On how to model Content Engineering in a Semantic Web environment

Lutz Maicher

Department of Information Science, University of Leipzig
maicher@informatik.uni-leipzig.de
http://www.informatik.uni-leipzig.de/~maicher

The technologies of the Semantic Web demand complex publications which themselves are the result of complex production processes. The complexity of these publications is cause and effect of more sophisticated communication processes possible through the Semantic Web. We introduce a model which briefly describes these processes on a solution-independent level by using a market perspective. The model is based on the assumption of a market interaction between content offer and demand, but is independent of the existence of real content markets with financial transactions. We emphasise the need of structured guidelines for Semantic Web Content Engineering Processes. Furthermore, the model represents a foundation for their development.

Introduction

The Semantic Web will provoke more and more complex publications, which are context-sensitive and readable by man and machines. The results of Semantic Web research will drastically contribute to the complexity of the production processes of the publications themselves. Such processes are called Content Engineering Processes (CEP). Publications will be "enriched" with sophisticated metadata to improve the information retrieval process. Furthermore, each reader will use provided languages to describe the actual delivery context[1] and its desire for specific contents to personalize the communication process. In this paper, the more technical delivery contexts and the user's desire for specific contents will be merged into the term of "demands". These demands can be described in sophisticated ways in the semantic web, which has strong effects to the production processes of the publications. It is important to understand these processes from an economic viewpoint because of the strong impacts on a lot of different business processes in organisations.

Today, there are only few publications available for Semantic Web technologies. On the one hand, this is caused by immature technologies but, on the other hand, by

[1] Delivery contexts can be characterized in terms of specific user preferences and abilities, capabilities of the access device and available network resources [Osse+02].

the absence of structured guidelines for publication production in the Semantic Web environment. Unfortunately, the adjacent research communities Content Engineering (CE) and Semantic Web are still working relatively independently [Osse+02]. A closer cooperation in future will be necessary due to the importance of the publication production processes in enterprises.

In this paper, we introduce a model which is based on a market perspective. Publications are seen as a collection of offers for personalised content supply. The model is solution-independent and regardless of existing marketing buzz words. The model should serve as an interface between research in CE, Semantic Web and other related research areas like information retrieval, software engineering, business process engineering and service engineering.

A short summary of the paper's structure will conclude the introductory chapter. First, the idea of the Semantic Web and the term of Content Engineering is discussed in detail. Then, the model is described as being based on four assumptions which initiate its further development in the paper. Subsequently, the modelling of publications, which form the basis of the communication process in the Semantic Web is explained. Then, the communication processes are described in the introduced communication space, whereas the production space produces the publications for the communication space. Because the processes in the production space have to be realised in organisations they will be discussed in more detail. To conclude the paper applications of the model are discussed.

The Semantic Web

Berners-Lee et al. describe the Semantic Web as „an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" ([Bern+01]). Here, it should be emphasised that opportunities for cooperation should be improved for computers *and* man ([Osse+01]). The Semantic Web technologies not only can be deployed to improve information gathering and brokerage in the web, but also to present information most appropriate to each consumer ([Osse+02]).

The term Semantic Web pools an enormous amount of different approaches, which is illustrated by the vision of Berners-Lee et al. Thus, the Semantic Web does not define itself, but it only exists in communities which have made prior agreements concerning their approach to it.. At our abstract level we refer to the Semantic Web as a whole knowing that its realisations are only community depended occurrences of a research pool with own specific characteristics. See [Fens+03], [Hyvö+02] for a good introduction in aims, technologies, languages and applications of the Semantic Web research.

The task of the CEP in a Semantic Web environment is the construction and the maintenance of a system supplying sophisticated, goal-oriented communication processes of agents. These agents are located in different points of space and time (see [Romh98], p. 147; [Maic02], pp. 24). Figure 1 describes the basic characteristics of such a communication process in the Semantic Web.

