Analysis and Representation of Illocutions from Electronic Health Records

Julio Cesar dos Reis[1], Rodrigo Bonacin[2,3],
Edemar Mendes Perciani[2], and Maria Cecília Calani Baranauskas[1]

[1] Institute of Computing, University of Campinas, Campinas, Brazil
[2] Faculty of Campo Limpo Paulista, Campo Limpo Paulista, Brazil
[3] Center for Information Tecnology Renato Archer, Campinas, Brazil
{julio.dosreis, cecilia}@ic.unicamp.br,
rodrigo.bonacin@cti.gov.br, edemar.mendes.perciani@gmail.com

Abstract. Electronic Health Records (EHRs) store multiple patients' information, including medical history, diagnoses and treatments. Computer-interpretable representation of meanings and intentions in EHRs content might play a major role for decision making, as well as for medical system integration and information recovery. However, there is a lack of suitable representation models to specify the relations between semantic models and illocutions, which reflect the intentions of medical content producers. In this paper, we propose an analysis to understand how illocutions are expressed in EHRs. We aim to identify domain-specific terms to convey the different dimensions in which illocutions are classified. Furthermore, this research develops a model, based on Web ontology description languages, to encode and instantiate the illocutions in the medical domain. Obtained results point out that some illocution types and associated terms are predominant in the analyzed content. They highlight the potentiality of our model to explore illocutions in several computing tasks.

Keywords: Intentions, Illocutions, Pragmatics, Ontologies, Semantic Web, Pragmatic Web, Knowledge Representation, EHRs, Medical data

1 Introduction

Information and communication technologies are essential in complex contexts of medicine and clinical research. These domains are knowledge intensive and require appropriated methods and artifacts for enabling computational representation of knowledge. In this context, Electronic Health Records (EHRs) describe and store patients' information with a great volume of unstructured text. This hinders adequate integration, retrieval of information, as well as access to medical research data.

Aiming to deal with this problem, various studies have devised how to give well-defined meaning to information [1], emphasizing the construction of mechanisms for interpreting digital content through knowledge representation artifacts [2]. However, the communicated intentions affect the understanding of information content.

We understand "intention" as the effects of meanings with a determined purpose in a social context (e.g., a feeling or a judgment), as aligned with the discipline of

Organizational Semiotics (OS) [3]. We argue that dealing with both semantic and intentions (as part of pragmatics[1]) allow to improve medical knowledge management and sharing. Our assumption implies having an underlying model suited to formally represent and link semantic and pragmatic aspects of knowledge.

This model might support both human and machines to interpret and process information. Our research remains under the Semantic Web (SemWeb) vision, which suggests modeling the information in Web ontologies for enabling knowledge interpretation by artificial agents and people [1]. While considering the formal meaning of information may already improve several types of systems, in particular machine-to-machine interoperability, complex settings of medical information systems entail open issues related to human factors. Even simple concept representations may have their interpretation affected by contextual aspects such as intention, users' background and time. Complementarily, the Pragmatic Web (PragWeb) studies meaning negotiation [4], however, there are still open issues on knowledge representation techniques suited to model meanings in a social context with the implied intentions.

Although literature has examined the evolution of meanings and intentions in collaborative problem solving [5], several difficulties still exist to achieve the representation of a complete computer-interpretable model, including alternatives to define the way that intentions are manifested in domain-specific contexts. Moreover, there is a lack of experimentally grounded studies with the focus on investigations of how existing theoretical frameworks can be useful to deal with this issue.

This article proposes an original empirical analysis of dimensions and intention classes in real EHRs data. This investigation relies on Semiotic-based theories and frameworks that structure and classify intentions according to different dimensions of illocutions [3]. The work also contributes with the definition and refinement of an ontology model, specified with SemWeb technologies. The goal of this model is to represent the illocutions materialized in the domain.

This research collected and selected a set of EHRs, which were manually examined to classify the illocution dimensions in their content. On this basis, we performed quantitative and qualitative analyses, which were used as input to the design of a representation model. Our findings reveal the way that illocutions are expressed in the domain-specific text-based content of the medical field. This allows the extraction of commonly used domain terms to represent the illocution dimensions. The proposed model enables to formally explore intentions, in a structured way, in analytics tasks, as well as by information recovery systems. Our investigation demonstrates the potentiality of considering domain-specific terminologies for identifying and classifying illocutions, which represent relevant language expressions of the domain.

