A Clinical Information Management Platform for Semantic Exploitation of Clinical Data

Christian Seebode, Martin Trautwein, Matthias Ort, and Jan-Marcus Lehmann

***Abstract* — We present a platform that combines an approach to semantic extraction of medical information from clinical free-text documents with the processing of structured information from HIS records. The information extraction process uses a fine-grained linguistic analysis, and maps preprocessed terms to the concepts of domain-specific ontologies.**

***Index Terms* — clinical trial, knowledge management, nlp, semantic technologies, text mining**

I. Introduction

Hospital Information Systems (HIS) were designed to support administrative and logistic workflows. As their scope of application is limited, HISs barely satisfy the need for medical data in medical procedures and clinical investigations [14]. Nevertheless, HISs contain valuable *routine clinical data that should be available for healthcare delivery and/or clinical research*. Up to now, only few institutions were capable of taking full advantage of HIS data. In order to address this issue in a more advanced way, we need strategies to integrate HIS data with data produced by medical procedures. A reasonable strategy is to unify both kinds of data by interpreting and mapping their semantics. The co-existence and mutual dependency of medical data and knowledge are highly dynamic and put specific requirements on the development and architecture of mature information systems that aim at supporting all stakeholders in healthcare delivery. True semantic interoperability [1] of clinical information systems (including EHRs) is not a trivial task. Information models represent not only “what is”, i.e. the reality of the patient, but also “what is known”, i.e. the epistemic state of the health professional writing the documentation [2].

Manuscript submitted January 23, 2013. This work was supported in part by Technologiestiftung Berlin (TSB) and by Europäischen Fond für regionale Entwicklung (EFRE)

Dr. Christian Seebode is CTO of ORTEC medical, Am Sandwerder 37, 14109 Berlin, Germany (e-mail: seebode@bfg-berlin.de).

Dr. Martin Trautwein is with Vivantes – Netzwerk für Gesundheit GmbH, Oranienburger Str. 285, 13437 Berlin, Germany (e-mail: trautwein@bfg-berlin.de).

Matthias Ort is CEO of ORTEC medical, Am Sandwerder 37, 14109 Berlin, Germany (e-mail: mort@ortec.org).

Prof. Dr. Jan-Marcus Lehmann is professor at the HTW University of Applied Sciences, Treskowallee 8, 10318 Berlin, Germany

The *increasing demand for information about diseases and their clinical courses* puts more demand on the complexity of information processing. The knowledge-intense nature of healthcare processes [3][4] is not supported by mature information systems, while, by tradition, medical professionals exchange information mostly in textual form.

An efficient and comprehensive integration and exploitation of these data will be one of the success factors for improving health care delivery to individual patients, making health care services more cost-effective at the same time.

Evidence-based-medicine connects healthcare delivery to medical research. Both fields require an instant access to clinical data and a flexible approach to handle them. Current information systems are not prepared to offer this flexibility, while medical science and medical routine often seem to be separate domains when viewed from an IT perspective. Clinical trials make use of the information stored in electronic health records, but much of this information is encoded in free text rather than stored in structured records [5].

II. Method

We present a platform that addresses all mentioned requirements and combines an approach to semantic extraction of medical information from clinical free-text documents with the processing of structured information from HIS records. Existing clinical data are extracted from HIS records and clinical texts, semantically enriched to facts and stored into a semantic patient record. Information is extracted from texts using a fine-grained linguistic analysis, and preprocessed terms are mapped to the concepts of domain-specific ontologies. These domain ontologies comprise knowledge from various sources, including expert knowledge and knowledge from public medical ontologies and taxonomies.

Facts extracted from both clinical free texts and structured sources represent chunks of knowledge. Enrichment algorithms are the entities that produce medical facts, thereby reflecting the states of knowledge and data processing at a particular point of time. Enrichment algorithms produce facts in a particular version or instance that may evolve with time hence considering knowledge or data progression. The fact consuming entities are applications that support particular use cases like patient recruitment for clinical trials, feasibility studies or decision

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

support. Applications are supported by subscribing to an event stream that represents a selection or subset of facts required for that particular use case.

