

The Pinball Machine and the Cloud: Using Rules to move from Information to Knowledge

Gian Piero Pescarmona¹, Ferruccio Zamuner², Elena Giglia³

¹ Dipartimento di Genetica, Biologia e Biochimica, Università di Torino
Via Santena 5 bis, 10126 Torino (TO), gianpiero.pescarmona@unito.it

² NonSoLoSoft di Ferruccio Zamuner Via G. Di Vittorio 18, 10023
Chieri (TO), nonsolosoft@diff.org

³ Sistema Bibliotecario di Ateneo, Università di Torino
elena.giglia@unito.it

Abstract. The "Il Flipper e la Nuvola" project has been created within the course of Clinical Biochemistry at the Faculty of Medicine of the University of Turin. Structured in Reports, Rules, Items, Pathways and Tools (associated with the tag of the Web 2.0 or the PubMed MeSH terms), it allows students to create new routes from time to time according to different knowledge requirements. The critical points are then: 1) objectives 2) the criteria for the selection and the quality of information 3) the possibility of the use of information. Different objectives, different rules, and choice of different information. The explicit declaration of the rules and their falsifiability allows continuous adaptation of the rules to the latest issues, making the system evolutionary and adaptive. The methodological approach is context independent and manageable, as the rules, if properly chosen, don't need to be modified at the same rate of the Web information burden growth.

Keywords: Information, knowledge, objectives, rules, falsifiability, web 2.0, tag, e-learning, user experience

1 Introduction

The current issue of the development of knowledge often is not the scarcity but the overload of information in all its forms (books, Web resources, word of mouth, social networking, Encyclopedias etc). The information overload leads, in the absence of strong selection criteria, to an uncertainty feeling exiting either in random choices or in development of a passive attitude towards the experts view.

The passive acceptance of the experts view presents many drawbacks: 1) selection of information on the basis of the objectives of knowledge of the experts and not of the users 2) the possibility of conflict of interest 3) partiality of cataloging

information; the choices (in light of objectives not always explicit, often inherited from the past) of cataloguers may be optimal for a group of users and not for others.

The information landscape mainly includes two groups of players: information users and information suppliers. Information users are men/women with their wishes and targets, and their own language. Information suppliers range from Educational Institutions like University to sellers of any kind of item, to Groups of Social networking and each of them developed a subset of rules and a language.

Men's wishes are numberless and numberless the information available on the Web. How to match them? How to choose among them? On the basis of experts opinion? Or on the basis of users consensus? Are the knowledge targets really shared between experts and users? Between physicians and patients for example?

1.1 The Information

Information is no longer what it used to be. In the past raw data and their interpretation were supplied together by the authority. The Bible, the Greek philosophers for centuries have supplied the western culture with a set of information not to be discussed. The Textbooks manage to do the same: they express clear cut concepts and avoid the fuzzy one and contradictions. They support the idea that it is always possible to solve a problem, provided you study carefully the pertinent textbook.

Other form of information existed (like word of mouth, more reliable than the written texts in many cases) but they had a completely different way of diffusion.

The web now offers to everybody the possibility to publish his/her opinion, free of charge. Scholar papers and crazy chats containing the same key word can be extracted from the billions pages of the Web with the same relevance. No control can exist on the publishing activity, filters can be applied to the activity of the search engines.

Web cataloguing is even more difficult than books cataloguing, and librarians are still disputing on the topic. The Semantic Web is an evolving extension of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the Web as a universal medium for data, information, and knowledge exchange. That means that notations and formal specifications have to be introduced by experts, all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain.

A project such as Wikipedia, which relies on the creation and control of the global population of readers-writers, is very positive, but the results obtained are extremely variable. Even if as a whole a comparison between Wikipedia and Encyclopaedia Britannica online demonstrated that their accuracy is similar [1], some topics are heavily underdeveloped and their evolution is unpredictable.

When the knowledge domain is very specific it is relatively easier to build up data repositories like in the case of genetic data, DNA structure, genes and related diseases and so on as collected for example in the Entrez retrieval system, powered by NCBI, a service of the U.S. National Library of Medicine and the National Institutes of Health [2]. But diseases classification is much more uncertain; in some cases like

autoimmune diseases, where nobody exactly knows the mechanism of the disease, classification is just a descriptive list of symptoms [3].