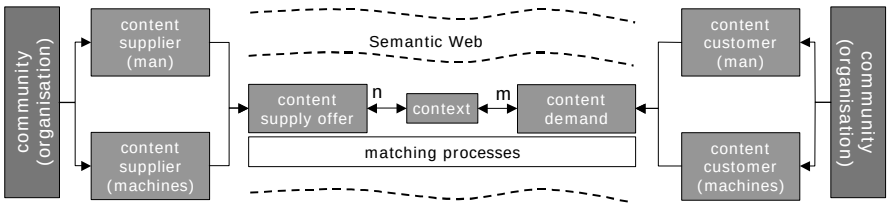

Fig. 1. The communication process in the Semantic Web

The Semantic Web develops its full potential in communication processes with the following characteristics:

- *a large number of content suppliers* (man and machines),
- *a large number of content customers* (man and machines),
- the communication partners are *not necessarily known* in advance,
- the communication is placed in *different contexts*,
- a *heterogeneous* and *decentralised* environment,
- based on *mass-data*,
- *trust and strategy* in and between *communities* are important.

Especially the cooperation of automatic information suppliers and customers leads to new challenges for the CEP. These opportunities will be leveraged by an appropriate application of context-sensitive publications. Besides, the communication in communities can be improved if the CEP is able to realise the benefits of the Semantic Web. However, communication based on Semantic Web technologies will only be able in communities which have made prior agreements about the technological parameters.

The Content Engineering Process

In the knowledge based economy content can be seen as a preliminary product, like screws and joints in an industrial process of manufacture. It is produced from the raw material knowledge and have to be refined in publications for the end-user's consumption. For a sustainable usage of the limited resources of organisations a reuse of knowledge, content and publications is necessary and is therefore to be integrated in the CEP.

We will regard Content Engineering as the industrial manufacture of publications (see figure 2) and use the term CEP as the goal-oriented and process-based *generation* (collection, production, storage), *transformation*, *aggregation* and *representation* of content in publications.

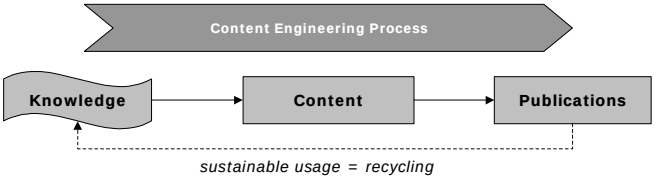

Fig. 2. The Content Engineering Process as an industrial process of manufacture

Modelling Content Engineering

The basis of the model from a market perspective is described by following assumptions:

1. Existence of offered content supply and (anticipated) content demand.
2. Each interaction between a content suppliers and a content customers is the result of a market matching process
3. Content suppliers and content customers try to maximise their utility.
4. The market perspective is independent from the existence of real, financial transactions.

We propose to distinguish between a *production space* and a *communication space*. The production space generates publications for the communication space. This process is modelled and referred to as CEP. Subsequently, the communication space is produced by the production space. Its inherent communication processes are to be analysed separately.

The Model of Publications

Before starting the development of the models of the production and the communication space publications as their connector have to be introduced. A publication will be called **m**. The characteristics of such a publication are (see [Roth+01], pp. 134):

- it is made for an (anticipated) demand formulated by the customer,
- it is produced for man and/or machine consumption and
- it is not necessarily persistent.

The discussion in research about adaptive hypermedia systems suggesting a separation of content and links ([Bech+01], [Staf97], [Wild+00]). We extend this approach by introducing a separation of the publication's content which is made for consumption, and a set of content offers which is made for the maintenance of the

communication process. Only these offers describe the possible content supply provided by the producer.

In the current web contents are produced for human consumption, like texts about enterprises or people, and are primarily written in HTML[2]. In the Semantic Web information will be described by sophisticated meta-data by using different languages (RDF[3], DAML+OIL[4], Topic-Maps[5]) and general or domain-specific ontologies. The set of all contents in a publication **m** will be called **c(m)**. The introduced market is rather based on a competition in **c(m)** than in **m**.

