Retrieving Business Information from the WWW

Dissertation
zur Erlangung des akademischen Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.)
an der Universität Konstanz
Mathematisch-naturwissenschaftliche Sektion
Fachgruppe Informatik und Informationswissenschaft

vorgelegt von

Gabriela Mußler

Tag der mündlichen Prüfung: 19.Juni 2002

Referenten:

Prof. Dr. Harald Reiterer, Universität Konstanz
Prof. Dr. Rainer Kuhlen, Universität Konstanz

Konstanz, im April 2002

ACKNOWLEDGMENTS

I would like to take this opportunity to thank all people supporting me throughout the period of creating this dissertation. Special thanks to Prof. Dr. Harald Reiterer for guiding me the last years throughout the development phase. Also I like to thank him and Prof. Dr. Rainer Kuhlen for their willingness to act as surveyors and their useful and most appreciated comments on it.

I wish to thank all current and former members of the working groups Information Systems and Information Science at the University of Konstanz, especially Thomas M. Mann, Georg Odenthal, Jörg Pfründer and Frank Müller. Also the input and support by the companies MIK AG (Konstanz, Germany) and Arisem (Paris, France), in particular Laurent Dosdat have been very much appreciated.

I would like to give a big thank you to Lynn Martin-Chambers for correcting my German English into British English.

My special thank goes to my partner Frank. This work would not have been realized as it is without his constant support and believe in me over the past years.

Diese Arbeit ist meinen lieben Eltern gewidmet.

Zusammenfassung

Im Rahmen dieser Arbeit werden die Retrievalaspekte eines Systems zur Suche und Gewinnung von Geschäftsinformationen aus dem WWW präsentiert. Das WWW wird als wichtige Quelle für Geschäftsinformationen angesehen. Innerhalb dieser Arbeit werden eine Reihe von Problemen und deren Lösungen präsentiert.

Einer Einführung im ersten Teil folgend, werden im zweiten Teil dieser Arbeit zunächst die Ergebnisse einer innerhalb dieser Arbeit durchgeführten Untersuchung, die das Ziel hatte, einen Einblick in den Informationsbedarf von Entscheidungsträgern in Firmen bezüglich externer Geschäftsinformationen zu schaffen, diskutiert. Die Ergebnisse zeigen, dass externe Informationen, insbesondere auch aus dem WWW, wichtig für diese Benutzergruppe sind. Aus dieser Untersuchung werden Handlungsempfehlungen für das Design eines Tools zur Gewinnung von Geschäftsinformationen aus dem WWW abgeleitet. Im weiteren Verlauf zeigt ein Abschnitt dieser Arbeit exemplarisch, welche Arten von Geschäftsinformationen im WWW vorhanden sind.

Der dritte Teil der Arbeit gibt zunächst einen Überblick über die Grundlagen von Business Intelligence Systemen. Die Ergebnisse der Untersuchung des zweiten Teils sowie das Heranziehen einschlägiger Literatur führte zur Entwicklung von INSYDER, einem visuellen Suchsystem für Geschäftsinformationen, das den Schwerpunkt dieses Teils bildet. Die Information Retrieval Aspekte stehen dabei im vierten und fünften Teil dieser Arbeit im Vordergrund, während die Visualisierungen in [Mann 2002] behandelt werden. Das vierte Kapitel schafft den theoretischen Rahmen für die eingesetzten Verfahren und zeigt zugleich deren Realisierung in INSYDER. Darüber hinaus werden verwandte Methoden und Ansätze diskutiert.

Ziel der Entwicklung von INSYDER war die Schaffung eines Systems, das dem Benutzer den größtmöglichen Mehrwert bei der Suche nach Geschäftsinformation bietet. Dafür wurden eine Reihe verschiedener Komponenten konzipiert und entwickelt. In einem ersten Schritt wird der Benutzer bei der Formulierung und insbesondere Erweiterung der Suchanfrage durch eine interaktive visuelle Suchanfrageerweiterung, der Visual Query, unterstützt. Dabei werden die Terme der Suchanfrageerweiterung einer dem System zugrundeliegenden Wissensbasis entnommen, die Visualisierung ist baumartig. Zu diesem Zweck werden Methoden der Graphenvisualisierung eingesetzt.

Für die Begutachtung und Analyse der gefundenen Dokumentenmenge wurde ein bestehendes Rankingverfahren weiter- und eine Kategorisierung der Treffer neu entwickelt. Das Ranking analysiert jedes Dokument sofort, ohne auf eine Dokumentenkollektion zurückgreifen zu müssen. Somit ist es möglich, vergleichbare Ergebnisse zu erzielen. Das Verfahren nutzt dabei die bereits angesprochene Wissensbasis. Die Kategorisierung ermöglicht es dem Benutzer, schnell zu erkennen, aus welcher Quelle die Dokumente stammen und um welche Art von Dokument es sich handelt. Dafür wurde ein Algorithmus entwickelt, der die Dokumente nach ihrem Inhalt, z.B. Linklisten oder Kataloge, unterscheidet. Die Kategorisierung der Quelle ist dabei an die jeweilige Einsatzumgebung anpassbar. Mit Hilfe der in [Mann 2002] vorgestellten Visualisierungen ist die Kategorienzuordnung schnell zu erkennen. Mit Hilfe der Relevance Feedback Option kann der Benutzer automatisch eine neue Suchanfrage erzeugen. Hierbei wird der Benutzer wie in der Suchanfrage durch die Visual Query Visualisierung unterstützt. Es ist ihm so möglich, die automatisch erzeugte Anfrage leicht selbst zu modifizieren.

Die Evaluierung der Rankingverfahren unter zu Hilfenahme der TREC Daten zeigt die Effektivität des Systems. Hierfür wurden zwei Evaluierungen durchgeführt. Zum einen eine online Evaluierung. Hier erfolgte die Evaluierung des Systems mit Dokumenten aus dem WWW und einer Suchanfrage, die den TREC Daten entsprach. Die Bewertung der

gefundenen Dokumente baute auf den vorgegebenen Bewertungskriterien der TREC Daten auf. Zum anderen wurde eine offline Evaluierung durchgeführt. Die zu bewertenden Dokumente entstammten hier der TREC Kollektion, die Bewertung ist für diese Dokumente durch TREC vorgegeben, so dass die Anwendung der Bewertungskriterien entfällt.

Als Resümee wird ein neues Business Intelligence System, bestehend aus den dargelegten Komponenten und Erweiterungen, skizziert.

Abstract

Within this work the retrieval aspects for a system to retrieve Business Information from the WWW are presented. The WWW is seen as an important source for Business Information. In this thesis a number of problems and their solutions are faced.

First the results of a study conducted within this work among business decision makers, with the objective to get an insight in their external Business Information need, are discussed. The results show that external information is very valuable for them and that Business Information from the WWW is seen to be an important source.

The outcomes of this study and the review of literature led to the development of a visual information seeking system for Business Information called INSYDER. The Information Retrieval aspects of this system are in the focus of this work, whereas the visualisations aspects are discussed in [Mann 2002].

It has been intended to develop a system giving a big added-value to the user. For this various components have been designed. The visualisation of the query for an interactive query expansion assists the user in the first step of the information seeking process. The proposed ranking and classification components support the user when reviewing results. Hereby the ranking analyses document by document on-the-fly. This way a comparable ranking, not relying on an overall document collection, has been achieved. For a redefinition of the initial query a relevance feedback option has also been included. The evaluation of the retrieval performance using TREC data shows the system's effectiveness.

As a résumé and outlook on future work the presented components and enhancements are rearranged in a sketch of a new Business Intelligence System.

Table of Contents

List of Figures

List of Tables

List of Formulas

List of Abbreviations

API	Application Interface
BI	Business Intelligence
BIS	Business Intelligence System
BSC	Balanced Scorecard
CEO	Chief Executive Officer
DSS	Decision Support System
DTD	Document Type Definition
DW	Data Warehouse
EIS	Executive / Everybody's / Enterprise Information System
GUI	Graphical User Interface
HTML	Hypertext Markup Language
INSYDER	Internet Système de Recherche
IPC	International Patent Classification
IR	Information Retrieval
IV	Information Visualization
MIS	Management Information System
MSS	Management Support System
OLAP	On-line Analytical Processing
OLTP	Online Transaction Processing
RDF	Resource Description Framework
ROI	Return of Investment
SGML	Standard Generalised Markup Language
SOI	Sphere-of-Interest
SQL	Structured Query Language
TREC	Text Retrieval Conference
URL	Uniform Resource Locator
WWW	World Wide Web
XHTML	Extensible Hypertext Markup Language
XML	Extensible Markup Language

Remark: Trademarks are respected but not explicitly marked within this work.

1 Introduction

1.1 Motivation and problem description

The benefits of using external information for business intelligence are significant. As markets become turbulent, the old way of doing business becomes less viable. Data from internal operational systems (e.g. enterprise resource planning systems like SAP R/3) are still very relevant to managing business, but the need for external information is ever increasing. An enterprise must know more and more about its customers, its suppliers, its competitors, government agencies, and many other external factors. The focus here is on electronic external information and not from personal communications. The information from internal systems should be supplemented with information about external factors.[1] This synergism of the combination creates the greatest business benefit for the enterprise. From a global perspective, the WWW is the most important resource for external business information. Valuable information about external business factors is readily available on the WWW and is increasing every hour. While a few WWW resources (e.g. direct feeds of stock quotes) are used as data sources, the immense resources of the WWW are largely untapped.[2] What is needed is a continuous and systematic approach to make use of these untapped resources. The problem is that users dealing with business intelligence systems are not trained on getting information from an Information Retrieval System, although having in mind that the relevant information is obtainable. And it is this relevant information that matters for business decision makers *"[…] delivering the basic input to executive decision making - usable and relevant information"* [Hoven van den 1996, p.5].

From a survey conducted within this work among knowledge workers[3], i.e. users of management information systems, it can be derived that there is an ongoing demand for external information [Mußler 2000]. Knowledge workers more and more often obtain information from external sources, which they need to proceed. Hereby the information obtained can not just be taken as it is, but many times has to be interpreted and formatted. This information then influence's for example reports for the management.

In the information society we face the problem of drowning in information instead of a shortcut of information. A lack of information is the result, as people are no longer able to process all the information they are given. People need information to fulfil their primary task [Preece 1994], supporting anyhow the company' s success. The overflow of information from inside the company is addressed by many companies developing various kind of information systems. These system are most of the time based on a data warehouse, consisting of well structured, mostly quantitative data. Still in this field many of the problems have to be solved, while a new problem of processing external information is already occurring. External information is often qualitative, not well structured, vague and can not be put in a data warehouse easily. There is a need to process it further. Within this work it is assumed that this further processing has to be performed by a knowledge worker in a company. So solutions for the knowledge worker in a company, e.g. in an information and documentation department, are needed to help him process external information, gaining the added-value [Kuhlen 1991] of it.

---

[1] See also [Mintzberg 1975, p.59] *"[..] into internal and external roles, for information from both sources must be brought to bear on the same decision."*

[2] http://webfarming.com/intro/intro02.html [2001-05-22]

[3] Remark: In this thesis the terms knowledge workers and of business decision makers are used vice-versa, meaning hereby those people working for a company dealing heavily with information to turn this into knowledge and somehow into a benefit for the company.

The commonly available search-engines for retrieving information from the WWW are designed for the use of a broad heterogeneous user spectrum. With this thesis I face the problem of providing knowledge workers in a company access to external information resources, in particular the WWW, having in mind the target user group: the users of Business Intelligence Systems. This is done by developing an information seeking system based on a new way to combine existing methods of research.

1.2 Solution

In the scientific community Information Retrieval has been having a long tradition on the processing of all kinds of information. Therefor this thesis combines two disciplines by applying information retrieval techniques to the sector of Business Intelligence Systems. On the one hand, various studies show that user interfaces for knowledge workers in a company must be very intuitive, which is the most important demand from these systems. On the other hand, traditional information retrieval systems have, thinking of intuitive use, often fairly poor user interfaces. Their goal is the optimisation of efficiency. A reason for this could be that the online time (connection time with database provider) is cost intensive. Within the literature numerous aspects of users interacting with IR systems for the WWW are discussed. Authors point out that users have problems formulating their information need [Pollock, Hockley 1997], [Nielsen 1997]. The presentation of information is often described as poor, e.g. just presenting a long list of search results [Zamir, Etzioni 1998], [Gudivada, Raghavan, Grosky et al. 1997], [Attardi, Marco, Salvi 1997]. According to a study by [Jansen, Spink, Saracevic 2000] of the user queries sent to the Excite[4] search-engine, the average query length is 2.21 terms, the use of query modification is not typical, 43% of the queries are identical, the number of pages viewed per user is 2.35 and only 58% access the first page of the result list. [Wang, Hawk, Tenopir 2000] identify in their study that the URL is the most used metadata element. [Choo, Detlor, Turnbull 1999] point out that users rarely make use of help pages. Apart from these problems addressed in the literature, valuable insight could be gained how business decision makers deal with external information by conducting a survey amongst the customers of a Konstanz based producer of Management Support Solutions.

The current work addresses some of these aspects, trying to find a solution for the problems shown. Within this work an Information Assistant [Kuhlen 1999] approach is proposed. The Information Assistant acts on behalf of the user, who is most likely to be an assistant himself (e.g. assistant to Financial Directors, member of an information and documentation department etc.). Various methods from the disciplines Information Retrieval and Human Computer Interaction have been investigated and combined for a new approach. This could have been implemented and partly evaluated in a research prototype within the EU project INSYDER[5]. The objective has been to maximise the added-value. For this the information seeking process is supported in various stages. In particular a visualisation for an interactive query expansion is discussed within this work and has been prototypically implemented, two ranking algorithms for an online ranking have been implemented and evaluated, a classification scheme for the result set and a relevance feedback option are also integrated into the system. Hereby the system is very much based on a pre-provided content. That is to say that a Knowledge Base and a knowledge of the application domain are the premises. This approach proposes a system capable of sharing information sources and using existing domain knowledge.

---

[4] http://www.excite.com [2001-09-03]

[5] The research project INSYDER was funded by a grant from the European Union, ESPRIT project number 29232.

Aspects concerning the analysis of search results are covered by [Mann 2002]. However this process is supported by the categorisation proposed in here (section 4.7).

The current solution addresses knowledge workers in a company, who are for example supported by the information and documentation department. Companies have a strong interest in sharing information resources and using the same language for their knowledge. Here knowledge is seen as the domain knowledge about the market segment the company acts in. To some extent this domain knowledge is externalised, e.g. when defining data models for business applications. The sharing of information sources and externalisation of knowledge is nowadays summarised with the term Knowledge Management, comprising the process of externalising knowledge, modelling it, finding ways to motivate the employees to share information and so forth [Probst, Raub, Romhardt 1997]. Within the present work this whole topic is out of focus, however the aspects of using existing domain knowledge and sharing the information are considered. The solution proposed to help business decision makers to retrieve relevant business information relies on the existence of a content based system. Hereby the content is for instance modelled in the Knowledge Base, containing terms and relations concerning the company and its environment (e.g. customers, suppliers, technology).

The information need of a business decision maker and the company he works for is likely to be sensitive. This leads to the demand of an independent solution for the satisfaction of the information need using the WWW as a resource. Otherwise it is (theoretically) possible to build up search profiles to use them for advertisement or in the worst case industrial spying.

A general requirement of an information seeker is that information is up-to-date and available. With the design of search-engines, this is not always possible as they rely on an index, which is updated from time to time, depending on the search-engine. Therefore users often find an out-of-date index, which results in the listing of a document that is no longer available or that it has been updated and is no longer available in the requested version. Search-engines like Google offer for these cases also a cached version of the retrieved document, highlighting the search keywords in that cached version.

A conclusion of all the above mentioned is to have a tool that does a *dynamic search*, which has been implemented with the INSYDER system. The idea is to use an own crawling and analysis to do an online search discovering up-to-date and available relevant information. A main advantage is that the current structure of the Web is searched, and not the index of a search-engine. The definition of heterogeneous sources, however relying on a hypertext structure like search-engines, Web directories, various Web sites, electronic market places and so on, build the entry-point for the link traversal. For example the query terms are submitted to pre-selected search-engines or catalogues (like Altavista, Yahoo! etc.) and the hyperlinks in the search results are then used for further crawling on the Web. All documents found are downloaded and analysed incrementally to find out how good they match the query. This way it can guarantee that the documents presented in the result views are up-to-date. A similar approach used by *Inquirus* [Lawrence, Giles 1998][Glover, Lawrence, Birmingham et al. 1999] also performs an online analysis and an own and therefore consistent ranking of documents found by search-engines, but it is designed to be a mere meta search-engine, as it does perform any further crawling starting with the documents found.

By design in the INSYDER system the query is not looked up in the own repository, which could give a first hint of results (with the disadvantage of knowing these would not be up-to-date). This means that it takes some time until the search results are available, al though they may be available in the own database (but maybe in an out-of-date version!). However the major advantage of this approach is that the query is

processed by an own system, that only specific sources might be queried, that the search results are up-to-date and comparable as all documents are analysed and given a consistent ranking. Unlike commercial search systems it has not been intended to crawl all the WWW and store its contents, but only dedicated parts, which are potentially relevant for a given query by a user. By this way of specialising the search by focusing the crawling, e.g. for a specific branch like CAD systems, it has been expected to increase the precision and recall compared to other meta search-engines, which rely on the results from the search-engines' indices.

A general important aspect of INSYDER thinking of its novelty is the fact that ideas and components from different fields were combined. It is certainly not new to combine visualisations and information retrieval, but nowadays systems performing a dynamic search with a metadata generation using a content based system and the different visualisations of this metadata and inherent document data are new. The approach aimed at getting the biggest added-value for the user combining components like dynamic search, visualisation of the query and different visualisations of the results (see [Mann 2002]) and information retrieval techniques (e.g. query expansion, ranking of results) in the context of Business Intelligence Systems.

1.3 Overview of this Thesis

This introduction is followed by the description of external information and their specific characteristics (chapter two). Describing the results of a study conducted among business decision makers the handling of information (external and internal) will be looked at in one particular example. Chapter three deals with Business Intelligence Systems. Beside various definitions an overview on related systems are given. As one tool in the suite of Business Intelligence Systems an overview is given on the EU project INSYDER, which has been used as the platform for the proposed solutions. From a process point of view Web farming as a process of integrating external information into data warehouses will be explained. Chapter four discusses the proposed methods to support the retrieval of business information from the WWW. Therefore the solution is embedded in the theoretical background of Information Retrieval. In chapter five the evaluation results of selected issues are presented. A summary and an outlook to future work in chapter six concludes this work .

Chapter	Page	Goal	Content
1	1	Introduction	Description of problem and proposed solutions
2	5	Overview on external information, MIK study and INSYDER	Explanation of external information, presentation of the results of study conducted within this work
3	42	Overview on Business Intelligence Systems	Description of systems, target user group, information obtainable from the WWW; INSYDER as a tool for integrating external information into Business Intelligence System
4	59	Getting to know an answer to problems addressed	Description of the proposed solution for the retrieval of business information from the WWW, discussion of theoretical background
5	152	Overview and Details of Evaluation	Description of evaluation and results of the ranking algorithms
6	175	Summary and Outlook	Summarising work, giving an outlook to future issues

Table 1-1: Overview on thesis

2 Need of External Business Information from the WWW

The development of the Internet and its value-added services like the WWW has changed the way business enterprises deal with business information. The necessity of using the Internet as an information source in decision making situations is increasing more and more. *"The Internet [...] is becoming the major supplier of external data for many decision situations"* [Turban, Aronson 1998, p.114]. However little is known how knowledge workers deal with external information. A survey from Herget and Hensler conducted in 1993 had the focus on the survey of using external information from online databases [Herget, Hensler 1995]. At the time this study was conducted it is clear that there could be no question dealing with the WWW. The study conducted within this work has been conducted to find out how business decision makers deal with information coming from external sources, focusing on the WWW. A number of actions are assumed, when thinking of dealing with information:

- analysis of sense
- analysis of sources
- analysis of content
- analysis of correlation and consequences.

To gain an insight a study amongst business decision makers was conducted in December 1999/January 2000. The study was carried out in co-operation with the MIK GmbH Konstanz[6]. The subjects have been the customers of the MIK GmbH.

After a discussion of the external information, possible sources and its impact on business uses, this chapter will present the results and conclusions of the study carried out.

2.1 External information

2.1.1 Introduction

In 1962 Machlup propagandised the knowledge sector, in which structures will change profoundly. Not the production of goods, but the production of information will be in the focus [Machlup 1962]. And Drucker identifies in 1969 knowledge as a key-factor. *"What matters is that knowledge has become the central ‚factor of production' in an advanced, developed economy."* [Drucker 1969, p.269]. In contrast to this, business decision makers experience today an information paradox. Commonly spoken, nowadays they receive too much information and therefore they can not distinguish between relevant and irrelevant information. They can not turn the information into knowledge, for they can not process all the input they receive. Or as Drucker states *"But what matters in the 'knowledge economy' is whether knowledge, old or new, is applicable, e.g., Newtonian physics to the space program."* [Drucker 1969, p.269]

The focus in the present work is on electronic external information and not on personal communications. Still these personal contacts have a great impact on the way the business decision makers do their job, but here are out of focus. Managers cultivate a variety of external personal contacts *"largely to find information. In effect, the liaison role is devoted to building up the manager's own external information system –*

[6] http://www.mik.de

informal, private, verbal, but nevertheless, effective." [Mintzberg 1975, p.55] Drucker demands the collection and organisation of outside information, claiming that *"All the data we have so far, including those provided by the new tools, focus inward. But inside an enterprise – indeed, even inside the entire economic chain – there are only costs. Results are only on the outside. […] to focus inward on costs and efforts, rather than outward on opportunities, changes, and threats."* [Drucker 1998, p.51f]

Business decision makers are primarily knowledge workers, scanning the organisations environment e.g. for competitors' performances, business opportunities, information on suppliers and customers [Drucker 1995]. Consequently they use knowledge in the decision making process [Turban, Aronson 1998]. [Picot 1989] sees information as the production factor, which is ahead of all other production factors. He points out that the chance of a successful company leading strategy lies in the ill-distribution of information, knowledge and skills in the business. Having more relevant information than the others gives the possibility to act on the markets for the wealth of the company more efficiently and more successfully. Under this point of view Picot classifies information management as an integral part of enterprise control and therefore business decision making.[7] So there is obviously a focus on information and knowledge.[8] It does not seem to be easy to define information, e.g. [Kuhlen 1989], [Losee 1997]. Kuhlen gives a terse definition: *"information is knowledge in action"*[9] [Kuhlen 1995, p.34], having the *"pragmatic primate"* in mind. Common human knowledge can only become information if the individual context of using it is taken into consideration [Kuhlen 1999]. Having no precise definition of information leads to its definition by attributes as for example found in [Wilson 1995], [Kuhlen 1989], [Picot, Scheuble 1997].