In a user centered world patients not satisfied by the scholar approach to their disease often create blogs or self-help groups where they exchange their experience, asking questions, trying to supply answers. Similar experiences are growing up as far as the drugs side effects are concerned. Up to now the patient complaining side or unwanted drug effect had to tell it the pharmacist or the physician, who should fill a form and send it to the National Health Service. The vast majority of the side effect is lost as can be expected. But now some local Health Services are predisposing an online form for self declaration of side effects. This practice has been widely applied in the past by the US Women associations that have been able to detect the drugs side effects in weeks instead of years.

1.2 The Knowledge

“Knowledge” is a multipurpose word like “mother”: anyone has his one, not comparable to the others. We already discussed the information classification problems and I don’t intend to afford them again. Let’s try to define knowledge operationally as the set of information I need to modify the reality according to my wishes. Suppose you succeeded in getting a date in fifteen days with the supposed man/woman of your life. You intend to cook a dinner by yourself: a pasta with the sauce you got in Vulcano last summer, the wine your grandfather used to produce when you still lived in a small country village, the orange flower bouquet pervading the evening like in his/her childhood in Sicily. Dreams or wishes? It depends on how much experts in findability were the Web providers of the objects you need to fulfil your requirements.

In short, knowledge is required only if you have targets, might you reach them or not. Knowledge targets therefore have to be verifiable; if I succeed in getting the sauce, the wine, the scent that’s to say that I knew how to manage to get the right information to modify the reality according to my wishes.

Unfortunately not all wishes are as simple as to buy a low cost fly ticket online, sometimes I would like to know the causes of most frequent and life threatening diseases with the hope of reasonably prevent or treat them.

But the structure of knowledge is always the same:

1. Wishes, dreams, targets
2. Collection of information
3. Use of information to modify the reality

As long as you will be able to get what you want, when you want, your knowledge will be up to date.

2 The Pinball Machine and the Cloud

“Il Flipper e la Nuvola” (The Pinball Machine and the Cloud) [4], instance of “Arancia” [5], is a Web application whose aim is an easier identification of the molecular basis of the diseases. It has been used since 2007 in a University medical class [6].

It is a Web application based on the Web 2.0 logic, which implies also a different attitude towards scholarly communication. It is structured in Reports, Rules, Items, Pathways and Tools referring and linking one another. The use of tags and/or controlled PubMed MeSH terms to categorize allows and fosters a free and personal use of information to create original knowledge.

Users can follow and open innovative paths each time answering a different question, re-combining the existing information. This is the richness added by the users: exploring a tag and its related material can lead to an unexpected point of view on the same symptom, or can change one’s perspective on a disease, or a clinical case. Change of view also means new targets, expectations or wishes: who is performing in my area that specific genetic or blood test? Where are hidden the local epidemiological data I cannot easily access to in my region while I can get everything about North Carolina so easily? The feeling of patients and physicians for the local Health Service is dramatically changed by these kinds of comparisons. «Think Globally, Act Locally» is more and more true.

The reality is represented by Reports, descriptions of clinical cases whose fate can be changed by a correct interpretation of symptoms, allowing a validation of the method. Each user edits with a simplified Wiki writing language its own Report and links it with the fitting involved Item or Pathway, and then can tag it or associate it to a MeSH term. That creates a tag cloud which allows unprecedented links and a critical reuse of the content. Comments are always possible. That shapes a multi-sided scholarly communication, far away from the traditional one-way descending pattern, both in vertical – teacher/learner - and peer to peer – learner/learner.

The Items and the Pathways mark out the framework of this innovative channel of communication, and the user generated content – dealing with diseases, drugs, proteins, metabolic paths and so on – consists of texts, images, links to scientific literature, links to biomedical websites, in a creative and critical approach as learned during classes. Tag clouds also apply to the most linked and handled Web sites, generating a sort of shared validation. The easiness and readiness both in submitting and in searching and retrieving the content creates such a participative environment that the user experience really results enriched.

The information, or better, an interpreted gateway to the information is collected in Tools, Items and Pathways, where the link to the contents is categorized with an indexing visually very similar to classical textbooks, but structurally based on a relational database and easily modifiable if needed. The Database is also searchable with the Google search engine allowing a search by argument independently from the type of indexing. Indexing itself carries a lot of information as different branches of learning usually aggregate differently the same set of contents.

3 The Rules

The most striking feature of the site is the chapter Rules, where the conceptual frame (the ontology) of the living organisms in health and disease is accurately described. Medical textbooks usually analytically describe specific symptoms, treatments, surgery by organs or class of diseases (infectious, cancer etc). But a robust definition of the disease itself is lacking, although it could help a lot anytime complexity arises in the reality the physician has to face: patients with multiple diseases, borderline symptoms involving more organs and so on.