The remainder of the paper is organized as follows: Section 2 reviews methods for representing meanings and intentions, and describes the adopted theories, frameworks and technical languages; Section 3 describes the methods and materials of the study; Section 4 presents and discusses the obtained results; Section 5 draws conclusion remarks and points out future work.

[1] In the context of this work, pragmatics is understood as described in the pragmatic layer of the semiotic framework [3].

2 Background

The Web has progressively evolved towards the SemWeb standing for an extension of the current Web that enables richer information share [1]. The SemWeb aims at making data more accessible and detectable by people and machines. SemWeb has predominantly focused on: (1) turning data machine-interpretable and making the semantic of information explicit at different degrees of expressivity, via formal representations; (2) providing metadata; and (3) integrating well-structured data.

Within the SemWeb proposition, ontology consists of a concrete syntactic structure that models the semantics of a domain [6]. In this paper, we adopt the term "Web Ontology" to refer to ontologies within the SemWeb field. This concept differs from the ontology concept used in OS field. Web Ontologies have specifically been designed to provide rich machine-decidable semantic representations and refers to a formal specification of some domain interpretable by machines. It specifies a conceptualization in terms of classes of domain objects, properties and relationships between classes [7]. This enables knowledge interpretation by artificial agents supporting the correct understanding of shared data. At the core of the SemWeb technology, computational languages, based on logic for knowledge representation and inference, have been designed specifically to define Web ontologies. In particular, according to the SemWeb architecture [1], there exists the *Web Ontology Language* (OWL), which relates to other Web languages, such as *Resource Description Framework* (RDF).

While semantics concerns the study of the meaning, independent of use and context, pragmatics regards the study of the meaning use in context and its purpose. In this perspective, PragWeb has originally been proposed as an extension or a complement of the SemWeb. PragWeb addresses shortcomings and challenges that purely SemWeb approaches fail to tackle with the aim of serving user's needs by making content more accessible. PragWeb emphasizes the relevance of context and purpose of information. Thus, it deals with research issues such as context and meaning negotiation between agents (human or artificial) [4] and issues related to intentions, interests and participation.

Few studies have attempted to represent and recognize intentions and other pragmatic aspects in computer systems. Our investigation indicates studies in natural language processing and computational linguistics that address pragmatic aspects. In the context of discourse analysis, Poesio and Traum [8] have studied a discourse model and different kinds of structure that play a role in conversation. They proposed a theory about the discourse situation, shared by the participants in a conversation, centered on information about the occurrence of speech acts [9]. Dam-Jensen and Zethsen [10] have conducted a linguistic corpus analysis considering pragmatic aspects via patterns. They investigated the relations between lexical meaning and the context where these meanings are inserted.

In addition to modeling pragmatics at the level of natural language, existing researches also emphasize the intentions modeling in other frames. The initial propositions examined logic-based models of intention formation applied to the multi-agent task [11]. They focused on the evolving intentions of agents and the conditions under which an agent can adopt and keep an intention. In contrast, Hawizy et al. [12] argue that the design of a model to produce clear representations of human intentions

requires the incorporation of communication studies, such as Semiotics, which encompasses verbal and non-verbal communication.

Some studies have focused on specific domains, e.g., health. Shahar et al. [13] have studied the representation of clinical guidelines, where intentions referred to action patterns or patient states that a system must maintain, achieve or avoid. Other investigations aim at analyzing the users' behaviors in collaborative environments using SemWeb technologies. Kanso et al. [14] present an approach to model intentions by analyzing the authors' acts, focusing on detecting intentions in scientific documents, while Angeletou et al. [15] have proposed a method to represent and compute behavior by inferring roles in online communities. Nevertheless, these behavior roles are not explicitly linked to intentions.

Our previous work proposed a semiotic approach to design ontologies [16]. The investigation adopted OS' concepts and methods to enrich the representation aspects of traditional Web ontologies. In particular, we used the notion of *Agents*, *Affordances* and *Ontological Dependences* from the Semantic Analysis Method (SAM) described as OWL classes. That proposal did not enable representation of intentions, which was initially proposed as a general ontology model to represent pragmatic aspects in a computational way [17]. It encoded the main concepts of the Pragmatics Communication Framework [3] using the OWL language, which included classes, and object and data properties to describe the model.

