Providing Data For Data Journalism

Requirements and Plans for Health Reporter, a Journalism-Focused Health Data Repository

Eric Busboom

20 July 2016

With the continued enthusiasm for telling stories with data, data providers are wondering how to get more of their data into news stories. Many health agencies, state departments of education and city open data projects have extraordinary data collections that are relatively unknown, and the data providers would like to see the data they have labored to release get used in data-driven stories.

So, what does it take to produce datasets that journalists can use? We're tackling this problem as part of the design phase for a project, funded by the California Health Care Foundation, to apply the lessons of Census Reporter to California health data. Our goal is to make health data as easy to use as Census Reporter does for demographic data.

How Data Is Used in Stories

First, it's valuable to examine how data is used in news stories. While there are a lot of ways to dissect data journalism, we've identified four major roles for data in new stories:

- **Informing a story.** Data is used to validate other information in a story, such as using demographic data to verify a hunch about why a political candidate is popular in one community but not another. The data usually doesn't actually appear in the story and may not be directly related to the topic, but rather gives the journalist confidence in the truth of other facts that are in the story.
- **Supporting a story.** Data is strongly related to the topic, and usually does appear in the story, such as the change in the rate of vaccinations in a story about a whooping cough outbreak.
- **Core of the story.** Data forms the backbone of a story, with other facts and interviews supporting the data driven narrative. This form is common in stories that start with a tip about financial corruption and are developed through analysis of financial records.
- **Source of the story.** Sometimes journalists start with a dataset and mine it for a story, such as a newly opened dataset of development permits that results in a series of stories. In these cases, the data is absolutely central.

In the first three cases, Informing, Supporting and Core, the journalist starts with the story, and then brings in the data, while in the Source case, the journalist starts with the data and the story follows from analysis.

If the story comes first, the journalist likely needs specific data, such as restaurant inspections in a particular community between two given years. The major complication is that the data may

This research was funded by the California Health Care Foundation.

not exist, or cannot be acquired, or doesn't directly address the topic. Occasionally, a different dataset can be substituted, and for the Informing and Supporting cases, the these substitutions may not affect the integrity of the story.

In Informing and Supporting cases, the journalist usually only needs a fact or trendline, but it must be for a specific time, place or subject. This is precisely the use case where Census Reporter is best suited.

The data process in the Informing and Supporting cases is very much like data-driven research for policy, marketing or epidemiology. In it's most complex form, the process starts by formulating a question, from which a set of potential datasets are identified. The datasets are evaluated for suitability, and one is selected that can best answer the question. Because this case most frequently uses datasets produced by agencies for analysis and research, the datasets are well-formed and relatively easy to analyze, so most of the effort is in identifying the correct datasets to use.

In its simplest and much more common form, the process involves the journalist visiting American FactFinder, getting murderously frustrated, remembering Census Reporter, and cutting and pasting a number from from there.

For the Core case, the journalist needs more than a fact, but rather an entire dataset, and the dataset must address the story topic directly. And for the Source case, the journalist not only needs the whole datasets, it usually much be a fairly large, detailed one for the exploratory analysis to yield any stories.

For the Source case, the right dataset already exists — it's the one the journalist already has — and the journalist needs to find a story in the dataset. This is an exploratory data analysis problem, and the skills required for exploratory data analysis can range from basic spreadsheet manipulation to advanced statistics or data science. Basic datasets, such as lists of expenditures or salaries can be easy for a journalist to analyze, but increasingly, journalists are getting very large datasets that require complex analysis. One example is a story by NBC 7 San Diego that condensed and linked over 10M GPS records from street sweepers to parking tickets, requiring both programming and data analysis.

What do Journalists Want?

While the question of 'What Do Journalists Want" can never be answered briefly, we can relate some of the lessons we've learned from interviewing journalists and working with them on data-driven news stories.

Data providers are most interested in seeing data used in enterprise stories, where journalists have a strong preference for distinctive stories, using novel datasets or a novel viewpoint. They don't want to re-write old stories, and the don't want to use an angle that others have used. They prefer stories that are relevant their readers, have a local connection, have substantial impact, or relate to topics of current interest.

These preferences lead journalists away from data analysis that has already been done or datasets that are commonly explored. For significant stories, they'd rather work with dataset sets that are exclusive, which is very contradictory with the data provider's goal of encouraging wide use. One way to generate exclusivity from an non exclusive dataset is to perform an expensive or difficult analysis, or a complicated visualization, that other journalists won't be able to do. This is a common way for large news organizations to generate stories from medicare or census data, but requires resources that are rare outside of the New York Times or Washington Post.

Beyond the largest newsrooms and the mid-size organizations that have made a decisive investment in data journalism, most journalists have a moderate level of data skill, most frequently limited to basic spreadsheet operations and occasionally Access database queries. Large newsrooms may have journalists with programming or data analysis skills, and sometimes dedicated technical staff, but most journalists are only able to work with small, well formed Excel spreadsheets.

