Representation of Biomedical Expertise in Ontologies: a Case Study about Knowledge Acquisition on HTLV viruses and their clinical manifestations

Kátia Cardoso Coelho[ac], Maurício Barcellos Almeida[b]

[a] *PhD Candidate, Federal University of Minas Gerais, Belo Horizonte, Brazil*
[b] *Associate Professor, Federal University of Minas Gerais, Brazil*
[c] *Hemominas Foundation, Belo Horizonte, Brazil*

Abstract

The process of obtaining knowledge from human experts – the so-called knowledge acquisition process – is a crucial activity in the construction of ontologies. In this article, we develop a set of methodological steps for knowledge acquisition and then apply them to the organization of biomedical information through ontologies. Those steps are tested in a real case of knowledge acquisition involving Human T Cell Lymphotropic Virus (HTLV), which causes myriad infectious diseases, as well as its clinical manifestations. Finally, we present results of the application of our methodology. We hope to contribute to the improvement of knowledge acquisition in ontologies with the aim of providing suitable knowledge representation of scientific domains.

Keywords:

Knowledge acquisition, Ontologies, Knowledge representation.

Introduction

The study of how knowledge produced by people and by groups of people can be converted in knowledge of a specialized scientific field is essential within the scope of scientific investigations. Historically, Information Science has studied the best ways to organize and represent knowledge for information retrieval [1]. However, the tasks involved in organizing and representing knowledge are not trivial, because of issues in communication, difficulties in comprehending scientific terminologies, to mention but a few.

This article investigates the activity of Knowledge Acquisition (KA) within the scope of biomedicine. In order to improve that activity, we propose procedures for knowledge acquisition, which adhere to some of the best practices found 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 was conducted within the scope of a biomedical project focused on the human blood domain. Results of the aforementioned list of steps to KA have been used for the development of a knowledge base, which aims scientific and educational applications related to Associated Myelopathy/Tropical Spastic Paraparesis I (HTLV-I)

Knowledge Acquisition: a multidisciplinary approach

Despite its importance in the context of information organization, the activity of KA is not a trivial matter since there is no reliable methodology for representing the cognitive processes of the agents involved [2]. Issues arise from both experts and from information professionals because of: i) lack of a suitable expression by experts, who do not always understand what is requested, what level of detail is required, how to present ideas logically and how to explain the jargon of their field; ii) the information professional's difficulty in understanding and recording what experts say, in maintaining concentration and comprehending new knowledge [3].

Different KA techniques have been proposed for acquiring knowledge from people or from documents. Whatever the approach selected, consulting experts is required every time one intends to organize information and knowledge of a research field [4]. Even though knowledge acquisition based on documented sources is also important for knowledge representation, the present paper emphasizes knowledge acquisition from people.

An overview of Knowledge Acquisition

Even though it appears under different denominations, the KA process is identified by a common concern present in several scientific fields – Knowledge Management, Computer Science, Librarianship and Information Science, to mention but a few – in capturing specialized knowledge for the aim of representation. All those fields have in common the difficulty in eliciting knowledge retained by an expert.

KA is a term employed since the 1980s to refer to the study of how people's expertise can be represented in computational systems [5, 3, and 6]. In the Knowledge Management, this sort of activity was, in the nineties, integrated into a set of pioneering strategies that aimed to capture individual knowledge and convert it into organizational knowledge [7]. In the development of ontologies, KA consists of a stage in the knowledge representation [8]. Within Librarianship and Information Science, KA activities are employed, for example, in the construction of controlled vocabularies to represent documents content in information retrieval systems. In that later context, KA also occurs during the interaction between librarians and experts as a mean of terminology endorsement, as a form of prospecting new terms, as well as confirming their effective use [9].

Whatever the context, the KA activity generally includes collecting, analyzing, structuring and validating of knowledge for representation purposes [3]. It is an activity composed of a set of tasks that employ computer-based and manual techniques [5-10, 6, 11-12]. A multitude of definitions for KA can be found in the literature

The theories and methods that support KA activities rely on diverse academic research fields, as well as on practice [18]. Ways of acquiring, representing and verifying knowledge come from Computer Science, Cognitive Science, Linguistics, Semiotics and Psychology. Each one of these research fields has contributed to the comprehension of KA.