Another view of information is to see the process of obtaining the information, e.g. reading a book, it is not the number of pages read which are received, but the mental process of understanding and integrating the read pages into our own personal knowledge structures [Wilson 1995]. The transitory attribute of information leads the user also to a new kind of information access: It is important to recall the search strategy which took the user to the information. This can either be the book and its location or more sophisticated a search and browsing activity in the WWW which leads by the serendipity effect [Kuhlen 1991a] to a previous unexpected information.

[Picot, Reichwald, Wigand 1996] argue that the cognitive features of humans are commonly not enough to process and analyse the presented information, e.g. tasks or solutions. [Picot 1989] stretches the fact that an information need can be objective or subjective (see also section 4.1). While the first is clearly defined by a task, the second depends on the business decision maker, not necessarily (often very different) from the objective information need.

2.1.2 Definition of external information

Following [Biethahn, Fischer 1994] internal information is information which originates inside different company departments. They are directly related to the objective of the company. Within this work external information shall be defined in delimitation to internal information. In contrast to the internal information, external information is produced outside the company, but as the internal they somehow also have an influence on the company's objective. Putting it this way we have a distinct separation. This way

---

[7] See also Mintzberg "*Information is not, of course, an end in itself; it is the basic input to decision making.*" [Mintzberg 1975, p.56]

[8] Already Mintzberg stated, that "*Information in turn, enables the manager to make decisions and strategies for his unit.*" [Mintzberg 1975, p.54]

[9] *"Information ist Wissen in Aktion"* (translation by authoress)

it is also clear that information from within a group will be treated as internal information. Hereby all management levels are concerned: the operational short term oriented management, the middle term oriented administrative management and the long term oriented strategic management. Figure 2-1 shows some of the possible information influences, quantitative and qualitative, like stock markets, data from competitors, Email, personal contacts by telephone or in conferences, printmedia or from the WWW (left to right).

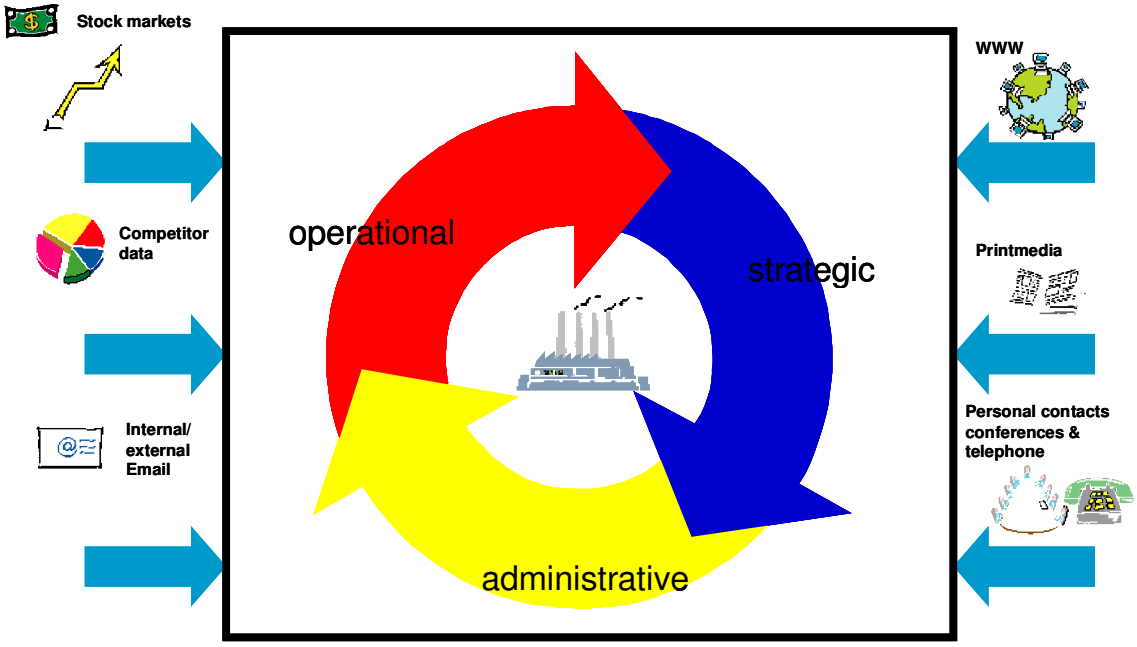


Figure 2-1: Internal and external information

Another classification is to look at information as hard and soft (also called fuzzy or informal [Lester 1989]) facts. While the former classification was discrete, a classification in hard and soft facts is continual. Much of the internal hard facts are extracted from internal existing databases, e.g. OLTP systems like SAP R/3. The use of hard facts is not straightforward. The information which is there needs further analysis, for reporting and updating cycles could vary, incompatible databases or misinterpretation etc. [Watson, Rainer, Koh 1992]. Soft facts often come from human sources, e.g. rumours, news items, explanations etc. [Behme, Mucksch 1999] and are often critical to understanding problems [Mintzberg 1975][10], [Zmud 1986], [Lester 1989]. Often soft facts provide an added-value *"[…] soft information enhances the understanding of past, current, and future events, often by adding value to factual data."* [Sprague, Watson 1996, p.302]. Soft facts are often not stored in the company systems, but a lot of the time just in the heads of the employees. Nowadays efforts are made to get more out of soft facts.[11]

With the development of the Internet the result is more and more external information. External information is now not only information from databases, which is in the case of commercial databases somehow structured, but is now often just unstructured information. The demand to integrate external data has existed for a long while [Runge 1988], [Meyer-Piening 1987]. *"Collecting, analysing, and entering these data [soft*

[10] See [Mintzberg 1975, p.51f], e.g. "*Managers seem to cherish 'soft' information, especially gossip, hearsay, and speculation. Why? The reason is its timeliness; today's gossip may be tomorrow's fact.*" [Mintzberg 1975, p.52

[11] These efforts can be subsumed under the term Knowledge Management.

facts[12]] to an EIS tends to be very labour-intensive but adds considerably to the richness of the information provided.“ [Watson, Rainer, Koh 1992, p.93].

The potential for innovation from external information is high and has to be seen equally to the management of internal information [Picot 1989]. External information on the WWW for example is *"just there"*. It has to be retrieved – assuming it is there and further processed. E.g. sales reports from concurrent companies showing a lot of (hard) data can not just be transferred into an own database as they are, but one has to have a close look at the numbers. For instance if the report periods (e.g. what period comprises the financial year) are the same, if the same measurement units (e.g. scales of meters or feet) are used etc. So here the original hard data becomes soft as further processing and interpreting steps are needed. Standards like XML (see section 4.4.5.1) and related aspects (see section 4.4.5.2) are an approach to help to overcome this problem.

Looking at external information it becomes clear that there are several problems with particular attributes. As information is a special kind of good (for a discussion see [Mußler 1997]), attributes like reliability come into focus. Reliability is strongly connected to trust and can be characterised by the source. Business decision makers are highly dependent on the information they receive and to make the necessary decisions they have to have trust in the source. However, there is a whole variety of sources especially when thinking of external information. Press releases from the ministry, newspapers, information brokers, libraries etc. By design the financial department is the traditional place in a group having an information (and in case of the controlling also a co-ordination) task [Hoitsch 1997]. However they can only work with the data that comes from underlying systems, so if anything goes wrong there, then the data is wrong and so any information derived from this data is also wrong. Therefore the information task is not only presenting the information, but also analysing it, e.g. in terms of its plausibility. Having to deal with external information is by far a more difficult task. [Drucker 1995] identifies four types of information (foundation, productivity, competence and resource-allocation information) which are required to enable executives to make informed judgements so that the objective of the company or group to create wealth and added value can be achieved. For Drucker, information (both the corporation's integrating systems as well its articulation) is also the new skeleton of companies which they are designed around.

Figure 2-2 shows a portfolio for the grade of reliability of hard and soft information, taking into account the source (extern / intern). While internal hard facts (e.g. data on stock, turnover) can be seen to be very reliable, the same information from an external source (e.g. some statement on some WWW site) is much less reliable. Having soft external information the degree of reliability is the smallest. An example for the latter is for instance a rumour, which may have a true kernel. Trusting the source of the rumour will have a great impact for the business decision maker on using this information. But the more tele-media-services, e.g. software agents (see section 4.9), are going to be used for this information tasks, the more difficult it is going to build up trust [Kuhlen 1999, p.111].

[12] Annotation by authoress

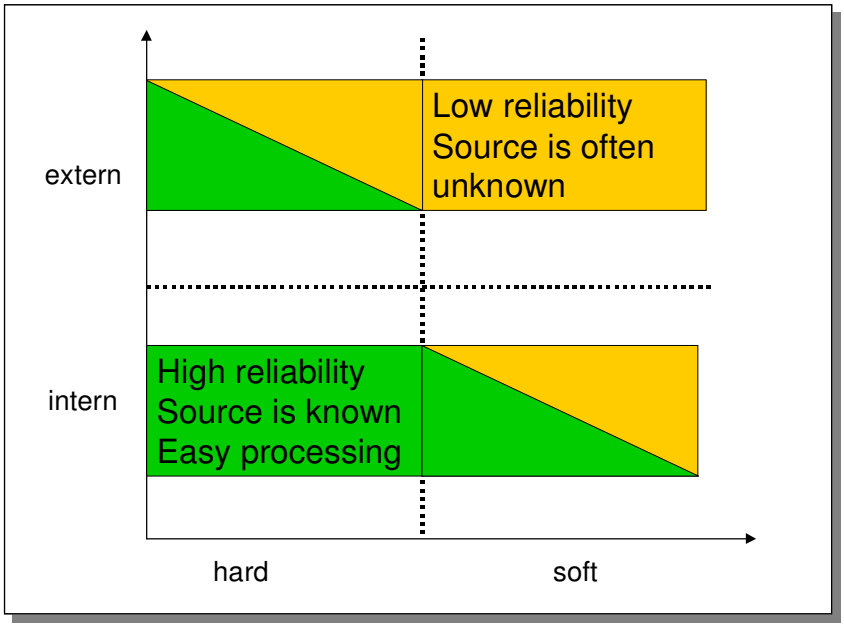


Figure 2-2: Information portfolio

2.2 A study amongst business decision makers

In November 1999 a survey was conducted with the aim to gain an insight of how business decision makers deal with information from external sources. Focus was placed on the handling of information from online sources, like the WWW. Hence three hypothesis were proposed:

1. External information is becoming more and more important in companies.
2. The pre-condition for the use of external information in a company, is the technical infrastructure to connect to the Internet and the open-mindedness towards the use of the WWW.
3. The demand for and the kind of external information is different in different branches and different departments. Whereas the difference in the departments is greater than the difference between branch-types.

For the evidence of these theses 528 questionnaires (paper versions) were sent out to the customers of MIK GmbH Konstanz[13] the partner of the working group Information Systems at the University of Konstanz in the VIEVAMIDES[14] project. *"MIK is a consultancy and software company helping companies in almost every industry and service sector to create advanced management information and control systems."*[15]

104 responses were sent back, that makes a total of 19,7 percent of all potential subjects. In the following the term subjects will represent these 104 answers.

2.2.1 Background methods and design of the survey

Several methods from the literature for the determination of the information needed are known (see e.g. [Schwuchow 1995, p.128-130], [Heinrich 1996]). [Krecmar 1997, p.56-59] gives a taxonomy following the nature of the methods: subjective, objective or a mix of the both. [Watson, Frolick 1992] describe several methods for the determination of information requirements with the focus on executive information systems. Their

[13] Since 1st March 2001 MIK has changed its status to MIK AG, a joint stock company

[14] Visualisation and Evaluation of Management Information for Decision Support.

[15] http://www.mik.de/WebSite/MIK-Site.nsf/E/Unternehmen [2001-04-05]

taxonomy is based on a portfolio determined by computer related versus noncomputer related and the type of source (see Figure 2-3).

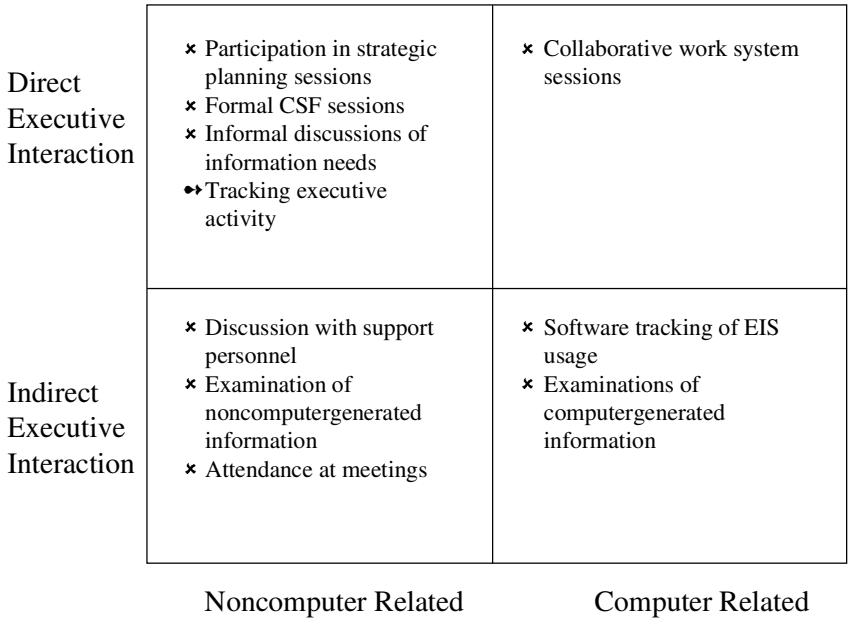


Figure 2-3: Taxonomy of [Watson, Frolick 1992] for determining the information requirement

Watson and Frohlick draw the conclusion that the determination of information requirement especially for executive information systems is in principle difficult, as executives do not have the time for such processes, that executives lack to formulate their information need and that the problem of every method is that trust is the base for any questionnaire in this field, as the company data is mostly sensitive. As the aim of the present study was to get an insight how business decision makers are dealing with external information, it has been chosen to do a mail survey and not selective interviews to gain a better insight in the whole target group [Laatz 1993]. A negative point about mail surveys could be the small number of responses. To maximise the return of questionnaires a number of methods are proposed to increase the number of responses. Most of these methods are based on the Total-Design-Method (TDM) of Dillman [Diekman 1995]. *"The credo of the TDB is, to design each aspect of the written survey in such a way that the quality of the answers and the rate of surveys returned is maximised."*[16] Also in the present survey these suggestions were taken into account, as they had been proven before to work out very well [Bohnert, Birkelbach, Grossman et al. 1997]. The guidelines include every aspect of survey design, in particular

- design of the envelope,
- design of the inquiry,
- formulation of the cover letter
- planning of additional mail actions to raise the number of responses.

For the checking of the questionnaire a pilot study with five users (three from the University of Konstanz, one from MIK and one financial controller, not taking part in

[16] [Diekmann 1995, p.442] *"Das Credo der TDM lautet, jeden Aspekt der schriftlichen Befragung derart zu gestalten, daß die Qualität der Antworten und die Rücklaufquote maximiert wird."* (translation by authoress)

the final study) has been conducted, which led to a partial redesign of the questionnaire, with the objective to eliminate errors and misunderstandings.

2.2.2 Results of the survey

This part describes the findings of the surveys in detail. The answers are organised in several groups. One directive when designing a survey is to put general questions as an easy entry-point to the survey at the beginning. The original questionnaire and the inquiry are attached in the Appendix.[17] The results presented hereafter have been translated most carefully by the authoress, in spite of this especially thinking of the classification used in the single questions there could be minor differences to the original German version. A detailed discussion of the results in German language are presented in [Mußler 2000].

2.2.2.1 General data

The originating country of the subjects was in most of the cases Germany (85%), followed by Switzerland (11%), Austria (2%) and Hungary (1%). 2% of the subjects did not specify the country.

Whereas 2% are working for agriculture and forestry businesses, 56% for the industry and 42% for commerce and service businesses (see Figure 2-4). Compared to the distribution of employees in these business sectors in Germany [Statistisches Jahrbuch 1999, S.21], where 2,9% work for the primary sector, 33,8% for industry and 63,3% for commerce and services, the deviation is remarkable. Still this is easy explained, as the customer structure of the MIK does not resemble this either. Also the greatest part of the potential subjects were suppliers of energy[18] (18%), therefore it is also no surprise that the main part of the subjects (13%) are also from this group.

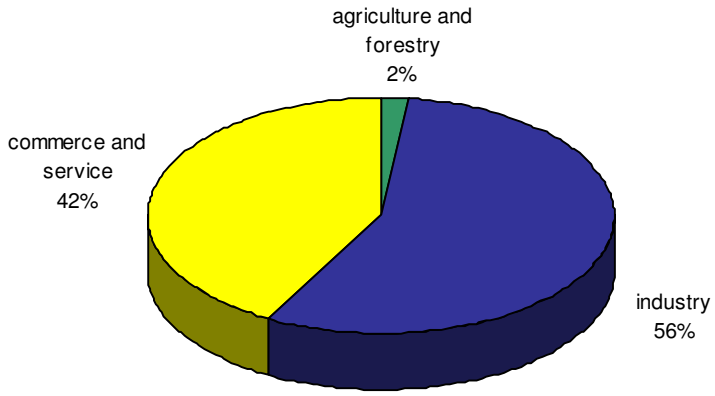


Figure 2-4: Distribution of subjects to business sectors (n=98)[19]

In most of the cases heads of departments, senior heads of departments, managing director and confidential clerks have filled in the survey (see Figure 2-5).

---

[17] Note: the survey's language was German

[18] *Stadtwerke (note of authoress)*

[19] n is the number of correct answers to that question and hence the basis for the percentage numbers

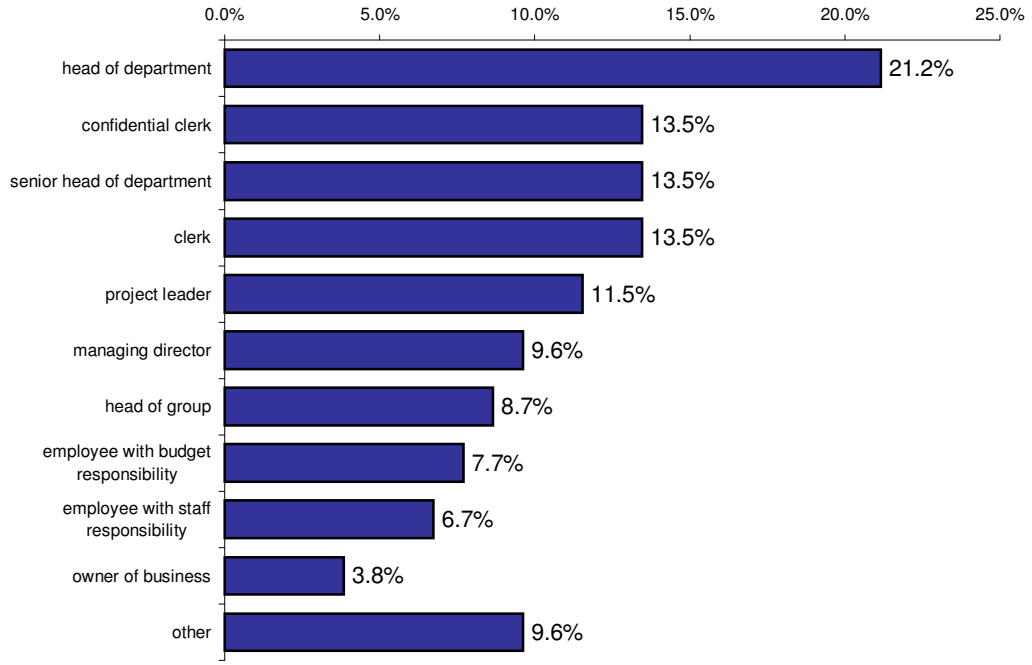


Figure 2-5: Position of the subjects in their companies (n=104, more than one answer was possible)

Most of the subjects work in the accounts department. This rather high number is because the target user group for the MIK product are accountancies. For the present survey this seemed to be no problem as they tend to be the user group for Business Intelligence Systems, too. Apart from accounting, managing directors answered the most. Therefore it can be stated that the target user group has been reached well.

