Towards Social Performance Indicators for Community-based Ontology Evolution *

Pieter De Leenheer[1,2], Christophe Debruyne[1], and Johannes Peeters[2]

[1] STARLab, Vrije Universiteit Brussel, Brussels 5, Belgium
[2] Collibra nv/sa, Brussels 12, Belgium

Abstract. The "living" ontologies that will furnish the Semantic Web are lacking. The problem is that in ontology engineering practice, the underlying methodological and organisational principles to involve the community are mostly ignored. Each of the involved activities in the community-based ontology evolution methodology require certain skills and tools which domain experts usually lack. Finding a social arrangement of roles and responsibilities that must supervise the consistent implementation of methods and tools is a wicked problem. Based on three technology-independent problem dimensions of ontology construction, we propose a set of social performance indicators (SPIs) to bring insights in the social arrangement evolving the ontology, and how it should be adapted to the changing needs of the community. We illustrate the SPIs on data from a realistic experiment in the domain of competency-centric HRM.

1 Introduction

While simple, the vision of the Semantic Web remains largely unrealised [14]. It resulted in a set of design principles, collaborative working groups, and a variety of enabling technologies that are becoming de facto formats for structuring and exchanging data and services. However, the "living" ontologies that will furnish the Semantic Web are generally lacking. In information sciences, ontologies are lexical representations that refer to context-independent and language-neutral concepts, relationships between these concepts, ontologically relevant instances, and axioms. Of those that are published on the Web, only some of them are actively maintained and thus reflect the current domain. Many others are rather outdated prototypes [10], not "usable and reusable" [11], and unworthily categorised ontologies as an agreement on the schema vocabulary is non-existing. The approaches to build and evolve *community-based* ontologies are unsatisfactory, both theoretically and as far as the quality of the results is concerned.

Despite the technological progress, in ontology engineering practice, the underlying *methodological* and *organisational* principles to involve the community are mostly ignored [2]. They systemically disregard the gap between socialisation among people at the community/social level; and information exchange between computer systems at the operational/technical level. A viable community-based approach considers socialisation as basis to identify interoperability needs in the community. Additionally,

* Thanks to our colleagues at Collibra and VUB STARLab. This research was partially sponsored by EC projects FP6 IST PROLIX (FP6-IST-027905) and FP7 TAS[3].

methodologies are indispensable to actually learn and enact the community to conduct, in a systematic and repeatable manner, the necessary activities in community-based ontology evolution. Such a paradigm shift would finally bridge the *gap* between the social and technical part of the community.

We define *community-based ontology evolution* as the co-evolution between three first-class citizens: (i) social interactions between people; (ii) the information systems[3] that support them; and (iii) the ontologies required to establish semantic interoperability between these systems. Each of the involved activities in the ontology evolution methodology (Sect. 4) requires certain skills and tools that domain experts usually lack [10]. Finding a social arrangement of roles and responsibilities must supervise the consistent implementation of methods and tools is a wicked problem. Therefore, we do not only log the evolution of the ontology itself, but also meta-level actions and discussions in order to shape the roles and responsibilities into an optimal social arrangement.

2 Social Performance Indicators

In [4], we introduced a *community metamodel* that includes a community's first-class citizens, namely *actors* (e.g., domain experts, knowledge engineers, concept stewards), *actands* (e.g., concept types), and *actions* (e.g., change operations). Actions are performed by actors on certain actands. Our approach here is to bootstrap a social arrangement of actors, but evolve it based on the social patterns that emerge from the individuals' actions on the community's actands, in this case concept types in the ontology.

As identified in [5], there are three main orthogonal technology-independent problem dimensions in ontology construction for which a social arrangement is critical. Based on these dimensions, we propose a (non-exhaustive) set of key *social performance indicators* (SPIs) that could bring insights in the social arrangement evolving the ontology, and how this arrangement should be adapted to the changing needs of the community. We illustrate the SPIs on data from a realistic experiment in the domain of *competency-centric HRM* [6].

