have used Yelp reviews to **track and respond to public health outbreaks**, and Penn State University researchers have used Wikipedia search and click data to **predict outbreaks of illnesses**. The uses of user-generated data will only continue to grow in the future as they offer an enormous supply of real-time data.

But not all communities are equally represented in these data sets. Unequal access to broadband service, variations in access to technology, disparities in the level of digital literacy, and a host of other factors can influence who is included in the data and who is not. When communities are not represented (or underrepresented) in the data, decisions made after analyzing this data may overlook members of these communities and their unique needs.

As a result, the amount of user-generated information produced in a particular area of the country can serve as a bellwether for how much that community is able to realize the benefits of the data revolution. As my colleague Daniel Castro describes in a **recent report**, these levels of inequality may even cluster geographically to give rise to "data deserts"—areas characterized by a lack of access to high-quality data that may be used to generate social and economic benefits.

As an initial attempt to measure data deserts, the Center for Data Innovation analyzed which areas of the United States contributed the most to Wikipedia. Since Wikipedia is an entirely crowd-sourced project, the scope, depth, and accuracy of its articles depends on the engagement and interests of its users. Uneven contributions have had bizarre results: For example, **the Wikipedia article "List of**

Back to Top ↑

Sign up for our weekly newsletter

Email Address

Subscribe

Popular

Recent



No, Algorithms Do Not Hijack Elections

by Joshua New | posted on September 22, 2015



After a Rocky Start, the Internet of Things May Finally Succeed in Chicago

by Joshua New | posted on September 15, 2015



5 Q's for Adam Bonica, Cofounder of Crowdpack

by Joshua New | posted on September 7, 2015



5 Q's for Nirmal Govind, Director of Streaming Science and Algorithms at Netflix

by Joshua New | posted on September 28, 2015



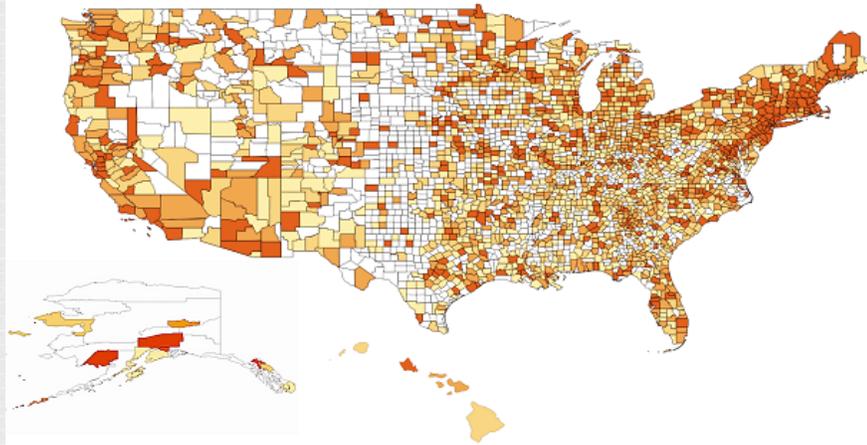
How to Promote Smarter Water Use by Giving Consumers Access to Their Consumption Data

by Daniel Castro | posted on September 6, 2015

**Advanced Dungeons & Dragons 2nd edition monsters** is longer than the article on **British literature**, and **the article on professional wrestler "The Undertaker"** has been revised **more times than has the article on global warming**. Similarly, a lack of data from particular communities might overlook important local insights and information known to members of that community.

To find Wikipedia's data deserts, we collected 134,958 edits to the English-language edition of the online encyclopedia Wikipedia made from July 19-23 and August 8-26 2014, and mapped them at the county level, adjusted for population. These edits represent public contributions from unregistered users in the United States. Since anyone can edit Wikipedia, gaps in this data set indicate areas where people have chosen not to participate in collaboratively editing the online encyclopedia.

## Findings



White indicates that no edits were recorded for that county, and darker color indicates higher edit rates.

