The Concepts of Semantic Heterogeneity and Ontology of the Semantic Web as a Background of the German Science Portals vascoda and sowiport

Jürgen Krause

IZ Social Science Information Centre of the GESIS,
University of Koblenz-Landau, Institute of Computer Science
Bonn and Koblenz, Germany
krause@bonn.iz-soz.de

vascoda (www.vascoda.de) is today the most important project to achieving a new innovative infrastructure in the field of scientific information in Germany. The aim is to integrate high-quality information from the deep and from the visible web by using search engine technology (FAST) and new concepts to integrate the data, not only technically, but to solve the problem of semantic heterogeneity at a high level of quality. In contrast to the ontology and semantic web approach that of semantic heterogeneity in the context of the shell model started from the invisible web, opening itself to the visible, not vice versa, and is putting the reuse of existing semantic knowledge (thesauri) in the foreground. The consequences of these differences and the common features of both approaches are in the focus of the paper.

Keywords: semantic heterogeneity, shell model, semantic web, ontology, vascoda, sowiport

1 Introduction

The middle of the 1990s are coined by the increased enthusiasm for the possibilities of the WWW, which has only recently deviated – at least in relation to scientific information – for a differentiated measuring of its advantages and disadvantages. Web information retrieval originated as a specialized discipline with great commercial significance.

Another line of thought has gained momentum in the last years. With Google Scholar it is becoming clear that the sciences are no longer just being seen as a windfall gain of a commercially rather uninteresting group of specialists. In this context it's also to be expected that a scientific approach – namely that of the semantic web on the basis of ontology – will be included in the discussion of its practical uses. Its development began a decade ago with a critical counter-position to the weak structuring of the web and the mostly lacking consideration of semantic information. The criticisms of the ontological approaches can be grouped with the likewise 90s-originated approach of the

ARD Prasad & Devika P. Madalli (Eds.): ICSD-2007, pp. 13-24, 2007

shell model (Krause, 2006), which called for a restructuring and new research approaches for digital libraries and specialized information providers. This was and is currently being implemented with the treatment of semantic heterogeneity in the science portal vascoda and in the social science portal sowiport (www.sowiport.de). The shell model argued from the viewpoint of the *invisible web* opening itself to the *visible web*, not vice versa like the semantic web.

Both, the visible and the invisible web, have one problem in common: As the world of scientific information providers is no longer centralized, but polycentric, groups who collect information in specialized areas can be found all over the world. A consequence of this is the lack of consistency: A term X can assume the most diverse meanings in such a system. In the narrowed field of scientific information, a descriptor X from a thesaurus that was determined with great intellectual and qualitative effort can often not be matched with term X delivered by an automatic indexing system from a fringe field. The fact that the librarians paradigm of homogenization through standardization would be at least partially sacrificed or that it could be complemented through an intelligent heterogeneous treatment procedure, was the actual challenge in constructing scientific subject portals like sowiport and vascoda (see chapter 2).

Conceptually, the science portal vascoda is built upon two building blocks: a governing science portal and the relatively independent-acting specialist portals of every academic subject[1]. The construction of them includes a second problem area in addition, but also connected, to that of semantic heterogeneity: Building up specialist portals like sowiport (for the social sciences) can be viewed as a process on many levels. It integrates national and international information of different types (metadata and full text) and offers them prepared for retrieval. At the same time, the connection to electronic publishing and discourse activities is established, which the search portals expand to communication platforms. In the long-term, this should lead to new forms and a higher quality of scientific working (see chapter 3).

2 Semantic Web, Ontology and the Shell Model[2]

Both approaches - the semantic heterogeneity components of the shell model as well as that of the semantic web with its ontology attempt (in their efforts to re-establish the lost homogeneity and consistency) to create new suitable information systems that can adequately and efficiently handle wide-spread

[1] In vascoda all important information providers of scientific information (2006: about 40 institutions) work together to integrate the largely distributed collections of scientific information in Germany and beyond.

[2] Chapter 2 is a short, summarized introduction to the essentials of (Krause, 2006).

distributed information beyond the traditional methods of librarians, but also beyond Google technology.

