ARGUMENTA OECONOMICA
No 1-2 (14) 2003
PL ISSN 1233-5835

Maria Mach[*]

INTEGRATING KNOWLEDGE FROM HETEROGENEOUS SOURCES

The paper is a survey of attempts to the problem of heterogeneous knowledge integration. The author shows the need for such integration in modem economic environment and in this context presents and comments both theoretical and practical solutions that can be met in the literature.

Keywords: knowledge integration, heterogeneity, unification, mediators

INTRODUCTION

The first aim of extracting and integrating information is to build synthesized, integrated descriptions of information coming from different sources. The second one is to provide the user with a uniform query interface, independent from the location of sources and the degree of their heterogeneity (Bergamaschi et al., 2001). Such problems appear in many systems, for example:

- temporal databases – sometimes there is a need for unifying information having different time granularity (Wiederhold et al. 1991),
- multi-agent systems – here we have to deal with the problem of integrating single agent's belief sources, as well as with the question of integrating all agents' beliefs (Wiederhold 1994; Maynard-Reid and Lehmann 2000),
- complex systems inferring on the basis of information coming from multiple sources (Subrahmanian 1994).

Till now, systems taking advantage of integrated knowledge have been used in practice in such areas as: application design (Calvanese et al. 1998), database systems (Li and Clifton, 2000), text integration (Fridman Noy and Hafner, 2000; Cohen, 2000), systems with hybrid knowledge base (Lu et al. 1996) and so on.

[*] Department of Artificial Intelligence Systems, Wrocław University of Economics

As the environment in which modern enterprises have to operate become more and more complex, so has the descriptions of this environment. And this in turn is strictly connected with the problem of heterogeneity. In order to perform coherent reasoning about possible environment features, it is necessary to unify descriptions that very often come in different forms and from different sources. In the literature there can be found so many attempts to solve the above problem, that it proves – in our opinion – the importance of the above question, not only in theoretical, but also in practical aspect.

The goal of this paper is to present the problem of knowledge sources heterogeneity and their integration, and to present selected approaches to the problem that can be found in the literature.

The paper is organized as follows: in Section 1 the notion of data sources heterogeneity and the genesis of the integration problem is presented, Section 2 presents the types of integration of knowledge coming from different sources, Section 3, divided into 2 sub-sections, contains a survey on the solutions concerning knowledge integration – both theoretical and practical ones. The last part of the article is devoted to conclusions and summary.

1. THE NOTION OF INTEGRATION. THE GENESIS OF THE INTEGRATION PROBLEM

As was mentioned in the introductory section, the notion of knowledge from heterogeneous sources came up when problems being solved by intelligent systems, as well as the environment of those systems became so complex, that knowledge gathered from only one source (e.g. from an expert) was not sufficient any more. The widely understood knowledge based systems started to use more and more sources of knowledge and this in turn triggered off the need for unifying information from those sources, to make further coherent reasoning possible. Knowledge unification, or integration, is understood differently by different authors, this question will be discussed in the next section.

In this section we would like to pay attention to the question, how heterogeneity (diversity) of knowledge sources is perceived.

Wiederhold et al.(1991) perceive knowledge sources heterogeneity in the context of time granularity differences. They discuss the problem of unifying temporal information coming from temporal databases when each of them is based on a different time model. Therefore the question of heterogeneity is connected here not only with time granularity, but also with heterogeneity of temporal representation formalisms.

The problem of intelligent systems is discussed by Wiederhold in (Wiederhold, 1994). Leaving behind the question of granularity, Wiederhold pays more attention to the ontology of knowledge domains and links it with the problem of heterogeneity. If there is a need to use knowledge from different domains, it may turn out that these domains differ in ontology, that is the nature and structure of reality depicted. This may have different reasons, as (Wiederhold, 1994):

- different types of attribute naming,
- differences in scope (domains covering different areas),
- differences in attribute coding,
- subjectivity of attributes meaning their scope.

Ontology unifying, or creating a common ontology for knowledge from different domains, would enable, for example, program agents' cooperation while minimizing the risk of misunderstandings.

A rather conventional and intuitive approach to heterogeneity is presented by Subrahmanian in (Subrahmanian, 1994). He simply assumes that knowledge sources heterogeneity comes into play when complex reasoning tasks require the use of information from several different sources, such as for example databases, knowledge bases, sensors, monitoring devices etc. It is obvious that each of those sources describes reality in a different way.

The already cited G. Wiederhold in his later works, e.g. (Wiederhold, 1999) goes on a higher level of abstraction and makes a difference between interoperating on information and integration of data. The basis of this difference is the subject of merging: if we merge different knowledge sources in one intelligent system, it is called integration, but if we merge only selected outcomes from those sources, it is called information interoperating. But it seems a problem appears here. It is obvious, that in the era of globalization and multinational corporations, making millions of different operations, it is almost impossible to build a single integrated system that would encompass all possible information sources. Therefore it is necessary – before the integration stage – to make a selection of outcomes from different sources. According to the terminology proposed by Wiederhold, this will be the case of interoperating. But is it purposeful to complicate terminology in this situation? It seems that it is better to use the notion of information integration, irrespective of whether it is tentatively selected or not, because the question of unification for further reasoning remains open.

It seems that opinions by different authors, cited above, make the problem of heterogeneous sources clear and there is no need to cite more opinions. The most important thing is that – as this short survey showed – the question of heterogeneity of sources of knowledge used in reasoning is crucial and will gain even more importance as the complexity of problems solved and the complexity of intelligent systems environment grows.

