SocietyByte
Wissenschaftsmagazin des BFH-Zentrums Digital Society
https://www.societybyte.swiss

Turning Big useless Data into useful Big Data

Autoren : Luciana Tricai Cavalini, Tim Cook

Datum : 15. Juli 2019

The S3Model approach can help organizations transform large amounts of disparate, unconnected data into meaningful data. This will save time and costs when attempting to exploit this data across a wide range of applications, write our authors.

Data is being collected everywhere. From the international conglomerate to the shop on the corner; from hospitals to all levels of government. These organizations often store data and do not use it. They often do not know why they are storing this data there is a general feeling that because everybody else in the business has a system to collect data, we must also do that.

Moreover, there are some questions which we need to ask them about all this data:

- Are some of the data points collected incomplete?
- How much time is devoted to massaging data?
- Do reports come from tables in databases that no one seems to understand?

- Does data from one department look similar to data from another; but no one is sure if or how it is connected?

If your organization answered yes to any of these questions, it has a data quality problem.

The terms Big Data, Internet of Things, Artificial Intelligence, Blockchain, Data Lakes and other panaceas are tossed around all over the tech industry and throughout the popular press. These technologies are often touted as the next big thing. But everyone is talking about the value of collecting all these disconnected data, but nobody is talking about its quality. The underlying data quality issues are as good as ignored: data quality is the elephant in the room.

When attempting to *fix things*, there's a tendency to dump everything into a data lake, without a schema, hoping that it will be sorted out later. This is an even bigger mistake. It might be easy to save everything but in the end, it won't be useful if its meaning is unknown.

To get a better quality of the data

With more and more data being committed into schemaless environments, data quality is becoming worse and we can expect it to continue to do so. It is now more difficult to determine the quality of the data for any particular purpose. There is often little to no governance information and there is no idea of the context of collected data.

The result is that there's a lot of data is being collected, but nobody does anything with it because of its questionable quality, meaning and manual labor required to make the data useful. How can we use this data, if nobody really knows how good or bad the information is? The manual intervention required in data preparation is too time-consuming to be cost-effective.

All of us should be concerned about data quality. Any reasonable proposal for improving it should include a way to properly capture data semantics. The semantics include not only the specific meaning of a given data point but also what it means in relation to the other data points captured at the same time and place. Linked data can help with retaining these semantics - when properly applied.

The use of semantic graph databases is the new trend. The utility of using semantic graphs for knowledge discoveryis also well known at some of the largest content companies. However, there is currently no implemented, systematic approach to creating semantic graphs that:

- will ensure data quality
- will provide the complete context of the data
- will provide a pathway to integrate existing data

The current process is expensive, slow, manual and error-prone. If you know of a different approach, let us know.

This is our "change my mind" meme challenge.

When determining that certain data should be captured for re-use, there are specific considerations that should be made. Starting with a very simple question: “What is the datatype of the data” and going on with:

- For quantitative data, is there a unit of measure?
- Is this data captured at specific locations only?
- What are the rules or guidelines in place that govern what this data means?
- Why do we want to capture it now?

Until we reach to the very broad “What does this data mean in the context it was captured”?

Data need surrounding context

When these questions are answered and recorded in a computable manner; the data becomes information. This information is what humans can use to make decisions. Simple data points are not very useful without the surrounding context. Since we build computers to emulate our decision-making process, then we must be able to encode this context in a way that the computer can interpret and process.

It is impractical to encode this context in every existing and future software applications that might need to process this information. A sharable, computable model solves this problem What happens in the current world is that new context and semantics may be added or changed so that the meaning of the original data may be lost or obfuscated. Each time the data is exploited in a new application its meaning may be changed kind of like the Telephone Game; you seldom get back what was originally intended.

But it is possible (although not trivial) to record and share contextual information using standards-based technology. Using standard XML, RDF and OWL we have defined an approach and process that we call the Shareable, Structured, Semantic Model(**S3Model**).

Data will be modeled for the user

S3Model is a new foundation for harmonizing*linked data*across information domains. It consists of a small core ontology of 13 classes and 10 object properties used to organize the components of the information model, which consists of nine base classes used to represent types of information. We also allow the definition of the spatial, temporal and ontological context of each individual data item.

The data modeling process in S3Model consists of arranging items as a document-like component. The content and structure of the component are determined by the person who is defining the data that will be collected. We call this person the *Data Model Designer*, who is someone that understands what must be included in the application because she is the final user or she knows very well the needs of the final user.

The final result of the work done by the Data Model Designer is a set of Data Models, which are

used to constrain the information in a data instance as an XML Schema (canonical expression) or another programming language such as Python, Go, Ruby, C++ or Java.

The Data Model Designer can express the appropriate semantics using popular open ontologies and vocabularies as well as any internal private vocabularies. Once defined in this model, the semantics and structure can be easily exchanged across all platforms without loss of fidelity. Data instances can be serialized as XML, JSON, and RDF to allow for the maximum cross-platform exchange and ease of analysis capabilities.

The S3Model

The key to widespread S3Model-based systems implementation is the tooling: the S3Model Tool Suitewas developed to allow domain experts, with minimal training, to become Data Model Designers. With the available online training, a domain expert can design a Data Model that is as technically robust as it is semantically rich.

In addition to building models from scratch, tools are available to convert existing data into semantically rich S3Model data: the open source S3Model Translatortool helps create a model for any Comma Separated Value (CSV) file such as a database extract or spreadsheet export. After being imported into the S3Model Translator, the tool guides the Data Model Designer through creating an S3Model-based Data Model, which can be used to validate data from the imported CSV file or any CSV file that has the same structure. This provides a pathway for your data from the flat, table-like world into the semantically rich linked data world.

The future of sharable, computable models for all datasets sounds wonderful. But there is no way that a rip and replace approach can ever work. That is why part of the design process was to ensure that there was a gradual pathway built into the S3Model ecosystem. Therefore the translator tool is a great step towards achieving this process. The tutorials and examples of the open source tools as well as our upcoming courses demonstrate in great detail the advantages of being able to connect data across models and across domains when the models are built by subject matter experts (SME) that understand how to use open vocabularies and ontologies to semantically markup their data.

The additional advantages of S3Model lay in the ability for an SME to model required governance requirements especially in domains such as healthcare, finance, etc. where there are specific legal constraints. These constraints can be for privacy or recording all contributors to data sourcing or editing throughout the workflow.

As examples and to demonstrate the broad capability and robustness of S3Model, we have translated all of the HL7 FHIR resources and all of the unique entries in the US NIH Common Data Element repository. We also did some proof of concept models for the NIEM models and XBRL.

We invite you to take a look at our open source offerings and participate in improving these tools so that we can all have better quality data.