Semantic heterogeneity: comparing new semantic web approaches with those of digital libraries

Krause, Jürgen

Empfohlene Zitierung / Suggested Citation:
Krause, J. (2008). Semantic heterogeneity: comparing new semantic web approaches with those of digital libraries. *Library Review*, *57*(3), 235-248. https://doi.org/10.1108/00242530810865501

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0024-2535.htm

Semantic heterogeneity: comparing new semantic web approaches with those of digital libraries

Jürgen Krause
GESIS-IZ, Bonn and Computer Science Department, University of Koblenz-Landau, North Rhine-Westphalia, Germany

Received 19 October 2007
Reviewed 19 October 2007
Accepted 20 November 2007

Abstract

Purpose – To demonstrate that newer developments in the semantic web community, particularly those based on ontologies (simple knowledge organization system and others) mitigate common arguments from the digital library (DL) community against participation in the Semantic web.
Design/methodology/approach – The approach is a semantic web discussion focusing on the weak structure of the Web and the lack of consideration given to the semantic content during indexing.
Findings – The points criticised by the semantic web and ontology approaches are the same as those of the DL "Shell model approach" from the mid-1990s, with emphasis on the centrality of its heterogeneity components (used, for example, in vascoda). The Shell model argument began with the "invisible web", necessitating the restructuring of DL approaches. The conclusion is that both approaches fit well together and that the Shell model, with its semantic heterogeneity components, can be reformulated on the semantic web basis.
Practical implications – A reinterpretation of the DL approaches of semantic heterogeneity and adapting to standards and tools supported by the W3C should be the best solution. It is therefore recommended that – although most of the semantic web standards are not technologically refined for commercial applications at present – all individual DL developments should be checked for their adaptability to the W3C standards of the semantic web.
Originality/value – A unique conceptual analysis of the parallel developments emanating from the digital library and semantic web communities.

Keywords Worldwide web, Digital libraries, Modelling, Information management
Paper type Conceptual paper

Alternative concepts to web search engines like Google: Semantic foundations and heterogeneity

In order to improve content indexing beyond that of Web search engines such as Google, one scientific approach is increasingly being discussed: the Semantic Web based on ontologies. The Semantic Web discussion (see for an overview, Fensel, 2001; Stuckenschmidt and Harmelen, 2005; Staab and Studer, 2004) began by focusing on the weak structure of the web and the lack of consideration for semantic information in content indexing. The research groups working on the Semantic Web maintained from the very beginning that the belief information retrieval models used by commercial search engines would lead to acceptable results – using automatic indexing of the websites in conjunction with an intelligent usage of hyperlinking structures – was not justified. Both groups began primarily from the "visible web". This was then further expanded with other sources of the "invisible web" (e.g. subject-specific databases, repositories, knowledge bases for companies, etc.).

The points criticised by the Semantic Web and ontology approaches match those of the "Shell model approach" (*Schalenmodell*, Krause, 2006; 2007). The Shell model, which appeared in the mid-1990s, had an emphasis on the centrality of its heterogeneous

Library Review
Vol. 57 No. 3, 2008
pp. 235-248

0024-2535
DOI 10.1108/00242530810865501

components. The argument of the Shell model did not begin with the web, but demanded restructuring and new research approaches for digital libraries and specialised information providers. In fact, the argument for the Shell model began with the "invisible web", which would then open to the visible web. In this case the actual challenge was that the paradigm of homogenisation through standardisation was partially sacrificed or was to be at least supplemented by intelligent processes for handling heterogeneity[1]. Here, just like with the Semantic Web approach, heterogeneity of semantic content analysis and indexing by various groups of providers was a central topic.

The Shell model and its "heterogeneity components" represents a general framework in which specific types of documents with differing content indexing can be analysed and algorithmically related (Krause, 2006). Key to this are intelligent transfer components between the different types of content indexing (a special form of cross walk or mapping) that can accommodate the semantic differences. They conceptually interpret the technical integration between individual databases with differing content indexing systems by relating the terminologies of the domain-specific and general thesauri, classifications, etc. to each other.

Semantic Web and ontologies

According to Fensel (2001, p. VI), an ontology is "a community mediated and accepted description of the kinds of entities that are in a domain of discourse and how they are related". Stuckenschmidt and Harmelen (2005, p. IX) define the problem to be solved by ontologies as "information sharing". This encompasses the integration of heterogeneous information sources and results.

> [T]he problem of information sharing . . . is only solvable by giving the computer better access to the semantics of the information . . . we not only need to store such obvious metadata as its authors, title, creation date, etc., but we must also make available in a machine accessible way the important concepts that are discussed in the documents, the relation of these concepts with those in other documents, relating these concepts to general background knowledge.

Clearly information and documentation centres and libraries with their classifications, standardised authority files and thesauri, have a long tradition of characterising the intellectual content of documents. Consequently Umstätter and Wagner-Döbler (2005, p. 54) argue:

> Ontologies in libraries, information science, and computer science are thesauri in which the basic meanings of word fields and their relations to one another are depicted in computers.

