Information Systems Management, 25: 132–148

ISSN: 1058-0530 print/1934-8703 online
DOI: 10.1080/10580530801941058

Taylor & Francis
Taylor & Francis Group

Management Support with Structured and Unstructured Data—An Integrated Business Intelligence Framework

Henning Baars and Hans-George Kemper
Universität Stuttgart, Stutgart, Germany

Abstract *In the course of the evolution of management support towards corporate wide Business Intelligence infrastructures, the integration of components for handling unstructured data comes into focus. In this paper, three types of approaches for tackling the respective challenges are distinguished. The approaches are mapped to a three layer BI framework and discussed regarding challenges and business potential. The application of the framework is exemplified for the domains of Competitive Intelligence and Customer Relationship Management.*

Keywords business intelligence, data warehouse, unstructured data, content and document management, analysis systems

Motivation

The concept of "Business Intelligence" (BI) is increasingly gaining in visibility and relevance within the business realm (Gartner Group, 2006). Originally coined by Gartner Group as a collective term for data analysis tools (Anandarajan, Anandarajan, & Srinivasan, 2004), "Business Intelligence" is now commonly understood to encompass all components of an integrated management support infrastructure. The increased importance of such infrastructures reflects three interacting trends: more turbulent, global business environments, additional pressures to unveil valid risk and performance indicators to stakeholders, and aggravated challenges of effectively managing the more and more densely interwoven processes (Kemper, Mehanna, & Unger, 2004). To meet the respective requirements, traditional management support systems have evolved to enterprise-spanning solutions that support all managerial levels and business processes: Envisioned are infrastructures for business performance management approaches that involve strategic, tactical, and operational managers alike. This calls for seamlessly interconnected functionality that enables continuous business process monitoring, in-depth data analysis, and efficient management communication. (Kohavi, Rothleder, & Simoudis, 2002; Golfarelli, Rizzi, & Cella, 2004; Eckerson, 2006; Kimball, & Ross, 2002).

Address correspondence to Henning Baars, Betriebswirtschaftliches Institut, Lehrstuhl für ABWL und Wirtschaftsinformatik I, Breitscheidstr. 2c, 70174 Stuttgart, Germany. E-mail: baars@wi.uni-stuttgart.de

Rooted within the tradition of classical Management Support, BI applications usually revolve around the analysis of "structured data." Structured data is here understood to be data that is assigned to dedicated fields and that can thereby be directly processed with computing equipment. The most salient tools in the current BI discussion are still "reporting," "data mining," and "OLAP" tools, which are primarily directed to the presentation and analysis of numerical business data. Reporting systems prepare quantitative data in a report-oriented format that might include numbers, charts, or business graphics (Kemper et al., 2004). OLAP stands for "Online Analytical Processing" and denotes a concept for interactive, multidimensional analysis of aggregated quantitative business facts (like budgeted costs, revenue, and profit). OLAP tools give the user flexibility regarding the choice of dimensions that describe the facts of interest (e.g. product, time, customer), the excerpt of facts to be looked at (e.g. March to December) and the level of detail (e.g. store, ZIP code, county, nation, region) (Codd, E.F., Codd, S.B., & Salley, 1993). Data mining tools support the identification of hidden patterns in large volumes of structured data based on statistical methods like association analysis, classification, or clustering (Hand, Mannila, & Smyth, 2001).

For many application domains this is not satisfactory, though (Negash, 2004): Numerous information sources are unstructured or at best semi-structured, e.g., customer e-mail, web pages with competitor information, sales force reports, research paper repositories, and so on. Most of this information is provided in the form of

electronic documents, here understood in the broadest sense as self-contained content items.

Especially in areas that reach beyond company borders, like Customer Relationship Management (CRM) (Cody, Kreulen, Krishna, & Spangler, 2002) or Competitive Intelligence (CI) (Vedder, Vanecek, Guynes, & Cappel, 1999), it becomes imperative to consider both structured and unstructured data to provide valid insights into current business developments (Mertens, 1999; Kantardzic, 2003; Weiss, Indurkhya, Zhang, & Damerau, 2005; Negash, 2004). Moreover the results from BI based analyses are usually at some point translated into an unstructured form (e.g., a PDF file) for distribution and archival purposes—the handling of these procedures is still considered unsatisfactory in many larger organizations (Alter, 2003).

This all leads to the requirement to couple "classical" BI infrastructures for management support with systems that are specifically designed to handle, refine, and analyze unstructured data.

The following paper proposes and discusses an integrated framework that binds respective state of the art approaches together, and thereby provides a structure for BI infrastructures that enables holistic decision support.

There have been several publications on BI frameworks in the past. One class of frameworks is build around the concept of the "data warehouse" and focuses on the technical processing of structured data (e.g. Devlin, 1996; Kimball et al., 2002; Inmon, 2005). A second approach to structuring BI is to take a broad organizational and demand-driven view. This naturally leads to the requirement of incorporating unstructured data—but without focusing on concrete components and solutions for the relevant integration tasks (e.g. Negash, 2004). A third class of frameworks concentrates at providing a (partial) structure for a specific approach, e.g. by discussing architectures for the integration of documents into OLAP environments (e.g. Sukumaran, & Sureka, 2006; Sullivan, 2001). This last group of publications does not provide a complete framework for BI, but they provide valuable insights into concrete solutions. The objective of this paper is to discuss how the diverse specific approaches of the third class fit together and how they can be embedded within an integrated, conceptual BI framework.

The course of the paper is illustrated in Figure 1: Based on a literature review three integration approaches are distinguished and discussed regarding their respective business potential (Section 2). These approaches are incorporated into a three layer BI framework that separates data, logic, and access-related BI components (Section 3). For each individual layer an overview of relevant components, issues stemming from the integration of unstructured data, and proposed solutions to tackle them is given (Section 4). The application of the framework and the integration approaches is illustrated for the domains of CRM and CI (Section 5). The paper concludes with a wrap-up discussion of the presented approaches and an evaluation of further research needs (Section 6).

Approaches to the Integration of Structured and Unstructured Data

Harnessing unstructured data for management support has been addressed from several angles. Case based publications often present pragmatic solutions that enable simultaneous access to structured and unstructured data (e.g. Becker, Knackstedt, & Serries, 2002; Priebe, Pernul, & Krause, 2003). Research with a strong focus on technical and algorithmic challenges mainly focuses on techniques for analyzing document collections based on an

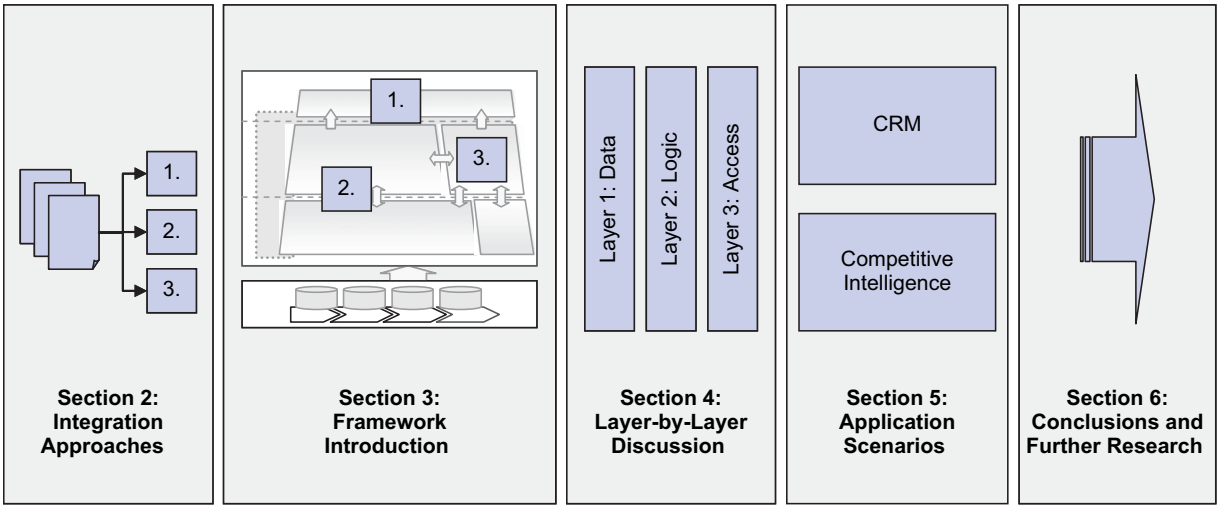

Figure 1. Course of the paper.

extraction of structured data from unstructured content (e.g. McCabe, Lee, Chowdhury, Grossman,, & Frieder, 2000; Mothe, Chrisment, & Dousset, 2003; Keith, Kaser, & Lemire, 2005; Sukumaran, & Sureka, 2006; Cody et al., 2002). Eventually, some authors are approaching the subject from a systems integration perspective and discuss the application of established tools for distributing unstructured content to effectively spread knowledge generated during the analysis of structured data (e.g., Klesse, Melchert, & von Maur, 2005; Baars, 2006). This leads to the following three main approaches:

1. integrated presentation of structured and unstructured content;
2. analysis of content collections; and
3. distribution of analysis results and analysis templates.

The three approaches are explained in further detail in the subsequent subsections.

Integrated Presentation

In this approach, structured data and unstructured content are simultaneously accessed via an integrated user interface. This basic idea leads to a wide spectrum of integration possibilities: Starting with a simple side-by-side presentation of contents up to firmly coupled systems with elaborately combined search and presentation functions (Becker et al., 2002; Priebe et al., 2003).

