

18 Knowledge Management in Large Organizations

John Davies¹ · Paul Warren²

¹British Telecommunications Plc, Ipswich, UK

²Eurescom GmbH, Heidelberg, Germany

18.1	<i>Scientific and Technical Overview</i>	739
18.1.1	Introduction	739
18.1.2	The Challenges for Organizational Knowledge Management	740
18.1.3	Finding Information and Organizing Information so that It Can Be Found	741
18.1.3.1	Defects of the Conventional Search Engine	741
18.1.3.2	Semantic Indexing and Retrieval	743
18.1.3.3	Storing Information for Easier Retrieval	744
18.1.4	Sharing Knowledge Across the Organization	745
18.1.5	Helping with Processes	746
18.1.6	Information Integration	746
18.1.6.1	The Challenges of Information Integration	746
18.1.6.2	Approaches to Information Integration in the Enterprise	747
18.1.6.3	Using Ontologies for Information Integration	749
18.1.6.4	Research Themes in Information Integration	750
18.1.7	Integrating Structured and Unstructured Information	751
18.1.7.1	The Need to Analyze Text	751
18.1.7.2	Combining the Statistical and Linguistic Approaches	754
18.1.8	Sharing Knowledge Between Organizations	756
18.1.9	Another Look at Ontologies	757
18.2	<i>Example Applications</i>	758
18.2.1	Semantic Search, Browse, and Information Storage	759
18.2.1.1	Squirrel: An Example of Semantic Search and Browse	759
18.2.1.2	SEKTagent: A Different View on Semantic Search	762
18.2.1.3	Semantic Filing: TagFS and SemFS	763
18.2.1.4	Commercial Activities	764
18.2.2	Semantic Information Sharing	765
18.2.2.1	Effective Document Sharing with Semantic Technologies	765
18.2.3	The Semantic Desktop: Supporting the User Throughout His Work	770
18.2.3.1	Sharing Information and Metadata Across Applications and Desktops ..	770
18.2.3.2	Understanding User Context	772

18.2.4	Graphical and Semiautomatic Approaches to Information Integration ...	773
18.2.5	Extracting and Exploiting Semantics from Unstructured Information ...	775
18.2.5.1	Software for Text Analytics	775
18.2.5.2	Extracting Information from the World Wide Web	776
18.2.6	Sharing Information Across Organizations	777
18.2.6.1	An Example from Medicine	777
18.2.6.2	The Web of Linked Data	778
18.3	<i>Related Resources</i>	779
18.3.1	Semantic Web Interest Group: Case Studies and Use Cases	780
18.4	<i>Future Issues</i>	781
18.4.1	Web2.0 and Ontologies	781
18.4.2	Integrating into and across Enterprises	781
18.5	<i>Cross-References</i>	782

Abstract: This chapter provides an overview of the knowledge management (KM) problems, and opportunities, faced by large organizations, and indeed also shared by some smaller organizations. The chapter shows how semantic technologies can make a contribution. It looks at the key application areas: finding and organizing information; sharing knowledge; supporting processes, in particular informal processes; information integration; extracting knowledge from unstructured information; and finally sharing and reusing knowledge across organizations. In each application area, the chapter describes some solutions, either currently available or being researched. This is done to provide examples of what is possible rather than to provide a comprehensive list. The chapter also describes some of the technologies which contribute to these solutions; for example, text mining for analyzing documents or text within documents; and natural language processing for analyzing language itself and, for example, identifying named entities. Most fundamentally, the use of ontologies as a form of knowledge representation underlies everything talked about in the chapter. Ontologies offer great expressive power; they provide enormous flexibility, with the ability to evolve dynamically unlike database schema; and they make possible machine reasoning. The chapter concludes by identifying the key trends and describing the key challenges to be faced in the development of more powerful tools to support knowledge work.

18.1 Scientific and Technical Overview

18.1.1 Introduction

This chapter is concerned with how semantic technologies can make a difference to managing knowledge in large organizations. That the management of knowledge in organizations is a problem, and also an opportunity, is of no doubt. The management scientist Peter Drucker has commented that “the most important contribution management needs to make in the 21st century is to . . . increase the productivity of knowledge work” [1]. He identified increased productivity of manual work as a major distinguishing feature of successful organizations in the twentieth century and saw increased productivity of knowledge work as a similarly distinguishing feature of successful organizations in the twenty-first century. To Drucker, knowledge work was work where “the task does not program the worker,” that is, where the worker himself or herself has to make choices about what he does. Writing at the very end of the twentieth century, he estimated knowledge workers, that is, those involved in this sort of work, as possibly already composing two fifths of the US workforce. Note that in all discussion about “knowledge work” and “knowledge workers” it is important not to assume too elitist a definition. The users of the technology described in this chapter are not limited to people who have a graduate-level education but include everyone who works with knowledge. Indeed, in the paper referenced, Drucker talks at length about what he calls “technologists,” that is, people who work with their hands and yet also perform knowledge work. As Drucker notes, these can range from surgeons to telephone repair technicians. As a management scientist, Drucker’s concern was with

management's contribution to increasing knowledge worker productivity. The related concern here is with technology's contribution.

Organizations most conscious of the importance of knowledge, and of managing knowledge, tend to be those in the business of selling knowledge, that is, consultancies. Such organizations usually invest a significant amount of money in KM technology and employ KM professionals to support the sharing and reuse of knowledge. However, all organizations experience problems in the managing of knowledge, and in general the larger the organization, the greater the problems experienced. Over the last few decades, a large amount of research has been undertaken into how to improve the management of knowledge. This research has been technological, organizational, and user-oriented. This handbook is primarily about technology and the focus in this chapter is on the application of semantic technology to KM. However, the authors of this chapter believe that the technological, organizational, and user aspects of KM cannot be seen in isolation, but rather that understanding their interaction is important to designing successful KM systems.

The chapter is entitled "Knowledge Management in Large Organization." In some places, it refers to "information management," in others "knowledge management" (KM). The former is a necessary precursor to the latter. A widely quoted articulation of the difference between information and knowledge is due to R. L. Ackoff [2]. Ackoff sees information as useful data, providing answers to the "who," "what," "where," and "when" questions. Knowledge, on the other hand, enables the application of information; it answers the "how" question. The chapter will adhere as far as possible to this distinction, although the choice of terminology will also be guided by what seems the more natural English usage in any given circumstance.

This chapter makes a number of references to research projects and also to commercial systems. In general, these are chosen as examples to illustrate possible approaches. This chapter is not an exhaustive review of such systems and the examples given are simply those known to the authors. Their inclusion here does not imply any particular merit over systems not described here.

18.1.2 The Challenges for Organizational Knowledge Management

For those concerned with the management of information and knowledge in an organization, there are a number of challenges:

- Enabling the user to find, or be proactively presented with, the right information to achieve a particular task. The information might be taken from a wide range of sources, including databases, an intranet or the Internet; or it might be an amalgam of information from various sources. Related to this is the need to organize information in a way in which it can be efficiently retrieved.
- Sharing knowledge across the organization. Here also, the knowledge may be in a database, intranet, or the Internet (explicit knowledge), or simply in an individual's head (tacit knowledge). The person who needs the knowledge, and the owner or creator of the knowledge, although colleagues, may even be located on different continents.

- Helping users to navigate the processes, often collaborative processes, of which their work is composed. Central to this is sharing metadata between applications, to support a particular goal. Also important is having an understanding of the user's current context, and what he or she is trying to achieve.
- The integration of associated information which is held in multiple databases across and outside the organization. Note that the concern here is specifically with information which is inherently structured.
- The integration of structured information held in corporate databases with unstructured information, for example, held on the corporate intranet. By merging information from all corporate sources, a complete picture of what the organization knows about a particular topic can be obtained.
- Organizations do not exist in isolation but collaborate commercially and for the purposes of research. That collaboration requires a sharing of information. Typically, different organizations will have different vocabularies for talking about their shared concerns. This creates an enlarged version of the enterprise database integration problem described above.

The importance of these challenges has been highlighted by an Economist Intelligence Unit report, which surveyed 565 executives from various industries [3] – 74% of respondents said “data gathering is a significant or very significant challenge” and 68% said the same about data-searching. In fact, 42% of the respondents could not find relevant information when needed, 58% rated the challenge of knowledge sharing and collaboration as 4 or 5 (on a scale of 1–5), and 52% similarly rated the challenge of data integration as 4 or 5. Further, bearing out the need for information integration, 54% said that “necessary information resides in silos.” Interestingly, users were more satisfied with the quality and quantity of information available than with the ease of access and ease of use of that information.

➤ [Sections 18.1.3–18.1.8](#) discuss these challenges in greater detail, explaining why systems which analyze information on the semantic level are important in solving these challenges; ➤ [Sect. 18.1.9](#) makes some remarks about the ontological approach to information management compared to that of relational databases. ➤ [Section 18.2](#) describes some applications of semantic technologies to the challenges previously discussed. ➤ [Section 18.3](#) lists some relevant resources. Finally, ➤ [Sect. 18.4](#) discusses future trends and unsolved challenges.

18.1.3 Finding Information and Organizing Information so that It Can Be Found

18.1.3.1 Defects of the Conventional Search Engine

The search engine has been the great success story of the World Wide Web. However, its use within the organization has been less successful and has created a degree of frustration.

An important reason for this is well known. The page rank algorithm, pioneered by Google, depends on the rich pattern of hyperlinks which exist on the Web but which are rarely to be found on the organizational intranet.

However, even at its most successful, the conventional search engine suffers from an approach based on text-string matching and consequent failure to interpret the semantics of a query or the semantics inherent in the documents being queried. In particular, the failure to identify polysemy; a similar failure to take account of synonymy and other forms of semantic connection between terms; an inability to make use of context; and less than optimal interpretation of results.

Polysemy

A difficulty with query terms is that they may have multiple meanings; this is called query term polysemy. As conventional search engines cannot interpret the sense of the user's search, the ambiguity of the query leads to the retrieval of irrelevant information.

Although the problems of query ambiguity can be overcome to some degree, for example, by careful choice of additional query terms, there is evidence to suggest that many people may not be prepared to do this. For example, an analysis of the transaction logs of the Excite WWW search engine [4] showed that Web search engine queries contain on average 2.2 terms. Comparable user behavior can also be observed on corporate intranets. An analysis of the queries submitted to BT's intranet search engine over a 4-month period between January 2004 and May 2004 showed that 99% of the submitted queries only contained a single phrase and that, on average, each phrase contained 1.82 keywords.

Synonymy and Semantic Links

Converse to the problem of polysemy is the fact that conventional search engines that match query terms against a keyword-based index will fail to match relevant information when the keywords used in the query are different from those used in the index, despite having the same meaning (index term synonymy). Although this problem can be overcome to some extent through thesaurus-based expansion of the query, the resultant increased level of document recall may result in the search engine returning too many results for the user to be able to process realistically.

In addition to an inability to handle synonymy and polysemy, conventional search engines are unaware of any other semantic links between concepts. Consider, for example, the following query:

“telecom company” Europe “John Smith” director

The user might require, for example, documents concerning a telecom company in Europe, a person called John Smith, and a board appointment. Note, however, that a document containing the following sentence would not be returned using conventional search techniques:

“At its meeting on the 10th of May, the board of London-based O2 appointed John Smith as CFO”

In order to be able to return this document, the search engine would need to be aware of the following semantic relations:

O2 is a mobile operator, which is a kind of telecom company.

London is located in the UK, which is a part of Europe.

A CFO is a kind of director.

Lack of Context

Many search engines fail to take into consideration aspects of the user's context to help disambiguate their queries. User context would include information such as a person's role, department, experience, interests, project work, etc. A simple search on BT's intranet demonstrates this. A person working in a particular BT line of business searching for information on their corporate clothing entitlement is presented with numerous irrelevant results if they simply enter the query "corporate clothing." More relevant results are only returned should the user modify their query to include further search terms to indicate the part of the business in which they work. As discussed above, users are in general unwilling to do this.

Presentation of Results

The results returned from a conventional search engine are usually presented to the user as a simple ranked list. The sheer number of results returned from a basic keyword search means that results navigation can be difficult and time consuming. Generally, the user has to make a decision on whether to view the target page based upon information contained in a brief result fragment. A survey of user behavior on BT's intranet suggests that most users will not view beyond the tenth result in a list of retrieved documents; only 17% of searches resulted in a user viewing more than the first page of results. Essentially, the requirement is to move from a document-centric view to a more knowledge-centric one (for example, by presenting the user with a digest of information gleaned from the most relevant results found as has been done in the Squirrel semantic search engine described later in this chapter).

18.1.3.2 Semantic Indexing and Retrieval

The previous section discussed the limitations of conventional textual search technology and indicated that these limitations were caused by a failure to interpret the semantics both in the query and in the textual corpus being interrogated. Chapter on [Semantic Annotations and Retrieval: Manual, Semiautomatic, and Automatic Generation](#) of this handbook has described techniques for the automatic creation of semantic annotations. As explained in [5], semantic indexing and retrieval can then be performed on top of the semantic annotations. Indexing can be done with respect to two semantic features: lexical concepts and named entities. In this way, a number of the problems discussed above can be overcome.

Lexical concepts are introduced to overcome the polysemy discussed earlier. Thus, a word with two different meanings will be associated with two different lexical concepts. Word-sense disambiguation techniques can be used to disassociate these meanings [6]. Similarly, knowing that two words or phrases are associated with the same lexical concept enables the system to

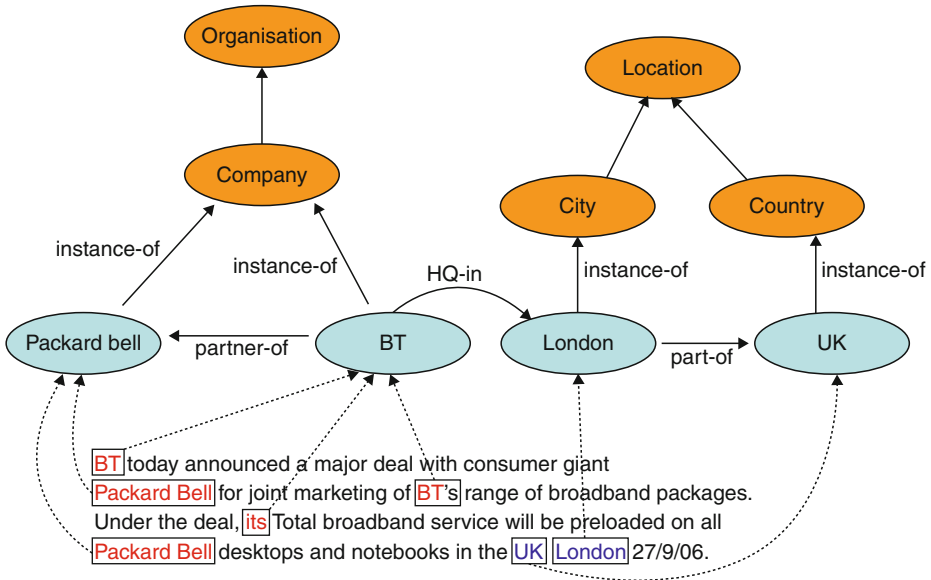
cope with synonymy. Moreover, the use of lexical concepts also enables hyponym-matching. A hyponym is a word of more specific meaning. Thus, referring to the example in ▶ Sect. 18.1.3.1, *CFO* is a hyponym of *director*. Hyponym-matching overcomes the problem that a search for *director* will not identify references to *CFO* which may be relevant.

