Natural Language Definitions for the Leukemia Knowledge Domain

Amanda Damasceno de Souza
Federal University of Minas Gerais, UFMG
Belo Horizonte, Minas Gerais, Brazil
amanda@ufmg.br

Mauricio Barcellos Almeida
Department of Theory and Management of Information
Federal University of Minas Gerais, UFMG
Belo Horizonte, Minas Gerais, Brazil
mba@eci.ufmg.br

***Abstract*— The creation of natural definitions is a phase of any methodology to build formal ontologies. In order to reach formal definitions, one should first create natural language definitions according to sound principles. We gather a set of principles available in literature and organize them in a list of stages that one can use to create good definitions in natural language. In order to test the set of principles, we conducted a case study in which we create definitions in the domain of cancer, more specifically, definitions for acute myeloid leukemia. After creating and validating the definition of this specific kind of leukemia, we offer remarks about the experiment.**

Keywords—Natural Language Definitions; Biomedical ontologies; Leukemia.

I. INTRODUCTION

In Information Science, ontologies have attracted the interest of researchers working on knowledge organization systems. Ontologies are employed to organize information and knowledge for purposes of information retrieval [1-2]. Ontologies also provide terminological standardization with the aim of organizing scientific knowledge and building repositories that are designed to foster interoperability between electronic information systems.

When building ontologies, an important activity is the formulation of definitions for domain-specific terms. The OBO Foundry [3] principles suggest that one should first provide natural language definitions and then provide the logical formulation [4]. Existing methodologies for construction of ontologies do not present satisfactory guidelines to create definitions.

The goal of this paper is to propose methodological principles in order to systematize the process of definition creation in biomedical ontologies. We describe a case study on creating natural language definitions for leukemia, a disease which includes as subtypes acute myeloid leukemia (AML), the myelodysplastic syndromes and the myeloproliferative neoplasm. Here, we focus on AML.

The present paper arises out of our work on the Blood Ontology (BLO) project, a resource that allows the exploration of relevant information for research in hematology [5]. The BLO encompasses terms representing hematologic neoplasia, including leukemia and lymphoma.

II. BACKGROUND: DEFINITIONS IN ONTOLOGIES

The formulation of definitions has been studied by philosophers and linguists since ancient times. It connects philosophical notions – for example relating to the natures or essences of things – with other constructs such as terms, concepts and meanings [6].

The search for a proper definition of terms used to represent biomedical entities is connected to the process of learning. In order to define a term, one needs to have previous knowledge about the subject, to know both the context in which a term is used and the associated technical jargon [7].

The difficulty in the activity of creating definitions in ontologies lies in the lack of trained people for the construction of ontologies. Indeed, this type of task gives rise to several theoretical and practical issues [8]. Definitions for terms in ontologies should be formulated according to familiar logical principles [9, 10, 11].

III. METHODOLOGY

In order to formulate proper definitions for terms related to Hematologic neoplasm, we devised a three-step procedure as follows.

A. The Sample Collection

First, the set of terms relating to AML was selected by identifying which terms in BLO deal with hematologic neoplasia [5], amounting to 43 classes: 25 AML terms, 6 myeloblastic syndrome terms and 12 myeloproliferative neoplasm terms.

B. Knowledge Acquisition

Second, we retrieved definitions in natural language for leukemia from traditional information sources such as the NCI Thesaurus [12]; the NCI Dictionary of Cancer Terms [13]; the Medical Subject Headings (MeSH) [14]; Medscape [15]; Ontobee [16]; the Disease Ontology [17]; the Gene Ontology [18]; and a text-book [19].

C. List of Stages to Create Definitions in Natural Language

Finally, we utilized the results of A and B to formulate a natural language definition, following well established practices identified in literature, such as: genus-differentia and

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.

essence [20]; formal relations [21,22]; necessary and sufficient conditions [11]; inheritance, intelligibility and circularity [9]; logical definitions [8]; issues in definitions and logical definitions [4]; biomedical definitions [32]; textual and formal definitions [23]. The mentioned stages are: 1) To elect the candidate term to be defined; 2) To obtain a preliminary definition; 3) To establish the superior genus; 4) To establish the essential characteristic; 5) To formulate the definition; 6) To check necessary and sufficient conditions; 7) To check non-circularity; 8) To check multiple inheritance.

