Using Stakeholder and Pragmatic Analyses to Clarify the Scenario of Data Sharing in Scientific Software

Alysson Bolognesi Prado(✉) and Maria Cecilia Calani Baranauskas

Institute of Computing, State University of Campinas, Campinas, São Paulo, Brazil
{aprado, cecilia}@ic.unicamp.br

Abstract. Scientific communities can be seen as highly focused organizations, composed of people performing strict patterns of behavior. The growing body of scientific data available digitally, as well as new infrastructure of distributed access, has given to funding agencies, politicians and scientists in general the foresight of novel possibilities of discovery and innovation reusing those data. Many stakeholders now expect the data to be released, although relevant sharing rates are not always verified. In this paper, we propose a method to bring forward and represent these interests. Applying this method, we investigated how the availability of software capable of data storage and sharing can act upon their users, and whether it makes them more suitable to share research byproducts. Results show that, although current software empowers the scientists to carry out their own research, it still does not create a path through which users can make their interests spread among other stakeholders.

Keywords: Scientific communities · Actor-Network Theory · Norm analysis · Stakeholder analysis · Data sharing · Collaborative systems

1 Introduction

Over the last two decades, scientists have increasingly relying on computers to store and manage research data. This trend has given rise to a whole field of knowledge called e-science [6]. Several software have been developed in order to support researcher activities, for instance, SEEK [27], openBIS [2] and PEDRo [10].

As a consequence of this higher availability of digitally stored data, some began to advocate the possibilities of other scientists to reuse these data [4]. However, data sharing did not reach the expected rates [1, 3]. Faniel and Zimmerman [9] have raised some questions to be answered in order to achieve data sharing and reuse in a larger scale. For instance, *"What other types of social interaction beyond that with the data producer can facilitate data reuse (e.g., colleagues, third party experts)? How can social exchange and documentation be combined to support data sharing and reuse on a large scale?"* [9, p. 61].

Scientific communities can be seen as highly focused organizations [26] performing tasks strictly conformed to their accepted methods. Therefore, we argue that the tools

M.C.C. Baranauskas et al. (Eds.): ICISO 2016, IFIP AICT 477, pp. 171–180, 2016.
DOI: 10.1007/978-3-319-42102-5_19

provided by Organizational Semiotics [19] are suited to capture the systemic features of scientific work, in particular, the ones related to intentions, interests and behavior, that is, the pragmatic level. Associated to a stakeholder analysis, it could also be used for understanding the cultural reverberations, either beneficial or disadvantageous, of an innovation among the involved parties [25].

In this paper, we outline and apply a method to capture the social and pragmatic aspects of the interaction of scientists with one another as well as with technological artifacts, and the interests involved with data sharing. Given the important social facet involved in use and reuse of scientific data, we draw upon the sociological framework provided by Actor-Network Theory [16, 17] in order to improve the understanding of how non-human entities, such as software and data but not limited to them, participate in the processes performed by scientists.

This paper is organized as follows: Sect. 2 brings a summary of other studies contributing to the theoretical and methodological support of our analysis, while Sect. 3 shows the sources of information that fed it. In Sect. 4 we propose a method to capture how the interested entities articulate in a social scenario involving people and technology in order to achieve mutual benefit, and apply it in Sect. 5 to existing software. In Sect. 6 we discuss the findings and add our final remarks in Sect. 7.

2 Theoretical-Methodological Background

The Actor-Network Theory (ANT) is a theoretical as well as methodological framework that emerged from Social Studies of Science and Technology. It proposes to see social phenomena as chains of associations distributed in time and space, dependent of the continuous agency of their participants on each other [17]. Its origins on empirical studies of scientists performing their daily activities make it very suitable to help understanding social aspects of science making.

An *actor* is defined as any entity, whether human or not, capable of changing a certain state of affairs Human actors encompass people involved and interested in a certain state of affairs, along with their embodied knowledge and know how. ANT claims that purely human relations are short ranged and fast decaying. Non-human actors, although not provided of intentionality, have the potential of mediation and interaction, either by physical or cognitive means. Participants of social activities create associations among themselves, with the intention to get support, propagate forces and interests as well as mobilize other partners to achieve their goals.

An actor is an *intermediary* in a chain of associations when propagate the actions received without change. The behavior of an intermediary is predictable and the outputs are determined by the inputs. On the other hand, an actor is a *mediator* when it modifies, distorts, amplifies or translates incoming stimuli, being creative and showing variability and unpredictability to act on others. During a scenario analysis, intermediaries often fade out whereas mediators stand out as solvers of asymmetries and conflicts between other actors.

