# Social Emergent Semantics for Personal Data Management

Cristian Vasquez

Semantics Technology and Applications Research Lab,
Vrije Universiteit Brussel, Brussels, Belgium
cvasquez@vub.ac.be

**Abstract.** In order use our personal data within our day to day activities, we need to manage it in a way that is easy to consume, which currently is not an easy task. People have found their own ways to organize their personal data, such as categorizing files in folders, labeling emails etc. This is acceptable to a certain degree, since we have to deal with have some (human) difficulties such as our limited capacity of categorization and our incapacity of maintaining highly structured artifacts for long periods of time. We believe that to organize this great amount of personal data, we need the help of our communities. In this work, we apply the emergent semantics field to personal data management, aiming to decrease our cognitive efforts spent in simple tasks, handling semantic evolution in conjunction with our close peers.

**Keywords:** Personal data management, Social Annotation, Distributed ontology evolution, Emergent semantics.

## 1 Introduction

Lots of effort has been put into improving our personal data management capabilities through a computer. One of the most common approaches is to distill structures from our resources, or just add or "attach" metadata to them. It is commonly agreed that we can use these structures to answer complex and precise questions, which incentives resource annotation. Some estimations says that more than the 12% of the Web is structured to some degree[14][12], corpus that is currently exploited by the most popular search engines.

These web resources have been "annotated" with meta-data expressed using published vocabularies, which have been constructed via standardization committees and other organized groups, or just driven by the market. This is useful, but in some cases it is not sufficient since we live in a heterogeneous world, where differences exist and can be observed. We can easily agree that we will use coordinates to individualize a geographic region, but is not easy to agree the political nature of that region. Regarding some topics, global interoperability just cannot be reached. In these cases we need to reach local agreements between communities of limited size, essentially constructing specific vocabularies for a certain world view. This is usually archived via agreement processes that multiple individuals and communities reach following a *bottom up* approach to support their

P. Herrero et al. (Eds.): OTM 2012 Workshops, LNCS 7567, pp. 23–32, 2012.

tasks, such as information sharing. This interplay have been studied in the field of emerging semantics [1], but haven't been widely exploited yet.

Within this work, we want to exploit local and global semantics to improve our personal data management capabilities. In this context, we will refer to personal data as all kind of digital information that is gathered by an individual over (long) periods of time, and stored in a Personal space of information (PSI) which "includes all the information items that are, at least nominally, under that person's control (but not necessarily exclusively so)" [7]. This data includes for example emails, files, image recordings, URLs of resources of the web, etc. We know that managing this kind of information cannot be addressed only by means of global vocabularies, since users have their own ways of categorizations, which can be completely different from person to person [6]. The main focus of this work will be to improve our personal data managing capabilities, taking into account the social discourses with our close peers. This will be reflected in individual and collaborative mechanisms that will be used together to improve our long term retrieval capabilities with the help of *semantics* .

The idea of using semantics to manage personal data is not new, we can find interesting research in the field of Semantic Desktops [22]. A semantic desktop is a collection of tools that allows us to store personal information using flexible data models, usually graphs that represent and relate digital information. A Semantic Desktop is built on top of the ontological knowledge that is generated starting from user observations and his own behavior. Some may think about a single user's Semantic Desktop as a building block of a global Semantic Web [21], while others will see it as some formal and semi-formal complement of the user's mental models, where a user stores its own "interpretations" of their observed world. Additionally, one common tool that allow users to build representations of their own mental models is the personal wiki. A wiki is essentially a collection of documents which is connected via hyperlinks, making use of simple syntax for editing content. Wikis are used in many areas, such as online encyclopedias, collaborative learning environments, personal and group knowledge management systems [26]

Through this study we will focus on how we can represent our digital data in order to be shared, and how we can profit from our social interactions to build and link dynamic artifacts that will represent our shared "understanding", which we believe, would be useful for long term retrieval. In order to study possible benefits of combining (i) our local context with (ii) the emergent semantics of our close communities, we have designed a preliminary experiment in order to observe how the distinct contributions of our communities can be used to improve our data management capabilities. We know that the more structure the users contribute to their descriptions, the better are the shared artifacts to find information efficiently, to promote re-usage and sharing with other users [24]. In order to provide structure, we incur into costs. Within this experiment, we aim to perform a cost-benefit analysis about the total modeling efforts which are spent for the construction of those shared artifacts.

