
Bridging the gap between data and semiotics: semantic data model

Tanja Sieber*

Department of Information Science,
University of Miskolc,
Miskolc-Egyetemváros H-3515, Hungary
E-mail: tanja.sieber@advan-ce.de

*Corresponding author

Matthias Kammerer

SAP AG, Dietmar-Hopp-Allee 16,
69190 Walldorf, Germany
E-mail: matthias.kammerer@sap.com

Abstract: In this paper we present the semantic data model that we have developed and discuss our experience in using this data model in the curriculum of an introductory semantic web course. The semantic data model bridges the existing gap between data and semiotics and is the first one covering a combination of different aspects regarding data.

Keywords: semantic data model; semantics; semiotics; semiotic triangle.

Reference to this paper should be made as follows: Sieber, T. and Kammerer, M. (2008) 'Bridging the gap between data and semiotics: semantic data model', *Int. J. Teaching and Case Studies*, Vol. 1, No. 4, pp.283–298.

Biographical notes: T. Sieber received her MSc in Electrical Engineering from the University of Karlsruhe in 1997, and her MSc in Technical Writing from Hochschule Karlsruhe in 2001. She worked as a Service Diagnostic Engineer for GM Europe GmbH from 1997 to 1999 and as Lecturer, Consultant and Trainer in the field of technical documentation and information systems from 2000. Presently, she is a Lecturer at the Department of Computer Science, University of Miskolc, Hungary and an individual PhD student with research interests including semantic technologies, documentation re-use analysis and measurement within technical documentation processing along product lifecycles.

M. Kammerer studied German studies, philosophy, and computational linguistics at the University of Heidelberg. The main focus of his research was semantics, lexicology, metalexigraphy, analytical philosophy, and syntax parsing. In 1995, he published his thesis on writing dictionary papers for computer use. From 1995 to 2000, he was employed as a Lecturer at the University of Heidelberg. In 2000, he published his doctoral thesis on lemma sign types of German verbs. In the same year, he started working at SAP AG in the documentation area.

1 Introduction

Teaching of the Semantic Web (SW) aims not only to give methods in how semantic technologies can be kept the best way in lectures but also how the core idea behind the semantic web can be taught in an effective way. Due to our experience of courses consisting of mixed audiences (programmers, students of information science, technical writers, and people with linguistic backgrounds) we realised the unsatisfactory situation where auditors use different terminologies to name basic ‘semantic-web-relevant-terms’. This effect is not only visible in groups made up of people having different educational backgrounds, but also in those being more homogeneously compound. This problem also appears regularly in mailing lists dealing with semantic web topics; at least every other month, discussions are tackling the questions: what *semantic* really means; what *ontology* is and/or why *semiotics* is important.

Our experience shows that people with an educational background in computer science are normally very enthusiastic about data definitions (see BMBF, 2006; Lackes et al., 1998; Gersdorf, 2003), that let *data* appear as signs, symbols, numbers, characters, images or whatever – always with the restriction that it is *without any meaning*.

How does this fit with the opinion and conviction of those people coming along with a strong linguistic background? For example, technical writers or computational linguists who refer to the semiotic triangle in order to explain their understanding of the world with the existing relationship between mental concepts, symbols, and referred objects? What makes the semiotic triangle, now in combination with the idea of the semantic web, interesting for those ‘data = something-without-meaning’ people?

We will describe our intention and motivation for the creation and development of the semantic data model in Section 2. We introduce the model in Section 3 and in Section 4 we describe a course outline, with special regard to the description, of how the semantic data model was embedded in a semantic web course. In Section 5 we describe our experience of integrating the semantic data model within a SW-introductory course.

2 Our approach

Most semantic web teaching materials available in the internet focus either on lightening the relevance of metadata and its semantic description, or on deepening ontologies and the different formalised ways to describe them (see Henze, 2005; Antoniou and van Harmelen, 2004). We recognised that this does not work when people have still no clear idea about what metadata is and what it means. One of the most frequently encountered statements about metadata is that it is ‘data about data’ (see Lackes et al., 1998; Gersdorf, 2003; Hahn, 2003; Tekom, 2005). However, when we are looking at a given environment, is there really any clarity about what metadata actually are, what function they serves and how they differ from ‘normal’ data? The confusing state of terms used in connection with the (meta-) data jungle becomes clear if we look at some statements concerning metadata.

In addition to the ‘data about data’-approach mentioned above, there are also the opinions that:

- metadata are data that are never output
- metadata are contextual information (BMBF, 2006) or
- metadata are structured data (Tekom, 2005).