2. THE TYPES OF INTEGRATING KNOWLEDGE COMING FROM HETEROGENEOUS SOURCES

As was mentioned earlier, the notion of "knowledge integration" is understood differently by different authors. Therefore it seems purposeful to present types of integration, mentioned by diverse authors, to order them.

The understanding of integration notion is strictly connected with the context in which integration is considered. In the literature one can find such contexts, as: program (linked with the question of system construction), programmer, database, text integration, agents (belief revision).

The context called "program" we understand as a practical one, connected with system building. In this context Subrahmanian et al. (1997) mean domain and semantic integration. They link both types of integration to a so-called mediator system, the concept of which comes from work by Wiederhold (see e.g. Wiederhold et al. 1991; Wiederhold 1994; Wiederhold 1999), and by Nerode and Subrahmanian (Lu et al., 1996). And so, domain integration means, according to Subrahmanian et al., adding new data sources or reasoning systems to an already existing mediator system, in a way that resources of that new source/system added could be accessible for different mediators. By semantic integration the same authors understand a process of solving conflicts that appear during information pooling, during defining new, complex operations, based on operations possible in individual data sources that are integrated.

Calvanese et al. (1998) place the problem of integration in a similar context, that could conventionally be described as a strictly "programmer" one. They address their work to designing and maintaining applications that require integrating information from different sources, therefore they consider mainly data integration, which, according to them, can be virtual or materialised. We deal with virtual integration when an integrating system operates as an interface between user and data sources. Such a kind of integration is typical for open systems. Materialised integration, in turn, takes place when an integrating system maintains a replicated picture of data in sources. This kind of integration is typical for data warehouses.

Taking into account a subject towards which integration is directed (a context which could be called directional or target) the same authors speak of integration directed towards data sources and integration directed towards a client. The first one takes place when a new source (or its part) has to be taken into account, the second one – when a new query or a set of queries from the client application appears. It must be mentioned here that both kinds of integration can be at the same time virtual or materialised.

Also Li et al. (2000) discusses a semantic integration, but – contrary to (Subrahmanian et al., 1997) where the authors deal with a "programmer" context – Li and Clifton link semantic integration with databases, therefore the context is narrower here. And in this narrower context Li and Clifton understand semantic integration as identifying relationships between attributes or classes in different database objects. So in their opinion semantic integration concerns different database schemes merging.

The next two works that are worth mentioning here deal with integration in context of text integration. But, if Fridman Noy and Hafner (2000) link knowledge integration directly to the context of processing information in the form of electronic texts, Cohen (2000) treats data integration somewhat *per analogiam* as distributed text collections: according to Cohen, data integration differs from distributed text collections integration only in that sources being integrated are structured, while texts are not.

And last but not least the context that can be conventionally called an agent. Here come into play such questions as: aggregating agent's beliefs (coming from many sources), integration of information possessed by many agents, belief revision and update. In this context integration of data from heterogeneous sources is presented e.g. by Maynard-Reid and Lehmann (2000) and Liberatore et al. (2000).

Maynard-Reid and Lehmann (2000) deal with the problem of constructing the agent's state of belief, while that agent is informed by a collection of sources with a different degree of credibility, and with the problem of merging information from different agents. So in their opinion integration is identical with aggregation of information from different sources, that in addition have different degrees of credibility. Surely it is a very specific approach, because of the context. But it seems that such understanding of integration narrows this notion and only after making such an assumption is one entitled to use the notion of "integrating data from heterogeneous sources".

A similar approach is presented by Liberatore and Schaerf (2000). They begin with dividing integration into different categories, and speak of:

- belief revision – this is an integration of two information portions (fragments), one of which is considered 100% credible, while another can be partially wrong,
- actualisation – this is an integration of two information portions, both fully correct, but each of these portions concerns a different time point,
- merging – integrating two or more information portions with the same degree of credibility.

As it can be easily seen, there are many approaches and contexts, in which knowledge integration can be spoken of. Nevertheless, in all contexts and aspects the main problem remains the same: how to create a consistent description of information from heterogeneous sources and how to help a user to work with such a description.

3. SURVEY OF SOLUTIONS CONCERNING INTEGRATION OF HETEROGENEOUS SOURCES KNOWLEDGE

Even a cursory look at the literature allows to see that there are many proposals concerning solutions of the heterogeneous source knowledge integration problem. The reason for this seems to be simple – as it was shown above, the problems of knowledge integration appear in many aspects, are linked with many intelligent systems applications, and therefore these problems are very important. In further parts of this section there will be shown a diversity of practical domains, in which intelligent systems making use of knowledge integration mechanisms can be applied.

There are so many proposals of solutions in the literature, that it seems useful to classify them in a way, so to make the whole question clearer. In our opinion the simplest and the most intuitive classification consists of dividing integration solutions into theoretical and practical ones. Of course the criterion of this classification is based on the question, whether a given solution was implemented in a working system or not. Therefore we assumed that solutions that were implemented in a system, even if that system was not later put into practice, then such solutions we will call practical. On the other hand, solutions that were not implemented in any system will be called theoretical.

It is obvious that – as in the case of discussion on integration aspects or heterogeneity understanding – it is impossible to present all existing solutions. Therefore there was necessity to make a rather arbitrary choice, which nevertheless, in our opinion, shows well the abundance of approaches that can be found in the literature. Both in the theoretical and practical group, the solutions are presented chronologically, from the oldest to the newest.

3.1. Theoretical proposals

Our survey on theoretical proposals will begin with a solution by Wiederhold, Jajodia and Litwin (1991). They discuss the problem of unifying temporal information from temporal bases with a different time granularity. Generally speaking, they divide the processing of temporal data into three stages:

a) collecting new data,

b) converting events into histories – here we deal with unification,

c) searching for useful information *via* queries.