The consumer of a publication has to understand **c(m)**, since a publication comprises references to the used languages and ontologies as well as the original meaning of **c(m)**. As the result of this consumption process the consumer will formulate a new demand to continue the communication process. However, formulating its demand it can only use the offers for content supply provided by **m**. The publication provides these offers as a set of production rules called **r(m)**. These production rules can be seen as a language for the content demand description. To summarise, we will define a publication **m** as the following tuple **m=(c(m),r(m))**.

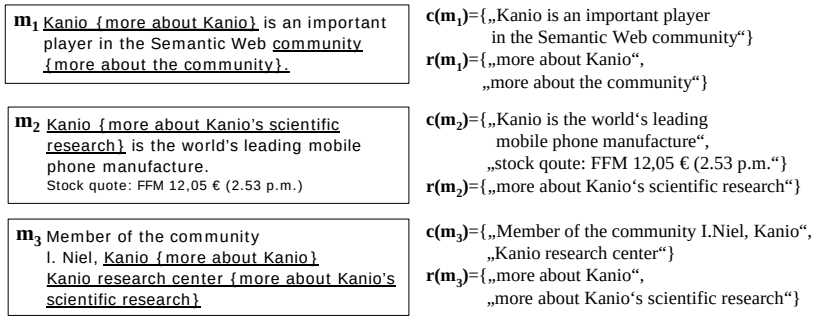

Fig. 3. Example of three publications

We will define the set of possible sentences of **r(m) as k(m)**. At present the main possibility to formulate a demand in the web is the usage of hyperlinks. For this kind of publications **m** the set of all these links is **r(m)**. As shown by the simple example introduced in figure 3, **r(m)** and **k(m)** are equal in this case. Although **r(m)** and **k(m)** are not equal at the entry site of a search engine. If **m** provides the possibility to enter a search query, **r(m)** describes only all valid letters and the valid length of this query. **k(m)** is the set of all possible search queries which can be produced dependent on the constraints formulated by **r(m)**.

In the Semantic Web the communication process will further be maintained by more sophisticated mechanisms than simple hyperlinks. Therefore, the web will alter

[2] see http://www.w3.org/MarkUp/

[3] see http://www.w3.org/RDF/

[4] see http://www.w3.org/TR/daml+oil-reference

[5] see http://www.topicmaps.org/1.0/

from a “web of links” to a “web of offers”. Already some features of the XML Linking Language[6] show possible further developments.

The Model of the Communication Space

After introducing publications the communication space will be modelled. It is a pair {**S**,**D**} and exists where published **m** and their consumers meet. This is based on the chosen market perspective. A publication supplies content according to a demand formulated by a customer. Furthermore, a publication offers new content supply to continue the communication process. If a content supply offer is published, the producer is forced to accommodate the induced demand. So, the production space is obliged to produce the offered content as long as the offer is available. This modelling approach based on transactions prevents the introduction of a time model.

The communication space consists of a supply system **S** and a fictive demand system **D**. The supply system **S={m}** is the set of all publications made by the producer which are available to customers. One has to bear in mind that each sentence **k(m)** is, on the one hand, a possible content supply offered by **m**. On the other hand, customers can only formulate their demand with one sentence of **k(m)**. This ambiguity of **k(m)** is important for the model. Because **m** is a tuple **(c(m),r(m))** and **k(m)** is the extension of **r(m)** the supply system is **S={(c(m),k(m))}** as well.

The demand system **D={(m,d)}** is a set of tuple **(m,d)**. *Each* consumer of a publication **m** is described by its real demand **d**. The real demand describes *all* about the consumer’s wants and possibilities. So **d** includes much more statements than expressible by **r(m)**.