Facts themselves may be subject to further enrichments that produce an update or just different facts (e.g. statistical analysis). This iterative refinement is necessary because

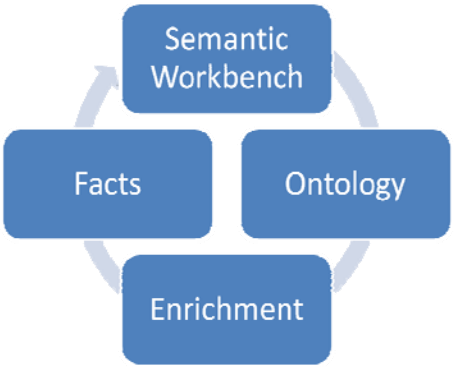

Fig. 1. Ontology Engineering Process

medical knowledge is constantly evolving, as are the information requirements of healthcare delivery and clinical science. Iterations that produce new facts may be triggered from both new data and/or new ontologies.

Physicians and medical experts are able to engineer ontologies and adapt the knowledge needed by an application or use case. For example, patient recruitment scenarios require knowledge that represents study criteria for case selection. The knowledge engineering process itself is always done by validating data and ontologies in an iterative fashion. Ontologies can always be validated and tested against the real data flowing into the system from structured sources or clinical texts. E.g. experts identify ontological concepts in texts by just scraping (?) above them.

III. Architecture

Facts are stored in a Clinical Data Repository (CDR) using a common document-oriented storage model, which takes advantage of an application-agnostic format, in order to support different use cases. It furthermore supports version control of facts reflecting the evolution of information. CDR clients are data-source adapters for storage of structured data or raw clinical texts, enrichment algorithms that produce semantically enriched facts. Facts represent statements drawn from original data like a clinical observation or symptoms but also complex statements that require analysis of other facts like a statistical analysis. The CDR generally separates information generation processes from information processing or consumption processes, and thus supports smart partitioning of data for scalable application architectures. The CDR supports a subscription mechanism where applications may register for a stream of events.

Applications are the information processing entities that support particular use cases like patient recruitment. Semantically, applications and adapters or enrichment algorithms share the knowledge needed for an application. Applications usually possess a web-based GUI and subscribe to the CDR event stream. But applications themselves may embed complete infrastructures that may harbor extensive algorithms running asynchronous to the event sources and user requests. The decoupled processing done by the platform entities allows them to adapt to consistency, availability or real-time requirements in a flexible way. Applications may decide to rely on consistent, time-critical data in order to provide real-time processing. Others decide to relax on consistency, ensuring availability.

The web-based application *StudyMatcher* maps study criteria to a list of cases and their medical facts. Trial teams

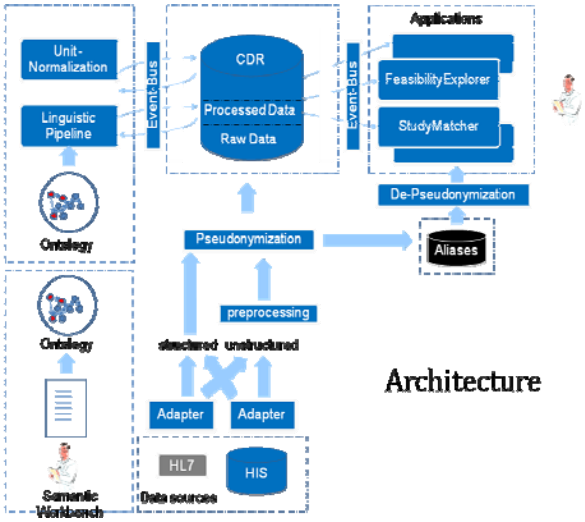

Fig. 2. Platform Architecture.

may define study criteria in interaction with the knowledge resources. The application automatically generates a list of candidate cases. Since the user interface links the facts extracted by the system to the original sources (e.g. the clinical documentation), users are able to check with low effort whether or not a fact has been recognized correctly by the system, and matched correctly with the given criteria. This strategy of combining automatic and supervised fact generation promises to be a reasonable approach to improving the semantic exploitation of data.

Data-source adapters are specific selectors and regulators of data traffic with HIS. A generic HL7 adapter integrates standardized HL7 traffic into the platform. HIS-specific adapters are currently available for AGFA Orbis and SAP I.s.h.med.