Although the achieved results already allow correlating the representation of communication acts with ontologies, which demonstrates the initial feasibility of the preliminary model, they have brought up theoretical and practical limitations, which are addressed in the current article. Our research aims to further explore the process of describing the terms by which illocutions are expressed in the domain content. In addition, we contribute with empirical analyses and an extension of the model, encompassing additional classes, terms and instances.

In our study, we adopt an OS view of intentions, which is based on Peircean Semiotics, aiming to represent and study intentions present in EHRs using an *interpretant* dependent communication model. In order to classify and structure types of intentions, this investigation relies on Liu's conceptual framework of Pragmatics Communication [3], which proposes classifying illocutions using three dimensions. Liu's proposal considers communication acts as the minimal unit of human communication. A complete communication act is defined as a structure consisting of three components: performer, addressee, and message. A message contains two parts:

1. The *content* part of a communication act manifests the meaning of the message. The meaning is determined by social construct or human behavior performed by the performer and by the addressee.
2. The *function* part of a communication act specifies the illocution, which corresponds to the intention of the performer.

One dimension distinguishes between descriptive and prescriptive "invention". If an illocution has an inventive or instructive effect, it is prescriptive, otherwise descriptive. Another dimension consists of affective and denotative "mode". If an illocution is related to the performer's personal modal state mood, we call it affective, otherwise denotative. The last one is the "time" dimension, namely past/present and future. The classification of the "time" dimension is based on when the social effects of the message are produced, i.e., in the future or the present/past. The three

dimensions result in eight different classes of illocutions including: *Proposal, Inducement, Forecast, Wish, Palinode, Contrition, Assertion* and *Valuation.*

3 Study Design

In this study, we conducted a five-step procedure to attain our objectives as follows:

1. Collect a set of real patients EHRs from hospitals. To make this research viable, we selected from the initial set of EHRs a subset according to a specific disease diagnosis. We only considered this subset of EHRs in our analysis.
2. Perform manual analysis of the EHRs according to the dimensions of illocutions, as proposed by our theoretical frame of reference. This step was performed by the researchers involved in this work with support of physicians. We assign the illocution dimensions to the sentences (messages) of the EHRs content. For this purpose, we consider *Zero* as past/present (time), description (invention) and denotative (mode); we denote *One* as future (time), prescription (invention) and affective (mode) (cf. Table 1).
3. Execute a quantitative and qualitative analyses over the illocutions. In general, we analyzed the occurrences of the dimensions (cf. Table 1) and the frequency of illocution classes detected (cf. Table 2). Furthermore, we examined the representative terms and keyphrases for each illocution class based on the EHR content (cf. Table 3). This allows us to state the domain terms that frequently represent the illocution dimensions.
4. Define an ontology model (cf. Section 4.2) based on previous analysis. This model represents illocutions related to domain terminologies. To this end, we rely on the reuse of previous models proposing further extensions and refinements.
5. Instantiate the model with the EHRs contents.

This research considers a set of EHRs available in a public hospital from "Águas de Lindóia" in São Paulo State, Brazil. The total amount of EHRs accounts ~10.200 and all patients' data are anonymous. Our manual analysis effectively used 26 cases regarding the diagnosis of "Dengue fever" disease. We considered free-text notations in pre-consultation and patient's history case, where physicians report on symptoms according to patients' statements, exams results, and suggest treatments.

4 Results and Discussion

In this section, we first present the results concerning the conducted analysis of illocutions in EHRs (Section 4.1). Afterwards, we describe a Web Ontology model for representing the expression of illocutions from EHRs content (Section 4.2). The results are then discussed in Section 4.3.

4.1 Analysis of illocutions in EHRs

Table 1 presents the occurrence of the values *Zero* and *One* of each dimension according to defined methods (cf. Section 3.1). The analysis of 26 EHRs resulted in the identification of 201 illocutions. The results point out that around ninety percent of the illocutions are in present/past tense, are descriptive and are denotative. The affective mode is present in less than eight percent of the messages.

Table 1. Distribution of occurrences regarding dimensions of time, invention and mode.

	#Time	#Invention	#Mode
Zero	182 (90.55%)	182 (90.5%)	186 (92.54%)
One	19 (9.45%)	19 (9.45%)	15 (7.46%)

Table 2 presents the frequency of illocution classes in the EHRs' texts. The majority of the illocutions are assertions within 84.58% of the messages. Proposals (7.96%), valuation (5.97%) and inducements (1.49%) classes are also present in the analyzed messages. Nevertheless, forecast, wish, palinode and contrition did not occur in the analyzed EHRs.

