Incompleteness, indeterminacy, and holism in the construction of crosswalks, or, why there's no substitute for knowing your data

Patrick Herron
INLS 150 Spring 2005
Reflection 2

Part 1. Innocence: A prolegomenon[1]

> The difficulty [with claims as to the functional nature of things] is not that physically possible organisms don't have functional organizations, but that they have *too many*. …there is a sense in which *everything has every functional organization*. When we are correctly described by an infinity of logically possible "functional descriptions," what is the claim supposed to *mean* that one of these has the (unrecognizable) property of being our "normative" description?[2]
>
> - Hilary Putnam, *Representation and Reality*

1. Functional organization, the essence of computational description, is indeterminate.

2. The identification of the meaning of a word is the association between two stimuli such that at least one of the stimuli is the word itself, either spoken or written.[3]

3. The description of any stimulus is incomplete.

---

[1] Author's note: In lieu of a traditional introduction I have instead derived or borrowed and subsequently presented several crucial philosophical insights by influential thinkers in the hopes that the points may act as a loose framework embodying some of the very spirit of the discussion. It is nearly impossible for me to discern which of these points I am making come from Quine, or from Putnam, from Godel, or from Wittgenstein, from Nelson Goodman or Walter Benjamin or from my friend Murat Nemet-Nejat, or from my own experience or thinking on the subject, however feeble that may be.

[2] Putnam, xv.

[3] See Quine's *Word and Object*.

4. No naturalistic means (as in, "assessing the true nature of x") of acquiring functional descriptions is useful; there fore we must denaturalize formal descriptions

5. Translation of linguistic information is radically indeterminate; there is simply no correct one.

6. There seems to be no way to connect a word to a physical object. Any capture of a meaning of a word is incomplete, and when such meaning is placed in a functional description, there's no end to the possibilities, and no obvious way to pick the right one.

7. The development of lexical databases, semantic tools, ontologies, controlled vocabularies, crosswalks, etc. are subject to incompleteness and the complete lack of objective evaluative criteria.

8. Understanding is possible, despite the challenges of incompleteness and indeterminacy.

> The sign (the sentence) gets its significance from the system of signs, from the language to which it belongs. Roughly: understanding a sentence means understanding a language.[4]
>
> - Ludwig Wittgenstein, *The Blue and Brown Books*

Part 2. Experience

In 2000 I found employment at a small telecommunications startup as the fifth member of the IT department. The company specialized in providing telework solutions to Fortune 100 companies; by providing telework solutions I mean that we as an organization helped

[4] Wittgenstein, 5.

companies like Nortel and BellSouth in their efforts to provide opportunities to their own employees to work from home or other out-of-office locations. From a data management perspective, our customers by that time were not the companies as wholes but instead individual employees within those companies. Avaya's Bob Smith in Tuscon AZ needs a second telephone line in his home; HP's Sally Johnson's network connection to her office is currently down. And so on. By the end of 2000 our department, then 18 strong, had implemented a customer service database that put in place the storage infrastructure for maintaining customer information. At the same time XML was emerging as a means for business-to-business data exchange. The time was ripe for integrating with our corporate clients' databases. All we needed was the plumbing to connect our databases to our customers', that, and a map that represented the translation of the data between corporate client databases and our own.

In early 2001 I became technical lead of all things extensible in our XML-to-XML integration project. In other words, I was responsible for creating, implementing, and maintaining the DTD for the XML-to-XML system as well as the XSLT that dictated the function of the mapping. Our system allowed us to integrate our customer db with employee databases for BellSouth, Southwestern Bell, Avaya, HP, Nortel, and a number of other companies with quite large employee populations. That integration meant that, for example, if someone in our call center received a call from an Avaya employee, and in addition to a help request perhaps we learned that the employee had a new phone number, the information from that call would not only update our internal records but also Avaya's records.

The model for exchange we designed required that each company produce their data in XML format, it also required that we translated their XML into an internal XML format. Mapping our data structure to each client structure took some time, but after the creation of a 30 page DTD, we were successful. What was interesting was that, in practice, there were two never-ending problems, problems that at the time I considered fatal but I now consider necessary and even indicative of success. The first problem was that the relationship in models—in schemas—between our structure and theirs was in constant need of revision usually coming in the form of extension. The second problem was that of harmonization: how to get the other companies in line with the new structures. Simply having the new DTD readily available simply wasn't enough. Changes would need to be implemented for each change in each company.

The lesson was that no translation could ever be complete, and that improving such a translation by expansion does make things seem more complete, yet it never led to being closer to completeness. There was no such thing as complete. Certain attributes might need to be decomposed, while others might need to be brought together in order to make our "router" field, for example, better fit with HP's "networking equipment" field. And so on. You could update as often as sanely possible, stop updating in the realization that no models were ultimately more complete than others, or teeter dangerously between the two. It seemed at the time work on the DTD would never be complete, and as such it felt like a failure.

But it wasn't a failure—in retrospect I should have assumed that the translation could never be complete but rather could be treated as if it were in constant need of continual expansion, and that treatment is essentially an arbitrary practice. Setting the expectations such that everyone involved would understand that the translated set of data should be *minimally* expressive across all organizations would have likely been the best way to handle the situation. Lowering expectations serves only in cases where there's no hope of coming up with a cross-organizational single schema, of course, but even picking a single overarching structure only temporarily holds off the infinite refinement problem resulting from the indeterminacy and incompleteness.