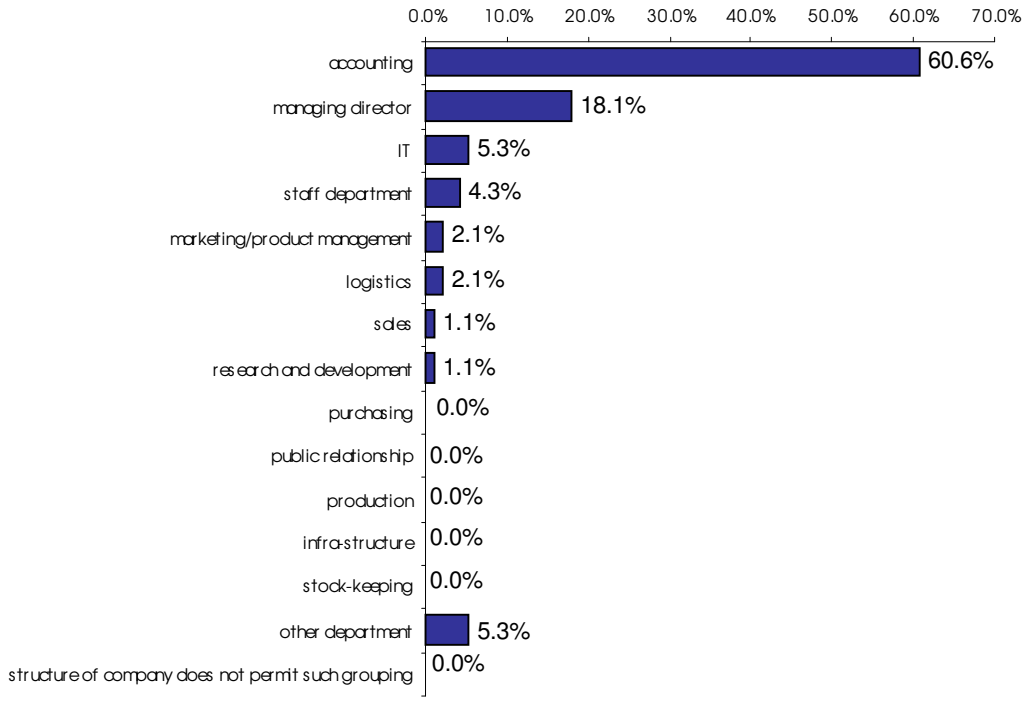


Figure 2-6: In which departments do the subjects work (n=94)

The number of employees is in 63,7% of all cases less than a thousand. In 24,5% of the companies the number is between 1.000 and 10.000 and 11,8% of the correct returned

questionnaires have more than 10.000 employees (see Figure 2-7). In comparison to the numbers for Germany (West) from 1987[20] (99,87% of all companies have less than 500 employees, 0,13% have more than 499 employees), the number of subjects in the current study working for companies with more than 500 employees is very high (55,9%). This is due to the circumstance that in a German wide survey the primary sector would play a much greater role than in the current survey (see also Figure 2-4, for the distribution of the subjects and the business sectors). Another point is that for the use of a Management Support System (as provided by MIK) in most cases a special size of the company is presumed.

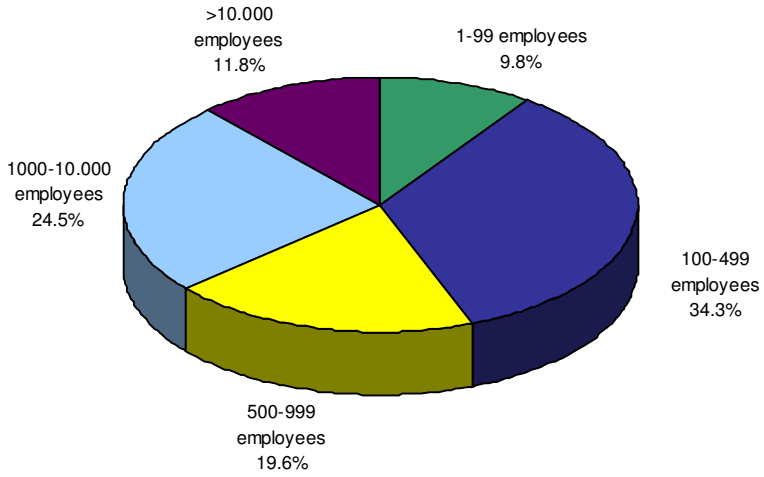


Figure 2-7: Number of employees of the companies the subjects work in (n=102)

[20] The data has been taken from [Statistisches Jahrbuch 1999, S.130] which unfortunately does not list newer data.

2.2.2.2 IT experience and infrastructure

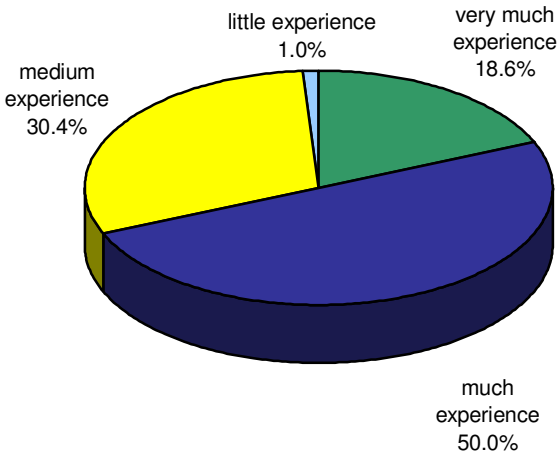


Figure 2-8: IT experience (n=102)[21]

Nearly all the subjects are at least at an intermediate level when thinking of IT experience. 68,6% judge themselves to be experienced users. Only 1% state to be a novice user.

93% of the companies the subjects work in have a company wide network (n=104). From these 93%, 65% of the subjects say it is an Intranet, 3% do not know if it is and the remaining 32% state that their network is not an Intranet.

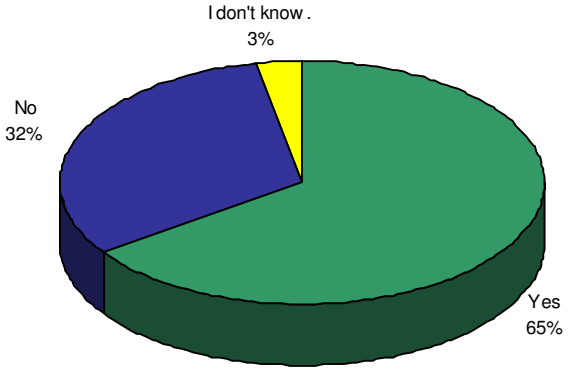


Figure 2-9: Is there an Intranet in your company? (n=95)

2.2.2.3 Using the WWW

At the time of the study 69% of the subjects have used the WWW at work. 13% say that they use it but only at home. 18% are not using the WWW (see Figure 2-10).

In a study conducted by the German TV stations ARD and ZDF from 1999[22] 49% of the cases state that the reason to have a WWW connection privately is, that they need the

[21] Remark: 0% stated very little experience.

[22] http://www.das-erste.de/studie/ [2001-04-12]

WWW for their job. In comparison: in a European study it was found that Germany has the highest percentage of Net-connected companies (74%), followed by the United Kingdom (70%) and France (55%) [IBM 1997].[23]

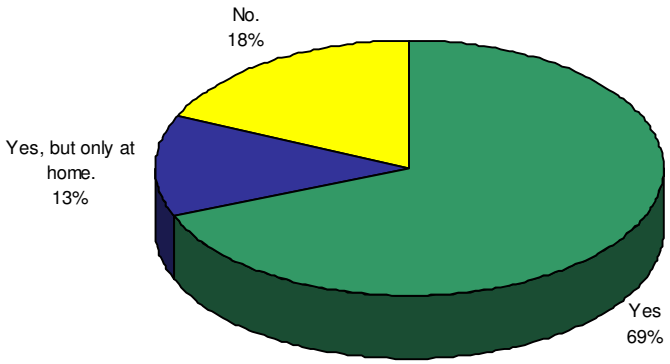


Figure 2-10: Subjects using the WWW (n=103)

When it comes to the frequency of using the WWW, 35% state that they use the WWW daily, 39% more than once a week and 17% more than once a month. Only 7% use it once a month. 2% cannot give an estimation, which is luckily a rather small number (see Figure 2-10).

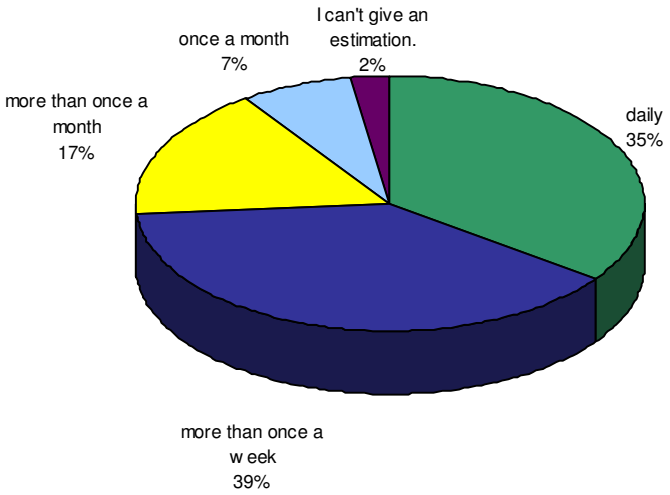


Figure 2-11: Frequency of using the WWW (n=83)

Asked about which browser the subjects use, 75,6% state to use always the same browser. Thereof 52,7% use a browser from Netscape and 47,3% a version of Microsoft (see Figure 2-12). This question was an open question, but no other browsers were mentioned by the subjects.

[23] European business on the Net http://www.ibm.com/news/wsj/s1.html [2001-05-09]: the survey was conducted by telephone in April 1997 among board-level executives from companies in the UK (n=200), France (n=203) and Germany (n=200). Participants came from different industries.

In a study of Zona Research of April 1999 the result is that 62% of the companies use Microsoft's Internet Explorer as a standard browser, while only 38% use a browser of Netscape [InformationWeek 1999, p.10]. The study of Zona Research shows also that 69% of the responding companies have an explicit browser strategy. Having in mind that the study was conducted in 1999 it is also interesting to compare the results with recent studies from 2001, showing a leadership in browser usage of the Microsoft Internet Explore. One possible explanation is the browser strategy of companies, another the politics by Microsoft, that some of its products are strongly related to the browser software (e.g. Netmeeting 3.01 is not installable until the MS Internet Explorer is installed).

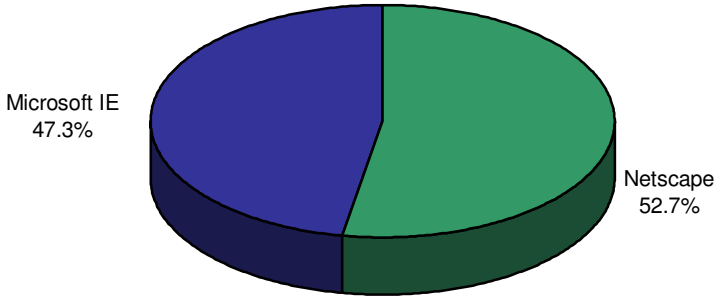


Figure 2-12: Used browser (n=55)

In most of the companies (almost 70%) the connection to the WWW is only for selected employees (Figure 2-13). Unfortunately the design of this question was misleading. The intention of the question was to ask about the WWW connection availability for business decision makers, while the answer showed "all employees" instead of "all business decision makers".

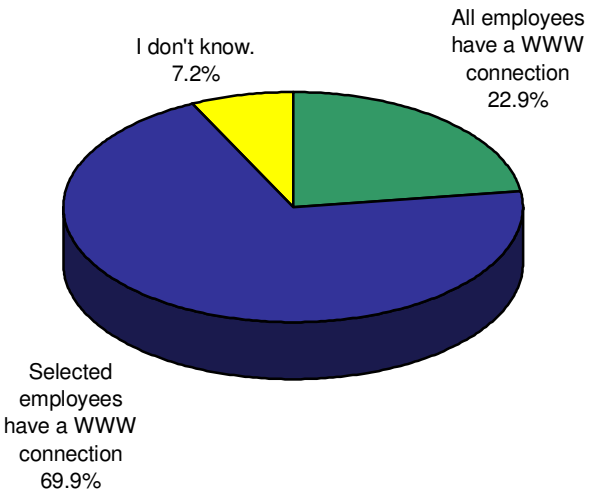


Figure 2-13: WWW connection policy of the companies (n=83)

Summarising the results of the WWW survey section, it can be said that the major part of the subjects have good experience in using the WWW. Whereby nearly three quarters (74%) of the subjects should be very familiar with the WWW as they use the WWW on a regular basis. Interestingly the number of subjects who think that they are experienced users (both very experienced and experienced) in IT are 69% and only a little less than the numbers of using the WWW.

2.2.2.4 Information Supply

The section about information supply had the aim to ask about sources and their importance of information to the business decision makers. Other studies like [Becker 1997], [Staudt, Bock, Mühlemeyer 1992] report printmedia, fairs, informal contacts and external information brokers as the main information sources for companies.

Asked about the use of WWW pages with costs, 20% of the present study answered that they were using such sites, while 76,5% answered that they were not using them (see Figure 2-14). When asked about the kind of information the business decision makers obtain from the WWW the subjects answered bank and stock information, access to commercial databases (e.g. Hoppenstedt, Genios, STN, Dialog, Datatar), newspapers and their archives, news services (e.g. Reuters).

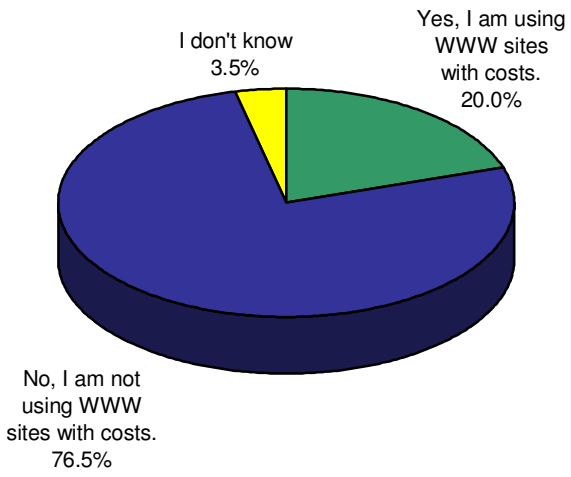


Figure 2-14: Use of WWW offers with costs (n=85)

2.2.2.4.1 Importance of sources generally

The following figures show how the subjects ranked the importance of sources depending on the decisions they have to make. The order is the same as in the questionnaire. A green colour indicates an importance of the source, yellow is undecided and blue shows that this source is either less or unimportant. If the source is not used at all this is marked in red.

Printmedia seems to be a very important source of external information (see Figure 2-15). Especially specialised journals (85% state that they are very important or important) and textbooks (71% state that they are very important or important) are outstanding. Also important are German newspapers (14% of the subjects state very important, 51% important). Foreign and weekly newspapers play a minor role, 35% (foreign newspapers) resp. 10% (weekly newspapers) state that they do not use them at all. For 32% of the subjects weekly newspapers are less important or unimportant, while foreign newspapers are to 36% of the subjects less important or unimportant.

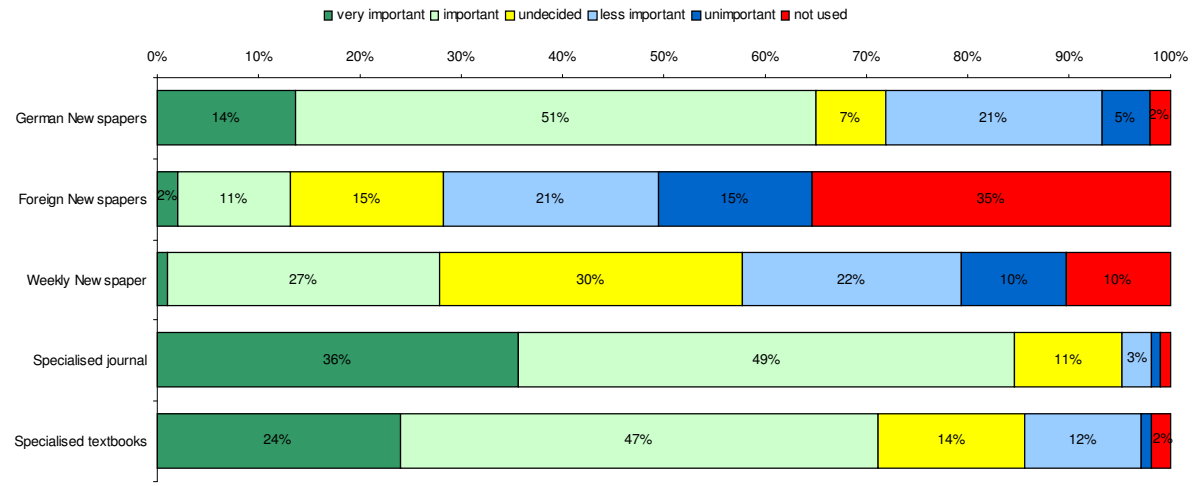


Figure 2-15: Importance of printmedia[24]

Being asked about the importance of personal information (see Figure 2-16), colleagues turned out to be the most important: 94% say that they were very important or important. Conferences are in 60% of all valid cases very important or important, while fairs are only to 31% most important or important. Nearly a third of the subjects are undecided about conferences. Remarkably 44% state that fairs were less important (25%), unimportant (10%) or not used (9%).

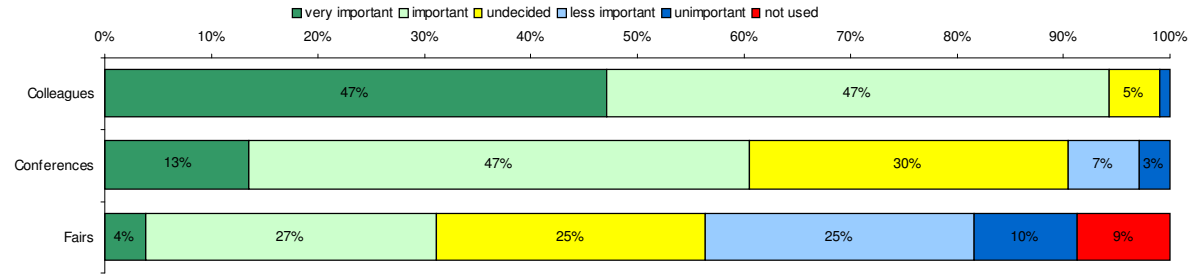


Figure 2-16: Importance of personal contacts

Press announcements from various institutions are seen diversely (see Figure 2-17). Most important are the announcements of federations (60% state these as very important or important) followed by announcements of competitors (accordingly 57%). Remarkable is the fact that information from the chamber of commerce is either not used at all (16%) or seen as less or not important (with the highest percentage of 35%). It is the only type of announcement, which is not stated as being very important. An explanation could be the circumstance that the subjects are often not belonging to small and medium sized enterprises (SME), who are normally in the focus of the information delivery of the chambers of commerce.

[24] Remark: In Figure 2-16 until Figure 2-21, 1% values are not labelled in the charts.

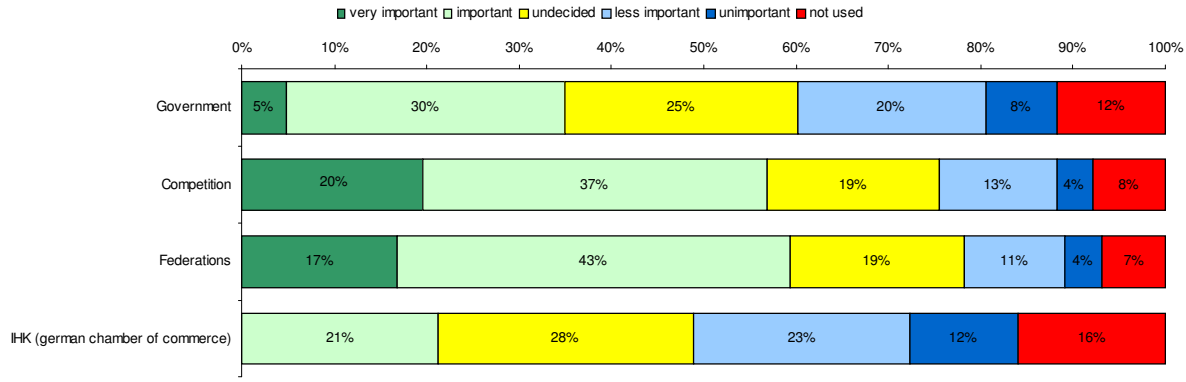


Figure 2-17: Importance of press announcements

Among the information service providers the companies internal information service is seen as the most important (see Figure 2-18). Nearly half of the subjects (49%) state that their service is very important or important for their decision making. External libraries and external information bureaus (brokers from banks etc.) play a minor role with a quarter of the subjects answering to use these not at all. An explanation for this could be that the internal information service works very well and covers tasks that are normally performed by the external institutions. However this could also mean that the internal information department orders some services from the external organisations.

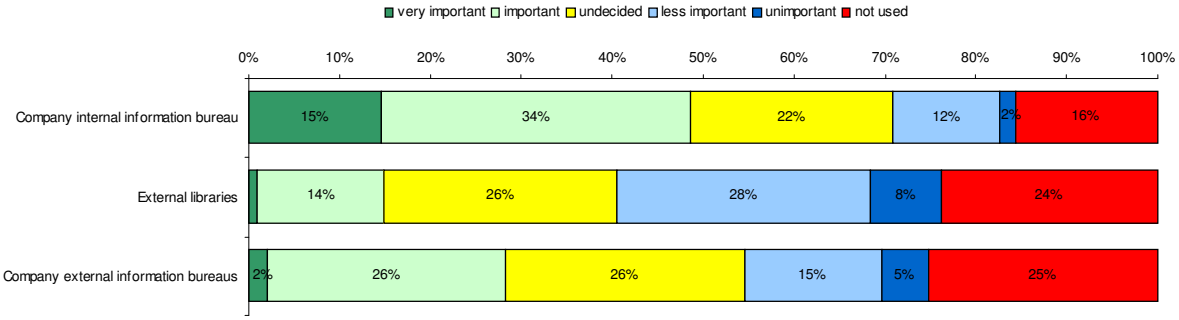


Figure 2-18: Importance of information services

Overall seen Email plays an important role for the business decision makers (Figure 2-19). The importance of internal Email is with 76% not remarkably higher than the importance of external Email (65%).