Table 2. Frequency of illocution classes

Illocution classes	Frequency (Percentage)
Assertion	170 (84.58%)
Proposal	16 (7.96%)
Valuation	12 (5.97%)
Inducement	3 (1.49%)

Table 3 presents terms and keyphases used in the messages that indicate their illocution classes. The terms "Refers", "Exhibits" and "Reports" are present in the total of 66 assertions. From a qualitative view, typically, these terms were used to confirm patients' symptoms, characteristics or disease, which are important for medical diagnosis. The terms "Denies" and "Lacks" are present in the total of 85 assertions. Physicians frequently used these terms to report the absence of symptoms or diseases related to the diagnosis. Time related keyphases/expressions (e.g., "There is 'x' hours/days") are present in 24 assertions. The time expressions are frequently used to refer to the presence or absence of a symptom or disease in the days or hours prior to the consultation. Other terms are present in 7 assertions (one occurrence each).

As presented in Table 3, the terms "Requests" and "Advices" are present in 15 proposals. Typically, physicians use these terms to give instructions for the patients. The term "Refers", "Complaints" and "Improves" are present in 7 valuations that consist in subjective judgments about the patients' symptoms, characteristics and conditions (e.g., to say that they are feeling better or to complain about a symptom). The term "Denies" is also present in 2 valuations, which refers to subjective judgments about patients' conditions. The term "Advises" was used in 3 inducements made by physicians to warn patients.

Table 3. Analysis of representative terms and keyphases by the illocution classes[2]

Terms	#Assertion	#Proposal	#Valuation	# Inducement
"Refers"	48	-	3	-
"Denies"	82	-	2	-
"Exhibits"	12	-	-	-
"Reports"	6	-	-	-
"Requests"	-	12	-	-
"Advises"	-	3	-	3
"There is '*x*' hours/days"	24	-	2	-
"Lacks"	3	-	-	-
"Complain"	-	-	2	-
"Improves"	-	-	2	-
Others	7	1	2	-

4.2 Web Ontology for representing domain-related illocution terminologies

Our Web Ontology representation is based on previous studies and the *Communication Act Ontology* (*CactO*) [17]. The first version of *CactO* was constructed in OWL using the *Protégé tool*[3] with the objective of representing aspects related to intentions in messages of collaborative problem-solving systems. This Web ontology was evaluated in information retrieval scenarios from discussion forums. Despite the promising results substantiated by overall good objective measures of evaluation, the first version of *CactO* is limited when we consider the complexity of medical texts. Thus, based on our reported analyses, we propose a new version of *CactO*, which we named *MedCactO*.

This version of the *CactO* aims at representing intentions in text from EHRs. In *MedCactO*, the *function* part of a communication act is more detailed, including the specification of dimension values and terms used to express these dimensions. The *MedCactO* also links behaviour patterns to standard medical terminologies and existing *Knowledge Organization Systems* (KOS).

Fig. 1 presents an overview of *MedCactO*, including the classes inherited from *CactO*. The *Agent* class comprises who (*HumanAgent)* performs (or is the addressee) a communication act. *Physicians* and *Patients* might be subclasses of *HumanAgent.* The *Behaviour_Pattern* class represents patterns that delineate the actions performed by an agent (including meaning interpretation). The *communicationAct* is performed when an agent write a text. One *communicationAct* has a message, which has the function and content parts (cf. Section 2).

In *MedCactO*, a *Behaviour_Pattern* is linked to concepts modeled in existing medical KOSs (terminologies, taxonomies, ontologies, *etc*.) (top of Fig. 1). The *MedCactO* also includes the *FunctionAct* class. This class is associated with an illocution type, which has the *Dimensions* of time, invention and mode. Each *Dimension* is described with a value (between 0 and 1) and it is expressed by a *Behaviour_Pattern,* which is linked to terms specified by the existing medical KOSs.