2.1 Bi-sortality

Regarding the *bi-sortal* nature of the Web, a viable ontology should be useful for both human knowledge sharing and automatic information processing. In order for an ontology to define a common understanding for both human as well as for software agents, concept URIs should dereference to a formal and informal description of the semantics. The *informal* part describes the meaning of the element in terms of consensual terminology. E.g., the English version of Wikipedia defines a collection of more than 850,000 hyperlinked URI-referred entries, which means it holds unique identifiers for 850,000 concepts [10]. If there is more than one meaning for a term (e.g., "person"), a special disambiguation page provides a set of alternative semantics depending on the context (grammatical, legal, or natural sense) of the term. *Formal* semantics are used to support automatic processing and explicitly exclude unwanted interpretations. E.g.,

[3] Actually we refer to the computerised components of information systems.

a DL-axiom can be defined to exclude objects to instantiate both class `Man` and `Woman`. Finally, orthogonal to the informal and formal parts, there is the meta-part consisting of actions for adding, changing and deleting concept stewards, and discussions that led to the concept type's current version and hence should be considered as an important part of the concept URI [10]. To preserve a balance between efforts spent on the three respective parts of the ontology representation, following SPIs are considered:

- **SPI 1.1.** If there is a lot of discussion about a concept type, but almost no formal or informal actions in response to the discussion, there may be a lack of expertise or the discussion may be a waste of time. Therefore, we **observe the number of change requests, comments, or suggestions that did not receive any respons or answer**. The derived social network could indicate an underperformance or under-expertise in the current social arrangement, and the need to introduce more specialised domain experts to reconcile those unelaborated concept types in scope. We distinguish following discussion-related actions: a request (REQ) asks for an refinement action on a concept type; a question (QUE) concerns a concept type; a comment (COM) is a note about a concept type, independent of any REQ, QUE or other COM; a respons (RES) is a special type of COM on a QUE or COM; and an answer (ANS) acts on a QUE about a concept type.
- **SPI 1.2.** If the purpose of the resulting ontology is mainly for human knowledge sharing (resp. automatic data processing), then not much effort should be put in the formal (resp. informal) representation of a concept type, e.g. as in a canonical data model (resp. business glossary). Therefore, we **observe the balance between the resources spent on the respective parts of the representation of individual concept types through time.** This may indicate the need to adapt the social arrangement accordingly.
- **SPI 1.3** Highly connected actors within the derived social network may also indicate **emerging sub-communities or elites**.

2.2 Impedance Mismatch

An ontology is a *community contract* [10], and therefore should be defined by its community not by a single developer [6]. Stakeholders may have diverging perspectives on the the representation of the domain, that consequently serve as input for the alignment of next version of the shared ontology. Consider, e.g., the *impedance mismatch* between business analysts and engineers, which is one of the key issues in business/IT alignment. E.g., perspectives on following aspects of concept types may be different:

- *terms*: engineers tend to use camel-case or abbreviated labels for concepts as opposed to business analysts that prefer well-elaborated terms with proper adjectives. Following is an example of one term in XBRL[4] that includes too much context. `AdditionsOtherThanThroughBusinessCombinationsCopyrightsPatents-AndOtherIndustrialPropertyRightsServiceAndOperatingRights` Ideally, a term should only refer to the essential meaning (i.e., gloss, synset), the rest of its meaning is contextual, incl. classification, relationships, and/or rules. Moreover, multilinguality may be a problem to understand each other.

[4] http://www.xbrl.org/

– *relationships*: depending on the domain, the properties attributed to a concept type may be different. E.g., the term `Car` can be uniformly articulated with gloss[5] "a motor vehicle with four wheels; usually propelled by an internal combustion engine". However, the relationship `Car has Price` in the domain of car sales, is of no "use" in the car registration domain. On the contrary, `Car has Registration Plate` in the latter domain may be irrelevant for car sales.
– *epistemology*: Term `Customer` may refer to a reified concept type in a CRM datamodel. However, from a business perspective a `Customer` may be indirectly modelled in terms of circumstantial facts about a `Person`. E.g., a `Customer` may be seen as a `Person` that purchases `Goods` with a certain `Frequency`. `Frequency` determines the degree of "customership", the type of `Goods` determines the type of `Customership`, etc.

In order to bridge the impedance mismatch between different stakeholder perspectives, we observe the following SPIs.

– **SPI 2.1.** Optimally, the definition of a concept type should be an agreement resulting from a mixed discussion between both technical and non-technical experts, instead of an isolated decision biased by one or two of either side. Therefore, we may **analyse the social networks derived from the discussion threads (see also SPI 1.1).** Maximising the explicit consideration of different perspectives on the evolution of the ontology will ultimately increase stakeholder satisfaction [4].
– **SPI 2.2.** Sometimes it is difficult to categorise contributing actors as technical or non-technical. Therefore, **we may derive certain patterns in the actions of actors, and eventually identify which role these actors fit better in order to optimise the social arrangement.** This shifting of responsibility may have to be accredited by additional training.
– **SPI 2.3.** Additionally,we could **automatically identify possible association between action types and how they appear in clusters**.

