On the Weaving of Documents and Formal Knowledge, a Personal Knowledge Approach

Pierre-Antoine Champin, Yannick Prié

Université de Lyon, CNRS
Université Lyon 1, LIRIS, UMR5205, F-69622, France
[pierre-antoine.champin|yannick.prie]@liris.cnrs.fr

Abstract. In this problem analysis paper, we address the problem of weaving documents and formal knowledge. With the emergence of increasingly integrated technologies for handling conjointly these different kinds of information, we claim that this question deserves attention and careful study from the knowledge engineering perspective. We advocate for a global framework that would help to conceptualise, in an integrated way, the interrelations between documents and knowledge. Furthermore, we sketch such a framework from the perspective of *personal knowledge*, which we claim is the place where the weaving actually occurs at the user level. We propose to consider three *poles* (taking an electromagnetic metaphor) in personal knowledge: data, schema and presentation. Elaborating one's personal knowledge is a matter of deciding how to dispatch information around those poles. But it also requires to continuously manage their evolution and *circulations* from one pole to another while maintaining consistency, in order to fulfil the needs at hand during the user's activity. Such an approach appears promising both for analysing existing work, and building applications which enable the circulation of personal knowledge, such as our video annotation system Advene.

1 Introduction

As the digital world gets increasingly connected, information gets both more fragmented and sharable, and computerised devices act more and more as memories and supports to various activities. Technologies such as those related to the Web and the Semantic Web are leveraging the way formal knowledge and data on the one side, and documents on the other side are considered and weaved in various applications.

In this problem analysis paper, we claim that such weaving of documents and knowledge is actually a fundamental question that deserves attention and careful study from the knowledge engineering perspective. We advocate for a global framework that would help to conceptualise, in an integrated way, the interrelations between documents and formal knowledge. Furthermore, we sketch such a framework from the perspective of *personal knowledge*, which we claim is the place where the weaving actually occurs and continuously evolves. We propose to consider three *poles* (taking an electromagnetic metaphor) in personal

knowledge: data, schema and presentation. Elaborating one's personal knowledge is a matter of deciding how to dispatch information around those poles. But it also requires to continuously manage their evolutions or *circulations* from one pole to another while maintaining consistency, in order to fulfil the needs at hand during the user's activity. We will show that this approach is promising both for analysing existing work and building personal knowledge circulation aware applications, such as our video annotation system Advene.

In the section 2, we will further develop our motivations, introducing the document / knowledge weaving question and identifying related works. In section 3, we will define personal knowledge and propose its modelling in three poles, and its evolution as knowledge circulation around these poles. We will then discuss in section 4 the efficiency of that model to analyse solutions proposed in the literature and to propose new ones, before we conclude.

2 Motivations: Knowledge and Document Weaving

In this section, we first present the general context of our proposal, which relates to knowledge and document weaving, before addressing three specific cases where this weaving actually occurs.

2.1 Knowledge and Documents

Knowledge-Based Systems. One historical goal of artificial intelligence has been to capture knowledge into a digital form that can be handled by computers, that are in turn expected to mimic the reasoning of the original expert. This has lead to the definition of a wealth of formal models for knowledge representation. Among those, conceptual graphs [36] and description logics [3] have largely inspired recent efforts in the domain of the Semantic Web [5].

The problem of knowledge-based system is the high cost of acquiring and maintaining a knowledge base. Indeed, the first bottleneck of knowledge-based system is the elicitation of knowledge representing large and complex domains, from numerous and heterogeneous sources. This difficulty is stressed by the very terms traditionally used to describe that task, such as "knowledge extraction" or "knowledge capture". Furthermore, the application domain is in most situations a moving target; knowledge about it is constantly created, revised or deprecated. That cost is therefore not limited to the bootstrapping of the system, but enduring for its whole life-cycle.

Although knowledge based-systems have their success stories, those difficulties have been a significant hindrance to their wide adoption, and even of the Semantic Web as it was originally envisioned.

Knowledge Management and Documents. Documents are the main vector of knowledge and, as such, are central to organisations. Another trend in knowledge management has therefore been to study documents, and how they convey meaning and practices through their structures and genres [9, 16, 35]. The goal

of tools such as digital libraries and search engines is to provide users with the knowledge they need by retrieving the appropriate document (rather than by implementing that knowledge directly in the system).