Named entities are items such as proper nouns (denoting, for example, persons, organizations, and locations), numbers, and dates. One study found that named entities were a common query type, in particular people's names, while "general informational queries are less prevalent" [7]. Such named entities can be identified as instances of a predefined ontology. A typical ontology for such purposes would need to have information about people, geography, company structure, etc. One such ontology is PROTON [8], which was developed by Ontotext (<http://www.ontotext.com>) and used within the SEKT project (<http://www.sekt-project.com/>) as the basis for several semantic search and browse tools. In fact, PROTON also includes a world knowledge base. The word "knowledge base" is used to describe a set of instances and instantiated relations conforming to an ontology. Thus, the PROTON world knowledge base is a set of instances and instantiated relations, which are used to pre-populate the ontology. This initial knowledge base can then be extended through analysis of the textual corpus. Of course, this approach, while highly accurate, can lead to error. Therefore, information in the knowledge base is flagged to indicate whether it is predefined or whether it is learned from the document database. The PROTON ontology is itself extensible, any particular domain can develop its domain ontology as an extension to PROTON.

▶ Figure 18.1 illustrates how sentences can be analyzed and the named entities related to the classes of an ontology. Packard Bell and BT have been identified as instances of companies, while London and UK have been identified as instances of city and country, respectively. Once identified, these instances then form part of a knowledge base. Note that "its" has been identified as being equivalent to BT in this particular sentence. The identification of words such as pronouns with the words or phrases which they stand in for is known as anaphora resolution. Software to achieve this textual analysis is described in ▶ Sect. 18.2.5.1.

18.1.3.3 Storing Information for Easier Retrieval

Quite apart from the problem of finding information on the Web or corporate intranet, many people find it difficult to retrieve information they have stored on their personal computers. This subject has been extensively studied, for example, by William Jones [9, 10]. One reason for the difficulty is that people frequently do not have a consistently defined folder structure. In fact, even an entirely consistent structure can lead to ambiguity and questions such as "are the company financial results for 2008 in the folder *2008*, perhaps in a sub-folder *finance*, or in the folder *finance* in a sub-folder *2008*." Again, the problem is that the system is unable to understand semantics which are relatively obvious to a human, and which make it clear to the human that the paths *2008/finance* and *finance/2008* are likely to lead to related information. One proposed solution is the use of tags rather than folders. Reference [11] discusses the advantages and disadvantages of the two approaches.



■ Fig. 18.1

Relating the named entities in a sentence to an ontology

18.1.4 Sharing Knowledge Across the Organization

Sharing knowledge across large organizations is a notoriously difficult problem. Sometimes, the need is to make an employee aware of a document created by a colleague; at other times the need is to put the colleagues directly in touch. In any case, the colleagues may be completely unaware of each other and located geographically far apart. Of course, a useful document might have been created some time ago, by an employee who has moved on to other work or left the organization.

As already observed, consultancies such as Ernst and Young [12, 13], frequently take this subject most seriously. Typically they have a combination of part-time knowledge management enthusiasts in their operating units and full-time knowledge management specialists in a central unit. They use a platform, such as Lotus Notes, for document storage; a typical such document might be a customer proposal, which might be partially reused for other customers. In some cases, users may simply enter a document directly into the repository. In others, the document is vetted for quality by one of the knowledge management team. In both cases, the user will be required to describe the document using metadata compliant with a predefined taxonomy. Depending on the experience of the user and the particular document, this can take a significant amount of time and inhibit information being entered into the repository. A similar problem applies in reverse. To retrieve information, a user needs to understand the taxonomy, and of course the original metadata need to be accurate. Information may be missed, or the complexity of the system may again deter its use. What is needed is to analyze the documents as they are entered

into the system, so as to automatically create semantic metadata, which can be used for document retrieval. Automatic metadata creation also provides a consistency which may not occur when metadata are manually created.

Systems also exist for identifying people within the organization with a particular expertise. These may rely on employees inputting their information directly, with the result that the information is often not present or not up-to-date. Alternatively, they may use information collected by the human resource department, which has similar problems. What is really required is to understand a person's expertise by semantic analysis of the documents, e-mails, etc., which he or she creates and reads.

18.1.5 Helping with Processes

Current productivity tools offer basic support for processes, but little proactive help. Within Microsoft Outlook, for example, calendar and contact facilities provide tools for the user. However, all the intelligence needs to be supplied by the user. When the user types "phone John Smith" at a given time in his diary, there is no automatic link to the contact book entry for John Smith.

In addition, what information the system does have is routinely lost. Imagine the user receives an e-mail with attachments from John Smith as part of the customer X bid proposal process. He saves the attachments in a folder. Then the link between the attachments and John Smith, or customer X, is lost. If the user wants to find all information sent by John Smith or about customer X, then there is nothing associated with the saved files to help him. When he or she is working on the customer X proposal process, there are no metadata associated with those files to indicate their relevance to customer X.

Moreover, current systems have no idea of the context in which the user is working or the process currently being followed. For example, if the user is a patent lawyer with six different patent filings under consideration, the system has no idea which one is currently the focus of his attention. Nor does it know whether the user is creating a patent, reviewing a colleague's proposed filing, or searching for prior art. Yet such information would enable the system to proactively help the user. What is missing are metadata shared between applications and linked to the context of the user's work and the processes he or she performs.

18.1.6 Information Integration

18.1.6.1 The Challenges of Information Integration

McComb [14] suggests "that at least half the cost of integrating systems comes down to resolving semantic issues." Integration is a challenge in all organizations, but particularly where mergers and acquisitions have led to the need to rationalize different systems. Even without the stimulus of mergers and acquisitions, organizations often need to rationalize their information. For example, separate product lines may have different customer databases and this creates difficulties for cross-selling.

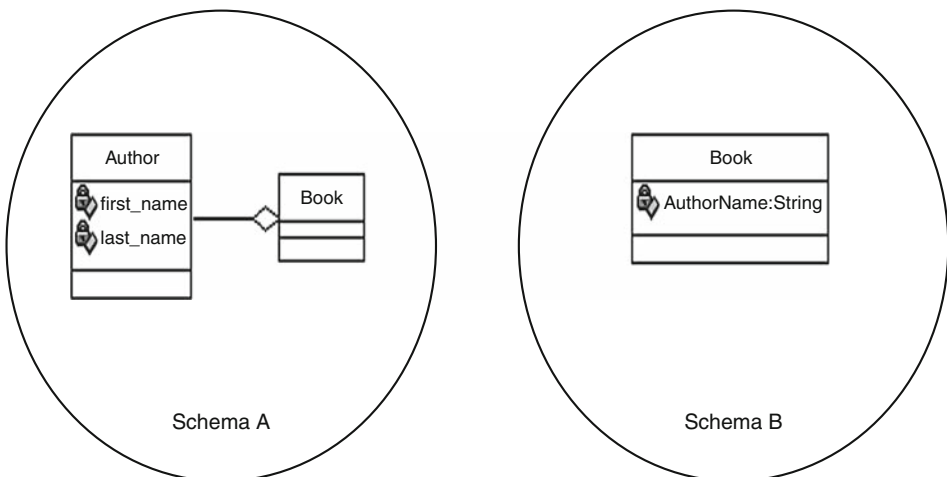
The problems of polysemy and synonymy, discussed earlier in a different context, arise again here. Different database schemas use the same terms with different meanings and different terms with the same meaning. Different database schemas may also use different structures and different values to represent the same information. This is illustrated in [▶ Figs. 18.2](#) and [▶ 18.3](#), which are adapted from [15]. These conflicts are very frequent, occurring as a natural consequence of data modeling – whether due to isolated development, changing needs, organizational or structural differences, or simply the different approach of two human data modelers.

As McComb points out, non-semantic issues such as language mismatch and platform boundaries, rarely cause surprise and can be planned for. It is the semantic mismatches which create the real problems in systems integration.

18.1.6.2 Approaches to Information Integration in the Enterprise

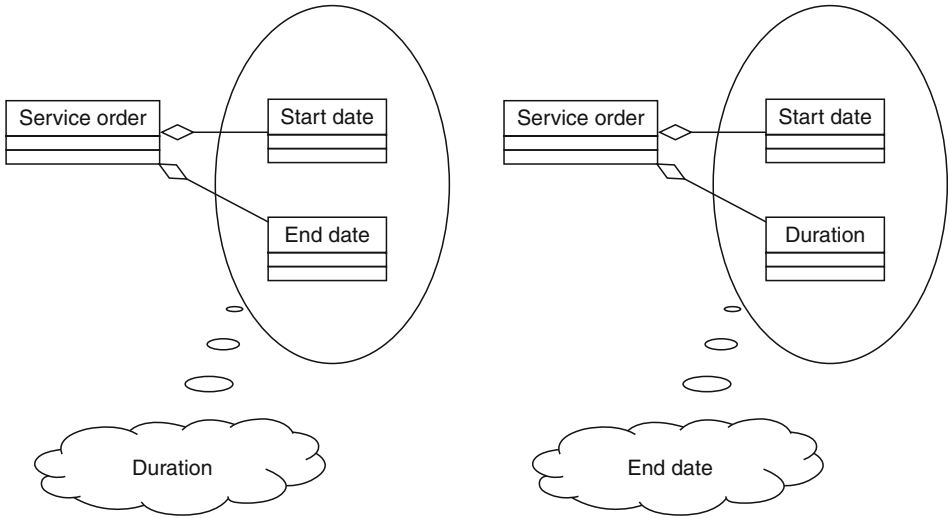
Information integration has been recognized as a significant problem in enterprises for some years, certainly well before the Semantic Web was conceived, and before the use of ontologies were a major subject of research in Computer Science. It is a problem of considerable economic importance. Based on a number of papers in the literature, Bernstein and Haas, claim that IT departments spend about 40% of their budget on information integration [16]. Their paper classifies information integration into a number of strands: data warehousing; virtual data integration; message mapping; object-to-relational mappers; document management; and portal management.

A *data warehouse* consolidates data from multiple database sources so as to allow querying to provide a comprehensive view of, for example, a customer. This consolidation



■ Fig. 18.2

Aggregation conflict – difference in structure



■ Fig. 18.3

Value representation conflict

is achieved through the use of Extract-Transform-Load (ETL) tools. The source databases are likely to have different schemas, and the warehouse database schema needs to permit mapping from each of these source schemas.

Virtual data integration avoids creating an actual warehouse of data yet also provides an integrated view. This is done by a query mediator, which translates the user's query into queries on the individual databases. Such an approach is referred to as Enterprise-information Integration (EII).

Message mapping uses message-oriented middleware to “integrate independently applications by moving messages between them.” Where a broker is used, this is called enterprise application integration (EAI) and where all applications use the same protocol this is called an enterprise service bus (ESB).

Data warehousing, virtual data integration, and enterprise application integration all involve mapping between database schema, which is the subject of this section.

Document management may be concerned with integration on a superficial level, for example, making documents available on a single *portal*. However, integration may also mean combining information from documents to create a new document or database.

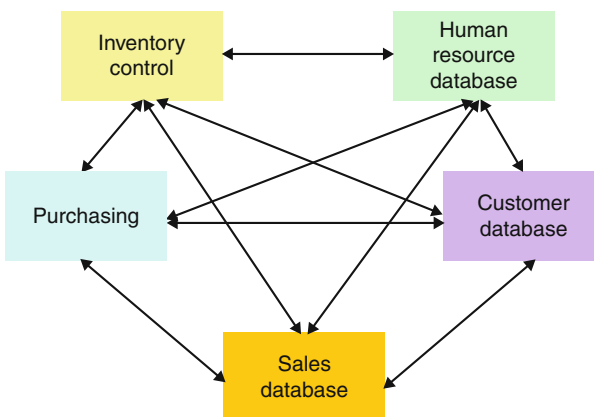
Bernstein and Haas, in their overview, also make the point that information integration was originally conceived as a predefined problem, that is, integrating a number of enterprise databases. More recently, the problem has widened, increasing to personal information management, creating a link with a theme of [Sect. 18.2.3](#).

For the personal views of a number of practitioners in the field of information integration, see [17]. Amongst the multiple authors, one (Pollock) argues strongly that EII will in the future make use of formal semantics. As Pollock sees it, the problem with database integration is that the structures contain no explicit formal semantics. Draper

stresses the importance of data modeling, and the need to model “the relationships and meaning of data separately from the aspect of when and where it is computed.” Rosenthal calls for, not just semantic integration but also “semantics management.” Within this, he includes guiding (e.g., enterprise managers) as to what concepts should be used, either to describe existing systems or for newly built systems. He sees this as a compromise between totally centralized and peer-to-peer systems. Bitton, arguing why EII will never totally replace data warehousing, draws attention to the performance implications of query processing in EII. Performance implications will remain an issue in the more sophisticated ontological approach proposed below. Finally, Sikka calls for a common semantic framework for information retrieval from structured and unstructured sources. This points again to the theme of [Sect. 18.2.3](#).

18.1.6.3 Using Ontologies for Information Integration

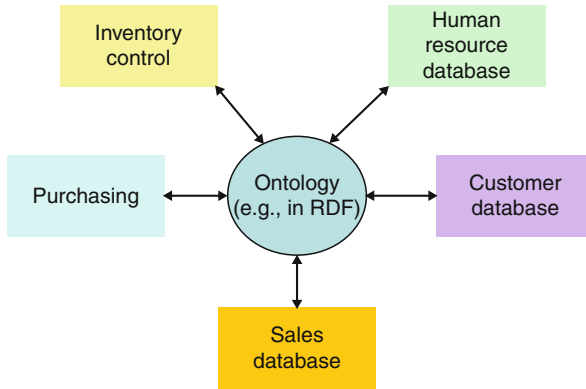
The value of ontologies in information integration stems from the ability to create an overarching ontology which can subsume multiple database schemas. The current state of the art in information integration is illustrated in [Fig. 18.4](#). To achieve integration at the semantic level, mappings are created between each database. These might be databases internal to one organization, for example, order processing and stock control databases; or the mappings might be across organizations, for example, between databases held by separate companies working together in a joint venture or supply chain. In any case, the problem is that the number of mappings increases quadratically with the number of databases.



N.B. This shows the situation within one organization. The problem multiplies when organizations wish to work together and link up their databases.

Fig. 18.4

Information integration today – many one-to-one mappings



■ Fig. 18.5

Illustrating the use of a central broker

► *Figure 18.5* illustrates the use of a central broker to reduce the number of mappings to that of the number of databases. Of course, the idea of a central hub is not new in systems integration. The innovation here is that the integration is at the semantic level, and is achieved through the use of a central overarching ontology based on open, lightweight standards. Note that the mappings are potentially both between schema and instances. To take two trivial examples, in the case of schema, a mapping is required which identifies “first name” and “forename” as the same; in the case of instances, the mapping must equate “Paul William Warren” with “Paul W Warren.”

Information about tools and techniques for creating semantic mappings is given later in ► [Sect. 18.2.4](#). An example of using this approach in the supply chain is described in [18]. The example shows how a number of Internet service providers can integrate their heterogeneous operational support systems with those of a telecoms operator, in this case BT. The approach reduces costs and time-to-market while, in particular the use of ontologies, enables a reuse of services.

18.1.6.4 Research Themes in Information Integration

There has been significant research activity into the use of ontologies for semantic integration. As long ago as 2004, there was a special issue of the ACM SIGMOD Record on Semantic Integration. In the introduction, the editors drew attention to three research activities, which remain challenges today [19]:

- Extending the scalability of schema techniques to large schemas.
- Designing the interaction with the user. It is generally accepted that a schema matching system will never be completely autonomous, and hence user interaction is required. The user interface has its own scalability problems. Moreover, schema matching may be part of a larger task, hence the schema-matching user interface needs to be embedded into some larger system.

- Mapping maintenance. Schemas change frequently, and therefore mappings need to be maintained.

The editors also noted the need for measures to establish similarity between schemas; similarity measures remain an active area of research.