2.3 Co-evolution

Ontology evolution may happen either in an engineering-like approach, where ontology evolution activities are all conducted by a small elite in a top-down fashion; or in a more pragmatic bottom-up manner where all community stakeholders are invited to render their perspective on every step of ontology evolution. Community-based ontology evolution is a hybrid approach where the community is involved in all stages, but responsibility is enforced by *policies* and *roles* (see [6, 4] for details). In order to realise co-evolution, we consider following SPIs:

– **SPI 3.1.** In an agile approach the list of concept types under development is continuously rescoped regarding the changing needs of the community. Domain experts need to divide their efforts between these concept types, and at the same time on newly scoped concept types. Gradually, the community may end up with a large concept base under construction, where none of the concept types is mature enough

[5] e.g., from http://wordnetweb.princeton.edu

to be applied. Therefore, we may get **an overview on the efforts spent on each concept type, cluster them accordingly, and analyse the lag that emerges with respect to the time the concept types were first scoped.**

– **SPI 3.2.** In a collaborative setting many domain experts contribute to many concept types concurrently. This makes it difficult to check the different steps of the methodology. In order to determine the maturity of a concept type, it is important to **observe the gradual maturing of a concept type from creation to unification.** This may lead to the identification of actors that lack skills or do not fully understand their responsibility.

3 Related Work

Based on a representative sample from Wikipedia, Hepp et al. [10] concluded that wikis are promising platform for community-based evolution of structured knowledge. To this end, MyOntology [15] is an Austrian project that built an infrastructure and culture of Wikis as an ontology editor that fosters collaborative, community-driven ontology creation and maintenance. Via concept URIs, it facilitates the use of multimedia elements to improve the expressiveness and disambiguity of informal concept definitions. They regard it as beneficial if the definition of a concept is not separated from the discussion that lead to shaping the intension of this concept. OntoWiki [1] is a free, open-source semantic wiki application, meant to serve as an ontology editor. In contrast to most semantic wikis, OntoWiki is form-based rather than syntax-based, and thus tries to hide as much of the complexity of formal representation formalisms from users as possible. AceWiki [13] applies controlled natural language ACE so that the formal statements of the wiki are shown in a way that looks like natural English.

Efforts are made on augmenting wikis with semantic annotations. E.g., Semantic Wikipedia [12] does not focus on semantic reconciliation of fact types but applies them to annotate the content of normal Wikipedia.

The SIOC[6] Ontology[7] focuses on the integration of online community socialisation by augmenting it with semantics. SIOC is used in conjunction with the FOAF[8] vocabulary for expressing personal profile and social networking facts. In the context of a discussion, `Forum` topics can range from conceptions that must be added to the ontology to meta-concept types (beyond actor, action, and actand) that constitute the community metamodel itself. By semantically augmenting discussions, we will be able to produce SPIs more correctly.

4 Methodology and Tool

For our experiment, we adopted the Business Semantics Management (BSM) methodology that is defined by two iterative cycles, each grouping a number of activities (detailed

[6] Semantically-Interlinked Online Communities
[7] htttp://www.sioc-project.org
[8] Friend Of A Friend

in [6]). The first cycle is *semantic reconciliation*. It is concerned with the rendering (create, refine, articulate activities) and unification of diverse perspectives on the representation and meaning of a *scoped* set of concepts. The second cycle *semantic application* concerns the activities to select and commit unified semantic patterns for automatic information processing. The activities in BSM are implemented by a set of change operators (see [7]) for each of the parts of the ontology. For the formal part: ALE: add lexon[9]; DLE: Delete Lexon; and RLE: Refine Lexon. For the informal part: ART: Articulate Concept; ASY: Add Synonym; ASO: Add Source; DSO: Delete Source; DSY: Delete Synonym; RGL: Refine Gloss; and RSO: Refine Source. For the discussion part: CLE: Clean up; COM: Comment; CST: Change Steward; AST: Add Steward; RST: Remove Steward. QUE: Question; REQ: Request; RES: Respons; ANS: Answer. Finally there is CRE: Create Concept; DCO: Delete Concept; and MOV: Move Concept. Changes were logged with following attributes: time, actor, action, actand.