This has become all the more efficient that documents have gradually mutated from material objects to digital ones, hence an abundant literature studying the nature, structure and specificity of digital documents. For instance, since IBM's Generalized Markup Language, there has long been a clear distinction between a document logical and physical structures. The former focuses on the content of the document, while the latter focuses on how the document will be presented to a reader, and both can be handled with a relative independence.

Nowadays, documents are a major component of organisations' information systems, managed jointly with databases and, in some cases, formal knowledge bases. The latest trend aims at integrating users' *social networks*, in order to retrieve not only useful documents and data, but also knowledgeable people, that could be helpful for a user's task at hand.

Weaving Knowledge and Documents in the (Semantic) Web. Although the research communities studying documents, on the one hand, and formal knowledge, on the other hand, were relatively separate for some time, those fields have been brought closer recently under the influence of the Web and Semantic Web.

Primarily focused on documents (HTML and XML being descendants of SGML [21]), the Web has also drawn the attention of the knowledge representation communities (with technologies such as RDF or OWL [8, 22]). Since its beginning in the early nineties, and much more vividly with the emergence of the so-called Semantic Web — which has been slowly but strongly maturing for more than a decade — the Web has been an intent to integrate various information systems into a unified framework, where humans and machines can consume or create information.

Documents and formal knowledge on the (Semantic) Web are therefore becoming fundamentally weaved with each other. Moreover, with the growing momentum of mobile computing, where information is preferably stored "in the cloud", this aspect of the Web is pervasively extending to a wide range of applications: office productivity tool, social network, access to or population of knowledge/data bases, activity streams, etc.

2.2 Three Cases of Document and Knowledge Weaving

From Documents to Knowledge: Annotations and Extraction. One of the first proposed used of Semantic Web technologies was the semantic *annotation* of the content of existing documents in order to elicit the knowledge they contained [25, 27]. Those efforts faced the problem of maintaining the consistency of annotations with changing documents.

Other efforts intend to help the author include formal knowledge in their documents while they are writing them. The WiCK project [12] instruments existing office productivity tools to produce knowledge-rich documents; SALT [24] pursues the same goal for LaTeXdocuments, proposing to annotate the rhetoric

relations between document parts; semantic wikis [4, 29, 34] aim at leveraging knowledge formalisation with the ease of use of wikis.

Standing midway between annotation and edition, microformats [28] propose to overload HTML attributes with a specific semantics (e.g. business cards, events). RDFa [38] extends this concept to include arbitrary RDF in HTML documents, allowing author to annotate their document with any kind of formal knowledge for which an ontology exists. [38] propose a WYSIWYG interface in which users copy and paste RDFa annotated texts, and seamlessly organise formal knowledge by standard document manipulations.

Annotating documents is not the only way to express their semantics. Knowledge engineering and Information Retrieval has a history of using natural language processing, from the recognition of named entity in a document [18] to the building of an ontology from a whole corpus [20, 10]. Another trend is to use the regularity of existing document format to extract their semantics. This is the purpose of the GRDDL recommendation [15], standardising a way to convert documents in legacy formats to the more reusable RDF format, using transformation languages such as XSLT [14]

From Knowledge to Documents: Presentation. Presenting formal knowledge to the user has always been a problem faced by knowledge based systems and, to some extent, a hindrance to their acceptance. This subject has been revisited in the field of Semantic Web, trying to reiterate the initial success of the web by making information easy to publish and to *consume*.

The main problem faced by tools presenting knowledge is to hide the complexity of the underlying data. One way is to propose *faceted browsing* [26] where relevant filtering options are dynamically created from the data and proposed to the user. Another way is to propose dedicated views for specific kinds of data, like the FOAF pane in the Tabulator semantic browser [6] or brainlets in DBin [39]. Presentations framework for RDF [13, 30] provide ways to define such specific views (as lenses or templates).

An interesting trend in that domain consists in proposing hybrid languages playing different roles in the life-cycle of documents and knowledge. XTiger [17] is at the same time a template language, a schema language and transformation language. XSPARQL [1] on the other way, can be used either to query or present data from both XML and RDF sources.

Semantic Desktops. The field of semantic desktop [23, 33], as an extension to classical Personal Information Management (PIM), is probably the most prominent one addressing the weaving of formal knowledge and documents in order to sustain personal activity. The main idea is to help people tackle with the heterogeneity of their personal information; this is achieved by keeping track of the links that exist between different kinds of information, using semantic web technologies. This linking information can be provided by instrumenting the underlying applications [31] or by inferring it from the observation of the user's behaviour [37].