Before anyone can talk about metadata, some light has to be shed on the term *data* itself. Also, besides the usage within the semantic web context, we can often recognise that the term *data* is used without any exact terminological definition or a detailed description concerning its semantic content. The effect is that the term still remains confusing, sometimes even contradicting the definitions of the term presented. The usage of the term seems to be counter intuitive and often leads to unacceptable consequences. The term is not used in a uniform way in papers and lectures; consequently, misunderstandings occur from the diversity of the introduced and used definitions.

This is astonishing, particularly in the area of semantic web discussions, as the main idea and vision of the semantic web deals with exact terminological working and the reference to uniform identifiers. In spring 2006 we developed for our internal usage a semantic data model with clear term definitions. As we both come from different professional areas, it was a terminological foundation for our common understanding and to enable us to work on our research on semantic enrichment of documentation processes. As we used this data model later in discussions with people within SAP AG, students, PhD students, and professors of different areas, we recognised that this serves as a well-thought base for the term *data* for any discussion about data or metadata.

We will present the semantic data model in this paper in the context of implementing this model within a lecture held in the summer term of 2006 at the *informatica femminile* (Germany). This was an introductory course with inhomogeneous auditors: professionals working in computer science, freelance professionals working in the field of technical writing, students of computer linguistics, students of information science, biologists and chemists. The target of the course was to give a solid background about the idea and vision of the semantic web, a deep terminological foundation and an introduction to SW technology.

3 Semantic data model (Sieber/Kammerer)

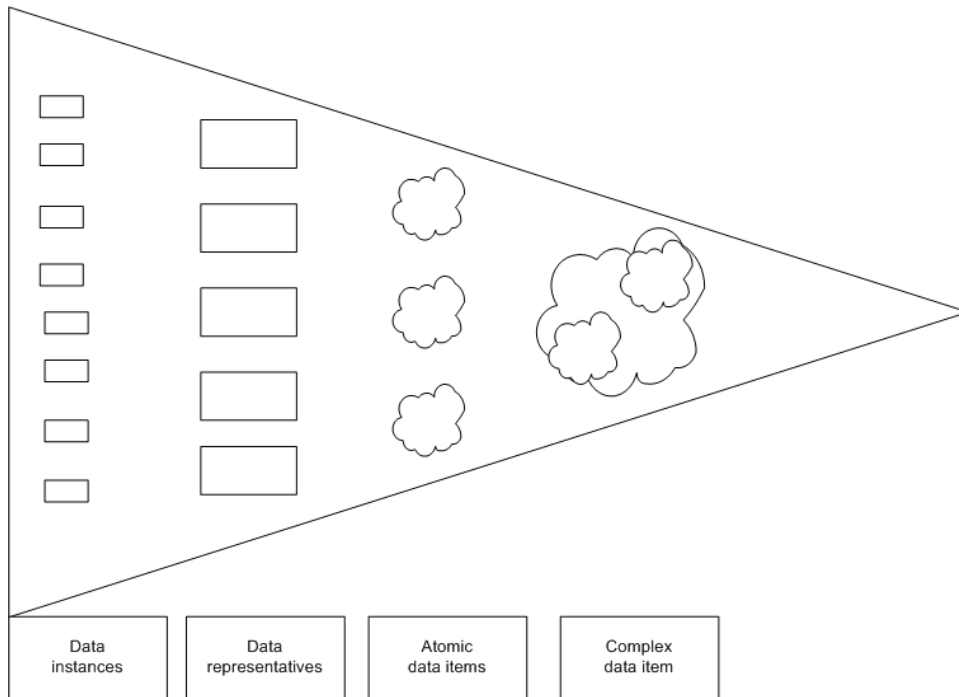
3.1 Form, representation, and meaning of data

According to the semantic data model (see Figure 1), data penetrate through various levels:

- at the level of form, data appear concretely as so-called *data instances*
- at the level of representation, data appear as the so-called *data representatives*, which can be assigned to particular data representation systems
- at the level of meaning, data appear again as abstractions, which implies that a semantic system is allocated to the data.

Furthermore, *atomic data items* should be differentiated from *complex data items*. Atomic data items have the property that they can not be split any further into smaller meaningful components. In Section 3.2, we list the definitions of all key terms appearing in our semantic data model.

Figure 1 Semantic data model (Sieber/Kammerer)



3.2 Definitions: data item, data instance, and data representative

Definition 1: *Data instance*

A *data instance* d_i is the concrete (extra-individual) occurrence of a semiotic entity. This means simultaneously: Every semiotic entity is a data instance.

Note 1: In terms of Peirce, a data instance can, therefore, appear on all three semiotic levels: as an icon, index, or symbol.

Definition 2: *Data representation system*

A *data representation system* d_{rs} is a set of predicates p .

Definition 3: *Data representative*

A *data representative* d_r is the (intra-individual) abstraction set on all data instances of one arbitrarily chosen data representation system.