Matching processes occur in the communication space. Each costumer consumes **c(m)** of a specific publication **m**. While consuming the publication its real demand **d** alters. If the customer wants to continue the communication process, it has to formulate a new content demand. The publication provides only the limited language **r(m)** and the customer describes its demand by choosing one sentence from **k(m)**. In most cases this leads to a loss of information. Furthermore, one has to pay attention to the strong interrelations between the concepts “demand” and “context” already discussed. In the case, other customers do not want to maintain the communication process while the state transformation they can alternatively choose the empty set. This is formalised in expression 1:

$$P_d : (m, d) \rightarrow k(m) \vee \{\varnothing\} \qquad (1)$$

The matching between supply and demand is realised by $\mathbf{P_d}$. The production space interprets the chosen sentence from **k(m)** and transforms the communication and production space according to its goals. This transformation function **T** is subsequently discussed in detail.

[6] see http://www.w3.org/TR/xlink/

The state transformation function **R** is formalised in expression 2:

$$R : D \xrightarrow{P_d} k(m) \xrightarrow{T} S \qquad (2)$$

Figure 4 illustrates an example. There are three publications available to the enterprise Konia (see figure 3). In publication $\mathbf{m_3}$ only the decision between "more about Konia" and "more about Konia's scientific research" is possible formulating a content demand. In the demand system **D** four customers consume publications in the state *i*. Publication $\mathbf{m_3}$ is consumed by customer *4*. While consuming the publication $\mathbf{d_4}$ of this customer alters. Perhaps it needs more information about Topic Maps and the Semantic Web. Yet, $\mathbf{m_3}$ provides only something "about Konia's scientific research". In this case, the customer formulates its demand with this sentence. The customer expects a publication which maximises its expected utility at the state *i+1*. When the production space receives the new demand, the communication (and production) space will be altered according to the goals of the producer. A considerable difference between the customer's expectation and the producer's deliveries is caused by the wide range of possible interpretations of the demand.

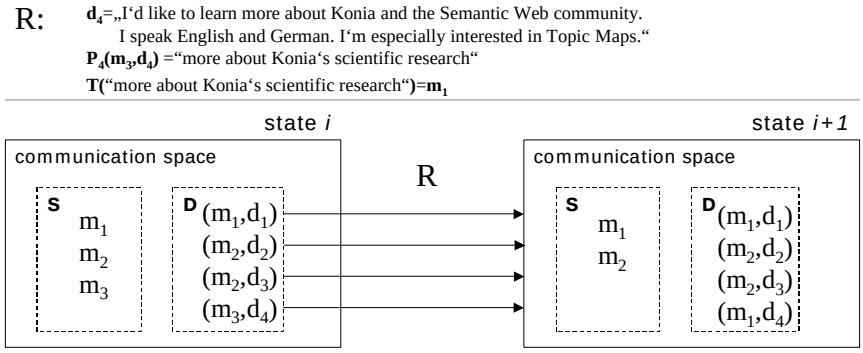

Fig. 4. Example for the communication space

In the example the other customers did not want to continue their communication processes at the given state. They chose the empty set. The result in the example is the reduction of the supply system. According to the demand, the production space has provided publication $\mathbf{m_1}$ to customer *4*. In the new state *i+1* the production space is only to provide content supplies which are offered in publication $\mathbf{m_1}$ and $\mathbf{m_2}$.

The Basic Model of the Production Space

In the production space the CEP is implemented. The CEP is the collection and production of a set of (raw) content objects **q** (texts from internal and external authors, external or internal databases, ERP-systems, RSS-feeds[7]) and their transformation in a set of customised publications **m** which will be published in **S**.

[7] see http://www.mnot.net/rss/tutorial.

Each content object **q** can be either atomic, i.e. a special text from an author, a set of content objects or a function. Domain and co-domain of these functions are content objects too. For example, a function can be a service which provides the stock quote for a given enterprise in a specified content object. These extensions allows the encapsulation of intelligence in the content objects.