The results of these preferences, paradoxically, is that for the kinds of stories that data providers most want to have their data used in, the journalists would prefer to use data sources that they are least likely to have the skill to use.

Implications

The primary implications from the intersection of the four roles of data, journalists preferences for stories, and their average level of data skill are:

- When data is Supporting or Informing a story, the journalist will often use data for research and doesn't care much about novelty. This form of data is frequently single facts or metrics, such as are easily found on sites like Census Reporter.
- When data is the Core or Source of the story, journalists prefer datasets that are exclusive, either by being difficult to acquire or difficult to analyze, but they rarely have the skill or time to do such analysis.

Together, the conclusion is that there are two different ways to support data-driven journalism:

- **Quick Easy Facts**. To support Informing and Supporting stories, more data sets should be digested into easy to access facts about regions and demographic groups. Ideally, there should be a way to load any dataset into software like Census Reporter, so journalists can use a high-quality interface to get a broader range of information.
- **Frictionless Data Experts**. For Core and Source stories, journalists should be able to get data help on demand at a cost within discretionary budgets. It should be as easy for a reporter at the Idaho Register to get expert data help on a 1 hour turnaround as it is for a reporter at the Chicago Tribune.

Quick, Easy Facts For Informing and Supporting Stories

The first way to support data-driven journalism is the subject of our Health Reporter project, which has the goal of providing California health data in the same form as Census Reporter provides Census data. Conceptually, the project isn't hard, nor is it novel — we're building what has long been known as an Indicator website. But there are two aspects that are hard:

- Designing a user interface for quickly exploring the datasets and finding sets that are relevant to a story.
- Ensuring we can manage the system with a very large collection of data at a reasonable, sustainable cost.

To be useful, and to be used, the indicator website must have a lot of datasets. If users' aren't fairly certain that they will find what they want, they will stop coming. Either the site must answer questions about a specific topic, or it must have a very large number of datasets. Census reporter follows the first case, but Health Reporter must follow the second. Practically, we should expect to include 500 to 1000 indicators in the site.

Having a large number of datasets produces two problems: it is expensive to manage that much data, and as the number of datasets increases, it is harder to find the one you want. This tension has been a primary problem for most data access sites; with few sets, the system doesn't have what they want, and they stop coming. With too many, the system is expensive and while the collection has what they users' want, they can't find it in the pile.

For Health Reporter, making the system sustainable requires us to solve two problems: reducing the cost of managing data and developing a new interface for visually exploring datasets.

Reducing Management Cost

Creating and operating an indicator website is expensive. Census Reporter, versions 1 and 2, cost about $500,000, and other indicator websites have costs ranging from $250K to $800K. Building one indicator site is primarily a matter of money and time and skilled programmers. Building dozens of them, for different types of data in different regions, is a complicated cost reduction challenge. As a result, the most important innovations of our project involves how to make it much less expensive to load data into an indicator website.

The most significant task for maintaining Health Reporter is regularly importing and updating datasets. We'll need to have a lot of data, and it will change on a regular basis. So, there must be a good process to load data inexpensively.

To manage the data process, we have developed an open source data management system, Ambry. Using this software, we can package data in a format that allows it to be directly importable into Health Reporter. This method allows for cost reductions, because the process is well defined and has good software support, so the labor is much less expensive that would be required without Ambry.

However, in the long run, it would be much less expensive if data producers improved their metadata, so dataset could be programmatically imported, with less human intervention. This effort is the subject of another project, which is currently has an early draft specification and interest from a range of state, federal and private organizations.

Fast Visual Exploration and Searching

When journalists are looking for datasets from which to extract facts for stories, they prefer data that is:

- **Local**. Journalists prefer local data to state or county data, with neighborhood, ZIP code or Census Tract level aggregations being most valuable.
- **Has Many Dimensions**. Stories often have a demographic angle that requires data with demographic factors such as age, race or income.
- **Accessible**. It must be easy to get to the data required, or to determine that the data is not available.The interface should be dedicated to fast facts, not creating custom visualizations.

For Informing and Supporting stories, journalists approach searches for data in two ways. If the reporter has a well-developed need — a “hole” in the story to be filled with data or a data-driven fact — it is usually possible to use a text query to start the search. It’s easy to type “disability rates in San Diego vs Los Angeles” into a search engine and eliminate a lot of browsing and clicking. However, if the data need is not well developed, the journalists may prefer to browse for the data.

The requirements for a good data search system have already been analyzed in a previous project, so this project will focus on developing a good browsing system.

The interface for the data repository should emphasize basic visualizations and simple tables, but it does not need to have a chart builder or table builder for custom datasets; either the journalist uses the provided charts and tables, or can get a reference to the original dataset to do custom analysis, but there no special features for modifying the charts and tables.

Because locality is important, the interface should have strong support for identifying what data is available at the local level, or for a particular locality, what is the most fine-grained data available.

Because demographics are important, particularly race, ethnicity and poverty category, the interface should have search features for those demographic categories, when they are available in the data.

