# Knowledge Representation and Management: Towards Patient Health Self-management

**A.-M. Rassinoux, Section Editor for the IMIA Yearbook Section on Knowledge Representation and Management**
Information Systems Division, Geneva University Hospitals, Geneva, Switzerland

## Summary

***Objectives:*** To summarize excellent current research in the field of knowledge representation and management (KRM).

***Method:*** A synopsis of the articles selected for the IMIA Yearbook 2012 is provided and an attempt to highlight the current trends in the field of health management is sketched.

***Results***: Among the five selected papers, two confirm the benefit of exploiting open-source language toolkits for the automatic extraction of medical concepts, assertions and/or relationships from clinical texts. One paper aims at exploiting domain-specific terminologies to improve the parsing of biomedical noun phrases, and another one aims at discovering rare diseases associations embedded into disparate textual sources. Finally, the last paper describes a collaborative search approach integrated into a homegrown EHR search engine.

***Conclusions***: This selected set of papers confirms that natural language processing, as well as knowledge extraction, discovering and retrieval, are still active and fruitful research fields. Although these papers are not directly focusing on personal health informatics applications,, important features are highlighted and tailored to fit the requirements of patient health self-management. Delivering timely, friendly and secure access to functional, accurate, up-to-date and sustainable personal health records is a significant challenging task for supporting self-managed healthcare.

## Keywords

Open-source language toolkits; Knowledge extraction, discovering and retrieval; Health self-management; Health behavior.

**Yearb Med Inform 2012:126-9**

## Introduction

"Personal Health Informatics" is the theme of the 2012 Yearbook. These last years, personal health informatics applications, based on health self-management principles, show promise for allowing individuals to manage their health and medical conditions. This is especially challenging for patients living with chronic illness. They are even more motivated to improve their quality of life and health outcomes by making "informed decisions" [1]. These are reasoned decisions jointly shared between clinicians and patients, about the right actions or interventions to be carried out in accord with the individual's beliefs. This collaborative health behavior is also enhanced by the growing development of interactive communication tools such as web-based and mobile-based applications [2].

Assisting patients in health self-management requires the use of electronic personal health record (PHR) systems [3, 4]. A PHR can be seen as an extension of the physician electronic health record (EHR) [5] where the individual patient can manage, share and control its healthcare data. For this, it is necessary to incorporate new technologies that appropriately empower patients or any consumers. In particular, the following features must be taken into account when designing personal health informatics applications: the definition of user requirements, the adoption factors for PHR use, the process of acquiring consistent and accurate healthcare information, the broad and timely access to relevant personal health information resources, the authentication and authorization processes that guarantee individual privacy and security of health information, as well as a streamline communication with care providers, to name a few.

## Best Paper Selection

For the section Knowledge Representation and Management (KRM) of the IMIA Yearbook 2012, five papers were selected, following a comprehensive review process. These elected articles, listed in Table 1, are summarized in the appendix of this synopsis. Although these papers are not directly focusing on personal health informatics applications, some key points are highlighted and argued to address the specific requirements of patient health self-management.

Two papers [6, 7], among the five selected, reported their participation to the 2010 fourth i2b2/VA Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data [8]. Three tasks were identified on clinical texts [9]: 1) extraction of medical concepts including medical problems, treatments and tests; 2) classification of assertions made on medical problems; 3) classification of relationships between medical concepts. Focusing on the first task, D'Avolio et al [6] describe a tooling chain, integrating open-source frameworks such as clinical natural language processing (NLP) systems and machine learning classifiers, to automatically extract concept-level information. Besides, Islamaj Doğan et al [7] have built an end-to-end system that focuses on identifying clinical relationships in patient records while addressing also the first two tasks of the 4th i2b2 challenge. These two papers emphasize the use of language toolkits [10, 11], embedding scalable, reusable and robust NLP components, that allow information-retrieval pipelines to be rapidly developed, evaluated and deployed. This contributes to minimizing the burden on system developers and end users. This strategy was already emphasized in the KRM synopses of the previous yearbooks [12, 13]. It is also worth noting that clinical assertions [14], whose detection was the aim of the second i2b2 challenge task, as well as temporal relationships [15] are critical attributes when considering variables in health self-management. Indeed, these features guarantee that the information is accurate and up-to-date and therefore exploitable for further decision making.