Note 2: A data representative and – a fortiori – a data instance always belongs to a specific data representation system.

Definition 4: *Data item*

A *data item* is the (intra-individual) abstraction set on all data representatives of the same meaning.

Note 3: Definition 4 implies that data following our understanding and the semantic data model bear a meaning, otherwise no set constitution would be possible.

Definition 5: *Atomic data item*

An *atomic data item* is a data item that can not be analysed in further data, which contribute to the meaning of the data item.

Note 4: An atomic data item requires mapping rules in the form of rules assigning the chosen (intra-individual) data item to the chosen (intra-individual) data representation system and the chosen (intra-individual) data representation system to concrete (extra-individual) data instances.

Definition 6: *Complex data item*

A *complex data item* is a data item which can be analysed as further complex and/or atomic data items.

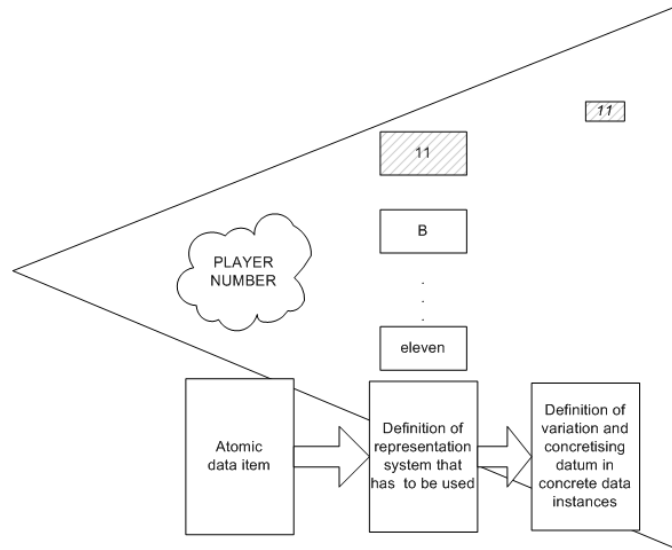
Note 5: A complex data item requires, in addition to the mapping rules, syntactical rules in form of a grammar for the correct syntactical composition of the atomic data items on the level of form and meaning.

3.3 *Concretion process: data item → data instance*

In the semantic data model the data representation system also plays a key role, besides the meaning of the data. This can be seen in an example showing the concretising of a data item taking place: For instance, visualise the atomic data item PLAYER NUMBER – that appears on a water polo player's cap, you will have various possibilities for concretising this (see Figure 2).

- you can decide to use the decimal system and Arabian numerals, and to represent the values between 1 and 15 in this system (for example, '11')
- but you can also use Roman numerals and map the value range mentioned (for example, 'XI')
- or, in the hexadecimal system, the Latin alphabet together with Arabian numbers and map the value range similarly (for example, 'B')
- of course, you could just as well spell out the corresponding numbers in English, using the Latin alphabet (for example, 'Eleven')
- this can be extended as desired.

Figure 2 Example – concretising data item ‘player number’



One thing which takes place almost unconsciously while concretising a data item is the choice of a particular data representation system, and thereby a data representative for your data item. In the following (and similarly, mostly unconscious) step, you will select a certain formal representation for the selected data representative, and create a concrete data instance, such as typing and formatting ‘11’.

Base process models for concretising an atomic data item and a complex one are shown in Figures 3 and 4.