This document is organized as follows: Section 2 describes how annotations are interpreted in this work. Section 3 will explore the notion of description. Section 4 will explore the notion of personal context. Section 5 will describe the role of emergent semantics in this experiment. Section 6 will describe the experiment. Section 7 presents our conclusions and future work.

## 2 The Notion of Annotation

Currently, there is no consensus about what an annotation is. One of the most common definitions is that an annotation is an object that describes or says something about other object of information, constituting descriptive information of any type about objects. Usually this is understood as a piece "attached" to other piece of information, but an annotation is not limited to that interpretation. We can also see the process of annotating as the process of building *relationships* between descriptions. We can note that this last interpretation of annotation is plausible since the annotations are objects of information which can be seen as equivalent[1] to the objects being annotated; since they are both descriptions.

To focus on relationships rather than in "attached data" is a convenient interpretation for some cases. Particularly in this work we refer to descriptions and relations between information objects as the same.

Annotations can be hard to capture, yet worth to keep. It is unclear to a person creating an annotation, when and in which context that particular annotation will be relevant again. Because of this uncertainty, we want to keep contextual information in form of meta-data about the relationships between the descriptions. Which contextual information is the best to use, remains an open question, the most used are the five "who, what, when, where, how" or the "who, what, when, where, why" dimensions [19]. We will treat these "associations" as first class citizens, used to discover, understand and manage our stored objects, exploiting annotation processes as the association of things, complementing categorization or selecting resources to "attach" meta-data such as descriptions, properties, type etc. We expect that focusing on the representation of "associations" instead of the attached descriptions potentially will leverage the complexity of our underlying models that focus on personal data management. For example, metadata about relationships can directly support associative browsing, where we can analyze and cluster the information to find related items.

## 3 The Notion of Description

But how we describe our reality through a computer? People naturally refer to their "reality" using distinct languages and notations. Those languages allow them to say things about their perceived world and finally share information with others. When multiple persons want to agree about some observed

[1] Although the annotated object and the annotation itself may be represented in a different way.

subject, they may decide to interact sharing descriptions collectively, trying to converge into useful representations, which can improve their communication means. Within the Web, multiple artifacts (documents) are currently used to "hold" those descriptions, which are used to incrementally convey into shared concept definitions or vocabularies. These artifacts can vary from wikipedia web pages to web ontologies, depending on the degree of formalization or "precision" required to describe the observation subjects. It is important to note that these artifacts are constructed gradually and usually are based on the combination of previous description artifacts.

When we talk about describing our perceived world we frequently refer to semantics. Semantics classically concerns a triadic structure comprising a symbol (how some idea is expressed), a conceptualization (what is abstracted by someone from reality) and a referent (the observed thing or concept) [13]. According to Peirce, this triadic structure is indivisible, meaning that people cannot be left out of this process.

In the web domain, sophisticated systems have been constructed aiming to capture and represent the "semantics" that are relevant to a web community. To better describe our observation subjects we choose distinct mechanisms, just to name a few we can make use of multimedia content, tagging or formal models, which allow us to reach distinct degrees of expressiveness or "precision". Additionally we can refer to the observation subjects using universal identification schemes such as URLs, which we use to retrieve documents from the web in a decentralized environment. This is a powerful and unique mechanism that allows us to get distinct representation variants starting from a symbol, and probably it is one of the most important steps towards the semantic web. We can say that today we count with symbols and identified referents which allows us to share concept descriptions across the Web.

However, we still have difficulties to model simplifications of the third component of peirce's triadic structure, which refer to the conceptualizations. Usually, in computer science literature the importance of human interpretation is left out, which is natural since we don't count yet with logic systems capable of characterizing the interpretation made by a human. However, interpretivist research [4][27] can work together with the predominantly positivist research tradition in information systems, which is not necessarily sufficient to deal with data from one individual.