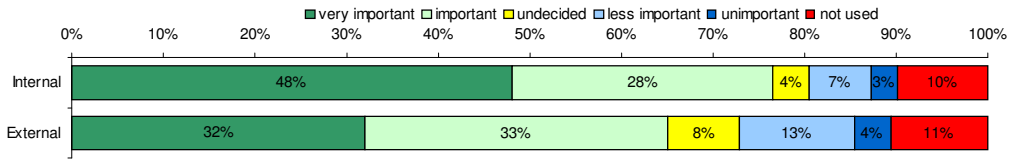


Figure 2-19: Importance of Email

Electronic media has been divided in eight categories (see Figure 2-20). Within these the Internet is seen as the most important medium for the business decision makers (61%). Electronic newspapers and magazines as available in the WWW play a minor role. However it can be that some of the subjects subsumed these two categories under the Internet category not being aware of the different kind of sources the WWW offers. The distinction against the online databases has been made on purpose to see their potential from the view of the subjects (see Figure 2-21). Here the importance is very

low, in more than a third of all cases they are not used at all. Summarising the importance of electronic media and online databases it is remarkable that at the time the study was conducted the Internet is already the most important electronic medium. It is presumed that under the term Internet many sources are subsumed by the subjects (by having the same user-interface as a WWW Gateway), which could mean the importance of online databases is higher.

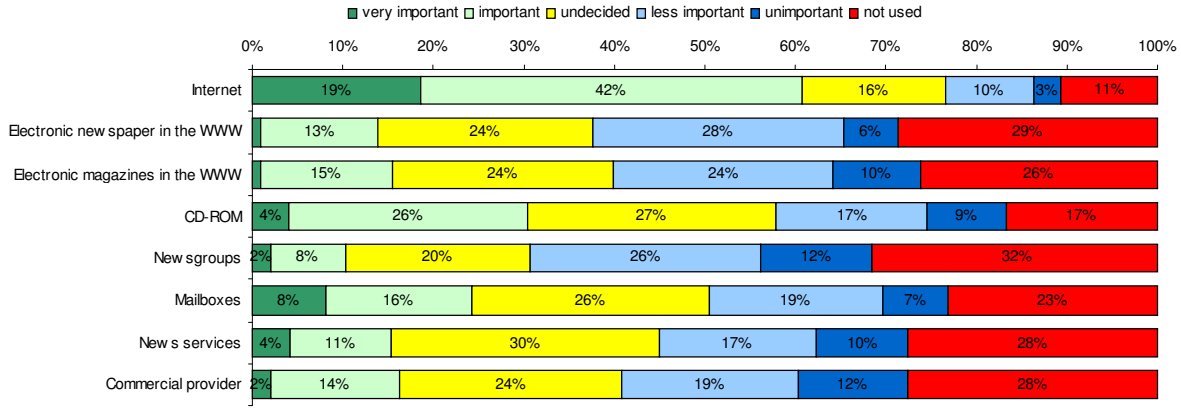


Figure 2-20: Importance of electronic media

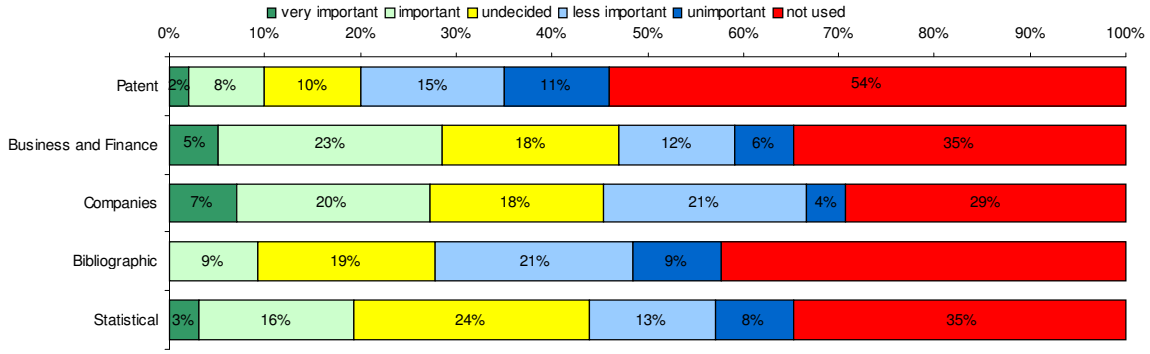


Figure 2-21: Importance of online databases

2.2.2.4.2 *Importance of sources ranked*

After the detailed view of the importance a general view shall be given in this section. Therefore two rankings are presented: the first (see Figure 2-22) showing the sources which are *very important* for the business decision makers and the second (see Figure 2-23) showing those which are either *very important* or *important*. The distinction of very important and very important/important has been made to show the assessment of the electronic sources more precisely. As can be seen in Figure 2-22 internal Email is seen to be more *very important* than personal contact with colleagues. The positions of external Email and the Internet in this ranking scheme are also remarkable. External Email ranges just behind the colleagues and specialised journals, while Internet ranges before the internal information services.

A similar distribution can be seen when looking at the ranks for the sources being classified as *very important* or *important*. Here the Internet ranks also very high in the seventh position. As with the *very important* ranking internal and external Email rank higher. Obviously (comparing the very important and the very important/important

rank) specialised textbooks and German newspapers have an important, though not very important status.

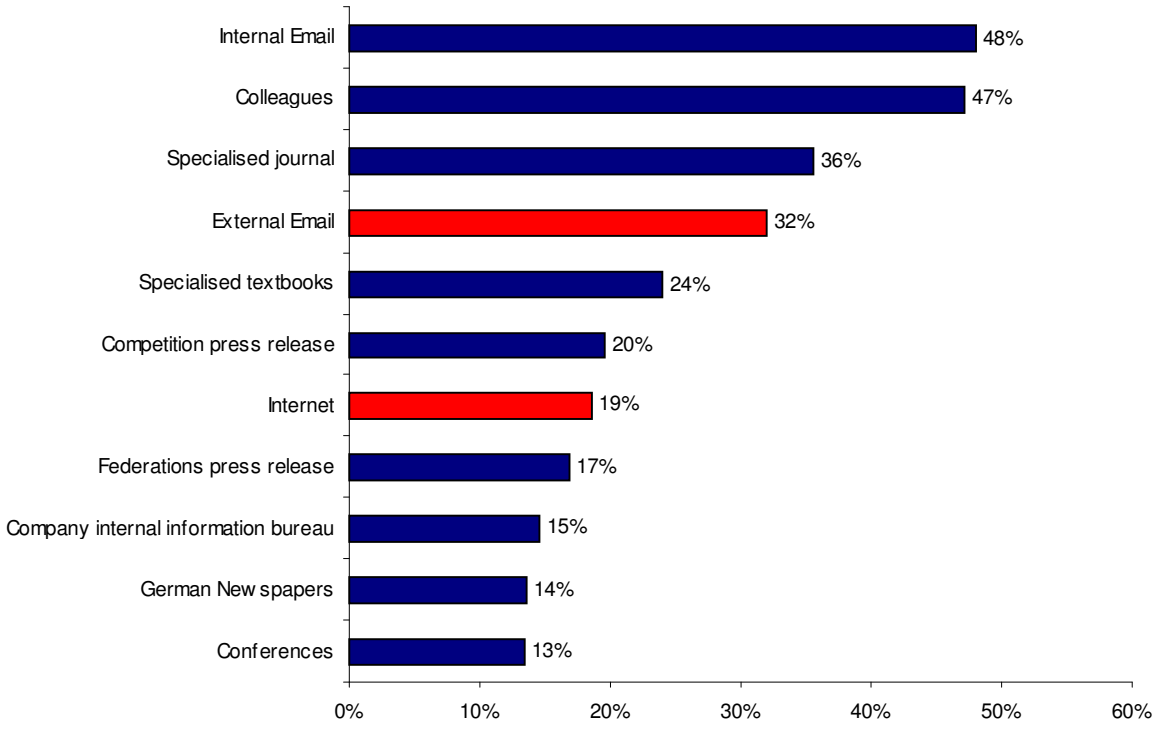


Figure 2-22: Source is considered **very important**, external electronic information sources are highlighted in red.

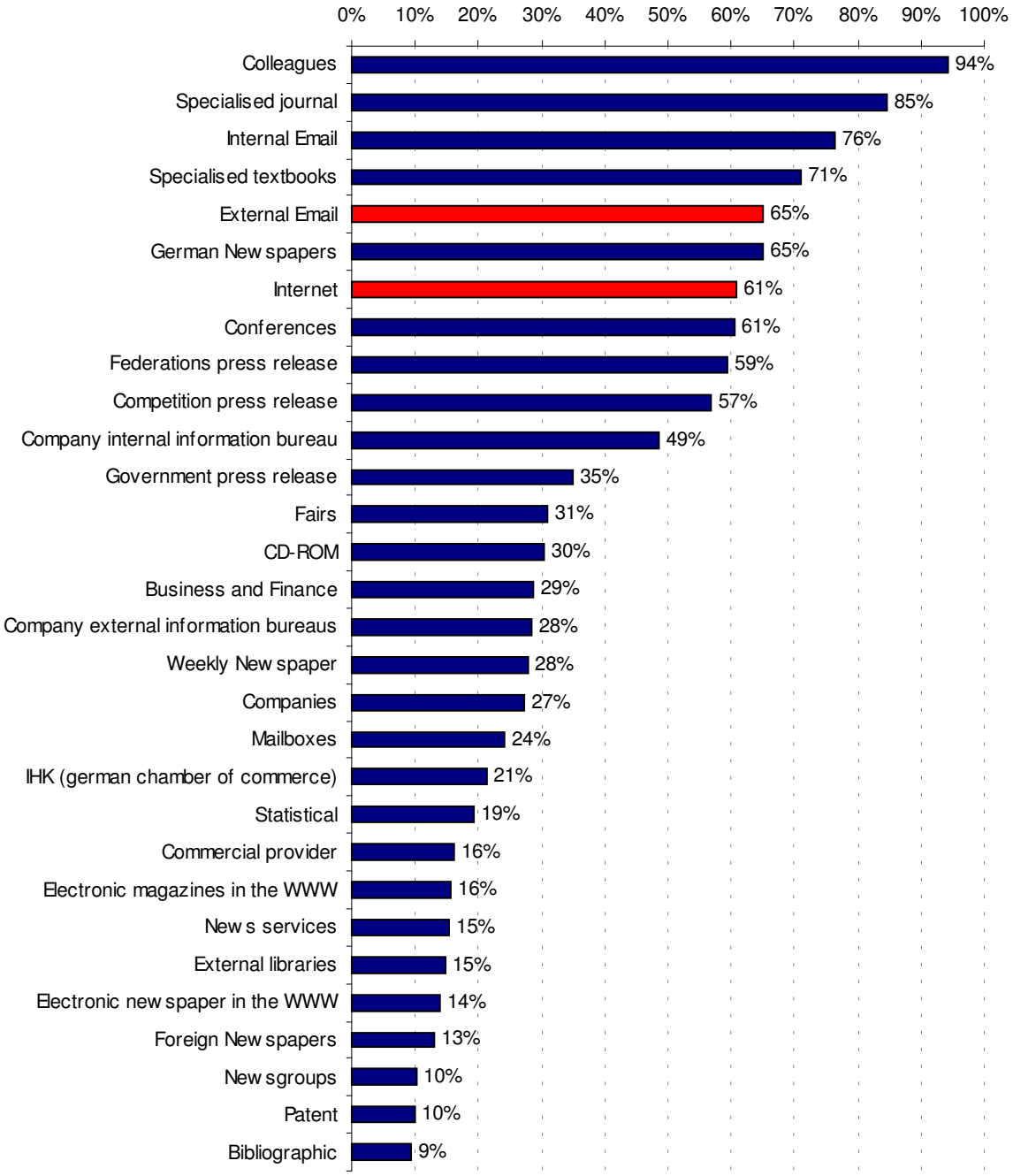


Figure 2-23: Sources are considered **important** or **very important**

In summary it can be said that the results of this section show the same trend as has been found in the European study cited before: the most frequently cited reason for using the Internet has been information gathering (85%)[25] as a base for decision making.

2.2.2.5 Use of external information

The aim of this section was to obtain indicators for the frequency of use of external information.

Nearly half of the subjects (48%) state that they use external information either often or rather often, 39% say that they use it medium often and only 10% rather seldom. None of the subjects use external information seldom or never (see Figure 2-24).

[25] [IBM 1997] http://www.ibm.com/news/wsj/s2.html [2001-05-09]

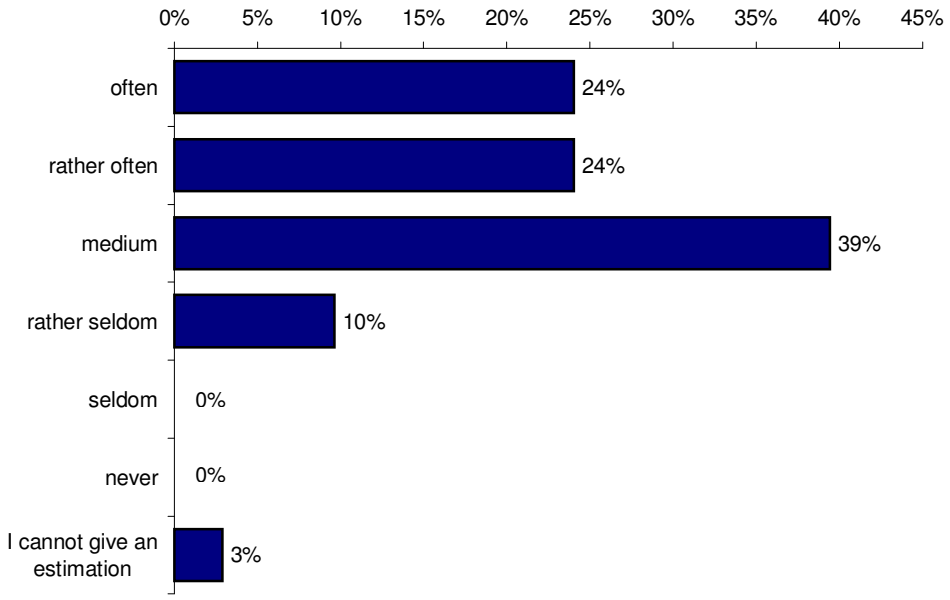


Figure 2-24: Use of external information (n=104)

There seems to be a tendency nowadays that business decision makers use external information much more often than in earlier times. 36% believe that they use external information more often than before, 28% are sure to do so and 27% think that they do not (see Figure 2-25). As there has been no follow-up survey the tendency is only stated from the subjective statements given.

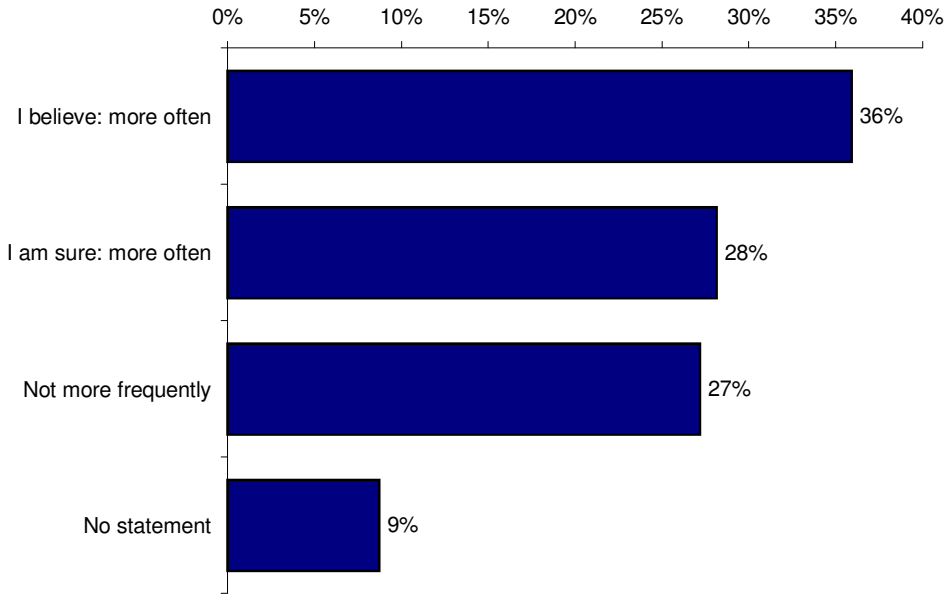


Figure 2-25: Use of external information more often than formerly (n=103)

2.2.2.6 Information overload

An indicator for an information overload could be the number of decision revisions due to incorrect or unavailable information. When asked about this most of the subjects responded that they had to make little or very little decision revisions (67%). Only 1% stated that they have to make very many revisions due to wrong information (see Figure 2-26).

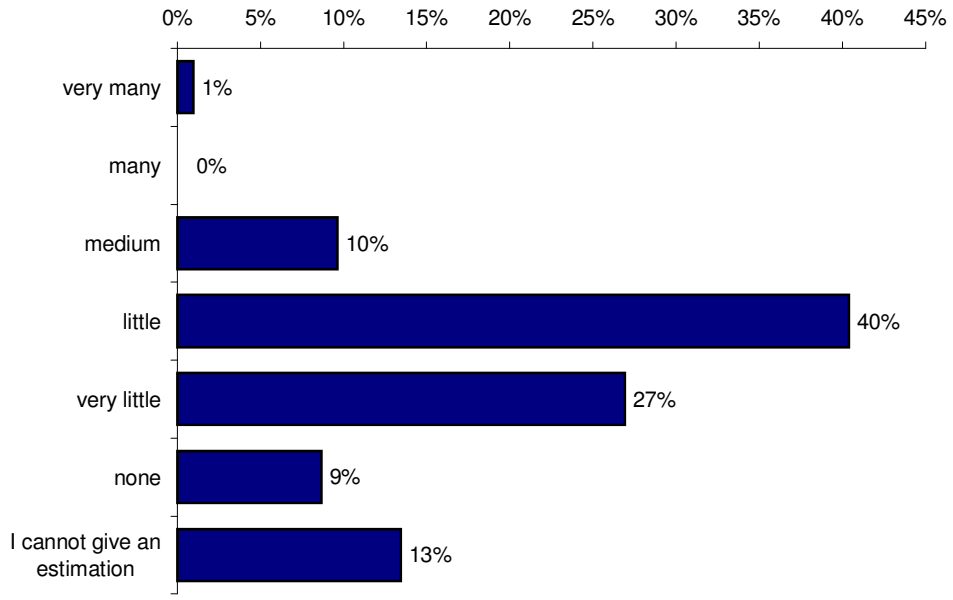


Figure 2-26: Revision of decisions because of wrong information (n=104)

However, when asked about postponing a decision because of missing information the impression given is different. 29,8% of the subjects answered that they had to postpone a decision at least indefinitely often and 9,6% said often. Still the majority stated that they had to do this seldom (45,2%) or never (5,8%). A rather large number of 9,6% could not give an estimation (Figure 2-27).

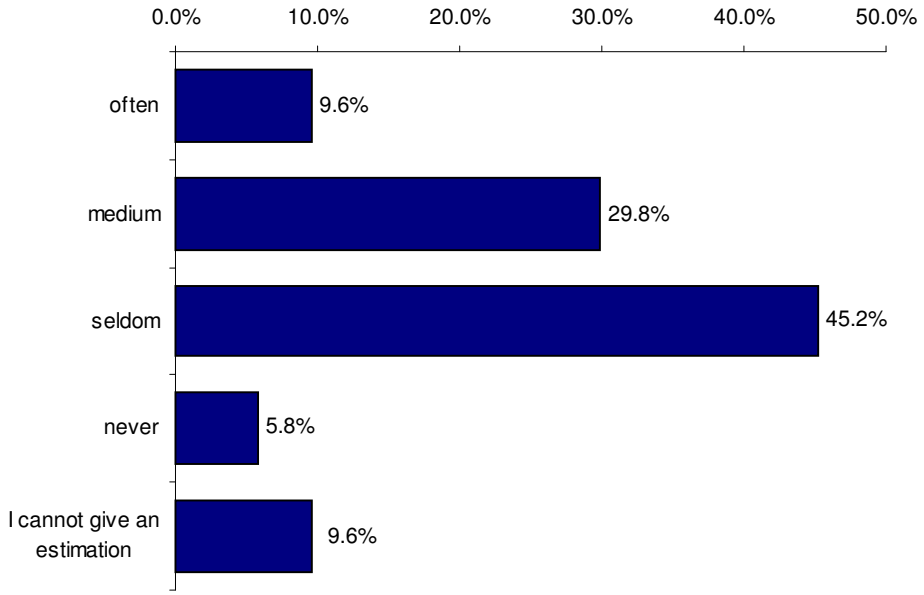


Figure 2-27: How often had the decision makers to postpone a decision because of missing information (n=104)

When talking about information overload the subjective relevance (see section 4.1) comes to mind. Information already known to a user, can be interpreted as being not relevant. Therefore the findings of the next question, where the subjects were asked about the frequency how often they receive information they have already received is worth noting (see Figure 2-28). 36% of the subjects often receive information which

they have already received. Less often and at a medium level are 43%. Only the minority receive such information rather seldom (11%) or seldom (2%).