Figure 3 Base process model – concretising an atomic data item

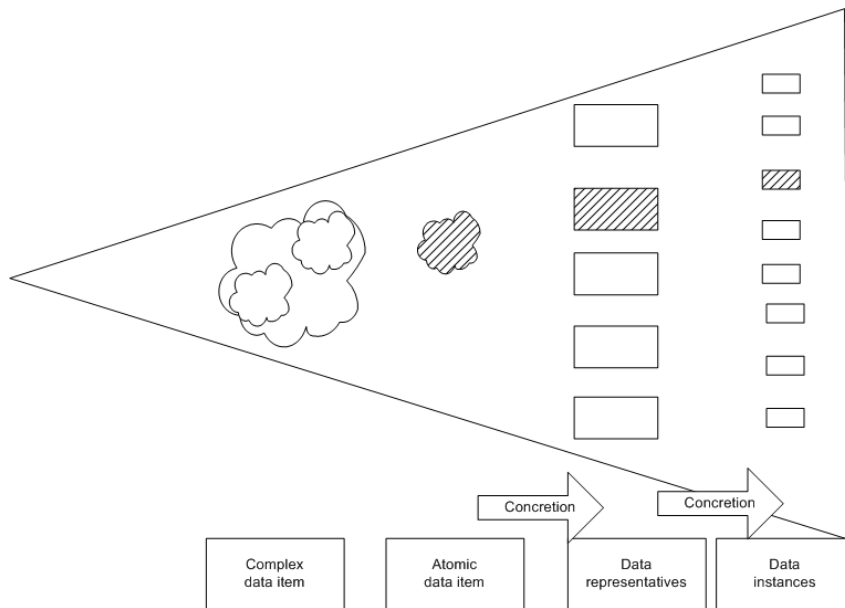
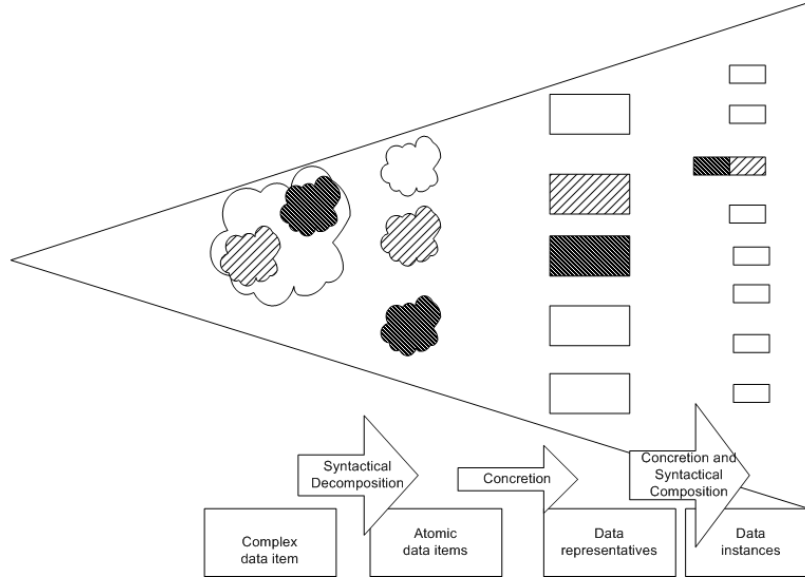


Figure 4 Base process model – concretising a complex data item



3.4 Abstraction process: data instance → data item

Now, data items that have been concretised in this manner confront another person in the form of a data instance. This person, in turn and again unconsciously, tries to understand the data. This is done by abstraction.

A base process model of an abstraction process is shown for an atomic data item in Figure 5, and for a complex data item in Figure 6.

Figure 5 Base process model – abstracting an atomic data item

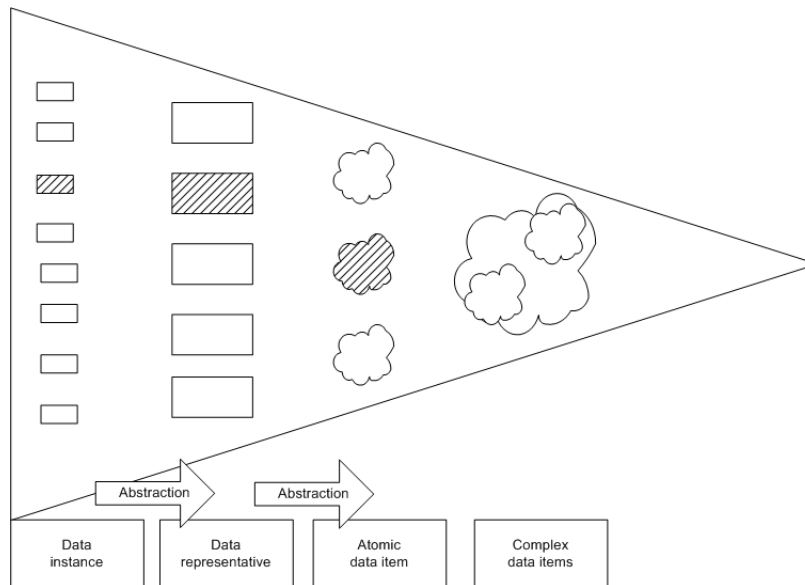
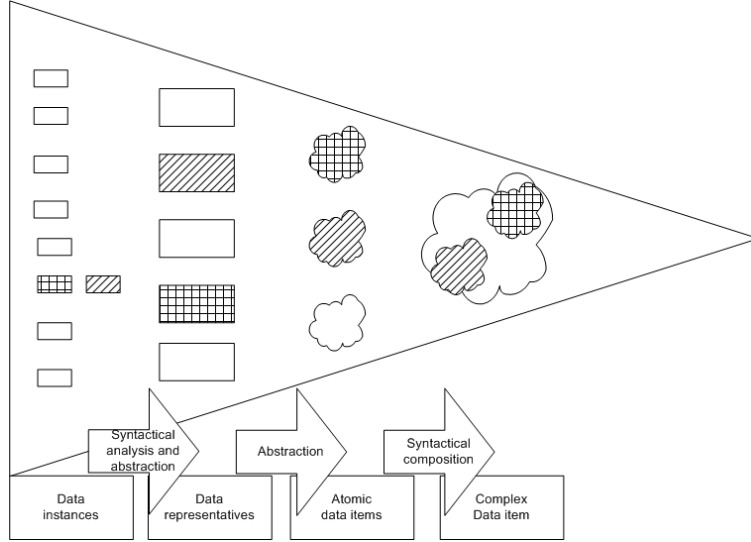