IV. Results

We describe the list of stages, explaining how to proceed in each stage to create natural language definitions of AML using the example of the first class AML:

1) To elect the candidate term to be defined

First, one should choose the candidate term to be defined according to techniques of knowledge acquisition [24]. In our case study, we followed the list of stages presented in section IIIC for the term "*Acute Myeloid Leukemia*".

2) To obtain a preliminary definition

In this stage one should perform a search in specialized literature to obtain information about the term. The sources may be textbooks, papers, dictionaries, encyclopedias, thesauri and ontologies. In our case study, a librarian helped to select the suitable bibliography and to extract information required to establish the genus and differentia of the term. The main source used to obtain a preliminary definition of AML was the NCI-Thesaurus [12], from which we obtain the following definition:

"*AML is an aggressive (fast-growing) disease in which too many myeloblasts (immature white blood cells that are not lymphoblasts) are found in the bone marrow and blood. Also called acute myeloblastic leukemia, acute myelogenous leukemia, acute nonlymphocytic leukemia, AML, and ANLL*".

3) To establish the superior genus

We determined the genus by seeking to identify a common feature of the selected term. In the case of leukemia, the common feature is the existence of an abnormal derivation of the myeloid lineage that occurs in each AML. So, we established the basic relation: *Acute Myeloid Leukemia* <is-a> *Hematopoietic Neoplasm*.

4) To establish the essential characteristic

We used the notion of differentia in order to define the essential characteristics that mark the distinction of the entity under definition from other entities in the hierarchy. These essential characteristics are often difficult to be found. So, we first studied the domain, then we sought support from a cancer expert. Thus, we analyzed the characteristic that best represented that type of pathology.

The differentia between each class of leukemia was obtained from a diagnosis based on morphological criteria (cell type), immunological criteria (ICD 13, ICS 33, etc.) and cytogenetic criteria (abnormalities t8, t21, q22, q22-PML RARA.), as well as based on the lineage and the maturation degree.

In the domain of cancer, these three diagnostic aspects – morphological, immunological and immunophenotyping – were essential to perform the task of defining. However, we learned from experts that morphology still represents the central criteria in distinguishing leukemia types.

In cases where several features were defined by the diagnosis, another criterion was required in order to find the essence, for example, a criterion based on a prognosis. We reviewed what characteristics induced a prognosis of the disease and the most important was considered the essential characteristics. The participation of an expert was crucial to confirm the essential feature.

"AML derives from an uncontrolled proliferation of the myeloid lineage and their precursors"

5) To formulate the definition in the form

In this stage, the information gathered from stages 1 to 4 was applied to formulate a preliminary version of the definition. The definition has the form: S = Def. a G which is Ds (where "G" (genus) is the superior term of "S" (species) and "S" is the term under definition; and "D" is an essential characteristic, that is, the differentia). So, "S" is the class of leukemia to be defined, "G" is the general class and "D" is the differentia that characterizes an instance S of D in the context of leukemia.

The formulation of the definition was initiated by writing the term to be defined followed by its genus (stage 3) and its differentia (stage 4). Then, we corrected the first version definition from a grammatical point of view, adding or removing parts of the text obtained in the first stages. We also chose preferential terms and eliminated redundant words. After some changes in the first versions, the result of stage 5 was:

"*An acute myeloid leukemia is-a hematopoietic neoplasm that derives from an uncontrolled proliferation of the myeloid lineage and their precursors*".

6) To check necessary and sufficient conditions

The verification was performed through the following expression: to be an *A* is a necessary condition to be a *B*, then each *B* is an *A*; to be an *A* is a sufficient condition to be a *B*, then each *A* is a *B* [11]. This expression means that A" represents the essential characteristic of the definition, and "B" represents the term under definition.

To be an AML is a necessary condition to "derive from an uncontrollable proliferation of a myeloid lineage and its precursors, that is, each AML derives from an uncontrollable proliferation of a myeloid lineage and its precursors".

To be an AML, a sufficient condition is to "derive from an uncontrollable proliferation of a myeloid lineage and its precursors, that is, each AML derives from an uncontrollable proliferation of a myeloid lineage and its precursors".

