THE WALL STREET JOURNAL.
WSJ.com

February 18, 2015, 3:47 PM ET

It's Way Too Late Not to Know Where Your Data Is

By Thomas H. Davenport

One of the paradoxes of IT planning and architecture is that they have made it more difficult for people to find the data they need to do their work. For generations now, companies have created "data models," "master data models," and "data architectures" that lay out the types, locations, and relationships for data that they will have in the future. Of course, those models rarely get implemented exactly as planned. So organizations have no guide to what data they actually have in the present and how to find it.

One other reason why companies don't create simple catalogs of their data is that the result is often somewhat embarrassing and irrational. Data are often duplicated many times across the organization. Different data are referred to by the same term, and the same data by different terms. A lot of data that the organization no longer needs is still hanging around, and data that the organization could really benefit from is nowhere to be found. It's not easy to face up to all the informational chaos that a cataloging effort can reveal.

Perhaps needless to say, however, cataloging data is worth the trouble and shock at the outcome. A data catalog that lists what data the organization has, what it's called, where it's stored, who's responsible for it, and other key metadata can easily be the single most valuable information offering that an IT group can create.

Given that IT organizations have been preoccupied with modeling the future over describing the present, enterprise vendors haven't really addressed the catalog tool space to a significant degree. There are several catalog tools for individuals and small businesses, and several vendors of ETL (extract, transform and load) tools have some cataloging capabilities relative to their own tools. Some also tie a catalog to a data governance process, although "governance" is right up there with "bureaucracy" as a term that makes many people wince.

At least a few data providers and vendors are actively pursuing catalog work, however. Enigma has created one for public data, for example. The company has compiled a set of public databases, and you can simply browse through their catalog (for free if you are an individual) and check out what data you can access and analyze. That's a great model for what private enterprises should be developing, and I know of some companies (including Tamr, to which I am an advisor) that are developing tools to help companies develop catalogs.

In some of those private enterprises, interesting work along these lines is already taking place despite the lack of great tools. However, given the sensitivity about data these days (particularly

in highly-regulated industries), most people don't tend to want to go on the record about it. But an IT executive at a biotech company in my snowy hometown of Cambridge, Mass. told me that they are working on a catalog of all their data. In his previous job at a research institute, he said "we lost a lot of data because we simply couldn't find it." At his new company, he declared it a priority for the organization to "know what we have, where it is, and what can and can't be used." The company's datasets in the catalog include genomic data, clinical trials data, and "real world" datasets after products have been introduced to the market. He said his company saved several hundred thousand dollars at one point just by showing someone who was about to buy a database that the company already had it.

Another executive I interviewed comes from a financial services company, where he is head of IT innovation. He often gives a presentation inside and outside his company called "Do You Know Where Your Data Is?" He finds that the vast majority of companies he talks with don't know the answer to that question, and hence can't use their data very effectively. He's working on a project at his own organization to identify the data that's important to make business decisions, who owns it, who decides access, how long to keep it, its data quality, and so forth. This is a bit more than a catalog, but it's the same general idea.

In both of these industries—biotech and financial services—you increasingly need to know what data you have. And it's not only so you can respond to business opportunities. Industry regulators are also concerned about what data you have and what you are doing with it. In biotech companies, for example, any data involving patients has to be closely monitored and its usage controlled. In financial services firms there is increasing pressure to keep track of your customers' and partners' "legal entity identifiers," and to ensure that dirty money isn't being laundered by them.

But if you don't have any idea of what data you have today, you're going to have a much tougher time adhering to the demands from regulators. You also won't be able to meet the demands of your marketing, sales, operations, or HR departments. Knowing where your data is seems perhaps the most obvious tenet of information management, but it has thus far been among the most elusive.