In *Computing Science*, examples are the pioneering works of Newell & Simon and Compton & Jansen [19-20]. From this perspective, the ability to acquire and represent knowledge in a computer-readable format is obtained through the *physical*

symbol hypothesis [21]. Such hypothesis postulates that knowledge would be constituted of symbols that represent reality. Thus, intelligence would correspond to the ability to logically manipulate symbols and relationships between them.

In *Psychology*, the basis of the KA process can be found in seminal works like the *Personal Constructs Theory* [22]. In that context, the transference of knowledge constitutes the foundation for KA: people transfer expertise so that others may be able of replicating their performance. In *Cognitive Science*, the same line of thought can be found in the *Hawkins model,* which identifies the transference of *expertise* as a means of knowledge elicitation [23].

Semiotics contributes with the triad perspective exemplified by Ogden-Richards. That triad is cited by Campbell as a theoretical base for KA [24-25]. Semiotics, roughly speaking, consists of the "study of signs" and how the meaning of these signs is understood individually or by a group of people [2]. The semiotic triad is represented by a triangle in which the vertexes are: i) *symbols, which* are specializations of signs and, indeed, conventions used to represent an object or entity; ii) *referents*, which are the objects or entities of reality themselves; and iii) *references*, which are representations of the understanding of an agent using knowledge.

In the field of *Linguistics* there are initiatives for the extraction of knowledge by analyzing collections of texts. The analysis of the underlying language is one of the approaches to KA. This theoretical basis was described by Harris in his work on the nature of the use of language in highly specialized domains characterized by a regular and reproducible grammatical structure [26]. Patterns can be discovered through the application of recognition patterns, either manual or automatic, to a large linguistic corpus extracted from a knowledge domain.

It is worth noting that studies of KA make references to the existence of various types of knowledge that can be represented [4, 18, 27]. However, it is beyond the scope of this paper to list classifications for kinds of knowledge.

The several KA approaches put forward so far emphasize the complexity of the activity, the different agents involved, and the need of selecting suitable methods according to the context.

Materials and Methods

The methodological steps are developed from a comprehensive literature review and tested in a real case of knowledge acquisition about HTLV, which causes infectious diseases, as well as its clinical manifestations.

We present here a case study in which medical knowledge is acquired and validated systematically. The remainder of this section details our ongoing research and outlines the list of steps for KA as an attempt to systematize the process in biomedicine.

Case study: context and domain

The project is taking place in a medical public institution responsible for hematology and blood transfusion research and that offers healthcare services for a population of around 16 million people in Minas Gerais, Brazil. The institution´s experts that participate in the project are members of the Interdisciplinary HTLV Research Group (GIPH). Knowledge about the infection pathogenesis by HTLV is recent, even considering that the virus is endemic in several regions of the world. In general, genetic and immunological factors are the cause of the associated clinical manifestations, which may be divided into three main categories: neoplastic, inflammatory and infectious. HTLV-Associated Myelopathy / Tropical Spastic Paraparesis (HAM/TSP) and Adult T-cell leukemia/lymphoma (ATL) stand out as the first diseases associated with the virus [28]. Indeed, various diseases have been related to the virus, but there is no ordered initiative for recording epidemiological, physio-pathological and therapeutic knowledge about it.

Methodological steps for KA

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: *survey, elicitation, validation and refinement*.
The first phase is the *survey phase, which* consists of activities performed before the interview with experts. The goal of that activity is to obtain candidate terms for the construction of the ontology, and then the first task would to figure out the *scope proposed for the ontology* under construction. Once the scope is known, the second task is to *obtain basic concepts of the domain* under study. In order to do this, the information professionals involved analysis in the specialized literature and material supplied by the medical research group. This task provides familiarity with the subject. The third task in the survey phase is to identify the *expertise* of the physicians involved by capturing data such as: education, main activities, main expertise, main research interests, articles published, to mention but a few.