Hahn and Schulz (2004, p. 134), who take medical thesauri as a starting point for constructing "true" ontologies, clearly state that advocates of ontologies generally view these relations as insufficient:

> [UMLS Metathesaurus] [Their] semantics are shallow and entirely intuitive, which is due to the fact that their usage was primarily intended for humans . . . there is no surprise that the lack of a formal semantic foundation leads to inconsistencies, circular definitions, etc. . . . This may not cause utterly severe problems when humans are in the loop [as] its use is limited . . . [to] document retrieval tasks. Anticipating its use for more knowledge-intensive applications such as medical decision making . . . those shortcomings might lead to an impasse.

Ontologies should therefore enable more than just the homogenisation of search terms during database searching, which is why they require the possibility of automatic deductive processes.

> Ontologies are formal structures supporting knowledge sharing and reuse (Fensel, 2001, p. 1).

> An ontology provides a domain theory and not the structure of a data container. In a nutshell ontology research is database research for the twenty-first century, where data needs to be shared and does not always fit into a simple table (Fensel, 2001, p. 10).

Heterogeneity among multiple ontologies

Given the aforementioned ontological comments and definitions, it is clear that it is the formal deductive processes that solve the problem of heterogeneity between various different information sources. This was a central component of the ontology approach for the Semantic Web from the very beginning.

> Local models must be interwoven with other models, such as the social practice of the agents that use ontologies to facilitate their communication needs ... We no longer talk about a single ontology, but rather about a network of ontologies. Links must be defined between these ontologies and this network must allow overlapping ontologies with conflicting – and even contradictory – conceptualizations (Fensel, 2001, p. 4).

Stuckenschmidt and Harmelen (2005) discuss two approaches for carrying out information sharing among multiple ontologies in a network:

- Each information domain utilises its own vocabulary and develops an ontology (the multiple ontology approach). There is no arrangement regarding a common vocabulary or a minimal global ontology. Stuckenschmidt and Harmelen (2005, p. 33) clearly reject this approach:

> In reality the lack of a common vocabulary makes it extremely difficult to compare different source ontologies ... the mapping has to consider different views of a domain ... the mapping is very difficult to define, because of the many semantic heterogeneity problems which may occur

- Hybrid approaches allow different ontologies for different information domains, but are based on a jointly utilised (and agreed upon) vocabulary, which facilitates deductive mapping. The drawbacks are equally obvious:

> The drawback of hybrid approaches, however, is that existing ontologies cannot be re-used easily (Stuckenschmidt and Harmelen, 2005, p. 34).

Problem areas from the perspective of information and documentation

Just like the Shell model, ontologies in the framework of the Semantic Web focus on the inevitable heterogeneity that arises when attempting to exploit different information sources, a difficulty that needs to be overcome. They also emphasise the value of an in-depth semantic indexing in comparison to approaches used by general web search engines. Two observations are important when making comparisons to the Shell model and its heterogeneity components:

- First, the difference between ontologies and thesauri is not so much about the concept, but about indexing depth.
- And second, the semantic foundation of content analysis and indexing improves and brings with it the hope for better search results.

Knorz and Rein (2005, p. 1) illustrate the dilemma of this approach:

> This hope can be justified but not very well verified. Usages of ontologies can be extensively ordered along two ends of a spectrum ... these usages of ontologies cover either a broad area

> and work with semantic, barely differentiated relations, or they work with respect to attributes and relations in a very differentiated way with simultaneous limitations in a mini-world. In the first case the result is barely better than what a conventional thesaurus delivers, in the second case . . . the essential result can be obtained via conventional databases.

Multiple ontologies are specifically allowed in the Semantic Web. This means that the homogeneity paradigm of the LIS community to enforce standardisation through homogeneity appears to have been overcome. However, preferring the hybrid approach indicates that, instead, ontology modelling relies ultimately on a common vocabulary and defined relations between both terminological systems. In turn, this means that the direct utilisation of the invisible web – used as a basis up to now – is problematic, at least in the specialised science context.

The Shell model and bilateral transfer modules

Essentially then, ontologies are also focused on standardisation. This standardisation assumes a new form by considering various perspectives to be integrated. Principally, they try to do the same as the centralist approaches of the seventies in the information and documentation domain, but on a different level. The models of both groups coexist and agree on cooperation, and do so without means for hierarchical implementation. The classic demand for comprehensive standardisation makes sense and is, per se, not wrong: No heterogeneous components are needed if everyone uses the same thesaurus or the same classification. As long as everyone knows that efforts at standardisation can be only partially successful, then everything speaks in favour of initiatives such as these. But, no matter how successful they are in one subject area, the remaining heterogeneity, for example, with respect to various types of content analysis and indexing (e.g. automatic, different thesauri, various classifications, differences in categories included.) would be too great to ignore. The broad access to the web speaks against a centralist doctrine of content indexing *per se*.

The question then remains for both approaches – for the Shell model and for ontologies. In contrast to ontology research, the Shell model emphasised the usage of existing semantic knowledge from the very beginning. Thesauri and classifications have been further refined over decades and connected through the process of intellectual indexing, specifically with the high quality information sources of the "invisible web". Their intelligent usage promises, in the mid-term, the greatest progress vis-à-vis those of Google, or even traditional specialised databases and library catalogues.

