The Power of Babble

**EXPECT TO BE
CONSTANTLY
AND PLEASANTLY
BEFUDDLED**

PAT HELLAND

Metadata defines the shape, the form, and how to understand our data. It is following the trend taken by natural languages in our increasingly interconnected world. While many concepts can be communicated using shared metadata, no one can keep up with the number of disparate new concepts needed to have a common understanding.

English is the lingua franca of the world, yet there are many facets of humanity and the concepts held by different people that simply cannot be captured in English no matter how pervasive the language. In fact, English itself has nooks, crannies, dialects, meetups, and teenager slang that innovate and extend its permutations with usages that usually do not converge. My personal idiolect shifts depending on whether I am speaking to a computer science audience, my team at work with its contextual usages, my wife, my grandkids, or the waiter at a local restaurant. Different communities of people extend English in different ways.

Computer systems have an emerging and increasing common metadata for interoperability. XML and now JSON fill similar roles by making the parsing of messages easy and common. It's great we are no longer arguing over ASCII versus EBCDIC, but that's hardly the most challenging problem of understanding.

As we move up the stack of understanding, new

*It's Not Your Grandmother's
Database Anymore

subtleties constantly emerge. Just when we think we understand, the other guy has some crazy new ideas!

As much as we would like to have complete understanding of each other, independent innovation is far more important than crisp and clear communication. Our economic future depends on the “power of babble”.

THE APOCALYPSE OF TWO ELEPHANTS

To facilitate communications, the computing industry, various companies, and other organizations try to establish standard forms of communication. We see TCP, IP, Ethernet and other communication standards as well as XML, JSON, and even ASCII making it easier to communicate. Above this, there are vertical specific standards (e.g. health care and manufacturing standards). Many companies have internal communication standards as well.

Dave Clark of MIT observed that successful standards happen only if they are lucky enough to slide into a trough of inactivity after a flurry of research and before huge investments in productization (figure 1). This observation is known as the Apocalypse of the Two Elephants (although Clark actually didn’t name it that).¹

Standards that happen in this trough are effective and experience little competition. If a standard doesn’t emerge here or the trough is squished by the two humps overlapping, it’s a much murkier road forward.

*The best de jure standards are rubber stamps
over de facto standards.*

1

FIGURE 1: THE APOCALYPSE OF TWO ELEPHANTS

If there's no de facto standard to start from, then the de jure standard typically contains the union of all ideas discussed by the committee. Natural selection relegates these standards and their clutter to history books.

THE DIALECTS OF THE BUSINESS WORLD

Computer systems and applications tend to be developed independently to support the special needs of their users. In the past, each system would be bespoke and support detailed specifications. Increasingly, shared application platforms are leveraged, either on premises or in the cloud. In these common apps, there is common metadata—at least as far as the apps have a common heritage.

When applications are independently developed, they have disparate concepts and representations. Many of these purchased applications are designed for extensions. As the specific customer gloms extensions onto the side

Whenver there are two representations of data, either somebody adapts or the fidelity of the translation suffers.

of the app, this impacts the shape, form, and meaning of its internal and shared data.

When there's a common application lineage, there's a common understanding of its data. Popular ERP (enterprise resource planning), CRM (customer relationship management), and HRM (human resources management) applications have their ways of solving business problems, and different companies that have adopted these solutions may find it easier to interoperate.

INTERNECINE INTEROP

Still, challenges of understanding may exist even across departments or divisions of the same company. A large conglomerate may sell many products, including light bulbs, dishwashers, locomotives, and nuclear power plants. I would hazard a guess that it doesn't have a single canonical customer record type.

Of course, mergers and divestitures impact a company's metadata. I know from personal experience how hard it is to change my mailing address with a bank or insurance company. They can't seem to track down all the systems that record my address even over the course of a year. It's not a big surprise that they have a hard time managing their metadata.

WHAT'D YOU SAY?

Whenever there are two representations of data, either somebody adapts or the fidelity of the translation suffers. In many cases, the adaptation is driven by economic power. When a manufacturer wants to sell something to a huge retailer, it may be told exactly the shape, form, and

semantics of the messaging between the companies. To get the business, the manufacturer will figure it out!

The dog wags the tail. In any communication partnership, the onus to adapt rests on the side that most needs the relationship to work.

Translating between two data representations may very likely be *lossy*. Not all of the information in one form can be moved to the other form. It's highly likely that some stuff will be nulled out or possibly translated into a form that doesn't precisely map.

Each translation is lossy. By the time the translation occurs, a loss of knowledge has occurred. The best results will be from dedicated transformations designed to take exactly one source and translate it as best as possible to exactly one target. This is the least lossy form of translation. Unfortunately, this results in a boatload of translators. Creating a specific conversion for each source and destination pair results in great conversion fidelity but also results in N^2 converters (see figure 2).

