

To the Utah Transparency Advisory Board:

My name is Aaron Shafovaloff. I live in Draper, Utah. I am a software engineer. I would like to briefly describe some frustrations of mine in retrieving complete data sets from utah.gov/transparency. I would then like to offer some proposals for moving forward.

There are four ways I have unsuccessfully tried to get complete data sets from the site.

1. The first is to do so manually. On the site I can drill down to a level, entity, period, and type. From here I can filter it down even more and then click on a button to download a CSV file of records. This is a daunting process if one is attempting to download complete data sets. In fact, it makes comprehensive retrieval of the data essentially impossible. There are simply too many variations, too many entities, and too many filters that one has to go through. Most notably, and the available data through this method are truncated to a limited number of records, so one doesn't get everything.
2. The second way was to click on the link, "Download Full Transparency Lists By Year", and to select a level, a year, a type, and then a specific file for each entity. One is charged a dollar for each file download. This doesn't seem like much at first, but take the number of levels, years, types, and entities, and we are talking thousands of dollars. This is unfeasible and inhibiting. It seems unfitting to charge so much for public domain data in 2014.
3. My third attempt was to write what is called a scraper. Think of a scraper as a program that simulates the actions of a user. It can automate repetitive tasks. I wrote a scraper in an attempt to get complete data sets from the site. I was unsuccessful for a number of reasons. Also previously mentioned, the number of records in such open downloads is truncated, and there is no apparent way to paginate the data. What compounded the difficulty here was that the site does not expose a publicly consumable API. In fact, it seemed I had to use something called HTTP referrers to indicate which subset of data I was trying to get. From a technical perspective this in particular was concerning — it made retrieving the data more difficult than necessary, even for a scraper. I doubt this was intentional, but it was exasperating.

4. In my fourth attempt, I e-mailed transparency@utah.gov, where I was pointed to Michael Rice of Utah Interactive. I spoke briefly on the phone with Michael, suggesting ways I could retrieve the data. I was willing to drive over with an external hard drive. I was also willing to use a mechanism of transfer such as BitTorrent Sync. I later received an e-mail from Darrell Swensen from the state on August 21st, 2014:

"After meeting with Utah Interactive yesterday and discussing your request, the Division of Finance has determined that a full data dump of the data in the Transparency website cannot be provided because it contains private information which needs to be redacted. Funding is not currently available for the Division of Finance to pay for a feasible means to exclude that data in a full data dump. The process to exclude this private information is already in place in the website downloads and therefore this is the means at this time that needs to be used."

I have some five proposals for going forward:

1. Re-use the same queries used for the web site (which already redact private information) to make a complete data dump, and to make those exported files available on storage service such as Amazon S3.
2. If the cost of storage and transfer is a concern, there are solutions available that are not cost-prohibitive. In 2014 this simply should not be a problem. It is an era of content distribution networks, NoSQL databases, Amazon Web Services, archive.org, and BitTorrent.
3. Offer a consumable, paginated API that both developers and users of the web client can use. An API is essentially a service that makes the data useful and accessible to other computer programs. I encourage the state to insist that a format be chosen that is widely known and used among software engineers, such as JSON (see also the JSON API standard: <http://jsonapi.org/>). Using this format, the API could serve up not only records, but aggregate data associated with the records, such as totals that account for records not included because of privacy.

4. Have, as a part of this API, an endpoint for notifying the public about updates made to data. One open source project fitting for this would be <http://dat-data.com/>, which is sponsored by the U.S. Open Data Institute and the Alfred P. Sloan Foundation. It provides a way for computer programs to subscribe to updates made to the data.

5. Insist that software contractors of the state make maximal use of open source software, and to release the code of state-specific software projects in a standard open source venue such as GitHub.

In general, I am asking that complete transparency data sets be easily and inexpensively available to citizens who are web developers and software engineers. Then those interested can build a better experience for search and visualization, as well as do meaningful analysis.

With respect and appreciation, thank you.

Aaron Shafovaloff

Draper, UT

(801) 502-9269

aaronshaf@gmail.com