Conversion of events into histories is a critical stage. The cited authors suggest a special history operator, *H*, which allows to specify each transformation. They introduce also a second operator – *I*, which gives information on an object in a given time point, therefore it is the most often used to get current information. A detailed formalization of the proposal can be found in (Wiederhold et al., 1991).

Gio Wiederhold is an author of a mediator concept, that was later used in many works. A mediator is a "device" which specifies how the intended integration is to be performed (Wiederhold, 1992; Wiederhold, 1993). Wiederhold is also the author of the next theoretical proposal – multidomain algebra (Wiederhold, 1994), which would allow to create systems encompassing many domains, where domain is understood as an area of science or computer program products having a common ontology (Gruber, 1993). Unifying information from many domains would allow for example a co-operation of many agents, while minimizing the risk of misunderstandings.

Next, Subrahmanian (1994) proposes an amalgamation logic for integrating knowledge from heterogeneous sources. The logic is based on a group of logics that are an extension of logic programming, in which atoms are marked explicitly with values that can be perceived as confidence coefficients, degrees of certainty etc. (Adali et al., 1995). This group of logics, so-called annotated logics, was also introduced in the late eighties by Subrahmanian, and was intended to constitute logical framework for deductive databases containing incoherent, conflicting or contradictory information. These logics have no algebraic semantics (Bowers et al., 2000).

Generally speaking, knowledge bases amalgamation means that there exist some local knowledge bases and a superior (meta)knowledge base (in some works this is called a mediator) which defines, among others, in what way the local knowledge bases are to be merged. The metabase must be expressed in a language that allows for reasoning about local bases and for their manipulation. An integrated base that is a result of the above operations, is called an amalgam. The solution proposed by Subrahmanian has the following features:

- a user can work directly with an amalgam, and at the same time formulate queries concerning bases that are components of the amalgam,
- it is possible to examine relationships between semantics of local knowledge bases.

As can be therefore seen, this solution allows to merge different knowledge bases and data structures (e.g. relational, object, spatial and temporal ones).

Adali et al.(1995) refer to the concept of mediators proposed in works by Wiederhold, cited above. They treat mediator as a program written in a special language, operating on information from different sources. Usually these sources are application and program packages already existing. Adali and Emery propose MPE – Mediatory Programming Environment and understand it as an "interpreter" that executes programs written in a mediating language and communicates with external programs. A general structure of MPE is shown in Figure 1.

The theoretical framework described above was used practically in the HERMES system, described by Subrahmanian et al. (1997), which will be discussed in the next section.

Konieczny et al. (1998) an axiomatic characteristics of merging operators and a logic of merging propose. The authors do not propose one specific method of information merging, they try rather to define the characteristics of such methods that show what postulates should be met by a merging method. They also propose a set of postulates for a merging operator if it is a (so-called by them) rational one. This approach is further discussed and developed by Konieczny (2000).

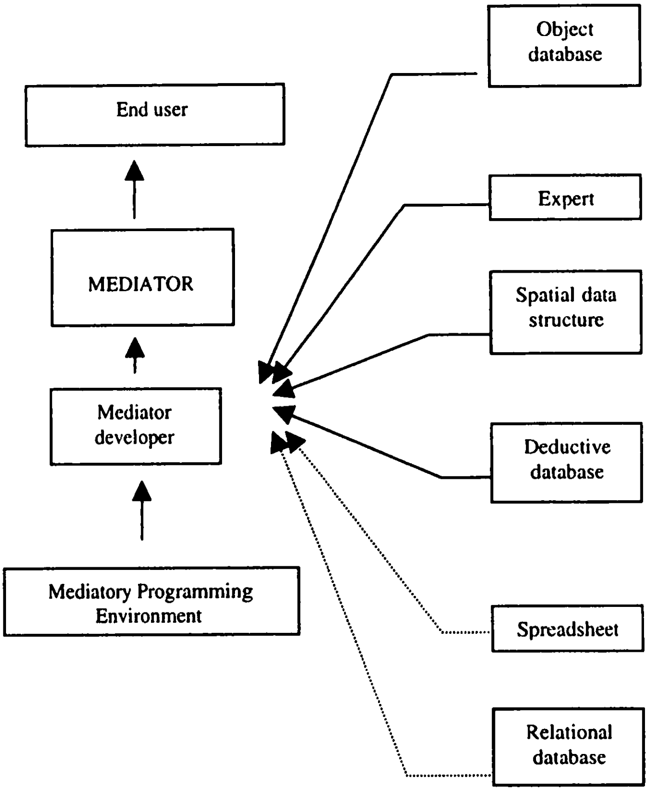

Fig. 1 MPE concept
Source: Adali et al., 1995

Hsu et al. (2000) propose a concept that is somewhat "competitive" to Wiederhold's concept of mediators. Their proposal is called a semantic query optimization. The goal of such an optimization is to optimize query plans, both global and local (local means optimizing queries searching for data in each individual database source). Query optimization is divided into two main stages:

- first, an optimizer locates relevant semantic knowledge and on the basis of it proposes a sequence of one or many query reformulation operations, while retaining query's semantics,

- during the second stage the proposed reformulations are evaluated and the best query plan (according to cost criterion) is performed.

The above mentioned "competitiveness" of this proposal in relation to mediators concept by Wiederhold lies in that the former contains a possibility of reducing costs linked with processing queries generated by mediators, while Wiederhold did not take this question into account.

Maynard-Reid et al. (2000) address a very interesting aspect of integrating knowledge from heterogeneous sources. They discuss multiagent environment, in which agents are informed by different sources. In this context, they address the following questions:

a) How to represent agents' common beliefs?

b) How to construct agent's belief state, having to aggregate information from sources with a different degree of reliability?

c) How to merge information given iteratively by many agents?