The second phase of KA, which is called *elicitation phase*, consists of holding interviews and applying KA techniques to experts, physicians, biologists and researchers. During the course of the *interviews,* sorting and matrix techniques are applied. The cycle that characterizes the clinical process, ranging from the development of an infectious disease through its treatment, is adopted to guide the approach taken with experts. The three major stages that comprise that cycle, which are depicted in figures 1, 2 and 3, are: *etiological process, course of disease* and *therapeutic response. We provide a brief description of each of those stages.*

In the stage called *etiological process* (FIG.1), one should consider that there is a healthy human body with characteristics that seems to be normal according to medical parameters. In the pre-clinical manifestation of a disease, the body develops disorders, which are bearers of dispositions. Dispositions are naturally associated with the course of entities' existence, for example, the disposition of the human body to get sick, the disposition of fruit to ripen, and so forth [37]. A patient may have already noticed changes in his/her organism even though there are no signs or symptoms yet.

Figure 1: etiological process, disorders and disease (disposition)

The *course of disease* phase starts with the clinical manifestation of a disease (Fig. 2). At this moment, the disorder manifests itself through symptoms, which the patient is able to identify. Then, a physician identifies the disease signs through a physical exam. In this phase, it is possible to determine a clini-

cal phenotype, that is, the main observable characteristics of that disease.

Figure 2: disease, pathological process and abnormal conditions

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 (see FIG.3).

Figure 3: Signs, symptoms and interpretative process

In order to apply the described rationale, a template was created in Protégé-Frames[1]. Protégé is a software package created in the 1990s for biomedical KA (FIG.4).

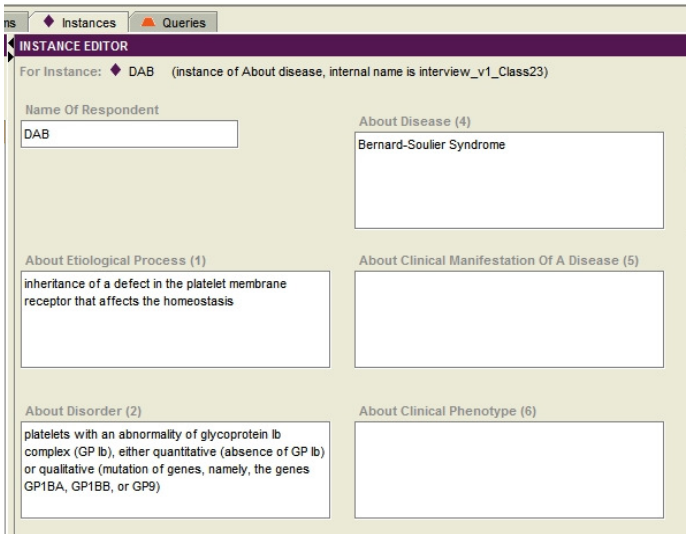

Figure 4: data on blood disease in Protégé-Frames template example

The third stage of the proposed list of steps, called the *validation phase,* make use of *wiki science* tools for collaborative validation of the ontology candidate terms. After elicitation phase, according to the knowledge obtained, candidate terms are transferred to a wiki in order to be validated by experts online. Fig. 5 shows a *template* created using the semantic wiki[2] with the validation done.

[1] Available at <http://protege.stanford.edu/>. Access on September, 12- 2014

[2] Available at <http://mbaserver.eci.ufmg.br/BLO-wiki/>. Access on September, 14 2014