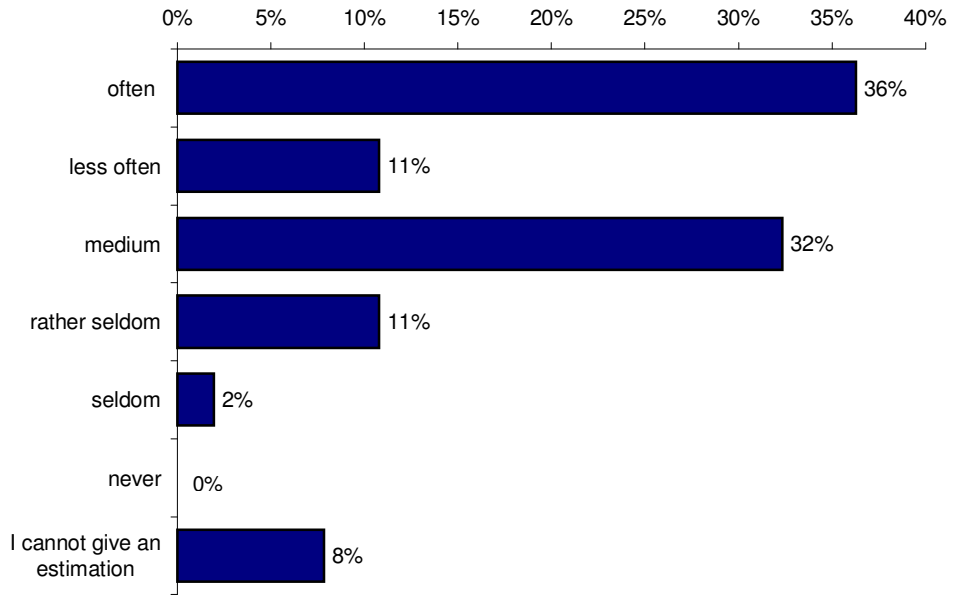


Figure 2-28: How often do the decision makers receive information they already have (n=104)

2.2.2.7 Dealing with information

That this information is not just "there", but that it has to be organised, analysed etc. (see section 2.1) can be derived from the findings in this section.

The subjects were asked how often they verify information they received by using external sources. 39,8% of the responses were either very often (8,2%) or often (31,6%). Still 23,5% do this medium often, 7,1% frequently and 28,6% less frequently (17,4%) or seldom (11,2%). Only 1% of the answers were "never" (see Figure 2-29). The interpretation of information also plays a major part in business decision makers work. 60,8% responded that they receive information which they have to interpret (see Figure 2-30).

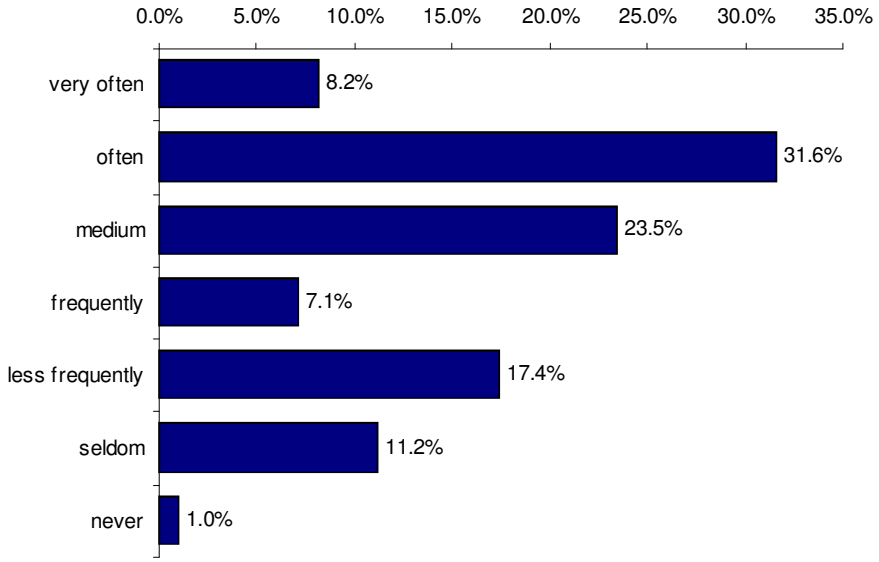


Figure 2-29: Verification of relevant information by using external sources (n=98)

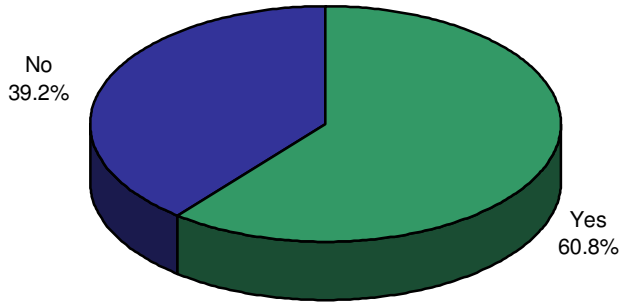


Figure 2-30: Do the business decision makers receive numbers they have to interpret (n=97)

2.2.2.8 Sources of information for special tasks

When asked about the sources used for special tasks the trend is generally that traditional sources like colleagues, printmedia and press are still in the lead (see Figure 2-31). However, as an information source for specific questions the WWW is also high and can generally be compared with the press as an information source. The WWW seems to play an important role especially for information about conferences, state-of-the-art, training and basic knowledge

Figure 2-31: Sources of information for special tasks[26]

Interesting is a comparison of information for specific tasks derived from both non-electronic and electronic sources. This is shown in Figure 2-32. Hereby the information tasks are ordered according to the use of electronic sources. As with the detailed view in Figure 2-31 it can be seen that the non-electronic sources are used in more than 50% of the given tasks. However for the competitor analysis and monitoring, electronic sources are in 43% resp. 40% of the cases used. For information about politics electronic sources are used the least.

[26] Multiple answers were possible, therefore the numbers show the absolute values.

Figure 2-32: Information supply for specified tasks depending on non-electronic sources and electronic sources

2.2.2.9 Statistical Analysis

The previous chapters described the results of the univariate analysis. This section is the second step of the analysis, examining the relationship of variables for the testing of the thesis. For this test the SAS statistic program JMP[27] Version 3.1.6.2 has been used.

As most of the scales used in the questionnaire were nominal scales, the CHI Square Test was used to analyse dependencies between variables. The null hypothesis is that the response probabilities are the same across the subgroups. As a significant level the p-Value is listed. A p-Value (Prob > ChiSq) of less than 0,001 is a hint for a high significance [Sall, Lehmann 1996, p.220]. The smaller p is the more significant given that the variables are not independent. All cases listed here have a significant relationship, still only one of them is highly significant. Also in most cases the table is sparse, meaning that more than 20% of the cells have expected counts less than 5. In these cases the Pearson chi-square tends to behave better than the likelihood ratio chi-square [Sall, Lehmann 1996, p.216].

To gain an insight into different categories of dependencies, several indices were constructed. The construction of the indices has to be carried out very carefully, otherwise the analysis of the statistical data may lead to wrong statements. For instance having school marks for the different categories with 1 being the best and 5 the worst, it is clear that a summarisation of all marks over the different categories will lead into a number that does not say anything.

[27] http://www.jmpdiscovery.com/ [2001-09-26]

PrintIndexImportance: Constructed from question 9, sources are print-media; the index is calculated by

$$\frac{\left(\frac{1}{Newspapers}+\frac{1}{Foreign\ newspapers}+\frac{1}{(weekly\ newspapers/magazines)}+\frac{1}{specialized\ journals}+\frac{1}{specialized\ textbooks}\right)}{5}$$

Using the inverse, the more important the marks, the more weight they have in the index. A mark of six for not using a category (e.g. newspapers) has therefore little influence on the overall index. The higher the index the more important the category. The same pattern for the index construction is used for the personal contacts index (personalIndex), press releases (pressReleaseIndex), information services index (infServiceIndex), e-mail index (emailIndex), electronic information index (e_infoIndex) and the index for the importance of commercial databases (commDBIndex). The normalisation gives the possibility to compare the different indices. The indices are also rounded to the first digit, resulting in personalIndexRounded, used for the analyse of dependencies.

2.2.2.10 Application to thesis

2.2.2.10.1 Thesis 1: External information get more and more important in companies.

As electronic markets grow more and more information is easily available for companies. Speaking of electronic markets, it also becomes clear that acting on electronic markets tends to be acting on global markets [Kuhlen 1995, p.78]. As information can be seen as immaterial, good electronic markets can be compared to punctual markets. Information is more and more easily obtained, therefore companies will be more and more managed and controlled by taking into account market observations. In former times it was more difficult to get this information and most of the times it was both cost and labour intensive. With the development of the WWW as an interface to many information resources this whole process is much less cost intensive. A study conducted in 1993 by [Herget, Hensler 1995] shows that external information becomes more and more important to companies. This survey was only carried out on the basis of the use of commercial databases, but the trend was clearly there.

From the present survey the thesis can not be verified definitely as no comparison study at a later time has been conducted. Still trends can be found from the results of the survey. Using the univariante analysis, it can be seen that 64% of the subjects believe or are sure to use external information more than before (see Figure 2-25). Nearly half of the subjects (48%) use external information either often or rather often (see Figure 2-24). These two statements are certainly not statistical evidence but a trend from this survey. Looking at the variables, frequency of using external information, postponing decisions, revision of decisions, receiving information already received, verification of relevant information by using external sources and their dependencies, one can see that there is a strong relationship of the frequency of using external information and the postponing of decisions, of receiving information already received and of the verification of relevant information by using external sources (see Table 2-1).

X	Y	p Prob>ChiSq <0,001
Frequency of using external information	Postpone a decision	0,0025
Frequency of using external information	Getting information already known	0,0337
Frequency of using external information	Verification of information by external source	0,1187

Table 2-1: Relationship of thesis 1 variables

2.2.2.10.2 Thesis 2: The pre-condition for the use of electronic external information in a company, is the open-mindedness towards the use of the WWW, this implies the installation of the technical infrastructure.

As it has been expressed in Thesis 1, external information is becoming more and more important to companies. It is trivial that the pre-condition for using the external information provided in the WWW is the technical infrastructure, however this does not mean that having the technical infrastructure leads to the use of electronic external information. More important is that it is predicted that the use of electronic external information is depending on the open-mindedness of the users towards the WWW. Hereby the open-mindedness is interpreted as the use of the WWW.

Analysing the relationship of the variables representing the use of digital external information and the open-mindedness towards the WWW and technical infrastructure using the Pearson ChiSquare Test (p) leads to Table 2-2.

X	Y	p Prob>ChiSq <0,001	Type
Network installed	Use external information more frequent, as before.	0,0344	Technical infrastructure
Network installed	Use of Internet/WWW	0,0983	Technical infrastructure
Network installed	Index: pressReleaseRounded	0,1326	Technical infrastructure
Network installed	Importance of electronic magazines in WWW	0,1370	Technical infrastructure
Do you use the WWW?	Importance using Internet	**<0,0001**	Use of WWW
Do you use the WWW?	Importance external Email	0,0002	Use of WWW
Do you use the WWW?	Importance electronic magazines in WWW	0,0004	Use of WWW
Do you use the WWW?	Importance electronic newspapers in WWW	0,0010	Use of WWW
Do you use the WWW?	Use external information more frequent, as before	0,0111	Use of WWW
Do you use the WWW?	Index: e_infoIndexRounded	0,0259	Use of WWW
Frequency of using the WWW?	Index: infServiceIndexRounded	0,0363	Use of WWW
Do you use the WWW?	Importance internal Email	0,0640	Use of WWW
Do you use the WWW?	Importance external Email	0,0640	Use of WWW
Frequency of using the WWW?	Index: personalIndexRounded	0,0669	Use of WWW
Do you use the WWW?	Index: EmailIndexRounded	0,0675	Use of WWW
Do you use the WWW?	Importance of using Newsgroups	0,0955	Use of WWW
Frequency of using the WWW?	Use external information more frequent as before.	0,1133	Use of WWW
Do you use the WWW?	Importance of commercial database providers.	0,1927	Use of WWW

Table 2-2: Relationship by infrastructure and use of the WWW

Classifying the relationships in technical infrastructure and use of the WWW (see the last column of Table 2-2) one can see that it seems that especially the use of the WWW as hypothesised seems to have the greatest influence on the use of electronic external information (see Table 2-2). The p-Values for the type "Technical Infrastructure" are generally much higher than the "Use of the WWW". The relationship of the importance of the Internet use and the use of the WWW is as expected highly significant. Also interesting is the importance of external Email depending on the use of the WWW is with a value of 0,0002 significance. The relationship of the importance of electronic magazines in the WWW depending on the use is also as expected significant (p = 0,0004). The thesis is very much supported in the sense that the open-mindedness towards the WWW is the basis for the use of external electronic information. Especially when thinking of external Email, it is not obvious that it is significantly related to the use of the WWW as many Email clients exist and can be used independently from the WWW.

It becomes clear against the thesis that the use of external information does not depend on the technical infrastructure, but on the use of the WWW (which however presumes that some technical infrastructure is present).

2.2.2.10.3 Thesis 3: The demand for and the kind of external information is different in different branches and different departments. Whereas the difference between the departments is greater than the difference between branch-types.

Classification of branch-type also accumulated as many users did not specify exactly their branch, e.g. Chemistry.

The general statement of this thesis is that the effects of external information influence nearly every branch. The organisation of the companies within these branches are undergoing a general change. This change will affect the departments and their structure, processes and organisation more than the branches themselves. E.g. a retailer of furniture can now easily communicate with manufacturers all over the world. Using telemedia-services like electronic-conference tools, products can easily be viewed before a first real meeting and the final quality control. This way the sales department of the manufacturer has to be prepared to use external information, e.g. what do customers in the destination country prefer, what cultural conditions have to be taken into account etc. However, it would also be the same for a sales department of a finance-service company. They also have to have the external information about their potential customers, e.g. what governmental laws to consider etc. Both the furniture manufacturer and the finance-service provider do not even belong to the same sector, still the need for external information will not differ much.[28][29]

For proving this thesis the dependencies between variables concerning the dealing with information and the branch type resp. department were analysed. To sum up, the types of branches the first two digits of the branch code [NACE-CODE] were taken into account. In the following the results are presented, where a significant relationship has been found (either by branch-type or department). Table 2-3 gives an overview of the significant relationships using the Pearson ChiSquare Test.

Classifying the relationships also according to their type, (see Table 2-3 column four), that is to say if the variable expresses an information need (IN), the dealing with information (DI) or information overload (IO), one gets the impression that the information need is more related to the branch-type than to the departments (see Table 2-3), while the dealing with information tends to be related to the department and not to the branch-type. The information overload is more significantly related to the department than to the branch-type. A conclusion from this is that the demand and the use of information are diverse looking at the branch-type and different departments. From the analysis it can be seen that the diversity of the information demand has to be classified into categories to get a clearer view of the type of demand. Therefore the thesis can not be validated as the resulting demands of information are rather heterogeneous.

---

[28] C.f. also [Kuhlen 1995, p.78] for changes to result from electronic marketplaces, see the criterions defined

[29] [Mintzberg 1975, p.59] states that depending on the kind of work (sales, production, staff), the managers tend to spend relatively more time on various roles (e.g. interpersonal, decision making, informational).

		By Branch-Type Prob>ChiSq <0,001	By Department Prob>ChiSq <0,001	Type
1	**Use external information more frequent as before**	0,5271	0,0933	DI
2	**Postpone a decision**	0,1483	0,0678	DI
3	**Numbers that need interpretation**	0,8779	0,0428	DI
4	**Use of WWW sites with costs**	0,0051	0,8703	IN
5	**Other commercial information**	0,0350	0,7666	IN
7	**Index: Importance of Printmedia (PrintIndexWichtigRounded)**	0,8136	0,0085	IN
8	**Index: Importance of Email (external and internal) (EmailIndexRounded)**	0,9124	0,1528	IN
9	**Index: Importance of commercial databases (CommDBIndexRounded)**	0,1564	0,9882	IN
10	**Revision of decisions due to lack of information**	0,0005	0,8369	IN
11	**Receiving information already received**	0,9087	0,0596	IO

Table 2-3: Summary of significant relationships (IN = information need, ID = dealing with information, IO = information overload)

2.2.2.11 Summary of the results

The present survey has shown that external information is becoming more and more important within companies. Still personal contacts play the major role as an external information source, but the use of Email and the Internet are also considered to be very important, and overall there is a trend that electronic external information is used more and more, e.g. 40% use it often or very often to verify relevant information. Receiving information already received seems to be a general problem (47% get such often or less often). However it also became clear that a lot of time the business decision makers are confronted with soft (qualitative and quantitative) information, which needs further interpretation.

The discussion and analysis of the variables for the proposed thesis show that the use of the WWW is an important factor when talking about the use of external information. For the subjects in this study, business decision makers, internal and external Email and the Internet are the most important electronic sources. The discussion of thesis 3, whether information demand and kind differs rather in branch-types or departments shows that the behaviour is more related to the departments than branches.

The discussed results lead to the following propositions which should be taken into account when designing and implementing a system for the retrieval of business information:

Technology:

- It can be assumed that the technical infrastructure and personal readiness for using the WWW is given. 99% of the business decision makers of the study have at least an intermediate IT experience level. Even at the time of the study in 1998 nearly 70% have already used the WWW, of whom nearly three quarters daily or more than once a week.
- Design the system browser independently, however the browser strategy of the companies should be taken into account.

Soft information:

- Integrate commentary for the interpretation of soft information
- Integrate Email functionality to share information

Proposed sources:

- Proposed electronic sources for business information are: internal and external Email and the WWW.
- More and more federations, newspapers, conferences etc. also present their information electronically. Integrate these as business information sources.
- Frequently the same sources of external information (e.g. newspaper, magazines etc.) are used regularly. Therefore it is not necessarily important to search the whole of the WWW, but only pre-defined sources.
- The proposed sources have to be adapted to the branch type as it has been found that the information need is related rather to the branch-type than to the department.

General:

- General system design can be independent from branch-type but must be adaptable to department, e.g. by a given post inside the companies, like the internal information services.
- The study shows that internal information services are not so important for the business decision makers as other external sources. Maybe this is because some of the services of this department are not obvious to the user or that the internal information departments need to redesign their services. A supporting function for a business information system could be one step in that direction.
- Design the system open-minded to be able to include other value-added services

Within the current work a solution for such a system is presented. Rather than focusing on all the aspects listed, a system for the retrieval of business information from the WWW is proposed. After having done this first design decision, the next step is to find a way to obtain the information available in a way that e.g. allows further processing, integration into the organisations' information systems, supporting the information role of the business decision maker [Mintzberg 1975]. On a broader view this was the objective of the research project INSYDER, that the working group Information System of the University of Konstanz have been involved in.

Knowing that there is a demand, the next section will present selected Business Information sources of the WWW and show what kind of information is available, either with or without cost.

2.3 Sources for external Business information

It is undisputed that business information can be retrieved from the WWW [FAZ 1997], [Behme, Mucksch 1999] and in this section selected examples will be given. Still with the WWW being an open medium to everybody in the sense of everybody can easily publish, using it one has to keep in mind that the quality of the information obtained might be doubtful. *"The Web is a global bulletin board where the wise and the foolish have equal space".*[30] When thinking of quality at least the terms reliability, actuality and truthfulness come up, and are an important factor when judging that this is the kind of source the information is from. As people are also (or should be) careful in daily life

[30] http://webfarming.com/intro/intro05.html [2001-05-22]

with information they receive (e.g. reading an article about a competitor in the yellow press and in serious magazines), the same is demanded for dealing with business information from the WWW. However information from the WWW is volatile and under constant change. The WWW has a great variety of obtainable information. Figure 2-33 shows a taxonomy by Hackathorn of information under the dimensions quality versus coverage. Hereby commercial data is seen as information provided by the classical online database providers, like DIALOG or STN. Governmental data is information released by the governments, while corporate data is the information companies or federations provide.

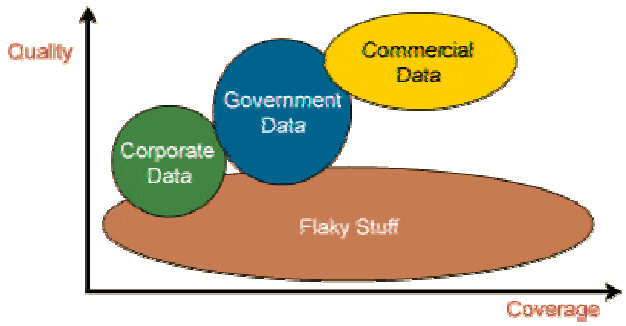


Figure 2-33: Quality versus Coverage [Hackathorn 1999, p.16]

An example for information from ministries and federations is the German Chamber of Commerce (DIHT)[31], the Department of Trade and Industry of the United Kingdom (DTI)[32] or the Eurochamber[33]. Typical obtainable information from these institutions are for example surveys, political (e.g. tax laws) or business news, contact information or special services for start-ups. Depending on the offered information it is free of charge or with a charge. The European Union also offers a wide range of business information.[34] Other business information is often found at the company online presentation themselves. For example looking at the British online presentation of Bosch the visitor of this site gets information on facts, e.g. sales volume, number of employees and so forth (see Figure 2-34).