As mentioned in the introduction, the Shell model and its heterogeneity components represents a general framework in which specific types of documents with differing content indexing can be analysed and algorithmically related. Intelligent transfer components between the different types of content indexing (a special form of cross walk or mapping) that can accommodate the semantic differences are central to its operation. They conceptually interpret the technical integration between individual databases with differing content indexing systems by relating the terminologies of the domain-specific and general thesauri, classifications, etc. to each other.

So far, this approach of handling semantic heterogeneity has been mainly implemented with the German science portal *vascoda* (www.vascoda.de) and the social science portal *sowiport* (www.sowiport.de), developed at the GESIS information centre in Bonn, Germany.

vascoda is currently the most important project for achieving a new and innovative infrastructure in the field of scientific information in Germany. This new science portal merges controlled and qualitative collections from more than 40 providers. Its aim is to

integrate high-quality information from both the deep and the visible web by using search engine technology (FAST) and new concepts to integrate the data, not just technically, but to also solve the problem of semantic heterogeneity for achieving a high level of quality.

Conceptually, the science portal *vascoda* is constructed upon two building blocks. These include a governing science portal, and the relatively independent-acting specialist portals of the individual academic subjects. Their construction also entails an additional problem area connected to that of semantic heterogeneity. Building up specialist portals like *sowiport* (for the social sciences) can be viewed as a multi-level process. It integrates national and international information of different types (metadata and full-text) and makes it available prepared for retrieval. This system offers the possibility for electronic publishing and discourse activities (i.e. discussion components), which is expanded to communication platforms via the search portals. In the long-term, this should lead to new forms of and higher quality of scientific working (Krause, 2007; Depping, 2007; Stempfhuber, 2006).

Two different types of transfer modules have been implemented in *sowiport* and *vascoda*:

- *Cross-concordances*: the different terminology systems of classifications and thesauri are analysed in their context of usage and the terminologies are intellectually mapped to one another.
  This concept is not to be confused with metathesauri. Cross-concordances of the Shell model only contains those parts of the vocabulary where there are general semantic connections. A lot of terms can therefore remain unrelated. Cross-concordances only cover the static part of the transfer problem. This also differentiates them from ontological approaches.
- *Quantitative-statistic approaches:* the transfer problem can generally be modelled as a vagueness problem between two content description languages. Various methods (e.g. probability methods, fuzzy approaches, rough set theory and neural networks) were suggested for the vagueness in information retrieval between user terminology and the database content that were also used for the transfer problem.

What is essential is that the transfer modules bilaterally operate on the database level (see Krause, 2004 for more details), connecting the different content description terms. None of the approaches carries the burden of transfer alone. They are entwined with one another and act in unison. This is both conceptually, as well as in practice, somewhat different from the traditional handling of vagueness between the user query on the one side and the document content of all databases on the other.

A transfer module (e.g. V2) can be applied bilaterally between, for example, the content of a document that was indexed with a general key word list – as with SWD[2] – and a second document whose indexing is based upon a domain-specific thesaurus (like the social science thesaurus produced by the GESIS Social Science Information Centre (GESIS-IZ)) through qualitative processes such as cross-concordances and/or by another type of transfer module (Figure 1). The search algorithm can then establish a connection to the user terminology (V1) with a probabilistic method (Figure 2).

In comparison to current information retrieval solutions, the important distinction is the possibility of using a specific type of transfer according to circumstances, and not just dealing with the problem of different terminology systems in an undifferentiated fashion.

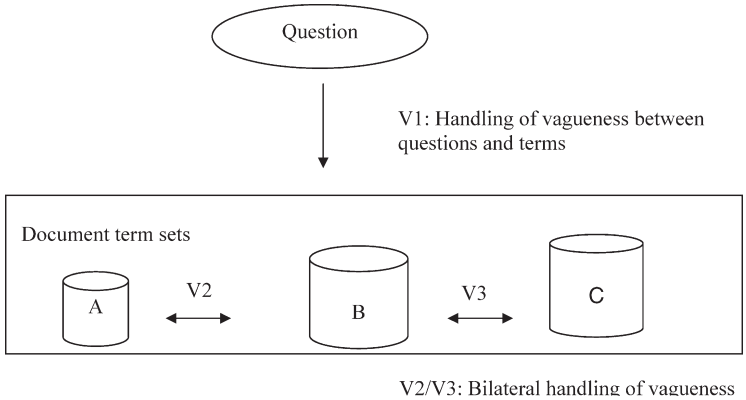

Figure 1.
Bilateral transformation

Bilateral transfer modules are – from a model building perspective – a very simple and basic building block that can quickly become very complex. The interplay of the transfer modules can be easily analysed and managed in the context of previous applications as is currently the case with *sowiport* or *vascoda*. Things can be expected to change quickly if the huge number of variations found on the visible and invisible web are taken into account. For this reason, the suggested model also needs more abstract modules that operate on a higher level of integration. The Shell model should accomplish this by complementing the bilateral transfer modules with a number of additional assumptions. For example, different levels of content indexing and document relevance are integrated into shells that are interconnected with each other through higher level transfer modules. The concept of bilateral transfer modules is now far enough advanced that it can be applied practically and promising initial empirical results are available (Krause, 2006).