## 4 The Notion of Context

But how we can characterize a conceptualization? how can we represent our interpretations in a way that a computer can make use of them?. In this work we don't aim to answer those questions, instead we will experiment with models which make use of the notion of *personal context*, in order to model artifacts which can be used and observed to reach efficient retrieval of our personal data, using knowledge representations that make sense to us.

We can observe in the work of Santini et al. [20] that we cannot encode the semantics of a document independently of the act of human interpretation, which is a major limitation. As an alternative, the notion of context is introduced as an essential ingredient to determine the meaning[2] of a document. Santini et al. proposes to observe and measure how certain document *changes the context* of the user, and use that information to archive personal context-based retrieval. Here, context is formalized to a certain degree using a similar technique to the semantic maps WEBSOM [8]. The semantic maps represent a context by means of self organizing maps in the euclidean space of words, which are used to reformulate user queries by means of deforming the queries according to this context. After the query is transformed, it can be "projected" in traditional information retrieval representations. This approach was tested with promising results [20] showing an approach that will be tested in the early stages of this work, in conjunction with emerging semantics.

## 5 The Notion of Emergent Semantics

If we are not isolated from the online world, to more or less degree, we can influence and we are influenced by others. We can say that many of the distinct terminologies or categorizations that we use to organize our objects are learned in community. Moreover, we copy, transform and combine things that we get from others. It seems convenient to use those organic and common ways of organization to manage our own personal data. One problem is that to follow the dynamics of online communities is not inherently simple, specially when we see these online communities as groups of people determined by their own interactions. In this work we don't want to address explicitly the dynamic nature of those interactions, instead we want to capture to certain degree the terminology or "semantics" used within our interplay by means of incrementally trade knowledge within our communities. Interesting work have been found in the field of *emergent semantics*, where Cudr-Maurux [2] presents several techniques to allow observation and capture of semantics from interacting agents in the wild. In this approach, global interoperability is seen as emerging from collections of dynamic agreements between autonomous self-interested agents. Those agents interact following a Peer-to-Peer paradigm, where agents are allowed to create mappings in order to interact. These mappings are the ones that determine the formation of semantic neighborhoods of agents.

In our experiment, instead of simulating an interaction network of agents, we will make use of simple artifacts, which we call web *semantic blackboards* [23], which are freely used by users as main interface. These blackboards are extensions of a semantic wiki web page, which is used as a playgrounds where the participants are allowed describe some observation subject collaboratively, making use of distinct description mechanisms and formalisms. We aim to

[2] According to Santini et al, it is only possible to formalize meaning only in the extent that is possible to formalize context.

exploit this description diversity to improve custom query results and exploratory search, using structured descriptions in conjunction with others more suitable for result recognition (i.e photograph). In general terms, a participant is allowed to subscribe to multiple blackboards and contribute content, aiming to converge into acceptable conceptualizations "agreed" by his peers. A user can "collect" distinct blackboards, constructing a network that he can he bind directly with his own personal data in form of folders, files, mails etc. Participants of each blackboard are allowed to (i) contribute content from their own private spaces, building implicit relationships. In the same way, participants can (ii) borrow descriptions from others or (iii) relate external descriptions to their own descriptions (iv) or relate blackboards using relationships such as causality, location, function etc.

Within this interplay, we expect to count with a increasing number of relationships between observation subjects, forming a network of blackboards. Having a considerable amount of transitive or symmetric relationships such as "part of", "is a", "same as" etc, can be beneficial in order to observe emerging semantics. In early stages of this work we will experiment with some of the techniques presented in [3], which include cycle analysis and probabilistic message passing. We expect that analysis of transitive closures between blackboards would be a valuable tool to provide feedback to the communities, increasing their awareness. For example, if some path of multiple "is a" relations forms a cycle within a network of blackboards, we can suspect that the involved blackboards are sharing similar semantics. We can mark these blackboards for semantic reconciliation, potentially leading to agreed inter-blackboard meaning convergence. This can result into *merging* multiple blackboards into one, or *branching* a new observation subject(blackboard) to be described, for example an "abstraction" of the blackboards of the cycle. In the same way, failed convergence processes can give us clues about sub-optimal relations or conflicting descriptions within "semantic neighborhood" which can result in blackboards that diverge(or branch) into new ones. Other example could be the generation of "composite blackboards" for blackboards with multiple incoming "part of" relationships.