The other three articles [16, 17, 18] depict fairly heterogeneous studies focusing on active research fields such as NLP, knowledge discovering and search engine. Incorporating terminology semantics in the building of an operational grammar was the goal of the study conducted by Fan et al [16]. The results reported an improved performance for the parsing of noun phrases that are commonly used in biomedical reports. Besides, Holmes et al [17] describe a system for discovering rare disease associations that are embedded into diverse textual sources, such as EHRs, journal articles or web sites. These heterogeneous sources convey different levels of expressivity, specificity and complexity of medical language. However, to be accessible by a large population, including laypersons, an easy-to-understand language must be adopted for expressing functional information stored in PHRs. Finally, the interactive communication era, promoted by the growing use of email, SMS/MMS messages, mobile applications and social networks, is also influencing health behavior [19]. The development of computerized systems, that ease clinical workflows through convivial user interface, should enhance user participation as depicted in the collaborative search approach chosen by Zheng et al [18]. Indeed, they have successfully developed a friendly EHR search engine, based on search-terms bundles that are shareable and reusable across users.

**Table 1** Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2010 in the section 'Knowledge Representation and Management'. The articles are listed in alphabetical order of the first author's surname.

Section
Knowledge Representation and Management

- D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD. Automated concept-level information extraction to reduce the need for custom software and rules development. J Am Med Inform Assoc 2011 Sep-Oct; 18(5): 607-13.
- Fan JW, Friedman C. Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies. J Biomed Inform 2011 Oct;44(5):805-14.
- Holmes AB, Hawson A, Liu F, Friedman C, Khiabanian H, Rabadan R. Discovering disease associations by integrating electronic clinical data and medical literature. PLoS One 2011;6(6):e21132.
- Islamaj Doğan R, Névéol A, Lu Z. A context-blocks model for identifying clinical relationships in patient records. BMC Bioinformatics 2011 Jun 9;12 Suppl 3:S3.
- Zheng K, Mei Q, Hanauer DA. Collaborative search in electronic health records. J Am Med Inform Assoc 2011 May;18(3):282-91.

## Conclusion and Outlook

The final five articles, selected for the KRM section of the IMIA Yearbook 2012, corroborate the significant and continuous efforts on exploiting open-source language frameworks and/or available sources of clinical information, such as terminologies or the rich content of EHRs, to analyze, discover, extract or search for relevant pieces of clinical information. While the meaningful use of EHRs within and among care providers is acquired, moving beyond the view of traditional medical record to focus on the consumer's needs is one of the most challenging goals of the emerging PHR. However, a lot has still to be achieved in tailoring operational personal records that would help both healthcare providers and patients make health-related decisions. Issues such as the quality, timely access and individual privacy of healthcare data, must be clarified. Moreover, how PHRs should be designed and who should manage them remain open questions. Finally, the fact that individuals take an active role in their healthcare should motivate positive health behavior change.

### Acknowledgement

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

## References

1. Bekker H, Thornton JG, Airey CM, Connelly JB, Hewison J et al. Informed decision making: an annotated bibliography and systematic review. Health Technol Assess 1999;3(1):1-156.
2. Chi EH. Augmented social cognition: Using social web technology to enhance the ability of groups to remember, think, and reason. Proceedings of the 35th SIGMOD international Conference on Management of Data (SIGMOD '09). New York: ACM Press; 2009:973-84.
3. Kaelber DC, Jha AK, Johnston D, Middleton B, Bates DW. A research agenda for personal health records (PHRs). J Am Med Inform Assoc 2008 Nov-Dec;15(6):729-36.
4. Archer N, Fevrier-Thomas U, Lokker C, McKibbon KA, Straus SE. Personal health records: a scoping review. J Am Med Inform Assoc 2011 Jul-Aug;18(4):515-22.
5. Hoerbst A, Ammenwerth E. Electronic health records. A systematic review on quality requirements. Methods Inf Med 2010;49(4):320-36.
6. D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD. Automated concept-level information extraction to reduce the need for custom software and rules development. J Am Med Inform Assoc 2011 Sep-Oct;18(5) 607-13.
7. Islamaj Doğan R, Névéol A, Lu Z. A context-blocks model for identifying clinical relationships in patient records. BMC Bioinformatics 2011;12 Suppl 3:S3.
8. I2B2: Informatics for integrating Biology & the Bedside; Available from: http://www.i2b2.org/NLP/Relations/Main.php.
9. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011 Sep-Oct;18(5):552-6.
10. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17:507-13.
11. McCallum AK. MALLET: A Machine Learning for Language Toolkit. 2002: [http://mallet.cs.umass.edu].
12. Rassinoux AM. Decision Support, Knowledge Representation and Management: Transforming Textual Information into Useful Knowledge. In: Kulikowski CA, Geissbuhler A, editors. Yearb Med Inform 2010:64-7.