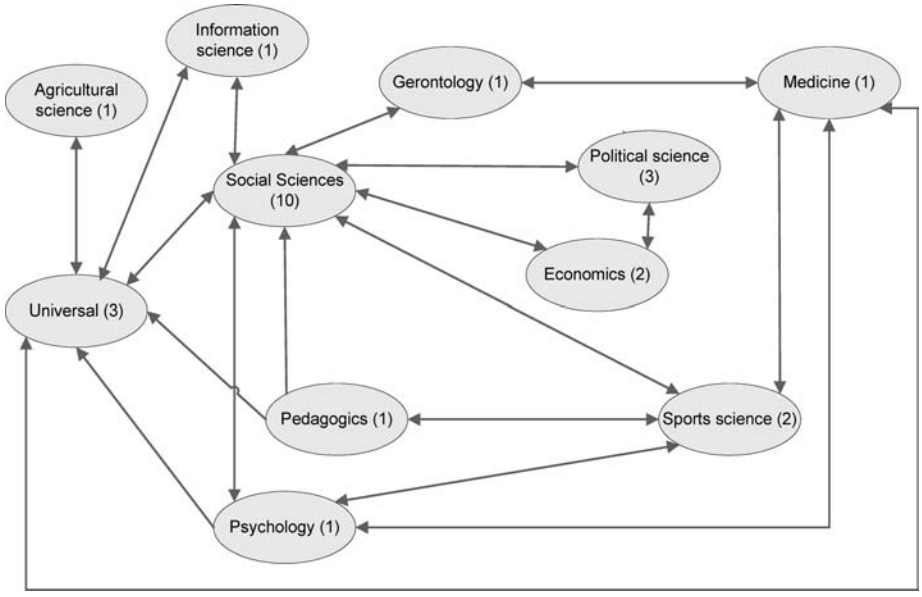

Figure 2.
Subjects and controlled vocabularies connected (by Philipp Mayr, GESIS-IZ, Bonn)

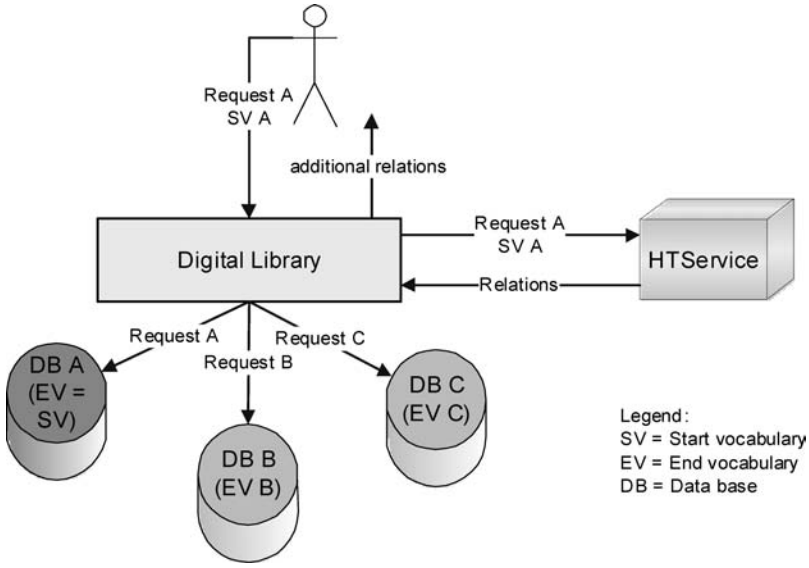

Figure 3.
Heterogeneity Service (HTS, by Philipp Mayr, GESIS-IZ, Bonn)

In the context of the GESIS–IZ project "competence centre modelling and treatment of semantic heterogeneity" (financed by the Federal Ministry of Education and Research, BMBF), 64 cross-concordances (including 513,000 term relations, based upon 25 controlled vocabularies) were developed (for more details see www.gesis.org/en/research/information_technology/komohe.htm, Mayr and Walter, 2007a, b) between 2004 and 2007. The project is the largest terminology mapping effort in Germany.

A heterogeneity service (HTS) was built up and implemented in *sowiport* using the social science cross-concordances. The HTS will be integrated into *vascoda* over the next six months. The service supports two scenarios: direct transformation of submitted user terms into equivalence relations and the presentation of the additional relations (of the cross-concordances) for users (Figure 3).

The HTS translates between the various content indexing vocabularies. This system currently functions only if the vocabularies are interconnected with one another via cross-concordances. Users oftentimes choose their terms freely, meaning that these are either not, or only accidentally, part of controlled vocabularies. For this reason the improvement of an added service has been developed that reformulates a users' freely chosen terms into a term(s) from a suggested controlled vocabulary. This so-called search term recommender increases hit accuracy (Mayr *et al.*, 2008).

