Knowledge Acquisition in the construction of ontologies: a case study in the domain of hematology

Fabrício M. Mendonça[1,*] Kátia C. Coelho[1] André Q. Andrade[1,2] and Mauricio B. Almeida[3]
[1] Graduate Program in Information Science, Federal University of *Minas Gerais*, Brazil
[2] Institute for Medical Informatics, Medical University of Graz, Austria
[3] Department of Information Theory and Management, Federal University of *Minas Gerais*, Brazil

ABSTRACT

The activities of organizing knowledge recorded in texts and obtaining knowledge from human experts – the knowledge acquisition process – are essential for scientific development. In this article, we propose methodological steps for knowledge acquisition, which have been applied to the construction of biomedical ontologies. The methodological steps are tested in a real case of knowledge acquisition in the domain of the human blood. We hope to contribute to the improvement of knowledge acquisition for the representation of scientific knowledge in ontologies.

1 INTRODUCTION

Ontologies have been proposed as an alternative for creating representations of reality suitable for computers. At least four activities are essential in the development of ontologies: specification, knowledge acquisition, conceptualizatio and formalization. In knowledge acquisition (KA), the experience available in the literature of diverse fields mentions difficulties in communication between experts and professionals who deal with information (Boose, 1990).

This article investigates the activity of KA within the scope of biomedicine. In order to explore the activity, we propose procedures for KA employing the best practices referenced in the literature. We systematize these procedures in a list of methodological steps with the aim of testing their feasibility in a real case.

The empirical research is conducted within the scope of a biomedical project, focused on human blood. The knowledge acquisition results have been used in the development of a knowledge base for scientific and educational applications related to the human blood. Descriptions of different stages of research are provided as examples throughout the article. The main contributions are the aforementioned list of steps and observations made in real situations with the aim of improving the KA performance.

The remainder of this paper is organized as follows: section 2 reviews the literature on KA. Section 3 explains the theoretical rationale, the systematization and tools that compose the KA methodology. Section 4 presents comments of interest during the next phases of the research. Finally, section 5 puts forward our final remarks.

2 BACKGROUND

2.1 An overview of Knowledge Acquisition

The KA activity generally includes the collection, analysis, structuring and validation of knowledge for representation purposes (Hua, 2008). It is an activity composed of a set of tasks that employ computer-based and manual techniques (Gaines, 2003; Boose & Gaines, 1989; Shadbolt, 2005). A multitude of definitions for KA can be found (Shaw & Gaines, 1996; Scott & Clayton, 1991; Payne et al, 2007) and the theories and methods that support KA activities rely on diverse academic research fields. Ways of acquiring and representing knowledge come from Computer Science (Compton & Jansen, 1989), Cognitive Science (Hawkins, 1983), Linguistics (Campbell et al, 1998) and Psychology (Harris, 1976).

2.2 Classification of KA techniques

KA techniques can be classified into manual techniques and computer-based techniques (Boose, 1990). In general, the manual techniques are rooted in Psychology (Kelly, 1955) and computer-based techniques are classified as automatic or semi-automatic. KA can be classified according to the knowledge obtained in the process. The assumption that different methods of elicitation result in different types of knowledge is known as the differential access hypothesis (Hoffman et al, 1995). In addition, KA can be classified according to application methods such as protocol-generation techniques, protocol-analysis techniques, matrix-based techniques and sorting techniques (Shadbolt & Swallow, 1993).

Protocol-generation techniques include interviews. The most well-known technique for interviews is the *teachback* technique (Hua, 2008; Shadbolt, 2005). Protocol-analysis techniques are used in the transcription of interviews in order to identify different knowledge types. Matrix-based techniques involve the diagrammatic organization of problems. The most well-known technique is the *repertory grid* (Hua, 2008; Shadbolt, 2005). Sorting techniques are techniques in which the domain entities are classified in order to check how an expert classifies the knowledge. The most well-known technique is *card sorting* (Hua, 2007; Hoffman

[*] To whom correspondence should be addressed: fabriciommendonca@gmail.com

et al, 1995). The Diagram-based technique consists of the creation and use of network representations, such as conceptual maps (Corbridge et al, 1994). A methodology for KA that combines card sorting and laddering can be employed in the construction of ontologies (Wang et al, 2006).