Based on related work, we have chosen a wiki which is a simple collaborative system for creating and maintaining hyperlinked collections of Web pages. A wikipage defines the workspace for a concept type, it provides: (a) a non-intruisive interface to describe conceptions in natural language, augmented by multimedial, without the need to understand or locate the underlying physical file structure; (b) a built in mechanism to track changes, to compare different versions, and to revert to a previous version; (c) the use of URIs to identify concepts (cf. future Internet requirements); (d) a discussion forum as important part of the concept's evolutionary representation; (e) basic role mechanism for social arrangements of concept stewards and concept watchmen. Figure 1 illustrates the wiki page for the concept type `Resume`, including the different aspects: articulation (gloss), synset, and lexons. Despite its success in Wikipedia, using a wiki to construct an ontology in the context of a professional organisation requires additional policies and management control to ensure appropriate quality control and governance of concept types and domain experts. This forms an extra motivation to use SPIs for a careful configuration of actor roles and responsibilities, which may be initially bootstrapped, but should be adaptable if opportunities for improvement are observed.

5 Experiment Results

The change logs for the analysis results from a realistic case were a community of 14 actors collaboratively developed an ontology base of 180 concepts in the domain competency-centric HRM, over a period of ca. 12 weeks. The goal was to build an ontology base that can be used for exchange of competency information. The HR-XML standard was the main ontological resource. Due to space limits, we only report on a selection of SPIs. However, the reader can analyse the change logs via our public portal[10].

[9] A lexon, as defined in [7], is a quintuple (header term, role, co-role, tail term, resource) representing a possible binary relationship. It corresponds to two RDFS triples, one for each reading direction.

[10] http://starpc11.vub.ac.be/∼chrdebru/OISExperiment/

SPI 1.1 Fig. 2 illustrates the derived social network. Circles denote actors P0x ($0 \leq x \leq 14$), boxes denote concept types, and edges indicate discussion actions (ANS, RES, REQ, COM). For the concepts outside the circle, e.g., `Skills`, `Skill Level`, and `Skill` the REQs, RESs and COMs were never answered (ANS). It turns out that for most of the concept types outside the circle, the number of concept refining actions is low as well, resp. 7, 5, and 6 for the examples. An exception is, e.g., `Resume` that was actand of 70 actions, but none of them were discussion-related, and therefore it is not in the figure. Most of the discussion were about taxonomical classification and reification of lexons. For **SP1.3**, in Fig. 2, a sub-community (indicated in grey) consisting of $P2$, 3 and 9 is emerging from the higher concentration of shared discussions.

SPI 1.2 In Fig. 3 the actions are grouped per part of the ontology: $G0$ for the discussion part; $G1$ for the formal part; and $G2$ for the informal part. $G3$, 4 and 5 resp. for creating, deleting and moving concept pages. The graph shows three moments (i.e., 3/26; 4/2; and 4/23) where all groups peak. These moments indicate (i) an intermediary deadline for a new ontology version to be accepted, and (ii) and consequently a point where the domain is rescoped for another iteration of the ontology evolution cycle, resulting in a temporarily higher production. The initial scoping peak is the largest, while the following two peaks become gradually smaller. This indicates the ontology reaches a fixpoint as the final deadline approaches, as more concepts covering the domain become mature. There are two isolated peaks of actions on the formal parts in the second iteration: 29 actions on 2009-04-09 and 22 on 2009-04-16. This shift of balance between formal and informal actions is the result of a general request by the core domain expert to spent more resources on formalisation of core concept types.

SPI 2.2 The graph in Fig. 4 shows per person the distribution of effort (in $\%$) for different action types. We zoom in on the discussion actions (ANS, REQ, RES, COM, QUE). All other actions are grouped in $G5$. $P9$ and $P14$ are clearly more involved in driving the discussions. They take initiative by making change requests (REQ): 8% and 25% of their time resp. $P09$ is more engaged in quality control by giving answers (ANS) 8% of its time. On the contrary, while spawning RES and REQ, $P14$ does not answer, or contribute to the other parts of the ontology at all. This may indicate spam. Analogously, we could focus more on the informal or formal actions and identify the formal ontologist(s) in the community.

