Failing universal classification schemes from the Renaissance to the Semantic Web

ANIMALS THAT BELONG TO THE EMPEROR

FLORIAN CRAMER

The weapon with which European search engines intend to beat Google is semantic information processing: pattern recognition in media in the case of Quaero, Semantic Web technology in Theseus, its German cousin. Originally, the Quaero project was a French-German collaboration, funded by both governments, until the German Theseus project split off from Quaero to pursue its own vision of future Web search. This vision is twofold, involving a number of classical holy grails of computer science: (a) provide search on the basis of semantic tags, (b) have software recognize the contents of web pages in order to automatically apply those tags. While the second point is utopian enough and something that Artificial Intelligence research hasn't achieved in decades, I would like to show why even point (a), in other words: the Semantic Web, is doomed to fail. (I leave it to the audience to draw its own conclusions why the Theseus projects receives high public funding nevertheless.)

The "Semantic Web" itself is a highly misunderstandable term and project. It was founded and is pursued by Tim Berners-Lee, the creator of the World Wide Web in the early 1990s. In 2004, prior to Quaero and Theseus, the German federal government subsidized research on the Semantic Web with 13.7 million Euros, reasoning that as a "semantic technology", it would allow people to phrase search terms as normal questions, thus giving computer illiterates easier access to the Internet. But, alas, this is not at all what the Semantic Web is about, or what it would even implement; it was, in another words, a 13.7 million Euro misunderstanding.[1] Indeed, natural language parsing is another holy grail of search engine development, from "Ask Jeeves" (which renamed itself Ask.com, and deemphasized its initial concept) to . . . , recently brought up by Geert Lovink on the Nettime

[1]Ich hab irgendwie den Eindruck dass unser Bundesforschungsministerium in der irrigen Annahme ist, das 13 Millionen Euro eine Software schaffen die es jedem Computer-Analphabeten ermÃ¶glicht, ganz ohne den "Extra Effort" seine "Pisa-Versagen vermarkten und als hochinnovative Rettung des Wissens- und Wirtschaftsstandorts Deutschland (wers glaubt . . .),

mailing list. Natural language parsing falls into the category b I mentioned before, as the Holy Grail that artificial intelligence research for couple of decase has consistently claimed to have almost, but just not quite reached – while critical A.I. researches like Luc Steels say that it won't ever be reachable with current algorithmic computer architectures, regardless their speed. And in reality, natural language search systems are no more than inefficient interface wrappers around classical Boolean search expressions with logical AND, OR and NOT operators.

The Semantic Web does not fall into this trap because it does not involve any automatic interpretation of meaning. Berners-Lee is quite outspoken on this, saying that his concept "does not imply some magical artificial intelligence which allows machines to comprehend human mumblings" – much in contradiction to Quaero and Theseus. Instead, his Semantic Web is a universal markup or "tagging" schemes. In Berners-Lee's words: "Instead of asking machines to understand people's language, it involves asking people to make the extra effort". This effort, semantic tagging is well-known and a popular device on sites like Flickr, digg.com and delicious. It simply means that users attach keywords to texts, images and other information, so that this information can be searched by its keywords or particular keyword combinations. On Flickr, for example, the search keyword combination "birthday", "children" and "clown" results in a list of pictures of clowns appearing at children's birthday parties, not because of any Quaero-style computer recognition of the image contents, but because of the keywords assigned to images by Flickr users.

From a Semantic Web perspective, this system is flawed though, because there are no nomenclatures for tagging. For example, a user might have tagged an image with "kids" instead of "children", so it won't turn up in the results. And the tags lack systematization: for example, children could be classified as a subset of humans, humans as a subset of mammals; birthdays as a subset of celebrations etc.etc. Then one would also find pictures marked up with "birthday" and "children" in a more general search for pictures of human celebrations. This is why unsystematic, ad-hoc, user-generated and site-specific tagging systems like on Flickr are referred to as "folksonomies".[2]

[2]Wikipedia: Folksonomy (also known as collaborative tagging , social classification, social indexing, social tagging, and other names) is the practice and method of collaboratively creating and managing tags to annotate and categorize Content. In contrast to traditional subject indexing, metadata is not only generated by experts

The Semantic Web promises to overcome those folksonomies with one, unified and standardized keyword tagging system that can applied to anything. In other words, it is a universal classificatory description system, a grand unified hierarchical metatag design.

For somewhat mysterious or at least idiosyncratic reasons, Berners-Lee calls this classifactory system an "ontology", making his project particularly confusing for people with backgrounds in philosophy and humanities – because it is not an ontology, but a cosmology.

Just as cosmologies are by no means new, so are universal classification and tagging systems of the world. In his essay and short-story "The Analytical Language of John Wilkins", Jorge Luis Borges writes about the English 17th century scholar that

> "He divided the universe in forty categories or classes, these being further subdivided into differences, which was then subdivided into species. He assigned to each class a monosyllable of two letters; to each difference, a consonant; to each species, a vowel. For example: de, which means an element; deb, the first of the elements, fire; deba, a part of the element fire, a flame." [...]