From the ANT perspective, normative behavior can be seen as the sum of social forces generated, stored and replied by actors and conducted through the associations between them. One actor acts upon the others affording some behavior, trying to mold

it to his own interests and benefit. The more these forces are stable, the more community's behavior seems homogeneous [18]. Knowing the source and path of these influences, as well as the reservoirs of the rules, examples, laws and models [17], is fundamental when someone is interested in understanding or changing them.

Prado and Baranauskas [23, 24] proposed a method for representing the social forces involved in a social phenomenon, introducing the idea of *intended norms*, that is, desired or expected patterns of behavior. These norms can be scored using the following proposed syntax, where each one receives an identification (ID) and the source of the norm, that is, the actor who is acting upon the other, shaping its behavior, is identified: **Norm** <norm-id>**: whenever** <condition> **if** <state> **then** <target-actor> **is** <obliged | permitted | prohibited> **by** <source-actor> **to do** <action>**.** These IDs can be applied in a graphical representation of the actors and their relationships, labeling arrows that show the path of mediators and intermediaries they travel.

3 The Conundrum of Sharing Research Data

Borgman [3] studies in depth the intricate subject of data sharing among scientists, producing a clarifying discussion about the involved interests, benefits and the difficulties to overcome. For her, sharing covers a variety of acts as varied as announcing the existence of data, posting them on a website, or contributing them to a richly curated repository.

There are four main rationales driving the requests for sharing research data: reproducibility of experiments, publicity of the outcomes of public funded research, reuse of data for asking new questions, and innovate the way science is made. The reproducibility is desired, for instance, by the peer reviewers of publications, who can make better judgments of the submitted papers. Other scientists are also interested in reproducibility, since they can validate the references on which they are basing their own research. Publications add value to data and vice-versa [20].

Public funded research is a target of legislators, representing the taxpayers, who wants to make available all research data produced using governmental grants, as a direct return to the society of the invested amounts. Researchers are also willing to have access to third-party data to ask new questions over the existing datasets, particularly when those are expensive or difficult to obtain. The reuse of data raises the question of assessing the veracity and integrity of a given dataset, and the need for documentation [4, 8]. At last, it has been argued about the existence of a scientific "fourth paradigm", a new way of doing science where algorithms for data crunching and mining are applied to massive datasets to produce scientific knowledge, therefore being highly dependent of availability of data.

However, there are also reasons for limiting data sharing. For instance, researches involving human subjects must be concerned with privacy issues and not all data could be disclosed. Scientists may also be unwilling to provide their data to other researchers, particularly when they are not related to the same project or institution. That happens because researchers compete for grants, jobs, publication venues, as well as for

students, and access to data is a competitive advantage in this scenario. In private funded research projects, a lower rate data sharing could be observed.

Studying how scientists interact with one another, with technology and with nature during the production of a scientific fact, Latour [16] proposes to understand this process as a progressive effort to strengthen claims by means of mobilizing other entities. These may be other person, but mostly are non-human of semiotic or material nature, which provide support to the arguments, allowing certain statements to be held against inquiries.

One of the main allies a scientist can resort is the previously published scientific literature, because claims are harder to be refuted when adequately associated to citations widely accepted. The same rationale can be applied to inscriptions and other visual records produced by laboratory instruments or derived from their data, which can go along with texts to permit their authors to sustain their point of view about a subject. The validity of these arrangements, however, cannot be measured by its intrinsic characteristics. Once approved by the scrutiny of other scientists, the hypothesis can be gradually strengthened as a scientific fact as it is used by others, who become interested in its correctness.

4 A Method to Trace Interests of Stakeholders

Understanding that the behavior of users towards a piece of software can be affected by the resultant of the forces propagating on the network of associations that surround and reach them, we need to clarify the intricate set of influences each actor exerts and receives in this scenario. To perform this task, we envisioned a method that identifies the involved parties and represents their relationships and interactions, as well as the patterns of behavior they expect or desire from the others.

The Stakeholder Diagram [15] is widely used in problem articulation, serving as a good starting point to clarify the scenario under study. It provides visual representation of the roles of the participants, also showing how closely related they are to the system under study. Pouloudi et al. [22] suggest the conjoint use of stakeholder analysis and ANT as a generic, context free, guidance to identify the involved humans, as well as other relevant non-human actors.