> " … the next generation 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, 2004, p. V)

According to Fensel 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" (Fensel, 2004, p. VI). For information providers and libraries, there is a long tradition of dealing with classifications and thesauri to represent the content of a document. Consequently some information scientists argue:

> "Ontologies in library science, information science and computer science are thesauri in which the basic meanings of semantic fields and their connections to each other are represented in computers" (Umstätter & Wagner-Döbler, 2005, p. 54, translation).[3]

Essentially, ontologists are attempting the same thing as the centralized information and documentation approaches of the 70s, if on a different level, in a new way and in observance of different aspects. Both gear their models towards cooperation agreements without nowadays having the power to force implementation. The classical demand of librarians and other information providers for overarching standardization efforts suggests itself and is logical: if everyone uses the same thesaurus or the same classification, heterogeneity components won't be needed. As long as it is clear that standardization efforts will only partially be successful, everything is in favor of these types of initiatives. Yet no matter how successful they are in a particular field, the remaining heterogeneity will be too great to neglect, for instance when dealing with different types of content indexing (automatic vs. intellectual indexing, different thesauri, classifications and metadata schemes). The question then is what model can be developed for the remaining portion of heterogeneity, after all standardization efforts have been exhausted.

In contrast to ontological research, the shell model puts the re-use of existing semantic knowledge in the foreground. Thesauri and classifications were constantly refined over decades and directly connected through intellectual

---

[3] That the thesauri relations are usually seen as inadequate by ontology advocates, shows the following example from using a medical thesauri as starting point:

> "[UMLS Metathesaurus] Its semantic is 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." (Hahn & Schulz, 2004, p. 134).

indexing processes with high-quality information sources. Their intelligent use promises – in the mid-term – the greatest advancement in comparison to the Google search.

2.1 Bilateral Transfer Modules as Part of the Shell Model

The semantic heterogeneity component of the shell model – briefly outlined in the following – represents a general framework in which specific types of documents with differing content indexing can be analyzed and algorithmically related. Key are intelligent transfer components between the different types of content indexing that can accommodate the semantic-pragmatic differences[4]. 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.

Essential is that the postulated transfer modules bilaterally operate on the database level (Krause, 2004 for more details)[5].

So far, two approaches have been implemented. None of the approaches carries the burden of transfer alone. They are entwined with each other and act in unison.

- Cross-concordances
  The different terminology systems of classifications and thesauri are analyzed in their context of usage and the terminologies are mapped to each other intellectually.
  The concept may not be confused with the one of metathesauri. A new standardization of existing terminology worlds is not intended. Cross-concordances of the shell model only contain that part of the vocabulary where general semantic connections between the existing terminology systems exist. A lot of terms can remain unrelated. This also differentiates them from ontological approaches. Cross-concordances only cover the static part of the transfer problem.
  In vascoda up till now 18 cross-concordances were developed.
- Quantitative-statistic approaches
  The transfer problem can generally be modeled as a vagueness situation between two content description languages.

[4] The bilateral transfer modules can also be conceptualized as agents (Krause, 2001).

[5] This is also 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 The distinction in comparison to current information retrieval solutions is the possibility of using a specific type of transfer in each case, according to the circumstances and not just encounter the problem of different terminology systems undifferentiated as general vagueness of the information retrieval process.

Various methods[6] were suggested for handling the vagueness in information retrieval between the user terminology and the database content. They were also used for the semantic heterogeneity components of the shell model (Zhang, 2005).

The concept of bilateral transfer modules is so well-advanced nowadays that it can be applied practically. Also the first promising empirical results are available (Marx, 2005).

2.2 Conclusion

Both approaches, the semantic heterogeneity component of the shell model and the ontology of the semantic web, share basic premises that target the nowadays weaknesses of content indexing methods: The semantic foundation of the content analyses and the acceptance of diverse methods to integrate heterogeneous information sources are essential building blocks for *information sharing*, which includes search processes. There is also no opposition between the shell model's semantic heterogeneity components and ontologists striving for a more in depth semantic indexing than is nowadays provided by general web search engines like Google or Google Scholar. The former allow them to be interpreted as a sublevel of ontology with reduced demands on the depth of the semantic indexing and limited deductive features – insofar as it only applies to the retrieval components. The theoretical basis of such an approach is that the information retrieval portion remains partially unanalyzed, because the user supplements these portions through human intelligence without difficulties. Natural language – partially not understood by the machine – serves as the transport medium. With this as a basis, the semantic knowledge of the thesauri and the classifications with the help of bilateral heterogeneous components can be used without blocking the way for sub-areas in which more in depth and logically more precise – but also more complex – ontological approaches are required.

3 sowiport as part of vascoda

The construction of a specialized portal like sowiport and its embedding in national (vascoda) and international development leads to a worldwide network of all social science-relevant information, without accepting the inevitable quality losses of general search machines. As a process, it can be viewed on several levels:

- The first level comprises the integration of various documents and data types all the way to text-facts integration (Krause & Stempf-huber, 2005), and their preparation for query;
- The second comprises the specialized field-overlapping query.