> *Example:* When navigating in sales data with an OLAP application the selection of analysis dimensions (e.g. "time" and "product group"), of the data subset (e.g. "only East-Asian outlets"), and of the granularity (e.g. "results per quarter") automatically triggers a parallel search for fitting content in a document repository (e.g. documents with results from market research on customer requirements in the Pacific-Asian region). The OLAP data and the selected documents are presented side-by-side.

Figure 2 visualizes the approach by depicting the independent systems with the integrated presentation layer on top of them.

The main benefits of this approach can be traced back to a more convenient handling of the respective functionalities to bolster their combined usage: Functions to access structured and unstructured data can be used together in an efficient and straightforward manner and users have to get accustomed to one system with one user interface only. Moreover, an automatically generated juxtaposition of search results can uncover and visualize otherwise neglected interrelations between structured and unstructured content (Klesse et al., 2005; Baars, 2005).

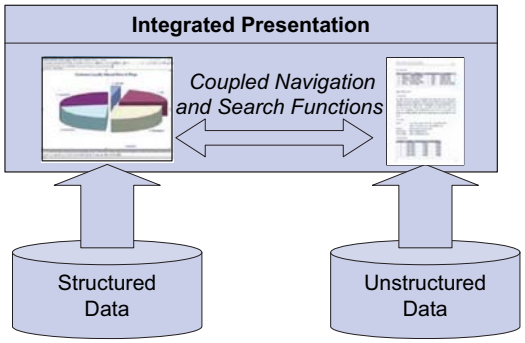

Figure 2. Approach 1—integrated presentation.

Analysis of Content Collections

Based on a structured description of content items with metadata (e.g., author, date of creation, length, and addressed product) it becomes possible to analyze large collections of unstructured data: Identifiers of the content items are treated as facts that are subject to analysis, whereas metadata fields are used for classification purposes and thereby act as analysis dimensions. This especially makes it possible to associate individual documents with numerical facts directly, based on shared dimensions and to investigate document frequencies—e.g., the number of documents that cover a certain subject and are connected to certain organizational units (Gregorzik, 2002; Cody et al., 2002; Inmon. 2007; Sullivan, 2001; Keith et al., 2005; McCabe et al., 2000; Sukumaran, & Sureka, 2006).

After its extraction, the metadata-based content descriptions can be handled just like any other structured data source and be stored in an integrated data repository alongside other relevant data. Such repositories can be accessed with all analysis tools known from "classical" management support, especially with data mining or OLAP tools. By combining metadata based descriptions of content items and structured (numerical) data from other sources it becomes possible to conduct joint analyses combining both information types (Cody et al., 2002; Sukumaran & Sureka, 2006; McCabe et al., 2000; Mothe, Chrisment, & Dousset, 2003). The processing steps necessary for this approach (extraction of meta data, integration into a structured data repository, integrated analysis) are illustrated in Figure 3.

Regarding their origin, relevant metadata can either be entered manually in the source systems by end users, for example, with fields to identify customer satisfaction levels. Interesting metadata can also result from access and usage logs or from search queries, for example, to identify the demand for content and to identify gaps in information coverage. Eventually "text mining technologies"

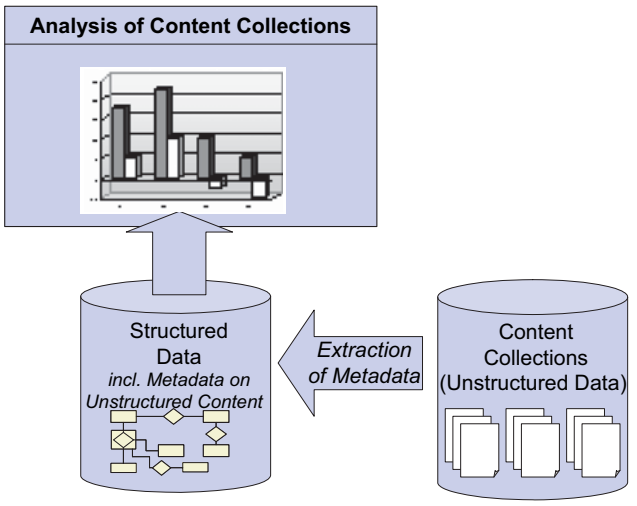

Figure 3. Approach 2—analysis of content collections.

can be used to automatically extract metadata on the semantics of unstructured content items (Baars, 2005; Inmon, 2007; McCabe et al., 2000). Examples for analyses of content collections follow.

- Content can be based on metadata for trouble tickets collected at a help desk: What types of problems arise with what kinds of products in what regions at what times? Can certain types of complaints be linked to recent profit developments? A
- Content can be based on service reports: Are there clusters of problems that can be traced back to certain product configurations?

As the examples above illustrate, this approach enables completely new types of analyses and thereby in-depths insight into business interrelations (Cody et al., 2002; Sullivan, 2001; Sukumaran & Sureka, 2006).

Distribution of Analysis Results and Analysis Templates

Stemming from the realm of "Knowledge Management" (KM), a wide variety of mature systems is available on the market that addresses the issue of distributing unstructured or semi-structured electronic content within an organization (Maier, 2004). The respective functionality can also be used for the diffusion of relevant knowledge generated with BI tools (Klesse et al., 2005; Baars, 2006).

This approach presupposes that there indeed is BI knowledge that can be efficiently shared and that is of some relevance for a sufficient number of users. This holds true for direct analysis results of general interest that remain relatively stable over time, e.g. a cluster analysis that groups customer types according to their buying behavior or results from a shopping basket analysis.

Nevertheless, even if the concrete analysis results are too specific for immediate reuse, the process on how they have been derived might nevertheless be of interest for sharing: What selected data sources have been selected? Which analysis model has been applied? What have been the selected parameter values? In what sequence has the analysis been conducted? Which visualization was chosen?

An analysis of profit contributions for different products of an individual business segment might not be germane for other segments—but the knowledge on how to achieve significant results and how to present them effectively certainly is. Ideally, this knowledge can be stored and applied automatically with templates that can be used to pre-configure the relevant BI applications (Baars, 2006).

> *Example:* To identify the root causes of diminishing cash flows in a certain business segment, several analyses are conducted in an iterative fashion. It turns out that stock turnover has been diminishing significantly recently for certain inventory items. Although those concrete results might not be of interest to the other segments, the analysis unraveling these trends should be applied in those segments as well to check for similar developments.

Figure 4 shows the steps involved in the third approach: Data extraction, data analysis and refinement, and transfer of analysis results or analysis templates to KM tools for distribution.

The main benefit of this approach lies in the facilitation of a more effective and efficient application of analysis systems and methods—especially pertaining to more complex ones that require skillful application as well as extensive calibration and parameterization to provide comprehensive and meaningful results.

A Multi-layer Framework for Business Intelligence

The approaches to integrate structured and unstructured data for management support are embedded in a framework that can be used as a vendor-neutral conceptual reference for BI solutions. It maps relevant logical BI components and visualizes their core interrelations.

The basic structure of the framework has been developed over the course of several years in tight interaction with practitioners both from the supply and the application side. The groundwork of the framework encompasses results from several empirical studies and research projects. With the addition of components to handle and analyze unstructured data, it takes the trend

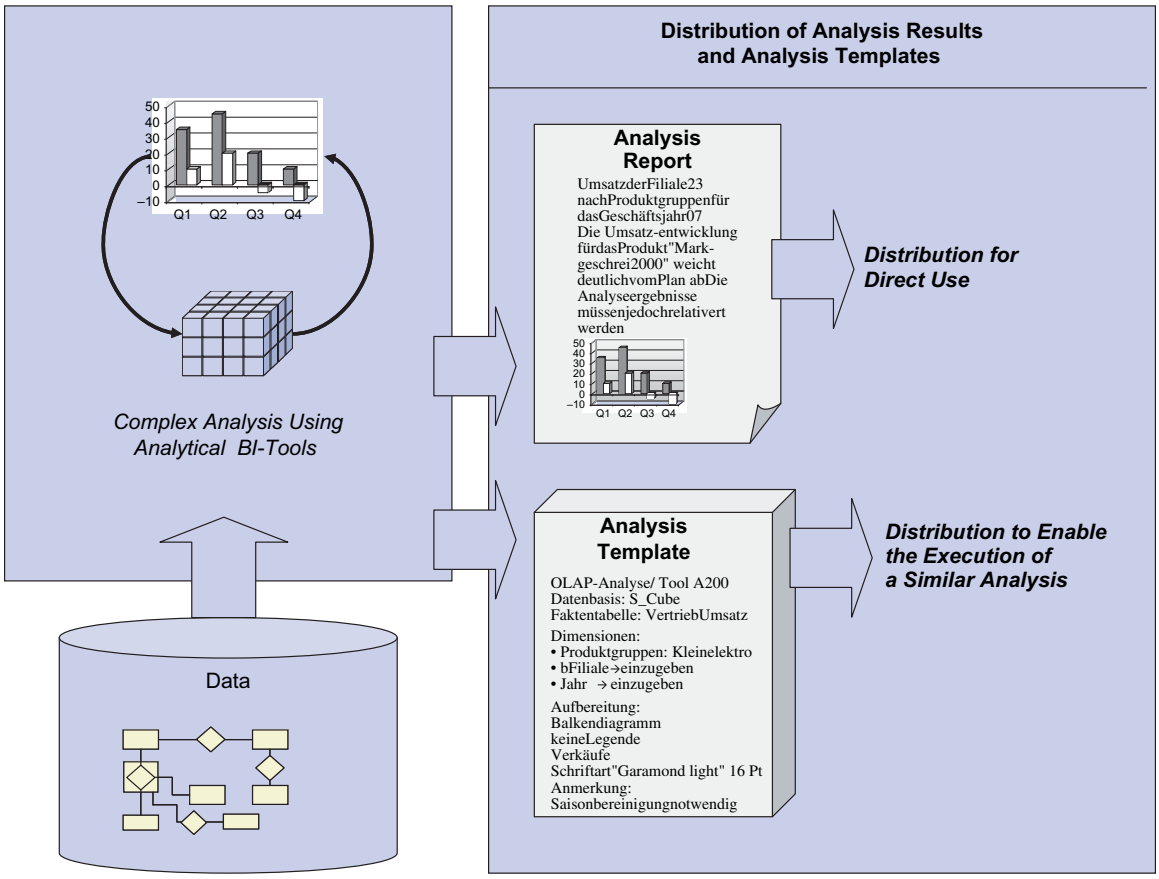

Figure 4. Approach 3—distribution of analysis results and analysis templates.

to holistic management support infrastructures into account.

