HB 598: The Open Government Data Initiative

Let a thousand real time dashboards bloom!

The Open Government Data Initiative advocates for laws that force government to proactively publish data in machine-readable formats.

It is non-partisan ; we support exposing *all* data, whether that might support the policies of the left, the right, both, or neither.

It does not advocate for government-provided *tools* to view or read data; it advocates only for the *publishing* of data.

(The OGDI believes in The Unix Philosophy: small tools, loosely coupled. Make the government provide the data and the free market will do the rest!)

Finally, OGDI is based in New Hampshire and advocates first and foremost for state-level laws, but we support activists in other states and look forward to sharing model legislation and discussing best practices.

The problem

> “Right to know” is
>
> the right for people to “participate in an informed way in decisions that affect them, while also holding governments and others accountable”. It pursues universal access to information as essential foundation of inclusive knowledge societies.

At the federal level we have FOIA (the Freedom of Information Act).

In New Hampshire we have RSA 91-a.

These are better than nothing, but inadequate.

Both of these require:

1. that a citizen know what information they want
2. that a citizen send in a request to the government
3. that the citizen wait for the government (often there's a 'deadline', but (a) there are no penalties when the government breaks the deadline, (b) the deadline can be satisfied by the government merely replying by the specified date "this may take up to three years").
4. that the citizen pay for reproduction costs

Further, the data

1. does not come back in a machine-readable format
2. is published, at most, once, and not repeatedly (e.g. annually, quarterly, weekly…)

Contrast this to the private sector where every 1-person e-commerce business, or 4-hour-per-week software side project has a dashboard.

A typical 1-person project has a dashboard that displays in real-time (update cycle of < 1 second) and shows things like

- number of site visitors
- percent of site visitors that spend more than 1 second on site
- percent of site visitors that put an item in a shopping cart
- percent of site visitors that click the checkout button
- number of orders received
- total value of orders received
- inventory levels
- packages shipped per day
- etc.

This has been state-of-the-art in the private sector for almost 30 years.

Now compare this to even a small responsive state like New Hampshire, where almost no data is proactively published, and what is published is

- not machine readable
- often out of date

For an example of both problems, look at the NH Department of Education data for Weare, NH's high school. The most recent data is seven years old, and is presented in crappy HTML format.

What sort of data do we want to see?

All of it!

Let's reframe that. What data do *you* want to see?

Pick the issue that you care about:

Children

- How many families did DCYF investigate last year?
- How many children were taken into state custody?
- How many of those children were placed in foster homes within one week?
- …within two weeks?
- …within a month?
- How many of the children placed in foster homes were seen by therapists within a week?
- …a month?
- …ever?
- How many of those children placed in foster homes had the appropriate paperwork from DCYF to attend local schools?
- …and how many didn't, because DCYF couldn't get its act together in time, and children instead stayed home?
- How many children in DCYF custody were shipped to out-of-state facilities?
- How many children in DCYF custody were sexually abused?
- How many children in DCYF custody died?

and further: how do the figures for 2023 compare to 2022, 2021, 2020, etc? Is NH DCYF getting better at getting children needed care? …or worse?

and further: how does NH DCYF's numbers compare to, say, Vermont's ? or Maine's? Are NH children who have survived trauma seeing therapists in half the

time that Vermont children do? Or twice the time? Are we paying twice as much per child processed?

Police and Crime

- How many drunk driving arrests were there state-wide this year? Last year? 10 years ago?
- What percent of those arrests led to conviction?
- What percent of those convicted went to jail?
- What percent of those arrested ended up not charged at all?
- Are we sending more or fewer drunk drivers to jail now vs then?
- Are conviction rates heading up or down?
- …out of how many drivers, and driver-miles (we want to be able to scale statistics appropriately, not have them be swamped by secular trends / exogenous events)?
- How many robberies were reported in Hillsborough county last year?
- …how does that compare, per capita, to the same county five years ago? To Strafford county this year?
- How do robberies correlate with population density? With demographics like age of the perpetrator?

Colleges

For any given college in the state:

- How many people applied to the college? What were their average SAT scores?
- How many people were accepted? What were their average SAT scores? What were the scores broken down by race (suggestive of institutions abusing their non-profit status to illegally discriminate against citizens)?
- What percent of students successfully completed each degree program?
- What percent of students successfully graduated?
- What was the average debt load of drop-outs?
- What was the average debt load of graduates?
- How long does it take a graduate of, say, a chemical engineering program to pay back their average student loan at by spending 25% of the average salary of a graduate of such a program?

- How long does it take a graduate of, say, a gender studies program to pay back their average student loan at by spending 25% of the average salary of a graduate of such a program?

