PublicBits

Improving the accuracy of analyses built on open data by helping data-driven analysts find, collect, and share the open data they use.

By Karissa McKelvey and U.S. Open Data

1. How has the organization defined the problem it is trying to solve?

Over the past two years, we have had conversations with data-driven researchers and open data experts at GovLab, Freeman Lab, FiveThirtyEight, California Civic Data Coalition, Indiana University, Trillian Project, The Wurm Lab, USGS, Ocean Health Index, International Monetary Fund, Data & Society, BetaNYC, Rackspace, Google, Digital Ocean, and Jupyter Project. Through these relationships, we have identified a significant pain point that spans the open data landscape centered around the discovery and distribution of data.

Organizations ranging from government institutions, cities, newspapers, scientific journals, non-profit organizations, and research labs have begun deploying websites called “data portals” (aka “data repositories”). In theory, data portals aggregate and share information that otherwise would be difficult to collect, enabling transparency and supporting data-driven decision making. As the demand for these systems increased, various actors stepped into the market to design, build, and deploy them. Unfortunately for data consumers, each open data portal provider has their own standards and pitfalls regarding discovery, distribution, provenance, and availability. This has fragmented the open data landscape, making it difficult for the typical user to navigate. The theoretical benefits of these portals prove to be time consuming and error-prone, at scale.

The first thing an open data researcher must do is go to a series of siloed data portals and, in each, enter their topic in the search bar, select the correct dataset, and download it to their local computer. A single data portal might point at relevant datasets on a great many websites. For example, a search on a data portal for “city water” could require that somebody visit and download data from dozens of websites. Once the data is downloaded and integrated into an analysis, policy paper, news article, or visualization, it is often necessary to check before publication whether the data has been updated, which requires repeating the search and download process again, usually manually. This

is a difficult process even for that tiny number of people who know how to create automated data-updating systems.

2. What is the overall challenge being addressed? What is the proposed approach? What evidence is there that this approach will work?

The aggregation of data into data portals has failed to improve the discovery process substantially. There are over a dozen software packages that are used widely to power data portals, and there is no common API or even a unifying interface practice among them. There have been efforts at creating a universal interface (e.g., the U.S. CIO's Data.json standard and Socrata's "Open Data Network"), but those have only succeeded in creating a series of competing standards.

This funding will support the development of an aggregation and search system called PublicBits, which will allow users to search across data portals. The search engine will crawl supported data portals and index their metadata in one place, enabling a powerful system for directing people original data sources. GitHub has proven that a centralized location for source code accelerates open source development, and we believe that open data will likewise benefit from centralization. Unlike source code, however, storing the world's data in one place would be prohibitively expensive, so we plan to host just the metadata.

We will accomplish this centralization by standardizing data portal distribution mechanisms with a suite of data translators. Each data translator is a bit of code that converts a data portal's inventory into a common format, facilitating indexing and comparison. The translators are designed to require *nothing* from vendors or portal operators—no work or even awareness is required of them. We have already written two prototype data portal translators for two widely-used data portal packages, Figshare [1] and CKAN [2], using Open Knowledge's Data Package standard [3]. This worked brilliantly, unlocking hundreds of thousands of datasets to be incorporated into PublicBits. We're confident that we can apply this technique to the other major data portals in the space.

To accompany this search engine, we have already began building a desktop application that uses Dat to manage data downloads [4]. (Dat is our data synchronization tool, which supports version control and forking to facilitate collaboration. It was created with a Knight Prototype Fund grant, and is supported by funding from the Alfred P. Sloan Foundation.) We imagine a user experience like Dropbox, which keeps track of data

sources automatically, notifying the user when their copy is out of date. After users download data via the desktop application, they will be able to offer their unused bandwidth to mirror their downloaded datasets. This system, based on BitTorrent, ensures that if the original source goes offline (e.g., the 16-day U.S. government shutdown of 2013 [5]), the data will still be downloadable through the PublicBits network. This technology exists within Dat, but is provided only as a JavaScript interface and limited command line tool. We propose to realize the promise of Dat by building upon its infrastructure, creating a tool that's usable by non-programmers.