The framework distinguishes between three relevant layers (Kemper, & Baars, 2006), as follows.

1. The "Data Layer" is responsible for storing structured and unstructured data for management support purposes. Usually structured data is kept in special data repositories—Data Warehouses, Data Marts, and Operational Data Stores, while unstructured content is handled with Content and Document Management Systems. The data is extracted from source systems which might include operational systems like ERP or SCM systems and external data sources e.g. with market research data. Before it can be analyzed in any valid way, the data usually has to be transformed in several steps (Kemper, 2000). The respective transformation steps are commonly subsumed under the acronym "ETL" (Extract-Transform-Load"). It has to be considered that especially external sources often deliver data in unstructured document form only (Mertens, 1999).
2. The "Logic Layer" provides functionality to analyze structured data or unstructured content and supports the distribution of relevant knowledge. The analytical functionality of the Logic Layer includes OLAP and data mining but also functionality to generate (interactive) business reports, ad hoc analysis, and to implement performance management concepts like the Balanced Scorecard (Kaplan & Norton, 1996), or Value Driver Trees (Rappaport, 1998). For the distribution of knowledge, tools from the Knowledge Management domain are applied, e.g. workflow support or tools for information retrieval.
3. The "Access Layer" allows the user to conveniently use all relevant functions of the Logic Layer in an integrated fashion—within the confines of defined user roles and user rights. Usually the Access Layer is realized with some sort of "portal software" that provides a harmonized Graphical User Interface (Priebe et al., 2003).

The management of BI components and BI contents entails the need to document information about both technical configurations; for example, regarding the connection to the source systems, and the business related background, e.g., the semantics behind the data and the transformation steps (Vaduva, & Vetterli, 2001).

This is done with metadata storage repositories, which can either come along with individual components (decentralized metadata repositories) or be realized in a centralized fashion (central and federal metadata repositories). The complete framework with its three layers and its components and the metadata repository is shown in Figure 5.

Commercial BI tools do not always fully "comply" with the conceptual structure of this framework. For example, some analysis tools bring along their own data repositories and some "end-to-end-solutions" cover every layer. Usually such tools are equipped with interfaces that allow interlinking them at exactly the component borders depicted in the framework, however.

Furthermore, it needs to be noted that a concrete BI solution of an individual company does not necessary need to include all components shown in the framework. Using the framework, it becomes possible to identify gaps in the current BI infrastructure, however, and to sketch a migration path for further developments. The three approaches to integrate structured and unstructured data can be located within the framework outlined below.

1. Approach 1—"Integrated Presentation"—resides primarily on the Access Layer as it addresses user interaction only without directly combining contents or systems. It can usually be realized within a portal environment that might be enhanced with an add-on-component for joining navigation and search requests.
2. Approach 2—"Analysis of Content Collections"—spans both the Data and the Logic Layer. On the Data Layer, selected collections of unstructured content have to be stored and handled—this is usually done with content and document management systems. These systems also serve as additional source systems from which metadata is extracted that is fed into a data warehouse or a data mart. The extraction process itself might include some sophisticated text mining tools. After its storage, the metadata-based content descriptions can be analyzed with analysis systems from the Logic Layer.
3. Approach 3—"Distribution of Analysis Results and Analysis Templates"—connects different components on the Logic Layer, namely analysis systems and components for knowledge distribution. This can be either handled manually or with the use of a middleware between the respective components. Usually the approach also utilizes some types of content stores from the Data Layer for persistence. The discussed mapping of the three approaches with the BI framework is shown in Figure 6.

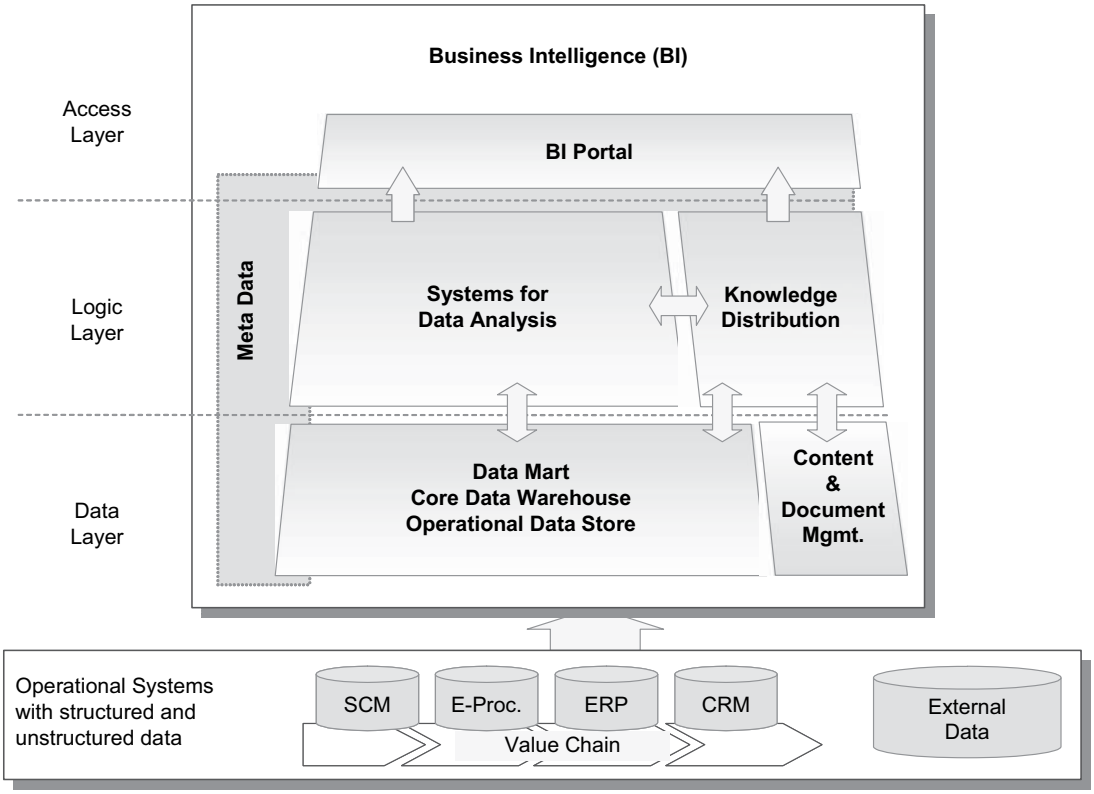

Figure 5. Business intelligence framework.

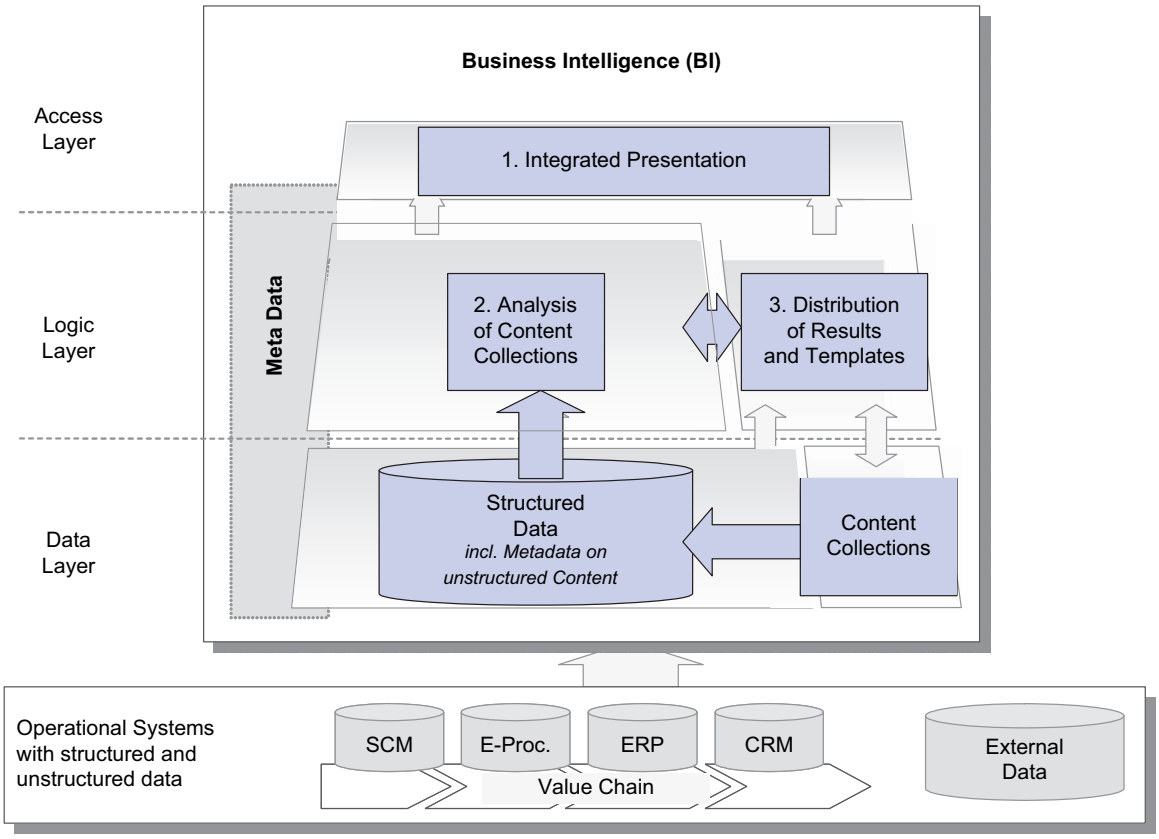

Figure 6. Business intelligence framework and integration approaches.

