

Capture, Indexing & Auto-Categorization

Intelligent methods for the acquisition and
retrieval of information stored in digital archives

AIIM Industry White Paper on Records,
Document and Enterprise Content Management
for the Public Sector



AIIM
International



© AIIM International Europe 2002
© DLM-Forum 2002
© PROJECT CONSULT 2002

All rights reserved. No part of this work covered by the copyright hereon may be reproduced or used in any form or by any means – graphic, electronic or mechanical, including photocopying, recording, taping, or information storage and retrieval systems – without the written permission from the publisher.

Trademark Acknowledgements

All trademarks which are mentioned in this book that are known to be trademarks or service marks may or may not have been appropriately capitalized. The publisher cannot attest to the accuracy of this information. Use of a term of this book should not be regarded as affecting the validity of any trademark or service mark.

Warning and disclaimer

Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied. The author and the publisher shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book.

First Edition 2002

ISBN 3-936534-00-4 (Industry White Paper Series)

ISBN 3-936534-01-2 (Industry White Paper 1)

Price (excl. VAT): 10 €

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Printed in United Kingdom by Stephens & George Print Group

Capture, Indexing & Auto-Categorization

Intelligent methods for the acquisition and
retrieval of information stored in digital archives

AIIM Industry White Paper on Records,
Document and Enterprise Content Management
for the Public Sector

AIIM International Europe
Chappell House
The Green, Datchet
Berkshire SL3 9EH - UK
Tel: +44 (0)1753 592 769
Fax: +44 (0)1753 592 770
europeinfo@aiim.org

DLM-Forum
Electronic Records
Scientific Committee Secretariat
European Commission SG.B.3
Office JECL 3/36, Rue de la Loi 200, B-1049 Brussels - Belgium
Tel. +32 (0) 2 299 59 00 / +32 (0)2 295 67 21 / +32 (0)2 295 50 57
Fax +32 (0) 2 296 10 95
A/e: dlm-forum@cec.eu.int

Author
Christa Holzenkamp
Innovationspark Rahms
53577 Neustadt / Wied - Germany
Tel.: +49 (0)2683 984-214
christa.holzenkamp@ser.de

Executive editors and coordinators
Dr. Ulrich Kampffmeyer
Silvia Kunze-Kirschner
PCI PROJECT CONSULT International Ltd.
Knyvett House, The Causeway
Staines, Middlesex TW18 3BA - UK
Tel.: +44 (0)1784 895 032
info@project-consult.com

Published by PROJECT CONSULT, Hamburg, 2002

Industry White Papers on Records, Document and Enterprise Content Management	Series	ISBN 3-936534-00-4
(1) Capture, Indexing & Auto-Categorization		ISBN 3-936534-01-2
(2) Conversion & Document Formats	HP	ISBN 3-936534-02-0
(3) Content Management	FileNET	ISBN 3-936534-03-9
(4) Access & Protection	IBM	ISBN 3-936534-04-7
(5) Availability & Preservation	Kodak	ISBN 3-936534-05-5
(6) Education, Training & Operation	TRW/ UCL/ comunicando	ISBN 3-936534-07-1

Preface

The Information Society impacts in many different ways on the European citizen, the most visible being the provision of access to information services and applications using new digital technologies. Economic competitiveness of Europe's technology companies and the creation of new knowledge-rich job opportunities are key to the emergence of a true European digital economy. Equally, the Information Society must reinforce the core values of Europe's social and cultural heritage – supporting equality of access, social inclusion and cultural diversity. One important element in ensuring a sound balance between these economic and social imperatives is co-operation between the information and communication industries and public institutions and administrations. Over the past 5 years, the European Commission in co-operation with EU Member States, has worked to create a multi-disciplinary platform for co-operation between technology providers and public institutions and administrations. The Forum aims at to make public administration more transparent, to better inform the citizen and to retain the collective memory of the Information Society. These objectives are at the heart of the eEurope Action Plan adopted by the European Summit in Feira on June 2000. I welcome the way the DLM-Forum has evolved over this period as a platform for identifying and promotion concrete solutions to many of the problems facing our public administrations.



In 1996 the initial focus of the DLM-Forum was on the guidelines for best practices for using electronic information and on dealing with machine-readable data and electronic documentation. More recently, at the last DLM-Forum in Brussels in 1999 a challenge was made to the ICT industries to assist public administrations in the EU Member States by providing proven and practical solutions in the field of electronic document and content management.