Another addition is a complementary service, simultaneously employing two procedures borrowed from scientometry and network analysis. This service allows the re-ranking of the results list according to structural criteria. This entailes ranking the hit list according to core periodicals (so-called Bradfordising), and ranking according to the author's significance within author networks (Mayr *et al.*, 2008). The exceptional quality of these complementary services is gained through their connection to the other components. They specifically focus both on the search and the results (re-ranking), positively influencing one another[3].

The following section demonstrates that both approaches – the ontologies of the Semantic Web and the bilateral transfer module of the Shell model – supplement rather than exclude each another.

The Shell model and its heterogeneity transfer modules vs ontologies of the Semantic Web

Both the Shell model and the Semantic Web approaches acknowledge the weaknesses of the content analysis and indexing procedures employed by general web search engines. Both note the semantic foundation of information and its heterogeneity, and accept that heterogeneity is an acknowledged component of information sharing.

In principle there is no conceptual discrepancy regarding the deeper semantic indexing which the ontologies aspire to. The German language example of searching for "Befinden von Kindern berufstätiger Mütter" (well-being of children of working mothers) in infoconnex (www.infoconnex.de) already poses a classic difficulty in the literature search. It reflects the criticism that thesauri only yield very limited relations, predominantly broader terms, narrower terms, similar terms and synonyms. Most references found with the above mentioned search terms yield documents dealing with the well-being of working mothers, not on the well-being of the child. The irrelevant hits can only be avoided through a more comprehensive explanation of the complex relations between the terms used in the thesauri. Since this would lead to a dramatic increase in the indexing work, thesauri usually do not contain these more complex relations. The underlying thesis is that the precision gained in some cases does not justify the increased effort. The conceptual background to information retrieval in this context is that the semantics will always remain unanalysed with regard to users' search terms and descriptors for indexing. In contrast to that, ontologists require a formal semantic foundation with the possibility of formal deductive processes. Advocates of currently used thesauri argue that human intelligence supplements these non-analysed parts in man-machine interactions (Kuhlen, 2004)[4]. The question in traditional information retrieval and with regard to the Shell model is not whether parts of the semantics remain unanalysed, but rather whether the (intelligent) user is able to supplement these parts with acceptable (minor) effort when searching.

In Krause and Stempfhuber (2005), one finds an indication for an increased semantic foundation with respect to the development of a social science portal, specifically relating to the "text-fact" integration of social science survey data and literature documents. In other words, there was a strong indication that users are typically unable to compensate for the missing semantic analysis when dealing with survey data rather than text. The fact is that the necessary compensation by users' intelligence and research context fails, and not just in individual cases: It is the rule, rather than the exception.

This example demonstrates that, in principle, no conceptual discrepancy exists between the bilateral transfer components of the Shell model and the ontology approach. The first may be interpreted as the lower level of an ontology, with reduced requirements regarding the depth of semantic indexing and limited deductive capabilities. The theoretical basis for such an approach is the aforementioned thesis; that within information retrieval contexts semantics may remain unanalysed since it is assumed that they can be compensated by user (human) intelligence. The natural language – partially not understood by the machine – serves, in this case, as a transport medium. Based on this, the semantic knowledge of thesauri and classifications may be used with relatively simple procedures and with the help of bilateral heterogeneity components, without blocking the possibility for areas in which more in-depth and more logically precise – but also more labor intensive ontology approaches – may become necessary.

The discussion in the first section therefore suggests there is a common goal, but the approach of the Semantic Web may serve neither practically nor conceptually as the

exclusive basis of a digital library application such as *vascoda*. The following section demonstrates that this statement may be correct in the medium-term, but that newer developments in the Semantic Web suggest that the approaches may grow together over the longer term.

Newer Semantic Web developments significant for the digital library

Sharing ideologies vs sharing technology and standards

Berners-Lee *et al.* (2001) articulate one of the most well known visions of Semantic Web, indicating that it should be ". . .an extension of the current [web], in which information is given well-defined meaning, better enabling computers and people to work in cooperation". Herrmann (2007) and Ankolekar (2007) no longer see these visions as binding; instead, they replace this with a series of practical goals:

The Semantic Web gives

- common, interoperable standards for machine processable data and metadata markup;
- data interoperability across knowledge sources, applications, communities, organizations;
- architecture for interconnected vocabularies and communities;
- reasoning about distributed web information . . . (Ankolekar, 2007, slide 8).

It is no longer about images such as the now famous "silver bullet"[5], but about standardisation work and joint development of general tools which can be used to realise the various visions[6].

Web ontology language (OWL) vs simple knowledge organization system (SKOS)

The Semantic Web community at W3C has become receptive to other theoretical approaches. This has been most evident in the use of ontology. Their differences with respect to the philosophy of thesauri and classifications have thus far represented a clear obstacle for digital library applications; however, there is a close connection with the Semantic Web community: that the deep and invisible web are important.