SPI 3.2 When creating a concept (CRE), the methodology requires to add a steward (AST) and articulate (ART) that concept with a gloss as well. Fig. 5 given the unbalanced distributions of CRE ($G0$, left), AST ($G1$, middle), and ART ($G2$, right) resp. for actors $P2$, 3, and 9. $P02$ introduces 21 concepts, but appoints a steward in only 13 cases, and articulates these concepts in only 16 cases. $P03$ articulates almost every concept he introduces (28 out of 29), but does only adds a steward in 5 cases. $P09$ was already mentioned (in SPI 2.2) as an initiative taker and quality controller rather than somebody who works on the definitions (created 0 concepts). The third distribution confirms this. Moreover, it seems that $P09$ also fixes problems by adding stewards (3 times) and articulations (6 times) for concepts he did not introduce himself.

6 Discussion and Future Work

For the SPIs we did not report on, we give some further pointers. For SPI 3.1., we refer to Martin Hepp's ontology engineering lags [9]. For SPI 3.2, we refer to the Mature[11] project that focuses on ontology maturing during learning processes. Quick analysis is possible with wikimedia's special pages: e.g., in our case it turns out that there are 17 orphan concepts, and 8 dead-end concepts. Moreover, there seems to be hold power law distribution between concepts and revisions: there are many concepts with few revisions, while only a small amount concepts have many iterations. Concept types with only few iterations are usually value types that are agreed on easily. On the other hand, core concept types, like `Resume` have many iterations (38 refinements) as they have to reflect the many stakeholder perspectives present in the community.

In this paper the SPIs were identified starting from the three problem vectors of ontology construction. Currently, we are developing a reference framework which will allow an exhaustive identification of SPIs for more purposes than only ontology construction. E.g., SPIs to develop and maintain the *reputation* of actors in the community. In this work, $P09$ popped up repeatedly for its particular reputation of quality assessor by requesting changes, answering questions, and fixing stewardships and articulations that were supposed to be done by the actors that created the concept. Reputation management then boils down to tracking somebody's actions and other actors' opinions about those actions [3].

Another, objective of SPIs would be to analyse underperformance, and finally trigger the right incentives to the right actors to take action. This could start from the factors that lead to the success of Wikipedia. We refer to the European project Insemtives[12] for more on incentive management. SPI analysis may be supported by several techniques. Related to SPI 2.3, we are investigating the use of *association mining*. Also we will incorporate our ontology reuse mechanisms [8] and define SPIs to analyse their effect on the ontology evolution cycle. Currently we are co-developing our SPI framework in several government and industry cases. This will provide a more heterogenous set of actors from both business as IT needed to assess SPI 2.1.

References

1. S. Auer, S. Dietzold, and T. Riechert. Ontowiki – a tool for social, semantic collaboration. In *5th International Semantic Web Conference*, LNCS 4273, pages 736–749. Springer, 2006.
2. J. Cardoso. The semantic web vision: Where are we? *IEEE Intelligent Systems*, 22(5):84–88, 2007.
3. T. Coenen. *Knowledge Sharing over Social Networking Systems*. PhD thesis, Vrije Universiteit Brussel, Brussels, Belgium, 2006.
4. P. De Leenheer. *On Community-based Ontology Evolution*. PhD thesis, Vrije Universiteit Brussel, Brussels, Belgium, 2009.
5. P. De Leenheer and S. Christiaens. Challenges and opportunities for more meaningful and sustainable internet systems. In *Proc. of the International Future Internet Symposium*, LNCS, Vienna, Austria, 2008. Springer.