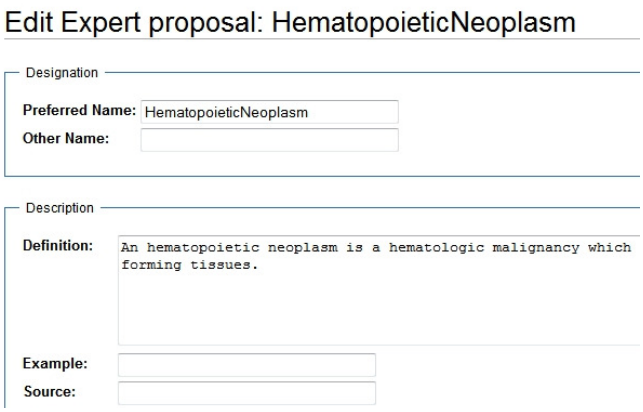

Figure 5: screenshot of BLO-WikiKnowledge Acquisition Environment example

The fourth stage of the proposed list of topics, called *refinement phase*, uses a second *template*, also created using *Protégé-Frames* (depicted in FIG.6). The goal of that template is to record information about how to integrate the different levels of granularity required in understanding a disease and its manifestations. That integration involves obtaining the relationships between parts of the body affected by certain diseases, the related genes and the related proteins. Following the approach, it is possible to characterize both the disease and the processes involved, as well as to foster interoperability. Interoperability is favored because of links created between the ontology under construction and foundational international ontologies, such as the *Gene Ontology*, the *Protein Ontology*, the *Foundational Model of Anatomy*, and so forth.

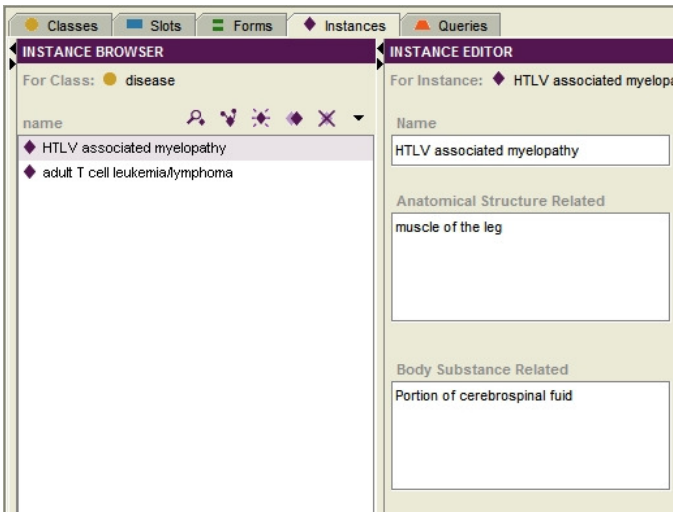

Figure 6: Protégé template with data for HAM/TSP, in refinement phase

Finally, the steps put forward so far are gathered together, thus creating the list of steps for KA (depicted in TAB. 1). Considering the lack of systematic approaches to KA in biomedicine, the list of steps may be useful in other medical fields related to infectious diseases.

Table 1: KA Script proposed for the Biomedicine domain

Phase	Task Objective	Description of Task	Instrument used
(1) Survey	1.1 know the context	Figure out the scope of ontology	Project data
	1.2 know the foundation	Obtain basic concepts of	Basic literature of the

		domain	area
	1.3 identify expertise	Identify the *expertise* of the experts involved	Researcher directories
(2) Contact	2.1 obtain knowledge	Hold Interviews with experts	*Template Protégé-Frames*
	2.2 know the terminology	Identify information organization issues	Matrix Techniques
	2.3 see *ad-hoc* organization	Understand how experts order concepts	Sorting techniques
(3) Validity	3.1 validate knowledge	Obtain approval on terms and definitions acquires	Semantic wiki page
	3.2 update	Update data after each validation	Semantic wiki Page
(4) Refinement	4.1 integration of granularities	Characterize related genes, proteins, body parts.	*Template Protégé-Frames*
	4.2 connection with top-level ontologies	Connect diverse granularities through top ontologies	*Template Protégé-Frames*

Results

After the preliminary organization of the terms, the results were presented to the main researcher, in order that she could validate them. In this step, the expert had to accept what is presented or suggest changes, which will be recorded for new future evaluations. The aim here was mainly to check if such terms, as proposed in the structure representing the HAM domain, corresponds to what was presented by the domain expert during the elicitation phase. Table 2 depicts the final set of terms after validation of the expert.