According to those authors, there are several types of sources aggregation:

a) aggregation of sources having equal ranks - this may lead to conflicts because intuitive treatment of all sources as equally important is justified,

b) aggregation of strictly ordered sources – sources that are higher in hierarchy by reliability replace sources with lower rank. We use the latter only if in a given situation "higher" sources are indifferent (neutral),

c) general aggregation (general case) – if there are several ranks and many sources having those ranks. In this case Maynard-Reid and Lehmann propose an aggregation operator, which qualifies the set of source beliefs, before reasoning about new beliefs is performed.

The above questions concerned one agent case, where the agent has to "form an opinion" on the basis of heterogeneous information sources. There is also a case of multiagent fusion, that is a case of aggregating belief states of many agents, while each of the agents has his own set of information sources. Maynard-Reid et al. (2000) address the question, whether it is possible to calculate a state of beliefs resulting from agents' fusion only on the basis of their initial belief states, therefore not taking into account the sources of individual agent's beliefs. Such a calculation would be useful because of the cost of storing and transmitting all states of source beliefs. This is possible if all sources have equal ranks, but it is a very rare case. If in turn sources have different ranks, but are totally pre-ordered according to reliability, it is not necessary to store all of them; for each opinion it is sufficient to store the source with the highest rank.

An interesting approach to knowledge integration can be found in Olszak et al. (2003). They discuss the problem in context of business intelligence (BI) systems. BI systems are designed to make use of many different kinds of data and knowledge. As knowledge in an enterprise originates from many different, and therefore heterogeneous sources (such as information systems, internal documentation, corporate databases, the web etc.), the need to integrate this knowledge is obvious. Olszak et al. propose an integrated approach to build and implement BI systems. They distinguish four basic dimensions of such a system and propose to take them into account during the design and implementation of the system. Their approach is summarized in Figure 2. The approach may be considered interesting, as the previous approaches were not linked with the BI technology.

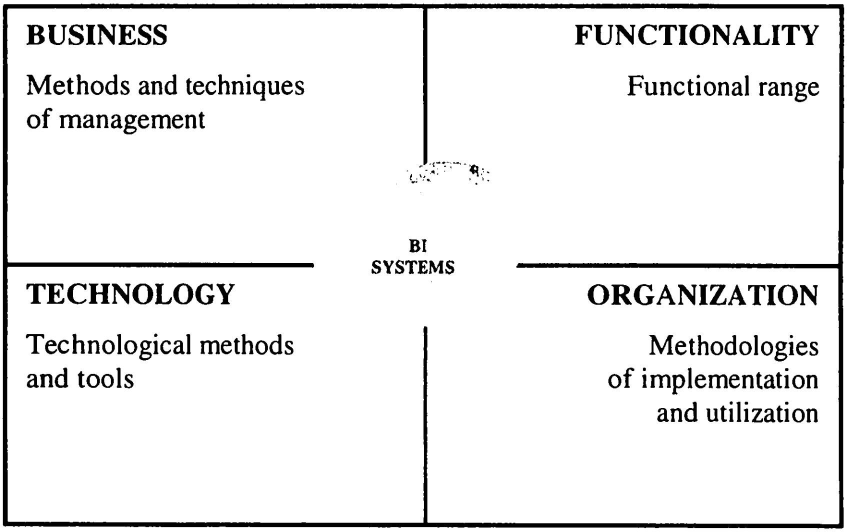
BUSINESS
Methods and techniques of management
FUNCTIONALITY
Functional range
BI SYSTEMS
TECHNOLOGY
Technological methods and tools
ORGANIZATION
Methodologies of implementation and utilization

Fig.2 Integrated approach to build and implement BI systems
Source: Olszak et al. 2003

The next theoretical proposal that we would like to present – by Dudycz and Sierocki (2003) – is also connected with the context of the BI systems. Again, the authors point out the diversity of knowledge types in an enterprise. They propose to distinguish a new class of BI systems, and they call it AIAS – advanced information-analysis systems. According to Dudycz et al. (2003), AIAS systems are understood as a broad collection of applications and technologies, connected together, that enables collecting, merging, selecting, analysis of knowledge from heterogeneous sources, as well as a comprehensible presentation of this knowledge. They discuss the basic features that an AIAS system should possess and the functionality of such systems. Unfortunately, the discussion is rather general, they do not present their own concept of an AIAS system architecture, instead they present the already existing solutions that in their opinion may be considered AIAS systems.

And the last theoretical proposal to be presented here – the one by Bonifacio and Molani (2003). We have chosen it because we consider it controversial. Why? Simply because Bonifacio and Molani claim that there is no need for integration, on the contrary, heterogeneity of knowledge sources should be preserved. In their opinion, the different, alternative "knowledges" of an enterprise constitute a so-called cognitive source that may enable to perceive the economical environment in many perspectives. And this in consequence allows the enterprise to better adapt to changing circumstances. Therefore, the heterogeneity of knowledge sources constitutes an opportunity, not a limit and there is no need to integrate the knowledge.

We cannot agree with such a concept. In our opinion the heterogeneous sources of corporate knowledge are very valuable, nevertheless there is also a need to integrate them. Each knowledge source may be – and has to be – used separately, but all the sources integrated together may create a new knowledge, even more valuable thanks to the synergy effect.

Summing up the above survey on theoretical solutions it must be said that – regardless of approaches and contexts diversity – it was the concept of mediator proposed by Wiederhold which had the maximum influence on other authors. This concept and its varieties are the most often seen in the literature.

