SCIENTIFIC AMERICAN™

Guest Blog

What's Wrong with Open-Data Sites--and How We Can Fix Them

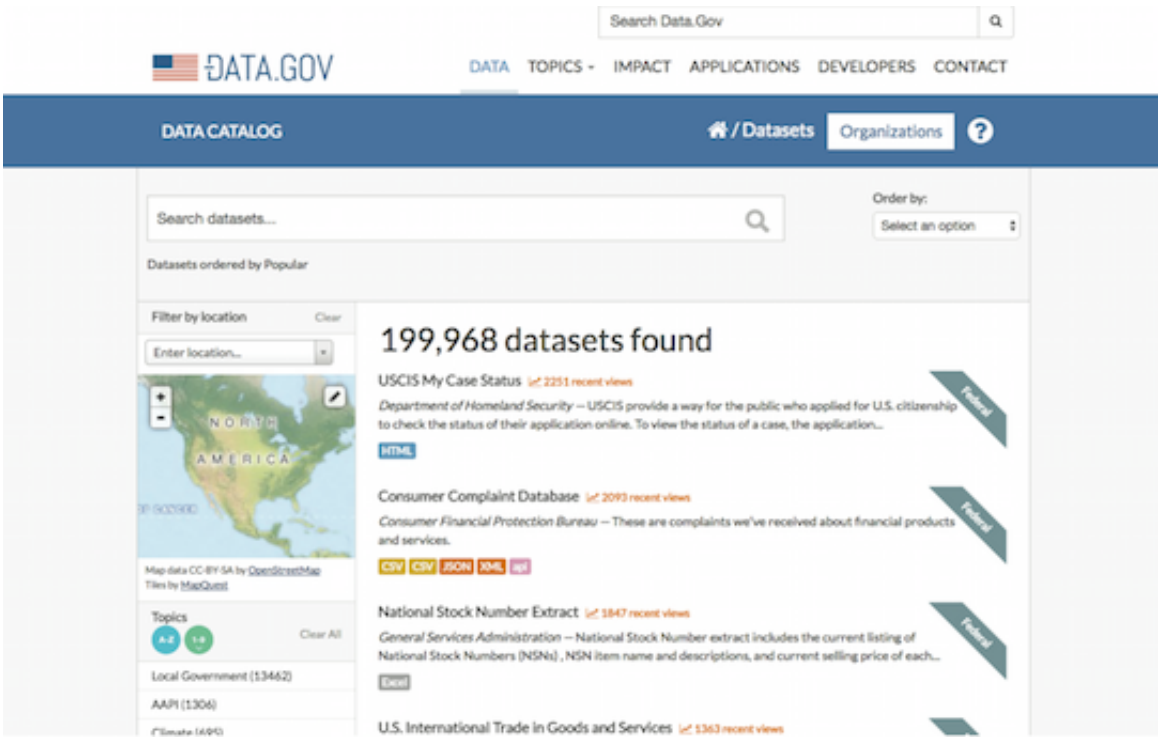
Vast amounts of useful information can be found on government Web sites, but it's often impossible to make sense of it

By César A. Hidalgo on May 2, 2016

Image by luckey_sun/Flickr under Creative Commons 2.0 license

Imagine shopping in a supermarket where every item is stored in boxes that look exactly the same. Some are filled with cereal, others with apples, and others with shampoo. Shopping would be an absolute nightmare! The design of most open data sites—the (usually government) sites that distribute census, economic and other data to be used and redistributed freely—is not exactly equivalent to this nightmarish supermarke. But it's pretty close.

During the last decade, such sites—data.gov, data.gov.uk, data.gob.cl, data.gouv.fr, and many others—have been created throughout the world. Most of them, however, still deliver data as sets of links to tables, or links to other sites that are also hard to comprehend. In the best cases, data is delivered through APIs, or application program interfaces, which are simple data query languages that require a user to have a basic knowledge of programming. So understanding what is inside each dataset requires downloading, opening, and exploring the set in ways that are extremely taxing for users. The analogy of the nightmarish supermarket is not that far off.



THE U.S. GOVERNMENT'S OPEN DATA SITE

The consensus among those who have participated in the creation of open data sites is that current efforts have failed and we need new options. Pointing your browser to these sites should show you why. Most open data sites are badly designed, and here I am not talking about their aesthetics—which are also subpar—but about the conceptual model used to organize and deliver data to users. The design of most open data sites follows a throwing-spaghetti-against-the-wall strategy, where opening more data, instead of opening data better, has been the driving force.

Some of the design flaws of current open data sites are pretty obvious. The datasets that are more important, or could potentially be more useful, are not brought into the surface of these sites or are properly organized. In our supermarket analogy, not only all boxes look the same, but also they are sorted in the order they came. This cannot be the best we can do.

There are other design problems that are important, even though they are less obvious. The first one is that most sites deliver data in the way in which it is collected, instead of used. People are often looking for data about a particular place, occupation, industry, or about an indicator (such as income, or population). If the data they need comes from the national survey of X, or the bureau of Y, it is secondary and often—although not always—irrelevant to the user. Yet, even though this is not the way we should be giving data back to users, this is often what open data sites do.

The second non-obvious design problem, which is probably the most important, is that most open data sites bury data in what is known as the *deep web*. The deep web is the fraction of the Internet that is not accessible to search engines, or that cannot be indexed properly. The surface of the web is made of text, pictures, and video, which search engines know how to index. But search engines are not good at knowing that the number that you are searching for is hidden in row 17,354 of a comma separated file that is inside a zip file linked in a poorly described page of an open data site. In some cases, pressing a radio button and selecting options from a number of dropdown menus can get you the desired number, but this does not help search engines either, because crawlers cannot explore dropdown menus. To

make open data really open, we need to make it searchable, and for that we need to bring data to the surface of the web.