The importance of providing public access and long term preservation of electronic information is seen as a crucial requirement to preserve the “Memory of the Information Society” as well as improving business processes for more effective government. Solutions need to be developed that are, on the one hand, capable of adapting to rapid technological advances, while on the other hand guaranteeing both short and long term accessibility and the intelligent retrieval of the knowledge stored in document management and archival systems. Furthermore, training and educational programmes on understanding the technologies and standards used, as well as the identification of best practice examples, need to be addressed. I welcome the positive response from the ICT industries to these challenges and their active involvement in the future of the DLM-Forum, for example in the event proposed in Barcelona in May 2002, to coincide with the EU Spanish Presidency.

The information contained in the following pages is one of a series of six ICT Industry White Papers produced by leading industry suppliers, covering the critical areas that need to be addressed to achieve more effective electronic document, records and content management. I am sure that the reader will find this information both relevant and valuable, both as a professional and as a European citizen.

A handwritten signature in dark ink, appearing to read 'Erkki Liikanen'.

Erkki Liikanen

Member of the Commission for Enterprise and Information Society

Preface Sponsor

We are currently faced with an ever-increasing overload of information and must decide how we will go about mastering it. An individual can read approximately 100 pages per day, but at the same time 15 million new pages are added to the Internet daily. Our limited human capabilities can no longer filter out the information that is relevant to us.

We therefore need the support of a machine which facilitates the exchange of knowledge by storing information and enabling individual, associative access to it through the lowest common denominator in human communication: The common human index is natural written and spoken language.

All other types of indexing are limited aids which humans must first learn to use before they can employ them. To sum it up, the standard has already been set and recognised as language, but where are the systems, which have adapted this standard?

A handwritten signature in black ink, appearing to read 'Gert J. Reinhardt', with a stylized flourish at the end.

Gert J. Reinhardt

Table of Content

1.	Introduction	6
2.	The importance of safe indexing	7
2.1	Description of the problem.....	7
2.2	The challenge of rapidly growing document volumes.....	9
2.3	The quality of indexing defines the quality of retrieval.....	9
2.4	The role of metadata for indexing and information exchange.....	12
2.5	The need for quality standards, costs and legal aspects.....	12
3.	Methods for indexing and auto-categorization	14
3.1	Types of indexing and categorization methods.....	14
3.2	Auto-categorization methods.....	16
3.3	Extraction methods.....	19
3.4	Handling different types of information and document representations.....	22
4.	The Role of Databases	25
4.1	Database types and related indexing.....	25
4.2	Indexing and Search methods.....	28
4.3	Indexing and retrieval methods using natural languages.....	30
5.	Standards for Indexing	35
5.1	Relevant standards for indexing and ordering methods.....	35
5.2	Relevant standardisation bodies and initiatives.....	39
6.	Best Practice Applications	41
6.1	Automated distribution of incoming documents Project of the Statistical Office of the Free State of Saxony.....	41
6.2	Knowledge-Enabled Content Management Project of CHIP Online International GmbH.....	46
7.	Outlook	53
7.1	Citizen Portals.....	53
7.2	Natural language based portals.....	54
	Glossary	55
	Abbreviations	58
	Authoring Company	59

1. Introduction

According to a Berkeley University study, 12 exabytes of information have been generated during the past 100,000 years. This information has been stored on the most various types of medium like cave walls, clay tablets, papyrus, paper, magnetic discs and tapes as well as laser discs. The information contained on these storage devices was first delivered by messengers within a limited area and then later with faster and wider reaching means of transportation – ship, automobile, rail, airplane, fax, radio and Internet. In that past two and half years, a further 12 exabytes of information have been produced and the next 12 exabytes are forecasted for within the next year. Even if one assumes that much of the information contained in this amount is repeated or at least much of it is basically the same in different words, the remaining amount of information is still much too extensive for a single person or an organisation to gain an overview.

Every individual sets up their own personal structure so as to avoid having to search through piles of random information. This structure is perfectly understandable for the individual, but can someone else locate information within this structure without understanding it? Only after having been trained to use the structure would someone else be able to find what they are looking for. Such structures are set up to store the information of individuals, departments, groups, companies and government offices. As long as each individual holds to the set structure and everyone knows how to use this structure then there is at least a basis for accessing the stored information.

Filing structures are not the only thing, which need to be agreed upon however, but also the method, which is used to make the information available. If we think about this in terms of each EU citizen, then the accessing method has to be so simple to use that everyone would be able to use it without requiring lengthy instruction.

There are two challenges, which have to be met to achieve this:

- The growing volumes of information have to be kept under control and
- the storage of and access to this information has to be completed in such a way that everyone is able to work with it.

These requirements can be answered by applying the already agreed upon method for dialog between humans – it is based on their natural language. If the EU governments and other public administration adapted such an approach for their internal and external communication then they would be well positioned to demonstrate their citizen friendliness.