The Three Layers of the Business Intelligence Framework

In the following section, the three layers of the framework are discussed in further detail. For each layer its components and their interactions are described. Special emphasis is given to components that handle unstructured data and to the challenges of implementing and using them.

Data Layer

A consistent base of semantically and syntactically harmonized data pertinent for management support is a main pillar for a solid BI solution (Kemper, 2000). Regarding structured data, the central component to accomplish this task in larger BI installations is the "data warehouse." According to Inmon (2005), a "data warehouse" is defined as a "subject-oriented, integrated, time-variant, and non volatile collection of data in support of management's decision-making process." Many current realizations of data warehouses are based on so called "core data warehouses", which are dedicated components for the storage of all management support data. core data warehouses are usually not used as a direct source for analysis systems, but rather act a source for individual "data marts" that keep application specific data, i.e. data prepared for the support of a single business process or business function (Inmon, 2005). The use of a core data warehouse in combination with dependent data marts is known as the "hub-and-spoke approach" (Ariyachandra, & Watson, 2006).

More recently, there has been a shift towards data warehouse infrastructures, which also feed operational systems, and thereby support real-time data monitoring and analysis. In such an environment, it might be of use to introduce an "Operational Data Store" (ODS) that keeps real time data on a transactional level for time critical tasks (Kimball et al., 2002; Inmon, 1999).

To transfer data into the Data Warehouse "ETL tools" are needed to support the extraction and transformation with data from the source systems. The transformation encompasses filtering out syntactical and semantic errors, harmonizing data from different sources, aggregating data, and enriching it by calculating additional business metrics (Kemper, 2000).

A core requirement for realizing the integration of unstructured data into BI infrastructures is to include components that handle large volumes of such content. In many companies, there are already document and

content management solutions in place for exactly this purpose (Maier, 2004): "Document management" encompasses functions for input, indexing, archival, versioning, and provision of digitized paper as well as genuine electronic documents. A slightly different focus can be found in "content management systems", which primarily aim at bringing together various information sources of different format and origin. By decoupling content, structure and layout these systems allow for a consistent storage, administration, and distribution of the administered content. Recently the two categories have begun blending: Document Management Systems are increasingly enhanced with features rooted in the Content Management realm and vice versa (Gronau, Dilz, & Kalisch, 2004).

Next to the provision of access to unstructured data, content and document management systems also take the role of sources of structured (meta) data for integrated data warehouses or data marts. This is a cornerstone of Approach 2 (analysis of content collections).

A major challenge of gathering metadata from content and document management systems lies in its efficient extraction. While harvesting and filtering data from manually entered metadata or from log files is usually straight forward and widely practiced—especially for analyzing web content (Giovinazzo, 2002), automatically extracting semantic metadata is still challenging. The respective procedures are usually understood as parts of a "text mining" process (Sullivan, 2001). When extracting metadata with text mining software two tasks need to be separated conceptually (Cody et al., 2002; Dörre, Gerstel, & Seiffert, 1999) as detailed below.

1. Describe and cluster content to uncover relevant dimensions.
2. This step is necessary if relevant dimensions for the description of content items are not fully known beforehand, e.g., when analyzing patent documents or news articles on yet undiscovered trends. In this case, one needs to explore what subjects the items cover, as well as what metadata fields might characterize them comprehensively. This task can be supported with text mining tools designed for describing and clustering documents.
3. Analyze content and generate metadata according to predefined dimensions.
4. As soon as the dimensions are known all individual content items need to be classified accordingly, i.e. metadata has to be generated. This can be done with tools for text classification. An example would be the classification of customer complaint e-mails based on the referred products.

As soon as the metadata is generated, it needs to be transformed and included in the data warehouse (e.g., with ETL tools). The complete process with the distinction between direct extraction of meta data and a text mining based pre-processing is illustrated in Figure 7.

So far the extraction of semantic metadata based on text mining algorithms is still research in progress—a valid classification of unstructured content usually requires a considerable amount of manual interaction and can by no means be fully automated (Cody et al., 2002). If the application of such tools is considered, the tools should be thoroughly scrutinized for their cost/benefit-ratio.

Logic Layer

The Logic Layer focuses on the compilation, processing, and distribution of management support data. Two types of systems are distinguished at this layer: systems for data analysis and components for knowledge distribution. Systems for data analysis are collecting and processing data from the Data Layer and convert them into a form pertinent for a presentation to the user. Here two groups of systems are distinguished (Kemper et al., 2004).

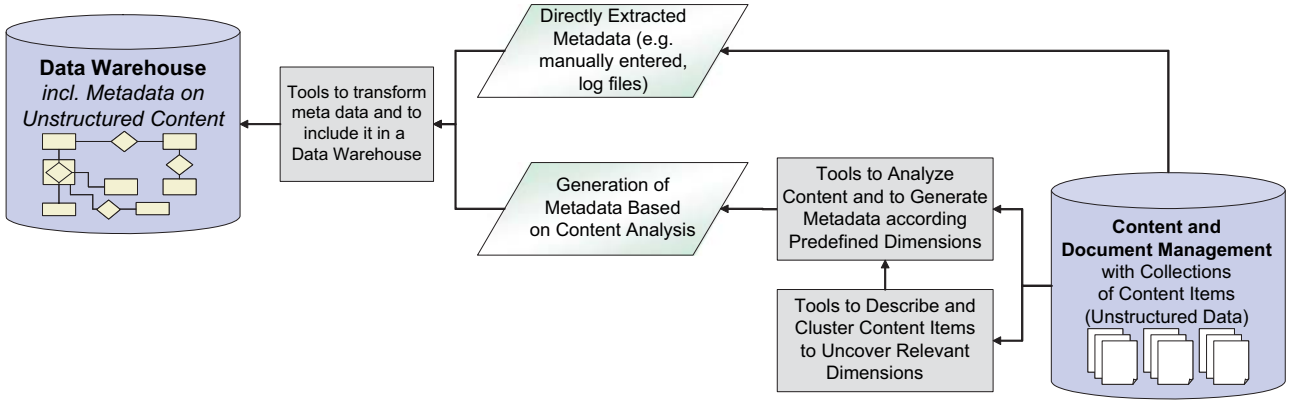

Figure 7. Extracting metadata from content and document management systems.

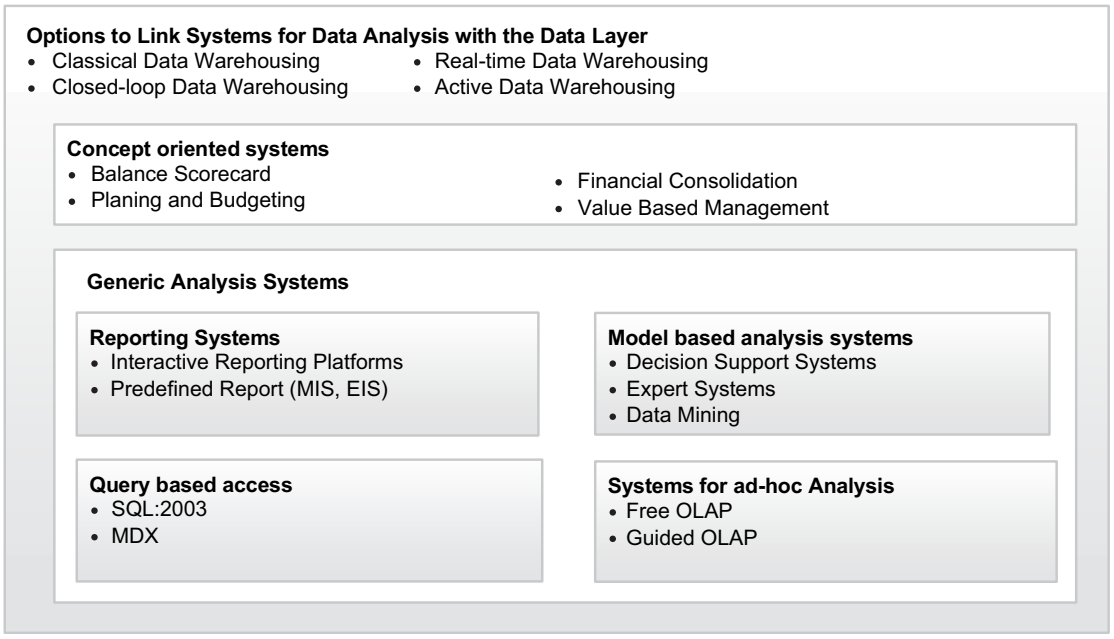

Figure 8. Types of data analysis systems.