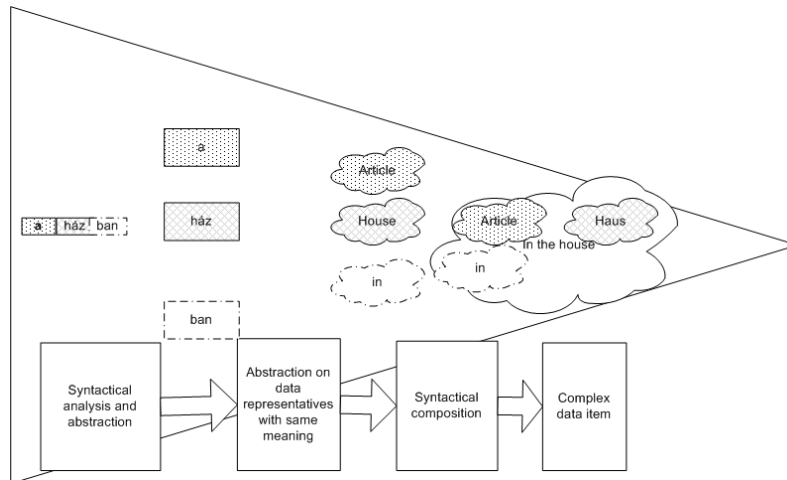
Figure 6 Base process model – abstracting a complex data item



A person seeing the concrete data instance ‘a házban’ of the example shown in Figure 7 is confronted by several questions during the process of abstraction:

- What is the syntactical decomposition of the given complex data instance?
- What is the variation used in concretising the chosen data representatives?
- What is the underlying data representation system?
- What is the meaning of the atomic data items?
- What is the meaning and the syntactical composition of the complex data item that was originally the basis for the concretisation?

Figure 7 Example – abstraction of complex datum ‘in my house’ from concrete data instance ‘a házban’



Abstraction towards the ‘correct’ (in terms of meaning) complex data item can happen only if the corresponding mapping rules are defined or can be derived without ambiguity. In the case of complex data items, syntactical rules have to be given in addition to the data representation systems. In the example given above, the syntactical rules (grammar of the Hungarian language) are applied for the decomposition of the complex data instance and for the composition of the complex data item ‘in the house’.

4 Integration into curriculum

The lecture’s outline for the first compact day of an introductory course is:

- today’s web: problems
- the semantic web approach
- difficulties in human – (machine-) human communication
- data – information – knowledge
- meaning triangle
- semantic web – I know what you mean!

As we avoid ‘PowerPoint poisoning’ in our lectures and training, we present a maximum of two slides per hot spot and work with exercises, and sometimes even games, in order to give a practical experience of using the knowledge. We present, as follows, short summaries of these hot spots showing where exercises or games can be integrated, and referring to a suitable order for introducing the semantic data model into the lecture’s curriculum.

4.1 Today’s web

Meanwhile a core part of our daily life has been affected by the way the internet has changed the way in which ‘information’ (see Figure 8) is transmitted, stored, and accessed.

Although libraries and archives will never be replaced totally, there is no doubt that the internet becomes more and more important with regard to information retrieval. In the web context, the term *content* is widely used instead of *information*.

Today’s typical web uses are:

- seeking and making use of information
- looking for and getting in touch with other people
- reviewing catalogues of online stores
- ordering products by filling out forms.

Summarising, we can say that the intention of the web is bringing together

- people who have content
- with people who want content.

Figure 8 Internet as a source of information?

The first guided exercises go here. To give an experience of the existing difficulties of today's web query results (see also Figure 8) it is highly recommended to prepare some examples for those search- and find-problems. In German lectures and trainings we had good experiences with using queries for the distance from one town to another. Most returned results are hotel recommendations and city information for one of the two cities, but not the actual distance between the cities or even a link to a website where we could find the distance.

As a first summarised point the participants get an understanding of existing problems in today's web. Most of today's web content is suitable for human consumption and not for machine comprehension.

Human comprehension. For a user of the internet the meaning of content presented in the different resources is quite clear, as he is able to read the context, in which the content is embedded.