Concerning these assumption the set of all *1..I* source objects **q** is the knowledge base **Q** of the CEP. The set of all *1..L* offered publications **m** is **M**. The set of all possible publications which can be produced from **Q** is $\mathbf{M_Q}$. Formally, the whole CEP is the production of **Q** and of a function **T** which chooses for all content supply offers published in **S** a publication from $\mathbf{M_Q}$:

$$\forall m \in S \quad \forall l \in k(m) \quad T(l) \in M_Q \tag{3}$$

According to the market perspective each publication **m** should be producible on demand if the offer is published. The demand is one possible sentence of **r(m)**. The production room's task is the interpretation of these sentences and the development of the according part of the function **T**. Here, one has to bear in mind that the producer tries to maximise its utility. One of the main design issues of the Semantic Web is that it has to handle inconsistent data[8]. So, we can imagine scenarios where the users' demands will intentionally not be accommodated, because the produce tries to maximise its utility and he has the technological possibilities.

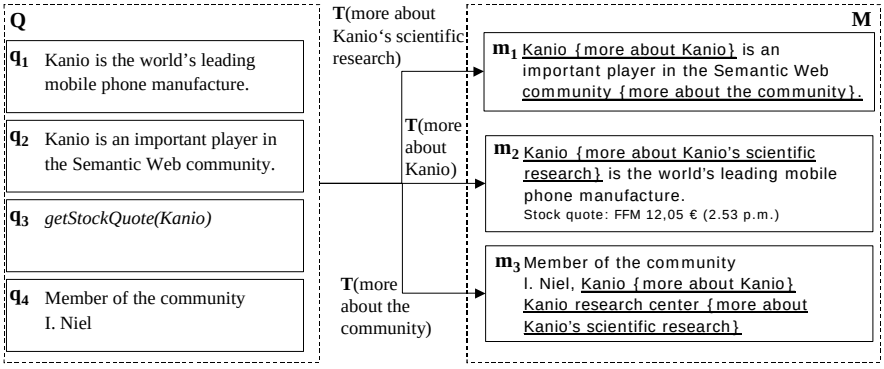

Fig. 5. Production space of the example

Figure 5 shows the production space of the introduced example. Four content objects will be collected to produce the system. The first object describes the enterprise, the second its scientific activities, the third is a service which provides its actual stock quote and the fourth characterizes the research community. If the production space receives the demand "more about Konia's scientific research", publication $\mathbf{m_1}$ from $\mathbf{M_Q}$ will be selected. There are only few opportunities in the publications to explicit a new demand, e.g. only two links are provided in $\mathbf{m_1}$ for further information.

[8] see http://www.w3.org/DesignIssues/Inconsistent.html

The detailed Model of the production space

The number of possible demands and $\mathbf{M_Q}$ extremely grows in the Semantic Web. The introduced function **T** is a relevance function which chooses the right publication from the large set $\mathbf{M_Q}$. At an abstract level we adhere to this approach, but at an practical level **T** is a transformation function, which transforms **Q** into **M** in a pipeline processing model. In order to handle the remaining complexity we propose the separation of the production space in *a source system*, *a concept system* and *a publication system*. The separation shown in figure 6 structures the already introduced tasks of the CEP.

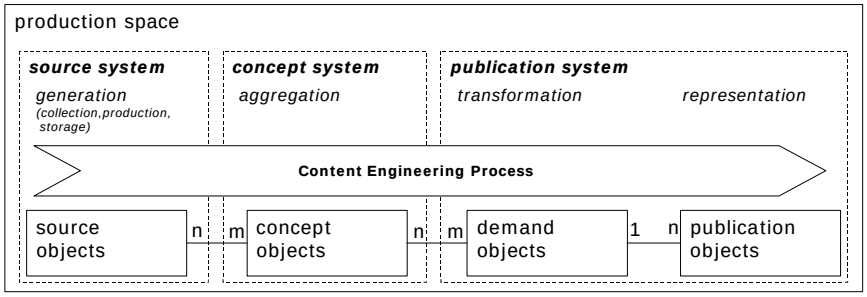

Fig. 6. Separation of the production space

The source system is **Q**, the set of all source objects **q**. All objects with relevance to the objectives of the CEP's owner have to be *produced*, *collected* and *stored*.

