A Medical Informatics Perspective on Health Informatics 3.0

Findings from the Yearbook 2011 Section on Health Informatics 3.0

P. Ruch, Section Editor for the IMIA Yearbook Section on Health Informatics 3.0
University of Applied Sciences Geneva, Dept. of Information and Library Sciences, Geneva, Switzerland

Summary

Objectives: To summarize current advances of the so-called Web 3.0 and emerging trends of the semantic web.

Methods: We provide a synopsis of the articles selected for the IMIA Yearbook 2011, from which we attempt to derive a synthetic overview of the today's and future activities in the field.

Results: while the state of the research in the field is illustrated by a set of fairly heterogeneous studies, it is possible to identify significant clusters. While the most salient challenge and obsessional target of the semantic web remains its ambition to simply interconnect all available information, it is interesting to observe the developments of complementary research fields such as information sciences and text analytics. The combined expression power and virtually unlimited data aggregation skills of Web 3.0 technologies make it a disruptive instrument to discover new biomedical knowledge. In parallel, such an unprecedented situation creates new threats for patients participating in large-scale genetic studies as Wjst demonstrate how various data set can be coupled to re-identify anonymous genetic information.

Conclusions: The best paper selection of articles on decision support shows examples of excellent research on methods concerning original development of core semantic web techniques as well as transdisciplinary achievements as exemplified with literature-based analytics. This selected set of scientific investigations also demonstrates the needs for computerized applications to transform the biomedical data overflow into more operational clinical knowledge with potential threats for confidentiality directly associated with such advances. Altogether these papers support the idea that more elaborated computer tools, likely to combine heterogeneous text and data contents should soon emerge for the benefit of both experimentalists and hopefully clinicians.

Keywords

Medical informatics, International Medical Informatics Association, yearbook, semantic web

Yearb Med Inform 2011: 30-2

Introduction

Out of the five selected papers this year, it is worth observing that two of them propose to articulate together semantic web challenges and text analytics. Interestingly, such an a posteriori obvious complementarity was clearly not seen at all by pioneers of the semantic web. Access to literature and Text mining are seen as both enabling technologies for the semantic web, as well as first beneficiary of semantic web technologies. Thus, Cheung et al. do not hesitate to base today's biomedical semantic interoperability challenges on top of these two pillars. Somehow related to the same pillars (i.e. connecting structured triples with unstructured textual contents), Samwald and Stenzhorn propose to combine RDF contents as available in structured biomedical databases with short „evidence-based" passages extracted from the literature. The combined information item would augment structured data with source evidences as found in scientific articles. Further, two papers are dealing with core semantic web priorities, which are naming conventions and access algebra for semantic web endpoints. Thus, Shaw proposes a language definition model for the semantic web, while Patterson et al. propose to federate the whole biological web around a subset of entities such as organisms using universal nomenclatures of species. The fifth paper can ultimately be regarded as illustrating an unfortunate side-effect of the power of the semantic web – hypothetically further amplified by text analytics. Wjst shows how fully de-identified genetic data samples can be potentially re-identified with the risk of revealing confidential clinical information and outcomes for patients participating in genetic studies.

A brief content summary of the selected best papers can be found in the appendix of this report.

Conclusions and Outlook

The best paper selection for the Yearbook section on future trends of the semantic web can by no means reflect the high level of activity that we observe in this new and emerging field. More than other fields, the semantic web is a work area with intrinsically heterogeneous subfields. The author of this section is quite confident that the semantic web should find its way to a „normal science" (Kuhn 1962). Whether such a step is already passed or not is difficult to claim, as some distance is needed to estimate the stability of the new paradigm, in particular if we hypothesize that Kuhn's epistemological model do not apply anymore to entities such as *omics* sciences. However, it seems today clear that semantic web researches, in particular when combined with text analytics models, can radically renewed the field of information sciences (Hjørland 1992). At least one out of the five selected papers shed light on some special aspects deserving particular attention as they concern key methodological questions for the future of the

field. The aggregation power of the semantic web questions open data sharing models, which are at the very foundation of the methodology of scientific experiments as experimental settings must be reproducible. Such a radical question must be explored by researchers, including research clinicians, whose professional activities, workflows and methodologies could directly be affected.

Acknowledgement
I greatly acknowledge the support of Martina Hutter and of the reviewers in the selection process of the IMIA Yearbook.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2011 in the section 'Health Informatics 3.0'. The articles are listed in alphabetical order of the first author's surname.

Section
Health Informatics 3.0
▪ Cheung KH, Samwald M, Auerbach RK, Gerstein MB. Structured digital tables on the Semantic Web: toward a structured digital literature. Mol Syst Biol 2010 Aug 24;6:403.
▪ Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP. Names are key to the big new biology. Trends Ecol Evol 2010 Dec;25(12):686-91.
▪ Samwald M, Chen H, Ruttenberg A, Lim E, Marenco L, Miller P, Shepherd G, Cheung KH. Semantic SenseLab: Implementing the vision of the Semantic Web in neuroscience. Artif Intell Med 2010 Jan;48(1):21-8.
▪ Shaw M, Detwiler LT, Noy N, Brinkley J, Suciu D. vSPARQL: a view definition language for the semantic web. J Biomed Inform 2011 Feb;44(1):102-17.
▪ Wjst M. Caught you: Threats to confidentiality due to the public release of large-scale genetic data sets. BMC Med Ethics 2010 Dec 29;11:21.

References

1. Hjørland, Birger. Epistemology and the Socio-Cognitive Perspective in Information Science. Journal of the American Society for Information Science and Technology 2002;53(4):257-70.
2. Kuhn, Thomas S. The Structure of Scientific Revolutions. Chicago, IL: University of Chicago Press; 1962, 1970, 1996.