3.2. Practical solutions

Before we start to present some concrete solutions, it must be first said that all of them are depicted only roughly, to let the reader make an opinion on them. The details on each solution can be found in the literature cited.

Our survey on practical proposals will start with the HERMES system, described by Subrahmanian et al. (1997). It is a system in which the concept of mediators (already presented above) is used, a system based on a HKB (Hybrid Knowledge Bases) theory by Nerode and Subrahmanian (Lu et al., 1996). HERMES allows for the gradual integration of new systems with the already existing mediating system. Versions of HERMES for PC and DUN/Unix platforms were developed. HERMES integrates the following types of sources (Subrahmanian et al., 1997):

- relational data of different formats, encoded in text ASCII files,
- relational databases Paradox and Dbase V,
- spatial data,
- rough text data,
- pictorial data (GIF format files).

Topographical and engineering centre of the American Army implemented, on the basis of the HERMES system, a route planning tool, that searches for the optimal cheapest path between two points. HERMES was also used to build a face recognizing tool.

Wiederhold (1999) proposes a system with an architecture also based on the mediator concept. The architecture can be discussed in two dimensions: horizontal and vertical ones. In the horizontal dimension there are three system layers: client application, mediating service modules and base servers. The vertical layer of the mediator, in turn, is divided into many domains. Of course this vertical division of the mediating layer is done on the basis of expert domain knowledge.

The above discussed architecture of a system with a mediator is presented in Figure 3:

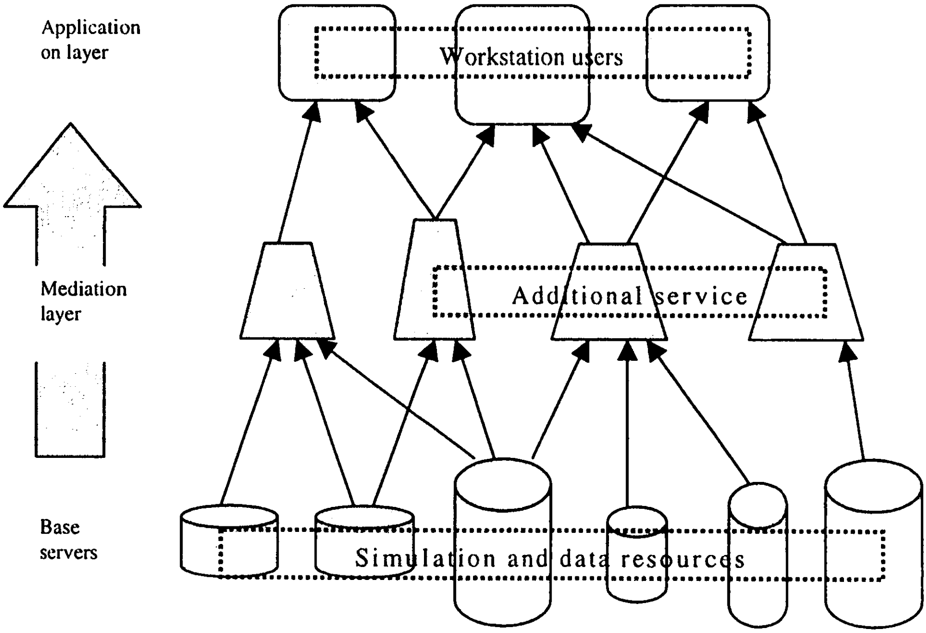

Fig. 3 An architecture of a system with mediator
Source: Wiederhold, 1999

The next practical solution to be presented is the M-LaSIE-II system by Azzam et al. (1999) for multilingual information extraction. The basis for this system's construction was an assumption that it is possible to develop a representation of notions important for a given domain, that would be independent from any language. As it is easy to guess, M-LaSIE-II system performs mainly semantic integration.

Palopoli et al. (2000) also propose semantic integration, but this time it concerns database schemes. The goal of this integration is to create a global notion scheme on the basis of heterogeneous initial schemes. This kind of integration can be termed as semantic, because the algorithms used for it take into account object contexts, their semantic relevance and they examine semantic relationships between scheme objects. The authors of the work cited tested their solution – or more precisely, both algorithms of semantic integration – in Italian central administration offices, now they are working on implementing the algorithms in a more general system, called DIKE.

The next example of a practical solution in which we meet semantic integration, is the SEMINT system, depicted with details by Li et al. (2000). Integration is understood here similarly to the previous example, that is as identifying relationships between attributes or classes in different database schemes. The SEMINT system is based on neural networks. Its authors distinguish three levels of metadata that can be automatically extracted from databases:

- attribute names (vocabulary level),
- scheme information (field specification level),
- data contents and statistics (data contents level).

Neural networks are used for system training, during which the system "learns" how the metadata characterize attribute semantics in a concrete domain.

The main task of the SEMINT system is to help the database administrator in finding corresponding attributes in heterogeneous databases of large organizations.

The goal of research presented by Craven et al. (2000) was automatic creation – on the basis of information from the Internet – a knowledge base "understandable" for a computer. Such a base would contain assertions in a symbolic form. To create this base the authors of the work cited propose to use machine learning methods, thanks to which it would be possible to create methods of information extraction for each of the types of knowledge desired. The project, known as "Web-KB", encompasses learning algorithms of 1st order for classifying web sites and for identifying relationships between them. As the authors of the project claim, their approach can be practically used to:

- improve the process of information search on the Internet,
- use the Web as the aid for knowledge based reasoning and for problem solving,
- aid knowledge based intelligent agents.