Although semantic desktop tools have been proposed for several years, they are still largely outnumbered by "classical" PIM tools. This can be explained by the mitigated results of user studies aiming at measuring the actual benefit of those tools [19, 7]. In [40], the authors analyse this situation and propose eight requirements for PIM tools, among which is the need to smoothly integrate with user's working habits. The recent integration of the Nepomuk with the KDE desktop environment (`nepomuk.kde.org`) can be seen as a move in that direction.

2.3 Problem Statement

As we have seen, the current convergence and weaving of documents and formal knowledge, catalysed by the pervasiveness of the Web, has been tackled in many different ways by different communities (document engineering, information retrieval, knowledge representation, *etc.*). Although some proposals are crossing the boundaries of those fields, it seems to us that there still lacks a *unifying framework that would allow to consider the document / knowledge weaving question as such*. Such a framework would not only provide hints to compare and assess those different efforts, it would also help identifying concepts for imagining new kinds of tools and applications integrating the weaving in their core.

In the remaining of this article, we will sketch a first proposal towards that objective by considering *Personal Knowledge* and *Personal Knowledge Management* as interesting angles for tackling the question of the interrelations between documents and knowledge at the very locus where they are read, created and manipulated, that of the individual.

3 A Proposal for Personal Knowledge Modeling

3.1 Personal Knowledge

If we consider knowledge to be any information supporting the achievement of one's task, it is then trivial to state that knowledge is almost everywhere in the digital world. Ontologies and knowledge bases are examples, of course, but so are digital documents, databases or log-files. Different tools allow to handle those different kinds of knowledge, and users are generally not too concerned about the differences —as long as the tools efficiently fulfil their needs [11]. Whenever they do not, users will link, move or copy information across tools, changing the nature of the underlying information, creating an intricate network of knowledge.

As a definition, let us state that *personal knowledge encompass the more or less structured data that results from the one's activity and at the same time sustains it*. Any system that can be used to define, organise, visualise and share such data fits in the category of personal knowledge management tools. That includes semantic web and semantic desktop applications, but plain address books, file or document management systems, personal mobile assistants also fit in that category. Moreover, personal knowledge management is focused as much

on the way those tools are used as on the way information is linked, moved or copies across them.

In this section, we propose the basis of a framework for considering different kinds of personal knowledge, how they relate to each other and how knowledge evolves and *circulates* between those different kinds. The next section will be devoted to discussing the usefulness of our proposal regarding the problem of document and knowledge weaving.

Let us consider in Figure 1 a user (Alice) managing her personal information through a computer application (either desktop or web-based). As we pointed out, that information is composed of a mix of structured data, documents and formal knowledge. As any digital (immaterial) information, it is not handled directly; presentation and editing are respectively mediated by processes (represented by the gears on the figure). Those processes are themselves parametrised by dedicated meta-data. On the figure and in the following of the paper, we will use the term *style-sheet* to identify meta-data used to present information to the user, and *forms* to identify meta-data used to feed the underlying data with input from the user. Although such terms may vary across contexts and applications, the concept of presentation- and input-specific meta-data is generally valid.

Figure 1 also depicts a third process, looping back to the data. Such processes are exploiting the structure of the data, according to another kind of meta-data, that we will call *schemas* (the term "ontologies" is also popular in the field of formal knowledge). Schema-based processes are used to check or even transform the data, according to the intended semantics of their structure. They can for example check integrity constraints, add default values to unspecified fields, or even infer new data from the one that is given. Note that this family of processes defines the *operational semantics* of the data. For ontologies, it is expected that this semantics is complete and sound with respect to the declarative semantics of the formalism in use [3].

Also note that, in many situations, the meta-data used by those processes is either built in the application or enclosed in configuration settings, out of control of the user. Thus, they are not part of that user's personal information. However, we consider them as part of the personal *knowledge*, as they influence the way users produce and consume their personal information. Indeed, even if users may not have a complete insight of those meta-data, they are nevertheless acquainted to them. This acquaintance is acquired through using the application and interacting with personal information *through* the related processes.