[2] A same illocution can be related to more than one term/keyphase.
[3] protege.stanford.edu

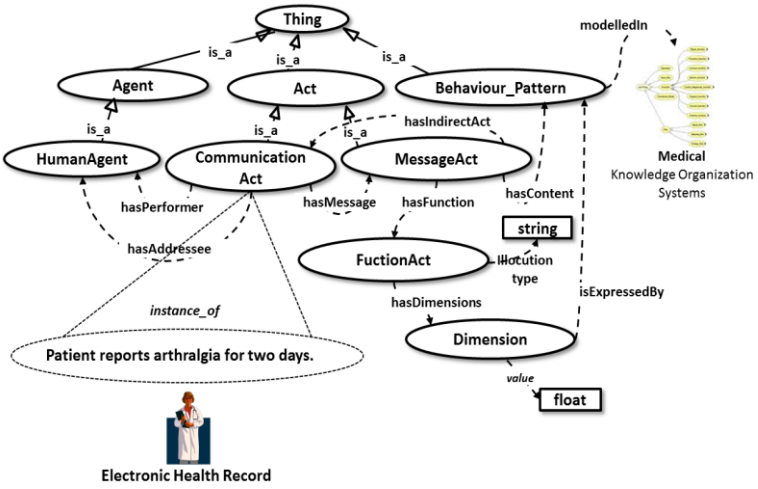

Fig. 1. Overview of the *MedCactO*

Fig. 2 presents an example of communication act modelled according to *MedCactO*. We extracted the following text fragment from the analyzed EHRs to illustrate the Web ontology instantiation: "*Patient reports arthralgia for two days*". As shown in Figure 2, the communication act (*c_act4*) was performed by an agent (*uid_2*) and this act has a message (*m_act4*). Such message has the content part linked to a behaviour pattern (*b_pattern5053*), which is related, for instance, to an UMLS[4] *Concept Unique Identifier* (*C0003862*). The message also contains a function (*illoc_4*), which has an illocution type (*assertion*) and three dimensions (mode and invention dimensions were omitted in the figure for reliability purposes). The "time" dimension (*me_illoc_4*) has the value 0 (i.e., present/past) and is expressed by a behaviour pattern (*b_pattern51*), which is modeled externally in UMLS (for instance) by the concept *C0449238*. The value associated to this concept is *two days*.

4.3 Discussion

The analysis of the EHRs revealed interesting aspects of free-text annotations regarding illocutions in medical records. In general, the analyzed texts are concise and impersonal. This can be observed by the higher incidence of assertions, and the mode 0 (denotative) in more than 90 percent of the messages. The majority of the analyzed texts also remains descriptive and in the present/past tense. These characteristics differ from our previous studies [5] in "special education domain", in which there are a wider range of incidence of other illocutions types, including the affective mode.

[4] *Unified Medical Language System* - Available in <www.nlm.nih.gov/research/umls>

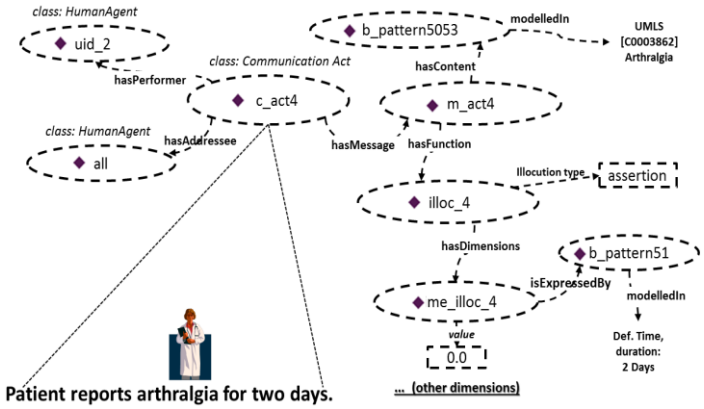

Fig. 2. Example of a communication act modeled using *MedCactO*

The analysis also revealed the importance of the terms or expressions that indicate the dimensions of an illocution. For instance, a *coughing* for one day is a different indicator that a coughing for one month. The modeling and computational interpretation of these aspects may result in improvements of medical information tools, including information retrieval tasks and applications for analytical purposes.

Inspired by these results, we proposed the *MedCactO*, which aims at representing the illocutionary acts and linking them to standard terminology and models of the medical field. This Web ontology was constructed using the OWL language, which enables the use of reasoning tools and other Semantic Web technologies that can be used to develop several integration and information retrieval tools.