IV. Related Projects

We evaluated a couple of related projects [6][7][8][9][10][11][12] to learn about the strengths and weaknesses of our platform. *I2B2* [6] and *SHRINE* [7] offer a distributed architecture for federated queries. Though our platform in its current version doesn't offer a completely federated architecture, the architecture itself is scalable. *I2B2* doesn't offer NLP or semantic technologies in the first place. *Stanford SCCI* offers 'extracting useful information from typed or dictated text within medical records' [10] and the STRIDE Cohort Discovery Tool addresses the same problem domain as StudyMatcher. *EHR4CR* [12] is a European consortium that uses the *I2B2* architecture for

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

applications in clinical research, but only with data from structured sources.

V. Future Applications

StudyMatcher and *FeasibilityExplorer* are the first applications built on top of that platform that support clinical trial uses cases. Other will follow that provide mobile access to clinical data and decision support. Patient-centered applications or infrastructures are feasible because of the possibility of attaching semantics to patient data, thereby leveraging rich dialogues between the stakeholders in healthcare delivery.

VI. Conclusion

In contrast to traditional HIS and clinical information systems, we present a platform that is not limited to specific use cases. It supports the flexible and just-in-time exploitation of clinical data from various sources, including clinical texts, to leverage unlimited clinical or patient-centered use cases. Knowledge is represented by ontologies and engineered by the domain experts themselves in a straightforward manner in a semantic workbench. Semantically enriched facts are processed by domain-specific applications that support specific use cases. Semantic exploitation of clinical data from structured HIS records and clinical texts provides an up-to-date evidence base for medical research and healthcare delivery.

Platform and applications are developed in cooperation with Europe's leading healthcare providers Charité and Vivantes, and will be rolled out in January 2013.

References

[1] Semantic Interoperability for Better Health and Safer Healthcare, S e m a n t i c H E A L T H R e p o r t, 2009, http://ec.europa.eu/information_society/activities/health/docs/publications/2009/2009semantic-health-report.pdf

[2] Bodenreider O, Smith B, Burgun, A.: The Ontology-Epistemology Divide: A Case Study in Medical Terminology, Proc. FOIS-2006, Torino, Italy (2004).

[3] Marjanovic, O.: Improving Knowledge-Intensive Health Care Processes beyond Efficiency. In: ICIS'11 (2011)

[4] Ammon,D.,Homann, D., Jakob, T., Finkeissen, E., Detschew, V., Wetter, T.: Management of Knowledge-Intensive Healthcare Processes on the Example of General Medical Documentation. In: BPM Workshops (2008)

[5] Shah AD, Martinez C, Hemingway H, The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records, BMC Med Inform Decis Mak. 2012 Aug 7;12:88. doi: 10.1186/1472-6947-12-88.

[6] I2B2, Informatics for Integrating Biology and the Bedside, https://www.i2b2.org/

[7] SHRINE, Shared Health Research Informatics Network, https://www.i2b2.org/work/shrine.html

[8] GAIN, Genetic Association Information Network, http://www.genome.gov/19518664

[9] REMIND, Reliable Extraction & Meaningful Inference from Non-structured Data, http://www.medical.siemens.com/webapp/wcs/stores/servlet/CategoryDisplay~q_catalogId~e_-1~a_categoryId~e_1031969~a_catTree~e_100010,1008631,1029622,1031969~a_langId~e_-1~a_storeId~e_10001.htm

[10] SCCI, Stanford Center for Clinical Informatics, https://clinicalinformatics.stanford.edu/

[11] THESEUS/MEDICO, http://theseus-programm.de/de/920.php

[12] EHR4CR, Electronic Health Records for Clinical Research, http://www.ehr4cr.eu/

[13] J. Frankovich, M.D., C A. Longhurst, M.D., and SM. Sutherland, M.D., Evidence-Based Medicine in the EMR Era, N Engl J Med 2011; 365:1758-1759

[14] Marsolek, I., Sander, H., Backhaus, C., Friesdorf, W., Haake, K., Petersen, R., Werners, M., Höhn, T. (2006). Workflow Process Support for the Clinical Routine – A Comparative Assessment of Hospital Information Systems (HIS). Journal of Clinical Monitoring and Computing, 20 (2), 118-119

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)