Machine comprehension. Content stored in the internet is represented as a sequence of characters mostly without any meaning (to the machine!). Machines can crawl through millions of web pages and find matches of keywords, but the concerned knowledge about the context and the background of the terms is not allocated in the pages in a machine-comprehensible way. Even web content that is generated automatically from databases is usually presented without the original structural information found in databases. All these facts cause existing difficulties in seeking and finding relevant content.

4.2 The semantic web approach

This is the right time to present a slide with the semantic web approach (Antoniou and van Harmelen, 2004) and not how it is very often practiced, to present the layer architecture that confuses most people first at all. It is very important to give the

participants a well-based understanding of what we are talking about when we say ‘semantic web’, i.e.,

- represent web content in a form that is more easily machine- understandable
- use intelligent techniques to take advantage of these representations
- gradually evolve the semantic web from the existing web; it is not a competition to the current internet.

4.3 Sensitiveness for SW-relevant terminology

We arrange the course in small groups with a maximum of four persons and present them their tasks.

Input: Questions concerning software related terminology, e.g.,

- Data – Information – Knowledge – Content: what are we talking about?
- Meaning triangle: is this a funny game?
- Where can we find the following within the meaning triangle: *data, information, knowledge* and *content*?
- *Syntactics, semantics, pragmatics*: semiotic relicts or important for semantic web technologies?
- Metadata: where can we find it? What is it good for?

Sources: internet, some useful URLs and literature provided as PDFs in the participants' materials.

Desired output: short 10 minutes presentations from each group.

The following discussion can be guided with the introduction of the semantic data model (detailed description see Section 3). The objective for the exercises, the discussion and the introduced model, is to get to the following conclusion: data appear at different levels: form, representation, and meaning. To get to the next important conclusion it is useful to highlight the communication processes.

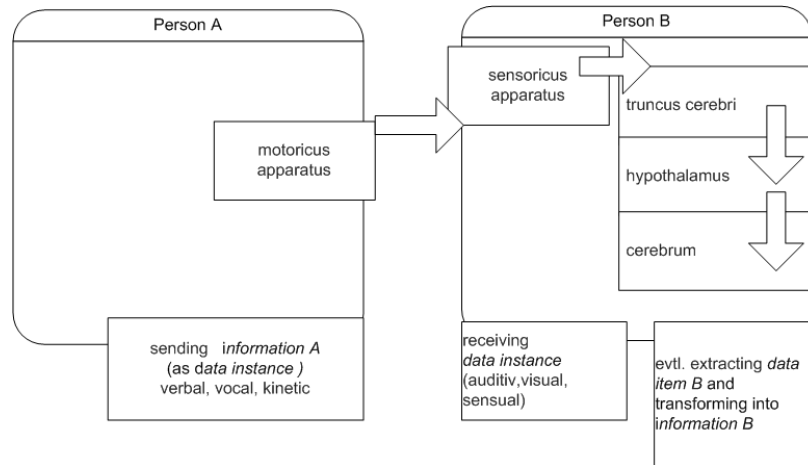
4.4 Roundtrip: from data to meaning

Communication between a sender and a recipient does not happen at the ‘same level of knowledge’ (see papers of Sieber in *technische kommunikation*), especially in human-machine communication. The machine does not know what you mean, and can only evaluate whatever it has been programmed to evaluate. This state of mutual non-comprehension is definitely not restricted to human-machine communication alone: rather, it is equally common in the exchange of information between people (see Figure 9).

A data instance sent by person A and perceived by person B should first cross two areas of the brain that can not be influenced consciously, before it reaches his cerebrum.

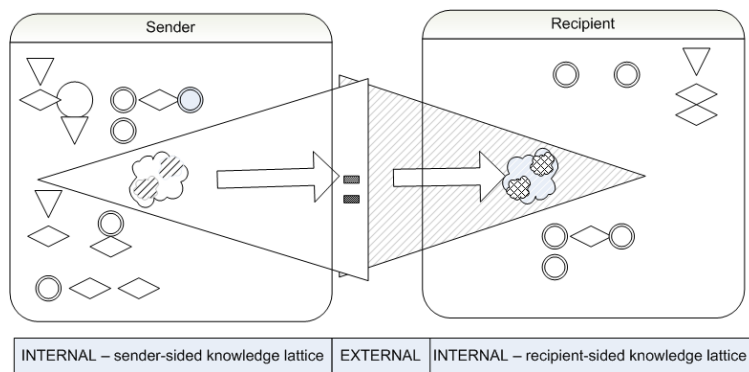
Now, before a piece of data reaches the cerebrum, which is the only part of the brain that can compile information to form a complex picture, it is unconsciously influenced by the emotions and instincts of the recipient.

