On Open Government, Education Data, and PDFs

Late on Christmas Eve, the City of Newark released dozens of emails (as PDFs) relating to Facebook CEO Mark Zuckerberg's $100 million donation to its schools back in the fall of 2010. It did so to comply with a Superior Court order, following a lawsuit by the ACLU on behalf of local parents.

The Star Ledger has posted the PDFs to its website — PDFs which contain some interesting conversations between various Facebook, Gates Foundation, and city execs as they hash out the details of how Zuckerberg's millions should be spent.

As part of the recent Race to the Top-District competition, 16 school districts will split some $400 million in funding in order to "personalize and deepen student learning, directly improve student achievement and educator effectiveness, close achievement gaps, and prepare every student to succeed in college and their careers." The applications, along with the scores and comments they received, are posted to the Department of Education's website.

I am deeply curious about the contents of the Newark emails and the relationships and politicking they reveal. Why was Facebook COO Sheryl Sandberg so deeply involved in the negotiations with the city about how Mark Zuckberg's private donation would be spent? What policies Bill Gates expect the city's schools to enact in order for him to make a matching donation?

I am deeply curious about the contents of the RTTT applications. What technologies are being built and utilized by these districts to (ostensibly) "personalize" students' education? What ed-tech companies will benefit?

I will have to read my way through the PDFs in order to tell you — not a particularly fun task (neither grant-writing nor email writing are genres I really enjoy). I will have to read my way through in lieu of asking a computer to do so — to identify common words, to search for key terms or figures, to compare phrases and funding. I will have to do so because there's no easy way for a machine to parse the PDFs.

As I've written before, I hate PDFs.

The Portable Document Format is almost 20 years old now. Once Adobe's proprietary format, it's become a de facto standard for distributing electronic documents, in part because it nicely "fixes" things onto the digital page as they were meant to be seen on the printed one. In other words, the PDF is a file that contains all the text, fonts and graphics necessary to render digitally what appears on a piece of printed paper.

My problems with the PDF are severalfold, and I'd contend is a wildly unsatisfactory way to release government data (and not just because as the Star Ledger [later discovered](#) that passages that had been redacted by the city could in fact be read by simply opening the PDFs in Adobe Acrobat Pro and deleting the redaction boxes) — that is, if you really want people to read and use it.

These problems with PDFs include the file size (often larger than HTML or plain text) as well as the high cost of buying the Adobe software that was once necessary to create one. (The Reader has always been free.) The PDF is also incredibly frustrating as it's a digital document version of Hotel California: once you check in, you can never check out. While you can save a file as a PDF to make it supposedly "portable," it's very difficult to get data or a document back out of that format.

PDFs. It's where data goes to die.

-----

2. Executive Committee Roles and Responsibilities
    i.   Grant oversight, support, and compliance;
    ii.   Establish and execute a strong Consortium conflict of interest policy;
    iii. Establish appropriate competitive Request for Proposal (RFP) criteria and processes for procurement and access of funds that ensure rigorous proposal review and sound investment decisions;
    iv.   Establish rigorous leverage and sustainability requirements for all Consortium investments;
    v.   Approval of disbursements of investment funds and any procurements made via competitive process;
    vi.   In cases where districts would be required to develop proposals to access Consortium investment funds the Executive Committee will require evidence of Education Association involvement in such proposal development, and no proposal will be considered by the Executive Committee without the signatures of the local superintendent and local Education Association president;

(What looks like a PDF of a photocopy of a photocopy, the Puget Sound ESD's PDF is barely human-readable, let alone parsable by a machine.)

I suppose you can argue the release of these documents is good news for open government (although let's not forget it did take a court order to get the Zuck-related emails). The data dump from Newark and from the Department of Education can offer us a bit of transparency, a glimpse into what negotiations and please are made to get money, to give money, to shape and respond to policy.

But in both cases, that transparency is clouded by the fact that the documents have been released as PDFs. And while government can shrug and say that it's done its part by making this information available, it needs to do better and release information in open, machine-readable, parsable formats. If not, why, it's almost as if folks in the government don't really want the public to read and analyze this information...

Written by

Audrey Watters

Published 03 Jan 2013