2.3 KA in Biomedicine

Natural Language Processing (NLP) techniques are common in the biomedical domain (Hersh, 2009; Verspoor et al, 2006). These techniques can be divided into two main streams: the rule-based approach (Friedman et al, 2004; Hahn, Romacker & Schulz, 2002) and the statistical approach (Taira & Soderland, 1999; Sebastiani, 2002).

A comparison between the two methods involved the testing of systems using both approaches to the automatic categorization of MEDLINE abstracts (Humphrey et al, 2009) and found comparable results for most evaluated items. The results favored the statistical approach, though the authors suggested the combination of both approaches.

3 METHODS

3.1 Case study: knowledge context and domain

This work explores the best practices in an ongoing KA scenario applied within the scope of the Blood Project (Almeida, Proietti & Smith, 2011), an information organization initiative in hematology. The project is taking place in a medical institution responsible for hematology and blood transfusion research and that offers healthcare services for a population of around 20 million people.

3.2 Methodological steps

In this section, we describe the list of steps for KA. Then, we present a synoptic table summarizing the tasks involved and systematizing the steps in the list, which was divided into four main phases: *extraction, elicitation, validation and refinement*.

In the *extraction phase* we applied NLP techniques and tools in order to obtain candidate terms for the ontology. KA from texts consists of three main activities: construction of a corpus related to blood transfusion, codification of this corpus and information retrieval from the corpus.

The subset of the corpus related to blood transfusion uses the manual of the American Association of Blood Banking (AABB) as a source. From the AABB website[1] we downloaded thirty-two chapters that comprise the seventeenth edition of the manual. From this material, twenty-seven chapters were processed by the tool used for codification. This material was select as a sample according to the stage of the research underway when writing this paper. Certainly, in future works, diseases processes and clinical finding will be considered.

In the activity of codification we employed Sketch Engine[2], an online tool for the creation and analysis of linguistic corpora. The fragmentation of the text into morphemes and the identification of the grammatical classes are automatically performed.

After the codification activity, we proceeded with the information retrieval from the corpus with the aim of identifying terms used to describe blood transfusion procedures. In order to do so, we used word suffixes common of medical terms (Lovis, Baud & Rassinoux, 1998) as such *-apheresis, -centesis, -desis, -ectomy, -opsy*, to mention but a few. Then, we built regular expressions using the Sketch Engine corpus query language, in order to retrieve terms related to procedures, as well as the absolute frequencies that occur in the corpus.

As a final task of the extraction phase, we analyzed the morphological productivity of the terms obtained using the British National Corpus (BNC)[3] as a reference. The analysis consisted of comparing the frequency of each term in the corpus with its frequency in the reference corpus. In order to proceed with the morphological productivity analysis we used the AntConC[4] tool.

In the *elicitation phase*, we made use of the terms obtained in the extraction phase, which were employed as guidelines to start the contact with experts. This phase consisted of holding interviews and the application of KA techniques with experts, doctors, biologists and researchers. During the course of the interviews, sorting and matrix techniques were applied. The cycle that characterizes the clinical process, ranging from the development of an infectious disease through its treatment, was adopted to guide the approach taken with the experts. For modeling the domain, we adopted the disease as disposition approach, as proposed by (Scheuermann, Ceusters & Smith, 2009). The three major stages that comprise that cycle are: etiological process, course of disease and therapeutic response. In order to apply the described reasoning so far, a template was created in Protégé-Frames.

In the stage called *etiological process*, there is a healthy human body with characteristics that are normal according to medical parameters. In the pre-clinical manifestation of the disease, the body develops disorders, which are bearers of dispositions. Such dispositions are naturally associated with the entities' existence, for example, the disposition of the human body to get sick (Smith, 2008). There are changes in the patient already, but not noticed. The etiological process stage can be represented as follows: ETIOLOGICAL PROCESS => *produces* => DISORDER => *bears* => DISPOSITION.