Generic Analytical Systems

"Generic analytical systems" that enable accessing, combining, and analyzing data without a build-in ex-ante specification of business logic. Generic analytical systems encompass the following data analysis techniques:

- "Query based access" permits the user to read from the data layer using languages like SQL or the multidimensional query language MDX.
- "OLAP systems" provide functions to navigate data in a multidimensional manner as described in Section 1.
- "Reporting systems" present data by combining text, numbers, and business graphics. Current reporting systems allow the definition of interactive reports where the users can switch views on the data.
- "Model based analytical systems" offer algorithmic means to perform complex analytical operations for uncovering complex patterns and relations. They include "decision support systems", "expert systems", and the already introduced "data mining systems". Decision support systems usually provide methods for defined problem areas, e.g. to optimize lot sizes or to select optimal investment alternatives. Expert systems use a set of rules to derive conclusions in defined application domains.
- "Concept oriented systems" are designed to implement complex business concepts such as the Balanced Scorecard (Kaplan, & Norton, 1996), Value Based Management (Rappaport, 1998; Martin, Petty, & Petty, 2000), planning and budgeting, or financial consolidation in form of standard applications.

Based on this categorization Figure 8 gives an overview on systems for the analysis of structured data. Note that all of the described systems—in particular OLAP and data mining systems—might be applied on document metadata to uncover hidden patterns in huge document repositories (Weiss, Indurkhya, Zhang, & Damerau, 2005)—and can therefore be used for the implementation of Approach 2—Analysis of Content Collections.

Components for Knowledge Distribution

For handling unstructured data, components for knowledge distribution are needed. Required are functions for identifying single information objects ("information retrieval"), and for the navigation in large knowledge bases (Weiss et al., 2005; Maier, 2004). The relevance of such functions increases with the volume stored in the content and document management systems on the Data Layer. To extract information from such repositories, elaborate text-mining tools have been proposed that help identifying, summarizing, linking, and grouping content items (Fan, Wallace, Rich, & Zhang, 2006). Note, that some of the text mining technologies for clustering and categorization that have been discussed in the preceding section can be applied here as well—in both cases, issues of understanding and describing unstructured data are addressed. On the Logic Layer the respective functions are directly made accessible to end users, though, and the tools are used in their own right and not part of as a metadata extraction process.

The components for knowledge distribution can be used to spread results and templates generated with data analysis systems, specifically; thereby, they can be used for the implementation of Approach 3. This enables the utilization of mature functionality found commonly in the fields of content and document management, which encompasses) the following elements (Gronau et al., 2004; Götzer, Schneiderach, Maier, Boehmelt, & Komke 2001; Baars 2006.

- Workflow management for the implementation of multi-stage editorial processes.
- Information Life Cycle Management to ensure no obsolete results or templates remain in the content base.
- "Check-In, Check-Out" to handle concurrent access.
- Bundling of several documents (document collections) to efficiently deal with sets of interrelated documents.
- Versioning to track changes on documents and to reuse older versions.
- Information Retrieval, to identify relevant documents.

Regarding the exchange of results and templates between Analysis Systems and the knowledge distribution components, efficiency gains may be attained by the introduction of a respective middleware. Such a middleware can act as a hub between several systems and include functions to convert and translate formats. This idea is shown in Figure 9. Reports and templates that are generated with the diverse analysis systems (on the left side of Figure 9) are converted and transferred via the "middleware hub" to the knowledge distribution components (on the right hand side). When templates need to be applied, they can be transferred back to the analysis systems.

The realization of such "hub components" is still subject to research and development. With the continuing trend towards tools with standardized web service interfaces and the ongoing standardization of exchange formats in the BI area, interface related issues could be expected to become less and less severe in the near future (Baars, 2006).

Besides this, issues of information quality and motivation have to be considered carefully. Compiling comprehensible result documents or analysis templates requires extra efforts on end user side and has to be supported with additional incentive systems and quality assurance processes. In consequence, this approach introduces recurring costs.

Access Layer

The Data Access Layer is responsible for bringing all relevant components and functions from the Logic Layer together and to present them to the user in an integrated and personalized fashion. It is common to use "portal systems" for the implementation of this layer. Portals provide an integrated user interface for different content

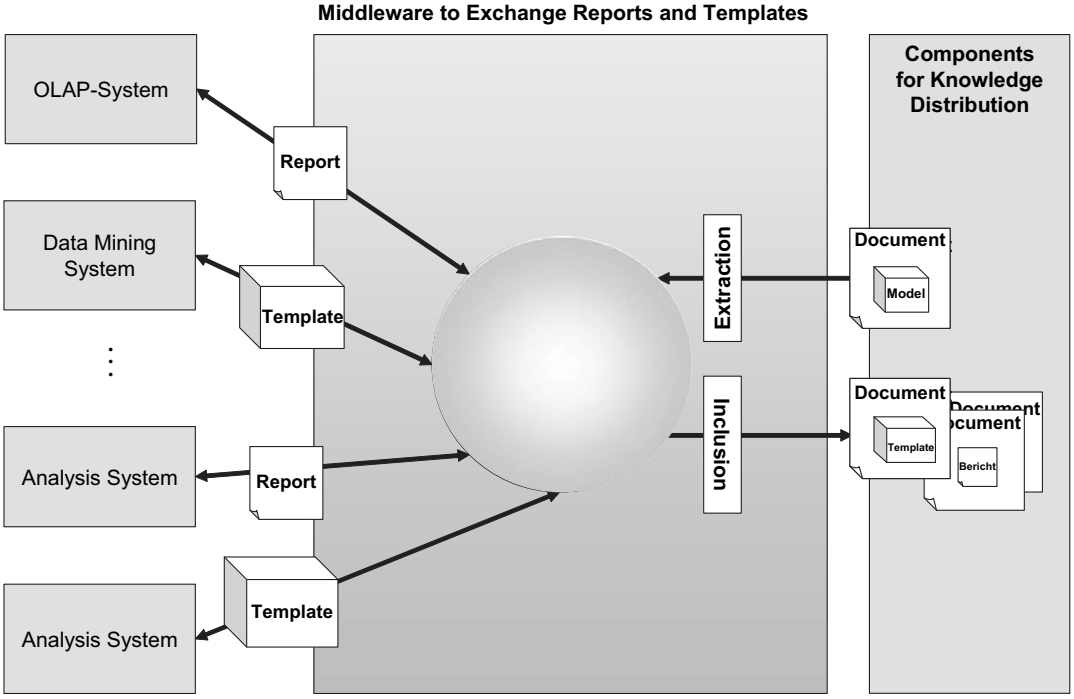

Figure 9. A middleware hub for the exchange of analysis results and analysis templates.

sources and application systems. This goes along with a consolidated user navigation and an integrated access and user rights management ("*single login*"). Applications are included in the portal by the use of "portlets." "Portlets" are independent software components that bring along their own Graphical User Interface, occupy a defined space of a portal web page, and communicate with the portal via defined interfaces (Davydov, 2001; Priebe et al., 2003).

Portals are particularly useful to realize simultaneous and homogeneous access to both of structured and unstructured data (Approach 1—Integrated Presentation). It has been shown that a portlet-based environment can be used to realize joint search and navigation functionality as discussed in Section 2. For example, it is possible to couple the navigation in an OLAP application with search functionality that selects fitting documents from a content management system. Necessary is a component that intercepts user events from the portlets, translates them, and hands them over to the respective complementary systems (Cody et al., 2002; Priebe et al., 2003). A possible architecture for such an approach is shown in Figure 10. A main advantage of this approach is that it does not incur recurring costs besides portal administration and maintenance and that it does not induce additional efforts on end user side.

Application Scenarios

Use and relevance of the presented framework and the three integration approaches are illustrated for two application domains:

- CRM; and
- CI.