3.2 Three "Poles" for Personal Knowledge

We can therefore simplify Figure 1 into Figure 2 : the circle labelled D represents the Alice's data (documents, personal information, etc.); the circle labelled S represents the schemas outside her control; the circle labelled P represents the presentation-related meta-data (style-sheets and forms), also outside her control. The links between the circles represent the dependencies between the different

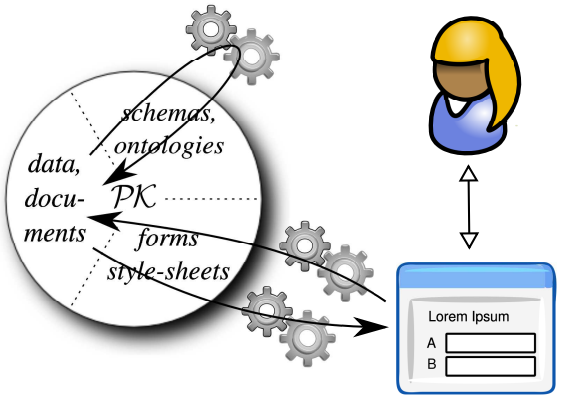

Fig. 1. Personal knowledge workflow: the user interacts with data through style-sheet- and form- based processes while schema-based processes take advantage of the structure of the data

kinds of information : Alice's data are constrained by the schemas, and, to some extent, by the presentation (a form may not allow her to input any data structure that would be acceptable by the schema). In addition, the P and S meta-data are also inter-dependant : style-sheets and forms are designed in conformance with the target schema.

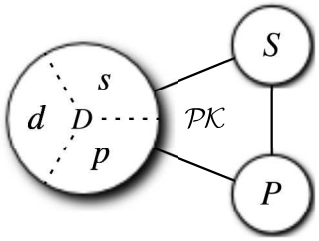

Fig. 2. Personal knowledge and the three poles

Not all schemas, style-sheets or forms are read-only, though. Many applications give users the opportunity to customise the way data is presented, processed or even input. For example, Web Browsers allow users to override the presentation of pages, from changing the zoom to using a personal CSS style-sheet. Typical word processors make it possible to define templates, which act as schemas (providing their own styles), style-sheets (defining how styles are presented),

or even forms (preventing the edition of some parts of the document). Finally, in more complex applications like MS-Access (`office.microsoft.com/access`) or Protégé (`protege.stanford.edu`), users are free to define how their data are structured (with schemas / ontologies), processed (with queries / axioms), presented (with views and reports / forms) or input (with forms).

As a consequence, Alice's data as contained in D is not absolutely homogeneous. Taking an electromagnetic metaphor, we consider D to be "polarised" by the way that data might be used through the processes of Figure 1. We name those three poles d, s and p, after the knowledge container (D, S or P) to which they relate.

Note that these poles do not split the user's data and documents into crisp categories, that would tentatively represent their inherent nature. For example, a document containing Alice's resume is closer to the d pole, as it is not meant to be used as meta-data for any of the described process. However, should this resume be described in PDF or HTML, it would tend to be on the p "side" of that pole, as those formats are oriented towards presentation. On the other hand, the same resume annotated with a dedicated microformat, or expressed in a dedicated XML form, would rather be on the s side, being more structured and presentation-independent. Hence, the poles merely represent the fact that some data are *prone to be used* in certain ways but, as we will see in section 3.3, they may have *several* typical uses and can even be diverted from those typical uses.

3.3 Knowledge Evolution and Circulation

Let us now consider a CSS style-sheet, that Alice would have copied from the web for applying to her on-line HTML resume. Its primary use is as presentation-related meta-data, so it is closer to the p pole. At some point, her resume does not look like it should; by inspecting the CSS style-sheet, Alice understands that she should have used a `<div>` element rather than a `<p>`. Alice therefore tweaks the configuration of her text-processor in order to enforce this new constraint, turning part of the presentation knowledge embodied in the style-sheet into schema knowledge. Later, Alice decides that using `<div>` was a better idea after all, and decides to change the CSS style-sheet. At that point, the CSS style-sheet has temporarily moved towards the d pole, becoming a document of its own, presented through other style-sheets (syntax colouring) and its syntax being checked through its own schema.

This example stresses the facts that 1) a user's piece of data is not linked to a single pole, and 2) evolves in time regarding how it relates to the three poles. We extend our electromagnetic metaphor by calling those evolutions *knowledge circulations*. Most of the related work described in section 2 focus on a particular kind of circulation. We argue that our framework allows us to systematically explore the different kinds of circulation. This is synthesised in Figure 3.