A concept[9] in knowledge representation is each real or abstract "thing of interest" whereas statements exist ([Reim91], p. 14). The concept system is the set of concept objects. A concept object is also a content object and represents a view of **Q** which pools all statements (content) according to a special concept, i.e. specific processes, roles, employees or products. The concept objects will be *aggregated* from **Q** according to the relevant concept. We put forward an object-centric knowledge representation approach (see [Reim91]) to represent the whole concepts' knowledge.

A separation of concerns between logic, content and layout (see [Roth+01]) is applied in practice. Especially in the Semantic Web the same content can be published in various ways. See as an example [LeGr+01]. According to this heuristic the publication system is separated into the set of all demand objects and the set **M** of all publications.[10] A demand object represents the demand-based view on the concept system. The production space interprets the demand and *transforms* all relevant

[9] A number of man-centered approaches to the Semantic Web are based on concepts as a central design criteria ([Avel+02], [Scha02], [Thom02]). While developing concept-based solutions the conceptual hypertext system research should be used ([Osse+01], [Bech+01]).

[10] The number of all demand objects and publications is not equal because to one demand object different representations can be produced (HTML, PDF, WML, VOXML, SVG, VRML, RTF, Plain Text).

contents from the concept system according to the goals of the producer. The task of the layout production is the development of layout transformations for each demand object. These transformations linearise the demand objects for *representation* in a medium.

How to apply the model?

The proposed model helps to structure the research in Semantic Web Content Engineering. We propose to use "Cocoon"[11] from the Apache project to simulate all ideas concerning the production space. Cocoon is a completely XML-based publishing framework which allows the separation of content, logic, style and management with a sophisticated pipeline processing model. Furthermore we propose the following research efforts:

Development of the model and the formalism

The following limitations of the model should be lifted in further research:

- integration of goal systems for producer and customer, which helps to evaluate the quality of the publications (difference between customers' expectation and producers' realisation),
- further enhancement of "demand" as a union of context and desire and
- further discussion of concepts in connection with the Semantic Web.

Development of IT-systems

Concerning the development of content-based applications the following ideas should be discussed in detail:

- Development of a data-type „content" as an encapsulation of content and intelligence, which can be used to model and realise content-based applications at different abstract layers.
- The content exchange in a content commerce scenario should be supported. This has strong advantages if content will be supplied as web services with defined service level agreements.
- If customers are machines, $\mathbf{P_d}$ is an automatic matching between two formal languages. This fact touches research in collaborative ontology design and usage ([Hols+02]).

Processes and Organisation

The CEP is a main business process in knowledge based enterprises. The model helps to describe requirements for the development of the following (business) processes:

- initial definition of **T** and **Q**,

[11] see http://cocoon.apache.org/2.0/

- manual or automatic development of **T** and **Q** according to the development of $\mathbf{P_d}$ and to the communication and production space,
- development of processes for goal definition and enforcement,
- identification of communities which already use Semantic Web technologies (which understand **c(m)** and use **r(m)**),
- integration of results from service engineering research.

Strategy and Trust

Strategic thoughts are necessary in the Semantic Web, especially in enterprises where each publication is the result or the preparation of an value-adding business process. In this case, the communication space represents real markets with strong impacts to the real world. Nevertheless, these markets have anomalies. Disinformation and deceit will frequently occur, which implies the use of game theory in connection to CE:

- Which parts of the demand should (not) be satisfied? Which parts should be ignored or definitely used?
- Do recipients deviate from benevolent strategies? Does the demand represent the honest transformation of **d** or does it only represent a construction to manipulate the results of the communication process?
- Do producers deviate from benevolent strategies? Does **m** represent a honest transformation according to **d** or an intentional manipulation?
- Which communities are trustworthy in which context?

Conclusion

The necessity of an integration of research in Semantic Web technologies and Content Engineering has been shown and emphasised.