SYNDIKATE system, presented by Hahn et al. (2000) is devoted to solving the same problem as the Web-KB project. More precisely, it is a whole family of systems which execute the task of understanding natural language texts, acquiring from them knowledge in the form of facts, complex sentences and evaluating propositions, and finally moving this knowledge to formal representation structures, that is to a text knowledge base.

Cohen (2000) addresses the problem of heterogeneous databases that do not share common object identifiers. He points out that integration of data from heterogeneous bases is a problem analogical to the one of integrating distributed text collections, it only differs in that database sources are structured. As a solution to the problem Cohen proposes WHIRL – a database management system, that allows to integrate – by queries – data on web pages, that is to integrate text data. Finding an answer to a query is treated in WHIRL as an optimization problem, meaning that query processing is perceived as data space search.

An environment that has heterogeneity as a typical feature is software development environment. One should even introduce "environments" as they encompass diverse tools, user interfaces, data repositories. The Chimera system, presented by Anderson et al. (2000), is an open hypermedia system. Its main task is to aid at software development in a heterogeneous environment, by already extending environments with hypermedia services (a combination of hypertext and multimedia techniques) without a need of modifying already existing clients, objects or repositories. The system makes use of a client-server architecture. The first prototype of Chimera was constructed in 1992 (Anderson et al., 2000, p. 226). The system was tested in the army, where it was used to develop aviation software.

Liberatore et al. (2000) address the problem of integrating knowledge from different knowledge bases. They distinguish three main conceptual approaches to the problem:

- belief revision,
- merging,
- update.

The BReLS system presented by them constitutes a framework in which the three approaches meet, and thanks to this it is possible to formalize complex domains in which information can be of different certainty and can appear in different time points.

Knowledge in the BReLS system is expressed in the form of proposition formulas, completed with positive integers denoting a degree of certainty. Time is expressed by a syntax borrowed from Sandewall (1994). There are two kinds of models in the system: static and dynamic ones. With the aid of BReLS framework it is possible to link belief revision, knowledge fragments merging and knowledge update together.

Domenig et al. (2000) present their SINGAPORE system (SINGle Access Point for heterogeneous data REpositories). Data integration in this system is performed via a unified interface, without affecting data sources. Users formulate queries to data sources repository in a global query language (the language is described in the work cited). The SINGAPORE system has a three-layer architecture, consisting of: user layer, mediator layer, and sources layer. It must be said here that in our opinion this solution has one very important practical disadvantage: user formulating a query has not only to be familiar with the relevant query language, but also has to possess knowledge e.g. on types of conflicts in heterogeneous bases.

The next practical solution that is in our opinion worthy mentioning in this short survey, namely MOMIS system, described by Bergamaschi et al. (2001). The task of this system is to integrate information extracted from structured and semi-structured sources, to build synthesized, integrated descriptions of information from different sources, and to provide a user with a unified query interface, independent from source location and the degree of heterogeneity of sources. The MOMIS system, like HERMES (Subrahmanian et al., 1997) and Wiederhold's system (Wiederhold, 1999) described above, is based on the concept of mediators. It performs integration on the semantic level.

Another problem connected with knowledge integration concerns verification of unified knowledge. This question is addressed by Ochmańska and Owoc (2001) and by Owoc (2001).

In the first of the works cited, the authors deal with the problem of classifying heterogeneous knowledge that may be found in a knowledge base, the problem of appropriate classification criteria, and the problem of finding a universal approach to the verification of knowledge bases containing knowledge in different forms. They experiment with the PROLOGA tool, which they consider universal with respect to different knowledge types, and they claim that using PROLOGA makes heterogeneous knowledge base consistent.

In the second work again the question of verifying heterogeneous knowledge bases is addressed. Three tools are tested in context of this task, namely PC-Shell, Kappa and again PROLOGA. The author distinguishes two approaches to heterogeneous knowledge verification:

a) verification before knowledge is transformed (integrated),

b) verification after integration (where integration is understood as unifying knowledge by using decision tables).

He proposes his own approach which could be called a mixed one, as it combines the above two approaches. We agree with the author, that such a mixed approach is much more powerful and effective. Unfortunately, the concept of the mixed approach is only roughly sketched. We are convinced that a final form of the concept may be very interesting and worthy of attention.

Now let us recall the "Web-KB" project discussed earlier as the last of the four practical proposals that we would like to present, also concern integrating knowledge gathered from the web.

Abramowicz and Kalczyński (2001) present the concept of automatic building collections of documents filtered from the web. The aim of this process is to build organizational data warehouses. They experiment with their HyperSDI batch filtering system. The main features of HyperSDI are as follows:

a) it allows for pre-filtering of web documents,

b) it enables automatic filtering,

c) it enables establishing a non volatile collection of documents filtered from the web.

Thanks to the HyperSDI system, a data warehouse is created in which a semantic linking of structured and unstructured content is possible, therefore we have here another example of semantic integration.

Vetulani (2002) addresses the question of getting information from the web in an user-friendly way. He discusses a practical technology called Question&Answering (Q&A for short). The Q&A technology is based on integrating different techniques for text understanding, information searching and information retrieval in the artificial intelligence context. The technology is still under construction and is intended to cope with:

- information sources heterogeneity,
- different data formats collecting,
- merging information from sources having different degrees of credibility.

The next project linked with the question of web resources collecting is called Hyperguide and is described by Paci and Canali (2002). The Hyperguide is an interactive push XML application for digital collection access, and was designed specially for web resources. Its main aim is to facilitate identification of selected web sites of a heterogeneous nature, to identify certain web sites, and to describe their information contents in a dynamic framework. The project is still under construction, the authors plan to transform it to a completely developed tool called TOOL2KNOW. Therefore we can only indicate here an interesting research direction that can be found in the literature.