[1] Available at: <http://www.aabb.org>. Accessed: July 23, 2010

[2] Available at: <http://www.sketchengine.co.uk/>. Access: Dec. 15, 2010

[3] Available at: <http://www.natcorp.ox.ac.uk/>. Access: Nov. 30, 2011

[4] Available at: <http://www.antlab.sci.waseda.ac.jp>. Access: July 23, 2011

The course of disease stage starts with the clinical manifestation of the disease (disposition). At this moment, the disorder manifests itself through symptoms, which the patient is able to identify. Then, a doctor identifies the disease signs through a physical exam or through a report of the patient. In this stage, it is possible to determine the clinical phenotype, that is, the principal observable characteristic of that disease. The course of disease stage can be represented as follows: DISPOSITION => *realized in* => PATHOLOGICAL PROCESS => *produces* => ABNORMAL BODY FEATURES.

In the therapeutic response phase, a sample is taken from the infected part of the body in order to perform laboratory tests. At this point, it is possible to establish a treatment plan so that the body may return to normality. The plan is the result of a diagnosis founded in the interpretative process of a clinical framework. The clinical framework is composed of symptom representation records as well as physical and laboratory exam results. The therapeutic response stage can be represented as follows: ABNORMAL BODY CONDITION => *recognized as* => SIGN AND SYMPTOM => *used in* => INTERPRETATIVE PROCESS.

The third phase of the proposed list of steps for KA, called the *validation phase*, uses wiki science tools for collaborative validation of candidate terms for an ontology. After the elicitation phase, according to the knowledge obtained, candidate terms are transferred to a wiki to then be validated by experts online.

The fourth stage of the proposed list of topics, called the *refinement phase*, uses a second template, also created using Protégé-Frames. The goal was to record information about how to integrate the different levels of granularity required to understand a disease and its manifestations. This integration involves obtaining the relations between parts of the body that a certain disease affects, the related genes and the related proteins.

Finally, the steps put forward so far are gathered together, thus creating the list of steps for KA.

Phase	Task	Description	Resources and people involved
(1) Extraction	1.1 build a corpus	Create a corpus from texts	-Medical texts -K. engineer
	1.2 codification	Automatically fragment texts	-Sketch Engine tool -K. engineer
	1.3 information retrieval	Obtain terms through suffixes	-Sketch Engine tool - K. engineer
(2) Contact	2.1 obtain knowledge	Hold interviews with experts	-Template Protégé and teachback; -K. engineer, experts
	2.2 know the terminology	Identify experts' rationale	-Matrix Techniques -K. engineer and expert
	2.3 see *ad-hoc* organization	Understand how experts sort concepts	-Sorting techniques - -Experts
(3) Validation	3.1 validate knowledge	Obtain approval of terms acquired	-Wiki Page -Expert
	3.2 updating	Update data after each validation	Wiki Page K. engineer
(4) Refinement	4.1 integration between granularities	Characterize related genes, proteins, etc	-Template Protégé -K. engineer
	4.2 connection with top-level	Connect data with other ontologies	-Template Protégé- - K. engineer

Table 1: KA list of steps proposed

4 RESULTS

One evident result is the methodological list of steps described in the previous section, which has been tested and improved over the course of the research (Table 1).

In the codification activity (extraction phase), from the texts selected 369,741 tokens were automatically identified and related to parts-of-speech. Subsequently, in the information retrieval phase, 57 terms related to blood transfusion procedures were identified. Table 2 depicts the top-five terms from the set of 57 terms retrieved, which were used as a basis for starting interviews with experts:

Term	Frequency
apheresis	124
phlebotomy	32
cytometry	20
cordocentesis	16
plasmapheresis	15