Correspondence to:
Prof. Dr. Patrick Ruch
University of Applied Sciences Geneva
Department of Library and Information Sciences
Geneva, Switzerland
Tel: +41 22 388 17 81
E-mail: patrick.ruch@hesge.ch

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2011, Health Informatics 3.0*

Cheung KH, Samwald M, Auerbach RK, Gerstein MB

Structured digital tables on the Semantic Web: toward a structured digital literature

Mol Syst Biol 2010 Aug 24;6:403

The growth of data in biomedical databases has increased in direct relationship with the growth of published literature. The extraction if high-quality information out of digital libraries documents remains challenging although text-mining developments. To address this challenge, the authors suggest the creation of a structured digital table as part of an overall effort in developing machine-readable, structured digital literature. Those large tables could be then stored in standardized triples stores using Semantic Web infrastructures. Three main types of tables are proposed: 1. Tables conveying information about properties; tables storing information about networks, and finally tables for concept hierarchies. The authors demonstrate show how more complex tables can be constructed from these three basic types. The authors propose that authors could create tables initially using the structured triples for canonical types. At publication-time, the triples could be visualized. Further, the authors present examples for transforming representative tables into triples. Finally, a discussion is initiated in an attempt to automatically link literature as stored in digital libraries and structured biomedical data repositories to improve semantic interoperability between these complementary sources of knowledge.

Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP.

Names are key to the big new biology

Trends Ecol Evol 2010 Dec;25(12):686-91

The authors present the Global Names Architecture. An infrastructure based on semantic web technologies to answers big and broad questions encompassing entities related to taxonomy, evolution and ecology. The emerging name-based infrastructure builds on the availability of comprehensive phylogenetic trees and organisms terminologies to build a hub of interconnected contents over the World Wide Web. The proposed virtual infrastructure should further expand as more data are shared; thus unifying biology into a homogeneous *big science*. Further, computer artefacts likely to exploit this new resource are likely to renew life sciences' perspectives by making explicit currently unseen associations and trends in order to support new scientific discoveries.

Samwald M, Chen H, Ruttenberg A, Lim E, Marenco L, Miller P, Shepherd G, Cheung KH

Semantic SenseLab: Implementing the vision of the Semantic Web in neuroscience

Artif Intell Med 2010 Jan;48(1):21-8

Information technology can potentially speed up scientific discovery by help-

* The complete papers can be accessed in the Yearbook's full electronic version, provided that permission has been granted by the copyright holder(s).

ing accessing and generating information for researchers. In such a context, several projects attempt to use semantic web technologies for such a purpose. Nevertheless, successful applications in that domain, which have been able to reach the end-users, remain marginal. The aTag ('associative tags') framework that is described in this paper aims at facilitating the penetration of semantic web technologies. The aTags resource consists in short snippets of HTML+RDF with embedded RDF/OWL. The overall architecture is based on the Semantically Interlinked Online Communities (SIOC) vocabulary, as well as domain ontologies and taxonomies, such as the Open Biomedical Ontologies and DBpedia. aTags has a very simple: a short piece of human-readable text that is associated with a set of relevant ontological entities. This paper reports on efforts to develop a set of software and data resources around aTags. Some of the prototypes results are available at http://hcls.deri.org/atag. Conclusion: the aTags convention can help the rapid delivery of diverse, integrated datasets and semantically interoperable contents and services. The usability and scalability of the overall approach need to be tested in more use-case scenarios; therefore the adoption of the convention by other groups is encouraged.

Shaw M, Detwiler LT, Noy N, Brinkley J, Suciu D

vSPARQL: a view definition language for the semantic web

J Biomed Inform 2011 Feb;44(1):102-17

The authors present the vSPARQL system, a view definition language for triple stores. The innovative solution serve as basis to help designin translational medicine applications using biological and biomedical ontologies, terminologies, controlled-vocabularies, as well as various data sets available on the semantic web. The proposed language borrows from relational database views to propose a new algebra for RDF datawarehouses. vSPARQL, allows applications to generate unambiguous contents regarding definition, structure and lifecycle. The stored content can then be accessed using vSPARQL. The expression power of the proposed view definition language is assessed using a set of practical use cases that the authors also compare with existing query languages.

Wjst M

Caught you: Threats to confidentiality due to the public release of large-scale genetic data sets

BMC Med Ethics 2010 Dec 29;11:21

Genetic data are becoming more and more commonly used and shared between research groups. Some of the datasets are also made available on the internet for secondary experiments or as supplementary materials to support published reports. Patients, whose data are used, are commonly not informed because it is generally assumed that such data are anonymous, when nominal data are deleted. However, such an assumption is questionable because genetic data are per se self-identifying. In particular, there exist at least two re-identification methods: the "Netflix" and the "profiling" method. Netflix requires a small set of 100 single nucleotide polymorphisms (SNPs) together with a personal identifier. This second source of data can be obtained from a different clinical source, such as forensic tests. Once associated with the small set of genetic data, it then becomes possible to re-identify all samples of the patient. More critically, the author shows that even when no personally-identified data is available, it is possible to profile a sample collection in order to extract potentially de-identifying information. It is thus possible to start identifying information such as ethnic subgroups, various body characteristics, including predictions of pathologies. It is concluded that there is a relatively good chance that at least a few individuals can be identified from an anonymized genetic data set. Any de-anonymization initiative of genetic data could potentially threats integrity of participants as such an effort is likely to release potentially sensitive data such as disease-related risks.