7) To check the principle of non-circularity

In this stage, we scrutinized the definition for circularity. Roughly, circularity is a situation in which a term is employed to define the very same term.

8) To check the principle of multiple inheritance

All kinds of AML descent from the myeloid lineage because we are dealing with a clonal disease. So, we could define AML without multiple inheritance.

Once the eight steps have been presented, brief considerations regarding the data validation by experts are in order. In our case study, the validation was performed by a pediatric oncologist specializing in leukemia. We first asked the expert whether the derivation from an uncontrollable proliferation of a myeloid lineage (and its precursors) is the main characteristic of an AML. The oncologist reported:

"*To define leukemia we need three characteristics: morphological, cytogenetic and immunophenotyping. This specific case has a unique parent, namely, the myeloid lineage since it is a clonal disease. In cases that a leukemia is biphenotypic or bi-lineage, it has two origins since it presents several clonal cell populations. However, this kind we are evaluating presents only the myeloid lineage. Even if the descent from the myeloid lineage was minimal, it presents only one differentiation*."

V. DISCUSSION

In this section we present some remarks about our experience of creating definitions of AMLs.

After our search for definitions in healthcare and biology information sources, we analyzed the definitions found according to a set of criteria that considers: multiple definitions, lack of proper characterization, intangible definitions, circular definitions and presence of technical terms [4, 9].

In analyzing the definitions of leukemia presented in the literature, we recognized that the definitions found were not satisfactory. For example, definitions found in the Disease Ontology [17] are too general, and Ontobee [16] lacks natural language definitions for several leukemia types. In the NCI-Thesaurus [12] we found definitions in natural language for all classes of our sample. It offers criteria of clarity, consistency, coherence, and extensibility [25,26].

Some classes were so complex in definition that our first attempt to classify them resulted in circularity. With respect to technical terms, it was necessary to clarify their meaning. For example, to understand how cells suffer changes, one should consider three types of mutation: translocation, deletion or inversion. This requires a deep knowledge of the domain in order to establish the genus and the differentia.

As leukemia is a clonal disease, meaning the lineage could be either myeloid or lymphoid, we defined the myeloid inheritance according to well-known classification criteria [27,28,29]. However, there are types of acute leukemia that have both myeloid and lymphoid lineages. Thus, these cases are considered mixed, hybrid or biphenotypic cases. Nevertheless, we classified these cases as subtypes of AML derived from a myeloid lineage following the FAB classification.

We applied ontological principles in the formulation of definitions in order to test the proposed method. After testing, we noticed issues regarding the methodological steps and validation by the domain expert. For example, we realized that to obtain definitions of leukemia from only one dictionary was not enough. So, we resorted to additional sources such as pathology and hematology textbooks [19,30], leukemia classifications [27,28,29] and scientific papers.

All selected sources had general definitions of leukemia, but we detected categorization issues described by Seppälä and Ruttenberg [4]: circular and intangible definitions, use of technical terms and multiple definitions for the same term. In order to soften these issues, we decided in some case to explain the meaning of other terms, for example, genetic mutations, then creating more definitions. In order to understand leukemia, we conclude that one must explore relationships to areas of pathology, cancer diagnosis, etiology, and so forth.

As mentioned previously, the need for guidance from a domain expert is noteworthy. The definitions presented to the expert for validation and her observations were used both to amend the definitions and to review the method. The crucial aspect of the validation process was finding the essence of acute myeloid leukemia classes. The process was conducted through personal interviews with a pediatric oncologist, who employed her experience to confirm the essential characteristics and to determine necessary and sufficient conditions. Our case study relied on only one expert, but we are certain that a true validation should consider several specialists.

The essence of entities was based on the diagnosis criteria of the FAB classification [27] and the WHO [28,29]. The essence was mainly defined by morphological characteristics except in the case of the class "*acute myeloid leukemia with recurrent genetic abnormalities*". In this particular case, the essence was based on cytogenetic abnormalities. The WHO [29] decided to maintain cytogenetic abnormalities as the main characteristics for this class. Therefore, when reviewing the necessary and sufficient conditions, we recognized the influence of leukemia classifications issues. For example, in the just mentioned class "*acute myeloid leukemia with recurrent genetic abnormalities*" the requisite characteristic is that every AML was a carrier of a genetic mutation.