3. How does this initiative fit into your organization's priorities for the upcoming year(s)? If this is a partnership please answer for all the partners. If this is a proposal from an individual please answer in the context of your current career path.

We are the team that has been developing and testing Dat, a project of U.S. Open Data, an organization focused on building the capacity of open data in the United States. Waldo Jaquith, advisor to Dat and the Director of U.S. Open Data, recently received an individual Shuttleworth fellowship supporting his work improving the production and use of open data by government.

Dat is currently funded by the Alfred P. Sloan Foundation to focus on reproducing scientific research. To meet this goal, Dat focuses on guaranteeing data replication across computers and tracking how the data has changed over time.

PublicBits will allow Dat's tools to be applied to U.S. Open Data's larger goals, beyond the sciences, and allow the Dat team to solve problems in the civic tech space, a long-time goal of theirs. The project is a natural combination of the skillsets of the U.S. Open Data / Dat team members, and deeply relevant to their interests.

4. What are the key project activities? (Please include an estimated timeline.)

Phase 0

Ramp up new team

1-2 months

- Hiring 2 full-time developers and 1 half-time designer, in addition to project lead Karissa McKelvey, for a total of 3.5 full-time people on the project. We have a large network reach in the open source/open data development world, and already have some promising candidates.

Phase 1

Fostering a PublicBits community during beta

6 months

- Beta launch of desktop application and PublicBits registry, all open source, in modular components, at http://github.com/publicbits/.
- Team will travel to New York City, Washington D.C., Portland, and San Francisco (data hotspots) to run all-day workshops with developers. We will perform user testing with them, and teach them to implement a small module to search and download from a not-yet-included data portal of their choice.

Phase 2

Prepare a 1.0 launch

6+ months

- After completion of the basic feature set, use community feedback to determine which features to experiment with.
- Consult with legal advisors about the legal pitfalls associated with rehosting data, creating appropriate terms of service, a privacy policy, and a DMCA takedown policy for any metadata on PublicBits.org.
- Explore what sustainability looks like. How will the project fund itself over time without relying on grant applications?

Activities	**Q1**	**Q2**	**Q3**	**Q4**	**Key actor(s)**
Build Team	**x**				U.S. Open Data
Design	**x**	**x**			Designer
User Research	**x**				Designer
User monitoring and reporting		**x**	**x**	**x**	Project lead
Beta Development	**x**	**x**			Development team
User Testing & developer workshops		**x**	**x**		Designer and development team
1.0 Development			**x**	**x**	Development team
Outreach	**x**	**x**	**x**	**x**	U.S. Open Data

5. What are the critical ideas/assumptions that will be tested and what indicators will be tracked? By whom? When?

- *That people want a centrally-searchable source of datasets.* Users will create accounts on PublicBits.org, which will allow us to understand our user base. We will ask them to enter their organization, email, and possibly complete a brief survey. We will also collect standard website traffic metrics to identify high-priority users, pain points, drop rates, and errors. Project lead Karissa McKelvey will study this data on a rolling basis, to ensure that we're creating something that people need and want.

- *That people want to know when datasets change.* It is difficult for us to test this assumption, due to the decentralized nature of PublicBits. The leader of our developer workshops will gather qualitative data about this, in the form of conversations with attendees, at each workshop. Separately, we will explore how to track use of this feature in the desktop application.

- *That it is plausible to create data translators for a sufficient number of data portal software packages to approach 100% coverage of datasets listed on public portals.* We will put together a brief survey of open data experts that collects information on the most popular data portals and couples that with existing compliations of data portals. We will create a public list of the most desired translators on GitHub and begin developing them during the beta phase. Because PublicBits is the first effort to aggregate a comprehensive list of data portals, there is no yardstick against which we can measure our progress. Instead, the developers will measure progress in terms of a) the number of supported software packages b) the number of supported individual portals relative to the number of known portals. This will be done constantly by McKelvey.