We adopted this method to elicit the participation of actors of both nature, and employed the concepts described in Sect. 2 to capture the interest each actor has in the behavior of the others. In our proposal, human actors will be drawn as round rectangles, while the non-human as the ones with straight corners. They are bound by lines whenever they have some kind of interaction or association. Norm statements, representing a behavior not necessarily observed but sometimes desired or intended, are attached to arrows depicting the path from the source actor to the target, that is, from the one who will benefit from the pattern of action to the actor who should perform it accordingly.

Whenever a non-human actor is software, we must inspect its user interface in order to capture some affordances for the users. As intended norms, these can be traced back to discover, or at least hypothesize, their sources. For example, consider a system that requires the user to select the project a given dataset belongs to, being otherwise denied to upload files. It may indicate a need or intent of the institution, which provides the

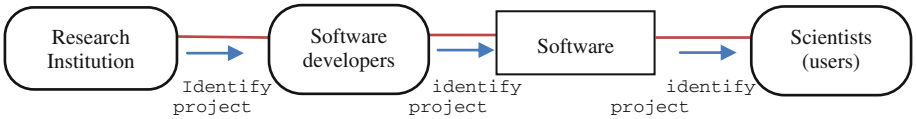

Fig. 1. Software and its developers acting as intermediaries, propagating the interests of the institution and molding the repertoire of actions of the users: they are obligated to identify the project a new dataset belongs. Arrows show the path of the influences and actions needed to promote such affordance

software to its scientists, to keep data files properly categorized. This can be captured as the following norm and diagram elements shown in Fig. 1.

Norm "identify project": **whenever** uploading a file **then** the scientists **are** obliged **by** the institution they work for **to** relate it to a previously registered project

In another example, consider the interests the general public have as taxpayers and as people being studied by scientists. They are stakeholders with distinct expectations about how the researchers should behave, expressing themselves through their representatives by means of legislation. The interest of the taxpayers is exerted by means of the taxes which willingness to be paid by the citizen can influence the legislators. On the other hand, the human subjects can only rely on their contacts to legislator and personal pressure. We must notice that the taxes are acting as a non-human actor – an intermediary forwarding the intended norm. In their turn, the legislators shall act as a mediator, finding a half term solution, for instance, creating legislation regulating the concession of grants. This can be recorded in the following set of norms plus the diagram sketched in Fig. 2.

Norm "funding": **whenever** funding research **if** with public taxes **then** the scientists **are** obliged **by** taxpayers **to** share their research products

Norm "privacy": **whenever** publishing data **if** it involves human subjects **then** the scientists **are** prohibited **by** human subjects **to** disclosure sensitive information

Norm "public/ethic": **whenever** funding researches **if** from public taxes **and** no sensitive data involved **then** the scientists **are** obliged **by** regulation law **to** make available their research products

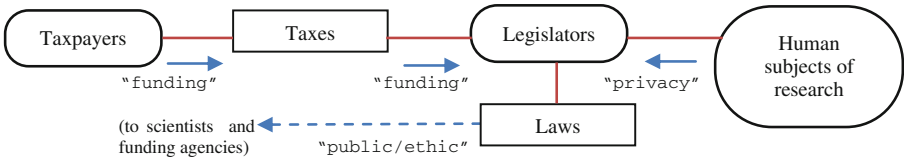

Fig. 2. Legislators acting as mediators of conflicting interests, since "funding" and "privacy" are balanced and issued as another distinct desired pattern of behavior

5 Inspecting a Software for Scientific Data Management

We have studied SEEK (v.0.16.3), a scientific web-based software intended to be used by systemic biology researchers, designed and developed by a team of e-science researchers funded by a consortium of research institutions. It supports the management, sharing and exploration of data and models [25]. The user can store biology-specific information, such as records of specimen, as well as research outcomes of generic type, such as publications and presentations. The user is always obliged to associate these items to one of the projects she is affiliated – as the "identify project" norm defined previously.