Cancers, especially leukemia, are complex diseases. A single morphological characteristic – for example, the presence of a percentage of blasts – is not enough for disease diagnosis and treatment. This fact confirms the relevance of defining a well-founded and robust formal vocabulary to represent entities in the leukemia field.

VI. FINAL REMARKS

This research gathered some principles already present in the literature of formal ontologies to propose a method with the aim of systematizing the process of creating definitions. After our practical case study, we recommend the method be reviewed and improved upon; the terminological complexity of leukemia made the work difficult and laborious.

As previously stated, the main challenge in our case was determining the essence of leukemia´s classes, since the domain approached is one that exhibits more diversity of phenotypic and genetics changes at diagnosis among cancer studies. Even the FAB classification [27] and the WHO [29] have difficulty in categorizing, defining and diagnosing subtypes of AML. Efforts to better categorize myeloid neoplasm exist, as pointed out by Varian, Harry and Branning [28]. Varian et al. [29] published an update to the WHO classification proposing the use of both morphology and immunophenotyping information to define leukemia, as well as catechetical, genetic and clinical characteristics.

With this experience in mind, we intend future research to further contribute to a theory of formulating definitions of ontologies as well the standardization of definitions in the field of cancer. In doing so, we contribute to the field of biomedical ontologies and healthcare. The complete results and findings of this work can be found on the thesis [33].

ACKNOWLEDGMENT

A. D. S. thanks to Dr. Joaquim Caetano Aguirre Neto, Md, that provided support in the textual definition validation, as well as to Dr. Dagobert Soergel, PhD, for his useful advices.

REFERENCES

[1] D. Soergel. “The rise of ontologies or the reinvention of classification”. J Am Soc Inf Sci.,vol. 50, n.12, pp.1119-20.1999.

[2] B.C. Vickery. “Ontologies”. Journal of Information Science, vol.23, n.4,pp.277-286. 1997.http://mba.eci.ufmg.br/downloads/recol/277.pdf

[3] B. Smith, , M. Ashburner, C. Rosse, J.Bard, W. Bug, W.Ceusters, L.J Goldberg, K.Eilbeck, A.Ireland, C.J Mungall, N.Leontis, P.Rocca-Serra, A.Ruttenberg, S.-A.Sansone, R.H Scheuermann, N.Shah, P. L Whetzel and S. Lewis. “The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration”.Nat Biotechnol. vol.25, n.11, pp.1251-5.nov. 2007.

[4] S. Seppälä; R. Ruttenberg. “Survey on Defining Practices in Ontologies : Report. in preparation of the International Workshop on Definitions in Ontologies”. International Conference On Biomedical Ontology (ICBO 2013), Montreal, Canada, http://definitionsinontologies.weebly.com/

[5] M. B.Almeida, A. B. Proietti, B. Smith, and J. Ai. “The Blood Ontology: an ontology in the domain of hematology”. [ICBO 2011; Buffalo, USA].

[6] A. Gupta. “Definitions”. The Stanford Encyclopedia of Philosophy (Winter 2008 Edition), Edward N. Zalta , Ed., UR. The Metaphysics Research Lab :Stanford. http://plato.stanford.edu/entries/definitions/.

[7] N. Swartz. “Definitions, Dictionaries, and Meanings”. This revision: November 8, 2010. "http://www.sfu.ca/~swartz/definitions.htm.

[8] G. Tsatsaronis, A. Petrova, M. Kissa, Y. Ma, F.Distel, F.Baader, and M. Schroeder. “Learning Formal Definitions for Biomedical Concepts”. [Srinivas, K;Jupp, S. (eds) Proceedings of the 10th OWL: Experiences and Directions Workshop (OWLED 2013), May 2013].https://ddll.inf.tu-dresden.de/web/LATPub509/en.

[9] J. Köhler, K. Munn, A. Ruegg, A Skusa,and B. Smith. “Quality control for terms and definitions in ontologies and taxonomies”. BMC Bioinformatics., vol.7, pp. 212. Apr. 2006.

[10] T.R.Gruber. “Toward Principles for the Design of Ontologies Used for Knowledge Sharing.” 1993. Stanford Knowledge Systems Laboratory. http://citeseerx.ist.psu.edu.