Seen from a non-visionary standardisation perspective, the W3C ontologies are about the following:

> Idea: extend web markup standards to go beyond syntax and express structure and semantics of the data
>
> (a) ambitious goal: to enable intelligent software agents to reason about information on the web;
>
> (b) less ambitious goal: to enable interoperability of data (Ankolekar, 2007, slide 6);
>
> . . . an ontology in computer science is essentially an engineering artifact . . .
>
> (c) a formal and machine-executable model of a domain of interest;
>
> - consists of a specific vocabulary to describe the domain of interest
> - a set of explicit assumptions regarding the intended meaning of the vocabulary in a logical language
>
> (d) representing a shared understanding of the domain of interest;
>
> - to help humans and machines reach common understanding
> - can import information from other ontologies (Ankolekar, 2007, slide 6)

OWL Full, OWL DL, OWL Lite

Since 2004, stable proposals have been established for ontology languages, fulfilling the requirement to support deduction processes. These are different in terms of potential deductive power (Ankolekar, 2007, slide 27).

The varies levels of OWL expressiveness ("Lite", "digital library (DL)", and "Full") exist as a result of the desire to have the necessary power of deductive languages, but to balance the resulting increased effort and difficulties of implementation (Figure 4). Language complexity and expressiveness therefore decreases from "Full" to "Lite":

- "Full" is the whole thing,
- "description logic (DL)" restricts Full in some respects,
- "Lite" restricts DL even more (Herman, 2007b, slide 111).

Simultaneously, there was a desire not to compromise on the ability to exchange information between all types of ontologies on the Semantic Web.

Simple knowledge organization system (SKOS)[7]

The advocates of Semantic Web expected a broad cooperative development of ontologies through the introduction of OWL standards; however, such a cooperative, voluntary (and unpaid) joint endeavor has not, in fact, developed in ontologies over the last ten years.

The idea underpinning the cooperative development of large information collections may indeed be successful on the Web; after all, Wikipedia is a fine example of this. However, writing partial ontologies which automatically connect (thanks to standardisation and tools) appears to be not such an interesting and stimulating activity for scientists. It can hardly be expected that this will change after ten years of intensive effort. This conclusion may be the main reason for Semantic Web advocates seeking alternatives.

Ultimately, the desire for stronger deductive processes has been partially sacrificed via the development of the SKOS (Miles and Brickley, 2005). The only promising strategy therefore remains utilising the existing thesauri and classifications, even if they are not "true" ontologies (also Miles, 2006). This is where the Semantic Web meets the semantic heterogeneity components of the Shell model and the cross-concordances

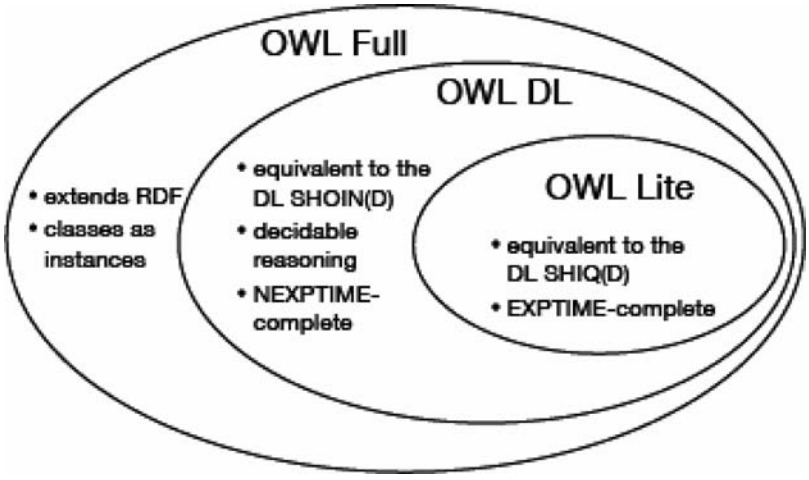

Figure 4.
The varying levels of OWL expressiveness

of the digital libraries. Integration of those is possible because the emphasis has shifted away from the previous visions and towards the standardisation and implementation of commonly used technologies.

The explanations for "webifying" thesauri and classifications have nothing to do with the theoretical explanations for employing classifications and thesauri in information science (see the following section):

> However: ontologies are hard:
>
> - a full ontology-based application is a very complex system
> - hard to implement, may be heavy to run. . .
> - . . . and not all applications may need it! (Herman, 2007b, slide 111)
>
> OWL's precision is not always necessary or even appropriate:
>
> - OWL is a sledge hammer/SKOS a nutcracker", or "OWL a Harley/SKOS a bike
> - they complement each other, can be used in combination to optimise cost/benefit
>
> Role of SKOS is:
>
> - to bring the worlds of library classification and Web technology together
> - to be simple and undemanding enough in terms of cost and required expertise" (Herman, 2007b, slide 128)
>
> Formal ontologies (like OWL) are important, but use them only when necessary:
>
> - you can be a perfectly decent citizen of the Semantic Web if you do not use ontologies, not even RDFS. . . (Herman, 2007a, slide 51)
>
> Languages should be a compromise between rich semantics for meaningful applications, feasibility, implementability (Herman, 2007b, slide 95)

SKOS-rationale vs information science rationale

The information science explanation for using classifications or thesauri instead of the richer semantics of ontologies, is different from that cited above by the W3C–Group.