Figure 3(a) describes circulations around the poles in the D knowledge. Some of them are illustrated in the examples above. We believe that personal knowledge elaboration heavily relies on such circulations: data, documents, and other

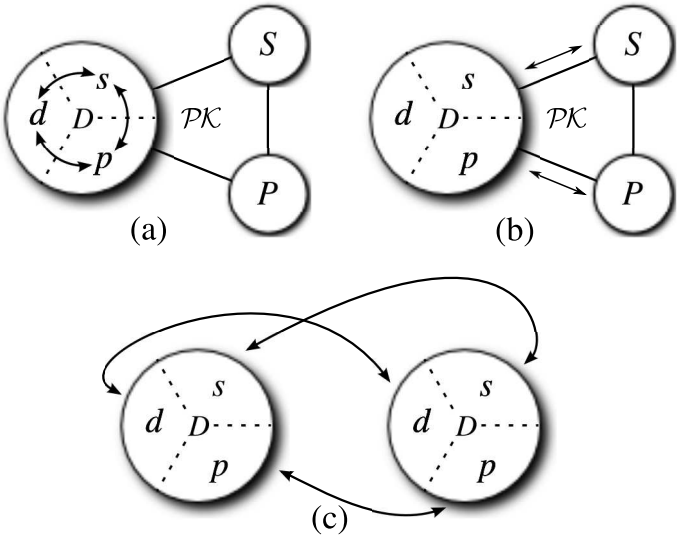

Fig. 3. Personal knowledge circulations

pieces of personal information are constantly adapted and put to new uses in order to fulfil the requirements of the task at hand. It is also important to note that those circulation may have very different *time scales*: Alice's style-sheet is only temporarily considered as a document, while its turning into a schema (for enforcing constraints expected by the style-sheet) may be much more durable.

Figure 3(b) describes circulations between D and the two other knowledge containers S and P. Circulations towards D represent the *appropriation* by the user of previously static meta-data, usually for the purpose of adapting it, possibly through different poles (as Alice did with the CSS style-sheet). Circulation in the other direction, on the other hand, may represent the *emergence* of common practices from individual ones —although this brings us outside the realm of personal knowledge *per se*, as we will discuss later.

Finally, Figure 3(c) describes the *sharing* of personal knowledge among individuals. We insist here on the fact that different tasks require the sharing of different kinds of personal knowledge. Alice may, for example, send her customised style-sheet to a friend wanting to write his own resume; on the other hand, she might send the resume (without the style-sheet) to a recruiting agent feeding his own application with the structured data, for further processing. Of course, this kind of inter-individual circulation is tightly coupled with other kinds of circulation. For example, a common practise would be for Alice to share her style-sheet together with her own resume, not for its content, but as a template for correctly using the style-sheet. Here, Alice's resume circulates into a kind of schema, for the sake of sharing the style-sheet.

We have presented our framework for considering personal knowledge and knowledge circulation. In the following section, we will discuss the benefit of this framework for analysing existing work related to knowledge / document weaving and proposing future research directions.

4 Discussion

4.1 Analysing Related Work

Most of the related work described in section 2 focus on a particular kind of circulation, mainly the kinds described by Figure 3(a).

Work in the field of semantic desktop are mainly focused on the s pole, attempting to make explicit the structures relating different items of personal information, from what is available in the d pole. Allowing users to define their own ontology (what [33] call the PIMO) in order to organise their personal information can also be seen as a circulation from d to s —although it is not clearly discussed how editing the PIMO can be leveraged by existing practices embodied in the actual data. It is interesting to see how these efforts are almost exclusively addressing clockwise circulation, favouring structured data and formal knowledge over presentation. This imbalance may explain the lack of adoption for semantic desktop tools pointed out by [40]. Some of the requirements they propose to solve that problem (transparency, respecting user's habits) can be seen as a plea for counter-clockwise circulation and the importance of the p pole. The text-based interfaces proposed in [7] is also a step in that direction, aiming to keep the presentation simple and easy to edit.

Semantic annotation can also be seen as sustaining circulation around the d pole from its p side to its s side. Other proposals (SALT, microformats, semantic wikis) sustain this circulation in both directions, as they can be used either to annotate existing documents, or as a presentation-friendly ways to describe formal knowledge from scratch[1]. Interestingly, the very principle of microformats can also be seen as a circulation from the p pole to the s pole, as presentation-related attributes were diverted from their main purpose to be attributed structural semantics. The same remark applies to XSLT, initially designed for style-sheets, and now used as a generic tool for structural transformation (e.g. by GRDDL).

Not all those proposals are dedicated to a unique kind of circulation, though. Versatile languages such as XSTL, XTiger and XSPARQL are flexible enough to express p or s meta-data, making it easier for data to circulate between these poles, but are not explicitly addressing the problem of circulation, nor are they adapted clearly adapted to the d pole (although XSLT makes it quite easy to make a template from an existing document). On the other hand, the semantic wiki proposed by [4] is based on the notion of a *knowledge continuum*, where presentation-oriented data and more structured formal data (all variants of d) co-exist and evolve with the users' needs, but where meta-data (s and p) are not explicitly addressed.