In the same issue, Noy provides another view of the use of ontologies in semantic integration [20]. She divides research into semantic integration into three “dimensions”: mapping discovery, representations of the mappings, and reasoning with the mappings.

To facilitate mapping discovery, Noy argues for using common upper-level ontologies. She argues that “if two ontologies extend the same reference ontology in a consistent way, then finding correspondences between their concepts is easier.” Of course, this describes a situation where one is starting from scratch and extending an upper-level ontology to create domain ontologies. Where there are existing legacy ontologies, mapping will be much harder.

Turning to mapping representation, she identified three ways of doing this. One can construct an ontology of mappings, in which case the individual mappings become instances of concepts in the ontology. Alternatively, bridging axioms can be defined in first-order logic to represent transformations. Finally, views can be used to describe mappings from a global ontology to a local ontology, that is, the global ontology is used to provide access to local ontologies.

Another paper in the same issue emphasizes the need for customizability to create an “industrial strength” schema mapping tool [21]. The authors argue that customizability is needed to select and combine the techniques appropriate to the particular schema-matching problem; to control scalability, for example, by trading off response time and quality of the result; and to enable extensibility so that new techniques can be easily added. The authors also emphasize that schema matching is the first step in automating the creation of mappings between schemas. The second step is query discovery, in which queries are obtained to translate instances of the source schema into instances of the target schema. More recently, two of the authors of this paper, both at Microsoft, have gone on to describe their work in model management [22]. Model management is designed to support schema matching, merging, translation, comparison, and mapping composition. It is not a user-oriented tool, but rather a reusable component to be embedded in user tools. One aspect of the direction of this research is an increased emphasis on the runtime system to support the execution of mappings. The paper contains a review and comparison of, on the one hand, the approach focused on the mapping designer, and, on the other hand, their approach of focusing research on the model management. In fact, they see the two approaches as converging.

18.1.7 Integrating Structured and Unstructured Information

18.1.7.1 The Need to Analyze Text

Conventional corporate information systems are built on relational database technology. This is true whether the systems are for customer relationship management, product

information, employee information, competitor information, etc. Section 18.1.6 has just discussed the problems of integrating such database systems. A further problem lies in capturing unstructured information and semi-structured information. By “unstructured” information is meant information for which no schema exists, for example, information in text on the intranet, in memos on personal computers, in e-mails, slide presentations, etc. By semi-structured information is meant information for which some kind of schema exists but for which the schema is not defined as rigorously as is the case in relational databases. This includes information in applications such as spreadsheets where schemas may exist in the form of row and column headings.

The claim has been made that over 80% of the data in an organization is unstructured [23]. Whether this claim is true, or even practically verifiable, is not important. It is a common experience that a great deal of valuable information in an organization exists in this form. What is needed is to extract this information and transform it into structured form to enable merging with the structured data. The problem is that structured data have defined semantics in the form of schemas. These semantics may be local to the particular application, rather than being expressed using shareable ontologies, but they are semantics nevertheless. The application knows, for example, that the *price* field in a relational database contains the price in an agreed currency. In unstructured data it could be argued that the semantics are still there. A human can detect when a brochure describes a product price. However, the semantics are no longer defined in a machine-interpretable way. The price can be anywhere in the document and can be introduced by many different kinds of language. Interpreting these semantics is a task which until recently has been regarded as requiring human intelligence.

If structured information could be extracted from unstructured data, then there are many applications which would benefit. A complete picture could be built up, based on all the information available to the enterprise, of, for example, any particular customer, supplier, or competitor. Instead of searching separately through e-mails, memos, corporate intranet and databases, a sales advisor would have a complete picture of a customer, based on all those sources.

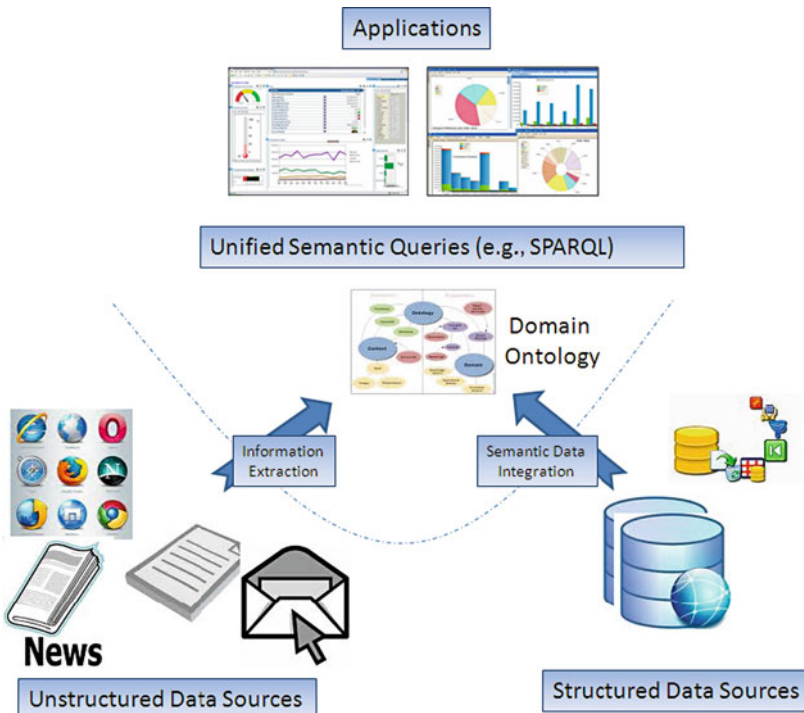
Added to the opportunity cost of not being able to use all the information potentially available to the organization, is risk associated with the regulatory environment. Organizations which do not disclose all relevant information to regulatory authorities may be seriously penalized. Yet the organization can only disclose information it knows it has. Information lost on corporate computers cannot be disclosed at the appropriate time – but will certainly be revealed if the organization is subject to a detailed forensic analysis of hard drives prior to a legal hearing. As an example, Forrester [24] describes a \$1.4 billion judgment against Morgan Stanley, arising from the latter’s inability to produce requested information.

The growing use of e-mail is one factor increasing the importance of unstructured information. AIIM (<http://aiim.org>), a nonprofit organization in the electronic content management industry, confirms that e-mail is a central means for business documentation [25]. Over 70% of the respondents to an AIIM survey reported exchanging

confidential or sensitive information via e-mail. AIIM found that e-mail is being used for critical processes such as contract negotiation, HR discussions, and invoice delivery. US public companies are also affected by the Sarbanes–Oxley Act, a US federal law enacted in 2002 which, among other things, sets enhanced reporting requirements for US public companies. Nearly one third of respondents reported that the Sarbanes–Oxley Act has affected the way their organization views e-mail.

All this points to a growing business need to understand the semantics of textual information, to extract such information from free text, convert into a structured form, and merge with preexisting structured information.

The overall goal is to combine structured and unstructured information and make the combined result available to a range of applications. This is illustrated in [Fig. 18.6](#) where information from a variety of unstructured sources is combined with information from databases to create information described in terms of an ontology. This can then be combined with domain-specific knowledge and business rules, and then operated on by semantic queries to input to client applications. Typical business rules would depend upon the application. For example, in sentiment analysis, where a company is interested in the perception of its products as expressed in blogs, etc., on the Web, then a rule might



■ Fig. 18.6

Combining structured and unstructured information

state that if customer perception for a particular product drops below a given level, then that product should be categorized as “at risk.” Combining information from structured and unstructured sources, a rule might say that if product sales have declined in the last month, compared with the month before, and customer perception has dropped below a particular level, then the product is in the “high-risk” category.

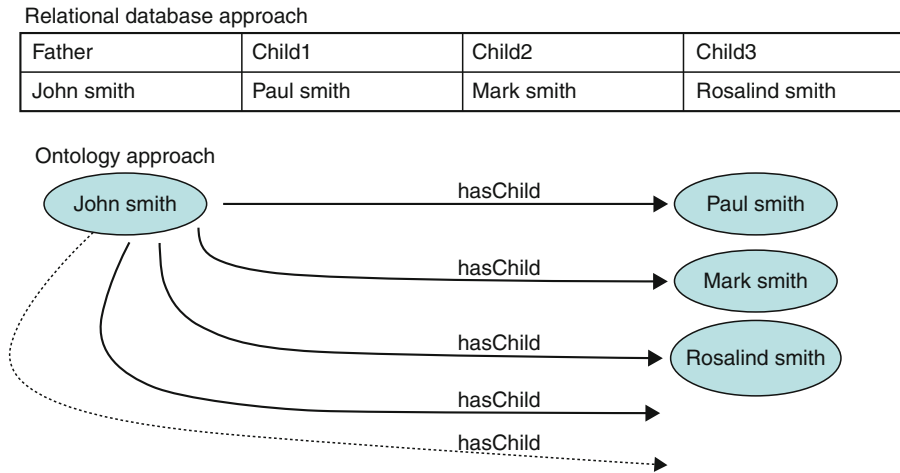
The essential challenge is to create some structure out of unstructured text. One way to do this is to create semantic metadata. HTML, the language which underlies the WWW and corporate intranets, is based on the use of metadata. However, the metadata in HTML are used to describe the format of data, for example, to indicate a heading or a bulleted list. The need here is to create semantic metadata, that is, metadata which provide information about the data.

Such metadata can exist at two levels. They can provide information about a document or a page, for example, its author, creation, or last amendment date, or topic; or they can provide information about entities in the document, for example, the fact that a string represents a company or a person or a product code. The metadata themselves should describe the document or entities within the document in terms of an ontology. At the document level, there might be a property in the ontology, for example, has Author, to describe authorship. Within the document classes such as Person, Company, or Country would be used to identify specific entities.

18.1.7.2 Combining the Statistical and Linguistic Approaches

The metadata could be created by the authors of the document. In general, this will not happen. The authors of Word documents or e-mails will not pause to create metadata. The need is to generate metadata automatically, or at least semiautomatically. There are two broad categories of technology which can be used for this: statistical or machine learning techniques; and information extraction techniques based on natural language processing. The former generally operate at the level of documents, by treating each document as a “bag of words.” They are, therefore, generally used to create metadata to describe documents. The latter are used to analyze the syntax of a text to create metadata for entities within the text, for example, to identify entities as Persons, Companies, Countries, etc. Nevertheless, this division should not be seen too starkly. For example, one of the goals of the SEKT project (<http://www.sekt-project.com>), a European collaborative research project in this area which ran from 2004 to 2006, was to identify the synergies which arise when these two different technologies are used closely together. An overview of semantic knowledge management, including these two approaches to creating metadata, is given in [73].

The metadata can create a link between the textual information in the documents and concepts in the ontology. Metadata can also be used to create a link between the information in the document and instances of the concepts. These instances are stored in a knowledge base. Thus, the ontology bears the same relationship to the knowledge base as a relational database schema bears to the information in the database. In some cases, the



In a database there are a predefined number of fields, e.g., here there are three “child” fields. An ontology is more flexible, e.g., one can have an unlimited number of instantiations of the “hasChild” property.

■ Fig. 18.7

Ontologies offer greater flexibility than database schema

ontology and the knowledge base will be stored together, in other cases separately. This is essentially an implementation decision.

Ontologies are particularly useful for representing knowledge from unstructured text because of their flexibility and ability to evolve. Once created, ontologies can be far more easily extended than is the case for relational database schema. ▶ *Figure 18.7* provides a simple illustration of how the ontological approach overcomes the limitation in databases of having a predefined number of fields. Here, the occurrence of new children simply requires new instantiations of the “hasChild” relation. This contrasts with a database design where one would need to decide at the beginning the maximum number of children a person might have. Moreover, it is not even necessary to decide initially what relations are needed. The ontology designer might realize at some stage that the “hasBrother” relation is useful in some cases. This can be added to the ontology far more easily than adding a new field to a database. This is not to say that the ontology-based approach will replace the use of relational databases. With increased flexibility comes increased computational expense. The ideal is to combine the two approaches.

Where the system identifies a text string as an instance of a concept in the ontology but which is not represented in the knowledge base, then that instance can be added to the knowledge base. For example, the text string “ABC Holdings” may be identified as a company, but one not represented in the knowledge base. The system can then add “ABC Holdings” to the knowledge base. ▶ *Section 18.1.3* has already discussed how entities in text can be associated with entities in the knowledge base; this was illustrated in ▶ *Fig. 18.1*.

Research is also in progress to use natural language processing techniques to learn concepts from text, and thereby extend the ontology. However, this is a significantly harder problem. For an example of the state of the art, see [74].

18.1.8 Sharing Knowledge Between Organizations

There are a number of motivations for an organization wanting to share knowledge with other organizations. One of the most obvious is to cooperate in a supply chain, where the information shared is contractual. Another is to undertake collaborative research, or simply to share research results. A discussion of knowledge sharing in the supply chain is properly the domain of eBusiness, which is discussed in the next chapter of this volume; while knowledge sharing for research is the domain of eScience, discussed in the previous chapter. However, there are situations where organizations need to collaborate together to achieve common goals, and where the activity might properly be regarded as knowledge management. One such is within the domain of medicine, where general practitioners and clinicians need to share information about patients, for example, describe their symptoms. Of course, the boundary between eBusiness, eScience, and knowledge management is somewhat fuzzy. In the medical example, the same vocabulary might be used in a clinical environment (knowledge management), to share information with an insurance company (eBusiness), or for research into illness (eScience).

In any case, the problems are similar. There is a need for a shared vocabulary, for example, for use within an industry sector or within a specialism. Usually these vocabularies, created and maintained by a standards body, are defined in a natural language, frequently English. Such informal definitions give rise to redundancies and even inconsistencies. They also give rise to misunderstandings when different parties interpret the natural language differently. What is required is a more formal approach based on knowledge representation techniques, for example, ontologies. The use of the informal approach is partly historical, some of these vocabularies have a long history going back before the use of ontologies was proposed. Even today, many of the people developing such vocabularies will not be skilled in knowledge representation and will use natural language. As a consequence, it is frequently necessary for ontologists to come along after the event and create a more structured approach out of what exists informally. This is true in eBusiness where Electronic Data Interchange standards such as ANSI ASC X12 (<http://www.x12.org/>) and the United Nations's EDIFACT (e.g., <http://www.unece.org/trade/untdid/welcome.htm>) have been in existence for some decades. An attempt to use an ontology to describe at least the syntax of X12, prior to “ontologizing” the semantics, is described in [26]. [▶ Section 18.2.6.1](#) described an example more properly from knowledge management, that of the use of ontologies in medical informatics.

An alternative approach to shared vocabularies is to use, for example, RDF, to create self-describing data and to make that data available to other organizations. If that data is made openly available on the Web, then this creates a Web of linked open data. This is

exactly what the *linking open data* initiative is in the process of achieving; this is described briefly in [▶ Sect. 18.2.6.2](#) and in more detail in [▶ Semantic Annotation and Retrieval: Web of Data](#).

18.1.9 Another Look at Ontologies

The constant theme running through this chapter has been the use of ontologies. An early, but still relevant, overview and categorization of the ways ontologies can be used for knowledge sharing is given in [27]. Here, the use of ontologies is categorized in a number of ways. Ontologies can be used in conjunction with conventional (i.e., nonintelligent) software or alternatively in conjunction with software employing AI techniques. The reference lists a number of principles which remain true: knowledge engineering needs to be minimized, as it represents an overhead; KM support needs to be integrated into everyday work procedures; and KM applications need to process information in an integrated manner. It describes a range of applications which remain important: knowledge portals for communities of practice; lessons learned archives; expert finders and skill management systems; knowledge visualization; search, retrieval, and personalization; and information gathering and integration.

