Baltimore Builds a Data Dictionary

Adam Stone | November 25, 2013

In Baltimore, CTO Chris Tonjes wants all the data in one place. All of it. He's looking at data on water usage, transportation, disease control, emergency services, snow removal.

"We are really looking to collect everything," Tonjes said. He isn't just hoarding. He's taking all that information, massaging it into a universally accessible format and then making it available to planners in 18 city agencies.

What Baltimore is doing citywide, Virginia is attempting throughout its educational system. By creating their "data dictionaries," each of these IT teams is attempting to turn disparate data into something that can be readily shared among a range of key stakeholders.

Information Sharing

Once completed, Baltimore's new consolidated database will give users unprecedented access to municipal information. Tonjes described a scenario: Suppose a police officer notices a surge in parking tickets on a particular block. By cross-referencing transportation data, that cop might save a lot of patrol hours by pairing recent tickets with reports of missing signs on the street.

In a more severe case, public health personnel might track a disease outbreak against water-related complaints that might indicate polluted water.

IT will make this possible. "It fulfills a pressing need for us because we have 45 legacy applications running on a very old mainframe," Tonjes said. "By pulling the data out from the mainframe and putting it into a relational database, we can do a lot of interesting things."

He expects to have a rudimentary warehouse established this fiscal year, with full-scale prototyping of analytic tools to follow soon thereafter. The department has no special budget for this: It's all being done with in-house resources.

Interns Assist

Those resources are being supplemented by computer science interns from Johns Hopkins University. The labor isn't free — it costs about $5,000 per six months' work for each of three interns — but Tonjes said it's worth it.

The interns are doing much of the heavy lifting. Since legacy systems don't sync today, it isn't possible to automatically tabulate the existing data inventory. Instead, the interns are gathering information from IT managers throughout the city, one-on-one.

"They are doing all the interviewing, creating the giant spreadsheet, talking to people and tabulating stuff. They are the ones who are doing the digging and the detective work, which is tremendously helpful," Tonjes said.

“It is no small feat, and we know that,” said Heather Hudson, Baltimore’s chief data officer. “It is going to take a lot of manual work, looking at the data and finding those relationships. Then once we see what we have, the next piece will be data governance rules and standardization.”

At Johns Hopkins, administrators said the project helps to further integrate the school into the life of its city. “It’s very positive for the university to be supporting the city with this expertise. It’s important that the university has this deep tie to the community and to city government,” said Randal Burns, an associate professor in the Department of Computer Science at the university’s Whiting School of Engineering.

The effort likewise helps to develop a future workforce. “Many of our courses are specifically designed to help people work in the data and cloud environment,” said Burns. “It’s great that Baltimore has made a place where they can use those skills.”

Broadening Agenda

This is not the city’s first effort to convert its manual spreadsheets into something more broadly usable. Rather, the data dictionary follows on the heels of an earlier program known as CitiStat.

Launched in 2009, that effort aims to make city government more responsible, accountable and cost-effective by pooling data from multiple agencies. The dictionary takes what began as a performance-management initiative and extends the concept to embrace a broad array of metrics throughout city agencies.

That wider agenda comes with certain technical hurdles. For one thing, the existing flat data must be converted into a relational database, something that will have to be done manually at first, though automated routines could eventually take over the task, said Hudson.

At the same time, data integrity can be sketchy in existing platforms. The year 2007, for example, can easily show up as 2077 thanks to human error. “So another benefit of doing this will be to improve integrity across the board, even in systems that don’t automatically have that integrity,” Hudson said. “Our data is going to get cleaned up as a natural side effect. It is technically feasible; it’s just going to require a large effort.”

An Edge for Education

Baltimore’s IT planners are not the only ones making that effort. At the Virginia Department of Education, Director of the Office of Educational Information Management Bethann Canada is looking to collate 777 data elements on the state’s student body, starting with the obvious elements such as gender, race, test scores, postsecondary enrollment, and drilling down from there.

The Virginia Longitudinal Data System, which went live in August 2013, is funded by a $17.5 million grant from the American Recovery and Reinvestment Act in 2010. Among other things, that money provided $2 million to help school districts improve the quality of their data, and also funded development of electronic transcripts in three universities and all 23 of the state’s community colleges.

Besides the Department of Education, players in the effort include the State Council of Higher Education for Virginia, Virginia Employment Commission and Virginia Community College System.

Planners seek to create fact-based information for researchers as well as for policymakers looking to develop a future workforce. By developing an easily accessible pool of hard data, “we can answer the questions policymakers are asking using real data,” said Will Goldschmidt, a project manager in the Department of Education and a lead developer of the data dictionary.

As in Baltimore, Virginia’s efforts have been hampered by the disparate nature of existing systems and the lack of a common format. Making things more complex, agencies want to retain those idiosyncrasies.

“One of our challenges has been that each agency has wanted to maintain its autonomy in terms of how we store and characterize data,” said Canada. As a compromise, each agency still creates data in its own format; it goes into the lexicon that way, and then the common elements are culled and linked, while leaving the basic structure intact.

This system helps overcome another concern: that student data should be fundamentally anonymous for the purposes of research. In making data linkable, IT managers simultaneously “de-identify” it, stripping away personal data. “You’re not going to have your name or Social Security number in there. But it’s not totally anonymous; it will still have things like your gender and race,” Canada said.

It took some skillful diplomacy to make this happen. “What you see with this federated model came out of 18 months of talking about the rules of engagement, creating agreed-upon business processes among various agencies,” said Goldschmidt.

In fact, the IT team went so far as to bring in a third-party facilitator to ensure a sense of fairness among all the players.

These efforts already are bearing fruit: The state’s education leaders have adjusted certain standards in science and history, based on the newly available data.

“There already have been policy actions taken based on this merged information,” Canada said.

http://www.govtech.com/data/Baltimore-Builds-a-Data-Dictionary.html