From Documents to Knowledge Models

Max Völkel
FZI, Karlsruhe, Germany
voelkel@fzi.de

Abstract. *This paper reviews the document concept and compares analogue and digital documents. It introduces knowledge models as a superset of documents and explains the benefits of using knowledge models in Personal Knowledge Management (PKM).*

1. Introduction

Due to the high degree of specialisation in our society, efficient knowledge organisation and sharing has become a critical success factor. Technological developments like written language, the printing press and finally the internet have lowered the costs and accelerated the process of information distribution many orders of magnitude. The packaging format of knowledge – documents – did not change much. Although documents are an established means of communication, their creation is costly, slow and not always needed. Often only small parts of a document are needed to answer a given information need.

This paper introduces knowledge models (KM) as a superset of documents. First the term Personal Knowledge Management (PKM) is defined (Sec. 2). Then I review the notion of a document (Sec. 3) and analyse which characteristics changed when documents became digital (Sec. 4). Knowledge models are introduced (Sec. 5) and compared with documents and related work. Finally the role of knowledge models in PKM is explained (Sec. 6) before I conclude (Sec. 7).

2. Personal Knowledge Management

Instead of defining the term knowledge precisely, I take the view of Haller [9] and focus on **knowledge cues**. A knowledge cue is any kind of symbol, pattern or artefact which evokes some knowledge in a person's mind, when viewed or used. Digital knowledge cues can be stored and retrieved on a computer – while knowledge may not.

Part of this work has been funded by the European Commission 6th Framework Programme in the context of the Knowledge Web - Network of Excellence project, FP6-507482 and the EU IST NEPOMUK IP - The Social Semantic Desktop, FP6-027705. Part of this work has been done in "WAVES -- Wissensaustausch bei der verteilten Entwicklung von Software", funded by BMBF, Germany.

Knowledge is fundamentally created by individuals [29, p.59]. Supporting individuals in their ***personal* knowledge management** (PKM) is therefore crucial.

Pollard [18] sees a trend from central content management to personal content management, shared in a peer-to-peer fashion. Instead of publishing one document for a broad audience, a knowledge worker today has to publish many documents and reports, targeted for small groups or even single persons. Oren [15] concludes in his *overview of information management and knowledge work studies* that one should focus on the individual and give individual users incentive and benefit before focusing on the social network.

3. Documents

A French team of over 50 researchers analysed the term document in depth [16] and gives three co-existing definitions of the term “document”: (1) *Document as form*, where a document is seen mostly as a container, which assembles and structures the content to make it easier for the reader to understand it. (2) *Document as sign*, which emphasizes the argumentative structure of the content. Also, a document that can be referenced acts as a sign for its meaning. (3) *Document as medium*, concentrates on the "reading contract", that is the intention or assumption of the author what will happen with the document.

I see a document as a knowledge artefact consisting of several layers. A document consists of information atoms. An **information atom** is the smallest unit of content which can be interpreted without a documents context (but of course requiring background knowledge). For text, these atoms are single words. Defining characteristics of documents are:

Packaging: A document contains a number of information atoms. A document packs a set of information atoms together, establishing a context for them. This influences the interpretation by the reader.

Reference-ability: Once a document is published, the reference can act as a placeholder for the content expressed within. A reference to a document can act as a meta-symbol on top of the symbols (information atoms) the document contains. The usage of document references as symbols allows a document to “participate” in conversations, which probably lead to scholastic methods and modern academia.

Process metadata: Each document is written by a number of authors for a certain audience with a certain goal. By sending this process metadata along with the document the reader has the ability to put the document in context and interpret it better.

Linearity: A document can typically be read from start to end by navigating through all contained information items.

Visual structure: A document is not only a stream of sentences, but uses type-setting, i.e. bold, italics, different font styles and size, and placement of figures. Using only the visual structure, references can only point to page numbers. They can change when the document is e. g. re-printed with a wider margin.

Logical structure: The visual structure is used to encode a logical structure consisting of i.e. paragraphs, headlines, footnotes, citations, and title. The logical structure makes it possible to reference smaller, meaningful parts *within* a document, i.e. “Sec. 4.2”.

Argumentative structure: On top of the linear content, a document follows an argumentative structure to convey its content to the reader. Argumentative structures appear on all scales. A typical structure is the “Introduction - Related work – Contribution - Conclusion”-pattern of scientific articles. On smaller scales, patterns like “claim-proof” and “question-answer” are used.

Content semantics: Documents content’s mean something. Building upon logical and argumentative structure, the author encodes statements about a domain within the content.

4. Digital Documents

Buckland argues [2], it's even harder to define the term "digital documents", e.g. in former times people used "log tables" to look up logarithmic values. Today, one would likely use a functionally equivalent software tool. The on-screen rendering of such a tool could be considered to be a document. Buckland sees a trend towards defining a document in terms of function rather than physical format. By following this trend, everything that behaves like a document *is* a document.