3. Translation and crosswalks: Language holism & pidgin schemas

Imagine we have two copies of Cervantes' *Don Quixote,* one in English and the other in Spanish. Both are accepted translations having been in circulation for some time. We might safely say that neither one is THE *Don Quixote* but rather, they are both *Don Quixote.* Each text taken as a whole is equivalent to one another in the sense that they tell the same classic story. Yet, interestingly enough, these wholes are so much more than the sums of their parts.

Imagine we take the English *Quixote* and, with schemas for English and Tajik and as well as an English to Tajik crosswalk we decide to create a Tajik version of the *Quixote.* What might we expect to see? Certainly we cannot expect to see the same *Quixote* come

out the other end. Translation work that takes this approach inevitably reads like translations in Babelfish; treated as the sum of its parts, any body of text will, in such a bottom-up translation approach, fail to remain the *Quixote.*

It seems that one of the ways in which translators work successfully is in knowing the two languages in which they work quite well. They get the whole picture of a work in one language and can construct that whole in another language, without having to resort to a functional/computational set of rules for doing so. There's not really any scientific way to express the equivalence of textual wholes, but when wholes are not equivalent, it is quite obvious. The translator acts as someone who is able to understand both language domains as wholes, wholes that cannot be decomposed into complete sets of interoperable parts, and use that understanding as a guide to recreate the whole in a secondary form without any allegiance to every element in the original form. Translation, it seems, is not functional but rather organic, perhaps even magical--an *art*.

Now, in our XML-to-XML case, there was never any hope that any one person would ever be knowledgeable about each company's data structures and cultures represented therein, for any countless numbers of pragmatic territorial reasons. Here we do not have the opportunity for any sort of multilingual translator.

With natural language it seems that in the absence of translators people with different languages do quite well in communicating with one another. Pidgin languages often form at the interfaces of two languages. These pidgin languages never approximate any

of the expressiveness of either of the languages they seem to conjoin, but they nevertheless seem to work for basic purposes, such as in the trade of goods. Parties using pidgin languages do not attempt nuanced descriptions of items but rather accept that they are in place for very necessary exchanges.

In the XML-to-XML example, it seems Edward Johnson's article on LinguaNet[5], "Talking Across Frontiers," provides an excellent example of what to strive for: exchange statements that are coarse, simplistic, yet get the salient and minimally essential points across. In Johnson's article we are presented with a problem such that law enforcement officials at either end of the English Channel Tunnel do not generally speak the same language, whether that language is the broader one (French vs. English) or a more specific one (law enforcement jargon speak). Any number of border transgressions (e.g., smuggling) or hazards (e.g., accident) in the tunnel need to be communicated from the British Law Enforcement on one end to the French on the other, and vice-versa. Yet the language barriers demanded the development of a sort of formal Pidgin language, to facilitate communication despite those barriers.

[5] http://www.prolingua.co.uk/talking.pdf

4. Understanding a sentence means understanding a language

> The mappings shown in this table are proposed and somewhat loose. The mapping does not provide all the information that would be needed to do an actual conversion of ONIX data to MARC 21. In some cases, interpretation and special processing would be required of the content of an ONIX data element to render data compatible with the content of the corresponding MARC 21 data element. Often a choice between more than one potential MARC 21 data element is required.[6]
>
> - ONIX to MARC 21 Mapping, Network Development and MARC Standards Office, Library of Congress

"Interpretation and special processing," the above passage reads, "would be required to … render compatible" ONIX elements with MARC 21 elements. In other words, we need either our master translator or a general acceptance of inequivalence and minimal expressiveness. There is not always a 1-to-1 relationship between two elements of different schemas, even if those elements are designed to describe the same wholes, as both ONIX and MARC21 are intended for capturing bibliographic information.

We should expect crosswalks to be minimally expressive, incomplete, *ad hoc*, and wholly dependent upon not any correct model of exchange but rather one that works. No crosswalk should attempt to perfectly integrate two differently-bound models. It is entirely more reasonable to expect that people working with the two original schemas should learn about each others' data and combine efforts to create a single schema covering both data sets—if the desire is to avoid data loss and cross-model integrity. If there is no opportunity for individuals to get a comprehensive knowledge of the

[6] http://www.loc.gov/marc/onix2marc.html#mapping

languages or models within the problem domain, then it should be anticipated that any working model should be simple, improvisational, and minimally expressive.

Bibliography

Johnson, E. (2000). *Talking across frontiers*. Paper presented at International Conference on European Cross Border Cooperation: Lessons for and from Ireland.

Network Development and MARC Standards Office. "ONIX to MARC 21 Mapping." *MARC Standards*. 07 Feb. 2005. Library of Congress. 21 Apr. 2005 <http://www.loc.gov/marc/onix2marc.html>.

Putnam, Hilary. *Representation and Reality.* MIT Press, Cambride, MA, 1996.

Quine, W.V.O. *Word and Object*. Wiley & Sons, New York, 1973.

Wittgenstein, Ludwig. The Blue and Brown Books. Harper Torchbooks, New York, 1965.