Another high-level view of ontologies, and specifically their use in achieving data connectivity, is given by Uschold and Gruninger [28]. They note that connectivity is required at three layers: physical, syntactic, and semantic. Great strides have been made in achieving connectivity at the first two layers. The challenge is now the third, and ontologies have a key role here. Semantic heterogeneity is a fact of life to be overcome – “there will always be sufficiently large groups for which global agreements are infeasible.” They present a spectrum of kinds of ontologies, defined by degree of formality. At the informal end, there are sets of terms, with little specification of the meaning, and also ad hoc hierarchies, such as in Yahoo!. At the formal end, there are, for example, description logics. At the informal end, some of these might not properly be called ontologies, for example, by members of the knowledge representation community. The point is that they are used in similar ways as some formal ontologies. Uschold and Gruninger compare ontologies with database schema; making the point that the mixing of types (concepts) with instances is a feature of ontologies which does not occur in database schema. In their view this is largely because of the much greater scale and performance requirements for database systems. Note that this is a computational feature; computationally database schema and database instances are treated quite separately. This is less the case in the ontological approach; indeed it can in some cases be a matter of design style whether an entity is represented as a concept or an instance. However, when one turns to implementation, the converse can be true. A database schema is embedded in the database; an ontology can exist in a separate physical implementation.

The authors of this chapter have prepared their own summary of the chief differences between the relational database and ontological knowledge base approach. This is summarized in [▶ Fig. 18.8](#).

	Relational database	Ontological knowledge base
Information model	Schema <ul style="list-style-type: none"> • Hard to evolve • Implemented with instances in database • Computationally separate from instances 	Ontology <ul style="list-style-type: none"> • Flexible • Can be implemented separately from instances • Computationally concepts and instances treated similarly
Information which can be retrieved	What you put in is what you get out	Information entered into knowledge base plus inferences from that information

■ Fig. 18.8

Comparison of relational databases and ontological knowledge bases

Uschold and Gruninger identify four ways in which ontologies help achieve a common understanding. Three are relevant to the theme of this chapter:

- Neutral authoring. Here an ontology exists for authoring purposes, and the results are then translated into a variety of target ontologies. Enterprise modeling is an example of this.
- Common access to information. Here, the ontology is used as a neutral interchange format, as discussed above. The objective is to avoid the need for $O(N^2)$ translators.
- Query-based search, that is, a sophisticated indexing mechanism with the added benefit of permitting answers to be retrieved from multiple repositories.

Uschold and Gruninger describe the first of these as using neutral ontologies, without describing formally what the adjective “neutral” means here. They go on to add that, in the case of neutral authoring, the ontology can contain only those features present in all of the target systems and that, in the case of providing common access to information, the neutral ontology must cover all of the concepts in each of the target systems. This, in a sense, provides a definition of what “neutral ontology” means in each of these two cases. In the latter case what Uschold and Gruninger call a neutral ontology is what others refer to as an overarching ontology.

They also identify the use of ontologies for specification in software engineering, which is beyond the scope of this chapter.

18.2 Example Applications

Building on the discussions in ▶ Sect. 18.1, this section describes example applications of semantic technology addressing each of the challenges described in the previous section. ◀ Sections 18.2.1–18.2.6 describe responses to each of these challenges: searching and finding information; sharing information within organizations; helping users to navigate processes, including by taking account of the user’s context; integration of structured data; extraction of structured information from unstructured data; and sharing information across organizations.

18.2.1 Semantic Search, Browse, and Information Storage

18.2.1.1 Squirrel: An Example of Semantic Search and Browse

Squirrel [29] provides combined keyword-based and semantic searching. The intention is to provide a balance between the speed and ease of use of simple free text search and the power of semantic search. In addition, the ontological approach provides the user with a rich browsing experience. For its full-text indexing, Squirrel uses software from the open-source Lucene suite, see <http://lucene.apache.org/>. PROTON is used as the ontology and knowledge base, while KIM [30] is used for massive semantic annotation.

The KAON2 [31] ontology management and inference engine provides an API for the management of OWL-DL and an inference engine for answering conjunctive queries expressed using the SPARQL syntax. KAON2 also supports the Description Logic-safe subset of the Semantic Web Rule Language (SWRL). This allows knowledge to be presented against concepts that goes beyond that provided by the structure of the ontology. For example, one of the attributes displayed in the document presentation is “Organization.” This is not an attribute of a document in the PROTON ontology; however, affiliation is an attribute of the Author concept and has the range “Organization.” As a result, a rule was introduced into the ontology to infer that the organization responsible for a document is the affiliation of its lead author.

Users are permitted to enter terms into a text box to commence their search. This initially simple approach was chosen since users are likely to be comfortable with it due to experience with traditional search engines. Squirrel then calls the Lucene index and KAON2 to identify relevant textual resources or ontological entities, respectively. In addition to instance data, the labels of ontological classes are also indexed. This allows users to discover classes and then discover the corresponding instances and the documents associated with them without knowing the names of any instances, for example, a search for “Airline Industry” would match the “Airline” class in PROTON. Selecting this would then allow the users to browse to instances of the class where they can then navigate to the documents where those instances are mentioned.

➤ *Figure 18.9* shows the meta-result page. This is intended to allow users to quickly focus their search as required and to disambiguate their query if appropriate. The page presents the different types of results that have been found and how many of each type for the query “home health care.”

➤ *Figure 18.10* shows a document view. The user has selected a document from the result set, and is shown a view of the document itself. This shows the metadata and text associated with the document and also a link to the source page if appropriate – as is the case with Web pages. Semantically annotated text (e.g., recognized entities) is highlighted. “Mousing over” recognized entities provides the user with further information about the entity extracted from the ontology. Clicking on the entity itself takes the user to the entity view.

➤ *Figure 18.11* shows an entity view for “Sun Microsystems.” It includes a summary generated by OntoSum [30]. OntoSum is a Natural Language Generation (NLG) tool which takes structured data in a knowledge base (ontology and associated instances) as

Matches for your query:

- [Journal Articles](#): 76
- [Conference Papers](#): 46
- [Periodicals](#): 257
- [Web Pages](#): 16
- [Library Topics](#) (60) including: [Home health care](#)(4), [Health care \(Technical\)](#)(135), [Rural health care](#)(1), [Mental health care](#)(4), [Long term health care](#)(2)
- [Organisations](#) (597) including: [National HealthCare Corporation](#)(PublicCompany), [National Home Health Care Corp.](#)(PublicCompany), [OhioHealth](#)(Company), [St. Luke's Episcopal Hospital](#)(Company), [Sunquest Information Systems, Inc.](#)(PublicCompany)
- [Knowledge Base](#) (673) including: [National HealthCare Corporation](#)(PublicCompany), [Home Health Care Services](#)(IndustrySector), [Home Health Care Services](#)(IndustrySector), [National Home Health Care Corp.](#)(PublicCompany), [OhioHealth](#)(Company)

Fig. 18.9

Meta-results page

Document: Helping employees stay healthy

Abstract: About 240 big companies (which together provide health insurance for more than 50 million Americans) have banded together to create a nonprofit organization called the **National Business Group on Health**. It advises large employers on health-care and benefits issues and has started bestowing a new honor - the **Best Employers for Healthy Lifestyles** award - on companies that are

Author: **Fish** National Business Group, Inc. is a Company located in United States (North America). Its webpage is <http://www.nbg.com>. National Business Group (NBG) knows it often takes networking to get a job. The systems

Topics: **Em** integrator specializes in LANs, WANs, remote access, and network security. NBG also resells networking and

Date Published: **Aug** connectivity equipment and offers network analysis, consulting, design, installation, technical support, and training services. The company purchases from such suppliers as Cisco Systems, and Microsoft, and Nortel

Full Text: **EV** Networks. NBG's customers include businesses in industries such as education, finance, manufacturing and technology as well as Federal, state and local governments. CEO **for** Richard Basich, who owns National Business Group, **fin** founded the company in 1987 with Jeff Malone, who heads **the** the company's sales efforts. Contact on: <http://www.nbg.com>. **iss** **con** Expenditures: are in the U.S. and of how worried e seen the fallout close to home, in the nd copayments. But corporate **America** is s (which together provide health insurance r to create a nonprofit organization called **Employers on health-care and benefits Employers for Healthy Lifestyles award-on thier.**


Some of the winning tactics include installing walking routes and hiking paths around workplaces and stocking cafeterias and vending machines with more fruit and other foods containing less fat and salt. NBGH hopes that highlighting such best practices will encourage other employers to copy them. Additional praised tactics:

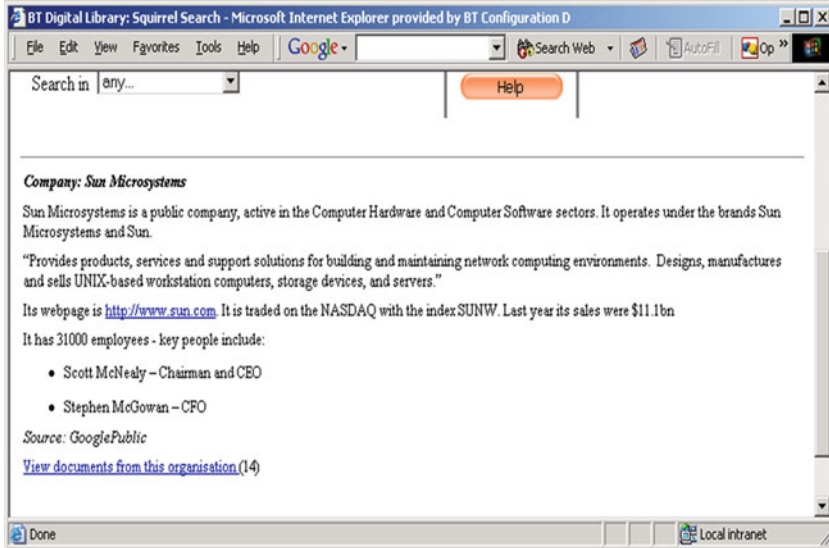
Fig. 18.10

Document view

input and produces natural language text, tailored to the presentational context and the target reader. NLG can be used to provide automated documentation of ontologies and knowledge bases and to present structured information in a user-friendly way.

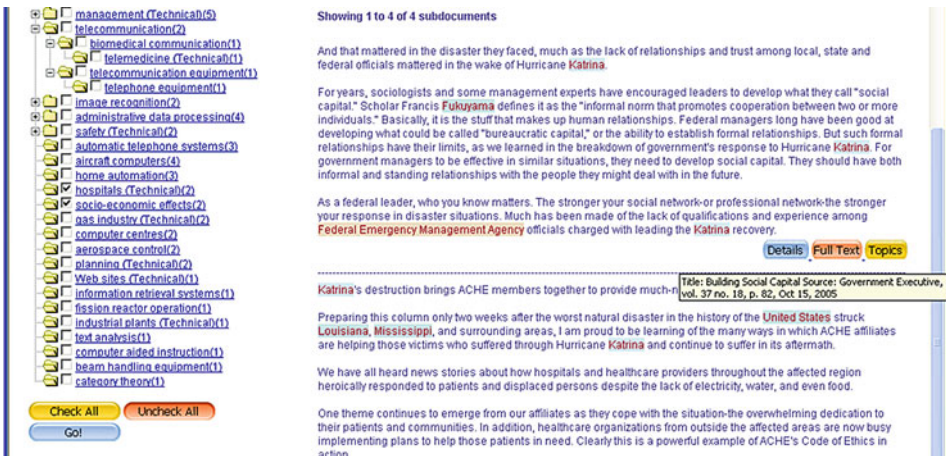
The summary displays information related not only to the entity itself but also information about related entities such as people who hold job roles with the company. This avoids users having to browse around the various entities in the ontology that hold relevant information about the entity in question.

Users can choose to view results as a consolidated summary (digest) of the most relevant parts of documents rather than a discrete list of results. The view allows users to read or scan the material without having to navigate to multiple results.  *Figure 18.12* shows a screenshot of a summary for a query for “Hurricane Katrina.” For each



■ Fig. 18.11

Entity view



■ Fig. 18.12

Consolidated results

subdocument in the summary, the user is able to view the title and source of the parent document, the topics into which the subdocument text has been classified or navigate to the full text of the document. The example of Squirrel shows that not only does semantic search offer the potential to improve search results, but also to improve the presentation of those results.

To gain an idea of how users perceive the advantages of semantic search over simply text-based search, Squirrel has been subjected to a three-stage user-centered evaluation process with users of a large Digital Library. Twenty subjects were used, and the perceived information quality (PIQ) of search results obtained. Using a seven-point scale the average (PIQ) using the existing library system was 3.99 compared with an average of 4.47 using Squirrel – a 12% increase. The evaluation also showed that users rate the application positively and believe that it has attractive properties. Further details can be found in [32].

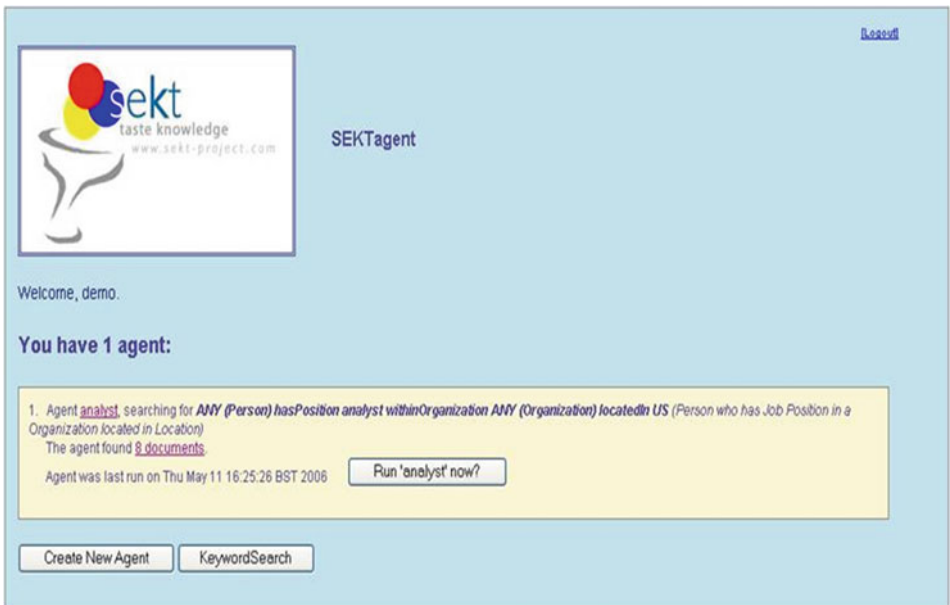
18.2.1.2 SEKTagent: A Different View on Semantic Search

Another approach to enabling semantic queries is exemplified by SEKTagent [29].

► *Figure 18.13* illustrates the basic approach by showing the following semantic query:

“ANY (Person) hasPosition analyst withinOrganization ANY (Organization) locatedIn US”

The query is looking for someone who is an analyst working in any US organization. This is quite different from a text query. Everything is stated at a conceptual level. The most concrete entity in the query is “US.” However, even this is not treated as a text string. The query may find a document referring to an analyst working in some city or state of USA, but not containing any reference itself to USA. The system makes use of the geographical knowledge in the knowledge base to determine that this is a relevant document.



The screenshot shows the SEKTagent web interface. At the top left is the logo for 'sekt taste knowledge www.sekt-project.com'. To the right of the logo is the text 'SEKTagent'. Below the logo is the text 'Welcome, demo.' and 'You have 1 agent:'. A yellow box contains the following text: '1. Agent analyst, searching for *ANY (Person) hasPosition analyst withinOrganization ANY (Organization) locatedIn US (Person who has Job Position in a Organization located in Location)*. The agent found 8 documents. Agent was last run on Thu May 11 16:25:26 BST 2006'. To the right of this text is a button labeled 'Run 'analyst' now?'. At the bottom of the interface are two buttons: 'Create New Agent' and 'KeywordSearch'.