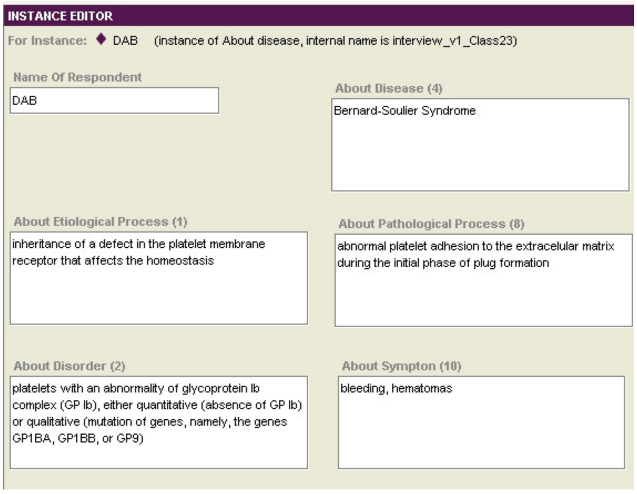
Table 1: top-five terms retrieved and absolute frequency

The rationale applied in the elicitation phase made it possible to understand the major stages of the disease manifestation. Table 2 presents an example of blood disease analysis following this rationale for Bernard-Soulier Syndrome:

Etiological process	inheritance of a defect in the platelet membrane receptor that affects the hemostasis
Disorder	platelets with a glycoprotein Ib complex (GP Ib) abnormality, either quantitative (absence of GP Ib) or qualitative (mutation of GP1BA, GP1BB, GP9)
Disposition	Bernard-Soulier Syndrome (A, B or C)
Pathological process	abnormal platelet adhesion to the extracellular matrix during the initial phase of plug formation
Symptoms	bleeding, hematomas
Signs	excessive bleeding, gingival bleeding, menorrhagia, purpura, epistaxis, gastrointestinal bleeding

Table 2: KA reasoning applied to a blood disease

An example of a Protégé-Frames template related to *Bernard-Soulier syndrome* is depicted in Fig. 4.



INSTANCE EDITOR

For Instance: ◆ DAB (instance of About disease, internal name is interview_v1_Class23)

Name Of Respondent: DAB

About Disease (4): Bernard-Soulier Syndrome

About Etiological Process (1): inheritance of a defect in the platelet membrane receptor that affects the homeostasis

About Pathological Process (8): abnormal platelet adhesion to the extracelular matrix during the initial phase of plug formation

About Disorder (2): platelets with an abnormality of glycoprotein Ib complex (GP Ib), either quantitative (absence of GP Ib) or qualitative (mutation of genes, namely, the genes GP1BA, GP1BB, or GP9)

About Sympton (10): bleeding, hematomas

Fig. 4. Protégé-Frames template with example about blood disease

Finally, it is worth mentioning that at the time this article was being written, the ontology developed in OWL had more than 300 classes and 50 properties, and practically all the methodological steps were up and running, providing data for different ontology parts.

5 DISCUSSION

In each stage of the KA process, as depict in Table 1, it is possible to identify issues to be discussed:

i) The extraction stage was undertaken mainly by a knowledge engineer using NLP tools applied to sources suggested by experts. As a means of producing a list of relevant terms in a domain, the extraction was useful in preventing the knowledge engineers from having to start interviews from scratch. In general, the terms selected were useful for describing the domain according to the opinions of experts.

ii) The contact stage is the heart of KA processes, since it is within this stage that experts share their knowledge. This stage was conducted as a cycle that involved interviews interspersed with attempts to understand the rationale used by experts to understand the phenomena in the domain. As part of this attempt, the knowledge engineer employed sorting and matrix techniques. Regarding the interview based on an ontological disease model, it is worth reporting that the results were very reasonable, insofar as the experts approved of the framework organized in the etiological process, course of disease and therapeutic response proposed by Scheuermann, Ceusters, & Smith (2009).

iii) The validation stage was conducted, in many cases, during the interviews, mainly in the beginning of the process when experts didn´t have experience with Wiki pages. In general, the validation confirmed the interviews and the teachback technique performed previously. It´s worth noticing that the difficulties in the validation stage did not occur among experts validating their own prior knowledge. Rather, the majority of cases of non-validity occurred when an expert evaluated the knowledge provided by another expert. However, the differences did not seem irreconcilable. In many cases, experts suggested referring to their own scientific publications to resolve outstanding issues.

iv) The refinement stage was conducted in the same way as the contact stage. Indeed, it was conducted as an interview merged with work to understand the rationale behind and organization of the experts'concepts. When analyzing the results, one can conclude that this stage provides useful insights into the building of ontologies in terms of interoperability. This is because the refinement stage is based on the premise of connection to top-level ontologies.