[31] http://www.diht.de/ [2001-04-03]
[32] http://www.dti.gov.uk/ [2001-04-03]
[33] http://www.eurochambres.be/ [2001-07-20]
[34] http://europa.eu.int/ [2001-07-20]

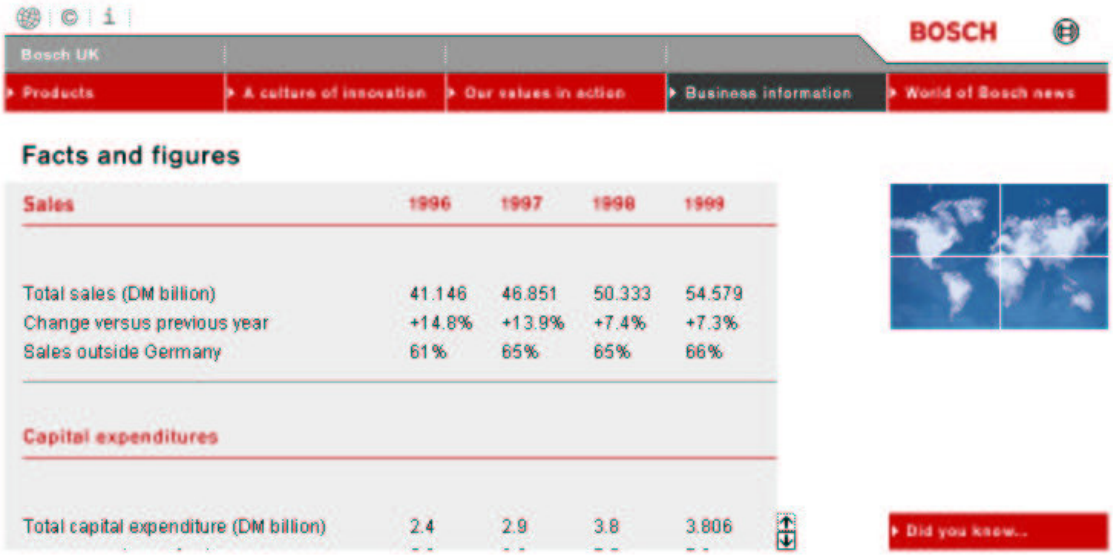


Figure 2-34: Business information from Bosch UK detail view[35]

Another kind of resource for business information are the traditional broadcasters of information, like newspapers, magazines or TV companies, radio stations etc. Information available, there are all kinds of news and stock charts. Most of the time they offer added-values, e.g. a user adaptation of the site, own portfolios or newsletters.[36] A typical online representation, the news channel CNN is shown in Figure 2-36. The obtainable information is mostly free of charge. As with other news services for market prices are not real-time values (common is a 15 Minute delay), if they are free of charge. Especially for stock information one lacks on overview of providers. Most of the banks offer such information as a service and also brokers[37], but also "exotics" in that field, like catalogues[38] or browsers.[39] Yahoo! has announced to launch a real-time stock information server for customers, who subscribe to the *"Yahoo Finance MarketTracker"* for a special fee of $9.95 per month.[40] A common service with the stock information is to also receive news about the company and its environment.

An example for the use of the *Webcast* technology for presenting business information is Reuters news service called *Reuters Television*[41] offering market reports and analysis. News reports are streamed live throughout the business day from 07:00 to 16:00 (London time) with the aim to support their customers in making their investment decisions (see Figure 2-35). In-between the live broadcasts (Figure 2-35 left), an update on stock exchange rates is given (Figure 2-35 right).

[35] http://www.bosch.co.uk/0530.html [2001-04-03]
[36] E.g. New York Times http://www.nytimes.com [2001-04-04]
[37] http:// www.consors.de [2001-04-04]
[38] http://www.yahoo.com [2001-04-04]
[39] http://my.netscape.com [2001-07-20]
[40] http://billing.finance.yahoo.com/ym/f/FinanceReal1 [2001-07-20]
[41] For the use of this service either the Microsoft's Media Player or Real Player have to be installed.

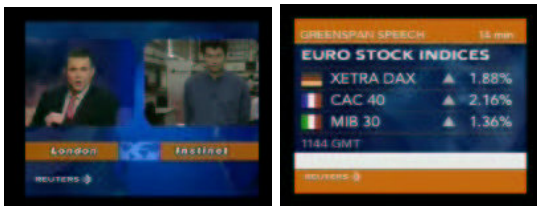


Figure 2-35: Reuters Television[42]

With relation to Information Visualization (see section 4.5.2) an example for a visualisation of stock information by Smartmoney is shown in Figure 4-19.

Figure 2-36: Screen-shot of business news presented by CNN[43]

There are a whole range of information providers who present information free of charge as a first step, while for detailed information costs are incurred. [Georgy 2001] presents an (rather abstract) overview how competitor analysis can be made using free of charge information from the WWW. *Kompass*[44] for example offers a *"business-to-business search engine"*. The information provided here is information about companies, such as address, number of employees, products, names of the board of directors, financial information etc. The data for some companies is free of charge, while others have a limited access and the rest is with a charge.

Business service providers (the traditional information brokers) present the business information mostly on a fee basis. Depending on the type of service, customers

[42] http://www.reuters.co.uk [2001-10-23]

[43] http://cnnfn.cnn.com/?s=6 [2001-04-03]

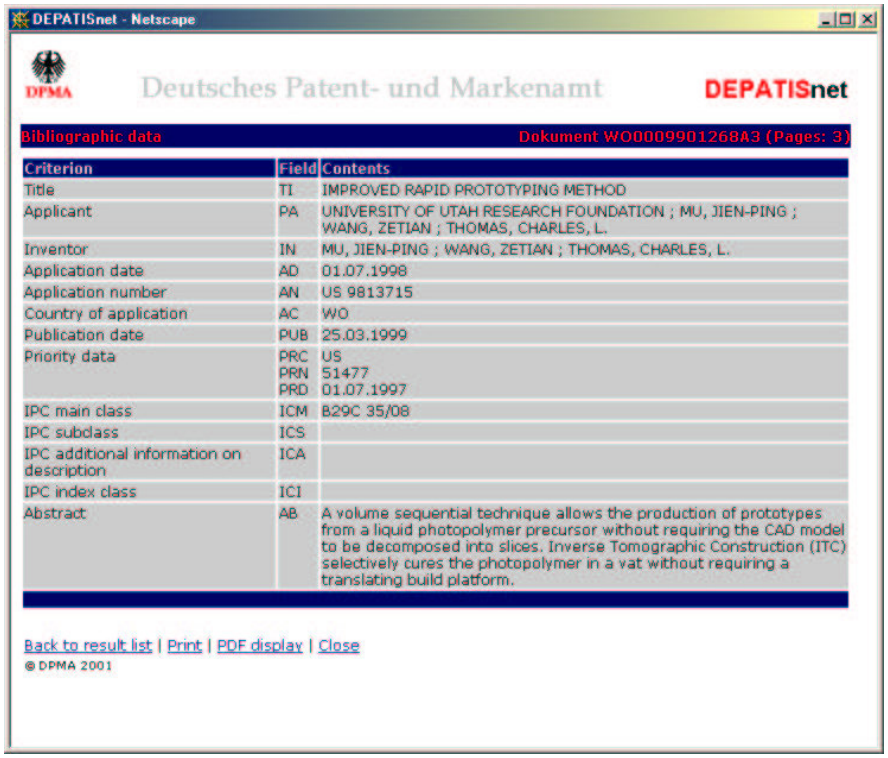
[44] http://www.kompass.com [2001-04-03]

subscribe to services or pay-by-demand. An important differential of these business information providers are the sources they use (e.g. national/international databases, print materials from newspapers and magazines, company reports etc.). According to the information need (e.g. extensive report on a competitor), the sources included play a major role when deciding upon which provider to favour. Also the way the information is presented differs greatly (push and pull service are both common, e.g. newsletter in html, text or pdf format by email, desktop integrating tools, service provided in the WWW). Selected examples are

- Factiva[45]
- BirdOnline (focus on business information for the UK)[46]
- Dun & Bradstreet[47]
- Global Access[48]
- Fuchsbriefe[49]
- ECOFIS (part of Creditreform)[50]
- Traditional database hosts and providers and business service providers, like Creditreform, DataStar, Lexis-Nexis and so forth.
- Portal B (Data Downlink Corporation)[51]
- Cutter Information[52]

Also a valuable source for business information is the source of patent information. Some of the patent databases are now searchable free of charge, e.g. Depatisnet[53] or IBMs Patent Server[54]. Figure 2-37 shows the abstract of an example patent about rapid prototyping. The full version of this document is available by using the *PDF display* option. Typical uses for patent information is to gain an insight in technical developments. However patents are also an interesting source of information about competitors (what do they research) and potential employees (who are the most listed inventors in a field, who did something extraordinary?).

[45] http://www.factiva.com/ [2001-04-03]
[46] http://www.bird-online.co.uk [2001-04-03]
[47] http://www.dnb.com/ [2001-04-03]
[48] http://www.primark.com/ga/ [2001-04-03]
[49] http://www.fuchsbriefe.de/ [2001-04-03]
[50] http://www.alleco.de [2001-04-03]
[51] http://www.portalb.com [2001-04-03]
[52] http://www.cutter.com/ [2001-07-20]
[53] http://www.depatisnet.de [2001-07-20]
[54] http://www.englib.cornell.edu/instruction/chemengr462/ibm1.html [2001-07-20], focus on U.S. patents and trademarks.



DEPATISnet - Netscape

DPMA Deutsches Patent- und Markenamt DEPATISnet

Bibliographic data Dokument WO0009901268A3 (Pages: 3)

Criterion	Field	Contents
Title	TI	IMPROVED RAPID PROTOTYPING METHOD
Applicant	PA	UNIVERSITY OF UTAH RESEARCH FOUNDATION ; MU, JIEN-PING ; WANG, ZETIAN ; THOMAS, CHARLES, L.
Inventor	IN	MU, JIEN-PING ; WANG, ZETIAN ; THOMAS, CHARLES, L.
Application date	AD	01.07.1998
Application number	AN	US 9813715
Country of application	AC	WO
Publication date	PUB	25.03.1999
Priority data	PRC PRN PRD	US 51477 01.07.1997
IPC main class	ICM	B29C 35/08
IPC subclass	ICS	
IPC additional information on description	ICA	
IPC index class	ICI	
Abstract	AB	A volume sequential technique allows the production of prototypes from a liquid photopolymer precursor without requiring the CAD model to be decomposed into slices. Inverse Tomographic Construction (ITC) selectively cures the photopolymer in a vat without requiring a translating build platform.

Back to result list | Print | PDF display | Close

© DPMA 2001

Figure 2-37: Example abstract of a patent retrieved using the Depatisnet

Besides these offers, a great number of specialised information providers exist, presenting their information on electronic market places or portals. Portals can be assumed as a sub-group of electronic market places, as they provide the basis for the information phase on electronic market places. The information phase is the first phase in a transaction on an electronic market place. The objective is that both the supply and demand side are provided with an overview of the market and the object dealt with [Langenohl 1994]. The information phase is followed by the agreement and transaction and after-sales phase. Characteristics as e.g. described by [Rösch 2000] can not be approved as they present nothing other than a subset of the characteristics of electronic markets presented before (e.g. see the definition by [Kuhlen 1996, p.6], and general discussion by [Kuhlen 1995], [Picot, Reichwald, Wigand 1996]). Examples of such information market places are often the sites provided by associations, e.g. VDA[55].

[Bates 1999] claims that value-added services (in her notion services which are not free of charge provided by information services, i.e. Dow Jones) are superior to simple Web searches for the retrieval of business information. Yet as her evaluation is sponsored by Factiva (which is a Dow Jones and Reuters company) and little is said about the experiment design (e.g. who performed the simple WWW searches), the findings (the Dow Jones Interactive performs best) can not be generalised. Still the general remark that most commonly Web-based sources are not well-suited for serious business research can be agreed to this extent and most of the time the appropriate tools to retrieve their (valuable) content are missing.

A complete survey of sources of business information available on the WWW is out of the focus of this work.[56] An extensive overview of business information available on the

[55] Association of the German automotive industry. http://www.vda.de

[56] In the WWW many starting points can be found for acquiring business, e.g. http://www.san-marino.k12.ca.us/~smpl/buslink.htm, http://www.europages.com/business/business-info-en.html, http://www.brint.com, http://personal.dis.strath.ac.uk/business/market.html (all [2001-07-20]) for a listing of providers of statistical, economic and market information.

WWW dating back to 1999 with a focus on German sites is given by [Lässing 1999]. [Meier 2000] and presents a taxonomy and examples of external information for business information. [Hackathorn 1999, p.235-278] also presents a comprehensive overview on resources for business information.[57]

2.4 Summary of this chapter

This chapter deals with the information need of business decision makers. Hereby the focus has been on their need for electronic external information. The first part of this chapter gives a general introduction to information as a good and distinctive attribute. A definition of external information as used in this work is given. Moreover the impacts and possible sources of electronic external information on the organisations are presented.

The second part of this chapter presents the findings of a study conducted in 1999 by the authoress in co-operation with a Konstanz/Germany based company. The results show that electronic external information is becoming getting more and more important within organisations and the use of Email and the Internet are considered to be very important. Overall there is a trend that electronic external information is used more and more. To summarise this, it becomes clear that organisations have a strong demand for information from WWW. Three proposed theses have been discussed and could partly be verified. Interesting is the fact that the results of the current survey propose that dealing with information differs from department to department more than from branch to branch. Altogether this gave the direction for the design and implementation of the current work. Within this chapter some of the design guidelines as derived by the survey results have been presented.

The third part of this chapter gives examples of and entry-points for business information on the WWW and shows what information can be expected.

After determining the information need of the business decision maker, focusing and investigating possible sources of the WWW for business information, the next step is to find a way to retrieve this information in a way that e.g. allows further processing, integration into the organisations information systems, supporting the information role of the business decision maker. With this objective in mind INSYDER has been developed within a research project as part of an overall approach to integrate external electronic information into Business Intelligence Systems. In the next chapter an introduction to Business Intelligence Systems is given, explaining components and how INSYDER fits into. There will be also a focus on the process of *Web Farming* [Hackathorn 1999], an organisational framework and method description of how information from the WWW can be integrated into Business Intelligence Systems.

[57] See also http://www.webfarming.com/service/resources.html [2001-05-23]

3 Business Intelligence Systems

This chapter describes Business Intelligence Systems (BIS) from a global view but also for this thesis. Thereby components, use, relations to other management information systems and examples will be described. A focus within this chapter is the integration and presentation of electronic external information in such systems and related approaches.

3.1 Definition and Overview

In the literature a variety of definitions of BIS can be found. All have in common the understanding that information has become a resource. While by business intelligence (BI) mostly the process of activity is meant,[58] Business Intelligence Systems (BIS) means the actual systems used to perform business intelligence tasks. Still, both terms have a broad scope of definitions. Gilad [Gilad 1998] focuses in his definition on the external aspect of business intelligence as a process:

> *"Business intelligence is the activity of monitoring the environment external to the firm for information that is relevant for decision-making in the company."*

Others still see it as a process but focus on the data management aspect, gathering management analysis and distribution of data [Taylor 1998].

BIS itself was created in the late 1980s, early 1990s,[59] trying to find

> *"a catch-all term to describe concepts, methods and processes to improve decision making in business through the use of facts and fact-based systems"*[Grise 1997].

A definition by [IBM 1998] sets the user and the variety of technologies and products in focus:

> *"A business intelligence system ... provides a set of technologies and products for supplying users with the information they need to answer business questions, and make tactical and strategic business decisions."*

[Grise 1997] points out the integration effect of BIS:

> *"... once the information is sourced, it is integrated with other relevant information before applying any further analysis.' It is the integration of this core source information with the relevant, or contextual, information that its at the heart of the matter."*

However the aspect of finding a good term to cover many concepts plays an important role. BIS is referred to many times, when MIS (see section 3.2.5) and OLAP (see section 3.2.4) technologies are joined [Chamoni, Gluchowski 1998].[60] Having the process of BI in mind, tools and technologies like data warehousing, executive information systems (EIS), decision-support systems (DSS), data mining, reporting and so forth are joined as the BIS.

> *"Many of the concepts of business intelligence are not new, but have evolved and been refined based on experience gained from early host-*

---

[58] Related topics are 'Competitive intelligence', 'Environmental scanning' [Correia, Wilson 1997], [Choo 1998], 'Economic intelligence' or 'Social intelligence' [Choo 1999].

[59] "*The term 'business intelligence' appears to have arrived on the scene circa 1989 and was described by Howard Dresner of the Gartner Group.*" [Grise 1997] see also [Taylor 1998]

[60] See also IMIS http://www.imis.de

based corporate information systems, and more recently, from data warehousing applications."[61]

Nevertheless IBM sees Business *Intelligence* Systems as a further step in the development of Business *Information* Systems. After the first generation, which were host-based query and reporting systems, the second generation of data warehousing focusing on the technology and its building, the third generation is formed by BIS, providing pre-packaged application solutions, focusing on the access and delivery of its contents, using a variety of tools from multiple vendors (see Figure 3-1). [Hall 2001] presents a state of the BIS market and presumes its further development, setting it in context to developments like XML, wireless techniques or data visualisation.

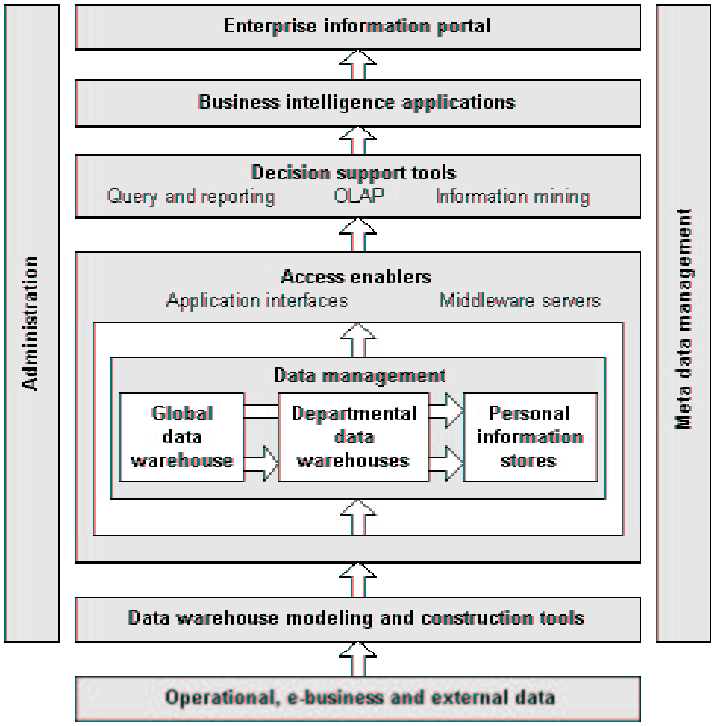


Figure 3-1: BIS structure[62]

A study among business decision makers (BDM) and technical decision makers (TDM) in October 1999 by [McDonald 2000] tries to clarify the state of BIS in the companies.[63] The main finding of the study is that BIS are mostly used for financial reporting and budgeting, while its use for financial consolidation and modelling was under 50%. OLAP (see section 3.2.4) is according to the author *"just getting a foothold"*, however 44% (BDM) and 25% (TDM) have no OLAP plans. XML (see section 4.4.5.1) is seen very differently by BDMs and TDMs, while 66% of the BDMs have no plans to use it, only 25% of the TDM have no plans, the result is similar (52% resp. 36%) for the plans and use of Balanced Scorecards (see footnote 82). Paper versions of financial reports are still in the majority (BDM 84%) but only 18% of the subjects asked, want these on paper, 56% want them as an Email and 26% available on the company intranet. This shows two things, first it seems that the new technologies are often just used as the

[61] http://www.software.ibm.com/data/pubs/papers/bisolution/index.html [2001-09-03]

[62] http://www-4.ibm.com/software/data/pubs/papers/bisolution/ [2001-05-15]

[63] Participants were readers of Business Finance, whose companies report annual revenues of more than $10 million taking part with a paper version (n=183 respondents) and visitors of the Business Finance Web site, to whom an online version was posted (n=1684 respondents).

familiar old ones and secondly that the possibilities of BIS are largely untapped, e.g. relating information etc. Reasons for this could be the lack of time among the business decision makers.

For an overview and comparison of vendors of BIS systems, the WWW pages of the "Business Application Research Center" (BARC)[64] are helpful.

3.2 Technologies and tools related to BIS

This section gives an overview on IT based management support systems as they are related to BIS (see section 3.1). Parallel to the invention of PCs in the 80s, spreadsheet applications have been established as an important tool for decision support systems. Still these had their limits e.g. thinking of multi-user applications or efficient data management. This led to the development of new technologies and tools to overcome these deficiencies. This section does not have the intention to provide a comprehensive overview on these information systems, but to present an insight and overview in their use and differences. For a detailed discussion, the reader is referred to the appropriate literature cited in each section.

3.2.1 Exemplary Architecture of a BIS

Figure 3-2 shows an architecture for a specific BIS. The idea behind this is to show how the information retrieval system, described in the current work, is technologically embedded in the overall BIS architecture. According to the definitions from the previous section a set of technologies is provided to assist users in satisfying their (business) information need. As can be seen in Figure 3-2 a variety of tools and technologies are combined. The internal data handling is shown on the right side. The techniques and tools will be explained later in this section. The left side of the architecture shows the complementary part for external information, like business information from the WWW. Here the INSYDER system could be used as an integral part within such a BIS. INSYDER offers value-added retrieval, categorisation and reviewing functionality, which is needed for the presentation of external information within a BIS. The architecture shows also how other sophisticated information systems like Knowledge Management and Management Support Systems (MSS) fit into the framework. Hereby both the Knowledge Management and the MSS component build the integral part of external and internal information.

The following sections present an overview on the basic information systems. This overview is then followed by a description of INSYDER, outlining how a BIS can benefit from its integration.

[64] http://www.barc.de [2001-07-23] fee-based.