Apart its definition the knowledge of the disease should include a thorough knowledge of the physiological mechanisms (health) underlying the correct functioning of the body. But often Medicine doesn't care of processes that work properly and don't fail (disease), but as a matter of fact the disease is a rare event if compared to the whole of the working organisms. The ontology of the living organisms has to include the concept that natural selection for billions years has been active in selecting the fittest behavior for the local environment and that many diseases may depend on the too fast changes of the environment, imposed by human activities.

The main property of the Rules described in this chapter is that they are revisable, in the sense that they are defined in a way they can be tested and eventually falsified so to be always valid in the context, according to Karl Popper [7]. At the very beginning they can be established by experts, but they will survive only unless not demonstrated false by users. The evolution of the rules is very similar to the biological evolution of molecules or whole organisms. Only the fittest to the environment survives; but if the environment changes also the pattern of survivors changes. As the pattern of information available change, also rules and methods of data mining have to change. Only the rules tested everyday are good rules and only if they are modified whenever they fail.

According to the same perspective – and to another Web 2.0 suggestion, «trust your users» - the concept of “reliable information” itself changes. Not the whole available information has to be validated: just the criteria to accept or reject them have to, depending on the target or the specific query. No doubt that information about adverse effects of a drug might be otherwise assessed by a patient association or a pharmaceutical company.

The underpinned logic is: different queries have different matching information sources. No expert can tell “which” source, the only valid criterion being the user's need and target. Web cataloguing and sites classification projects may help and avoid waste of time, but ultimately any target may require a specific pattern of search strategies to retrieve the information useful in a specific ontology.

4 Conclusions

After two years practice with “Il Flipper e la Nuvola” some conclusions can be drawn.

What the information users really need are data, not opinions or advertising. “Data” stands for the train timetable, the hospital FirstAid telephone number, the cancer epidemiology in my area, the access to gene-banks or to the drugs list with their commercial names, active principle, side effects. They should be correctly readable from the user and tagged to be retrieved and reused by all potential readers according to their needs. A faceted classification system that allows for dynamic categorization and taxonomy seems the best technique [8]. The best approach to define facets by far is a marriage of folksonomy and taxonomy, to combine users needs and the specific terminology of a discipline, as assessed by experts. Constructing multifaceted hierarchies from a large collection of databases, text or text-annotated objects may lead to too large hierarchies. The “Rules” are useful tools to impose constraints to the retrievable information.

The attainment of the target (to sell more wine bottles or to understand the causes of your patient symptoms) is the only way to evaluate a web application. The opinions driving your behavior (the Rules) must be very simple and always applied to select information, until they become useless and have to be modified. On the whole, accessing the data but filtering the information from the Web on the basis of each personal targets and rules seems a sound strategy for the access to knowledge and the development of new wishes and targets [9], keeping in mind that “Real human knowledge is, by nature, inconsistent, ill-defined, and unstructured” [8].

References

1. Giles J: Internet encyclopaedias go head to head. *Nature* 438: 900--901 (2005)
2. National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov/sites/entrez>
3. Autoimmune Disease on Wikipedia, http://en.wikipedia.org/wiki/Autoimmune_disease
4. Il Flipper e la Nuvola, <http://flipper.diff.org/>
5. Pescarmona GP, Zamuner F.: Arancia: l'e-learning 2.0 in Didamatica 2008, Taranto (2008), http://flipper.diff.org/static/files/1002/Arancia_Pescarmona_Zamuner.rtf
6. Mangioni M., Doctoral Thesis: Il Web 2.0 nell'educazione medico-scientifica. "Il Flipper e la Nuvola": un esempio applicativo nel campo della Biochimica Clinica (2008), http://flipper.diff.org/static/files/814/tesi_monica_mangioni.pdf
7. Popper K., *Congetture e confutazioni*, Il Mulino, Bologna, (1972)
8. Louie A.J., Maddox E.L., Washington W.: Using Faceted Classification To Provide Structure For Information Architecture, 2003 IA Summit, Portland, Oregon, http://depts.washington.edu/pett/presentations/conf_2003/IASummit.pdf
9. McLean R., Richards B.H. and Wardman J.I.: The effect of Web 2.0 on the future of medical practice and education: Darwikinian evolution or folksonomic revolution? *MJA* 187, 174–177 (2007).