Finally, Dreher (2003) proposes a method that enables to structure and access explicit knowledge, that in turn facilitates finding, accessing and structuring knowledge from the world wide web. The method is called the "Dreher Hypertext Development Methodology". It concerns only different textual forms of knowledge. It is platform independent and – in our opinion – it enables and facilitates the knowledge management process. The detailed algorithm can be found in the work cited.

The first conclusion from the above survey is that in practice, semantic integration is the most widely used and most popular. It results from the fact that – assuming that the main goal of integration is to create coherent descriptions of information from heterogeneous sources, and to make possible reasoning on the basis of such sources – it is obvious that semantic and conceptual unification, unification of notions is absolutely necessary and this is linked with semantic integration.

The next conclusion that comes to mind: there are many more practical solutions than theoretical ones. The reasons for this are obvious: as was mentioned in the introduction, the environments and domains of modern intelligent systems are so complex, contain so many different sources of information, that the integration of those sources becomes simply an essential step in system development. If we omit these sources of information, and therefore if we do not integrate information from them to make further reasoning possible, the intelligent system will not be up to the challenge of modern economic environments and enterprise needs.

CONCLUSIONS

The paper was devoted to the question of integrating knowledge from heterogeneous sources. We discussed such aspects of the problem as: the notion of knowledge sources heterogeneity, the types of integration, the approaches to knowledge integration that can be found in the literature and the solutions of the integration problem, both in theoretical and practical aspects.

The abundance of approaches to the problem of knowledge integration indicates the importance of the problem. In the paper we mentioned the reasons for which the question of heterogeneous sources integration is of crucial importance to modern enterprises. Here it is worth mentioning in short other domains, in which heterogeneous systems integrating different types of knowledge are used (for further details see Wiederhold, 1999):

- military applications (route optimization),
- state administration,
- large, heterogeneous databases administration,
- searching for data in the Internet,
- computer programs development,
- geographical systems.

It is obvious that as the world around us is more and more complex, there will appear more and more tasks requiring a coherent use of knowledge from different sources in a way to enable further reasoning. Therefore the role and importance of systems in which knowledge integration takes place will have a growing tendency.

REFERENCES

Abramowicz W., Kalczyński P. J., *On Supplying the Data Warehouse with Unstructured Contents Filtered from the Internet.* In: Baborski A. J., Bonner R. F., Owoc M. L. (Eds.), *Knowledge Acquisition and Distributed Learning for Resolving Managerial Issues.* Mälardalen University Press, 2001, pp. 133-144.

Adali S., Emery R., *A Uniform Framework for Integrating Knowledge in Heterogeneous Knowledge Systems.* Proc. of the Eleventh IEEE International Conference on Data Engineering, March 1995, pp. 513-520.

Anderson K. M., Taylor R. N., Whitehead E. J., *Chimera: Hypermedia for Heterogeneous Software Development Environments.* „ACM Transactions on Information Systems", vol. 18, no 3, July 2000, pp. 211-245.

Azzam S., Humphreys K., Gaizauskas R., Wilks Y., *Using a language independent domain model for multilingual information extraction.* „Applied Artificial Intelligence" vol. 13 no 7, October-November 1999, pp. 705-724.

Bergamaschi S., Castano S., Vincini M., Beneventano D., *Semantic integration of heterogeneous information sources.* „Data & Knowledge Engineering" vol. 36 no 3, March 2001, pp. 215-249.

Bonifacio M., Molani A., *The Richness of Diversity in Knowledge Creation: An Interdisciplinary Overview.* "Journal of Universal Computer Science", vol. 9, no 6 (2003), pp. 491-500.

Bowers S. E., Lewin R. A., Pigozzi D., *An Annotated Logic Defined by a Matrix.* http://www.mat.puc.cl/~rlewin/papers/paper_renato.pdf. 18 May 2000.

Calvanese D., Giacomo De, G., Lenzerini M., Nardi D., Rosati R., *Description Logic Framework for Information Integration.* Proc. KR-98: Sixth International Conference on Principles of Knowledge Representation and Reasoning. Morgan Kaufmann Publishers, Inc., 1998, pp. 2-13.

Cohen W. W., *Data Integration Using Similarity Joins and a Word-Based Information Representation Language.* „ACM Transactions on Information Systems", vol. 18 no 3, July 2000, pp. 288-321.

Craven M., DiPasquo D., Freitag D., McCallum A., Mitchell T., Nigam K., Slattery S., *Learning to construct knowledge bases from the World Wide Web.* „Artificial Intelligence" vol. 118, no1-2, April 2000, pp. 69-113.

Domenig R., Dittrich K. R., *A query based approach for integrating heterogeneous data sources.* Proc. CIKM-2000: Ninth International Conference on Information Knowledge Management, November 6-11, 2000, McLean, USA. ACM Press, 2000, pp. 453-460.

Dreher H., *Hypertext and Managing Knowledge.* Proc. Informing Science + IT Conference, June 24-27, 2003, Pori, Finland, pp. 27-33. ISSN 1535-07-03

Dudycz H., Sierocki R., *Przegląd funkcjonalności zaawansowanych systemów informacyjno-analitycznych [Survey of functionalities of advanced information and analytical systems].* In: Nycz M., Owoc M. L. (Eds.), *Pozyskiwanie wiedzy i zarządzanie wiedzą [Knowledge acquisition and knowledge management]* Prace Naukowe AE Wrocław nr 975, Wydawnictwo AE, Wrocław 2003, pp. 89-99.