[11] http://mature-ip.eu

[12] http://www.insemtives.eu/

6. P. De Leenheer, S. Christiaens, and R. Meersman. Business semantics management: a case study for competency-centric HRM. *Journal of Computers For Industry*, forthcoming, 2009.
7. P. De Leenheer, A. de Moor, and R. Meersman. Context dependency management in ontology engineering: a formal approach. *LNCS Journal on Data Semantics*, 8:26–56, 2007.
8. C. Debruyne, P. De Leenheer, and R. Meersman. A method and tool for fact type reuse in the dogma ontology framework. In *Proceedings of ODBASE-OTM 2009, forthcoming*, LNCS. Springer, 2009.
9. M. Hepp. Possible ontologies: How reality constrains the development of relevant ontologies. *IEEE Internet Computing*, 11(1):90–96, January 2007.
10. M. Hepp, K. Siorpaes, and D. Bachlechner. Harvesting wiki consensus: Using wikipedia entries as vocabulary for knowledge management. *IEEE Internet Computing*, 11(5):54–65, 2007.
11. M. Jarrar. *Towards Methodological Principles for Ontology Engineering*. PhD thesis, Vrije Universiteit Brussel, Brussels, Belgium, May 2005.
12. M. Krötzsch, D. Vrandecic, M. Völkel, H. Haller, and R. Studer. Semantic wikipedia. *Journal of Web Semantics*, 5:251—261, 2007.
13. T. Kuhn. Acewiki: A natural and expressive semantic wiki. In *Proceedings of Semantic Web User Interaction at CHI 2008: Exploring HCI Challenges, CEUR Workshop Proceedings*, 2008.
14. N. Shadboldt, T. Berners-Lee, and W. Hall. The semantic web revisited. *IEEE Intelligent Systems*, 2006.
15. K. Siorpaes, M. Hepp, A. Klotz, and M Waltl. myontology: Tapping the wisdom of crowds for ontology building. In *6th International and 2nd Asian Semantic Web Conference (ISWC2007+ASWC2007)*, pages 99–100. Springer, November 2007.

Resume

Articulation Distinctions are sometimes made between the terms "Resume" (also spelled as Résumé) and "Curriculum Vitae" (frequently abbreviated as "CV"). Distinctions between these documents are sometimes based on the length, content, and purpose of the document. A resume is frequently thought of as a summary (one or two pages) of a candidate's qualifications, work experience, and education. A CV may be longer and include details on talks, publications, inventions (patents), achievements, knowledge of languages, affiliations, and other details such as references and clearance (if required). For purposes of this specification, the term "Resume" will be used to include both the Resume and CV document types.

Synset CV, Curriculum Vitae, Résumé

Head Term	Role	CoRole	Tail Term	Resource
Resume	is about	writes	Person	HR XML Resume
Resume	contains	part of	Education	HR XML Resume
Resume	contains	part of	Qualification	HR XML Resume
Resume	contains	part of	Language Competency	HR XML Resume
Resume	contains	part of	Patent	HR XML Resume
Resume	contains	part of	Reference	HR XML Resume
Resume	has	is of	Achievement	HR XML Resume
Resume	has	is of	Security Credentials	HR XML Resume
Resume	contains	part of	Work Experience	HR XML Resume
Resume	contains	part of	Publication Experience	HR XML Resume
Resume	has	of	Service	HR XML Resume

Steward User:lhoste

Category: Concept

Fig. 1. Wiki page for concept type `Resume`.

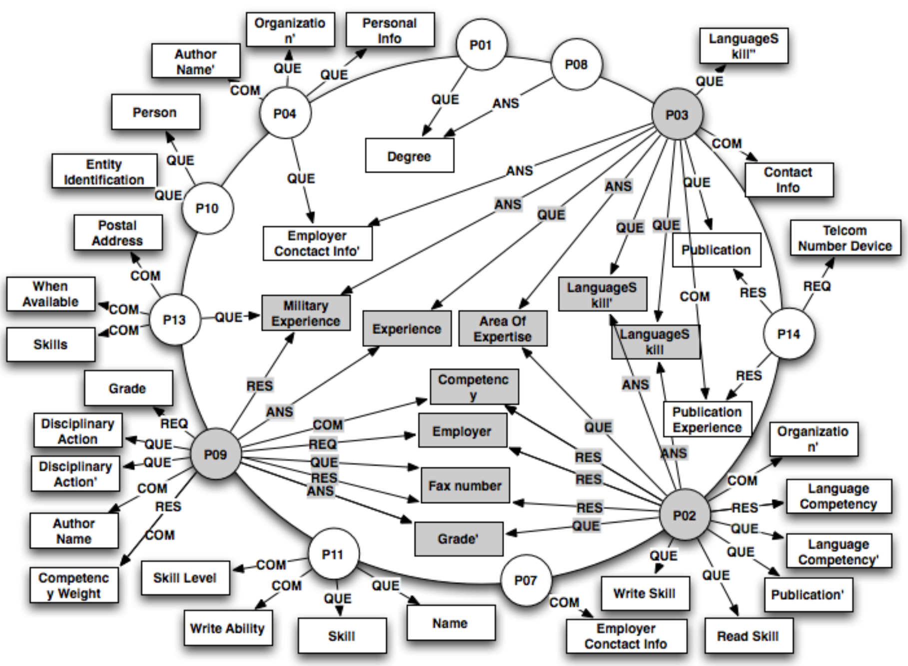

Fig. 2. The social network and sub-communities derived from the discussion threads.