SEEK allows users to standardize and organize their digital assets, as well as to define access permissions for each one, ensuring the scientists have the final word about who can access those files. When defining these access rules, developers provide a default value that allows anyone to know the existence of the file, but not its content. All uploaded items have a persistent URL that allows data citation, as well as a reference to its authors, rewarding the contributions of each individual. Put in norm syntax, this means:

Norm "control": **whenever** data is uploaded **then** scientists **are** allowed **by** developers **to** choose access rules

Norm "default": **whenever** data is uploaded **if** user does not control permissions **then** other projects' members, anonymous visitors **are** allowed **by** developers **to** view data summary

Norm "request": **whenever** viewing data summary **then** other scientists **are** allowed **by** developers **to** request full access to file owner

Norm "cite": **whenever** writing a paper **if** used data files **then** scientists **are** allowed **by** developers **to** add a link referencing data

Norm "access": **whenever** data is stored on SEEK **if** paper cites data **then** other scientists, peer reviewers **are** allowed **by** scientists **to** view, use or download the file

To describe the broader context in which the software operates, we used the previous studies summarized in Sect. 3, as well as the complete reports of Latour [16] and Borgman [3]. In addition to the actual examples presented in Sect. 4, we scored some other following intended norms, representatives of the main involved interests. The main goal of scientists is to produce scientific content, mainly in the form of publications, which are expected to be accepted and cited. All these interests are placed on the stakeholder diagram, and the result is shown in Fig. 3.

Norm "funding policies": **whenever** receiving grants **then** scientists, institutions **are** obliged **by** funding agencies **to** produce scientific knowledge, mainly in the form of publications, following public/ethic laws

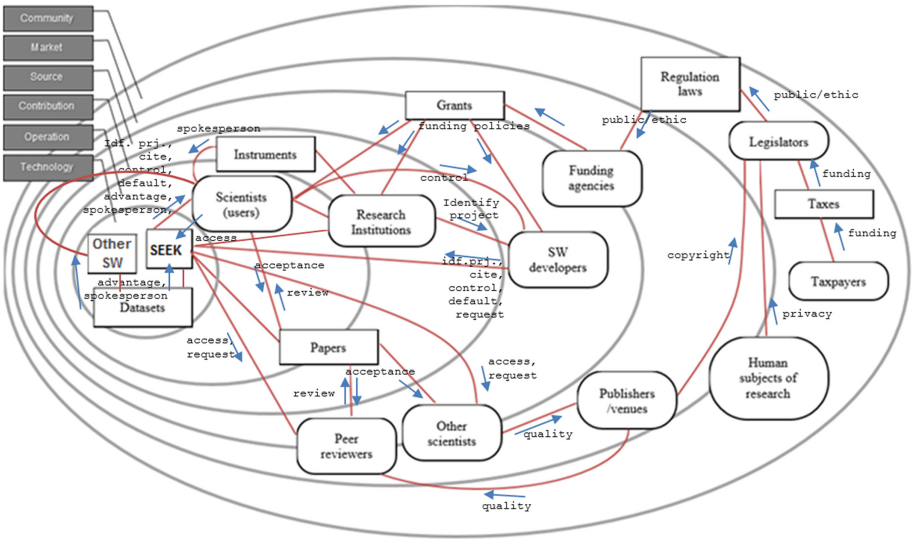

Fig. 3. Expanded stakeholder diagram eliciting the role of actors as the path of interests in certain patterns of behavior

Norm "copyright": **whenever** sharing resource products **then** scientists **are** prohibited **by** publishers **to** disclosure copyrighted material

Norm "quality": **whenever** publishing papers **then** publishers **are** obliged **by** other scientists **to** evaluate quality and veracity of its contents

Norm "review": **whenever** submitting paper for publishing **then** scientists **are** obliged **by** peer reviewers **to** provide related data

Norm "reuse": **whenever** possessing data difficult to be obtained **then** scientists **are** obliged **by** other scientists **to** publish data

Norm "advantage": **whenever** possessing data difficult to be obtained **if** they are kept private **then** scientists **are** allowed **by** those data **to** publish new relevant papers

Norm "spokesperson": **whenever** writing a paper **then** scientists **are** allowed **by** data, instruments, SEEK, other software **to** make claims and statements

Norm "acceptance": **whenever** writing a paper **then** other scientists, peer reviewers **are** obliged **by** scientists **to** accept their claims and statements based on data, instruments, references

6 Analyzing the Outcomes

The ANT rationale of the proposed method demands an inquisitive stance from the analyst of a given scenario, requiring the identification of in-between actors and finding or guessing how they receive and propagate interests. The search for a path for the influences, represented graphically by the arrows on the diagram, leads to a more complete set of involved entities and points out possible gaps. For instance, no path was defined for the norm "`reuse`", unless exists a missing link between the scientists and any – or several – other actor to promote some sort of influence. Other issues are brought forth by the arrangement provided by the stakeholder diagram; for instance, the funding agencies and the publishers are the main "bridges" to the more external social world, while the inner layers remain self-regulative. As a drawback, to guess the source of an interest can be a tricky task; for instance, "`default`" is attributed to the developers, only because they are the most probable responsible for that choice. Deontic operators also need to receive more attention, since the restrictions of "obliged" and "prohibited" have a different actuation than the possibilities given by "allowed".