■ **Fig. 18.13**

A semantic query in SEKTagent

for such tasks, said Gartner analyst Kimberly Harris-Ferrante. However, she said
nce company's lack of experience with large outsourcing deals.

■ **Fig. 18.14**

Extract from one of the results of a semantic query – showing entities in the knowledge base highlighted

► *Figure 18.14* shows an extract from one of the retrieved documents. Entities in the knowledge base are highlighted. In this case, there are three such entities: Gartner; analyst; Kimberley Harris-Ferrante. The first of these is a company, the second a position in an organization; and the third is a person. In fact, Kimberley Harris-Ferrante is the analyst, working in a US organization, who satisfies this query.

Moving the mouse over any of these entities displays more information about them. In the case of Gartner, for example, it provides the key facts about the company. Rather than just displaying raw information, natural language generation technology is applied to the relevant information in the knowledge base to create text, which can be easily read.

The example illustrates another important feature which differentiates the ontology-based approach from that of relational databases. In a database, the only information which can be retrieved is that which is explicitly input into the database. An ontology-based system can make use of a reasoner to perform inferencing over the ontology and knowledge base. In the example, the request was for someone performing a specific role in an organization in the USA. The information in the knowledge base could well be that the organization is located in some part of the USA, e.g., a city or state. However, the knowledge base associated with PROTON also has geographical information including states and major cities in the USA. Armed with this information, it is able to make the necessary inferences.

It should be noted that to identify any named geographical region (such as a county, state, region, district, town, village, etc.) with a particular country is in the general case a hard problem. However, a subset of the problem can be solved based on the knowledge available in the ontology. For example, the PROTON ontology contains, for all major cities, a link to the country in which they are located. It is relatively easy therefore to identify a major city in the query and link it with the appropriate country. In other applications, more domain-specific information may be required; frequently, it may be possible to draw on information already in structured or semi-structured form and thereby reduce the need for manual intervention.

Additional examples of semantic search are given in [18].

18.2.1.3 Semantic Filing: TagFS and SemFS

► **Section 18.1.3** discussed the difficulty which many people have in finding information which they themselves have stored, often on their own computers. One reason for this is that there is often more than one location where a file can logically be stored; yet users are

in general restricted to storing information in a single location. A partial solution to this is the use of tags. However, this loses the advantage of being able to travel through the tree structure of a hierarchical set of folders.

TagFS [33] merges the two approaches to obtain the advantages of both by using the tags to create a folder structure, which is dynamic rather than fixed. In TagFS, the organization of the resource is divorced from its location. The file is simply tagged. To take the example from the reference, in a conventional filing system, a user saving music files would first establish a directory structure, for example, *year/artist/album*. This would be quite distinct from a structure *artist/album/year*. In TagFS these three attributes, and any other which are appropriate are merely used to motivate tags. To find a file, it does not matter in which order you traverse the “directory”; the “directory path correspondingly denotes a conjunctive tag query which results in a set of files that fulfill all tag predicates.”

Apart from overcoming the need to specify folders in a specific order, tagging has the advantage that the user does not need to reach the end of a folder path before finding the required file. In addition, new tags can be added to describe a file in a way which new folders cannot.

TagFS is implemented using the SemFS architecture. SemFS provides mapping from traditional file system interfaces to annotation of information objects using RDF. Rather than interpreting directory structures as static storage hierarchies, as in a conventional file system, they represent dynamic views on information objects. In fact, TagFS makes relatively simple use of SemFS, in that the latter offers an arbitrary number of different views, while TagFS simply employs one called “hasTag.” The use of RDF enables integration with other semantic desktop applications, as described in [▶ Sect. 18.2.3](#).

18.2.1.4 Commercial Activities

There are a range of companies in this area, with new companies joining some established ones. In the domain of semantic search, there are companies such as Hakia, PowerSet (now acquired by Microsoft), Siderean, and Ontotext. In the information and process integration space there are, for example, Metatomix and Ontoprise. Turning to social networking and knowledge management generally, a company which has attracted recent interest is Radar Networks. In 2007, they announced their Twine semantic social networks offering. Twine mined fora, wikis, databases, and online newsgroups to identify relationships which were then expressed in RDF. Recently Radar Networks were acquired by Evri, and currently Twine is not supported. Evri themselves offer a “discovery engine” which identifies the currently most popular stories and trends.

Larger, more established vendors are also active, including Oracle with RDF support in Oracle 10g and ThomsonReuters making all their information available with semantic markup via their OpenCalais (<http://www.opencalais.com/>) service, which parses text for names, locations, organizations, and other entities.

In the search sector, PowerSet, mentioned above, was acquired by Microsoft for \$100m. Microsoft is believed to have incorporated aspects of PowerSet’s semantic

technology into its Bing Search engine. Yahoo! and Google have been more explicit in their use of semantic technology: Yahoo!'s Search Monkey platform allows developers to exploit semantic data (in RDFa or microformats). The idea is to make Yahoo! Search results more useful and visually appealing, and thereby drive more relevant traffic to their sites. In addition to the possibility for developers to create their own enhanced results, Yahoo! already provides a standard enhanced result for those sites providing structured data. Google followed suit with a similar initiative, known as Rich Snippets.

18.2.2 Semantic Information Sharing

▶ [Section 18.1.4](#) identified the importance, particularly acute in large organizations of being able to share information among colleagues. This applies both to knowledge explicitly written down and to tacit knowledge. In the former case, the need is to identify a document; in the latter case a person.

18.2.2.1 Effective Document Sharing with Semantic Technologies

Using Taxonomies for Knowledge Sharing

One way to share documents is simply to use the corporate intranet as a repository and provide employees with an intranet search engine. As already noted, search technology is not always fully effective. Even with the kind of advanced search technology discussed in ▶ [Sect. 18.2.1](#), a relevant document may be missed. One solution to make it easier to find and reuse documents is to require the author of the document to associate metadata with it when committing the document to a repository. Typically, the metadata relate to an agreed taxonomy.

As already discussed, the problem with this approach is that it can be time consuming for an author to save a document to the repository. The time taken will depend on how familiar he or she is with the system and the taxonomy, and also on the nature of the document. Frequently the time required is an inhibitor and the document will not be saved. A means of overcoming this is described in [\[34\]](#). Machine-learning techniques are used to automatically suggest metadata to the user, who can accept the suggestion, or make amendments or additions. The metadata can then be used by other users to search and browse the repository. Since this requires knowledge of the taxonomy, the system also offers a natural language search which requires no prior knowledge on the part of the user of how the information is classified.

A commercial example of a taxonomic system which offers support to the user is provided by Teragram (<http://www.teragram.com/>). The system employs linguistic technology. For example, an administrator is able to create rules to define which documents fall into each category of a taxonomy tree. Alternatively, the administrator can assign initial documents to each category and the system can then automatically make further assignments.

Using Ontologies

Taxonomies are limited in their descriptive power to describing hierarchical relationships. Ontologies are much richer in what they can describe. They offer an obvious basis for describing, and hence sharing information.

However, because of this increased richness, ontologies are in general more complex, and hence their creation and maintenance may be more time consuming. This depends, of course, on the tools available and the application domain. Similarly, from the user's viewpoint, the ontological approach will often be more time consuming than the taxonomic one. Once again, the kind of semantic annotation techniques described in [Semantic Annotations and Retrieval: Manual, Semiautomatic, and Automatic Generation](#) of this handbook can be used to automate, or at least partially automate, this process. The user wishing to retrieve information is then able to use the semantic search and browse techniques described in [Sect. 18.2.1](#). Reference [35] describes an implementation of this approach in a digital library. Here annotation is at two levels. Firstly, sets of topics are used to describe documents. Topics can have sub- and super-topics, to create a lattice structure. As a design decision, for reasons of computational tractability, topics are implemented as instances, not concepts. As a starting point, schemas used by proprietary information providers (e.g., Inspec: <http://www.theiet.org/publishing/inspec/>) provided the topics. Machine learning was used to refine these topics and to automatically associate documents with topics. Secondly, using natural language techniques, named entities within documents are identified and associated with concepts. These concepts are drawn from, for example, geography and business and include country, city, company, CEO, etc. The association of instances to concepts is illustrated by color coding, using the KIM system described in [36].

The creation and management of ontologies is required for many applications of semantic technology and is a significant research topic in itself. An overview of available methodologies is given in [37], which also describes a methodology, DILIGENT, for creating and maintaining distributed ontologies. In common with other such methodologies, the approach employs ordinary users, domain experts, and experts in ontology design. The approach is distributed in that different users may have slightly different versions of the ontology. Users refine a shared ontology on the basis of their experience, and these refinements are then fed back, as appropriate, to the shared ontology.

Tagging and Folksonomies

In parallel to the use of taxonomies in enterprises, and research into the use of ontologies, the hobbyist and consumer world has adopted the use of informal tagging to describe all kinds of information and media objects. Such tags are said to constitute “folksonomies.” Like wikis, folksonomies are part of the phenomenon of Web2.0, in which consumers of information are also producers. Such folksonomies are commonly represented by “tag clouds,” in which character size, font, or color are used to represent how much the tag has been used. Flickr (<http://www.flickr.com>) is an example of a website for sharing photos which uses this approach. Delicious (<http://delicious.com>) is another example where tags are associated with bookmarked pages. The website displays not just the most popular bookmarks, but also the most popular tags.

Some organizations now use similar techniques to encourage knowledge sharing and a McKinsey survey of the use of Web2.0 in companies has shown that many executives do believe that these techniques provide real business benefit [38]. Folksonomies have the advantage over taxonomies and ontologies in that they are easy to use. They do not have the development and maintenance costs associated with the use of the taxonomies and ontologies, that is, the cost of creating the taxonomy or ontology and then creating and updating the associated metadata.

McKinsey considered a range of Web2.0 technologies, including videosharing, blogging, RSS, wikis, and tagging. They looked at three broad areas of application: within organizations, which has been the theme of most of this chapter; in their dealings with suppliers and partners, which is outside the scope of this chapter; and in their relations with customers, which is similarly outside of the chapter's scope. They asked respondents to quantify the business benefit of using Web2.0 tools for each of these three areas. They found the median increase in speed of access to knowledge to be 30% and the median increase in speed of access to internal experts to be 35%. Other benefits were reduced communication and travel costs and reduced time to market. Similar responses occurred when respondents were asked about the effect of Web2.0 on collaboration with partners and suppliers. Not surprisingly, high technology and telecommunication companies reported the highest benefits with "business, legal, and professional services" also reporting a high level of benefits and manufacturing and financial further behind. Even so, in all industry sectors over 50% of respondents reported at least one measurable benefit from using Web2.0 technologies.

However, folksonomies lack descriptive power. In general, they possess no structure, usually not even the hierarchical structure present in a taxonomy. Moreover, the problems of synonymy and polysemy occur here; the same tag may be used with different meanings, or different tags may be used with the same meaning. Compared with ontologies, folksonomies are even more limited. They do not permit automated reasoning, nor the kind of search and browsing techniques described earlier. In general, the user is free either to use a preexisting tag or to use a new tag. The former has the practical value of encouraging convergence on a reasonable number of tags. However, it may lead to the emergence of dominant tags, representing particular views, and discourage the creation of new tags which may better represent a concept.

Nevertheless, after the success of tagging in the hobbyist world, it was natural to investigate the same approach in the enterprise. IBM's Dogear [39] is a bookmarking system in which bookmarks can be tagged. Once created, tags can not only be used for searching and browsing, but also to support social networks. The IBM designers of Dogear specifically chose to use real names, rather than pseudonyms. It is therefore possible for other users to see who has bookmarked a particular document. Knowledge of the particular bookmarks browsed by a user provides information about the user's expertise, or at least interests. This, in turn, enables the creation of communities of interest and potentially the identification of experts.

Another approach is to combine the ease-of-use of the folksonomic approach with the greater power of taxonomies. Reference [40] describes a proof-of-concept system which

suggests tags to the user by automatically selecting terms from a taxonomy. The user is, however, free to use other tags, and these are fed back to suggest new terms for the taxonomy. The authors call this a “taxonomy-directed folksonomy.” When users type a tag, they are prompted by a thesaurus which suggests terms which match the term they have entered. In principle, users could be given a choice of thesauri for tagging.

Heymann and Garcia-Molina [41] have developed an algorithm which converts a tag cloud into an hierarchical taxonomy. The starting point is to create a *tag vector* for each tag, of dimensionality equal to the number of objects, and such that the component in each dimension is the number times the tag has been applied to a particular object. From this, the cosine similarity between tag vectors is used to calculate the similarity between tags. These similarities are used by the algorithm to create a taxonomy.

Other work has combined user tagging and an ontology-based approach to the classification of information [42]. The goal of this work was to share information, in the form of bookmarked Web pages, and also to enable users to gain an awareness of others’ interests and expertise. Web pages are automatically classified, on the basis of their content, according to a preexisting library ontology. They are also tagged informally by users. A persistent problem with tagging is that different users will use different tags for the same concept, and the same tag for different concepts. In this work, equivalences are learned between different users’ tags on the basis of the content tagged. Moreover, the system recognizes relationships between pages, so that the user can browse from one page to a set of related pages. A fundamental intuition of the work is that Web pages bookmarked and tagged by the user’s close colleagues are more likely to be of significance than those bookmarked and tagged by people unknown to the user. Each user, when bookmarking a page, has the option of sharing to “self,” “team” (i.e., close colleagues), “community” (wider group of colleagues), and “everyone.” This is taken account of when ranking related pages. For example, those shared to “team” by one of the user’s team-members is ranked higher than a page shared by the same person to “community.”

Another approach [43] has proposed creating an ontological structure by combining a purely statistical analysis of folksonomies with a number of additional techniques:

- Terminological resources like WordNet (<http://wordnet.princeton.edu/>) are used, for example, to identify equivalence between tags.
- This is augmented by using Web resources such as Google and Wikipedia. The former is used to suggest alternative spellings, on the basis of the number of occurrences of the various alternatives. Wikipedia can be used to identify new terms which may not occur in conventional dictionaries. Moreover, Wikipedia *URLs* can be regarded as identifiers for many concepts.
- Ontology matching techniques are used, for example, to identify “relationships between tags, between tags and lexical resources, and between tags and elements in existing ontologies.”
- The preceding automatic techniques are enhanced by human intervention, to confirm the results of the automatic techniques, and to obtain information which could not be obtained otherwise.

The Semantic MediaWiki


The Semantic MediaWiki (http://semantic-mediawiki.org/wiki/Semantic_MediaWiki) represents a different approach to combining the power of formal semantics with the ease-of-use associated with Web2.0 [44, 45]. It builds on the success of wikis in enabling collaboration. Specifically, Semantic MediaWiki is a free extension of MediaWiki, the software used by Wikipedia.

Whereas conventional wikis enable users to collaborate to create Web pages, the Semantic MediaWiki enables collaboration to create a knowledge base to complement the Web pages. Conventional wikis have links between pages; a page describing London might contain a sentence “London is the capital of the U.K.” and a link to a page describing U.K. Syntactically this is done by writing `[[U.K.]]`. In the Semantic MediaWiki, the user can explicitly associate a relation with a link; so that the link between the London page and the U.K. page can have the associated relation “is capital of.” This is done by extending the normal wiki syntax and writing `[[is capital of::U.K.]]`.

This is entirely informal, in the sense that the user is free to choose any relation he or she likes, represented by any phrase the user likes. Of course, there is value in people using the same terms, and they can be encouraged to reuse existing relations; it is also possible to define equivalences between different terminologies (e.g., “knows about” can be equated to “is expert in”).