[11] B. Smith. “Introduction to the Logic of Definitions”. [International Workshop on definitions In Ontologie, DO 2013, July 7, Montreal, 2013]. http://ceur-ws.org/Vol-1061/Paper5_DO2013.pdf.

[12] NCI – Thesaurus. http://ncit.nci.nih.gov/

[13] National Cancer Institute. “NCI Dictionary of Cancer Terms: acute myeloid leukemia”. http://www.cancer.gov/

[14] MeSH. http://www.ncbi.nlm.nih.gov/mesh

[15] Medscape. http://www.medscape.com/

[16] Ontobee. “Leukemias”. 2014. http://www.ontobee.org.

[17] Disease Ontology (DO). “Leukemia”.2014.http://disease-ontology.org/.

[18] The Gene Ontology. http://www.geneontology.org/.

[19] K. Reichard, Wilson, Czuchlewski, Vasef, Zhang, and Hunt ."Myeloid neoplasm". In: Diagnostic pathology blood and bone marrow. Monitoba: Amirsys, 2012. cap.9.pp. 2-208.

[20] J.Michael, J.L. Mejino Junior, and C. Rosse. “The role of definitions in biomedical concept representation”.Proc AMIA Symp. pp. 463-7.2001.

[21] B.Smith, W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A,L,Rector, and C. Rosse. “Relations in biomedical ontologies”. Genome biology . vol.6, n.5, p.p.R46.2005.

[22] D. Soergel. “Knowledge Organization Systems: Overview”. 2014. http://www.dsoergel.com/SoergelKOSOverview.pdf.

[23] A. Petrova, Y. Ma, G.Tsatsaronis, M.Kissa,F.Distel,F.Baader, and M. Schroeder. “Formalizing biomedical concepts from textual definitions”.J Biomed Semantics., vol.6, pp.22. April. 2015.

[24] F. M. Mendonça, K.C. Cardoso, A. Q. Andrade and M.B. Almeida. “Knowledge Acquisition in the construction of ontologies: a case study in the domain of hematology”. Proceedings of the International Conference of Biomedical Ontologies (ICBO), 2012, Austria.

[25] M. Uschold. “Building Ontologies: Towards a Unified Methodology (1996)”. http://citeseerx.ist.psu.edu/

[26] T. B. Gruber. “A translation approach to ortable ontologies”. Knowledge Acquisition, vol. 5, n. 2, p. 199-220, 1993. http://www.dbis.informatik.hu- berlin.de/.

[27] National Cancer Institute. “FAB”. http://www.cancer.org/cancer/.

[28] J. W. Vardiman, N.L . Harris, and R.D . Brunning. “The World Health Organization (WHO) classification of the myeloid neoplasms”.Blood. vol.100, n.7, pp.2292-302. October. 2002.

[29] J.W Vardiman, J. Thiele,D.A.Arber, R.D. Brunning, M.J. Borowitz, A. Porwit, N. L.Harris, M. M. Le Beau, E. Hellstrom-Lindberg, A. Tefferi, and C. D. Bloomfield. “The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes”. Blood. vol.114, n.5, p.937-51. July. 2009.

[30] R. Hoffman, E.J. Benz Jr, L.E. Silberstein, H.Heslop, J. Weitz, and J. Anastasi. “Hematology: Basic Principles and Practice”. 5th Ed. New York: Churchill Livingstone,2008.pp.2560.

[31] C.Rosse, J.L.V.Mejino Junior. “A reference ontology for biomedical informatics: the foundational model for anatomy”.Journal of Biomedical Informatics. vol.36, pp.478-500.

[32] S.Seppälä, Y. Schreiber and A. Ruttenberg. “Textual and logical definitions in ontologies”. CEUR Workshop Proceedings, vol.1309, Houston, TX, USA, October 6-7, pp. 35-41.

[33] A.D.Souza. “Systematizing of the Methodology of Creating Formal Definitions in Biomedical Ontologies: an Investigation in Acute Myeloid Leukemia Domain”.[Thesis] Master in Information science. School of information Science, Federal University of Minas Gerais, Brazil, 2016. (in Portuguese)