[1] RDFa advocates insist on it being an alternative serialisation syntax for RDF, rather than an annotation language for HTML

4.2 Proposing New Research Directions

Towards Environments for Managing the Three Poles of Personal Knowledge. We already mentioned applications, such as MS Access and Protégé, allowing users to handle the three poles in a unified environment. Those applications already provide some kind of built-in circulations; for example, they automatically generate forms from given schemas. However, to the best of our knowledge, these circulations have not been the object of research work, unlike circulations in the other directions [2].

In developing Advene (`advene.org`), an application for video annotation, we implemented from the start the three poles. Since digital video is a relatively young medium compared to textual documents, with no well established practices, we were willing to keep the application as flexible as possible in order to let users invent their own practices. This made knowledge circulation relatively easy in Advene, and we have been studying how those circulation sustain the tasks of active readers [32], and how dedicated tools can be provided to supported the circulations they need.

For example, in Advene, a user may start to annotate a movie with free text *annotations*, then realise that she actually uses two different kinds of annotations: one to describe the actions of the characters, and the other one to describe the music. Querying the annotations already present (d pole), she can dispatch them in two sets which will be used to create two new *annotation types* (s pole). She may then *write a page* about emotions in that movie, dragging some snapshots from the movie into that page for illustrations (p pole), then realise that the instants of the movie depicted by those snapshots actually deserve to be more formally identified (i.e. not just as plain images), and turn them into actual annotations (d pole).

We are not implying that every tool should implement all kinds of circulations. However, we claim that the framework proposed in this paper is a valuable tool to systematically identify the gaps in the functionalities offered by a given tool, and for each of them, knowingly determine whether or not it is relevant to the users[2] and if so, how far it should be supported.

From Personal to Social Knowledge. As the term *personal* knowledge implies, our framework as no claim to capture the complexity of social or collaborative scenarios. The sharing circulations depicted in Figure 3(c) merely represent inter-individual circulation, and the tentative interpretation of that in Figure 3(b) as emergence still needs to be investigated.

This is one of the goals of the CineCast project (`cinecast.fr`) in which experience from Advene will be reused in the context of film libraries (both physical and on-line) and social networks of film amateurs that share film critics (p), annotations (d) and means of film description s. We will specifically

[2] Application designers should nonetheless be extremely careful before ruling out a kind of circulation, deemed "useless". A number of studies [11, 40] show that users are prone to use exitsing tools in unexpected ways.

study knowledge circulations within the context of various applications along the project.

Re-engineering and Evolutions of Applications. Another interesting direction suggested by Figure 3(3) is to consider the use an *application designer* could make of that kind of circulations.

We presented P and S as meta-data out of the control of the user; this encompass both explicitly shared read-only meta-data (as in the example from Section 3.3), but also implicit meta-data, built in the corresponding processes. In that latter case, appropriation (circulation from S and P to D) would be the extraction (through more or less formalised reverse-engineering) of built-in operational knowledge into a more reusable form. Emergence, on the other hand, would be the deep integration, in the code of the application, of meta-data provided by its users.

As system designers largely control the way their users are able to manage such circulations within their personal knowledge, they can take into account the actual uses of such circulations in further versions or iterations of the system. This can be done either by integrating users-provided meta-data, or on the contrary by opening meta-data that can potentially act as personal knowledge.

Though such ideas need further investigation, we are confident that our approach of personal knowledge can help conceptualise such kind of software re-engineering, which consists in changing the balance between explicit meta-data knowledge and implicit hard-coded knowledge. This might require to introduce a third level in Figure 2: personal data, shared meta-data and hard-coded (application) knowledge, with a greater variety of circulations between them.

4.3 Document and Knowledge Weaving as Personal Knowledge Circulations

It should be clear from what precedes that our tentative to address the problem of formal knowledge / document weaving deals with the consideration that documents are essentially composed by the presentation p of more or less formalised data d according to s, in the context of environments providing additional knowledge P and S. The weaving of documents and formal knowledge is then apprehended with the acknowledgement that 1/ a prolific approach to the question can be centred on personal knowledge, 2/ personal knowledge can be considered everywhere in the three poles we defined, and that documents indeed carry explicit manageable knowledge, and 3/ the associated notion of knowledge circulation is a powerful means of conceptualising the way the weaving actually occurs. We therefore believe that making applications more prone to personal knowledge circulation will help solving that problem.