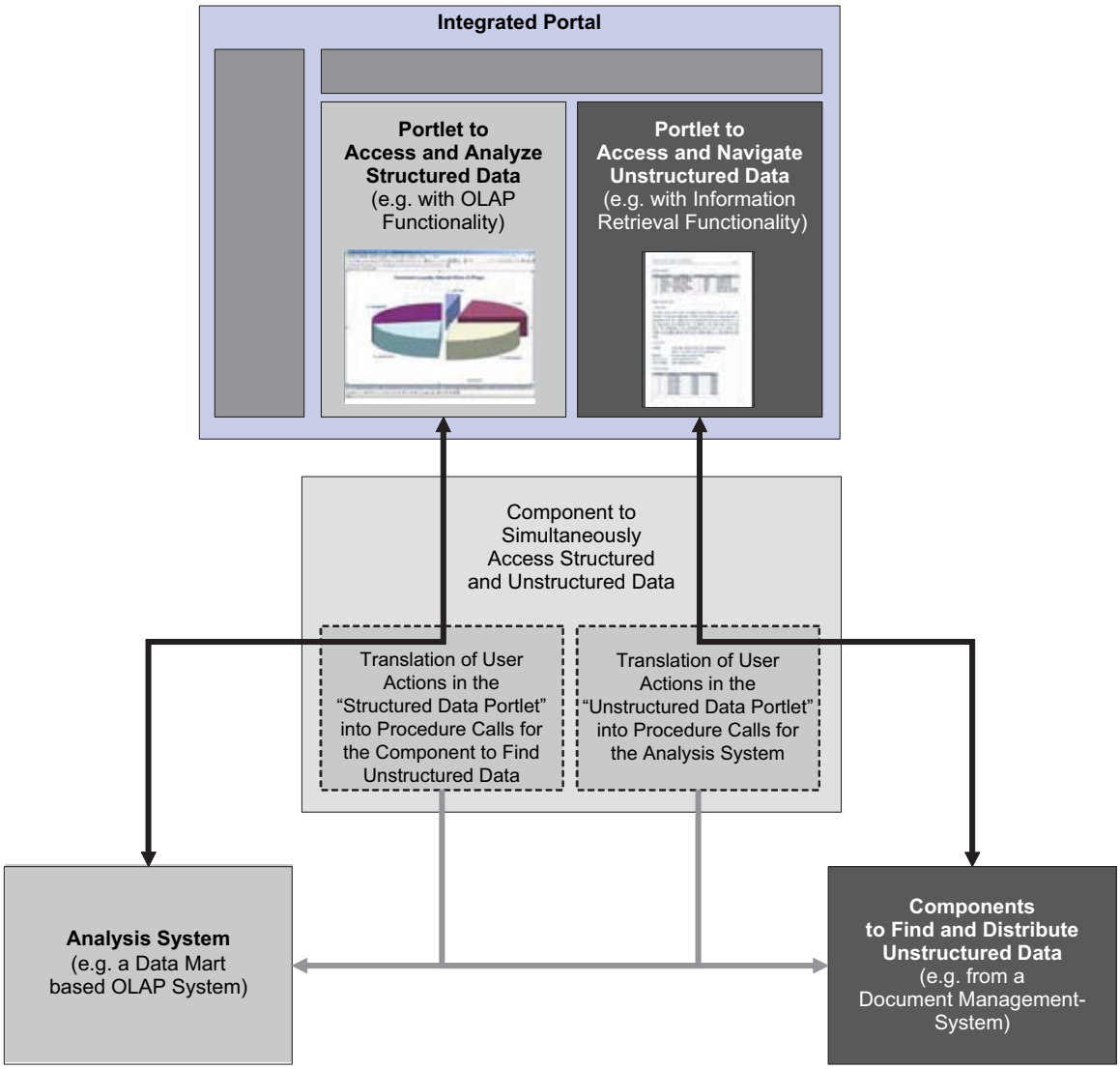

Figure 10. Integrated access to structured and unstructured data.

The following sections contain a brief characterization of the respective domains, an outline of their analytical demands and relevant systems, and a presentation of possible application scenarios that highlight the business potential of the three integration approaches.

Customer Relationship Management

The concept of Customer Relationship Management (CRM) suggests binding all customer-oriented activities together. In effect, this means an integration of all marketing, sales, and service processes. CRM advocates a consistent and complete view on the customer based on an integrated data pool—the Customer Data Warehouse (Payne, & Frow, 2005; Boulding, Staelin, Ehret, & Johnston, 2005). Usually CRM is divided into three main building blocks (Schierholz, Kolbe, & Brenner, 2005; Hippner, & Wilde, 2004).

1. *Operational CRM*, which encompasses all applications that support day-to-day operational activities in marketing, sales, and service.
2. *Collaborative CRM*, that addresses the integrated management of all customer touch points (e-mail, phone etc.) to enact a consistent and reliable communication with customers.
3. *Analytical CRM* for the provision for a consistent collection of customer data in a Customer Data Warehouse and its systematic analysis.

CRM is subject to BI, as it needs an effective and integrated management support infrastructure to cover the objectives associated with Analytical CRM successfully. Figure 11 visualizes the discussed building blocks and their interrelations.

Sales- and marketing-related data analysis tools have been subject to management support ever since. However, the highly integrated nature of the CRM concept and the proposed tight interconnection between analytical and operational systems impose special requirements on the respective BI infrastructures. First, the Customer Data Warehouse needs to be fed with a huge amount of data from a variety of historically grown and heterogeneous marketing, sales, and service systems. This demands for carefully crafted ETL processes and an adequately scaleable data warehouse. Second, CRM demands an environment in which analytical and operational systems need to be coupled tightly, e.g., to enable an automatic and immediate classification of customers entering a web shop or to support sales reps to identify cross- and up-selling opportunities (Furness, 2004; Reid, & Caterell, 2005).

A major obstacle on the way to the envisioned complete view on the customer lies in the unstructured nature of many relevant CRM data sources (Dörre et al., 1999; Cody et al., 2002; Mertens, 1999; Reid et al., 2005).

- A huge amount of the interaction with the customer is handled via unstructured media, e.g. letters, e-mails, or input forms that allow free text input.
- Independent sales and service reps often produce semi-structured reports at best.
- Marketing reports from third parties are in many cases only distributed in document form.

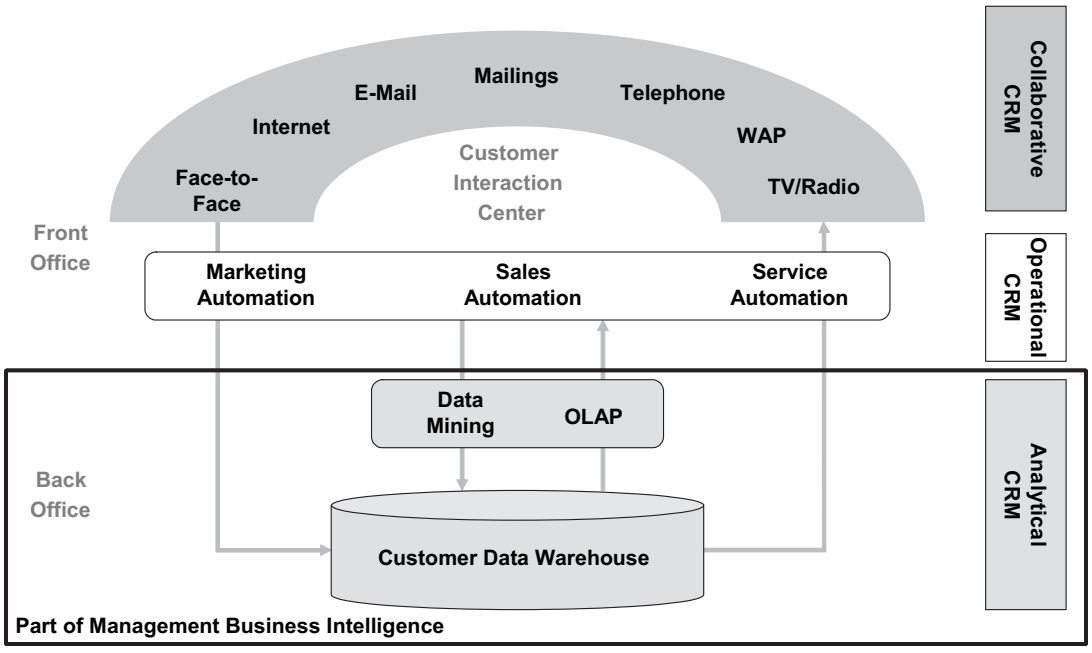

Figure 11. Customer relationship management and business intelligence (based on Hippner & Wilde, 2004).

- Frequently text input fields in operational systems designed to generate structured data de facto produce unstructured results because of a huge variety of possibilities to enter the same semantics.

All this demands for possibilities to store, navigate, search, and analyze unstructured content. The following application scenarios exemplify the potential of three discussed approaches in such CRM environments.

- While analyzing sales results with an OLAP application, corresponding market reports are selected and presented to the user. This way she can distinguish more easily between those that can be attributed to general market trends and those that are more likely the results of sales performance (Approach 1).
- All customer letters arriving at an organizational unit to handle complaints are scanned and stored digitally in a document management system. Relevant metadata is entered (manually) for each document, e.g. affected product, relevant product parts, occurrence time, problem type. The metadata is extracted and fed into the Customer Data Warehouse. Using analytical CRM applications an accumulation of defects in certain usage environments can be identified and be associated with a decline in sales. Based on these results sales representatives can optimize their selection of recommended products (Approach 2).
- Customer suggestions regarding product enhancements from a web form are classified with text mining software. Afterwards suggestions can be browsed and selected according to these classifications. Moreover, the classification metadata is fed into the customer data warehouse; thereby, it becomes analyzable with data mining software. For example, customer clusters can be generated under consideration of their suggestions (Approach 2).
- A data-mining expert conducts a complex analysis regarding up-selling possibilities based on a basket analysis. The respective result is adequately formatted, enriched with annotations, and made available to the sales and service staff in document form with components for knowledge distribution. This involves an editorial workflow to ensure intelligibility of the document from a target user's perspective and storage in a content management system. Sales and service reps can use effective search and navigation options that facilitating finding relevant reports whenever needed. Because the analysis will be obsolete in the foreseeable future due to changes in the product portfolio, the expert also generates a template which allows for a quick renewal of the analysis (Approach 3).

Competitive Intelligence

Despite the resemblance of their names, "Competitive Intelligence" (CI) and "Business Intelligence" have evolved independently of each other. The origins of CI can be traced back over two decades when Porter introduced the concept "Competitor Intelligence" (Porter, 1980). Nowadays CI is commonly understood to focus on all processes for gathering and analyzing information about the competition and the general market environment (Ghoshal, & Westney, 1991; Vedder et al., 1999; SCIP, 2005).

Up until recently, CI practitioners and researchers did not stress technological and infrastructural support (Vedder et al., 1999). That does not come as a surprise considering the heterogeneous and mostly external nature of the information sources used in CI; e.g., product brochures, conventions and conferences, customer and Delphi surveys, or patent databases (Dugal, 1998; Meier, 2004).