- *That people and organizations are willing to re-share datasets that they've already downloaded, for the benefit of strangers.* In collaboration with the Dat team, we will have connections to a variety of organizations such as universities and research labs that have extra bandwidth and a mission to share research data. We will track the health of the BitTorrent network of participating clients, measuring constantly how many people are sharing each dataset and how many people are downloading it. This will be automated. We have no baseline against which to measure, or concept of what ratio indicates "willingness." The real test

will be when a major data source disappears (e.g., another government shutdown), whether PublicBits proves to be a resilient source of that data.

- *That our dataset discovery interface is usable enough for even low-tech open data users.* During beta development, our designer/user researcher will conduct user testing. We believe this will help us promote the tool upon release and offer early indicators for a necessity to pivot on our designs.

6. What are the key challenges that could disrupt the project? What will be done to optimize for success?

There are thousands of data portals today that are powered by one of several dozen software packages. *The variety of data portals that need to be connected may be so vast as to present a significant obstacle to adoption.* We plan to combat this challenge by making it easy to connect new data portals, directly fostering an open source community around the application. It will be easy to build connectors that transform a data portal's functions to a generalized specification. As we gain popularity, data portals will have more incentive to write their own connectors to PublicBits.org or natively support the new specification because they will gain more users. We will optimize this possibility for success by focusing on making integration with the platform simple.

The desktop application must be cross-platform, and the search interface should be similar across Windows, Mac, Linux, and also the web. With a small team, this can be a difficult or expensive endeavor without reusing a significant amount of code. We have already begun to build a prototype proof-of-concept using Electron, which will allow us to build the frontend for all platforms in a single codebase using HTML, CSS, and JavaScript. The Dat team members have all built a variety of desktop applications using this approach, and we are confident in the stability of this toolchain. Will will hire a developer with experience in developing front-end applications using Node.js and/or Electron.

With a limited budget and only a year, we will have to make sure to find experienced developers and designer that understand the problem space well enough to be able to produce high-quality work in a short amount of time. Although we already have promising candidates, we will still interview candidates in the industry-standard fashion, relying on U.S. Open Data's guidance to ensure cohesion with our new team members. We also will focus on an open source, small-core, modular philosophy, enabling developers outside of the core team to contribute to parts of the project that can be

reused across the open source community. This has been a successful method for the robust development of a variety of projects we admire, including Dat, OpenDataCache [6], CityGram [7], Lightning-viz [8], and Flatsheet [9]. This ensures that even "failure" for PublicBits could still see the project succeed in unforeseen ways, by emitting components that can have impact on their own.

Endnotes

[1] McKelvey, Karissa. "CKAN search." *GitHub*. <https://github.com/karissa/ckan-search>
[2] McKelvey, Karissa. "Figshare.js." *GitHub*. <https://github.com/karissa/figshare.js>
[3] Open Knowledge Foundation. "Data Packages." *Data Protocols: Lightweight Standards and Patterns for Data*. <http://dataprotocols.org/data-packages/>
[4] PublicBits contributors. "PublicBits organization." *GitHub*. <https://github.com/publicbits>
[5] Desilver, Drew. "Federal government shutdown: The Data Casualties." *Pew Research Center.* Oct 2, 2013. <http://www.pewresearch.org/fact-tank/2013/10/02/federal-government-shutdown-the-data-casualties/>
[6] Krauss, John. "Open Data Cache." *GitHub*. <https://github.com/talos/opendatacache>
[7] Code For America and the City of Charlotte. "CityGram: Subscribe to your City." *Hacker League*. <*https://www.hackerleague.org/hackathons/accela-construct-app-challenge-2014/hacks/citygram-subscribe-to-your-city*>
[8] Freeman Lab. "Lightning." *Lightning Visualization Server*. <http://lightning-viz.org/>
[9] Vincent, Seth. "Flatsheet." *Knight Foundation.* <http://knightfoundation.org/grants/201450224/>