Taxes and lobbying expenses

For any given town in the state:

- how much was collected in tax dollars from citizens
- how much of this was spent on lobbyist organizations (e.g. towns sending taxpayer dollars to The New Hampshire Municipal Association which actively lobbies on many topics against the interests of citizens, or the New Hampshire School Boards Association or other lobbyists ?

Judges

- how many judges does NH have?
- how many cases did each of them judge last year? The year before?
- how many cases were overturned by the NH Supreme Court?
- what judge had the highest rate of cases being overturned? Who had the lowest?
- what areas did specific judges have more or less cases overturned? Is there some judge who has 2% of their zoning law cases overturned, but 50% of their criminal law cases overturned?

etc.

This list is just a start.

While the above list strives to be non-partisan, it's inevitable that certain prejudices have crept in … but others can expand the list with whatever issues they care about. Number of wetland development projects approved. Number of garbage dumps created / closed. Conviction rates for crimes against women. Spending on homeless veterans. Whatever it is that citizens care about, governments and non-profits operating under government regulation almost certainly already collect the data, and just aren't sharing it effectively.

What does getting this data achieve?

If government exists it is in our interest that it do the most good (or least harm) possible, for the lowest cost.

There is an efficient frontier in all things. We can have a government that spends $0 and prevents zero cases of a given bad thing. Or we can have a government that spends $1 million and prevents 1,000 cases of that bad thing. Both are defensible policies.

…but what's indefensible is a government that spends $1 million and prevents 5, or 4, or 0 cases.

Activists of all stripes believe that government is inefficient, wasteful, and focused on the wrong problems.

The problem is that, without access to data, their appeals are emotional and anecdotal, instead of evidence based.

…and further,

> "When you can measure what you are speaking about, and express it in numbers, you know something about it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts advanced to the stage of science."
>
> — Lord Kelvin

The Solution: Open Data

Before we can craft legislation that forces government to share data, we have to consider exactly what it is we're looking for (Remember the old adage: "be careful what you wish for").

Small tools, loosely coupled

Different citizens and groups will have different interests.

- one might care that per pupil spending is going up, and might pull from various data sources to assemble data to show expenditures per pupil, statewide, over time.
- another group might argue that high density urban areas are getting shortchanged in policing, and might pull for a different set of sources, and generate graphs to show policing results as compared with population density.
- etc.

Each citizen group should be encouraged to build tools that do the analysis and reporting that they care about.

Again, the OGDI believes in The Unix Philosophy: small tools, loosely coupled. Make the government provide the data and the free market will do the rest!)

To support this low-effort modular approach to data import and digestion, we need to think about …

Data formats

Data formats vary from human readable (a JPG of a graph, or a PDF file of a budget) to the machine readable.

Machine readable data formats vary from the opaque (a binary file, or a spreadsheet with zero column headings) to somewhat transparent (a spreadsheet in proprietary Microsoft Excel format), to the very transparent (a CSV file with column headers, or an XML file with a DTD (Document Type Definition).

Data varies from the easily comprehensible overviews(the list of NH town tax rates), to the complete (the full 1,284 pages of the New Hampshire 2023 budget).

There is a tension between the desire to get every scrap of data that's available, on the one hand, and drowning users in minutia that may make it effectively impossible to make use of the data (can you look at the almost 1,300 page

budget and - even using search in the PDF, find how much money goes to road building and maintenance?).

Additionally, there is the issue of …

Data format consistency over time

I recently did a small project importing NHLA spreadsheet data going back ~10 years into a Rails app, and because I could export each sheet as CSV, it was trivial to import … but even then, column headers changed from year to year, and I ended up doing a lot of special casing to, e.g., pull out canonical names which appear in 3 or 4 different formats over various years.

That's from great, clean, simple data only going back a few years.

You can imagine the challenges in trying to parse out the DCYF component of the state budget over 20 years, especially as divisions get renamed, programs come and go, etc.

So we'd like data that is easy to compare across time (especially given that one bureaucratic trick is to change measurements, or change the definition of terms over time, to intentionally thwart measurability). This pushes us in the direction of wanting to carefully specify, in legislation, exactly what must be provided.

…however, this is all in tension with the ideal of …

Re-use of existing data whenever possible

The state collects absolute mountains of data. Every town has a budget generated in a spreadsheet. Every school system, likewise. DCYF generates reams of internal reports. Every police department shares data with the FBI.

The key is to, when possible, discover existing data sources that the government already creates internally, and then mandate sharing.

So … would we rather a priori, when crafting legislation, specify the 12 fixed things we want, each year, in a time sequence … or would we rather get massive data dumps?

I think the answer is “both”.

This Bill

This bill is intended to create a bi-partisan study group that will identify pre-existing government data sources that can be exposed to the public at minimal or zero cost.

This study group will prioritize, at all times, the protection of PII (“personally identifiable information”), and include the priority of protecting PII in all recommendations it will make.