[6] Probability method, fuzzy approaches, rough set theory and neural networks.

Value-added services such as infoconnex – for education, psychology and social sciences (www.infoconnex.de) – categorize the intelligent and qualitatively premium shared query beyond subject borders. Particularly for the social sciences, it is difficult on the one hand to draw subject borders since there are many overlapping areas with other sciences; on the other hand, interdisciplinary research that goes beyond these overlapping areas has a special significance.

- The third comprises the integration of international information collections of external providers (such as CSA in sowiport) through intelligent networking;
- The fourth comprises the expansion of services for electronic publishing,
- And the fifth the combination of formal (static) and informal (dynamic) communication that – as a long-term vision – could generate a new quality of scientific working.

In the following we will discuss only the levels 4 and 5. With respect to the focus of the paper, the levels 1 – 3 are only variations of working with the concept of semantic heterogeneity as introduced in chapter 1. With 4 and 5 a new dimension is added to the underlying concepts of standardization, homogeneity vs. remaining heterogeneity and to the question of how to define high quality information und the information need of the scientific user.

3.1 Integration of Electronic Publishing

Electronic publishing has in recent years increasingly moved into the spotlight of the discussion of promotion and improvement of scientific work through the progression of information technology (IT). The use of the WWW as a communications channel brings a series of advantages with it:

- Electronic publications in comparison to print products are relatively inexpensive to produce and thus relatively easy to establish outside existing publishing structures.
- All components of the publication process can be carried out faster. Publishers, assessors and authors communicate over the Web; complex print processes are no longer a factor.
- The quantity and modality constraints of print products become non-factors. Multimedia elements can complement textual diagrams just as easily as large sets of primary data.
- In principle, anything electronic is available to all scientists worldwide today, an ideal precondition for knowledge reception and development. Claims such as those by Open Access advocates[7], transform this general possibility into a demand of scientists of politicians and infrastructure agencies.

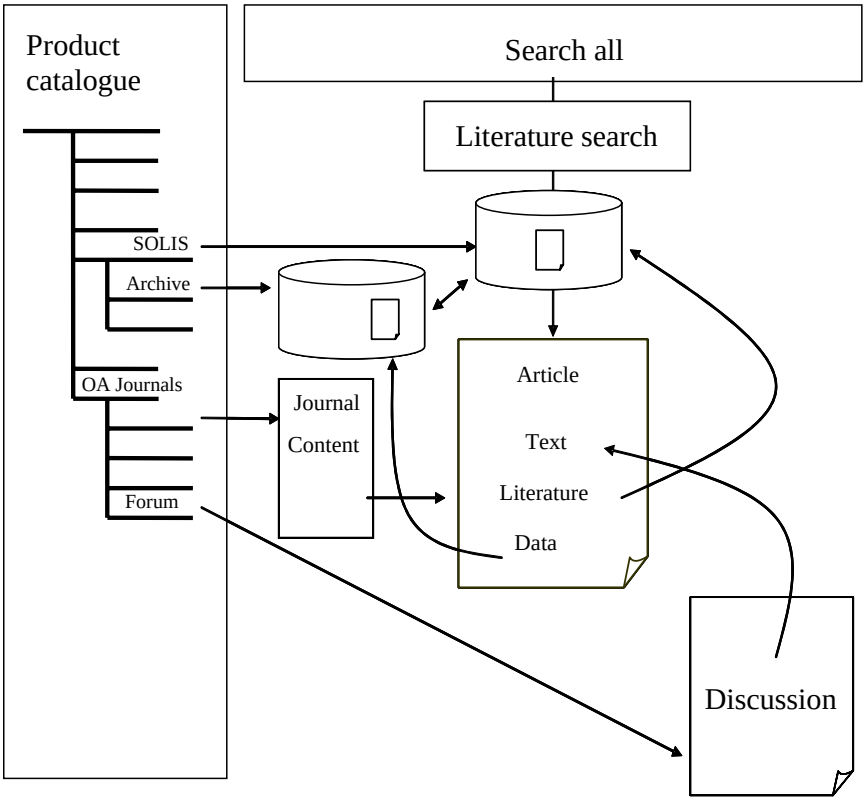
[7] http://www.soros.org/openaccess/ and

Along with this come advantages that connect electronic publishing directly with considerations on the improvement of inquiries in specialists portals like sowiport, using the new options of the search and networking. Most notably employed today are the integration of full-text searches in science portals and the cross-referencing of metadata in OPACs and specialized databases with full-texts, alternatively with borrowing services. Another obvious advantage is to network the literature details of publications directly with the full-texts when they are electronically available and automatically absorb the metadata of the electronic publications into the specialized databases.