With the increasing importance of the internet and of electronic documents, a trend towards integrated IT infrastructures for CI can be observed (Vedder et al., 1999). It becomes more and more important to understand CI applications as part of a wider management support IT infrastructure. CI is therefore nowadays seen as an application domain of business intelligence (Kemper, & Baars, 2006).

Considering its broad subject it does not come as a surprise that one finds a wide spectrum of heterogeneous CI analysis applications which differ widely in contents, frequency of usage, urgency of need, analysis depth, and target groups (Dugal, 1998). In CI, this is even more the case than in CRM, which, also subsumes highly different systems, is driven much more by an integrative orientation.

Because of the strong focus on external information, some important commonalities among all CI applications do exist, though. Most relevant CI information is coming from outside company borders: financial reports of competitors, government publications, patent databases, research publications etc. (Lux, & Peske, 2002). Most of those are nowadays available in electronic form and can be accessed via Internet technologies (Vedder et al., 1999). A particularity of these sources is that they usually vary in format, quality, and structure, and have to be gathered from a huge number of sites. Efficient data retrieval, quality assurance, and data consolidation is of central importance in CI.

When designing CI applications one needs to explicitly address the dominance of qualitative and unstructured data. Particularly data published on the Internet is (still) mostly only available in document form, e.g. as HTML or PDF files (Kemper, & Baars, 2006).

The discussed requirements result in dominance of systems for handling unstructured data, namely content

and document management systems, components for finding and distributing unstructured content, and portal based access systems. There are some interesting possibilities for a stronger connection of those systems with "classical" management support systems that handle structured internal data.

The following application scenarios illustrate this.

- An analysis of press reports on selected product groups reveals weak signals regarding shifts in public opinion. A portal environment offers functions for the selection of reports, which are thought to have high impact on public opinion—a selected report is automatically contrasted with sales data in the period around the report's publication date (Approach 1).
- By exploring patent document bases with clustering algorithms it becomes possible to identify technological trends and the strategic orientation of core competitors. After a classification of patent documents according to affected product groups, and a combined analysis with sales data, a possible influence on future sales can be estimated (Approach 2).
- Using imported market data from an external information provider, the impact of a new substitute product on sales is analyzed using trend analysis algorithms. The results are enriched with additional qualitative information regarding the substitute technology and made available for further use by product and marketing managers with knowledge distribution components (Approach 3).

Conclusions and Further Research

This paper discussed three approaches to the integration of structured and unstructured data for management support and positioned them in an integrated BI framework. The proposed tools and additional components to realize the three approaches and their localization within the BI framework are depicted in Figure 12: The portal add-on builds a bridge between the Access and Logic Layer (for Approach 1), the middleware hub for distributing analysis results and templates resides at the Logic Layer (for Approach 3), and the tools for content analysis and meta data extraction data integrate components from the Data Layer (for Approach 2) Table 1 provides a preliminary summery of the three approaches regarding their costs and benefits.

The presented framework can be used to structure BI initiatives which are designed to integrate unstructured data e.g. for CRM or CI applications. Some open issues remain that need special attention in further research, however, especially:

- the practicability of text mining technologies in integrated BI environments has to be evaluated

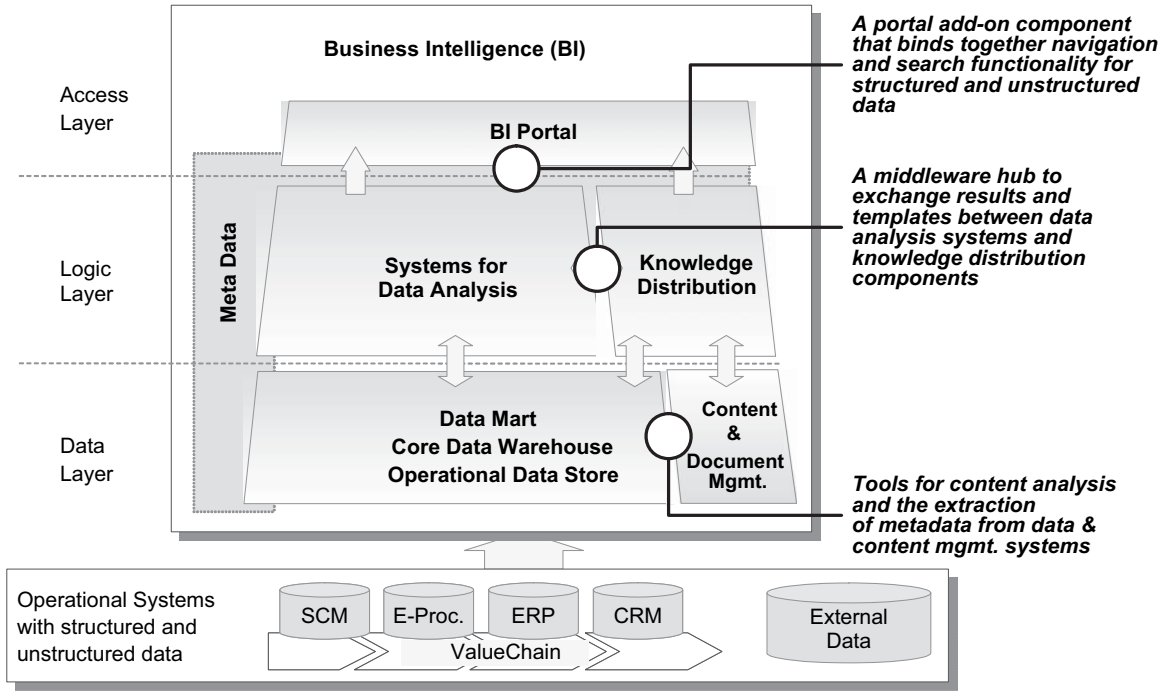

Figure 12. Proposed tools and components for the realization of the three approaches.

Table 1. Comparison of the Three Integration Approaches

	Integrated presentation	*Analysis of content collections*	*Distribution of analysis results and analysis template*
Complexity of implementation	▪ Rather low, primarily of technical nature ▪ Low learning effort on end user side	▪ Medium, if only *available* metadata is used ▪ Rather high, when metadata has to be extracted with text mining tools	▪ Depending on the interfaces of the involved systems: medium to high, primarily of technical nature
Costs of usage, maintenance and Support	▪ low on end user side ▪ costs for day-to-day portal administration (e.g. harmonizing access rights)	▪ Medium to high, depending on the complexity of the analysis (end user costs, support costs)	▪ High on end user side (contents must be compiled and enriched for distribution) ▪ Incentive systems and quality assurance
Benefits	▪ Higher acceptance to conduct simultaneous analysis of structured and unstructured data ▪ Uncovering hidden interrelations	▪ Enables new classes of business analysis and thereby more effective management	▪ Efficient reuse of BI knowledge ▪ Utilization of mature functionality for content storage and distribution

more closely. It especially has to be monitored whether the efforts to gather meaningful metadata with text mining tools are justified by their business value;

- approaches to the integration with further systems from the knowledge management domain for the exchange of BI content need to be elaborated in further detail, e.g. systems based on Semantic Web approaches or for collaborative filtering; and
- organizational operation models for BI environments with unstructured data have to be derived under special consideration of data quality issues.

Given the outlined business potential, it is to be expected that management support infrastructures, which seamlessly integrate components for handling structured and unstructured data, will receive vivid interest in both research and practice.

Author Bios

HENNING BAARS received his doctorate degree from the University of Cologne. He currently lectures and researches at the University of Stuttgart at the Chair of Information Systems I. His research focus lies on Business Intelligence applications in the industrial sector. He can be reached atbaars@wi.uni-stuttgart.de.

HANS-GEORG KEMPER is a Full Professor of Information Systems and holder of the Chair of Information Systems I at the University of Stuttgart. He has conducted extensive research in the fields of Business Intelligence and Management Support.

References

Alter, A. (2003). Business Intelligence—Are Your BI Systems Making You Smarter. *CIO Insight*, *05/2003*, 77–85.

Anandarajan, M., Anandarajan, A., & Srinivasan, C. A. (2004). *Business Intelligence Techniques*. Berlin: Springer.

Ariyachandra, T., & Watson, J. (2006). Which Data Warehouse Architecture is the Most Successful. *Business Intelligence Journal*, *11(1)*, 2–4.

Baars, H. (2005). Knowledge Management und Business-Intelligence. In *WM 2005 - Professional Knowledge Management - Experiences and Visions. Contributions to the 3rd Conference Professional Knowledge Management Experiences and Visions,* Althoff, K.D., Dengel, A., Bergmann, R., Nick, M. & Roth-Berghofer, T., Eds. (pp. 429–433). Heidelberg: Physica.

Baars, H. (2006). Distribution von Business-Intelligence-Wissen—Diskussion eines Ansatzes zur Nutzung von Wissensmanagement-Systemen für die Verbreitung von Analyseergebnissen und Analysetemplates. In, *Analytische Informationssysteme—Business Intelligence-Technologien und -Anwendungen* P. Chamoni & P. Gluchowski, Eds. (3rd ed., pp. 409–424). Berlin: Springer.

Becker, J., Knackstedt, R., & Serries, T. (2002). Informationsportale für das Management – Integration von Data-Warehouse- und Content-Management-Systemen. In *Vom Data Warehouse zum Corporate Knowledge Center—Proceedings der Data Warehousing 2002,* E. Maur & R. Winter, Eds. (pp. 241–261). Heidelberg: Physica.

Boulding, W., Staelin, W., Ehret, M., & Johnston, W.J. (2005). A Customer Relationship Management Roadmap: What Is Known, Potential Pitfalls, and Where to Go. *Journal of Marketing*, *69(4)*, 155–166.