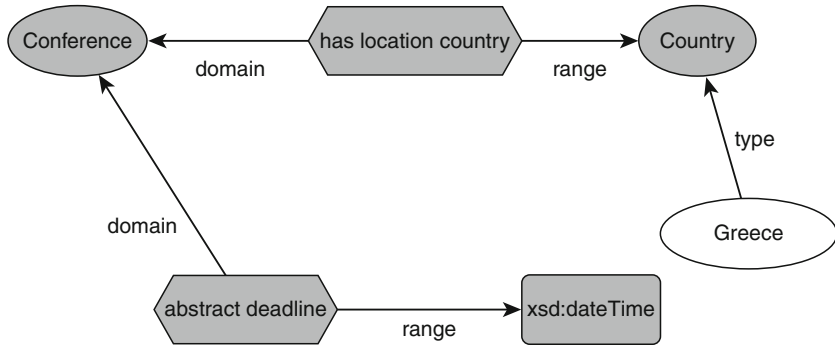
It is possible to use attributes to associate information with a page, other than that which can be represented by relations. For example, the U.K. page could have metadata associated with it describing its population. Syntactically, this can be achieved by writing `[[population: = 61,000,000]]`.

Once a knowledge base has been created using a Semantic MediaWiki, it can then be queried. This can be done using a syntax very similar to the annotation syntax. This is intended for use by the more computer-literate. However, the syntax can be used to create results pages (e.g., a table of the populations of various countries) which can be viewed by everyone. Alternatively, page authors can insert a query enclosed in the `<ask>` tag, so that the displayed page shows not the query but the result of the query.

More recently, an extension to the original Semantic MediaWiki enables forms-based input more suited to end users, see http://www.mediawiki.org/wiki/Extension:Semantic_Forms. Another initiative has generated the capability for nontechnical users of the wiki to create general queries in a relatively easy-to-use way, that is, without using a formal syntax. In this approach, textual queries can be translated into query graphs composed of concepts, relations, and instances in the ontology [46]. In the simplified example quoted in the reference, a user requires to know the deadline for submission to all (presumably forthcoming) conferences in Greece. He or she types the query string “conference Greece deadline.” The resultant query graph is shown in  Fig. 18.15. This is, in effect, a representation of a SPARQL query. The user is then provided with an interface for amending the query graph. He might, for example, wish to change “abstract deadline” to “submission deadline.”

A Lightweight Ontology Editor

The approaches described in [39–41] all in some way draw on the tagging behavior of a user or group of users in order to create or enhance a taxonomy or ontology.



■ Fig. 18.15

Query graph derived from “conference Greece deadline”

The objective is to create a synergy between the formal and informal approaches to knowledge representation. Another way to achieve the same goal is to provide users with an easy-to-use ontology editor, restricted to creating and editing lightweight ontologies. By lightweight ontologies are meant ontologies with relatively limited features, but nevertheless powerful enough for generic knowledge management applications.

Reference [47] describes the use of such an approach to create an ontology editor for the Semantic MediaWiki. The system supports both the import and export of OWL ontologies, and also the import of folksonomies. The latter feature allows a folksonomy dataset to be mapped to an ontology representation. Imported tags are compared with WordNet and Wikipedia, as in [43]. Tags are clustered, mapped to the SKOS knowledge-organization ontology [48] and then mapped and inserted according to the SMW ontology. Additionally, knowledge repair functionalities are provided that assist users with the discovery and mitigation of redundancies and inconsistencies within the knowledge base.

18.2.3 The Semantic Desktop: Supporting the User Throughout His Work

18.2.3.1 Sharing Information and Metadata Across Applications and Desktops

► Section 18.1.5 noted the need for metadata, shared between applications and linked to the context of the user’s work and the processes he or she performs.

One early initiative to address this challenge was the Haystack project [49]. The Haystack project aimed to provide more flexibility in personal information management, and to give the user more control over how information is recorded, annotated, and manipulated. Haystack is now a group at MIT which “develops tools for the web and desktop that can flex to hold and present whatever information a user considers

important, in whatever way the user considers most effective” (<http://groups.csail.mit.edu/haystack/>).

The original version of Haystack preceded RDF, but later RDF was adopted by the project. More recently, the adoption of RDF and semantic technologies has led to an initiative known as the *semantic desktop*. In general terms, the goal of the semantic desktop is to link information objects on the desktop, and on shared servers, through a shared ontology in much the same way as the Semantic Web aims to semantically link objects on the Web. A good description of the early work, along with a number of early references is in [50]. An important aspect of the vision is to allow users to create their own mental models, through the shared ontology. The reference talks about “trails,” which are “paths of resources that build a personal look on a topic.” Another important aspect is the emphasis on a P2P philosophy, so that information objects across desktops are semantically linked. Underlying all this is the need for personal knowledge management tools that can “integrate heterogeneous sources taken from the Semantic Desktop,” which in turn requires ontology mapping techniques.

In Europe, during 2006–2008, the Nepomuk project (<http://nepomuk.semanticdesktop.org>) was a major focus for work on the semantic desktop [51]. Consistent with the previous discussion, the goal of Nepomuk was to link data, and metadata, across applications and across desktops, using shared conceptualizations expressed in RDF. Specifically, the project set out to provide “a standardized description of a Semantic Desktop architecture, independent of any particular operating system or programming language.” A reference implementation of this architecture has been developed, known as Gnowsis. More recently this name has been adopted by a semantic desktop startup, see <http://www.gnowsis.com>.

The project employs an ontology-based approach and uses the *Personal Information Model Ontology* (PIMO) [52], originally developed to represent desktop sources in the EPOS project, which ran from 2003 to 2005 (<http://www3.dfki.uni-kl.de/epos>). Such an ontology allows different applications to share data, while at the same time avoiding the “n:n” problem, that is, the data models for each application map to the PIMO. This is essentially the use of ontologies for data integration, as discussed in [Sect. 18.1.6](#) below. PIMO uses a layered ontology approach, providing generic upper- and mid-level ontologies, and also permitting domain ontologies to be constructed, for example, for a particular company, and leaving users to create ontologies more specific to their needs. This enables users to create their own mental models building on a preexisting base. It avoids the so-called cold start problem where a lack of initial content deters use of the system and the construction of further content. Because the creators of PIMO did not believe that rules or description logic is required for personal information models, modeling is done in RDFS, rather than OWL. The model integrates some third-party ontologies such as the “Friend of a Friend” (FOAF) ontology (<http://www.foaf-project.org/>).

One of the products of Nepomuk was the SPONGE (Semantic Personal Ontology-based Gadget) software tool [53]. The tool “supports users finding, retrieving and annotating desktop resources ... plus seamless access to internet information.” Some

information and interaction is available via a small gadget, taking up limited space on the user's screen. More information is available via the user's browser. The reference claims that future work will extend the functionality with collaborative features. These include the ability to access remote desktops in a P2P topology and workspaces which will facilitate the sharing of resources.

18.2.3.2 Understanding User Context

One of the early goals of the semantic desktop was to understand how the users' information resources divide into a number of contexts, and to detect when a user switches between contexts [49]. This would enable information to be presented to the user, taking account of his or her current context. A number of current projects are investigating this theme.

The APOSDLE project (<http://www.aposdle.tugraz.at/>) is aimed specifically at informal eLearning, that is, at providing the user with small chunks of learning material just when required [54, 55]. This requires understanding the context of the user's current work. For example, in one envisaged scenario the user's actions are analyzed to determine that, for example, he or she is in the starting phase of a project. The user is then provided with information and guidance relevant to project start activity. The project is developing a number of widgets to enable user interaction. These include a context selector; a widget which displays resources relevant to the current context; a global search widget; and a "main" widget which presents the current selected or detected context and possible learning goals. There is also a "cooperation wizard" to guide users through cooperation processes.

APOSDLE is ontology-based. The user creates three types of models: a domain model; a task model describing the tasks which need to be executed; and a learning goal model. Modeling tools are provided, including a semantic wiki and plug-ins for the ontology editor Protégé. The user can also annotate parts of documents using the domain model.

A parallel but separate activity, involving some of the same researchers as in APOSDLE, is also developing a system for task detection [56]. The system is known as UICO, loosely an acronym from "an ontology-based User Interaction COntext model for automatic task detection on the computer desktop." The objective of this work is more general than eLearning, but much of the approach is similar to APOSDLE. An ontology-based user context model has been developed. The model is inspired by the Personal Information Model Ontology discussed above. The ontology, and modeling done in the ontology, form an input to the system's task detection software. To achieve this task detection, the project has developed the concept of the "semantic pyramid." At the bottom layer are events, resulting from "single user interactions with the computer desktop." Above this are event blocks, which are "sequences of events that belong logically together." At the top are tasks, which are "well-defined steps of a process, that cannot be divided into subtasks, and in which only one person is involved." Thus, from the user's viewpoint (rather than the computer's) tasks are essentially atomic. Key to the application of this concept is the delivery of resources relevant to the user's actions.

Another related project is ACTIVE (<http://www.active-project.eu>) [57]. ACTIVE has three main research themes:

- Information delivery guided by user context; this entails the system being able to detect a user's current context.
- The creation of informal processes by users, and the learning of these processes through observation of the user's interaction with his or her computer. By "informal" processes are meant processes designed by individuals to achieve their work-related goals, rather than the formal processes designed on behalf of the organization.
- Knowledge sharing through the synergy of an informal (Web2.0) and an ontology-based approach.

ACTIVE sees context and process as often orthogonal. For example, two of the case studies in the project (e.g., see [58]) are concerned in part with customer-facing people who spend a significant amount of time writing customer proposals. For these users, context will often, but not necessarily, equate to customer. The process, on the other hand, is that of writing a customer proposal, which can be enacted in a number of contexts (i.e., for different customers). As noted above, ACTIVE is seeking to identify both the user's context and his or her current process. The project has developed something similar to the semantic pyramid of UICO. Events as recognized by the machine level need to be combined through various stages to create an understanding of the processes which are intelligible to the users.

ACTIVE aims to impose a minimum of overhead on the user. The user is able to specify his or her set of contexts and to associate information objects with contexts. However, the project is also researching both how to automatically associate information objects with particular contexts and also learn contexts. The latter is a problem in unsupervised learning, that is, how on the basis of the user's actions and the information objects he or she accesses, can those information objects be partitioned into a set of contexts.

Contexts can be shared, that is, a group of users can share the same context; this encourages the sharing of information. Processes can also be shared. This encourages process reuse and also process improvement as colleagues are able to review and improve each others' processes.

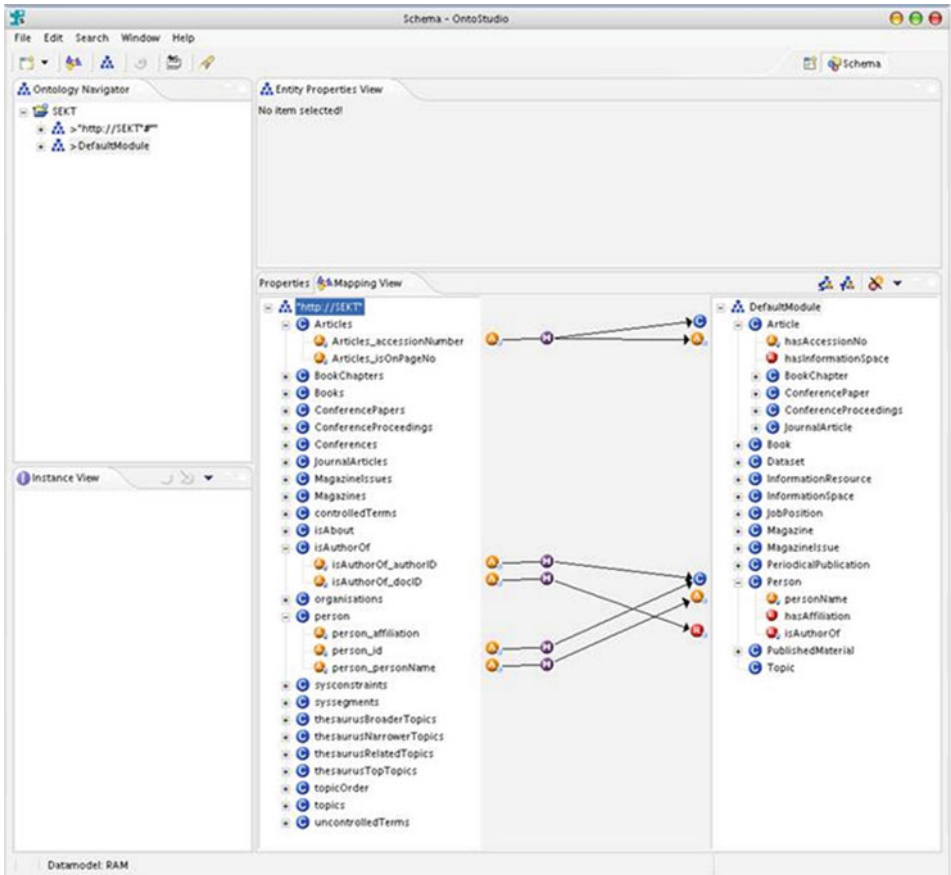
The third theme of ACTIVE is knowledge sharing. This includes continued development of the Semantic MediaWiki and the lightweight ontology editor discussed in [▶ Sect. 18.2.2](#). The goal here is to make use of ontologies in knowledge management, so as for example to be able to exploit reasoning, but in a way which is sufficiently user-friendly for casual, nonspecialist, users.

18.2.4 Graphical and Semiautomatic Approaches to Information Integration

The approach of [▶ Sect. 18.1.6](#) reduces the number of mappings needed, but they still do have to be created. One way to create mappings is to use a mapping language. This is fine

for specialist knowledge engineers but others need a more natural and intuitive approach which is easy to learn and use. A number of graphical mapping tools have been created for such users. One such has been developed by ontoprise GmbH (<http://www.ontoprise.com>) as part of their OntoStudio ontology engineering environment.

Simple drag-and-drop functionality is used to create and amend mappings. At the same time, the system undertakes consistency checks to ensure that the user's actions make sense. ▶ [Figure 18.16](#) shows a view of the mapping tool. The left- and right-hand side shows portions of two different ontologies, and the mappings are represented by lines between them. Mappings can even be conditional. Consider, for example, a mapping between two national transport ontologies. The definition of a “truck” differs in different countries, depending in some countries on the weight of the vehicle. This can be taken into account when creating the mapping.



■ **Fig. 18.16**

Ontology mapping in OntoStudio. Courtesy: ontoprise GmbH

Even greater gains can be achieved by automating, at least partially, the process of creating the mappings. This is an area of current research. A starting approach is to look for similarities in the text strings used to denote data fields by different schemas, for example, phone for telephone. This can even take account of different representations of similar sounds, for example, the use of “4” to represent “for.” Such an approach is frequently called syntactic matching. Some appreciation of semantics can be introduced by using a thesaurus, such as WordNet, to identify synonyms. Semantic matching can go further by taking account of the structure inherent in two schemas. For example, a product classification system can in general be represented as a graph. Structural similarities then enable the software to draw reasonable conclusions about the relationship between nodes (i.e., categories of products) in two classification systems. The software may propose equivalences between categories, or that a category in one system is a subset of a category in the other classification. Readers interested in the technical detail of one approach, based on the use of a form of logic known as propositional calculus, are referred to reference [59]. For a relatively recent overview of the state of the art in the area of ontology mapping generally, see [60].

Once these techniques have been used to create an initial mapping, it can then be loaded into a graphical editing tool and refined manually.

The end result is that it is possible to integrate heterogeneous databases, and provide the knowledge worker in an organization with a unified view across these databases. This is an important step in reducing the risk of significant information not being available, be it to better inform management decisions or to satisfy regulatory disclosure requirements.