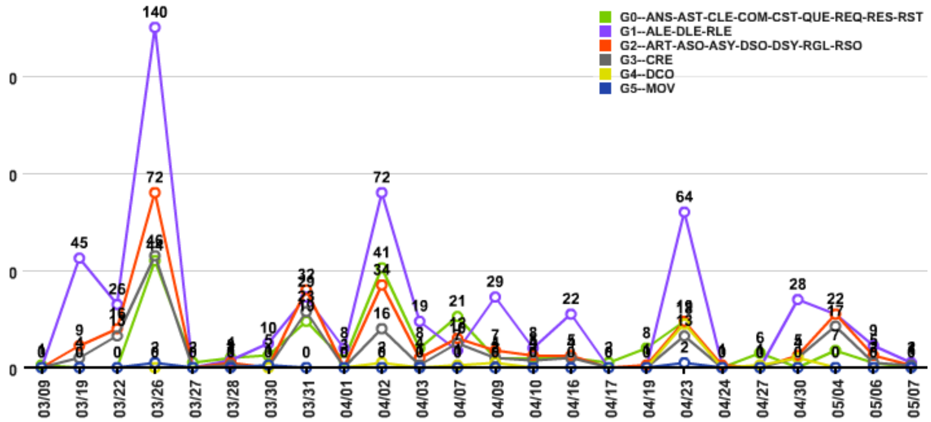

Fig. 3. Actions grouped per part of the ontology over time.

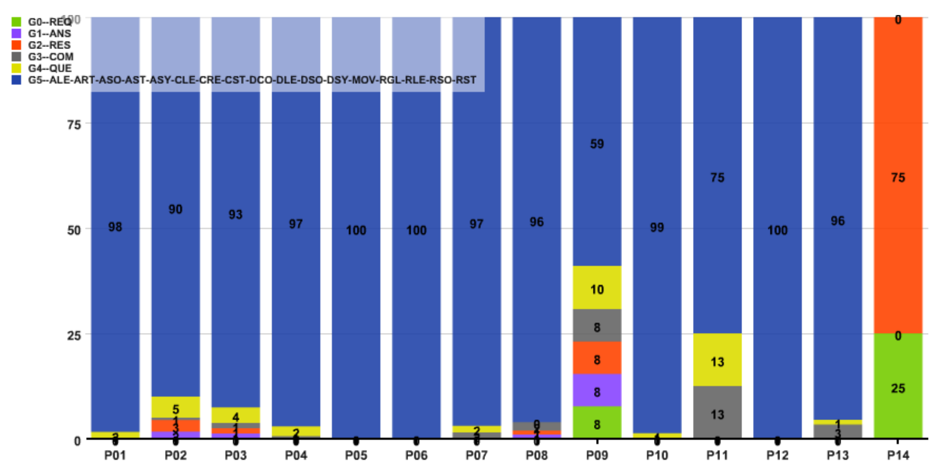

Fig. 4. The discussion effort distribution per actor leads to identification of initiative takers (REQ), answerers (ANS), and spammers (RES, REQ; but nothing else).

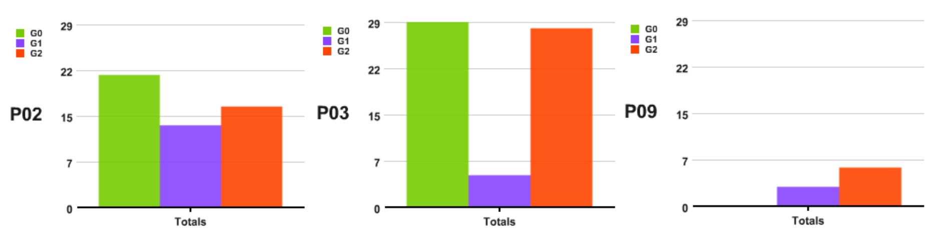

Fig. 5. Distributions of CRE ($G0$, left), AST ($G1$, middle), and ART ($G2$, right) resp. for actors $P2$, 3, and 9.