The market perspective characterises the communication processes in the Semantic Web, although it is independent from the existence of real financial transactions. Especially the introduction of “offers” and “demands” instead of links meets the requirements of the Semantic Web. The Semantic Web will be a “web of offers”. The formalism further provides consolidated semantics of methods and notions in CE independently of realisations and marketing terms. It is advisable to apply the results from other research areas to CE. The proposed model formulates the future requirements of these research areas , which can be integrated in the CE research.

References

[AVEL+02] AVELLO, Daniel Gayo; GUTIÉRREZ, Darío Álvarez: *The Cooperative Web. A Complement to the Semantic Web*. In: Proceedings of 26th Annual International Computer Software and Applications Conference (2002), S. 179-183.

[BECH+01] BECHHOFER, Sean; CARR, Leslie; GOBLE, Carole; HALL, Wendy: *Conceptual Open Hypermedia = The Semantic Web?* In: Proceedings of The Second International Workshop on the Semantic Web – SemWeb (2001).

[BERN+01] BERNERS-LEE, Tim; HENDER, James; LASSILA, Ora: *The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*. In: Scientific American. May (2001), S. 34-43.

[FENS+03] FENSEL, Dieter; HENDLER, Jim et al. (eds.): *Spinning the Semantic Web.* Cambridge, Massachusetts, London, MIT Press, (2003).

[HOLS+02] HOLSAPPLE, Clyde W.; JOSHI, K. D.: *A Collaborative Approach to Ontology Design*. In: Communications of the ACM 45, 2 (2002), S. 42-47.

[HYVÖ+02] HYVÖNEN, Eero (ed.): *Semantic Web Kick-Off in Finland. Vision, Technologies, Research, and Applications.* HIT Publications, (2003).

[LEGR+01] LE GRAND, Bénédicte; SOTO, Michel: *Visualisation of the Semantic Web: Topic Map Visualisation.* Available at: http://www-rp.lip6.fr/site_rp/

[MAIC02] MAICHER, Lutz: *Evaluation von Topic-Maps zur Beschreibung von Wissensstrukturen als Grundlage für das Wissensmanagement in IT-Projekten.* Diploma-thesis at university of Leipzig, August (2002).

[OSSE+01] OSSENBRUGGEN, J.R. van; HARDMAN, H.L.; RUTLEDGE, L.: *Hypermedia and the semantic web: A research agenda.* In: Technical Report INS-R0105, CWI, (2001).

[OSSE+02] OSSENBRUGGEN, J.R. van; HARDMAN, H.L.: *Smart Style on the Semantic Web.* In: Proceedings of the International Workshop of the Semantic Web, Hawai (2002).

[REIM91] REIMER, Ulrich: *Einführung in die Wissensrepräsentation.* Stuttgart, Teubner (2001).

[ROMH98] ROMHARDT, Kai: *Die Organisation aus der Wissensperspektive – Möglichkeiten und Grenzen der Intervention.* PhD-thesis at university of Geneva, (1998).

[ROTH+01] ROTHFUSS, Gunther Elgar; RIED, Christian (Hrsg.): *Content Management mit XML: Grundlagen und Anwendungen.* Berlin, Heidelberg, New York (Springer) 2001.

[SCHA02] SCHATZ, Bruce R.: *The Interspace: Concept Navigation Across Distributed Communities.* In: IEEE Computer 35, 1 (2002), pp. 54-62.

[STAF97] STAFF, Christopher: *HyperContext: Using Context in Adaptive Hypertext.* In: Proceedings of the Sixth International Conference in User Modelling; Vienna, (Springer), 1997.

[THOM02] THOMPSON, Bryan: *The Cognitive Web.* Presentation to the Semantic Web Interest Group. Available at: http://www.cognitiveweb.org/publications

[WILD+00] WILDE, Erik; LOWE, David: *From Content-centred Publishing to a Link-based View of Information Resources.* In: Proceedings of the 33rd Hawaii International Conference on System Sciences, IEEE Computer Society Press, (2000).