5 Conclusion and Future Work

In this article, we have first presented our motivating problem, that of formal knowledge and document weaving which occurs more and more in the context

of the (Semantic) Web which technologies provide a unifying substrate for both document engineering and knowledge representation. If numerous works aim at considering part of the question, from knowledge to documents or conversely from documents to knowledge, it appears to us that there still lacks an approach that could encompass those works in a global framework. A promising approach could be to consider the very locus where document and knowledge weaving occurs, what we defined as personal knowledge. We proposed to divide personal knowledge between three "poles" (data, schema and presentation) and to consider the constant evolution, or circulation, of the user's data (ranging from unstructured documents to formal knowledge) around these poles. Describing those poles and how they relate to each other, and to the processes allowing users to interact with their data, we showed that our framework can provide insight on existing tools, as well as guidelines to improve existing tools and develop new ones taking.

We have been applying our framework to the Advene application for video annotation and hypervideo sharing and we will continue to do so while carrying on our more theoretical investigation on our personal knowledge management framework to the document / knowledge interrelations question.

References

1. W. Akhtar, J. Kopecký, T. Krennwallner, and A. Polleres. XSPARQL: traveling between the XML and RDF worlds – and avoiding the XSLT pilgrimage. In *The Semantic Web: Research and Applications*, pages 432–447, 2008.
2. I. Astrova and B. Stantic. An HTML-Form-Driven approach to reverse engineering of relational databases to ontologies. In *Databases and Applications*, Innsbruck, Austria, Feb. 2005. ACTA Press.
3. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.
4. J. Baumeister, J. Reutelshöfer, and F. Puppe. Engineering on the knowledge formalization continuum. In C. Lange, S. Schaffert, H. Skaf-Molli, and M. Völkel, editors, *Workshop on Semantic Wiki at ESWC 2009*, volume 464 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.
5. T. Berners-Lee. Semantic web roadmap. Draft, W3C, Sept. 1998.
6. T. Berners-Lee, J. Hollenbach, K. Lu, J. Presbrey, E. Prud'ommeaux, and M. Schraefel. Tabulator redux: browsing and writing linked data. In *Linked Data on the Web Workshop at WWW08*, 2008.
7. M. Bernstein, M. V. Kleek, M. Schraefel, and D. R. Krager. Evolution and evaluation of an information scrap manager. Florence, Italy, Apr. 2008.
8. D. Brickley, R. Guha, and B. McBride. RDF vocabulary description language 1.0: RDF schema. W3C recommendation, W3C, Feb. 2004.
9. M. K. Buckland. What is a document? *Journal of the American Society for Information Science*, 48(9):804–809, 1997.
10. P. Buitelaar, P. Cimiano, and B. Magnini. Ontology Learning from Text: Methods, Evaluation and Applications. *Computational Linguistics*, 32(4), dec 2006.