Although in some places our findings track with population and population density estimates, despite adjusting for population, other areas diverge from those metrics. For example, the non-coastal western states exhibit an edit rate that far exceeds what would be expected based on population metrics alone, as do northern counties in New York, Vermont, New Hampshire, and Maine. Some well-documented technology hubs, like the San Francisco Bay Area, the Pacific Northwest, and the Washington, D.C.-to-Boston corridor, share a high edit rate with regions with less high-tech reputations, such as Florida and southern Arizona. Particularly prolific individual editors can come from anywhere, which introduces some noise into areas with generally few edits, but the barren vertical strip extending from west Texas up through western Oklahoma, Kansas, Nebraska, and the Dakotas still appears stark. These areas have a relatively **low population density** and **high median age**, which may contribute to the low per capita edit rates.

## Methodology

We created the data set of population-adjusted anonymous edits by county by downloading metadata for anonymous edits, geolocating the Internet protocol (IP) address associated with each edit, and identifying a county for each geolocation. We used the "recent changes" module of the **application programming interface** (API) for MediaWiki—the open source wiki software platform that contains the English Wikipedia as a subset—to download approximately 400,000 total edits. The API provides a range of information, including date, page title, and user data, for each edit. Edits submitted by anonymous users, i.e., those users who have not registered Wikipedia accounts, are also linked to an IP address. The IP address associated with each edit can be converted to a geolocation, i.e., latitude and longitude coordinates, using a range of online services. A very small fraction of the addresses we collected were formatted in IPv6, the newest version of the Internet protocol that **carries around four percent of Internet traffic** as of September, 2014. Available geolocation tools were not compatible with IPv6 addresses, so we filtered them out. We used freegeoip.net, a free API that takes IP addresses as input and outputs country, region, city, latitude and longitude, and other information. We filtered out IP addresses located outside the United States, as well as those addresses the geolocation tool was unable to resolve at the county level. IP geolocation is not exact, since some IP addresses reflect the locations of Internet service providers rather than users themselves. Still, geolocation services can typically place an IP address within a few miles of its origin, which is generally sufficient for

determining what county the address comes from.

After filtering out the unusable data, we then used the Federal Communications Commission's (FCC) **Census Block Conversions API** to match each latitude-longitude pair with a county, enabling mapping. We conducted all data processing in **R** and all mapping in **OGIS**.

Tags: [api](#), [counties](#), [data deserts](#), [wikipedia](#)

### About the Author



**Travis Korte** Travis Korte is a research analyst at the Center for Data Innovation specializing in data science applications and open data. He has a background in journalism, computer science and statistics. Prior to joining the Center for Data Innovation, he launched the Science vertical of The Huffington Post and served as its Associate Editor, covering a wide range of science and technology topics. He has worked on data science projects with HuffPost and other organizations. Before this, he graduated with highest honors from the University of California, Berkeley, having studied critical theory and completed coursework in computer science and economics. His research interests are in computational social science and using data to engage with complex social systems. You can follow him on Twitter @traviskorte.

### Related Posts



Fighting Child Labor with Open Data →



Addressing Inequality and the 'Data Divide' →



Wikipedia Releases Clickstream Data →



Create Maps of Wikipedia's Geotagged Articles →

1 Comment

Data Innovation Day

1 Login ▾

♥ Recommend

↗ Share

Sort by Best ▾



Join the discussion...



azael · 7 months ago

how can this help me.

1 ^ | ▾ · Reply · Share >

#### ALSO ON DATA INNOVATION DAY

WHAT'S THIS?

#### How Data and Analytics Can Help the Developing World

2 comments · a year ago

**DataDyne** — DataDyne's Magpi, originally funded by the United Nations Foundation but now a profitable social ...

#### How Can Policymakers Help Build the Internet of Things?

1 comment · a year ago

**John Alexander** — Daniel, The IoT can ONLY be harnessed properly if software is designed to manage the ...

#### After a Rocky Start, the Internet of Things May Finally Succeed in ...

2 comments · 23 days ago

**Joshua New** — Yes - this is good news! We've just updated the article to include this information.

#### Why Didn't Government Invent Uber?

1 comment · 7 days ago

**Hemant** — Fantastic question. High time atleast an Uber was viewed as an 'infrastructure app'. There is an even ...

✉ Subscribe

D Add Disqus to your site

🔒 Privacy

DISQUS

[Back to Top ↑](#)