Observations made over the course of all these stages allowed us to identify problems that occur in the KA process for which solutions have been sought as the research has continued. These problems are the result of the influence of the following factors:

i) factors related to the expert profile, such as: training, experience and previous participation in similar projects, limitations in expertise;

ii) contextual factors, such as: cultural, geographical, political and financial issues, lack of access to information sources and deficiency in organizational structure;

iii) factors related to the interaction between expert and knowledge engineer, such as: short-term outlook (KA is seen as "additional work") and domain complexity;

iv) factors that make recording results difficult, such as: non-approval by the expert of the results of the activity and constant advancement in the scientific field.

Concerning the proposed elicitation technique (section 3.2), which is based on Scheuermann, Ceusters & Smith (2009), one can argue that there is a methodological pitfall when using a formal disease model to acquire knowledge. It could be argued that relevant domain knowledge could be missed by doing so, because what would be acquired is something of a pre-conceived frame of meaning. However, we observed that some sort of structure was required to conduct the activity and save time, mainly considering the limited availability of the experts. According to our experience in this study of case, knowledge missed for this reason may be dealt with using complementary techniques. The interviewees were not constrained when talking and teachback techniques were employed to give them the chance to clear up misunderstandings and flaws. In addition, the ontological disease model was used only to organize the interview and to make notes, not in an attempt to formalize knowledge directly.

The NLP techniques applied aimed at collecting candidate terms for the ontology, instead of trying to populate it directly. In this sense, the use of those techniques was important to obtain a first list of candidate terms. Even though

NLP is not considered a good source for ontological knowledge, it may be useful when dealing with a large volume of material. Another issue when using NLP was the size of our sample: in order to build a significant corpus, one should have at least 10 million words, which are not available to us.

6 CONCLUSION

This article has proposed a list of steps for KA, which are based on techniques found in the literature. The steps in the list has been tested, proving their viability. The work described includes a project in which research was conducted to identify the best practices for and difficulties in performing the KA activities with hematology experts within the scope of creating an ontology. The list of steps is a partial result that has been improved based on direct observation.

One conclusion we could draw from the overall experience is that KA is a very time-consuming and expensive process. This may explain why it is neglected in many cases. In future work, we intend to further clarify in which context each technique is most suitable. This could be done with assistance from experts, taking in account their time limitations. Regardless, in this case study, some techniques were chosen, as was mentioned in last column of Table 1. The list of topics has been successfuly applied in other related domains. It appears to be a systematized alternative for creating ontologies using a rational means of approaching experts.

ACKNOWLEDGEMENTS

This work is partially supported by Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Governo do Estado de Minas Gerais, Brazil, Rua Raul Pompéia, nº101 - São Pedro, Belo Horizonte, MG, 30.330-080, Brazil. This work partially supported by CNPQ (Conselho Nacional de Pesquisa e Desenvolvimento)

REFERENCES

Almeida, M.B., Proietti, A.B., AI J., Smith, B. The Blood Ontology: an ontology in the domain of hematology. (2011) Proccedings of ICBO.

Boose, J.H. Knowledge acquisition tools, methods, and mediating representations. In: Japanese Knowledge Acq. for KBS, JKAW, 1. 1990.

Boose, J.H and Gaines, B.R. Knowledge Acquisition for Knowledge-Based Systems. 1989. Machine Learning, v. 4, p. 377-394.

Campbell, K.E. et al. Representing thoughts, words, and things in the UMLS. J Am Med Inform Assoc., v. 5, n. 5, p. 421–31, 1998.

Compton, P. and Jansen, R. A philosophical basis for knowledge acquisition. Knowledge Acquisition. European KA for KBS. 1989.