This mechanism of diverging and converging, provides a strong support for non-linear development which is already used widely in large scale scenarios such as collaborative software coding sites, benefited by distributed version control systems (DVCS[3]). Divergence of blackboards allows to follow an "organic" approach which doesn't rely principally in agreements. For example, it is useful when a group of participants have irreconcilable conceptualization (interpretations) conflict about some description subject (i.e. unicorns doesn't exist vs they exist), or having categorizations that are just not convenient for some participants. In these cases of divergence and convergence, participants may choose to keep a complete traceability, which is a valuable tool to improve management, measure formalization costs and understanding our own neighborhood evolutions.

[3] `http://en.wikipedia.org/wiki/Distributed_revision_control`

## 6 Preliminar Experiment

In order to study possible the benefits of combining (i) our local context with (ii) the observed emergent semantics of our close communities, we have designed an architecture which will allow us to perform a cost-benefit analysis about the total description formalization efforts within the blackboards. The configuration of this experiment is in early stages, although we have distinguished and constructed some of its main components.

The applicability and effectiveness of the experiment will depend largely on how we represent our information artifacts. We can represent them using a variety of technologies, but for the sake of simplicity we will limit this experiment to use only RDF[4] as data model, with a limited set of representations. We can categorize these representations in families according to their intended use and level of precision. These preliminary families are: (i) pragmatical representation family, which include vocabularies that support communication between users, such as the representation of dialogues (ii) empirical representations, that refer to vocabularies that constitute high precision artifacts to observe subjects, for example a photograph or geographical coordinates (iii) semantic level, which refer to the models or schemes created via human abstraction, such as a personal folder structures, or public "ontologies".

On the early stages we intend to support subject descriptions via structured note taking. The core process from notes to a document can be described as steps in a knowledge maturing process [10]. These activities will be integrated into a classical semantic wiki, because of its three defining characteristics: "easy contribution", "easy editing", and "easy linking" [26]. The tool itself needs to be extended in order to sort items, manage subscribed blackboards and link them to personal data items. In the same way we have to provide notification systems to increase user's awareness regarding detected semantic patterns. The overview about the "collected" blackboards of a user can be provided using a spatial layout. A good example of these layouts are IMaps [5], that can be seen as a large blackboard where smaller blackboards are positioned like post-its but also nested in each other. The preliminary technologies used include (i) DVCS implementations such as JGIT[5] to keep the traceability of the blackboard dynamics and to directly support divergence and convergence capabilities, (ii) RDF triplifiers[6] to extract structure from containers such as files or web pages and (iii) the nepomuk[7] framework to keep track of user digital "residues" within their desktops. With the use the nepomuk framework we benefit from a (iv) Personal Information Model Ontology (PIMO) editor component like in [18], where each user augments his own PIMO manually or by means of implicitly infering metadata from his resources, according to his own context. We can

[4] http://www.w3.org/TR/rdf-syntax-grammar
[5] http://eclipse.org/jgit
[6] http://code.google.com/p/any23/
[7] http://nepomuk.kde.org/