11. R. Capra. A survey of personal information management practices. In *Personal information intersections: What happens when PIM spaces overlap?*, Vancouver, BC, Canada, Nov. 2009.
12. L. Carr, T. Miles-Board, G. Wills, A. Woukeu, and W. Hall. Towards a Knowledge-Aware office environment. http://www.springerlink.com/content/auwh813xnajkjnu5, 2004.
13. P. Champin. T4R: lightweight presentation for the semantic web. In S. A. C. Bizer, editor, *Scripting for the Semantic Web, workshop at ESWC 2009*, June 2009.
14. J. Clark. XSL transformations (XSLT). W3C recommendation, W3C, Nov. 1999.
15. D. Connolly. Gleaning resource descriptions from dialects of languages (GRDDL). W3C recommendation, W3C, Sept. 2007.
16. P. Dourish. *Where the Action Is. The Foundations of Embodied Interaction*. MIT Press, oct 2001.
17. F. C. Flores, V. Quint, and I. Vatton. Templates, microformats and structured editing. In *Proceedings of the 2006 ACM symposium on Document engineering*, pages 188–197, Amsterdam, The Netherlands, 2006. ACM.
18. R. Florian, A. Ittycheriah, H. Jing, and T. Zhang. Named entity recognition through classifier combination. In *Proceedings of CoNLL-2003*, page 168–171, 2003.
19. T. Franz, S. Ansgar, and S. Staab. Are semantic desktops better?: summative evaluation comparing a semantic against a conventional desktop. In *Proceedings of the fifth international conference on Knowledge capture*, pages 1–8, Redondo Beach, California, USA, 2009. ACM.
20. L. Gillam, M. Tariq, and K. Ahmad. Terminology and the construction of ontology. *Terminology*, 11(1):55–81, 2005.
21. C. F. Goldfarb and Y. Rubinsky. *The SGML handbook*. Oxford University Press, USA, 1990.
22. B. C. Grau, B. Motik, I. Horrocks, B. Parsia, P. F. Patel-Schneider, and U. Sattler. OWL 1.1 web ontology language: Model-Theoretic semantics. Wroking draft, W3C, 2008.
23. T. Groza, S. Handschuh, K. Moeller, G. Grimnes, L. Sauermann, E. Minack, C. Mesnage, M. Jazayeri, G. Reif, and R. Gudjonsdottir. The nepomuk project-on the way to the social semantic desktop. *Proceedings of I-Semantics*, 7:201–211, 2007.
24. T. Groza, A. Schutz, and S. Handschuh. SALT: a semantic approach for generating document representations. In *Proceedings of the 2007 ACM symposium on Document engineering*, pages 171–173, Winnipeg, Manitoba, Canada, 2007. ACM.
25. S. Handschuh, S. Staab, and F. Ciravegna. S-CREAM - semi-automatic CREAtion of metadata. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pages 358–372. Springer-Verlag, 2002.
26. P. Heim, J. Ziegler, and S. Lohmann. gFacet: a browser for the web of data. In *International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW 2008)*, 2008.
27. J. Kahan, M. Koivunen, E. Prud'hommeaux, and R. Swick. Annotea: an open RDF infrastructure for shared web annotations. *Computer Networks*, 39(5):608, 589, 2002.
28. R. Khare. Microformats: the next (small) thing on the semantic web? *IEEE Internet Computing*, 10(1):68–75, 2006.
29. M. Krötzsch, D. Vrandečić, and M. Völkel. Semantic MediaWiki. In *The Semantic Web - ISWC 2006*, volume 4273 of *LNCS*, pages 935–942. Spinger, 2006.

30. E. Pietriga, C. Bizer, D. Karger, and R. Lee. Fresnel: A browser-independent presentation vocabulary for RDF. In *Lecture Notes in Computer Science*, volume 4273, page 158, Athens, GA, USA, Nov. 2006.
31. D. Quan, D. Huynh, and D. Karger. Haystack: A platform for authoring end user semantic web applications. In *The SemanticWeb - ISWC 2003*, pages 738–753, 2003.
32. B. Richard, Y. Prié, and S. Calabretto. Towards a unified model for audiovisual active reading. In *Tenth IEEE International Symposium on Multimedia.*, page 673–678, Dec. 2008.
33. L. Sauermann, G. Grimnes, M. Kiesel, C. Fluit, H. Maus, D. Heim, D. Nadeem, B. Horak, and A. Dengel. Semantic desktop 2.0: The gnowsis experience. *The Semantic Web-ISWC 2006*, page 887–900, 2006.
34. S. Schaffert. IkeWiki: a semantic wiki for collaborative knowledge management. *1st International Workshop on Semantic Technologies in Collaborative Applications (STICA'06), Manchester, UK*, 2006.
35. S. Schmid-Isler. The language of digital genres - a semiotic investigation of style and iconology on the world wide web. In *33th Hawaii Conference on System Science*, Maui, 2000. IEEE.
36. J. F. Sowa. *Conceptual structures.* Addison-Wesley Reading, MA, 1984.
37. S. Stumpf, X. Bao, A. Dragunov, T. G. Dietterich, J. Herlocker, K. Johnsrude, L. Li, and J. Shen. Predicting user tasks: I know what you're doing. In *20th National Conference on Artificial Intelligence (AAAI-05), Workshop on Human Comprehensible Machine Learning*, 2005.
38. R. Styles, N. Shabir, and J. Tennison. A pattern for domain specific editing interfaces using embedded RDFa and HTML manipulation tools. Heraklion, Crete, Greece, June 2009.
39. G. Tummarello, C. Morbidoni, M. Nucci, F. Piazza, and P. Puliti. Rich personal semantic web clients: Scenario and a prototype. In *Managing Knowledge in a World of Networks*, pages 246–255. 2006.
40. K. Voit, K. Andrews, and W. Slany. Why personal information management (PIM) technologies are not widespread. In *Personal information intersections: What happens when PIM spaces overlap?*, Vancouver, BC, Canada, Nov. 2009.