Similar classification schemes have been developed throughout the Middle Ages and Renaissance, by Ramon Llull, Giordano Bruno, and especially in 17th century encyclopedism of Johann Heinrich Alsted and Jan Amos Comenius in whose tradition Wilkins works and thinks. Since encyclopedias, before Diderot and d'Alembert, structured their knowledge systematically, not alphabetically, they developed increasingly complex tree-like classification systems of all things in the world as described in them. [picture] The so-called "ontology" of the Semantic Web does not only do something similar, but it does exactly the same again.

The Renaissance classificatory cosmologies could only work on the basis of a stable assumption of what the world is and how it is structured: for example, by the four directions, the four seasons, the temperaments, the seven virtues and vices etc. In other words, they were still embedded into the paradigm of Medieval scholasticist science which in turn was derived from Aristotle's system of categories that

---

but also by creators and consumers of the content. Usually freely chosen keywords are used instead of a controlled vocabulary."

broke up all things in the world into genres and species. The Semantic Web boils down to nothing else but technocratic neo-scholasticism, and a questionable if not dangerous belief that the world can be described according to a single, objective, universally valid viewpoint and classification – a blatant example of an engineer's blindness to ambiguity and cultural issues.

Although there was no Semantic Web yet in the 1940s, Borges pins down the issue in his essay. One is tempted to just replace the name John Wilkins with Tim Berners-Lee when he examines the former's categories to find out that stones, for example, are absurdly classified as either common, or modic, precious, transparent and insoluble, or that beauty is assigned to a "living brood fish". Borges' concludes that

> "These ambiguities, redundancies and deficiencies remind us of those which doctor Franz Kuhn attributes to a certain Chinese encyclopaedia entitled 'Celestial Empire of benevolent Knowledge'. In its remote pages it is written that the animals are divided into: (a) belonging to the emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies."

Although this is Borges' own fiction, it nevertheless reveals the arbitrariness of categories and classifications. It also had a thorough impact as a philosophical critique. The first chapter of Michel Foucault's "The Order of Things" is a discussion of precisely the above list of animals. Foucault confesses that for him, it "shattered all the familiar landmarks of his thought", opening his eyes on how the order of knowledge is culturally constructed, and may be thought differently. To understand Foucault's discourse theory, in order words, one just needs to read Borges.

But the order of things, and unified classification schemes, do not just break down in fiction. Sticking to the example of animals, it is obvious how Aristotelian philosophy continues to exist today, in the notion of gender and species [and even more questionably in the categorization of humans into biological races], we're still working within the paradigm introduced by Aristotle's zoology. But the problem is that does it not always work even in zoology itself. The prime

example is the platypus, an Australian animal that is a breastfeeding mammal, but it lays eggs, lives in the water and has a brill like a bird. If the platypus breaks genre and species classification of zoology, where does it fit the Semantic Web?

This situation is by no means new. In his book "Kant and the Platypus", Umberto Eco makes the animal a symbol for scholastic versus empirical science. A bit confusingly, he differentiates "cultural cases" – that means categorically defined phenomena – from "empirical cases", i.e. phenomena that are observed instead of predefined. "To be recognized as such," Eco states, cultural cases "need reference to a framework of cultural norms" (Eco 1997, p. 139). For Eco as a semiotician, this means that Being, or existence, is the frontier that systematic science cannot conquer. And this is what ontology means.

The innovation of modern science since Galileo, Newton and Descartes is that it operates without the reference to those norms. When Diderot and d'Alembert abandoned the old classificatory order of knowledge in encyclopedias and replaced them with a non-classificatory, non-systematic alphabetic order, they precisely followed the empirical paradigm, taking phenomena as they occurred and not as they fit. In other words: It was the innovation of modern critical science that it gave up "Semantic Web" schemes in its ordering of knowledge.

But to go back from academic discourse to folksonomies on the Internet, an even better example than the Platypus was brought up in a Web forum of the computer news site heise.de. Discussing the Semantic Web and its classification scheme, an anonymous poster brought up the hypothetical example "A Muslim is a potential terrorist" in order to show that a unified "ontology" cannot be built.[3] And this example scratches only the surface of the pending cultural problems. In other words, not the empirical, but the cultural cases bear the actual dynamite. The whole Semantic Web, and the search engines built upon it, rest on the illusion that there can be one objective assessment of the world. This is not only cosmology falsely named ontology, but also metaphysics disguised as physics.

On top of that, it is relying on the illusion of a culture where semantic tags wouldn't simply be used for spamming and search engine manipulation, which is why Google already ignores metatags. And while Berners-Lee is a realist enough to assess that tagging cannot be done by bots like those planned by the Theseus project, his Semantic

[3] URL

Web consequently implies a complexity nightmare of meta information overtaking information, i.e. where each piece of information would create at least twice as much work for its semantic markup than for its original creation.

It would be good if creators of so-called next-generation search engines would read up on Borges who concludes:

> "I have registered the arbitrarities of Wilkins, [and] of the unknown (or false) Chinese encyclopaedia writer [...]; it is clear that there is no classification of the Universe not being arbitrary and full of conjectures. The reason for this is very simple: we do not know what thing the universe is."