18.2.5 Extracting and Exploiting Semantics from Unstructured Information

▶ Section 18.1.7 identified the need to analyze text so as to create structured knowledge and merge with existing structured knowledge in, for example, relational databases. This section discusses some tools to help achieve this.

18.2.5.1 Software for Text Analytics

▶ Section 18.1 discussed the two approaches to creating metadata; one based on statistics and machine-learning and one based on an analysis of language syntax and grammar known as natural language processing (NLP). The term *text analytics* is used to describe both approaches.

The statistical and machine-learning approach is well represented by the Text-Garden suite of software tools (<http://kt.ijs.si/software/TextGarden/>) developed within the Jozef Stefan Institute in Ljubljana, Slovenia, and used within the SEKT project described earlier [61]. Text mining techniques are also provided as part of the open-source data mining software, Rapid Miner, which is available on SourceForge and supported by Rapid-I GmbH, <http://rapid-i.com>.

The NLP approach is represented by GATE, developed by the University of Sheffield in the U.K., and used in the SEKT project (<http://gate.ac.uk/>); and also by UIMA, originally developed by IBM. An early introduction to GATE is given in [62]; a slightly later, more comprehensive overview is given in [63] GATE is also covered in [▶ Semantic Annotations and Retrieval: Manual, Semiautomatic, and Automatic Generation](#). GATE provides an environment for creating NLP applications. It combines three aspects; it is an architecture, a framework, and a development environment for language engineering. GATE is open and includes a set of resources which others can use and extend. The architecture separates low-level tasks (e.g., data storage, data visualization and location, and loading of components) from data structures and algorithms. The framework provides a reusable design plus software building blocks. The development environment provides tools and a GUI for language engineering. It also provides an interface for text annotation, in order to create training corpora for machine learning algorithms. By an analysis of grammatical structures, such software can, for example, perform named entity recognition and deduce, with reasonable accuracy, to what nouns particular pronouns refer. Such applications are the basis for the semantic search techniques discussed in [▶ Sect. 18.1.3](#) and for the information extraction from text discussed in this section.

UIMA (an acronym for *Unstructured Information Management Applications*) [64] also provides an architecture for the analysis of unstructured text. Having originally been developed by IBM, it is now being developed by the standards body OASIS (<http://www.oasis-open.org>). Apache UIMA is an Apache-licensed open-source implementation of the UIMA specification, see <http://uima.apache.org>. The principle of UIMA is that applications are decomposed into components. The UIMA framework defines the interfaces between these components and manages the components and the data flows between them. As noted in the reference above: “The principal objective of the UIMA specification is to support interoperability among analytics.” This is divided into four design goals:

- Data representation – supporting the common representation of artifacts and metadata
- Data modeling and interchange – supporting the platform-independent interchange of artifacts and metadata
- Discovery, reuse, and composition of independently developed analytics tools
- Service-level interoperability – supporting the interoperability of independently developed analytics based on a common service description and associated SOAP bindings

GATE and UIMA are overlapping in scope and an interoperability layer has now been created between them; one view sees GATE’s advantage as a prototyping tool while UIMA’s advantages are in performance and scalability [65].

18.2.5.2 Extracting Information from the World Wide Web

The previous discussion has assumed that the information to be integrated resides within an enterprise. The rise of the World Wide Web has provided one motivation for

combining data from outside the enterprise. An approach to achieving this is described in [66]. Here an ontology is used to provide a view across information on the Web. In the future world of the Semantic Web, much information on the Web will be described using ontologies, and the problem will be to map from these into an overarching one. Today data on the Web exist in variety of forms, for example, unstructured or semi-structured HTML files. The first step is frequently to extract the desired data and to describe them in terms of the ontology. The next step is to undertake instance matching, that is, to identify equivalent instances. The paper proposes a scalable approach based on the use of a group of peers. However, more relevant to the interests of this handbook is the use of similarity metrics to construct the mappings between instances. The authors investigated three sets of features to characterize similarity: character level, word level, and ontological level. The first of these is determined by the number of character transformations to edit from one string to another (the so-called Levenshtein distance [67]) and the second is based on the “bag of words” approach common in information retrieval. The ontological similarity attempts to measure the distance between two concepts. For example, at the extremes, if two concepts are the same the distance is 0, while if they are disjoint the distance is infinite (represented in practice by a very large positive number). If two instances are known to instantiate two concepts, then intuitively the larger the concept distance, the smaller the probability of these instances being the same. The paper reports an experiment in which a method incorporating all three approaches had higher precision than other methods at “almost” all recall levels; although the higher the recall the less advantageous the incorporation of the ontological approach.

18.2.6 Sharing Information Across Organizations

▶ Section 18.1.8 talked about the need for shared vocabularies where organizations need to collaborate, and noted the problems which arise because such vocabularies are frequently informally defined. Two approaches to sharing data were noted. On the one hand, within a given domain, existing informal vocabularies can be formalized. This is the approach discussed in ▶ Sect. 18.2.6.1, where medicine is taken as an example.

On the other hand, where one is starting from scratch, self-describing datasets can be made available on the Web, and linked as appropriate. Where these datasets are made openly available, this creates a Web of linked open data. This approach is described very briefly in ▶ Sect. 18.2.6.2; much more detail on this topic is provided in ▶ [Semantic Annotation and Retrieval: Web of Data](#), this volume.

18.2.6.1 An Example from Medicine

In medicine and biology, the vocabularies are often very large and complex. This was a natural area, therefore, for the early application of ontologies. In fact, the most well known of all ontology tool suites, Protégé (<http://protege.stanford.edu/>), was originally

motivated by the needs of medical informatics, and this domain continues to influence its development. Today, there are a very large number of biomedical ontologies. Reference [68] gives a brief introduction, making the case for the development of virtual ontology repositories which could be browsed by potential users looking for an appropriate ontology, prior to downloading.

In the area of clinical medicine, the best-known example of a shared ontology is SNOMED-CT (Systematized **N**omenclature of **M**EDicine – **C**linical **T**erms). This was created, in 2002, by the merger of SNOMED-RT (Reference Terminology) from the College of American Pathologists and the UK National Health Service Clinical Terms. It is now maintained by the International Health Terminology Standards Development Organization (IHTSDO, see: <http://www.ihtsdo.org/>). SNOMED-CT is a very large vocabulary; by August 2008 it had 283,000 concepts.

SNOMED-CT was not originally designed as an ontology. However, as ontologies were being discussed in knowledge management, medical informatics was an obvious candidate for their application. Reference [69] is an early paper discussing how ontologies could be relevant to medical vocabularies such as SNOMED. The paper saw ontologies being applied in medicine to areas such as natural language processing, that is, to conceptualize language and serve an “interlingual” role; and supporting simulation and modeling, for example, in molecular biology; and knowledge sharing. They also note the difficulty that medical concepts are empirical rather than being perfectly defined. All this, of course, applies to many other specialist areas. The authors also lay down some principles for creating well-formed ontologies; this again is applicable to any domain area, not just medicine.

Despite not being originally conceived as an ontology, SNOMED adopted description logic as its representation language. Moreover, since the development of OWL1.1 it has been possible in principle to translate SNOMED into OWL. A discussion of what is involved in this is given in [70]; the barriers to achieving this are largely due to the size of SNOMED.

One valuable feature of the description logic approach is that of “post-coordination.” Whilst as already noted, SNOMED has a very large number of defined concepts, post-coordination helps reduce the number required. Post-coordination means that new concepts can be created from preexisting concepts, for example, by a clinician. Automatic consistency checking is required at the time the new concept is created.

18.2.6.2 The Web of Linked Data

The WWW as it has initially evolved is a Web of interlinked documents. Berners-Lee’s original vision, however, went beyond this to a parallel Web of Data. That this vision was not realized at the same time as the Web of Documents was probably due, at least in part, to the lack of finalized standards to describe data in the early years of the millennium, for example, the RDF standard was not finalized until 2004. However, in 2006 Berners-Lee returned to the subject of the Web of Data by publishing a set of principles for linked data [71].

The four principles which Berners-Lee enunciated are:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs so that they can discover more things.

The first two of these are familiar as a foundation for the Semantic Web. The third means that data will be self-describing. Finally, the fourth is a basis for the WWW, or indeed for any Web; through interconnectivity crawlers can discover all the data available.

The result is now called the Web of Linked Open Data, and is the subject of a W3C taskforce (<http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>). By July 2009, the Web of linked open data contained 6.7 billion triples and 149 million links [72].

As already noted, [Semantic Annotation and Retrieval: Web of Data](#) provides detailed information about the Web of Data, including descriptions of application areas; linked open data sources; and some of the available tools, such as Web of Data search engines.

Linked open data is provided by organizations who want to make information publicly available, for example, because they are government organizations with a mandate to do so, or because the information is about products which they market. Many of the consumers of such data will in turn make data openly available themselves. However, there are two broad scenarios where commercial organizations can make use of linked open data: firstly, to link internal and external data sources to add value to internal “own-use” applications and secondly, to build applications for customers based on public data. In the first category, one can imagine, for example, the use of public demographic data to enhance targeted marketing applications. In the latter category, an example would be to offer personalized location-based services by accessing public data about a particular location.”

The principles of linked open data and the technology developed for linked open data could equally well be used within organizations, to create data intranets, or between organizations to create data extranets. Again by analogy to the Web of Documents, links from these data intranets and extranets could reach out to the open data web, while links in the reverse direction would not, of course, be traversable.

18.3 Related Resources

An extensive list of references is given in the reference section. This section offers a non-exhaustive list of some of the key resources on the application of semantic technology to knowledge management.

“*Ontologies for knowledge management,*” Abecker, A., & van Elst, L. in “*Handbook on Ontologies,*” Studer, R. & Staab, S. (eds), Springer-Verlag, 2003.

This book chapter offers an excellent brief survey of the role ontologies can play in knowledge management systems. KM and the requirements on IT systems are introduced. The areas where ontologies can play a part in meeting those requirements are then discussed. An analysis of future practice and research, and the outlook for future trends and developments in ontology-based KM systems are given.

“*Towards the Semantic Web: Ontology-driven Knowledge Management*,” Davies, J., Fensel, D. & van Harmelen, F. (eds), Wiley, Chichester, UK, 2003.

Based on results from the OnToKnowledge project (one of the first European research projects looking at the relationship between semantic technology and knowledge management), this book covers basic research, tools, and case studies in ontology-driven knowledge management and offers a good overview of early work on this topic.

“*Ontologies for Knowledge Management: An Information Systems Perspective*,” Jurisica, I., Mylopoulos, J., Yu, E., *Knowledge and Information Systems*, Vol 6, No 4, Springer, London, 2004.

This paper surveys approaches to knowledge representation in Computer Science and categorizes them into four ontological categories: static ontologies, dynamic ontologies, intentional ontologies, and social ontologies. The benefits and drawbacks of the ontological approach are also discussed and the use of ontologies motivated at a foundational level.

“*Information Integration with Ontologies*,” Alexiev, A., Breu, M., de Bruijn, J., Fensel, D., Lara, R. & Lausen, H., Wiley, Chichester, UK, 2005.

This book describes how ontology technology can be used to manage dispersed, heterogeneous information assets more efficiently. The book compares the ontological approach with current EAI technology. One strength of the book is that examples are taken from an industrial application using real data sources from the automotive sector.

“*Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies*,” Davies, J., Grobelnik, M. & Mladenic, D., Springer, Berlin, 2009.

This book presents a framework, methods, and tools for semantic knowledge management, which it defines as the use of semantic technology for improved management of tangible knowledge assets. An interdisciplinary approach is advocated and discussed involving the use of knowledge discovery, ontology management, and human language technologies. Applications using the underlying technologies are described, along with a series of evaluated case studies showing the value of the semantic approach to KM in real-world settings.

18.3.1 Semantic Web Interest Group: Case Studies and Use Cases

The Semantic Web Interest Group (<http://www.w3.org/2001/sw/sweo/>), which has now closed, has produced a wide range of case studies and use cases, see <http://www.w3.org/2001/sw/sweo/public/UseCases/>. Here case studies refers to deployed systems while use cases refers to prototypes. They can be sorted along a number of dimensions, including application area and technologies used.

18.4 Future Issues

18.4.1 Web2.0 and Ontologies

The success of the informal Web2.0 techniques, discussed in [Sect. 18.2.2.1](#), is a challenge to semantic technologies. Can this success be further strengthened by combining these techniques with more formal techniques? This has been discussed in some depth. However, significant challenges remain.

The creation and use of ontologies needs to be simple and intuitive. It needs to be recognized that there are different constituencies to be catered for. There are some users who should not be aware of the existence of an ontology (or even what the word means), but who need straightforward tagging features, with software automatically creating and using an underlying ontology. At the other extreme there are power users, perhaps professionals in biomedical research, who will want to interact directly with the full power of ontologies – although even for them all interfaces should be as simple as possible, nothing should be more complex than it needs to be. There may be grades of users in between, requiring to understand something about ontologies and interact directly with them; although the language of ontologies may be too off-putting and other terminology may be more appropriate. There will also be people akin to database administrators who will create and maintain ontologies. Again, there will be a range of such people, depending in part on the nature of the applications. Some will have little formal training in IT; others will be IT professionals. The tools offered need to reflect this.

As far as is possible, the creation and maintenance of ontologies needs to be automatic. This requires the use of techniques from information retrieval (e.g., based on the “bag of words” approach) and natural language processing. There may be scope for combining these two approaches to provide increased user functionality. There are also user interface issues here. For example, there is a need to understand to what extent the process of metadata creation can be entirely automated and to what extent the user needs to confirm suggestions; and how this can be done in an unobtrusive way.

18.4.2 Integrating into and across Enterprises

McKinsey claim that “successful companies not only tightly integrate Web2.0 technologies with the workflows of their employees but also create a “networked company,” linking themselves with customers and suppliers through the use of Web2.0 tools” [38]. This highlights two challenges for applying semantic technology in the enterprise.

Firstly, there is a need to refine technologies such as those of the semantic desktop discussed in [Sect. 18.2.3](#) which integrate metadata across applications and with workflows. Integration of metadata with informal workflows created by information

system users, not just the formal ones created by the organization, is important to enable tools for improved productivity.

Secondly, building on the use of semantic technology to overcome heterogeneity within the organization, there is a need to use these technologies to address the even greater heterogeneity which exists when organizations work closely together. This may be in a supply chain or with customers; it may be for a relatively long period, or it may require that a collaboration infrastructure be created rapidly, used for a few months, and then withdrawn. Improved ontology mapping techniques will be required. As with the generation of automatic metadata, the need is to understand how to combine automatic and manual mapping techniques; and how to do this in a way which is natural for users who may not be IT professionals. Uschold and Gruninger [28] propose a methodology for making progress in research into achieving interconnectivity. They believe that working systems will require many assumptions and that research progress will be made by relaxing these assumptions one by one. Examples of such assumptions include use of a single ontology language, use of a single shared ontology, or use of a single shared upper ontology with distinct domain ontologies.