Figure 9 Registering and processing information



We know this for a fact: Exchange of ‘information’ often fails due to incompatibilities in the way data is processed. This is applicable to a great extent to the communication between human beings and machines, because the basic underlying basis of knowledge that enables us human beings to interpret data is unknown to the machine as such. This ‘information’ exchange can be visualised using the semantic data model (see Figure 10).

Figure 10 Registering and processing information visualised with semantic data model (see online version for colours)



From the point of view of the sender three steps in a communication are necessary:

- decision about what sort of ‘information’ the recipient needs
- decision about how that piece of knowledge can be ‘put into data items’
- the previously described concretising process of data items happens (see Section 3.3).

From the point of view of the recipient the abstraction process (see Section 3.4) takes place and it is not even certain if the abstracted data item is the data item concretised from the sender. Using the semantic data model and the sender-recipient-model from Figure 10 we come to the next important conclusion: everything outside of a human being are concrete data instances and only those can be perceived by another human being.

To deepen the understanding of problems in human-human communication, some exercises could be useful. We use these exercises dependant on timelines and the atmosphere, either as guided working with the semantic data model (which we prepared as triangles in advance) or as games. Both methods aim at combining the collection of experiences of failure possibilities in communication on different levels. The preparation of examples is therefore necessary, showing that even a form of a data instance can be the problem for a non-understanding of the underlying data item. Another example intends demonstrating the difficulty of defining exactly the underlying data representation system.

Finally, it is shown that for communication processes a sender has the possibility to use concrete data instances not only on the object but also on a meta level.

4.5 *Metadata*

Following the semantic data model the following is an explanation of what metadata are:

Metadata are data. Metadata are once again data and can, for their part, possess a formal, representative, and semantic level as in the semantic data model used above. That is, they appear as instances, are used in a system of representation and again have a meaning.

Metadata are relative (see Figure 11). Metadata are always data, for instance, about other data. Metadata are always found at a meta-level and the data about they say something lies at the object level. This meta-level can in turn be regarded as an object level, and this will give rise to metadata pertaining to the metadata.

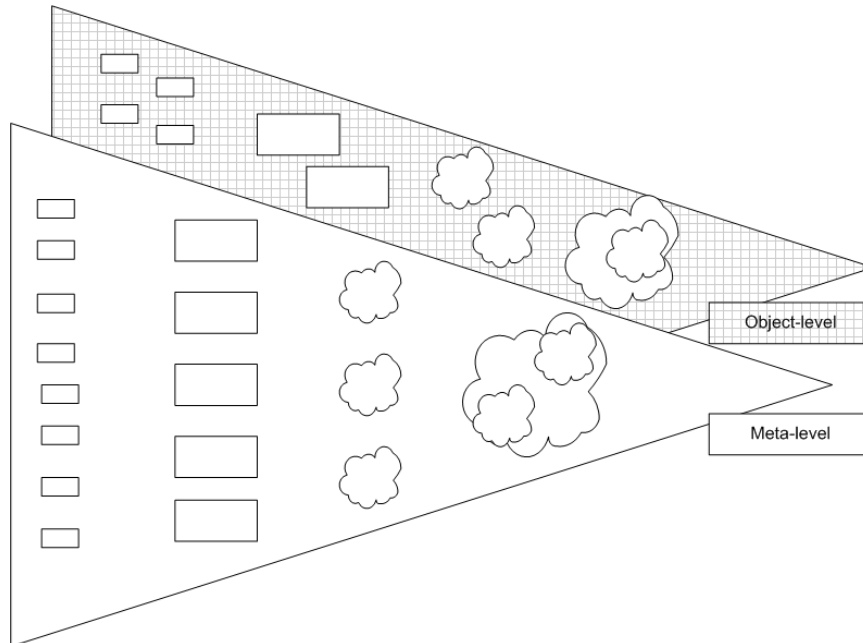
Metadata describe data. Metadata can be categorised largely into three categories, depending on their capacity to convey information about the descriptive data:

- syntactic metadata
- semantic metadata
- pragmatic metadata.

Metadata constitute a basis for bringing together data that are related in terms of content, and for processing them further. They can be understood as a pre-requisite for an intelligent and efficient administration and processing, and not least, as a focused, formal means of providing relevant data.

With these conclusions the first day of the introductory course ends with a slide that repeats once again the semantic web approach and the provocative slogan: I know what you mean!

Figure 11 Relative nature of metadata



5 Conclusion and future work

Using the semantic data model as a basic foundation for the SW-related terminology makes it easy to proceed on the course with the topics:

- *semantics*: why XML is not enough
- *semantic web*:
 - vision
 - overview of SW architecture – layers and wave
 - SW technology
 - explicit metadata/ontologies/logic
 - inference/agents
- semantic technologies: state of the art and future of SW.