Regarding the particular scenario in which the method was applied, the analysis reveals that software developers may play a major role as mediators. Despite of being often kept aside of the stakeholders list, their engagement to the subject in this case is not neutral. They capture the funding agencies' intentions issued as funding policies, the institutions' management requirements, the scientists' needs of using data, and finally their own interests of making some software capable of promoting data sharing; the resultant of mediation being expressed through the software affordances. Empirical validation of this finding will require further research, in situations with conflict of interests that are neither self-regulated prior to software development, nor solved during the analysis phase – for instance, using participatory design workshops.

Analysis also show that "other software" is an alternate path to provide to scientists certain possibilities of action – despite of these systems to have no social features, but instead good data visualization or manipulation – and therefore "SEEK" does not poses itself as an obligatory passage point [17] of interests. This raises as an interesting research question of how to better design these software, which should be addressed by research proposals such as the Human-Data Interaction informed by Organizational Semiotics [11]. Software has the capability of embed complex rules, also being able to act as mediator. Particularly in a social scenario where it can receive clues about the heterogeneous set of interests of the persons operating him [12]. For instance, "`request`" captures the interest of other scientists in datasets and mediates users, institutions, developers and other scientists' interests without exposing any file details.

There are other studies of the factors affecting data sharing and reuse among scientific communities, such as surveys based on questionnaires applied to scientists of different fields of knowledge such as health sciences [14], STEM – science, technology, engineering and mathematics [13], or social sciences [5]. They conclude that journal publishers have a statistically significant influence on such behavior, while normative pressure by other scientists and regulative pressure by funding agencies are not perceived. Our proposal complements these measures of perception by showing paths pressures travel, and pointing out possible inconsistencies. For instance, the value given to a published paper is coherent to the strength of the norms it mediates. On the other hand, the low perception of pressure by funding agencies is not compatible to

previous studies and discussions [3, 16], requiring a deep checking of the actual actuation of grants onto researchers. As stated by Eze et al. [7], the adoption of technology is not a one-off action that can be precisely captured by quantitative methods such as questionnaires; therefore a qualitative approach as provided by ANT also contribute to a broader analysis and comprehension of the phenomena.

7 Conclusion

The effectiveness of software when different interests – sometimes conflicting – are involved is not only a matter of its technical features but also a social [21] and pragmatic challenge. Scientific software intended for data sharing are not an exception, and despite of their capacity to store huge amounts of data, its publishing and reuse rates could be improved.

By asking who is potentially benefited from a certain behavior imposed or allowed by software, and the path this influence takes, responsible agents can be pointed out. Many of the patterns promoted by the system targets the users, while other external stakeholders seem barely influenced. As software is the direct point of contact of the scientists to data, its potential as mediator to resolve asymmetries and conflicts between converging interests could be better explored. Further research includes adding existing generic scientific social software such as Mendeley or ReserchGate to the diagram, to enrich the analysis with other possible social interactions between scientists mediated by technology.

Acknowledgments. We thank the Brazilian Research Foundation CNPq (Grant # 308618/2014-9). The opinions expressed in this work do not necessarily reflect those of the funding agencies.

References

1. Appelbe, B., Bannon, D.: eResearch – paradigm shift or propaganda? J. Res. Pract. Inf. Technol. **39**(2), 83–90 (2007)
2. Bauch, A., Adamczyk, I., Buczek, P., et al.: openBIS: a flexible framework for managing and analyzing complex data in biology research (2011)
3. Borgman, C.L.: The conundrum of sharing research data. J. Am. Soc. Inf. Sci. Technol. **63** (6), 1059–1078 (2012)
4. Carlson, S., Anderson, B.: What are data? The many kinds of data and their implications for reuse. J. Comput.-Mediated Commun. **12**(2), 635–651 (2007)
5. Curty, R.G.: Beyond "data thrifting": an investigation of factors influencing research data reuse in social sciences. MSc dissertation. Syracuse University (2015)
6. DeRoure, D., Jennings, N., Shadbolt, N.: Research agenda for the semantic grid: a future e-science infrastructure. Report Commissioned for EPSRC/DTI (2001)
7. Eze, S., Duan, Y., Chen, H.: Factors affecting emerging ICT adoption in SMEs: an actor network theory analysis. In: Khachidze, V., Wang, T., Siddiqui, S., Liu, V., Cappuccio, S., Lim, A. (eds.) iCETS 2012. CCIS, vol. 332, pp. 361–377. Springer, Heidelberg (2012)