18.5 Cross-References

- eBusiness
- eGovernment
- eScience
- Future Trends
- Multimedia, Broadcasting and eCulture
- Ontologies and the Semantic Web
- Semantic Web Search Engines
- Social Semantic Web

References

1. Drucker, P.: Knowledge-worker productivity: the biggest challenge. *Calif. Manag. Rev.* **41**, 79–94 (1999)
2. Ackoff, R.L.: From data to wisdom. *J. Appl. Syst. Anal.* **16**, 3–9 (1989)
3. Chadran, A.: Economist Intelligence Unit: enterprise knowledge workers: understanding risks and opportunities. www.eiu.com/knowledgeworkers (2007). Accessed 29 Dec 2010
4. Jansen, B.J., Spink, A., Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manag.* **36**(2), 207–227 (2000)
5. Davies, J., Kiryakov, A., Duke, A.: Semantic search. In: Goker, A., Davies, J. (eds.) *Information Retrieval: Searching in the 21st Century*. Wiley, London (2009)
6. Russell-Rose, T., Stevenson, M.: The role of natural language processing in information retrieval. In: Goker, A., Davies, J. (eds.) *Information Retrieval: Searching in the 21st Century*. Wiley, London (2009)

7. Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., Robbins, D.: Stuff I've seen: a system for personal information retrieval and re-use. In: Proceedings of 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), Toronto. ACM Press, New York (2003)
8. Kiryakov, A.: Ontologies for knowledge management. In: Davies, J., Studer, R., Warren, P. (eds.) *Semantic Web Technologies: Trends and Research in Ontology-Based Systems*. Wiley, Chichester (2006)
9. Jones, W.: Finders, keepers? The present and future perfect in support of personal information management. *First Monday* 9(3) (2004). http://131.193.153.231/www/issues/issue9_3/jones/index.html
10. Jones, W.: *Keeping Found Things Found*. Morgan Kaufmann, San Francisco (2007)
11. Civan, A., Jones, W., Klasnja, P., Bruce, H.: Better to organise personal information by folders or by tags? The devil is in the details. In: Paper Presented at the 68th Annual Meeting of the American Society for Information Science and Technology (ASIS&T 2008). Columbus. <http://www.asis.org/> (2008)
12. Davenport, T.H.: Knowledge management case study; knowledge management at Ernst & Young. Information technology management white paper. <http://www.itmweb.com/essay537.htm> (1997). Accessed 29 Dec 2010
13. Ezingard, J., Leigh, S., Chandler-Wilde, R.: Knowledge management at Ernst & Young UK: getting value through knowledge flows. In: Proceedings of the 21st International Conference on Information Systems (ICIS 2000), Brisbane, pp. 807–822 (2000)
14. McComb, D.: *Semantics in Business Systems: The Savvy Manager's Guide*, Chapter 12. Morgan Kaufmann, San Francisco (2004)
15. Pollock, J., Hodgson, R.: *Adaptive Information: Improving Business Through Semantic Interoperability, Grid Computing, and Enterprise Integration*. Wiley-Interscience, Hoboken (2004)
16. Bernstein, P., Haas, L.: Information integration in the enterprise. *Commun. ACM* 51(9), 72–79 (2008)
17. Halevy, A., Ashish, N., Bitton, D., Carey, M., Draper, D., Pollock, J., Rosenthal, A., Sikka, V.: Enterprise information integration: successes, challenges and controversies. In: Proceedings of the ACM SIGMOD 2005 International Conference on Management of Data, Baltimore, pp. 778–787. ACM Press, New York (2005)
18. World Wide Web Consortium Semantic web use cases and case studies. <http://www.w3.org/2001/sw/swseo/public/UseCases/>. Accessed 29 Dec 2010
19. Doan, A., Noy, N., Halevy, A.: Introduction to the special issue on semantic integration. *SIGMOD Rec.* 33(4), 11–13 (2004)
20. Noy, N.: Semantic integration: a survey of ontology-based approaches. *SIGMOD Rec.* 33(4), 65–70 (2004)
21. Bernstein, P., Melnik, S., Petropoulos, M., Quix, C.: Industrial-strength schema matching. *SIGMOD Rec.* 33(4), 38–43 (2004)
22. Bernstein, P., Melnik, S.: Model management 2.0: manipulating richer mappings. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD/PODS 2007), Beijing, pp. 1–12 (2007)
23. Moore, C. (vice president and research director Forrester Research): Information indepth, Oracle. <http://www.oracle.com/newsletters/information-insight/content-management/feb-07/index.html> (2007). Accessed 29 Dec 2010
24. Murphy, B., Markham, R.: *eDiscovery Bursts Onto the Scene*. Forrester, Cambridge (2006)
25. AIIM: Electronic communication policies and procedures. A 2005 industry study prepared jointly by AIIM and Kahn Consulting Inc. <http://www.aiim.org> (2005). Accessed 29 Dec 2010
26. Foxvog, D., Bussler, C.: Ontologizing EDI: first steps and initial experience. In: Proceedings of the International Workshop on Data Engineering Issues in E-Commerce (DEEC 2005), Tokyo, pp. 49–58. IEEE Computer Society, Los Alamitos (2005)
27. Abecker, A., van Elst, L.: Ontologies for knowledge management. In: Staab, S., Studer, R. (eds.) *Handbook on Ontologies*, Chapter 22, pp. 435–454. Springer, Dordrecht (2004)
28. Uschold, M., Gruninger, M.: Ontologies and semantics for seamless connectivity. *SIGMOD Rec.* 33(4), 58–64 (2004)
29. Duke, A., Heizmann, J.: Semantically enhanced search and browse. In: Davies, J., Grobelnik, M., Mladenic, D. (eds.) *Semantic Knowledge Management*, pp. 85–102. Springer, Berlin (2009)

30. Bontcheva, K., Davies, J., Duke, A., Glover, T., Kings, N., Thurlow, I.: Semantic information access. In: Davies, J., Studer, R., Warren, P. (eds.) *Semantic Web Technologies: Trends and Research in Ontology-Based Systems*, pp. 139–169. Wiley, Chichester (2006)
31. Motik, B., Studer, R.: KAON2: a scalable reasoning tool for the semantic web. In: *Proceedings of the Second European Semantic Web Conference (ESWC 2005)*, Heraklion (2005)
32. Thurlow, I., Warren, P.: Deploying and evaluating semantic technologies in a digital library. In: Davies, J., Grobelnik, M., Mladenic, D. (eds.) *Semantic Knowledge Management*, pp. 181–198. Springer, Berlin (2009)
33. Bloehdorn, S., Görlitz, O., Schenk, S., Völkel, M.: TagFS – tag semantics for hierarchical file systems. In: *Proceedings of the Sixth International Conference on Knowledge Management (I-KNOW 2006)*, Graz (2006)
34. Franz, J., Traphöner, R.: Semantic Web for knowledge reuse in business processes. In: Davies, J., Grobelnik, M., Mladenic, D. (eds.) *Semantic Knowledge Management*, pp. 215–229. Springer, Berlin (2009)
35. Warren, P., Thurlow, I., Alsmeyer, A.: Applying semantic technology to a digital library. In: Davies, J., Studer, R., Warren, P. (eds.) *Semantic Web Technologies: Trends and Research in Ontology-Based Systems*, pp. 237–257. Wiley, Chichester (2006)
36. Bontcheva, K., Cunningham, H., Kiryakov, A., Tablan, V.: Semantic annotation and human language technology. In: Davies, J., Studer, R., Warren, P. (eds.) *Semantic Web Technologies: Trends and Research in Ontology-Based Systems*, pp. 29–50. Wiley, Chichester (2006)
37. Sure, Y., Tempich, C., Vrandečić, D.: Ontology engineering methodologies. In: Davies, J., Studer, R., Warren, P. (eds.) *Semantic Web Technologies: Trends and Research in Ontology-Based Systems*. Wiley, Chichester (2006)
38. Bughin, J., Chui, M., Miller, A.: How companies are benefiting from web 2.0. McKinsey (2009). <https://www.mckinseyquarterly.com/>. Accessed 29 Dec 2010
39. Millen, D., Feinberg, J., Kerr, B.: Social bookmarking in the enterprise. *ACM Queue* 3(9), 28–35 (2005). <http://researchweb.watson.ibm.com/jam/601/p28-millen.pdf>
40. Hayman, S.: Folksonomies and tagging: new developments in social bookmarking. In: *Proceedings of the Ark Group Conference: Developing and Improving Classification Schemes*, Sydney. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.138.8884&rep=rep1&type=pdf> (2007)
41. Heymann, P., Garcia-Molina, H.: Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report 2006-10, Stanford University. <http://ilpubs.stanford.edu:8090/775/> (2006)
42. Kings, N., Gale, C., Davies, J.: Knowledge sharing on the semantic web. In: Franconi, E., Kifer, M., May, W. (eds.) *Proceedings of the Fourth European Semantic Web Conference (ESWC 2007)*, Innsbruck. *Lecture Notes in Computer Science*, vol. 4519, pp. 281–295. Springer, Berlin (2007)
43. Van Damme, C., Hepp, M., Siorpaes, K.: FolksOntology: an integrated approach for turning folksonomies into ontologies. In: *Proceedings of the Fourth International European Semantic Web Conference (ESWC 2007) – Bridging the Gap Between Semantic Web and Web 2.0*, Innsbruck. *Lecture Notes in Computer Science*, vol. 4519. Springer, Berlin. <http://www.kde.cs.uni-kassel.de/ws/eswc2007/proc/FolksOntology.pdf> (2007)
44. Krötzsch, M., Vrandečić, D., Völkel, M.: Semantic MediaWiki. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L. (eds.) *Proceedings of the Fifth International Semantic Web Conference (ISWC 2006)*, Athens, GA. *Lecture Notes in Computer Science*, vol. 4273, pp. 935–942. Springer, Berlin/Heidelberg (2006)
45. Vrandečić, D., Krötzsch, M.: Semantic MediaWiki. In: Davies, J., Grobelnik, M., Mladenic, D. (eds.) *Semantic Knowledge Management*, pp. 171–179. Springer, Berlin (2009)
46. Haase, P., Herzig, D., Musen, M., Tran, T.: Semantic wiki search. In: *Proceedings of the Sixth European Semantic Web Conference (ESWC 2009)*, Heraklion. *Lecture Notes in Computer Science*, vol. 5554, pp. 445–460. Springer, Berlin (2009)
47. Luger, M., Wölger, S., Bürger, T.: SMW ontology editor – features. Internal report. STI Innsbruck, University of Innsbruck
48. World Wide Web Consortium SKOS simple knowledge organisation system. <http://www.w3.org/2004/02/skos/>. Accessed 29 Dec 2010

49. Karger, D.R.: Haystack: per-user information environments based on semistructured data. In: Kaptelinin, V., Czerwinski, M. (eds.) *Beyond the Desktop Metaphor: Designing Integrated Digital Work Environments*. MIT Press, Cambridge (2007)
50. Sauer mann, L., Bernardi, A., Dengel, A.: Overview and outlook on the semantic desktop. In: *Proceedings of the First Workshop on the Semantic Desktop (SemDesk 2005)*, International Semantic Web Conference (ISWC 2005), Galway (2005)
51. Groza, T., Handschuh, S., Moeller, K., Grimnes, G., Sauer mann, L., Minack, E., Mesnage, C., Jazayeri, M., Reif, G., Gudjonsdottir, R.: The NEPOMUK project – on the way to the social semantic desktop. In: *Proceedings of the Third International Conference on Semantic Technologies (I-Semantics 2007)*, Graz, pp. 201–211 (2007)
52. Sauer mann, L., van Elst, L., Dengel, A.: PIMO – a framework for representing personal information models. In: *Proceedings of the Third International Conference on Semantic Technologies (I-Semantics 2007)*, JUCS, Graz, pp. 270–277 (2007)
53. Papailiou, N., Christidis, C., Apostolou, D., Mentzas, G., Gudjonsdottir, R.: Personal and group knowledge management with the social semantic desktop. In: Cunningham, P., Cunningham, M. (eds.) *Collaboration and the Knowledge Economy: Issues, Applications and Case Studies*. IOS Press, Amsterdam (2008). ISBN 978-1-58603-924-0
54. Lindstaedt, S., Mayer, H.: A storyboard of the APOSDLE vision. In: *Poster submitted to the First European Conference on Technology Enhanced Learning (EC-TEL 2006)*, Crete (2006)
55. Musielak, M., Hambach, S., Christl, C.: APOSDLE contextualized cooperation. In: *ACM SIGCHI: ACM Conference on Computer Supported Cooperative Work 2008*. Electronic Proceedings: CSCW 08 [CD-ROM], San Diego. ACM Press, New York (2008)
56. Rath, A., Devaurs, D., Lindstaedt, S.: UICO: An ontology-based user interaction context model for automatic task detection on the computer desktop. In: *Proceedings of the First Workshop on Context, Information and Ontologies (CIAO 2009)*. ACM International Conference Proceedings Series, Heraklion (2009)
57. Warren, P., Kings, N., Thurlow, I., Davies, J., Bürger, T., Simperl, E., Ruiz, C., Gómez-Pérez, J., Ermolayev, V., Ghani, R., Tilly, M., Bösser, T., Imtiaz, A.: Improving knowledge worker productivity – the ACTIVE integrated approach. *BT Technol. J.* **26**(2), 165–176 (2009)
58. Warren, P., Thurlow, I., Kings, N., Davies, J.: Knowledge management at the customer front-line – an integrated approach. *J. Inst. Telecommun. Prof.* **3**(4), 8–15 (2009)
59. Bouquet P., Serafini L., Zanobini S.: Semantic coordination: a new approach and an application. In: *Proceedings of the Second International Semantic Web Conference (ISWC 2003)*, Sanibel Islands. *Lecture Notes in Computer Science*, vol. 2870, pp. 130–145. Springer, Berlin. <http://citeseer.ist.psu.edu/bouquet03semantic.html> (2003)
60. Euzenat, J., Shvaiko, P.: *Ontology Matching*. Springer, New York (2007). ISBN 3540496114
61. Mladenic, D.: Text mining in action! In: *From Data and Information Analysis to Knowledge Engineering: Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Magdeburg* (2005)
62. Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V.: GATE: an architecture for development of robust HLT. In: *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, pp. 168–175 (2002)
63. Bontcheva, K., Tablan, V., Maynard, D., Cunningham, H.: Evolving GATE to meet new challenges in language engineering. *Nat. Lang. Eng.* **10**(3–4), 349–373 (2004)
64. OASIS: *Unstructured Information Management Architecture (UIMA) version 1.0, Working Draft 05* (2008)
65. Roberts, I.: GATE and IBM’s UIMA – interoperability layer. http://videlectures.net/gate06_roberts_giul/ (2006)
66. Wang, C., Lu, J., Zhang, G.: An ontology data matching method for web information integration. In: *Proceedings of the Tenth International Conference on Information Integration and Web-based Applications & Services (iiWAS 2008)*, Linz (2008)
67. В.И. Левенштейн (1965) Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий

- Hayk CCCP 163.4:845–848. Appeared in English as: Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Doklady*. **10**, 707–710 (1966)
68. Noy, N., Rubin, D., Musen, M.: Making biomedical ontologies and ontology repositories work. *IEEE Intell. Syst.* **19**(6), 78–81 (2004)
69. Burgun, A., Botti, G., Fieschi, M., Le Beux, P.: Sharing knowledge in medicine: semantic and ontologic facets of medical concepts. In: *Proceedings of IEEE International Conference on Systems, Man and Cybernetics (IEEE SMC 1999)*, Tokyo, pp. 300–305 (1999)
70. Spackman, K.: An examination of OWL and the requirements of a large health care terminology. In: *Proceedings of the Third OWL: Experiences and Directions Workshop (OWLED 2007)*, CEURWS, Innsbruck (2007)
71. Berners-Lee, T.: Linked data – design issues. <http://www.w3.org/DesignIssues/LinkedData.html> (2006). Accessed 29 Dec 2010
72. Bizer, C.: The emerging web of link data. *IEEE Intell. Syst.* **24**(5), 87–92 (2009)
73. Davies, J., Studer, R., Sure, Y., and Warren, P.: Next generation knowledge management. *BT Technol. J.* **23**(3), 175–190 (2005)
74. Cimiano, P., Volker, J.: Text2Onto - A framework for ontology learning and data-driven change discovery. In: *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB 2005)*, Alicante (2005)