The participants develop a deep knowledge of gain and experience in handling terminology. In our opinion this represents the most important prerequisite for clear and critical handling of semantic technologies. The possibility of naming the different levels of data was very effective for us in discussions, and not only in the semantic web context. As the definitions include constitution sets, it is possible to describe them in a mathematically formalised way, enabling further working in that area at a high technical level. Regarding the integration in a semantic web course, we put more emphasis on the cognitive experience, giving the participants a physical model to

work with (never forget to print out some triangles when using the semantic data model!). The greatest benefit we see is that this data model bridges the gap between data and semiotics, and is very helpful in every day business situations, allowing a common understanding of the communication when discussing semantic technologies.

Acknowledgements

We are grateful for the computing resources provided by ontoprise and *i*-views used in the second part of the introductory course at informatica femminile, which we do not handle here.

References

- Antoniou, G. and van Harmelen, F. (2004) *A Semantic Web Primer*, MIT Press.
- BMBF (2006) *Glossar*, URL <http://www.it2006.de/de/394.php> – checked 2006-05-02.
- Gersdorf, R. (2003) ‘Content Management für die flexible Informationswiederverwendung’, in Stahl, F. and Maass, W. (Hrsg.) (Eds.): *Content Management Handbuch: Strategien, Theorien und Systeme für erfolgreiches Content Management*, Universität St. Gallen mcm institute, St. Gallen, pp.61–74.
- Hahn, H-D. (2003) ‘Zauberwort Metadaten – elementares Handwerkszeug des Content- und Wissenmanagements’, in Stahl, F. and Maass, W. (Hrsg.) (Eds.): *Content Management Handbuch: Strategien, Theorien und Systeme für Erfolgreiches Content Management*, Universität St. Gallen mcm institute, St. Gallen, pp.165–176.
- Henze, N. (2005) *Semantic Web Lecture*, URL www.kbs.uni-hannover.de/%7Ehenze/semweb05/ – checked 2006-03-13.
- Lackes, R., Wolfgang, B. and Markus, S. (1998) *Datensicht von Informationssystemen: Datenmodellierung und Datenbanken*, Springer Bln, Berlin.
- Tekom (Hrsg.) (2005) *Effizientes Informationsmanagement durch spezielle Content-Management-Systeme: Praxishilfe und Leitfaden zu Grundlagen – Auswahl und Einführung – Systemen am Markt*, http://www.tekom.de/upload/alg/CMS_Studie.pdf

Bibliography

- Kammerer, M. (1998) ‘Kritisches zu Schnelles Applizierung einer ‘Logischen Semantik’ bei Wörterbüchern vom COBUILD-Typ’, in Mogensen, J.E., Pedersen, V.H. and Zettersten, A. (Eds.): *Symposium on Lexicography IX. Proceedings of the Ninth International Symposium on Lexicography April 23–25, 1998 at the University of Copenhagen*, Max Niemeyer, Tübingen, Lexicographica, Series Maior 103, pp.43–70, Manuscript: <http://www.matthias-kammerer.de/content/MS2000LogischeSemantik.pdf>
- Kammerer, M. (1999) ‘Zur framebasierten lexikographischen Bedeutungsbeschreibung von substantivischen Lemmzeichen’, *Lexicographica*, Vol. 15, pp.229–263, Article: <http://www.matthias-kammerer.de/content/Publ1999LexJb15.pdf>
- Kammerer, M. (2000) *Lemmzeichentypen für deutsche Verben. Eine lexikologische und metalexikographische Untersuchung*, Max Niemeyer, Tübingen, Lexicographica, Series Maior 104 [Especially chapter 3].
- Sieber, T. and Kammerer, M. (2006a) ‘Sind Metadaten bessere Daten?: Metadaten als Mittler zwischen Daten und Prozessen’, *Technische Kommunikation*, Vol. 5, pp.56–58.

- Sieber, T. and Kammerer, M. (2006b) *Daten, Wissen und Information: Eine Grundlagenanalyse unter besonderer Berücksichtigung der Technischen Dokumentation*, Ms. Miskolc-Egyetemváros, Walldorf.
- Sieber, T. and Kovács, L. (2006) 'Unbemerkt und unsichtbar: Metadaten als Triebfeder für intelligente Content-Automatisierung', *Technische Kommunikation*, Vol. 6, pp.42–45.
- Sieber, T. and Wolfgang, B. (2006) 'Ich weiß, was Sie meinen!: semantic Web – Kommunikation auf gleichem Niveau', *Technische Kommunikation*, Vol. 2, pp.56–59.