Corbridge C, Rugg G, Major N, Shadbolt N.R, Burton A. Laddering: technique and tool use in knowledge acquisition. J. of Knowledge Acquisition, 1994, p. 315-341. Available from: https://blog.itu.dk/SLR-F2010/files/2010/07/paper-1-pages-1-12-15.pdf

Friedman, C, Shagina, L., Lussier, Y., Hripcsak, G. Automated encoding of clinical documents based on NLP. Journal of the American Medical Informatics Association. 2004 Sep-Oct;11:392-402.

Gaines, B.R. Organizational Knowledge Acquisition. In: Handbook on knowledge management. Birkhäuser: Springer. 2003, 700 p.

Hahn, U., Romacker M., Schulz, S. Medsyndikate - a natural language system for the extraction of medical information from findings reports. International Journal of Medical Informatics. 2002;67:63-74.

Harris Z. On a theory of Language. The Journal of Philosophy, v. 73, n. 10, p. 253-276 1976.

Hawkins, D. An analysis of expert thinking. International Journal of Man-Machine Studies. v. 18, p. 1-47, Jan. 1983

Hersh, W. Information Retrieval: A Health and Biomedical Perspective 3ed: Springer 2009.

Hoffman, R.R., Shadbolt, N.R., Burton, A.M., Klein, G. Eliciting knowledge from experts. Organizational Behavior and Decision Processes. v. 62, n.2, 1995. pg 129-158.

Hua, J. Study on Knowledge Acquisition Techniques. 2nd Inter. Symp. on Intelligent Information Technology App. 2008.

Humphrey, S.M., Neveol, A., Browne, A., Gobeil, J., Ruch, P., Darmoni,SJ. Comparing a Rule-Based Versus Statistical System for Automatic Categorization of MEDLINE Documents According to Biomedical Specialty. J Amer Soc for Inf Sci and Tech. 2009 Dec; 60:2530-9.

Kelly, G.A. The psychology of personal constructs. New York: Norton, 1955.

Lovis, C., Baud, R., Rassinoux, A.M., Michel, P. A., Scherrer, J.R..(1998). Medical dictionaries for patient encoding systems: a methodology. Art. Int.in Medicine. Vol. 14, Issue 1, pp. 201-214.

Milton, N., Clarke, D., Shadbolt, N. Knowledge engineering and psycology. Int. J. of Human-Computer St., v. 64, n. 12, p. 1214-1229. 2006.

Payne P.R, Mendonça E.A, Johnson S.B, Starren J.B. Conceptual knowledge acquisition in biomedicine: a methodological review. J Biomed Inform. 2007. v 40, n. 5, p. 82–602.

Sebastiani, F. Machine learning in automated text categorization. ACM Comput Surv. 2002 Mar; 34:1-47.

Scott, A.C., Clayton, J.E., Gibson, E.L. A practical guide to knowledge acquisition. Addison–Wesle, 1991, 509 p.

Shadbolt, N. Eliciting Expertise. In: Evaluation of Human Work. Ed. Taylor & Francis. 2005.

Shadbolt, N., Swallow, S. Epistemics: Knowledge Acquisition. Shaw, M.LG. and Gaines, B.R. Soft.Engineering Journal, 1996. p. 149-165.

Smith, B. New Desiderata for Biomedical Terminologies. In Munn, K.; Smith, B. (Ed.). Applied Ontology. Frankfurt: Verlag. 2008. pp. 21-39

Scheuermann, R.H., Ceusters, W., Smith, B. Toward an ontological treatment of diasease and diagnosis. Proceedings 2009 Summit on Transl. Bioinf., San Francisco, CA, pp. 116-120.

Taira, R.K and Soderland, S.G. Statistical natural language processor for medical reports. J. Am. Med. Infor. Ass. 1999:970-4.

Verspoor, K., Bretonnel, C. K., Goertzel, K., Mani, I. Linking natural language processing and biology: towards deeper biological literature analysis. BioNLP '06: 2006; Morristown USA: ACM; 2006. p. iii-iv.

Wang Y, Sure Y, Stevens R, Rector, A. Knowledge elicitation plug-in for Protégé: Card sorting and laddering. In: The Semantic Web, ASWC, v. 4185, p. 552-565, 2006. Available from: http://dx.doi.org/10.1007/11836025_53