Fridman Noy N., Hafner C. D., *Ontological foundations for experimental science knowledge bases.* „Applied Artificial Intelligence", vol. 14 no 6, July 2000, pp. 565-618.

Gruber T. R., *A Translation Approach for Portable Ontology Specifications.* "Knowledge Acquisition" vol. 5 no 2, 1993, pp. 199-220.

Hahn U., Romacker M., *Content management in the SYNDIKATE system – How technical documents are automatically transformed to text knowledge bases.* „Data & Knowledge Engineering" vol. 35 no 2, November 2000, pp. 137-159.

Hsu Ch.-N., Knoblock C. A., *Semantic Query Optimization for Query Plans of Heterogeneous Multidatabase Systems.* „IEEE Transactions on Knowledge and Data Engineering", vol. 12 no 6, November/December 2000, pp. 959-978.

Konieczny S., *On the Difference between Merging Knowledge Bases and Combining them".* Proc. KR-2000: Seventh International Conference on Principles of Knowledge Representation and Reasoning, April 12-15, 2000, USA, Morgan Kaufmann Publishers, Inc., 2000, pp. 135-144.

Konieczny S., Pino-Pérez R., *On the logic of merging.* Proc. KR-98: Sixth International Conference on Principles of Knowledge Representation and Reasoning. Morgan Kaufmann Publishers, Inc., 1998, pp. 488-498.

Li W.-S., Clifton Ch., *SEMINT: A tool for identifying attribute correspondences in heterogeneous databases using neural networks.* „Data & Knowledge Engineering" vol. 33 no 1, April 2000, pp. 49-84.

Liberatore P., Schaerf M., *BReLS: A System for the Integration of Knowledge Bases.* Proc. KR-2000: 7th International Conference Principles of Knowledge Representation and Reasoning. Morgan Kaufmann Publishers Inc., 2000, pp. 145-152.

Lu J., Nerode A., Subrahmanian V. S., *Hybrid Knowledge Bases.* „IEEE Transactions on Knowledge and Data Engineering", vol. 8 no 5, October 1996, pp. 773-785

Maynard-Reid II, P., Lehmann D., *Representing and Aggregating Conflicting Beliefs.* Proc. KR-2000: Seventh International Conference on Principles of Knowledge Representation and Reasoning, April 12-15, 2000, USA, Morgan Kaufmann Publishers, Inc., 2000, pp. 153-164.

Ochmańska M., Owoc M. L., *Verification of Different Knowledge Bases.* In: Baborski A. J., Bonner R. F., Owoc M. L. (Eds.), *Knowledge Acquisition and Distributed Learning for Resolving Managerial Issues.* Mälardalen University Press, 2001, pp. 85-97.

Olszak C. M., Ziemba E., *Business Intelligence as a Key to Management of an Enterprise,* Proc. Informing Science + IT Conference, June 24-27, 2003, Pori, Finland, pp. 855-863. ISSN 1535-07-03

Owoc M. L., *Podejścia do weryfikacji heterogenicznych baz wiedzy [Approaches to veryfing heterogeneous databases].* In: Baborski A. (red.), *Pozyskiwanie wiedzy z baz danych [Knowledge acquisition from databases].* Prace naukowe AE Wroclaw nr 891, Wydawnictwo AE, Wroclaw 2001, pp. 186-198.

Paci A. M., Canali D., *Designing a Tool to Know Invisible Resources: the Hyperguide Project, an XML Storyboard for Digital Collections Access,* Proc. Informing Science + IT Conference, June 19-21, 2002, Cork, Ireland, pp. 1217-1220. ISSN 1535-07-03

Palopoli L., Pontieri L., Terracina G., Ursino D., *Intensional and extensional integration and abstraction of heterogeneous databases.* „Data & Knowledge Engineering" vol. 35 no 3, December 2000, pp. 201-237.

Sandewall E., *Features and Fluents.* Oxford University Press, 1994.

Subrahmanian V. S., *Amalgamating Knowledge Bases.* „ACM Transactions on Database Systems", vol. 19, no 2, June 1994, pp. 291-331.

Subrahmanian V. S., Adali S., Brink A., Emery R., Lu J. J., Rajput A., Rogers T. J., Ross R., Ward Ch., *HERMES: A Heterogeneous Reasoning and Mediator System,*1997. http://www.cs.umd.edu//projects/hermes/publications/abstracts/hermes.html

Vetulani Z., *Automatyczna interpretacja* pytań *i udzielanie odpowiedzi jako technologia multimedialna [Automatic interpretation of questions and giving answers as multimedia technology].* http://www.zsi.pwr.wroc.pl/zsi/missi2002/pdf/p01.pdf (2002, retrieved January 29th, 2003)

Wiederhold G., *Mediators in the Architecture of Future Information Systems.* "IEEE Computer", March 1992, pp. 38-49.

Wiederhold G., *Intelligent Integration of Information.* Proc. of the ACM SIGMOD Conference on Management of Data, pp. 434-437, 1993.

Wiederhold G., *An Algebra for Ontology Composition.* Proc. of 1994 Monterrey Workshop on Formal Methods, September 1994, http://www-db.stanford.edu/pub/gio.

Wiederhold G., *Mediation to Deal with Heterogeneous Data Sources.* Proc. Intcrop'99, Zurich, "Lecture Notes in Computer Science" vol. 1580, Springer, 1999, pp. 1-16.

Wiederhold G., Jajodia S., Litwin W., *Dealing with Granularity of Time in Temporal Databases.* Proc. 3rd International Conference on Advanced Systems Engineering, Trondheim, Norway, 15 May 1991, LNCS vol. 498, Springer-Verlag 1991, pp. 124-140.

received: October 2003, revised version: February 2004