Neverhteless, this study is limited in terms of size and scope. It was restricted to 26 EHRs of a specific disease (dengue fever). EHRs of other areas of the medical field must be investigated to verify the incidence of illocutions types. RDF/OWL models, as the adopted in this article, are also limited for representing *Agents*, *Affordances* and *Ontological Dependences*. Other methods from OS, such as, the Semantic and Norm Analysis may produce additional and relevant results. Despite these limitations, this study represents a promising and novel initiative for understanding and managing intentions in text-based content of medical records. In a long-term perspective, this might be useful and effective in the definition of further computational tools for supporting clinical research and better medical treatment plans.

5 Conclusion

Medical information, as text described in EHRs, requires adequate computational representation of meanings and intentions. This might be crucial to several organizational and computer supported tasks, including decision making and medical research. However, literature lacks formal methods and models to relate meanings and intentions systematically. Meaning representation and interpretation cannot be considered without context and intention, which can be classified according to several

dimensions. In this article, we made explicit how domain terms are used in illocutions related to intention classes. We conducted an analysis to investigate how illocutions are expressed in EHRs using domain terms. This was the basis to expand and refine a model, which formally describes illocution types using standard Web ontology languages. Our findings indicated relevant domain-specific expressions that refer to illocution dimensions and their adequate computer-interpretable representation. Future work involves experimentally investigating the use of the proposed model in larger scale and in specific computing applications such as information retrieval.

Acknowledgments. We thank the São Paulo Research Foundation (FAPESP) (Grant #2014/14890-0), and the CNPq (Grant # 308618/2014-9). The opinions expressed in this work do not necessarily reflect those of the funding agencies.

References

1. Berners-Lee T., Hendler, T.J., Lassila, O.: The Semantic Web. Scientific American 284(5), 34–43 (2001)
2. Hendler, J., Berners-Lee, T.: From the Semantic Web to Social Machines: A research challenge for AI on the World Wide Web. In Artificial Intelligence 174(2), 156–161 (2010)
3. Liu, K., Li, W.: Organisational Semiotics for Business Informatics. Routledge (2014)
4. Singh, M.P.: The Pragmatic Web. IEEE Internet Computing 6(3), 4–5 (2002)
5. Bonacin, R., Hornung, H., Dos Reis, J.C., Pereira, R., Baranauskas, M.C.C.: Pragmatic Aspects of Collaborative Problem Solving: Towards a Framework for Conceptualizing Dynamic Knowledge. LNBIP 141, 410–426 (2013)
6. Uschold, M.: Where are the semantics in the semantic web? AI Mag 24(3), 25–36 (2003)
7. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5, 199–220 (1993)
8. Poesio, M., Traum, D.R.: Conversational Actions and Discourse Situations. Computational Intelligence 13, 309–347 (1997)
9. Searle, J.R.: A Classification of Illocutionary Acts. Language in Society 5(1), 1–23 (1976)
10. Dam-Jensen, H., Zethsen, K.K.: Pragmatic patterns and the lexical system - A reassessment of evaluation in language. Journal of Pragmatics 39(9), 1608–1623 (2007)
11. Grant, J., Kraus, S., Perlis, D.: A logic-based model of intention formation and action for multi-agent subcontracting. Artificial Intelligence 163(2) 163–201 (2005)
12. Hawizy, L., Phillips, I.W., Connolly, J.H.: Intention Modeling: A semiotic view. In: Proceedigns of International Conference Applied Computing, pp. 478–482 (2006)
13. Shahar, Y., Miksch, S., Johnson, P.: An intention-based language for representing clinical guidelines. In: Proceedings of AMIA Annual Fall Symposium, pp. 592–596 (1995)
14. Kanso, H., Soulé-Dupuy, C., Tazi, S.: Representing Author's Intentions of Scientific Documents. In: Proceddings of ICEIS, pp. 489–492 (2007)
15. Angeletou, S., Rowe, M., Alani, H.: Modelling and Analysis of User Behaviour in Online Communities. In: Proceedings of Semantic Web Conference, Springer, pp. 35–50 (2010)
16. Dos Reis, J. C., Bonacin, R., Baranauskas, M.C.C.: A Semiotic-based Approach to the design of Web Ontologies. In Proceedigns of ICISO. Reading, UK, pp. 60–67 (2010)
17. Bonacin, R., Dos Reis, J.C., Hornung, H., Baranauskas, M.C.C.: An ontological model for supporting intention-based information sharing on collaborative problem solving. Collaborative Enterprise 3(2-3), 130–150 (2013)