Hey! What's New? 2024-137

Data Is Risky Business: Structured or Unstructured

Daragh O'Brien write, in an article in the latest issue of the *Data Administration Newsletter*, that "for the bulk of my career in data, there has been a simmering tension under the surface between what has traditionally been referred to as 'structured data´ and 'unstructured data.'" He notes that Wikipedia helpfully defines unstructured data as data that "does not have a pre-defined data model or is not organized in a predefined manner." Wikipedia unhelpfully defines structured data with a redirect from that search term to an entry about data models. "From this, we must infer that structured data is data that is defined in some form of relational model or, umm, structure."

O'Brien's view, the debate been exacerbated by the existence of two broad camps in the world of data and information management. "Historically, we have had the information managers, who have tended to come from the traditional records management and library science worlds, and the data managers, who have tended to come from the information technology and number-crunching side of the fence. In a manner reminiscent of a very nerdy reimagining of *West Side Story*, these two camps have not always seen eye to eye or had equal billing in the priorities of the organization."

O'Brien believes that there is a problem with this simply dichotomous view of the world. All "unstructured" data has some level of structure, he says.

Perhaps this distinction between "data that lives in a database" and "data that lives in a document" made sense in the past. Regardless, he points out, "the distinction makes little sense now in the age of machine learning and technologies that can extract and infer context and structure from text, audio, video, and other forms of recorded content and data. This ability for technology to take data in different formats and extract, apply or infer metadata means that all data is now structured as it is possible to identify concepts and entities in the data, to infer relationships between those concepts and entities and to then take that data and link it to other data selected from relational databases and present it to the knowledge worker or end customer."

For example, a standard machine learning process can identify and categorize content in a document, apply metadata to the content in the content management system and improve the searchability and findability of the data in that document by *putting it into a defined structure of metadata.*

According to O'Brien, "metadata is, after all, the data that defines other data. In the world of 'data that lives in a relational database,' that might be the data that defines the meaning of a business concept encapsulated in the data, or it might be the definition of the relationship between two entities. In the world of 'data that lives in a document,' metadata might define the file name, or it might define a business concept contained in the document, or it might define an important attribute of data described in the document. After all, whether we are looking at entries in an ERP database or copies of documents in a shared drive, one would hope that the concept of Customer, Account, Product and Transaction and all the attributes that

might be associated with these entities would be the same regardless of the type of bucket (database or document) that the data is being stored in."

But this raises another age-old challenge, he says. "As technology advances in its ability to figure out the metadata and structure of content in different containers, we need to ensure that there is management and governance of that data about data so that we humans can make sense of the world as seen through data."

So, he adds, "we can leave it to people to figure out for themselves, or we can start to invest in developing the power of the 'chief magistrate.' If we're developing a magistrate function for data, that sounds suspiciously like data governance."

But he notes, right now," we are, in effect, quite often leaving people to figure it out for themselves. 'The business will figure it out themselves' seems to be a common default setting in the IT change plan for content management. But this is replicated across the myriad of systems and solutions that are deployed at departmental or team level in organizations, or in the multiplicity of reporting dashboards and analytics models that spring up when knowledge workers are given access to boundless data in reporting tools such as PowerBI (or Excel)."

O'Brien concludes that, "if the 'secret sauce' to the egg-opening conundrum of 'data in a relational database' and 'data in a document' is the proper governance the metadata, then we need to embrace this key to enlightenment with both hands!"