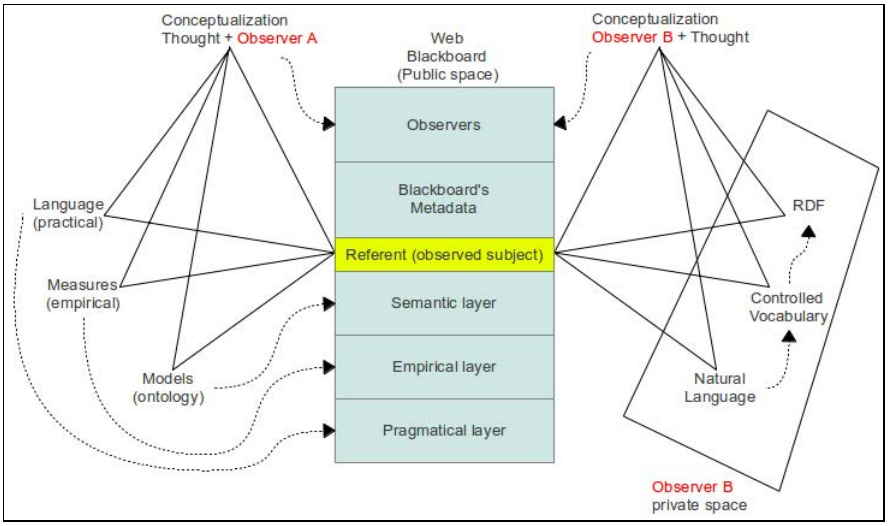


**Fig. 1.** Example of inter-subjective space between two participants, regarding some observation subject. "Conceptualizations" try to converge into a shared space, through a pragmatic, empirical and semantic levels. These levels can be seen as language, examples and models respectively. Fuzzy representations can also mutate into more formal ones, handling a continuous spectrum of description mechanisms which follow distinct degrees of structure.

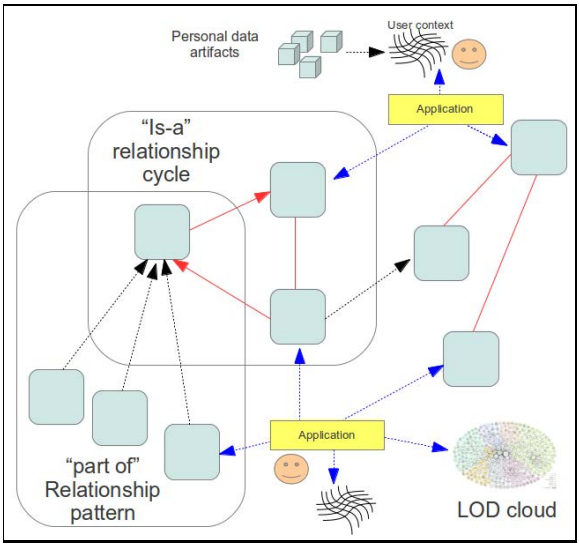


**Fig. 2.** Setup of an experiment where distinct observers commit to a set of interconnected blackboards. Patterns such as cycles are used to provide feedback to the users. Each application instance interact with blackboards, a personal context and external sources such as Linked Open Data.

automatically associate selected blackboards with the user PIMO concepts and the conceptual data structures [25] used in the nepomuk framework, in order to share distinct degrees of structuredness and formalization within a community.

## 7 Conclusions

Through this document, we explored notions such as personal context and emergent semantics, describing artifacts such as blackboards that can diverge and converge in order to support meaning evolution. We have presented the preliminary setup of a experiment which aims to measure possible benefits of using our (i) personal data context and (ii) the emergent semantics of our close peers, in order to retrieve our personal data. We can see this setup as a set of "semantic bridges", where our own semantics are "hooked" with the semantics of others. We expect that this mechanism facilitates the retrieval of personal data segments, since it will be organized in terms of the semantics of our communities.

Additionally we track the evolution of our shared descriptions, constituting artifacts which may provide significant insight about the evolution of our semantics, which is a valuable asset for the study of this approach.

In this work we don't aim to distill global semantics. Instead we want our own semantics, taking as hypothesis that they are incrementally constructed within our close communities. We will benefit from local schemes to reflect our "ideologies" or "realities". This approach can be considered as a single particular step within a wider production chain of knowledge, starting from the individual to its society and viceversa.

**Acknowledgments.** The research described in this paper was partially sponsored by the INNOViris Open Semantic Cloud for Brussels (OSCB) project.