As already mentioned, thesauri do without the richer semantic relations so that the effort needed for indexing remains small. The theory being that the additional gain in precision in individual cases does not justify the extra effort. The conceptual background is that in information retrieval, some semantic content remains unanalysed with regard to the search terms of the user and the descriptors used for indexing. Human intelligence supplements these non-analysed parts during machine interaction. The question in traditional information retrieval and in the Shell model is not whether parts of the semantic content remain unanalysed, but whether the (intelligent) user is able to supplement these parts with acceptable, minor effort. With SKOS, this can serve directly as a theoretical justification for the new W3C variant of ontologies. The newer developments of the Semantic Web, however, no longer requires acceptance of these theoretical thoughts as a necessary condition. SKOS can, but need not be, theoretically founded in a library and information science sense. In the Semantic Web, SKOS is currently accepted on a purely pragmatic basis as a way to achieve technical interoperability. In reality, there is hardly another possibility for the Semantic Web community than to permit this form of "weak" semantic foundation. One can therefore expect SKOS to rise in popularity.

Ontology sharing and statistical handling of vagueness

Following the acceptance of SKOS by the Semantic Web community, there remains no standardisation activities within the Semantic Web that revolve around the quantitative-statistical, non-deductive vagueness components in information retrieval.

At the International Conference on Semantic Web and Digital Libraries 2007 in Bangalore, Herman (2007b) reported that an incubation group had formed at the W3C. This group is working on alternative approaches based on description logic (fuzzy and statistical approaches). This development, combined with the existing Semantic Web portfolio, opens a path of enquiry which appears indispensable for digital libraries.

- Fuzzy logic
  - look[ing] at alternatives of description logic based on fuzzy logic
  - alternatively, extend[ing] RDF(S) with fuzzy notions
- Probabilistic statements
  - hav[ing] an OWL class membership with a specific probability
  - combin[ing] reasoners with Bayesian networks (Herman, 2007b, slide 46)

Conclusion

Recent developments in the Semantic Web mitigate many of the objections often raised by the digital library community against participation in its development. These recent developments also allow the Shell model to be reformulated, using these developments as a basis.

The advantages are obvious. Whereas standardisation, for instance with the science portal *vascoda*, had to be negotiated among all participating partners in the group, all offers for integration which adhere to the W3C standards, and ideally use their general tools, would be accessible without entering into further time consuming negotiations.

This line of development is currently not technologically refined for commercial applications, as initial applications of SKOS and associated tools have so far demonstrated. This is why it is not yet feasible to replace *sowiport*'s or the HTS with a SKOS model. Nevertheless, in the mid-term a reinterpretation of these approaches and an adaptation to the standards and tools supported by the W3C should be the best solution for the heterogeneity problem of digital libraries, like *sowiport* and *vascoda*. It is therefore recommended that all current developments should be checked for their adaptability to the W3C standards of the Semantic Web group. Ideally, test applications should be developed using the SKOS and SPARQL.

In September 2007 the executive committee of *vascoda* (representing about 40 participants and providers of scientific information in Germany) decided to support the thesis "*vascoda* goes Semantic Web". This decision may prove advantageous for all participating information brokers, libraries and information scientists. They will benefit from the above described advantages of a reinterpretation of *sowiport*'s and *vascoda*'s heterogeneity approach. But it is ultimately beneficial to the Semantic Web, which will gain a committed digital library development group, as well as a large interconnected collection of 62 crosswalks containing a major quantity of the semantic knowledge in libraries and documentation centres in Germany.

Notes

1. Generally, the shell model refers to the existence of various levels of indexing, quality and consistency into which documents may be grouped. The specific technique for

handling semantic heterogeneity between the different shells was developed as a refinement mainly since 2000. Today it is a central part of this model (Krause, 2006).

2. SWD is the keyword-norm data file created through the cooperation of German scientific universal libraries on the basis of the RSWK (rules for the keyword catalogue).
3. See Mayr *et al.* (2008) in this same journal for a detailed explanation of these added services.
4. This thesis has to be seen in connection with the empirical finding that the user typically repeats multiple queries on the same search in succession.
5. "... the next generation of the web, called the Semantic Web. To achieve even some of the promises for these technologies, we must develop vastly improved solutions for addressing the Grand Challenge of Information Technology, namely dealing better with semantics. . .. This challenge has been calling out for a silver bullet since the beginning of modern programming" (Fensel, 2001, p. V).
6. The symbol for this perspective is the oft-cited "Semantic Web Layer Cake" (Ankolekar, 2007, slide 9).
7. An overview of projects with SKOS can be found at: http://esw.w3.org/topic/SkosDev/DataZone

References

Ankolekar, A. (2007), Tutorial "Semantic Technologies for Digital Libraries", *Semantic Web & Digital Libraries, Proceedings ICSD, International Conference*, *Bangalore*, *21-23 February 2007*.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), "The semantic web", *Scientific American*, Vol. H. 5, pp. 34-43.

Depping, R. (2007), "vascoda.de and the system of the German virtual subject libraries", *Semantic Web & Digital Libraries, Proceedings ICSD, International Conference*, *Bangalore*, pp. 304-14.