So how do we that? The solution may not be simple, but it starts by taking design seriously. This is something that I've been doing for more than half a decade when creating data visualization engines at MIT. The latest iteration of our design principles are now embodied in DataUSA, a site we created in a collaboration between Deloitte, Datawheel, and my group at MIT.

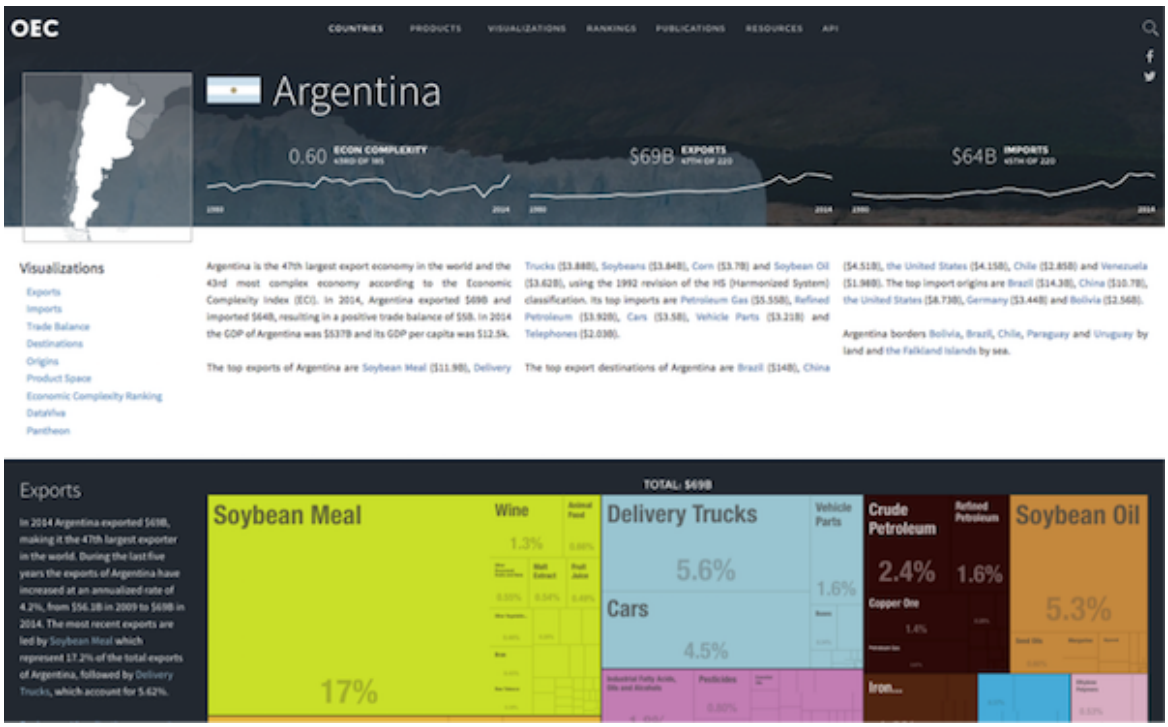
So what is design, and how do we use it to improve open data sites? My definition of design is simple. Design is discovering the forms that best fulfill a function. If they happen to also look pretty, that's a bonus, but usually good design will also be aesthetically pleasing since the forms that best fulfill functions express either the cleverness of simplicity or the mysteries of complexity. In the case of open data sites, what we want to make are tools that make data understandable to humans, but also, to the search engines that humans use to explore the web.'

Our solution so far has been the creation of sites that merge multiple datasets and transform them into stories. DataUSA merges data from the American Community Survey, the Bureau of Labor and Statistics, the Bureau of Economic Analysis, and the department of education, among other data sets, to create profiles that combine fast rendering visualizations and text for each state, county, metropolitan area, census designated place, industry, occupation, and university major.

The visualizations in DataUSA help humans understand what's inside each supermarket box of data. If the user then wants what's inside the box, they can now download that data directly or access it through our API. The text in DataUSA, which can also be read by humans, is intended primarily for search engines, and is partly written by algorithms that help bring the most important basic numbers of a dataset (maximums, minimums, and averages) to the surface of the web.

But does this approach work? In our experience the answer is a resounding

yes. Look at The Observatory of Economic Complexity (OEC), a tool we created for international trade data that was otherwise buried in the deep web. In its 3.0 version, the OEC now receives more than half a million visitors every month because it is the number one answer for searches such as “top exports of Argentina” or “what does China export?” By focusing on transforming data into stories, instead of hiding it behind dropdown menus, we brought trade data to the surface of the web and we now have a site that lives in symbiosis with search engines.



DataUSA 1.0 intends to do the same for some of the United States most important datasets. DataUSA surfaces data ranging from tuition costs, and wages, to data on commuting times, disease prevalence, and the linguistic and cultural origins of U.S. citizens and foreign born residents.

So going back to our supermarket analogy, what we have done with DataUSA is grab a heap of boxes and organized them into thematic aisles where each box is clearly labeled. If open data sites were Ikea, we have made sure to build the second floor. Our hope is to make the data shopping experience joyful, instead of maddening, and by doing so increase the ease with which data journalists, analysts, teachers, and students, use public data. Moreover,

we have made sure to make all visualizations embeddable, so people can use them to create their own stories, whether they run a personal blog or a major newspaper.

After all, the goal of open data should not be just to open files, but to stimulate our understanding of the systems that this data describes. To get there, however, we have to make sure we don't forget that design is also part of what's needed to tame the unwieldy bottoms of the deep web.

The views expressed are those of the author(s) and are not necessarily those of Scientific American.

ABOUT THE AUTHOR(S)