Table 2: validate knowledge by coordinator GIPH group

Etiological process	HTLV-1 infection
produces	
Disorder	Proliferation of provirus and viral particles in T cells, with production of viral substances like TAX, REX, ENV and other proteins.
bears	
Disposition	Human T-cell lymphotropic virus type 1(HTLV-1) / Tropical Spastic Paraparesis (HAM/TSP)
realized in	
Pathological process	Tax protein action in CD4 positive cells and CD8, dendritic cells and producing chronic inflammatory brain cells that affects the central nervous system leading to an irreversible degeneration of central nervous system cells
produces	
Abnormal features body	Atrophy of the thoracic spine with leptomeningeal thickening and spinal cord atrophy
recognized as	
Symptoms	spasticity; weakness of the lower limbs; bladder disorder; constipation; impotence; decreased libido; sensory symptoms (tingling, stinging and burning); low back pain radiating from the lower limbs. dysuria; intestinal disorders; trouble up and down stairs; alteration of reflexes; changing thermal sensations ocular changes; depression; difficulty walking; numbness.
Signs	Decreased vibratory sense; hyperreflexia of the lower limbs; hyperreflexia of the upper limbs; Hoffmann signs and positive Tromer; exalted mandibular reflex; clonus; Babinks positive signal. Peeling skin; alteration of reflexes. Increased reflexes; urinary incontinence; altered gait; spastic gait; hyperreflexia (increased spinal reflexes); altered gait. Bones and tendons (patellar) and aquiliana hyperreflexia; changes in sensitivity to the touch; altered sensitivity to pain; weakness in various muscle groups of the lower limbs.
used in	
Interpretative process	
produces	
Results	Diagnostic that the patient X has a neurological disorder characterized as a myelopathy and positive serology for HTLV, known as HTLV Associated Myelopathy - HAM /Tropical Spastic Paraparesis -TSP.

Discussion

The KA activity from experts as part of the process for developing ontologies can also be understood as a preliminary activity before automatic term extraction. It is required to have specialists to judge whether the extracted terms make sense in the domain. The biomedical vocabulary we come up with also has a relevant function: consensually define the meaning of terms used in medical practice and research. This is made possible precisely by considering directly knowledge acquisition from experts.

Conclusion

We hope that discussions presented in this paper contribute and enriches the search for new ways to aid both knowledge engineers and professionals involved with KA activity. Thus, we hope to make their work more effective and also help own expert's domain as physicians, biologists, biochemists, in their work.

However, some questions still remain open. It is worth mentioning, for example: can the use of collaborative tools such wikis allow the effective participation of all experts in the activity? Can the use of a wiki be considered an element that facilitates the directly creation of records? Should the expert take notes alone or together with information professional? How identify the validation occurred in practice? And whether there are questions by experts that required them to take notes? Or, perhaps, should the information professional write down what actually is obtained from the experts and record it?

The activity of knowledge acquisition, as shown in this paper, is a stage of construction of ontology in a given domain. This stage includes tasks of organization, delivery and sharing of knowledge, via ontologies, including domain specialists.

Acknowledgments

Work supported by the *Hemominas* Foundation, *Minas Gerais*, Brazil and Interdisciplinary HTLV Research Group (GIPH). This work is also partially supported by *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG), *Governo do Estado de Minas Gerais*, *Rua Raul Pompéia*, nº101, Belo Horizonte, MG, 30.330-080, Brazil.

References

[1] Bates MJ. The Invisible Substrate of Information Science. J Am Soc Inf Sci. e 1999; 50:1043-1050. Available from: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(1999)50:12%3C1043::AID-ASI1%3E3.0.CO;2-X/references

[2] Hayes-Roth F, Waterman DA, Lenat DB. Building Expert Systems. Massachusetts: Ed. Addison-Wesley. 1983, 350 p

[3] Milton NR. Knowledge acquisition in practice: a step-by-step guide. Cranfield: Springer, 2007. 176p.

[4] Hua J. Study on Knowledge Acquisition Techniques. Second International Symposium on Intelligent Information Technology Application. 2008. Available from: http://www.computer.org/csdl/proceedings/iita/2008/3497/01/3497a181-abs.html

[5] Milton N, Clarke D, Shadbolt N. Knowledge engineering and psycology: Towards a closer relationship**.** *Int J Hum Comput Stud* 2006; 64:1214-1229. Available from: http://www.sciencedirect.com/science/article/pii/S1071581906001212

[6] Boose JH, Gaines BR. Knowledge Acquisition for Knowledge-Based Systems: Notes on the State-of-the-Art. *Mach Learn* 1989; 4:377-394. Available from: http://www.springerlink.com

[7] Choo CW. The knowing organization: how organizations use information to construct meaning, create knowledge, and make decisions. New York: Oxford, 2006. 368 p.