In 1981, the Xerox Star Workstation, one of the first *personal* computers, was released [8]. It pioneered the WIMP-metaphor (window, icon, menu, pointing device) and placed digital documents, represented as little icons, in the heart of the user interaction. Files in the computer were modelled close to physical documents. Since then, documents remained the dominant paradigm for information exchange and archival. This is problematic e.g. when search results return references to long documents, instead of shorter (and maybe even referenceable and annotable) information atoms.

4.1 From analogue to digital documents

Prominent examples of digital documents are text processors files, hypertext documents and PDF files. Digital documents differ in many ways from analogue documents. In digital documents the visual structure is sometimes separated (e.g. via CSS[1]) from the logical structure. This makes it possibly to execute queries based on the logical structure and e.g. generate automatically a table of contents or return “all footnotes that contain a hyperlink”. Additionally, other documents can now deep-link into a document, e.g. by using named anchors. From a reader’s perspective, this effectively means that the granularity in digital documents is smaller compared to analogous documents.

A document has to be stable in time in order to become something reference-able. Only in this way people can cite the document without having to copy the content. This is not the case for all online documents and web pages: The content at any given URL can change at any point in time. Digital documents can therefore only replace or at least mimic classical documents in two ways: (a) an trustworthy source manages the web server and promises not to change the served content or (b) documents are sent as messages to the recipients, e.g. messages on email mailing lists. Nevertheless, with the advent of hypertext, the number of links between documents or parts thereof increased dramatically when documents became digital.

4.2 Augmented Documents

Documents can be annotated to make the argumentative structure of a document explicit [1]. In a similar manner the process metadata – who wrote the document when and why – can be made explicit. A system for annotating and relating documents in a visual way is described in [14]. It goes beyond showing annotations next to a document, as the annotations themselves are forming a document on their own.

Finally, the semantic features of the content – e.g. social networks between people mentioned in the document or the interplay of mouse genes in a medical publication – can be formalised and represented in a single knowledge model. A knowledge model can represent the content of *many* documents, links to external resources, persons, places and concepts. Query engines can answer domain specific questions about the documents content. Of course, this depends on the expressivity of the formal language used for the annotations.

[1] Cascading Style Sheets, W3C, http://www.w3.org/Style/CSS/

5. Knowledge Models

Knowledge models can be seen as a superset of documents and formal ontologies. *Definition*: A knowledge model is a set of items, linked by directed, typed relations. An item is either an information atom or an entity composed of other items. Each item can contain a small piece of text, such as a paragraph, a sentence, a single word, or a reference to an external web or desktop resource. Relation types are also labelled with items.

Relations can represent all structures listed in Sec. 3. Additionally, all user-defined relations can be within a single knowledge model. *Conceptual Data Structures* (CDS) [20], can be seen as a top-level ontology for personal knowledge model relations.

5.1 From documents to knowledge models

I see three trends from analogue to digital documents: a smaller content granularity, more interconnected content and more explicit structures. Extrapolating these trends we end at having knowledge models with (a) very small information atoms, such as single words, (b) richly connected items, and (c) explicit semantics for the links.

When using documents, a field study of Sellen and Harper [11] concludes that annotating documents is frequently a part of the document reading and understanding process. Phelps and Wilensky [17] describe the concept of "Multivalent Documents" for uniformly annotating different content types with rich annotation types. Annotated documents, stored together with their annotations, can be seen as a knowledge model.

5.2 From knowledge models to documents

Often research notes and references are already managed digitally. Recent works in PIM [12] and Semantic Desktop research have further stressed the need for *unified search and organisation* of a user’s personal files. Furthermore, "Almost all current documents have existed in electronic form at one stage in their life" [16]. I assume these “digital forms” include not only visual but also logical and maybe even argumentative structures (e.g. partially represented as outlines).

Using personal knowledge models could speed up the document creation process, by allowing an author to manage her knowledge digitally and refine it step-by-step into a document-like artefact. Then, a part of the personal knowledge model could be exported as a document. Bradshaw et al [19] argues that researchers would understand documents much quicker if they had access to annotations made by others. If this is the case, then how much quicker could a reader understand a document if he could get access to the knowledge model that the document was created from? Even better: What if not documents, but knowledge models would be exchanged between people? Creating a document is an expensive process - some of the cost could be saved by publishing knowledge models, not documents.

6. Knowledge Models in PKM

Oren [15] finds "an under-utilisation of the interlinked nature of the information". The fine-granular nature of knowledge models allows for precise and effective linking – and browsing.

People have problems in using strict hierarchies [15]. I thus propose to use classification methods like tagging and non-strict taxonomies. As a knowledge model represents the content of *many* documents, represented as many, interlinked small items, the classical document boundary is crossed: One item can be linked from many other items (like *transclusion* in hypertext research).

Another imperative from Oren is "keep the context". The networked nature of a knowledge model is more suited to represent contextual links than a set of documents.

As explained in [6] humans are good in using spatial information to find information. Research projects such as iMapping [10] aim to create intuitive, spatial semantics-based PKM tools. Spatial layout of items can be seen as a set of links in the knowledge model.