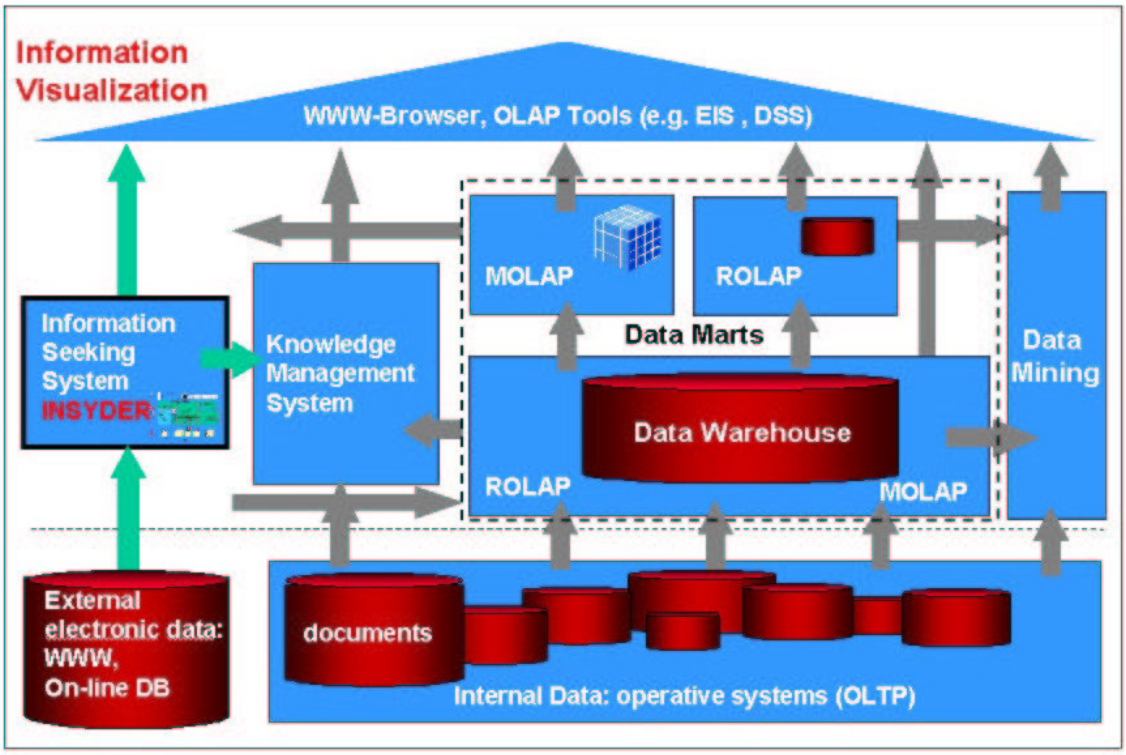


Figure 3-2: Architecture of a BIS adapted from [Gluchowski, Gabriel, Chamoni 1997]

3.2.2 OLTP: On-line Transaction Processing

OLTP is the automation of high-volume business processes focusing on transactions. These include short records of a few fields, with data being relatively simply coded (dates, numbers, short description). OLTP applications are typically characterised by online interaction of one to many users, manipulating shared data. Examples for the use of OLTP system are in the banking sector, financial services, and reservation systems [Tyler, Fisher 1995], [Black 1996].

3.2.3 Data Warehousing

> *"A data warehouse is the granular, corporate, integrated historical collection of data that forms the foundation for all sorts of DSS processing, such as data marts, exploration processing, data mining, and the like."* [Inmon 1999, p. 8]

Generally speaking a Data Warehouse builds the (data) base for information systems in an organisation. The purpose of a Data Warehouse is the establishment of a data repository to make operational data accessible, readily acceptable for further processing by for example decision support or executive information systems [Turban, Aronson 1998]. Hereby operational data comes normally from OLTP systems (see section 3.2.2) A term that occurs in context with Data Warehouse is 'Data Mart': *"A data mart is a subset of the enterprise-wide data warehouse"* [Turban, Aronson 1998, p.125].

The major characteristics of a Data Warehouse are [Inmon 1995], [Kimball 1996], [Turban, Aronson 1998]:

- subject orientated (representation of major subjects instead of processes, e.g. customers, vendors, products)
- integrated (consistency of representation, e.g. consistent naming conventions, consistent measurement of variables, consistent encoding structures)

- time-variant (data warehouse data represents data over a long time horizon, each key structure contains somehow a time element)
- non-volatile (the data once entered are not changed or updated)

Setting up a Data Warehouse is not easy. One has to deal with data quality problems, transformation problems, metadata definition problems and more [Martin 1997]. For a further discussion of the subject Data Warehouses the reader is referred to William H. Ilmon's comprehensive web page on Data Warehousing and related subjects.[65]

3.2.4 OLAP: On-line Analytical Processing

In delimitation to the term OLTP Codd created in 1993 the term OLAP (On-line Analytical Processing). [Codd 1993, p.7] defines

> *"OLAP is the name given to the dynamic enterprise analysis required to create, manipulate, animate, and synthesize information from exegetical, contemplative, and formulaic data analysis models […].This includes the ability to discern new or unanticipated relationships between variables, the ability to identify the parameters necessary to handle large amounts of data, to create an unlimited number of dimensions (consolidation paths), and to specify cross-dimensional conditions and expressions."*

This definition describes the demands for information system tools, which want to be referred as OLAP tools. In the same paper Codd defines 12 rules to evaluate OLAP systems.[66] Pendse from the OLAP report project[67] reduced and summarised the OLAP definition in 1995 to the acronym FASMI (Fast Analysis of Shared Multidimensional Information) with the intention of gaining a better basis in terms of a more applicable evaluation base for distinguishing OLAP and Non-OLAP tools.[68] In contrast to the OLTP systems, where the transaction is in focus, OLAP systems hold the data as data-cubes[69] with - according to the number of objects - n dimensions (data categories). The analysis of the data is possible by all dimensions. OLAP offers the possibility of browsing the cube by using slice and dice operations.

An example of an OLAP cube is to analyse the sales of a company. In this context the following dimensions are identified:

1. Periods
2. Products
3. Regions

---

[65] http://www.billinmon.com/index.html [2001-05-17]

[66] Extending them with another six rules in 1995. http://www.olapreport.com/fasmi.htm

[67] http://www.olapreport.com

[68] http://www.olapreport.com/fasmi.htm [2001-05-16]

[69] Also referred to as hypercubes.

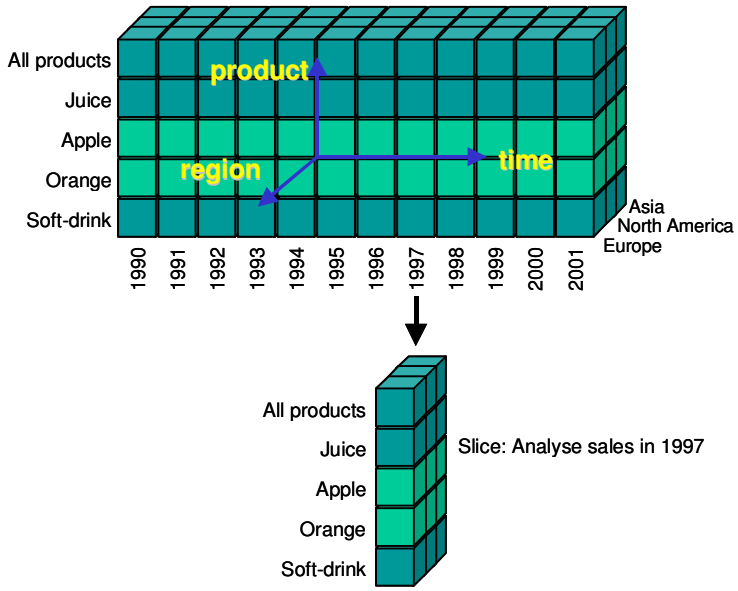


Figure 3-3: Example hypercube with three dimensions

This leads to a cube as shown in Figure 3-3. In the figure the product dimension is divided into further subcategories to show that dimensions can be hierarchical. For overview reasons the subcategories are omitted for the other two dimensions. Using the functions drill-down[70] and roll-up[71] enables navigation within the cube.

For the realisation of the OLAP technology two basic approaches can be identified [Koutsoukis, Mitra, Lucas 1999]:

- ROLAP: based on a Relational database with an OLAP server (low performance, high data volume)
- MOLAP: dedicated Multidimensional database (high performance, low data volume)

Besides the papers cited in this section several websites like the *OLAP Report project*[72], the *OLAP council*[73] or the *OLAPInformation*[74] offer information for further discussions of the subject.

3.2.5 MSS: Management Support Systems

Management information systems (MIS), decision support systems (DSS) and executive, enterprise or everybody's information system (EIS) are all types of an enhanced type of information system, having in common to rely on a corporate data. MIS was a first attempt to provide managers and other organisational staff with business information to support them performing their jobs more effectively and efficiently [Houdeshel, Watson 1992, p.15]. MIS are characterised by providing pre-configured queries to satisfy the information need of managers from different hierarchy levels. A typical objective of an MIS supporting a manager is the supply of reporting and query functionality. DSS are designed to help managers in decision making processes, where

[70] Going deeper in the hierarchy.

[71] Going up in the hierachy.

[72] http://www.olapreport.com [2001-05-17] focusing on case studies, product reviews.

[73] http://www.olapcouncil.org/ [2001-05-17] focusing on definition and benchmark studies. Website seems not to be well looked after, as the last news item is from January 1999.

[74] http://www.olapinfo.de [2001-05-18] is link collection with different categories (white papers, tools, newsgroups etc.).

data is badly-structured, solving part-tasks by providing models, methods and problem focused data [Gluchowski, Gabriel, Chamoni 1997]. In contrast to EIS, these two types of information systems for business decision makers have rarely been used by top executives [Houdeshel, Watson 1992, p.15]. [Wilson 1995] reports of two user groups regarding decision makers, perceptive and receptive. While the former wants to have data categorisation and exception reporting, the receptive type needs to access all historical data. As a conclusion of a discussion about information use and access in organisations [Wilson 1995, p.5] states that *"it is clear that cognitive […] styles do vary among individuals and that it is highly unlikely that a system designed according to one cognitive model will fit the behaviour and expectations of a user whose model is different."* The author refers back to Leavitt's diamond representing the organisational life, showing that all aspects affect each other. Wilson is also predicting that information systems for managers have to allow oral communication to achieve great dissemination.[75]

In the literature various functional types of EIS are described. As a result of a study among companies with EIS[76] [Iyer, Aronson 1995] developed five categories for EIS benefits:

1. Information: e.g. more timely information, faster access
2. Environmental scanning: e.g. better access to soft information
3. Improving executives' effectiveness: e.g. improved presentation of data
4. Meeting strategic objectives: e.g. increased span of control, improved decision making
5. Economy: e.g. cost savings, support downsizing of organization, less paper.

[Hoven van den 1995] strengthens in particular the environmental scanning requirement of an EIS as a benefit, concerning both hard and soft information, getting the external information either from professional services (information broker) or using the Internet as a source. The benefit criteria Hoven and Iyer/Aronson present the basis of Stein's benefit matrix. [Stein 1997] presents a benefit matrix of EIS functionality comprising seven categories, derived from literature (Strategic Objectives, Cost Saving, Flexible Information, Increased Productivity, Executive Effectiveness, Value Added Tool, Environmental Scanning). The objective of Stein was to find out how Australian executives rank these benefits. Selected findings were that timely and current information is crucial to the executives, environmental scanning was seen neutral with benefits from external EIS being seen as higher in lower organisation levels. Stein sees the advent and proliferation of the Internet, as an external source, as an explanation for this. Under the effectiveness category the data presentation and time saving benefits were dominating, however these criteria were more positively seen in the CEO[77] group than in lower organisational levels.[78] [Walstrom, Wilson 1997] describe four functions:

1. Improve information access
2. Improve communication
3. Solve problems
4. Monitor performance

---

[75] This prediction is based on several studies, showing that oral communication is the most used within organisation, e.g. by Mintzberg and Wilson himself.

[76] Mail survey to 215 firms (49 usable respondents)

[77] Chief Executive Officer

[78] Unfortunately not much can be found about the experimental setting.

Based on these functions of EIS [Walstrom, Wilson 1997] performed a study[79] with the objective to identify types of EIS users. As a result of their study the authors identify three EIS user types:

1. *Converts*, this user group has adopted an EIS as a replacement for previously existing systems.
2. *Pacesetters*, this group often makes use of EIS specific purposes, also using the EIS for communication and data analysis.
3. *Analysers*, this user group uses the EIS to perform analysis of data and ad hoc querying of the organizational databases.

EIS are typified by a intuitive GUI, providing high accumulated data, exception reporting functionality as well as navigational and analysis functions.

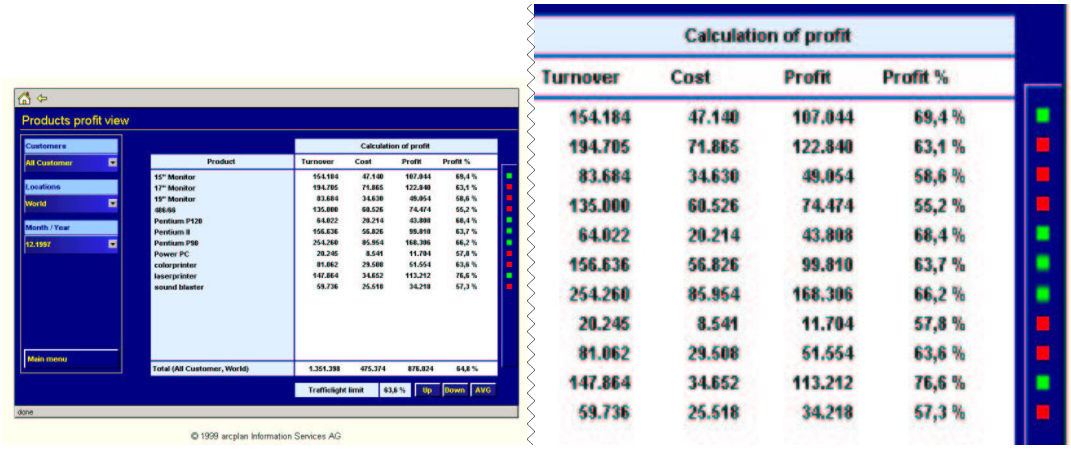


Figure 3-4: An example for the traffic light metaphor. The threshold have been selected by the user before.[80]

While the purpose of MIS is to analyse the past, e.g. number of sales per region, DSS are one level higher using sophisticated mathematical models and techniques like expert systems. Their objective is to look for alternatives, e.g. what would happen if? An example of this is the planning of routes for a transport company. EIS provide high accumulated data for the analysis of the current situation of the organisation.

Using exception reporting, e.g. metaphorically represented by traffic lights (see Figure 3-4, the user can easily get an overview of the status of the organisation (e.g. showing the main key figures, like turnover, ROI[81] and deviations using user defined thresholds),[82] using drill-down and roll-up the user can navigate through the data on different levels of detail. In recent years EIS stands not only for *executive* information systems, but also for *everybody's* or *enterprise* information system. More and more WWW browsers are used as the GUI for EIS (see Figure 3-5), e.g. presenting its

[79] Among 98 CEOs from the U.S. industry, of whom 44% use an EIS (this was checked by an entry question presenting a definition of an EIS).

[80] http://www.arcplan.de/dynademo02.htm [2001-05-21]

[81] Return of Investment

[82] In this context the term 'Balanced Scorecard' (BSC) is nowadays often in the literature. The BSC is a new type of management attitude. Not only the financial key figures of a company are in focus, but also other perspectives like customers, internal processes and learning and development. Each of the perspectives are defined as strategic objectives and potential key figures. E.g. for the learning and development perspective the number of visited educational events per employee could be a key figure. All these key figures are put in context, using cause and effect relations, early indicators and financial aspects (most late indicator). [Kaplan, Norton 1997], [Kaps, Nohr 2001], [Kaps, Nohr 2001a]

contents in the Intranet. Figure 3-5 and Figure 3-6 show an example of a demonstration of a browser integrated EIS. After selecting an initial point of interest (here *Sales Analysis*) on the entry screen the user then has the possibility to select dimensions (here *periods*) etc. to gain both an insight into and analyse the information provided.

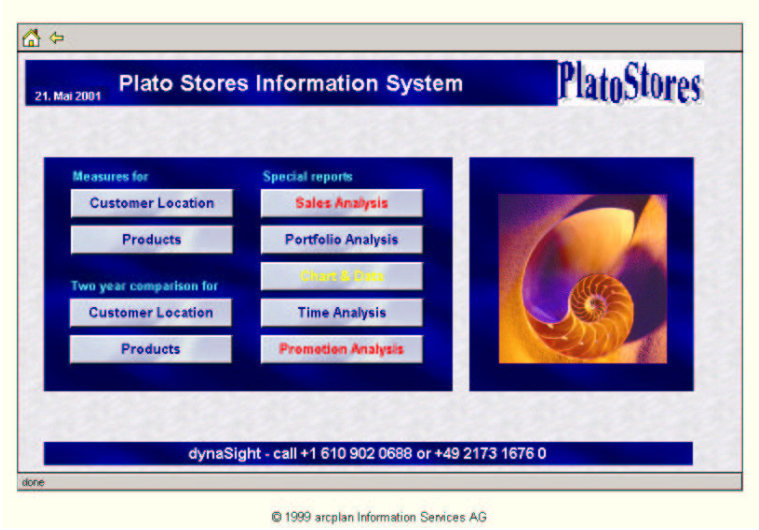


Figure 3-5: An example for a WWW based EIS with an GUI to the Microsoft OLAP Server (Demo) showing potential entry points for analyses[83]

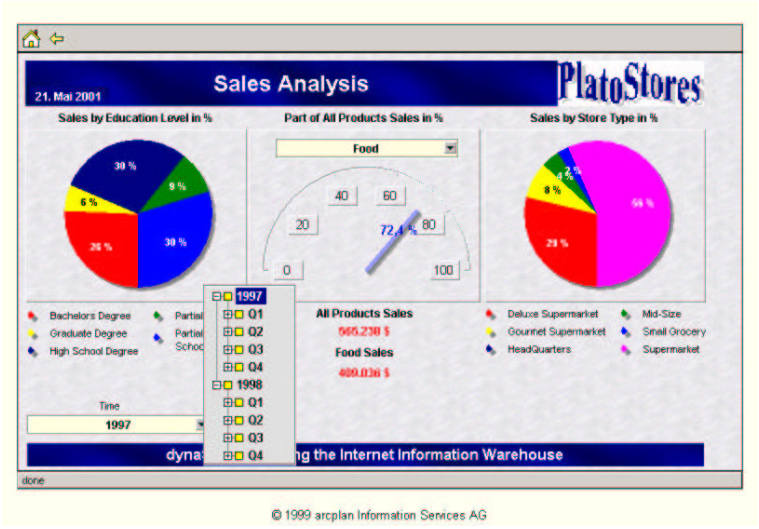


Figure 3-6: Example of a analyse with a drill down to food sales, showing a further navigation possibility by using the time dimension[84]

3.3 INSYDER as a component for retrieving External Business Information from the WWW

A demand for Business Intelligence Systems was the integration of external information (see section 3.1). According to [Drucker 1998], information systems focus too much on the internal information of enterprises rather than on the external, providing opportunities, changes and threats (see also section 2.1.1). Business Intelligence Systems try to meet this challenge by demanding the integration of external data.

[83] http://www.arcplan.de/dynademo01.htm [2001-05-21]
[84] http://www.arcplan.de/dynademo01.htm [2001-05-21]

However there are few systems and methods on the market to supply this integration. One example for such a system is the editoral workbench for the SAP Business Information Collection presented in [Meier 2000] (see also section 4.5.2.3). Thinking of the WWW as an information source even less is to be found. INSYDER as one example for such a business information retrieval system is presented within the current work. It has been a research project of the European Union with the objective of developing an information system to find, analyse and monitor business information on the Internet. It works as a kind of Information Assistant [Kuhlen 1999] on behalf of the user. Within this section its general functionality will be described, the retrieval aspects and its information assistant capabilities will be presented in chapter four. [Mann 2002] also gives a detailed description of the system with a focus on search result visualisations.

3.3.1 A content based system

The purpose of INSYDER is not to act as another search-engine, e.g. like AltaVista, Google etc., but to be a content based search assistant. This new way of designing a search assistant means, that the user first of all gets a pre-configured system, meeting his needs. It has been intended that the user gets predefined searches he can run again or use them for monitoring the WWW. The basic design of INSYDER is to give the user a sphere-of-interest (SOI), where he can organise all the information concerning his information need: searches (marked with a magnifier), watches[85] (marked with binoculars), news (marked with a notepad) and bookmarks (marked with a bookmark). The organisation of the SOI is subject-based, a sphere expresses an information need of the user, which might be manifold. For example if the user is a business manager in the sales department, one sphere-of-interest could be named *'customers*', including searches and watches about customers and bookmarks to their WWW presentation. Another sphere could be competitors, e.g. including searches for their products and the customers they serve. This way the business manager can easily keep track of these various information needs. In an internal study with users[86] from the project team[87], this concept has been well accepted.

Figure 3-7 shows a screen-shot of the INSYDER GUI. The example shows a possible SOI of a business decision maker working for a company using rapid prototyping technology for their product development. Hence he or she could be interested in patents on rapid prototyping (RP), CAD modelling services, leasing services for the equipment, research on CAD and RP and general information about manufacturers. These information needs are all reflected in the SOI, organised in the categories information services, RP equipment, RP related and CAD prototyping general (see Figure 3-7).

A further part of the content-providing within INSYDER is the pre-definition of sources. The sources are defined according to the need of the user. In the example of the sales manager, sources could be specialised electronically, available magazines about the product palette, electronic newsletters, portals, commercial databases, the Intranet, the local network, the own computer or just common search-engines. This way the searches can be restricted to an area of the WWW to return more precise results.

The INSYDER search mechanism basis is on an own knowledge base (see section 4.4.4), of which can be thought as a kind of thesaurus. This knowledge base provides a general view of the world, but can be adapted to the user, e.g. by classifying the

---

[85] Which is a monitoring, but the term used within the project was watch. See also section 4.10.

[86] Working as business managers themselves

[87] Not involved in the design of the SOI, nor the implementation

business focus in depth and in different languages. The content can then be used when formulating the query (see section 4.5.3) and for the analysis of documents found (see section 4.4).

One more content provision comes into view, when reviewing the results. Based on a pre-defined classification, the user can easily see from which source the documents result (see section 4.7). E.g. going back to the sales manager, he could obtain the information that the document he is examining comes from a competitor or a customer. Thinking of content providing it is clear that such systems have to be administrated to keep the content information up-to-date. It was intended to keep the content administration as easy as possible, e.g. the sources definitions are in XML (see section 4.4.5.1).