Thus it stands to reason to broaden sowiport with a toolkit for electronic publishing, which enables social scientists to issue an electronic newspaper and publish on the web. The technical infrastructure must be set up so that issuers and authors are not further encumbered with more IT-skills requirements than the accustomed work e.g. with office software entails.

The following diagram shows the connection to the product catalogue of sowiport and to the query process. The latter connection's objective in turn is to steer the compiling of metadata so that it is automatically transferred into the specialized database and the full-text is immediately available via the list of results of a portal search.

At the same time, the diagram points out the beginning of a new problem: the combination of informal and formal communication that represents the core of the considerations in the following level 5 of sowiport.



Level 4: Integration Query + Electronic Publishing
+ Beginning of informal communication (from M. Stempfhuber, IZ Bonn)

In the previous realizations of electronic publishing, informal communication is only addressed insofar that readers of the electronic product have the option to directly comment on articles and these comments are in turn accessible to all readers. Thus an informal discussion of every article is set into motion.

3.2 Long-term Vision: New Quality of Scientific Working

The considerations up to now attach themselves to the currently existing and respectively operating trials for the realization of electronic publishing. They especially use the inherent advantageous properties of the new medium, in particular in the form of the WWW.

Scarcely modeled as yet, but implemented in contrast are innovative design possibilities that clearly exceed the above-described direct correlation to the concrete advantages of the new medium's properties.

Some authors expect that the IT-transformation and the use of the innovative possibilities connected to this transformation will radically change the scientific style of working, the manner in which one works scientifically (Nentwich, 2003). Scientific findings will result that would not have been possible in a traditional context.

Innovative starting-points for a continued development of publishing, which are attributed to current print and electronic forms of formal communication, are produced in particular from a remodeling and the shifting of borders between formal and informal communication in an increasingly international, interdisciplinary and, in relation to time, asynchronous communications realm.

(Cronin, 1982) represents the formal communication through the characteristics of public access and permanent storage (= level 1- 4 in chapter 2). Not just this type is fundamentally changing today. Even clearer are the changes and new possibilities in informal communication, which include – in their traditional form – personal networks and here in the most prominent position, the *invisible colleges*. The previously dominating forms of time and location-based synchronous personal networks are being supplemented today through virtual group networks (email lists, discussion groups, video conferencing, etc.) and even partially replaced. Generally, any scientist can partake in them. If information in science portals and virtual specialized libraries is collected and networked according to results, then this is about a direct, dynamic and interactive knowledge exchange that leads to a networking of the players themselves.

The conceptual and technical basis of this information type is the direct communication. It is asynchronous, has a high timeliness in the content of its information and relevance for research and accelerates the rate of scientific innovation (Nentwich, 2003). This is also about the tackling of the offering-induced overload.

Scientists award informal communication a high significance. The desire of the scientist to not only absorb the quasi-static information of science portals and virtual specialized libraries, but at the same time receive intelligent support in the access and participation of relevant networks – which open up new horizons – seems clear. Especially the association with publishing promises new presentation possibilities that reach beyond the direct implementation and utilization of the conduit qualities of the Web:

- Publications on the web allow supplementation, both in the creation phase as well as after completion, through intelligent communications components.

- The former area includes all interactive activities of publication software in the pre-phase, such as article submissions and the review process.
- Electronic publications can be associated with the subsequent discourse and public (web) commentary process, which may stimulate professional discussions (e.g. alternative review services).
- Utilization of the new communication tools enable a change or supplementation through new forms of referencing and quality control:
- Supplemental or alternative assessment through open peer commentary.
- The exchange might become an ex post review instead of the traditional ex ante quality review through masked referees.
- Users can valuate publications (with rating systems, see for instance www.amazon.com).
- The quality can be measured by the amount of use (simple realization of citation analyses, complex *user tracking*).
  It is still largely unsettled today, how processes of this type can be sensibly integrated in the science world. The variance potential seems high and should be case-tested.
- The consequential use of the network with the interactive potential of the Web leads to new forms of publication such as *living documents* (permanent updating, e.g. with research synopses) or *skywriting* (development of a new publication form from previous discussions via email, discussion lists, etc.), that replace or sensibly supplement the one-dimensional, uni-directional knowledge proliferation of traditional publication forms.