13. Rassinoux AM. Knowledge Representation and Management: Benefits and Challenges of the Semantic Web for the Fields of KRM and NLP. In: Kulikowski CA, Geissbuhler A, editors. Yearb Med Inform 2011 121-4.
14. Clark C, Aberdeen J, Coarr M, Tresner-Kirsch D, Wellner B et al. MITRE system for clinical assertion status classification. J Am Med Inform Assoc 2011 Sep-Oct;18(5):563-7.
15. Savova G, Bethard S, Styler W, et al. Towards temporal relation discovery from the clinical narrative. AMIA Annu Symp Proc 2009;2009:568-72.
16. Fan JW, Friedman C. Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies. J Biomed Inform 2011 Oct;44(5):805-14.
17. Holmes AB, Hawson A, Liu F, Friedman C, Khiabanian H, Rabadan R. Discovering disease associations by integrating electronic clinical data and medical literature. PLoS One 2011;6(6):e21132.
18. Zheng K, Mei Q, Hanauer DA. Collaborative search in electronic health records. J Am Med Inform Assoc 2011 May;18(3):282-91.
19. Lau AY, Siek KA, Fernandez-Luque L, Tange H, Chhanabhai P et al. The Role of Social Media for Patients and Consumer Health. Contribution of the IMIA Consumer Health Informatics Working Group. Yearb Med Inform 2011:131-8.

Correspondence to:
Anne-Marie Rassinoux, Ph. D.
University Hospitals of Geneva
Information Systems Division
4, Rue Gabrielle-Perret-Gentil
CH-1211 Geneva 14, Switzerland
Tel: +41 22 372 6293
Fax: +41 22 372 8680
E-mail: anne-marie.rassinoux@hcuge.ch

## Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2012, Section Knowledge Representation and Management*

### D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD

### Automated concept-level information extraction to reduce the need for custom software and rules development

### J Am Med Inform Assoc 2011;18(5):607-13

This paper demonstrates that it is possible to deliver acceptable conceptual information extraction performance across various tasks with no custom software or rules development. Indeed, the burden on system developers and end users can be reduced thanks to the widespread adoption of open-source frameworks.

In this study, the automated retrieval console, called ARC, is used by non-technical end users to configure, conduct and review experiments on information-retrieval pipelines. First, the open-source clinical NLP pipeline, cTAKES, allows free text to be mapped over 90 different data types. Then, these data types and their associated values are used as features in an open-source supervised machine learning classifier that implements Conditional Random Fields (CRFs). Two algorithms for selecting appropriate feature types and their related values were assessed. The first algorithm, called 'Blast' algorithm, evaluates all combinations of the top five individual performing features. The second one, called 'All Scoring', calculates the performance of each individual feature, using 10-fold cross-validation. These two algorithms were evaluated using data sets and metrics available from the concept extraction portion of the fourth i2b2 challenge.

The algorithms explored as part of this study achieved F-measures that were competitive with the average performance of i2b2 challenge competitors without requiring tasks' customization. However, the authors conclude that no single feature type or unit of analysis is ideal for all retrieval tasks.

### Fan JW, Friedman C

### Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies

### J Biomed Inform 2011 Oct;44(5):805-14

Biomedical natural language processing (BioNLP) has seen an increased research interest in dealing with domain specificity and variation. Indeed, a robust BioNLP system should be able to cope with the huge lexical space of professional terms as well as to handle the grammatical characteristics of the biomedical language (i.e. prepositional phrases and conjunctions). The addition of semantic information has been shown to be necessary to resolve structural ambiguities and is commonly performed through a rule-based or a probabilistic grammar.

In order to reduce the human effort in grammar engineering and to yield a semantically transferable grammar to the biomedical community, the authors propose an automated method for deriving a probabilistic grammar based on a training corpus consisting of concept strings and semantic classes from the Unified Medical Language System (UMLS). Due to the nominal nature of the majority of biomedical terminological concepts, the designed probabilistic context free grammar (PCFG) strictly focuses on parsing noun phrases in biomedical texts. The generation of the semantic PCFG was performed in five steps. After selecting the strings for training noun phrase patterns, a POS-tagging and parsing were performed. Then, the parsed training strings were labeled with semantic types and the grammar rules were extracted from the semantically augmented parses. Finally, the probabilities of the rules were computed.

The resulting PCFG embeds more than five million rules with associated probabilities. Its evaluation on manually parsed clinical notes has permitted to highlight shortcomings that could be further addressed to improve performance. However, the precision of 0.644, recall of 0.737 and average cross-bracketing of 0.61 revealed better performance compared to a semantic-null control grammar. These outcomes promote the integration of terminology semantics into PCFG for the parsing of biomedical noun phrases.

### Holmes AB, Hawson A, Liu F, Friedman C, Khiabanian H, Rabadan R

### Discovering disease associations by integrating electronic clinical data and medical literature

### PLoS One 2011;6(6):e21132

Identification of rare disease comorbidities is not straightforward due to the small numbers of diagnosed patients. With its large sets of patients' longitudinal medical information, including coded diagnosis as well as textual discharge

* The complete papers can be accessed in the Yearbook's full electronic version, provided that the article is freely accesible or that your institution has access to the respective journal.

summaries, electronic health record (EHR) systems constitute a rich and reliable source for studying many diseases and their medical associations within a population. Besides, biological journal articles as archived in PubMed and web sites such as Wikipedia can also be exploited as they aggregate relevant information related to disease associations from around the world.