Codd, E. F., Codd, S. B., & Salley, C. T. (1993). *Providing OLAP to User-Analysts: An IT Mandate*. Retrieved September 16 2006 from http://dev.hyperion.com/download_files/resource_library/white_papers/providing_olap_to_user_analysts.pdf

Cody, W. F., Kreulen, J. T., Krishna, V., & Spangler, W.S. (2002). The Integration of Business Intelligence and Knowledge Management. *IBM Systems Journal*, *41(4)*, 697–713.

Davydov, M. (2001). *Corporate Portals and e-Business Integration*. New York: McGraw-Hills.

Devlin, B. (1996). *Data Warehouse: From Architecture to Implementation*. Boston: Addison-Wesley Professional.

Dörre, J., Gerstel, P., & Seiffert, R. (1999). Text Mining. Finding Nuggets in Mountains of Textual Data. In *Proceedings of the Fifth ACM SIFKDD International Conference on Knowledge Discovery and Data Mining,, San Diego, 1999* (pp. 398–401). New York. Association for Computing Machinery.

Dugal, M. (1998). CI Product Line—A Tool for Enhancing User Acceptance of CI. *Competitor Intelligence Review*, *9(2)*, 17–25.

Eckerson, W.W. (2006). *Performance Dashboards. Measuring, Monitoring, and Managing Your Business*. Hoboken, NJ: John Wiley & Sons.

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the Power of Text Mining. *Communications of the ACM*, *49(9)*, 77–82.

Furness, P. (2004): Techniques for Customer Modelling in CRM. *Journal of Financial Services Marketing*, *5(4)*, 293–307.

Gartner Group (2006). Gartner Survey of 1,400 CIOs Shows Transformation of IT Organisation is Accelerating. Retrieved June 19 2006 from http://www.gartner.com/press_releases/asset_143678_11.htm.

Ghoshal, S. & Westney, D. E. (1991): Organizing Competitor Analysis Systems. *Strategic Management Journal*, *12*, 17–31.

Giovinazzo, W. A. (2002): *Internet-Enabled Business Intelligence*. Upper Saddle River, NJ: Prentice-Hall.

Golfarelli, M., Rizzi, S., & Cella, I. (2004). Beyond Data Warehousing—What's Next in Business Intelligence. In *Proceedings of the DOLAP 2004*, pp. 1–6, New York, November 2004. Association for Computing Machinery.

Götzer, K., Schneiderach, U., Maier, B., Boehmelt, W., & Komke, T. (2001): *Dokumentenmanagement—Informationen im Unternehmen effizient nutzen* (2nd ed.) Heidelberg: Computerwoche.

Gregorzik, S. (2002). Multidimensionales Knowledge Management. In: U. Hannig (Ed.), *Knowledge Management und Business Intelligence* (pp. 43–51). Berlin: Springer.

Gronau, N., Dilz, S., & Kalisch, A. (2004): *Anwendungen und Systeme für das Wissensmanagement—ein aktueller Überblick*. Berlin: GITO Verlag.

Hand, D., Mannila, H., & Smyth, P. (2001): *Principles of Data Mining*. Cambridge, Massachusets: MIT Press.

Hippner, H., & Wilde, K. D. (2004): *IT-Systeme im CRM, Aufbau und Potenziale*. Wiesbaden: Gabler.

Inmon, W. H. (1999). *Building the Operational Data Store* (2nd ed.). J Hoboken, NJ: John Wiley & Sons.

Inmon, W. H. (2005). *Building the Data Warehouse* (5th ed.). J Hoboken, NJ: John Wiley & Sons.

Inmon, W. H. (2007). DW 2.0™—The Architecture for the Next Generation of Data Warehouse. Retrieved July, 24 2007 from the *Corporate Information Factory* website: http://www.inmoncif.com/

Kantardzic, M. (2003). *Data Mining. Concepts, Models, Methods, and Algorithms*. J Hoboken, NJ: John Wiley & Sons.

Kaplan, R. S., & Norton, D. P. (1996): The Balanced Scorecard—Translating Strategy into Action, Boston: Harvard Business School Press.

Keith, S., Kaser, O., & Lemire, D. (2005). Analyzing Large Collections of Electronic Text Using OLAP. In: *29th Conf. Atlantic Provinces Council on the Sciences (APICS 2005)*, Wolville (Canada), October, 2005.

Kemper, H. G. (2000): Conceptual Architecture of Data Warehouses—A Transformation-oriented View. In *Proceedings of the 2000 American Conference On Information Systems,* pp. 108–118, Long Beach, CA, August 2000. Association for Information Systems.

Kemper, H. G., & Baars, H. (2006). Business Intelligence und CI. IT-basierte Managementunterstützung und markt-/wettbewerbsorientierte Anwendungen. In, *Business & Competitive Intelligence—HMD—Praxis der Wirtschaftsinformatik 247,* H.G. Kemper, H. Heilmann & H. Baars, Eds. (pp. 7–20), Heidelberg: dpunkt.

Kemper, H. G., Mehanna, W., & Unger, C. (2004). *Business Intelligence—Grundlagen und praktische Anwendungen*. Wiesbaden: Vieweg.

Kimball, R., & Ross, M. (2002): *The Data Warehouse Toolkit. The Complete Guide to Dimensional Modelling* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Klesse, M., Melchert, F., & von Maur, E. (2003). Corporate Knowledge Center als Grundlage integrierter Entscheidungsunterstützung. In *WM 2003—Professionelles Wissensmanagement—Erfahrungen und Visionen* Reimer, U., Abdecker, A., Staab, S., & Stumme, G., Eds. (pp. 115–126). Luzern, April 2003. Bonn: GI.

Kohavi, R., Rothleder, N. J., & Simoudis, E. (2002): Emerging Trends in Business Intelligence. *Communications of the ACM*, *45(8)*, 45–48.

Lux, C, & Peske, T. (2002). *Competitive Intelligence und Wirtschaftsspionage—Analyse, Praxis, Strategie*. Wiesbaden: Gabler.

Maier, R. (2004). *Knowledge Management Systems—Information and Communication Technologies for Knowledge Management* (2nd ed.). Berlin: Springer.

Martin, J. D.; Petty, J. W. & Petty, W. J. (2000): Value Based Management: The Corporate Response to the Shareholder Revolution, Boston: Harvard Business School Press.

McCabe, C., Lee J., Chowdhury, A., Grossman, D.A., & Frieder, O. (2000). On the design and evaluation of a multi-dimensional approach to information retrieval. In *Proceedings of the 23rd Annual Int. ACM Conf. on Research and Development in Information Retrieval (SIGIR 2000)* (pp. 363–365). New York: ACM

Meier, M. C. (2004). Competitive Intelligence. *Wirtschaftsinformatik*, *46(5)*, 405–407.

Mertens, P. (1999). Integration interner, externer, qualitativer und quantitativer Daten auf dem Weg zum aktiven MIS. *Wirtschaftsinformatik*, *41(5)*, 405–415.

Mothe, J., Chrisment, C., Dousset, B., Alau, J., 2003. Doc-Cube: Multi-dimensional visualisation and exploration of large document sets. In *Journal of the American Society for Information Science and Technology (JASIST)*, *54(7)*, 650–659.

Negash, S. (2004). Business Intelligence. *Communications of the Association for Information Systems*, *13*, 77–195.

Payne, A., & Frow, P. (2005). A Strategic Framework for Customer Relationship Management. *Journal of Marketing*, *69*, 167–176.

Porter, M.E. (1980). *Competitive Strategy—Techniques for Analyzing Industries and Competitors.* New York: Free Press.

Priebe, T., Pernul, G., & Krause, P. (2003). Ein integrativer Ansatz für unternehmensweite Wissensportale. In *Proceedings der 6. Internationalen Tagung Wirtschaftsinformatik.* Dresden, September 2003. (pp. 277–291).. Dresden: Technische Universität Dresden.

Rappaport, A. (1998). Creating Shareholder Value—A Guide for Managers and Investors., (2nd ed.). New York: Free Press.

Reid, A., & Catterall, M. (2005). Invisible Data Quality Issues in a CRM implementation. *Database Marketing & Customer Strategy Management*, *12(4)*, 305–314.

Schierholz, R., Kolbe, L. M., & Brenner, W. (2005). Mobilizing Customer Relationship Management—A Journey from Strategy to Systems Design. In *Proceedings of the 39th Hawaii International Conference on System Sciences (HICCS-38)*, January 2005. (p. 1112c). Los Alamitos, CA: IEEE Computer Society.

SCIP (2005). Website of the Society of Competitive Intelligence Professionals. Retrieved on September 9 2006 from http://www.scip.org/

Sukumaran, S., & Sureka, A. (2006): Integrating Structured and Unstructured Data Using Text Tagging and Annotation. *Business Intelligence Journal*, *11(2)*, 8–17.

Sullivan, D. (2001): *Document Warehousing and Text Mining. Techniques for Improving Business Operations, Marketing, and Sales*. New York: Wiley&Sons.

Vaduva, A., & Vetterli, T. (2001): Metadata Management for Data Warehousing: an Overview. *International Journal of Cooperative Information Systems*, *10(3)*, 273–298.

Vedder, R. G., Vanecek, M. T., Guynes, C. S., & Cappel, J. J. (1999). CEO and CIO Perspectives on Competitive Intelligence. *Communications of the ACM*, *42(8)*, 109–116.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. (2005). *Text Mining—Predictive Methods for Analyzing Unstructured Information*. New York: Springer.