Low-Fi Visualizations for Fast Pattern Matching

To address these issues, we are developing a visual model for Health Reporter that presents high-dimensionality data with a combination of dimension reduction and interactive small

multiples, which can take advantage of pre-attentive processing to aid journalists in finding the right datasets.

The “Dimension Reduction” feature reduces the number of dimensions in a dataset through aggregation, sometimes altering the data in ways that preserve relationships but may reduce accuracy. Small Multiples are charts composed of many smaller charts. Small multiple charts can be interactive, to provide additional links between the individual charts. Together, these features allow for visualizing a whole dataset on one page, to quickly determine if it is appropriate to use for a story.

For more specific requirements and designs, refer to the Health Reporter design document.

Frictionless Data Experts For Core and Source Stories

A second problem, one which we are not solving (yet) is how journalists can get access to data experts for more complicated projects, especially for Core and Source stories. This issue deserves to be the subject of an entirely separate project, but there are a few use cases and requirement that we know now.

With Core and Source stories, the preferences of data journalists and the structure of their stories have some important implications for data providers who want to increase the use of data. The main implication is that Census Reporter is not the best model for these types of stories.

They’ll always want what you don’t have. Almost by definition, a highly available, accessible data library will be most interesting to journalists for supporting facts, but not as the core of enterprise reporting. To be valuable in enterprise stories, to satisfy the desire for exclusivity, either the dataset is novel or the analysis of the data is novel.

They won’t want what you do have. The types of data that are common on indicator sites like Census Reporter or DataUsa.io tend to be the datasets that are most thoroughly analyszed, and have the least potential for yielding interesting stories, so these data are most often used for supporting facts, not the core of a story.

Journalists can work magic with Excel. While journalists lack many of the important data manipulation skills, they can be very productive in analyzing spreadsheets, using pivot tables or making Access queries.

Good bespoke analysis is worth many stories. With a deep dataset subject to skilled exploratory analysis, a journalist can extract a series of stories. A few thousand dollars of data analysis on a well chosen dataset can yield a series of five or six stories.

These implications combine to a single central insight:

> *For Core and Source stories, the data the journalists can best use is in the form of a well structured spreadsheet that is derived from a complex, interesting dataset that few other journalists have access to.*

For instance, the San Diego Regional Data Library did the primary data work for a story about building permits. The City of San Diego released the data as a collection of XML files that provided an index to a web API, returning records in JSON. The Library's programmers scraped the API to produce specific spreadsheet files, which Library analysts explored for patterns. The output to the journalists were spreadsheets and plots that the journalist could sort and search, and easily use to create in-house visualizations.

This insight exposes a fundamental tension for data providers: *the data provider wants to have data be very available and be widely used, but the journalist wants to use data that is rare, restricted or novel.*

Resolving the Contradiction

To resolve the contradiction, we'll need a different model of how journalists interact with data providers, with the interface of that interaction being a well-formed, bespoke spreadsheet. There are many other possible interaction models that don't have a spreadsheet interface, but this specific problem requires cost reduction, for which specialization, standardization and defined process are time-proven techniques. Additionally, to fit in with the fast pace of a newsroom, the provision of that spreadsheet must be relatively fast and frictionless; for instance, the journalist cannot be expected to request budget to hire a data analyst.

The use case is fairly simple: A journalist, working on an enterprise story, requests help with a data project, and quickly, has a data expert working on a dataset. The expert returns a spreadsheet to the journalist, tailored to the story, that the journalist can manipulate (sort, search, extract or simple math) or create charts, graphs and tables from. Somehow, the journalist does not have to ask for budget approval, deal with hiring, or pay an invoice.

Obviously, the hardest part of this use case follows the "somehow." We'll need to solve a money, time and logistics problem, not a technical one.

Most likely the solution will be the result of special processes, specialization and standardization to drive costs down, an on-demand contracting service like UpWork, and a creative financing plan that combines subsidies, grants and small amounts of money from a large number of stakeholders.

While on-demand data experts are very helpful, they won't absolve journalists of acquiring some data skill. Data process that involve contractors work best when the journalist gets a data product that the journalist can manipulate to make it apply exactly to the needs of a story. Practically, this means that the data expert role is to turn a complicated data analysis or programming problem into a manageable spreadsheet, and the journalist can slice and dice the

spreadsheet. Journalists will still benefit from data skill training, but they should not need to learn the esoteric skills that a data expert has.

Thanks

Thanks to the journalists on whose insights this report is based, and who provided feedback during the development of the document.

- Emily Bazar, Senior Correspondent, Kaiser Health News
- Wendy Fry, Reporter, Investigative Journalist, San Diego NBC 7
- Lisa Halverstadt, Voice of San Diego
- John Howard, Editor, Capitol Weekly
- Rebecca Plevin, Health Reporter, KPCC Public Radio
- Paul Sisson, Health Care Reporter, San Diego Union Tribune
- David Wagner, Science & Technology Reporter, KPBS
- Joe Yerardi, Data Reporter, inewsource

Thanks are also due to the California Health Care Foundation for funding this project.