8. Faniel, I.M., Jacobsen, T.E.: Reusing scientific data: how earthquake engineering researchers assess the reusability of colleagues' data. J. Comput.-Support. Coop. Work **19**(3–4), 355–375 (2010)
9. Faniel, I.M., Zimmerman, A.: Beyond the data deluge: a research agenda for large-scale data sharing and reuse. Int. J. Digit. Curation **6**(1), 58–69 (2011). doi:10.2218/ijdc.v6i1.172
10. Garwood, K., McLaughlin, T., Garwood, C., Joens, S., Morrison, N., Taylor, C.F., Paton, N. W.: PEDRo: a database for storing, searching and disseminating experimental proteomics data. BMC Genomics (2004)
11. Hornung, H., Pereira, R., Baranauskas, M.C., Liu, K.: Challenges for human-data interaction–a semiotic perspective. In: Kurosu, M. (ed.) Human-Computer Interaction. LNCS, vol. 9169, pp. 37–48. Springer, Heidelberg (2015)
12. Jensen, C.J., Dos Reis, J.C., Bonacin, R.: An interaction design method to support the expression of user intentions in collaborative systems. In: Kurosu, M. (ed.) Human-Computer Interaction. LNCS, vol. 9169, pp. 214–226. Springer, Heidelberg (2015)
13. Kim, Y., Stanton, J.M.: Institutional and individual influences on scientists' data sharing behaviors: a multilevel analysis. In: ASIST (2013)
14. Kim, Y., Kim, S.: Institutional, motivational and resource factors influencing health scientists' data-sharing behaviours. J. Sch. Publ. **46**(4), 366–389 (2015)
15. Kolkman, M.: Problem articulation methodology. Ph.D. thesis, University of Twente, Enschede (1993)
16. Latour, B.: Science in Action: How to Follow Scientists and Engineers Through Society. Harvard University Press, Cambridge (1987)
17. Latour, B.: Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford (2005)
18. Law, J.: Actor-network theory and material semiotics. In: The New Blackwell Companion to Social Theory. Blackwell Publishing Ltd
19. Liu, K.: Semiotics of Information Systems Engineering. Cambridge University Press, Cambridge (2000)
20. Pepe, A., Matthew, M., Borgman, C., Sompel, H.: From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web (2010)
21. Pereira, R., Baranauskas, M.C.C., Silva, S.R.P.: A framework-informed discussion on social software: why some social software fail and others do not? In: 11th International Conference on Enterprise Information Systems, ICEIS (2010)
22. Pouloudi, A., Gandecha, R., Atkinson, C., Papazafeiropoulou, A.: How stakeholder analysis can be mobilized with actor-network theory to identify actors. In: Kaplan, B., Truex III, D. P., Wastell, D., Wood-Harper, A.T., DeGross, J.I. (eds.) Information System Research, vol. 143, pp. 705–711. Springer US, New York (2004)
23. Prado, A., Baranauskas, M.C.C.: Perspectives on using actor-network theory and organizational semiotics to address organizational evolution. In: The 15th International Conference on Enterprise Information Systems, ICEIS (2013)
24. Prado, A.B., Baranauskas, M.C.C.: Capturing semiotic and social factors of organizational evolution. In: Hammoudi, S., Cordeiro, J., Maciaszek, L.A., Filipe, J. (eds.) ICEIS 2013. LNBIP, vol. 190, pp. 264–279. Springer, Heidelberg (2014)
25. Rambo, K., Liu, K.: An organisational semiotics approach to multicultural requirements engineering: stakeholder's analysis of online shopping for Saudi Arabian female consumers. Int. J. Inf. (IJI) **4**(1/2), 473–483 (2011)
26. Wenger, E.: Communities of practice and social learning systems. Organization **7**(2), 225–246 (2000)
27. Wolstencroft, K., Owen, S., du Preez, F., Krebs, O., Mueller, W., Goble, C.A., Snoep, J.L.: The SEEK: a platform for sharing data and models in systems biology. Methods Enzymol. **500**, 629–655 (2011). PUBMED: 21943917