In the application of the varying dynamic communication components in combination with electronic publishing, the critical aspect is the permeability of the individual components and their connection to the static information collections of the science portals and virtual specialized libraries (specialized databases, OPACs, etc.). The interlocking of both information types requires a precise and intelligent coordination of all individual components.

4 Conclusion

A new type of infrastructure is emerging that is innovatively meeting the demands of technological change through the development of the web. The traditional models of library science and information science are being renewed. The components for the treatment of *semantic heterogeneity* are establishing an independent, theoretical alternative that is mediating the semantic-remote content indexing process of Google and the deep-rooted but complex world of ontology. In contrast to the semantic web and the Google world, the deep (invisible) web is the basis from where it is broadened through the information of the visible web (not vice versa) and the existing semantic knowledge of traditional thesauri work is reused.

In specialized portals like sowiport, it is not just about a broadening of the search area, about specialized field-overlapping searches (like www.infoconnex.de) and about the networking of national and international offerings: The combination of electronic publishing with offerings and an expansion of informal communication tools is just as significant. This expansion of the portal basic components contains the chance to achieving a new quality of scientific working. The historically-developed *artificial* differences between informal and formal information are being smoothed away[8].

The fulfilling of the objective is both political, organizational as well as technological and, in relation to the practical technological realization, difficult and only attainable in the long-term. Only a step-by-step process – independent of the data or information types or the institutional providers – will succeed in principally offering all the necessary information and tools qualitatively first-class, scientifically verified and user-friendly for scientists.

References

[1] Cronin, Blaise (1982). Progress in Documentation. Invisible Colleges and Information Transfer. *Journal of Documentation.* 38(3). 212 - 236.

[2] Fensel, Dieter (2004). *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. 2nd ed. Berlin, Heidelberg, New York: Springer.

[3] Hahn, Udo; Schulz, Stefan (2004). Building a very large Ontology from Medical Thesauri. Staab, Steffen; Studer, Rudi (Hrsg.) (2004). *Handbook on Ontologies*. Berlin, Heidelberg New York: Springer. p. 133 - 150.

[4] Krause, Jürgen (2001). How to Integrate Different Text Data and Fact Information: A Conceptual Transfer Problem in Digital Libraries and its Connection to Agent Theory. Smith, Michael J.; Salvendy, Gavriel; Harris, Don; Koubek, Richard J. (Hrsg.): Usability Evaluation and Interface Design, *Proceedings of the HCI International* 2001; Vol. 1. Mahwah: Erlbaum, 933 - 937.

[5] Krause, Jürgen (2004). Konkretes zur These, die Standardisierung von der Heterogenität her zu denken. *ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie*. 51(2). 76 - 89.

[8] Also in this problem area the adequate handling of semantic heterogeneity aspects will be crucial.

[6] Krause, Jürgen (2006). Shell Model, Semantic Web and Web Information Retrieval. Harms, Ilse; Luckhardt; Heinz-Dirk; Giessen, Hans W. (Hrsg.). *Information und Sprache*. Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern. Festschrift für Professor Dr. Harald H. Zimmermann. München: K. G. Saur.

[7] Krause, Jürgen; Niggemann, Elisabeth; Schwänzl, Roland (2003). Normierung und Standardisierung in sich verändernden Kontexten: Beispiel: Virtuelle Fachbibliotheken. *ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie*. 50(1). 19 - 28.

[8] Krause, Jürgen; Stempfhuber, Maximilian (2005). Nutzerseitige Integration sozialwissenschaftlicher Text- und Dateninformationen aus verteilten Quellen. König, Christian; Stahl, Matthias; Wiegand, Erich (Hrsg.). *Datenfusion und Datenintegration*. Bonn: IZ Tagungsberichte; Bd. 10 141 - 158.

[9] Marx, Matthias (2005): *Empirische Ergebnisse zur Evaluation semantischer Transformationen.* Bonn: IZ-Arbeitsbericht.

[10] Nentwich, Michael (2003). *Cyberscience. Research in the Age of the Internet*. Vienna: Austrian Academy of Sciences Press.

[11] Umstätter, Walther; Wagner-Döbler, Roland (2005). *Einführung in die Katalogkunde - Vom Zettelkatalog zur Suchmaschine*. Stuttgart: Hiersemann.

[12] Zhang, Xueying (2005). Concept integration of Document Databases using different Indexing Languages. *Information Processing & Management*. 42(1). 121 - 135.