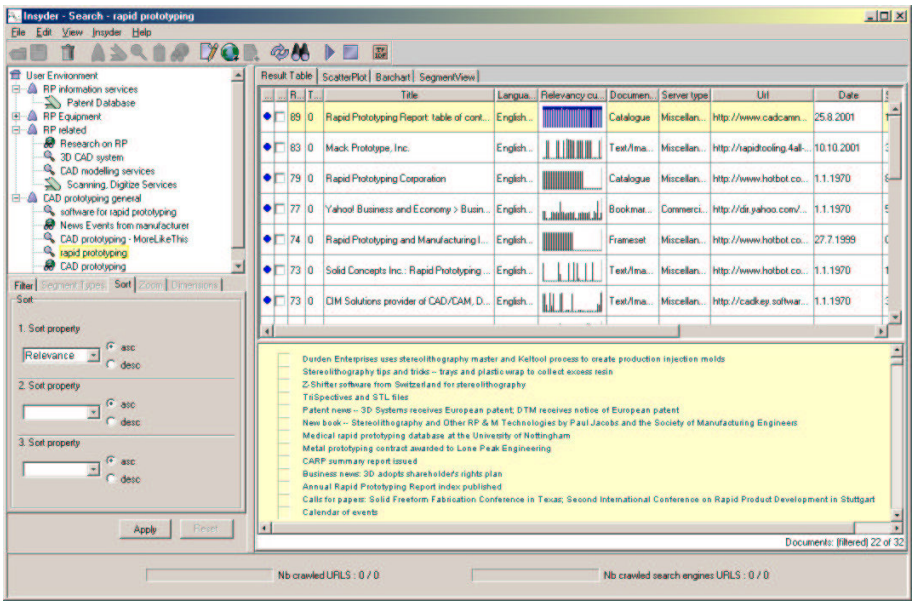


Figure 3-7: The INSYDER GUI showing the sphere-of-interest on the left

Thinking of the different steps of a search process (see section 4.5.1 for models of the information seeking process) and adapting the four-phase framework proposal by [Shneiderman, Byrd, Croft 1997] (see section 4.5.1.2), the user is in each step provided with content (see Figure 3-8, the preparation phase has been added explicitly, the processing is altered from the action phase).

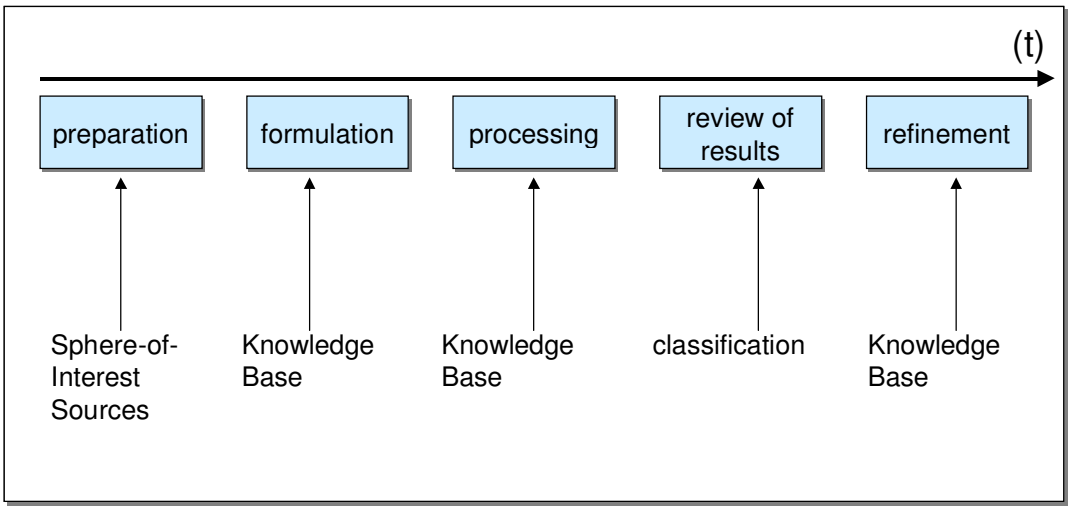


Figure 3-8: Content provision in the search process

3.3.2 INSYDER – a visual information seeking system

The user of INSYDER is also provided with comprehensive visual support, so that INSYDER can be seen as a visual information seeking system [Reiterer, Mußler, Mann 2001]. The visual support is found in various phases of the search process (see Figure 3-8). A visual query formulation component supports the user formulating and reformulating the query (see Figure 3-9, for a detailed description see section 4.5.3). Different search results visualisations using a synchronised multiple view approach support the user reviewing the results [Reiterer, Mußler, Mann 2000], [Reiterer, Mußler, Mann 2000].

Hereby the visualisation used are a result table as found from other applications. With the result table it is e.g. possible to change the organisation of the columns, to sort columns using different attributes, to size the columns or to make selections within the rows of the table. The relevance curve is the first visual element the user sees. It expresses the sections which match the query terms well. Still the length of the relevance curve does not give any information about the length of the document, therefore its use is to get a first insight on the document level. As can be seen in Figure 3-10 the visualisations are organised using a tabbed pane[88], each visualisation on one tab. The next visualisation is the Scatterplot (see Figure 3-11). Two axes define the appearance of the result set, whereby the two axes are either pre-defined or chosen by the user. This way it is easily possible to overlook a large result set, e.g. quickly finding all new documents from the result set that fit the query well.[89] Like the Scatterplot the Barchart also gives an overview of the total result set. Each document is presented using horizontal bars, where the length of the bar resembles the quality of the document found. Beside the overall view of the document in the first column, for each keyword of the query a bar is represented in different colours (see Figure 3-12), also expressing the quality of the document for the specific keyword. The fourth visualisation is the TileBars (see Figure 3-13), in contrast to the other visualisations the here the user gets a detailed view of the document level. Each document is represented by tiles, representing the sections of a document. The tiles are coloured according to the keyword quality occurring in that specific segment. This way it is easy for the user to see in which part of the document the most relevant sections are. The last visualisation is to present the result set in a "standard" static HTML list (see Figure 3-14), in a similar way the result sets are presented from the commonly available search-engines.

For a comprehensive description of the visualisations of the result sets and results from a user evaluation of these visualisations see [Mann 2002].

[88] See http://java.sun.com/products/jlf/dg/higk.htm#38176 [2001-05-14]

[89] Which means that they will have a high ranking

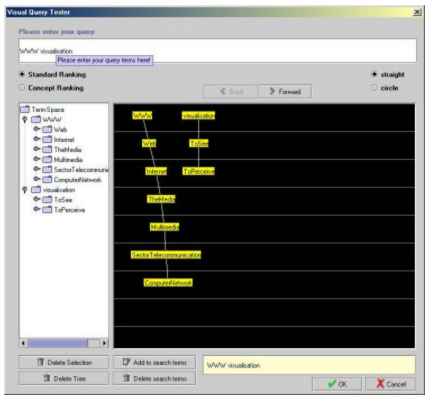


Figure 3-9: Visual Query

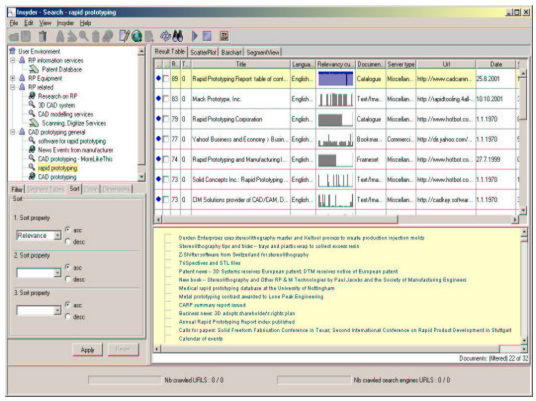


Figure 3-10: Result Table with integrated browser

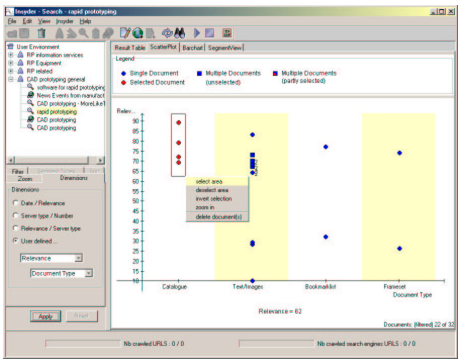


Figure 3-11: Scatterplot

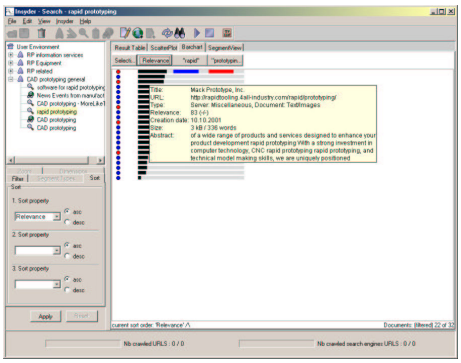


Figure 3-12: Barchart

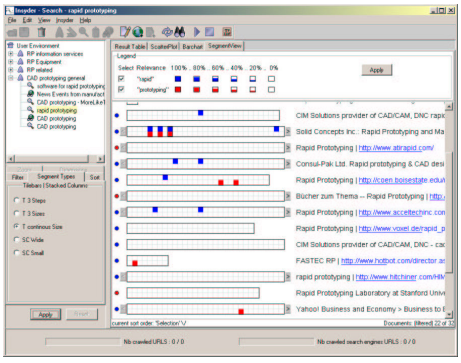


Figure 3-13: TileBars

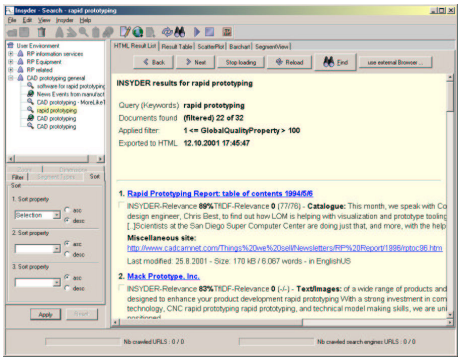


Figure 3-14: Static HTML List

3.3.3 Architecture of the system

The INSYDER system consists of several components, responsible for performing dedicated tasks within the system. Figure 3-15 shows the component view of INSYDER. The lower tier is the basis for the processing tier: the Knowledge Base (see section 4.4.4), the sources definition for the searches and the server-type definitions for the classification part (see section 4.7). For the storage of the meta-information of the search results the Microsoft SQL Server is used.[90] The documents themselves are stored using the operation systems flat file structure.

[90] http://www.microsoft.com/sql/default.asp [2001-05-14]

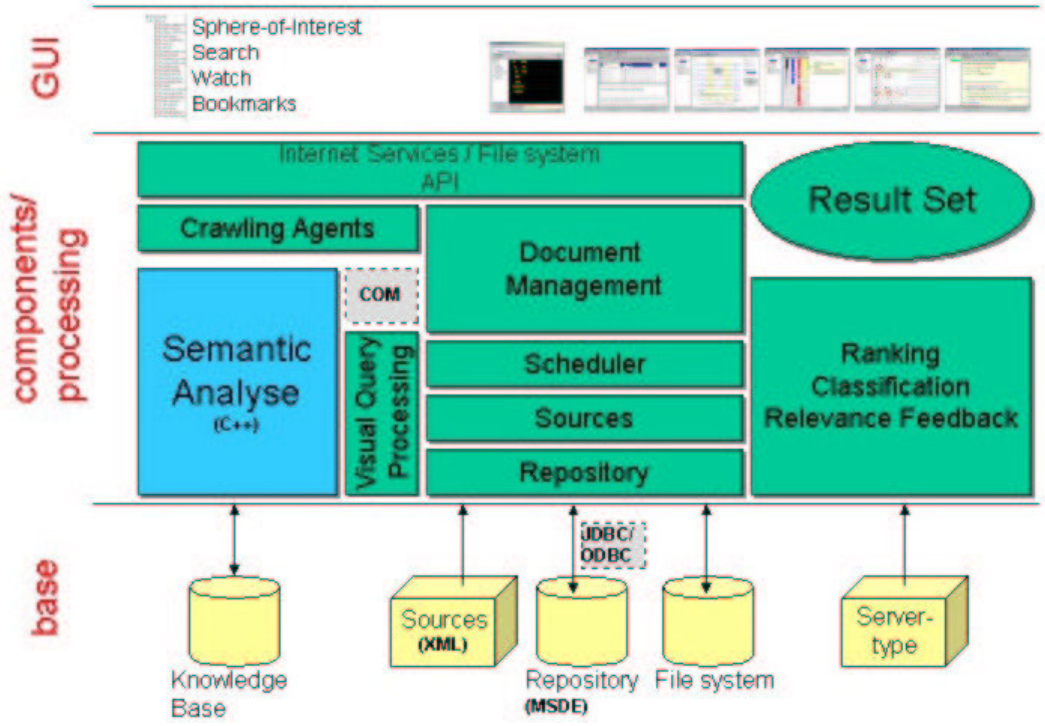


Figure 3-15: INSYDER architecture

Apart from the *semantic analysis* component, all components have been implemented using Java.[91] The semantic analysis is based on existing C++ modules from the project partner ARISEM [ARI], which could be reused and enhanced for the INSYDER project. The semantic analysis component communicates using a *COM* interface.[92] Its task is the analysis of documents. It also provides the interface to the *Knowledge Base*, which is used by the *Visual Query* (see section 4.5.3). The *document management* is the central component of the system. By document management all documents are accessible in all stages of their processing and further actions (calculation of meta-data) are triggered. The input for the system is either documents from Internet services, like the WWW or the local file system from the user's PC or network. Naturally both have their own implementation of an *API*. For the search in the WWW *crawling agents* are necessary, which use the Hypertext structure of the WWW, following the links of a distinct page, returning the new pages found to the document management. The *Ranking, Classification* and *Relevance Feedback* components analyse and classify the documents found. These components will be explained in detail in chapter 4. The *scheduler* is necessary when performing a monitoring of WWW sites, to trigger further events. The *sources* and *repository* components are the counterpart for the base components, processing the input (i.e. sources) resp. communicating with the database[93] and the file system[94]. In Figure 3-15 it can be seen that the *result set* component is drawn in a circle, this is to make it clear that the result set exists as an abstract construct, being substantiated by the different variations of the result visualisations (Result Table, Scatterplot, Barchart, TileBars, Static HTML List), which are on the GUI level of the system. All visualisations are based on the same result set, providing different views and interactions. The GUI of the visual query is the corresponding item to the Visual

[91] Microsoft J++ 6.0 with JDK 1.1

[92] "Component Object Model" software architecture.
See http://www.microsoft.com/com/tech/com.asp for details [2001-05-14]

[93] Using the JDBC data access API, see http://java.sun.com/products/jdbc/ [2001-05-14]

[94] Using rather simple I/O functions.

Query Processing component, using a graph network to visualise relationships of query terms to be used in the query (see section 4.5.3). The GUI has been implemented using the JFC[95] from SUN.[96] For the design and implementation of the GUI the Java Look and Feel Guidelines[97] were realised where possible (e.g. using the appropriate toolbar graphics, formatting etc.)[98].

3.4 Web Farming as a systematic approach for the Integration of external information in BIS

So far the integration of external information into a BIS has been looked at from a technical view. On the contrary to this with Web farming a systematic approach has been proposed for the integration of information from the WWW into a Data Warehouse.[99] It comprises several steps and applications to fulfil the integration. INSYDER can be seen as one tool within a Web farming framework. The architectural embedding is described in Figure 3-2.

The term Web farming has been used by Richard Hackathorn to describe a systematic approach for integration of data from the WWW into a Data Warehouse.[100] For Hackathorn *"The Web is the mother of all data warehouses."*[101] The objective of web farming is to deliver to the right people at the right time, information that is relevant to the enterprise. *"Web farming is the systematic refining (or cultivating) of information resources on the Web for business intelligence."* [Hackathorn 1999, p.10]. On a more detailed level, web farming has the following specific objects (see Figure 3-16):

1. *"To discover web content that is highly relevant to the business.*
2. *To acquire that content so it is properly validated within a historical context.*
3. *To structure the content into a useful form that's compatible with the data warehouse.*
4. *To disseminate the content to the proper people so it has direct and positive impacts on specific business processes.*
5. *To manage the previous steps in a systematic manner as part of the production operations of a data center environment."* [Hackathorn 1999, p.10]

---

[95] Java Foundation Classes. As the development has been undertaken with Microsoft Visual J++ Version 6, only the JFC version 1.1 have been used.

[96] http://java.sun.com [2001-05-14]

[97] http://java.sun.com/products/jlf/dg/index.htm [2001-05-14]

[98] http://developer.java.sun.com/developer/techDocs/hi/repository/ [2001-05-14]

[99] The following section refers to http://www.webfarming.com and [Hackathorn 1999], [Behme, Mucksch 1999], [Alpar, Leich 2000]

[100] Another use of the term is providing a 'web farm' that is a cluster of Web servers that collectively share the task of serving Web pages, see [Johnson 1999]

[101] http://webfarming.com/intro/intro02.html [2001-05-23]

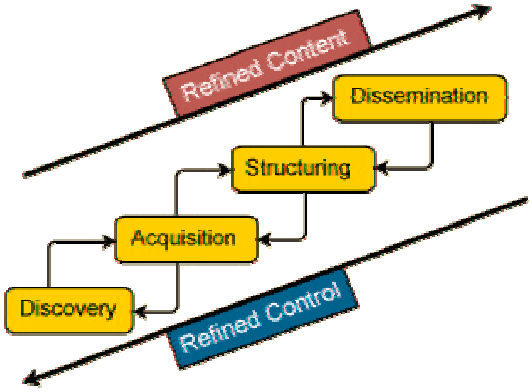


Figure 3-16: Web farming process[102]

As a methodology for introducing Web farming in a company Hackathorn proposes a four-stage approach:

1. Getting started: initiation. Putting the base for the serious phase by
   - Identifying the CEF[103] of the enterprise by profiling the company.
   - Formulation of a discovery plan for the constant monitoring and historical accumulation of information.
   - Identification of content providers (e.g. commercial databases, trade publications etc.).
   - Dissemination of information in the company to the right people by using appropriate formats.
   - Compilation of a business case with the aim to establish the Web farming within the company.
2. Getting serious: contagion.
   - Legitimise Web farming as an official part of the enterprise.
   - Build the infrastructure.
   - Refine the CEF list by determining entities (persons, places, organisations), indicators (measurement for character of the distinct CEFs) and events (great change of a CEF indicator, new competitor etc.)
   - Maintenance of historical context: keeping the historical context for each Web object e.g. by recording the deltas to the original one. Attach an appropriate time stamp. Think of notification method for changes.
   - Build an Intranet site or integrate in existing ones (e.g. from the Data Warehouse group). Implement notification method.
3. Getting smart: control.
   - Define and build selection and extraction filters to extract relevant information from a web page.
   - Build communication channels (pipelines) to information sources.
   - Structure and analyse the information, e.g. by fitting the information to conceived schemes of a Data Warehouse.

[102] http://webfarming.com/intro/intro06.html [2001-05-22]
[103] Critical External Factors

- Publishing of content to get the information used and to gain feedback about the information.

4. Getting tough : integration.

- Integrate the information with the Data Warehouse, e.g. by augmenting an existing dimension table or by creating new fact tables and associated dimensions.
- Link to other systems of the company.
- Mapping of the information by resolving entities.
- Establishment of checks to ensure the credibility of the web content.
- The Data Warehouse is now the resource centre, managing and disseminating business information.

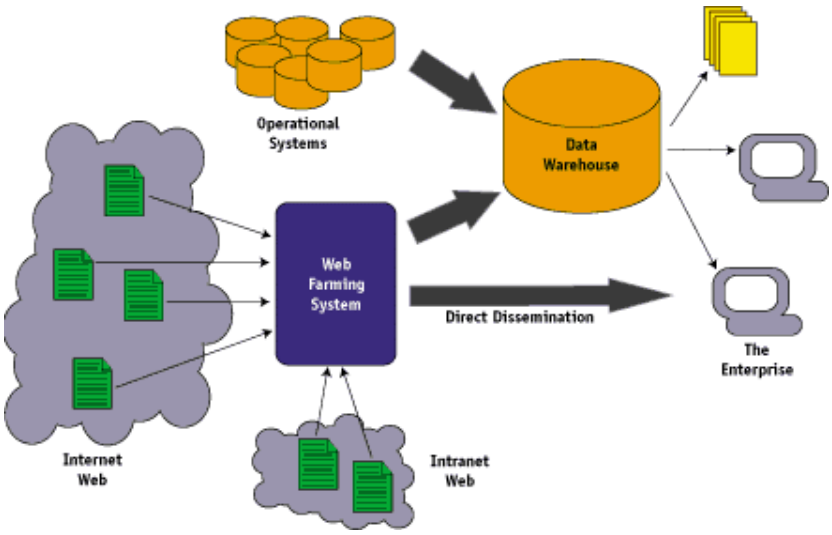


Figure 3-17: Web farming system[104]

Figure 3-17 shows an overview on the Web farming system and components. As can be seen from the four stage methodology, the implementation of a Web farming system is a major task for an enterprise and demands a variety of tools and knowledge to make it work. It is not just the use of a search-engine or the distribution of Emails. There is no application being the Web farming tool, it is more the combination of tools that help providing the added-value of the methodology.[105] INSYDER could be one element in such a tool suite.

[104] http://www.dmreview.com/master.cfm?NavID=216&EdID=1001 [2001-05-23]

[105] Unfortunately up to now no company has been found having implemented the Web farming process in the way Hackathorn proposes it. An email to R. Hackathorn himself asking whether he knows about such companies has unfortunately not been answered.