All these aforementioned disparate sources are exploited by ADAMS, an Application for Discovering Disease Associations using Multiple Sources. The authors apply ADAMS for investigating the co-morbidities associated with three rare diseases: Kaposi sarcoma, toxoplasmosis and Kawasaki disease. In this study, ADAMS is used to compare case and control disease cohorts within the New York-Presbyterian Hospital's EHR. In order to reduce the biases, inherent to the information present in the NYPH EHR, two control cohorts with different etiologies (post-traumatic stress and influenza) are chosen. Only ICD-9-CM coded data are considered by ADAMS for finding statistically significant positive associations. They are first generated within the case cohort and then organized into a visual network diagram that draws links between common associations among these terms. Then, ADAMS uses a fuzzy matching procedure to compare terms in the various considered data sources.

In a large discussion, the authors bring to light the ability of ADAMS for identifying both known relationships between the three rare diseases and other medical conditions that have already been reported in PubMed or Wikipedia, and novel associations with limited or no reporting in the existing literature. In particular, the authors report a statistically significant association between Kawasaki disease and diagnosis of autistic disorder.

### Islamaj Doğan R, Névéol A, Lu Z

### A context-blocks model for identifying clinical relationships in patient records

### BMC Bioinformatics 2011 Jun 9;12 Suppl 3:S3

In this paper, the authors present a successful end-to-end method for relationship extraction from clinical documents, as defined in the fourth i2b2 challenge. With the era of Electronic Health Records (EHR), the automatic recognition of both clinical concepts and the relationships that tie them together has become an active field of research motivated by clinical applications ranging from quality of care to hypothesis generation.

The system proposed by the authors starts with the recognition of concept phrases and then predicts possible relationships between two concepts that are found in the same sentence. Eight relations were highlighted between medical problems, treatments and tests. These latter were represented through a schema of five distinct context blocks, not necessarily consecutive, determined by the position of the two concepts in the sentence: [Introductory Block – 1st Concept Block - Connective Block – 2nd Concept Block – Conclusive Block]. A machine learning model, implemented through the MALLET language toolkit, was built to identify concepts and extract relationships. Conditional Random Fields (CRF) was successfully applied to recognize concepts, and Support Vector Machine (SVM) classifier was used for relationship identification. A set of 826 patient records from the fourth i2b2 challenge was used for training (349 documents) and evaluating (477 documents) the system.

The concept recognition system achieves a high accuracy with an F-measure of 0.870 for identifying the concept boundaries that are critical for the context-blocks relationship model. With an F-Measure of 0.775, the concept-block representation of relationships is more successful at identifying relationships than the traditional bag-of-words representation. The authors conclude that such a system may serve as a preliminary step for other discovery tasks in medical informatics.

### Zheng K, Mei Q, Hanauer DA

### Collaborative search in electronic health records

### J Am Med Inform Assoc 2011 May;18(3) 282-91

To help improve the quality and efficiency of information retrieval in healthcare, the authors implemented and evaluated a "collaborative search" feature, integrated into a homegrown electronic health record (EHR) search engine. Inspired by the success of many social information-foraging techniques used on the web, this study allows users to preserve their search knowledge and share it with others thus contributing to the transfer of search expertise across people and domains.

To address this issue, an Electronic Medical Record Search Engine (EMERSE) was built by the authors and successfully integrated in two institutions. This system provides a full-text search capability analogous to that of Google, in addition to features specifically designed to retrieving information embedded into unstructured medical data. In particular, 'search-terms bundles' were introduced to hold collections of words and phrases commonly used to express a concept in unstructured narrative documents stored in EHRs. On average, the search-terms bundles, dropped by the users in the system, contain 20 distinct terms. In EMERSE, the users have the option to convert a search query into a private or public search-terms bundle at any time. An empirical evaluation study, conducted over a 4-year period, recorded the interactions of 451 users (mainly academic researchers and medical practitioners) with the collaborative search feature provided in EMERSE. A social-network analysis was carried out to scrutinize the data. Five different bundle-sharing networks were built, highlighting various relationships on bundle usage, between the creator of the bundle and the consumer(s) of the bundle.

The outcomes reveal that the search-engine users were enthusiastic. About half of the bundle creators used the collaborative search feature to share their search knowledge with other users. Moreover, of nearly a million EHR searches processed by the system, about half were based on stored search-terms bundles, 35.8% of which utilized the shared knowledge made available by others in the user community. Thus, the knowledge of EHR search is expected to be collectively refined and distributed across people and domains.