What to do? Many times, we simply capture a canonical representation and do two data translations: first, a lossy translation into the canonical representation; then, a lossy translation from the canonical representation into the target representation. This is *double-lossy* and just doesn't supply as good a result.

Why do the translation to a canonical form? Because only $2N$ translators are needed for N sources, and that's a heck of a lot fewer than N^2 , as N gets large. Using canonical metadata as a common translation reduces the number of

2

FIGURE 2: **LEAST-LOSSY CONVERSION**

converters but results in a double-lossy conversion (see figure 3).

In most cases, people use canonical metadata to bound complexity but add specific source-to-target translators when the lossiness is too large.

WHAT COLOR ARE YOUR ROSE-COLORED GLASSES?

We all see stuff couched in terms of a set of assumptions. This is a worldview that allows us to interpret incoming information. This interpretation may be right or wrong, but, more importantly, it is right or wrong for our subjective usage.

Computer systems are invariably designed for a certain company, department, or group. The data is typically cast into a meaning and use that are appropriate for one side but lose their deeper meaning through the translation.

FIGURE 3: “DOUBLE-LOSSY” CONVERSION

Sometimes, the meaning and understanding of some data are deeply couched in cultural issues. Any translation to a new environment and culture simply loses all meaning. Reading about daily life in Medieval Europe doesn't help much unless you study the relationships between serfs and lord as well as between men and women. Only then can you understand the actions described in the book. Similarly, in any discussion of privacy, cultural expectations must be addressed. In North America and Europe, protecting against the damage that may result by disclosing a medical challenge is paramount. In India, the essential need to vet a prospective spouse for your child is deemed more

important than holding an illness private. Communication cannot take place without understanding the assumptions and interpreting through that lens.

The artificial language Esperanto was created in 1887 with the hope of achieving a common shared natural language for all people. Some folks grabbed hold and used it to write and share. Some say a few million speak it today.

The use of Esperanto has been waning, however. Each of the roughly 6,000 languages spoken by different communities in the world has its own flavor and nuance. You can say certain things in one language that you just can't say in another one.

DIVERSE AND HOMOGENEOUS

The words and phrases people use and the metadata that applications use follow a similar pattern. With a common codebase DNA and history, some meanings are the same. As time, evolution, and commingling occur, it's harder to understand one another.

New software applications either in the cloud or on premises sometimes offer enough business benefit that enterprises adapt their ways of doing business to fit the application. The new user adopts the canonical representation of data and business processes by sheer hard work. When the business value of the software is high enough, mapping to it is cost effective. Now the enterprise is much more closely aligned to the new approach and to interoperating with other enterprises sharing the new data and process.

Next, the enterprise will begin to extend the system using extensibility features. These extensions can then

become a source of misunderstanding, but they bring business value to the enterprise.

The United States, Canada, and many other Western countries have tremendous diversity in their populations. New arrivals bring new customs. They work to understand the existing customs in their new home. While there are many differences at first, in a few short years the immigrants fit in. Their children are deeply ingrained in the new country, even though they still like some of that food their mom cooked at home. That food becomes as American (or English or German) as pizza, tacos, and falafel. Similarly, the base metadata continues to move and adjust as it assimilates those new messages and fields that made no sense at all a short time ago.

RELISHING DIVERSITY

While not understanding another party is a pain, it probably means that innovation and growth have occurred. Economic forces will drive when and where it's worth the bother to invest in deeper understanding.

Playing loose with understanding allows for better cohesion, as exemplified by Amazon's product catalog and the search results from Google or Bing. Remember that in many cases, cultural and contextual issues will drive how something is interpreted. Extensible data does not have a prearranged understanding. Translating between representations is lossy and frequently involves a painful tradeoff between expensive handcrafted translators and even lossier multiple translations.

Personally, as the years have gone by, I've gotten much more relaxed about the things I don't know and don't

understand. A lot of stuff confuses me! As we interoperate across disparate boundaries, it would do us well to remember that the less stressed we are about perfect understanding and agreement, the better we will all get along. Moving forward, I expect to be constantly and pleasantly befuddled by the power of babble.

REFERENCES

1. Clark, D. 2009. The Apocalypse of Two Elephants, or “what I really said.” Advanced Network Architecture. MIT CSAIL; <http://groups.csail.mit.edu/anal/People/DDC/Apocalypse.html>.

Pat Helland *has been implementing transaction systems, databases, application platforms, distributed systems, fault-tolerant systems, and messaging systems since 1978. For recreation, he occasionally writes technical papers. He currently works at Salesforce.*

Copyright © 2016 held by owner/author. Publication rights licensed to ACM.