7. Related Work

The initial ideas of knowledge models, although that term was not used, can be found in [3, 7]. Ludwig [13] sees redundancy within and among documents as a hurdle to efficient information usage. He questions if documents are the best container for knowledge representations and proposes to work more direct with redundancy-free semantic knowledge management systems. In such a system, the traditional notion of a document is replaced by virtual documents, which render parts of the knowledge base as an interactive tree. Bernstein describes TinderBox, a "personal content management assistant" [4], which offers sophisticated HTML generation via templates. Both systems [4, 13] allow end-users to construct ontologies out of their linked information objects. The same direction can be observed in the larger fields of semantic desktop [5] and semantic wiki [21].

8. Conclusion: The Future of Personal Knowledge Models

PKM looks from the viewpoint of an individual and tries to support her in effectively storing, structuring, linking, formalising and retrieving knowledge cues. Second, the emerging knowledge models can be exchanged and linked with each other. In PKM, *content is first created for oneself, then shared with others* – in a way controlled by the user. Contrast this with the current situation where employees often either don't use a central database at all or where it is flooded with irrelevant, long documents.

For successful PKM, I believe the user must have freedom in: (1) granularity (size) of knowledge items, (2) degree of structure and formality between knowledge items: none, links, tags, and typed relations; (3) privacy: sharing on a per-item-basis with dedicated groups; (4) modelling: no schema is fixed; (5) annotation and expressivity: every item – including links between items – has to be annotable, and (6) choice of navigation: browsing hypertext or spatial layout. The systems should also allow powerful queries over the heterogeneous corpus as well as sophisticated interoperability options (import/export), in order to exploit the knowledge models added value.

As future research, an implementation of [20] will be created and user studies will be conduct using a wiki-like and a concept-map-style interface. The author plans to write one of the next papers using such a system and publish the knowledge model along with the paper.

Literature References

[1] C. Beckstein and H, Sack and H. Peter: “Tags and Dependencies: an Integrated View of Document Annotation”, 1st Semantic Authoring and Annotation Workshop at the ISWC 2006, Athens, GA, USA, Nov, 2006.

[2] M. Buckland: What is a "digital document"?, J. ASIS 48, no. 9 (Sept 1997): 804-809.

[3] V. Bush: As we may think aus: Atlantic Monthly, Juli 1945, Band 176, Nr. 1, S. 101–108.

[4] M. Bernstein: Shadows In The Cave: hypertext transformations. In Proc. of IVICA 2006, October 1–2, 2006, College Station, TX, USA.

[5] S. Decker and J. Park and D. Quan and L. Sauermann (eds.). The Semantic Desktop – Next Generation Information Management & Collaboration Infrastructure. Galway, Ireland, 2005.

[6] D. Elsweiler et al: Considering Human Memory in PIM. In [12].

[7] D. C. Engelbart: Augmenting Human Intellect. Research proposal, SRI, Ca, USA. Oct 1962.

[8] M. Friedewald: Der Computer als Werkzeug und Medium. Die geistigen und technischen Wurzeln des Personalcomputers, ISBN 3928186477, GNT-Verlag, 15 March, 2000.

[9] H. Haller: Mappingverfahren zur Wissensorganisation. Diploma thesis, Berlin, Germany, 2002.

[10] H. Haller: iMapping - a Graphical Approach to Semi-Structured Knowledge Modelling. In Lloyd Rutledge (eds), Proc. SWUI2006, Athens, GA, USA. Nov, 2006.

[11] K. O'Hara and A. Sellen: A Comparison of Reading Paper and On-Line Documents. In Proc. of CHI97, Steven Pemberton (eds), Atlanta, Georgia, USA, March 22-27 1997.

[12] W. Jones, Proc. of Personal Information Management, SIGIR 2006 Workshop, August 10-11, 2006, Seattle, Washington

[13] L. Ludwig: Semantic Personal Knowledge Management, Technical Report: Lion Project, DERI, 2005.

[14] Maier et al., Personal Information Enhancement (part of Sidewalk Project). Position Paper Poster. In [12].

[15] E. Oren. An overview of information management and knowledge work studies: Lessons for the semantic desktop. In Semantic Desktop (ISWC). Nov. 2006.

[16] R.T. Pédauque: Document: Form, Sign and Medium, As Reformulated for Electronic Documents, STIC-CNRS, Version 3, July 8, 2003.

[17] T. A. Phelps and R. Wilensky: Multivalent Documents: A New Model for Digital Documents. In CACM, 43(6): 82-90, 2000.

[18] D. Polard: Personal Knowledge Management (PKM) -- an Update. In "Dave Pollard's environmental philosophy, creative works, business papers and essays" [http://blogs.salon.com/0002007/], Nov 23, 2005.

[19] Bradshaw, Shannon, Marc Light, & David Eichmann, (Bee)Dancing on the Boundary between PIM and GIM. Position Paper Poster. In [12].

[20] M. Völkel and H. Haller: Conceptual Data Structures (CDS) -- Towards an Ontology for Semi-Formal Articulation of Personal Knowledge. In Proc. of the 14th International Conference on Conceptual Structures 2006. Aalborg University - Denmark, July 2006.

[21] M. Völkel and S. Schaffert (eds), Proc. of the First Workshop on Semantic Wikis - From Wiki To Semantics. Budva, Montenegro, 2006.