## References

1. Aberer, K., Ouksel, A.: Emergent semantics principles and issues. Database Systems for 2 (2004)
2. Philippe cudr E-mauroux. Emergent Semantics: Rethinking interoperability for large scale descentralized information systems (2006)
3. Cudré-Mauroux, P., Aberer, K.: Belief Propagation on Uncertain Schema Mappings in Peer Data Management Systems 1 (001935) (2003)
4. Falkenberg, E.D., Hesse, W., Lindgreen, P., Nilsson, B.E., Han Oei, J.L., Rolland, C., Stamper, R.K., Van Assche, F.J.M., Verrijn-Stuart, A.A., Voss, K.: A Framework of Information System Concepts - The FRISCO Report. International Federation for Information Processing WG 8.1 4 (1998)
5. Haller, H., Abecker, A.: iMapping A Zooming User Interface Approach for Personal and Semantic Knowledge Management. Knowledge Creation Diffusion Utilization 119–128 (Autumn 2010)
6. Hearst, M.A.: Clustering versus faceted categories for information exploration. Communications of the ACM 49(4), 59 (2006)

7. Jones, W., Phuwanartnurak, A.J., Gill, R.: Phuwanartnurak, and Rajdeep Gill. Don't take my folders away!: organizing personal information to get ghings done. In: CHI 2005 Extended Abstracts, pp. 1–4 (2005)
8. Kaski, S.: Computationally efficient approximation of a probabilistic model for document representation in the WEBSOM full-text analysis method. Neural Processing Letters, 139–151 (1997)
9. De Leenheer, P.: On Community-based Ontology Evolution. PhD thesis (2009)
10. Maier, R.: Characterizing knowledge maturing: A conceptual process model for integrating e-learning and knowledge management. In: Professional Knowledge Management, Wm 2007 (2007)
11. Marshall, C.C.: How people manage personal information over a lifetime. Personal Information Management (2007)
12. Mika, P., Potter, T.: Metadata Statistics for a Large Web Corpus, pp. 1–4 (2012)
13. Moore, E.C.: Writings of Charles S. Peirce: A Chronological Edition. University Press, Bloomington (1982)
14. Mühleisen, H., Bizer, C.: Web Data Commons Extracting Structured Data from Two Large Web Corpora. Distribution, 2–5 (2012)
15. Nadeem, D.: From Philosophy and Mental-Models to Semantic Desktop Research: Theoretical Overview. Proc. I-Sematics (2007)
16. Nonaka, I.: The Knowledge-Creating Company. Research Policy 26(4-5), 598–600 (1997)
17. Oren, E., Völkel, M., Breslin, J.G., Decker, S.: Semantic Wikis for Personal Knowledge Management. In: Bressan, S., Küng, J., Wagner, R. (eds.) DEXA 2006. LNCS, vol. 4080, pp. 509–518. Springer, Heidelberg (2006)
18. Papailiou, N., Apostolou, D., Panagiotou, D., Mentzas, G.: Exploring Knowledge Management with a Social Semantic Desktop Architecture. In: Wagner, R., Revell, N., Pernul, G. (eds.) DEXA 2007. LNCS, vol. 4653, pp. 213–222. Springer, Heidelberg (2007)
19. Ranganathan, S.R.: Elements of Library Classification. Asia Publi. (1945)
20. Santini, S., Dumitrescu, A.: Context as a non-ontological determinant of semantics
21. Sauermann, L.: The Gnowsis Semantic Desktop for Information Integration
22. Sauermann, L., Bernardi, A., Dengel, A.: Overview and Outlook on the Semantic Desktop
23. Vasquez, C.: Blackboard Data Spaces for the Elicitation of Community-based Lightweight ontologies. In: IEEE/ACM ASONAM 2012 (2012)
24. Völkel, M., Abecker, A.: Cost-benefit analysis for the design of personal knowledge management systems. In: ICEIS, pp. 95–105 (2008)
25. Völkel, M., Haller, H.: Conceptual data structures for personal knowledge management. Online Information Review 33(2), 298–315 (2009)
26. Völkel, M., Schaffert, S.: Personal Knowledge Management with Semantic Technologies. Technologies for Semantic Work, 1–20 (2008)
27. Walsam, G.: ISR emergence of interpretivism in IS research.pdf (1995)