Fensel, D. (2001), *Ontologies – A Silver Bullet for Knowledge Management and Electronic Commerce*, Springer, Berlin.

Hahn, U. and Schulz, S. (2004), "Building a very large ontology from medical thesauri", in Staab, S. and Studer, R. (Eds), *Handbook on Ontologies*, Springer, Berlin, pp. 133-50.

Herman, I. (2007a), "State of the Semantic Web", *Key Note Speech Semantic Web & Digital Libraries, Proceedings ICDS, International Conference*, *Bangalore*, *21-23 February 2007*.

Herman, I. (2007b), Tutorial "Introduction to the Semantic Web", *Semantic Web & Digital Libraries, Proceedings ICDS, International Conference*, *Bangalore*, *21-23 February 2007*.

Knorz, G. and Rein, B. (2005), "Semantische Suche in einer Hochschulontologie", *Information, Wissenschaft & Praxis*, Vol. 56 No. 5.

Krause, J. and Stempfhuber, M. (2005), "Nutzerseitige Integration sozialwissenschaftlicher Text- und Dateninformationen aus verteilten Quellen", in Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute (ADM), Arbeitsgemeinschaft Sozialwissenschaftlicher Institute (ASI), Statistisches Bundesamt, Wiesbaden (Eds), *Datenfusion und Datenintegration 6*, Bonn, S. pp. 141-58.

Krause, J. (2006), "Shell model, semantic web and web information retrieval", in Harms, I., Luckhardt, H.-D. and Giessen, H.W. (Eds), *Information und Sprache, Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern*, Festschrift für Professor Dr. Harald H. Zimmermann, K.G. Saur, München, S. pp. 95-106.

Krause, J. (2007), "The concepts of semantic heterogeneity and ontology of the semantic web as a background of the German science portals *vascoda* and *sowiport*", *Semantic Web & Digital Libraries, Proceedings ICDS, International Conference*, *Bangalore*, *21-23 February 2007*, pp. 13-24.

Kuhlen, R. (2004), Kapitel A 1: "Information", in Kuhlen, R., Seeger, T. and Strauch, D. (Eds), *Grundlagen der praktischen Information und Dokumentation, Band 1, Handbuch zur Einführung in die Informationswissenschaft und -praxis*, Saur, München, Ausgabe 5, pp. 3-20.

Mayr, P. and Walter, A.-K. (2007a) "Zum Stand der Heterogenitätsbehandlung in *vascoda*: Bestandsaufnahme und Ausblick", in Bibliothek & Information Deutschland (Ed.), *Information und Ethik 3. Leipziger Kongress für Information und Bibliothek*, 19-22 März 2007, Verlag Dinges und Frick, Leipzig, available at: www.opus-bayern.de/bib-info/volltexte/2007/290/ (accessed 19 December 2007).

Mayr, P. and Walter, A.-K. (2007b), "Einsatzmöglichkeiten von Crosskonkordanzen", in Stempfhuber, M. (Ed.), *Lokal-Global, Vernetzung wissenschaftlicher Infrastrukturen*, 12. Kongress der IuK-Initiative der Wissenschaftlichen Fachgesellschaft in Deutschland, Tagungsberichte, GESIS-IZ Sozialwissenschaften, Bonn, pp. 149-66, available at: www.gesis.org/Information/Forschungsuebersichten/Tagungsberichte/Vernetzung/Mayr-Walter.pdf (accessed 19 December 2007).

Mayr, P., Mutschke, P. and Petras, V. (2008), "Reducing semantic complexity in distributed digital libraries: treatment of term vagueness and document re-ranking", *Library Review*, Vol 57 No. 3.

Miles, A. (2006), "SKOS – requirements for standardization", *International Conference on Dublin Core and Metadata Applications*, *Mexico*, *3-6 October 2006*.

Miles, A. and Brickley, D. (2005), "SKOS Core guide", *W3C Working Draft, 2 November 2005 (Work in progress)*, available at: www.w3.org/TR/swbp-skos-core-guide/ (accessed 19 December 2007).

Prasad, A.R.D. and Madalli, D.P. (Eds) (2007), "Semantic web & digital libraries", *Proceedings ICDS, International Conference*, *Bangalore*, *21-23 February 2007*.

Staab, S. and Studer, R. (2004), *Handbook on Ontologies*, Springer, Berlin.

Stempfhuber, M. (2006), "Data integration in current research information systems", Magalhães, de S. T., Santos, L., Stempfhuber, M., Fugl, L. and Alrø, B. (Eds), *CRIS-IR 2006, Proceedings of the International Workshop on Information Retrieval on Current Research Information Systems*, *Copenhagen*, *Denmark*, *9 November 2006*, Minho, pp. 29-51.

Stuckenschmidt, H. and Harmelen, F. van (2005), *Information Sharing on the Semantic Web*, Springer, Berlin.

Umstätter, W. and Wagner-Döbler, R. (2005), *Einführung in die Katalogkunde - Vom Zettelkatalog zur Suchmaschine*, Hiersemann, Stuttgart.

Corresponding author
Jürgen Krause can be contacted at: juergen.krause@gesis.org