[8] Fernandez M, Gomez-Perez A, Juristo H. *Methontology*: From Ontological Art Towards Ontological Engineering, 1997. Available from: http://www.aaai.org/

[9] Lancaster FW. Vocabulary Control for Information Retrieval. 2nd ed. Arlington:Information Resources Press. 1986, 233 p.

[10] Gaines BR. Organizational Knowledge Acquisition. In: Handbook on knowledge management: Knowledge matters. Birkhäuser: Springer. 2003, 700 p.

[11] Wolf R, Delugach HS. Knowledge Acquisition via tracked repertory grids. Computer Science Dept. Univ. Alabama in Huntsville, 1996. Available from: www.cs.uah.edu/tech-reports/TR-UAH-CS-1996-02.pdf

[12] Shadbolt N. Eliciting Expertise. In: Evaluation of Human Work. Ed. Taylor & Francis. 2005. Available from: http://eprints.ecs.soton.ac.uk/id/eprint/14563

[13]Corbridge C, Rugg G, Major N, Shadbolt NR, Burton A. Laddering: technique and tool use in knowledge acquisition**.** *J Know Acq* 1994; 6: 315–341. Available from: http://www.sciencedirect.com/science/article/pii/S1042814384710168

[14]Shaw MLG, Gaines BR. Requirements acquisition. Sof. Eng J 1996; 11: 149-165.

[15]Scott AC, Clayton JE, Gibson EL. A practical guide to knowledge acquisition. Addison–Wesle, 1991, 509 p.

[16]Turban E. Expert systems and apllied artificial intelligence. Ney York: Macmillan Publishing Company, 1992. 804 p.

[17] Regoczei SB, Hirst G. Knowledge and knowledge acquisition in the computational context. In: The psychology of expertise.New York:Springer-Verlag, 1994, p.12-25.

[18]Payne PR, Mendonça EA, Johnson SB et al. Conceptual knowledge acquisition in biomedicine: a methodological review. J Biomed Inform 2007; 40: 82–602. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17482521

[19]Newell A, Simon HA. Computer science as empirical inquiry: symbols and search. 1976. Communications of the ACM 1976. Available from: http://www.cs.utexas.edu/

[20]Compton P, Jansen R. A philosophical basis for knowledge acquisition. Knowledge Acquisition. European knowledge acquisition for knowledge based systems. 1989. Available from: http://citeseerx.ist.psu.edu/viewdoc

[21] Nilsson N. The Physical Symbol System Hypothesis: Status and Prospects. 2007 Available from: http://ai.stanford.edu/~nilsson/OnlinePubs-Nils/PublishedPapers/pssh.pdf

[22]Kelly GA. The psychology of personal constructs. New York: Norton, 1955.

[23]Hawkins D. An analysis of expert thinking. *Int J Man Mach Stud* 1983; 18: 1-47, Jan. Available from: http://www.sciencedirect.com/ science/ journal/00207373

[24] Ogden CK, Richards IA. The meaning of meaning. San Diego: Harcourt Brace Jovanovich, 1989. 363p

[25]Campbell KE, Oliver DE, Spackman KA, et al. Representing thoughts, words, and things in the UMLS. *J Am Med Inform Assoc* 1998; 5: 421–31. Available from: http://www.ncbi.nlm.nih.gov

[26] Harris Z. On a theory of Language. J Philos 1976; 73: 253-276. Available from: http://www.jstor.org/stable/2025530

[27] Gandon F. Distributed artificial intelligence and knowledge management: ontologies and multi-agent systems for a corporate semantic web. 2002. 483f. – Doctoral School of Sciences and Technologies of Information and Communication. INRIA and University of Nice, Nice, 2002. Available from: http://www-sop.inria.fr/members/Fabien.Gandon/docs/PhD_FabienGandon.pdf

[28] Romanelli LC, Caramelli P, Proietti AB. Human T cell lymphotropic vírus (HTLV-1): when to suspect infection? Rev Assoc Med Bra 2010; 56:340-7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20676544

Address for correspondence

Kátia C.Coelho email katiaccoelho@gmail.com