John Mancini
AIIM International

2. The importance of safe indexing

By utilising automated capturing, categorization and indexing, organisations can create measurable benefits on different levels, mainly distinguished as quality, time and costs. Some of these benefits are:

Quality

- Enhanced and automatic checking and validation of information
- More information, more in-depth information and content-related details
- Better user specific response in communication
- Improved access to an enhanced information and knowledge pool

Time

- Process speed-up, elimination of up to 90 % dead-times
- Processing “on demand” becomes possible
- Reduction of manual processing steps (handling and capturing)
- Delivering results in real-time (deliver when ready)

Costs

- Reduction of costs for handling, capturing and maintaining data and information
- Planning of personnel resources closer to actual needs
- Better cost transparency and improved planning of costs

2.1 Description of the problem

Why is indexing important? When indexing, a “magic word” is set with which dedicated and extensive worlds of information can be opened and captured. Those who don’t know the magic word will not be able to gain access to the sought-after information. The magic indexing word is not just limited to access however, but is also used for every document in a set classification scheme so that these remain accessible with the least effort possible. The magic word is inevitably not a short “open sesame”, but instead drastically more complex. The word “complex” already gives a clue that a great amount of effort is required to create this complex magic word, to consistently assign it to all documents (conformity) and to learn to use it for accessing (interoperability).

Indexing challenges

If one considers everything that is implied with the term “indexing”, then one suddenly realises that there is not just one problem, but an entire network of problems that have to be considered. Just how multifarious the indexing problem is, can be seen in the following:

- Information volumes are growing faster in size and in scope than any individual or organisation is capable of controlling with a reasonable amount of effort and success. Even if indexing could be realised and automated, the problems caused by the sheer information volumes will be encountered, at the very latest, when accessing is

attempted. The technical means, such as internal and external network communication bandwidths or storage capacity, haven't even been considered because they no longer present a serious problem in regard to performance and costs.

- By its very nature, indexing is very multifaceted because all individuals have their own personal classification schemes and use them to organise both their professional and personal environments. If individuals do not introduce their own classification schemes, then they will be forced to subject themselves to the classification schemes of others. That again means an increased amount of effort because a person has to learn to operate a foreign classification scheme before they can use it.
- Indexing is completed using many various structuring approaches. If there is information to be classified, then the question of which classification criteria is to be employed must be answered first in order to keep the information retrievable.
 - Author / Source / Title / Topic / Contents / Period in which the document was created / Publication date and so on
- Indexing is currently approached with many conceptual methods:
 - Name / Digit series / Keywords, among others / Entire contents
- There are also many technologies which are employed for the indexing of documents:
 - (Alpha) Numerical keys / Full text, from key words to complete text / Natural language
- There are, in addition, technologies to be noted which are intended to enable exchange and transport, like XML as a neutralising format. Although Microsoft Office applications are considered de facto to be the desktop PC standard, they are still not the only such products. For this reason other document formats also have to be taken into account.
- Yet another technological background plays a role here, and that is the form of the document whose contents and index are to be saved or even archived. In general, the index is stored separate from the document, typically in a database. That means that there are two storage systems and locations to be maintained.

“Safe indexing” is not only restricted to the physical or logical quality of the storage process itself (= no loss of information, every item can be retrieved, ...). It can also be extended to mean “safe” in terms of “quality of business”. Does the indexing method used supply all information for storing the information and – later on – to make this information available including related content, which corresponds to the given case? There is a fundamental relationship to be noted: The quality of the index information determines the quality of the matching result.

There is one additional factor, which must be taken into account when talking about “safe indexing”. The regulations and laws initially applied to the retention of digitally archived paper documents have now been extended to include electronic documents like emails and any other web-based business interaction. “Safe indexing” also means being able to cope with these extended requirements.

2.2 The challenge of rapidly growing document volumes

The overall challenge is knowledge access from anywhere and at any time, and this despite information type and quality. Thanks to the today's electronic communication and networked organisations, an individual can contact an organisation from anywhere in the world at any time. This might be done by writing a letter, transmitting a facsimile or – with exponentially growing importance – by sending an email or by just accessing the desired website. In any case, the available time to give a high quality response by any means is reduced from weeks and days down to hours. Today we expect “my knowledge on demand”.

The challenge for an organisation is to be able to react at any time and to process incoming requests instantaneously. Consequently, tasks like categorization and capturing are no longer limited to an “off-line” process, which precede archiving but have to be applied spontaneously to all incoming requests and associated document items. International and inter-organisational communication also require multi-language environment support.

Manual capturing, indexing and categorization have worked quite well in the past – as long as document volumes to be processed remained small. Employees are able to “read” document by document and to attribute them to distinct categories. Employees are also able to capture a certain amount of dedicated information for building an index for storage and / or for following case treatment. This works quite well for a few hundred documents per person, which can be categorized by just taking a short glance (e.g. in-coming mail). It is also an applicable approach for a few dozen complex documents per person, which have to be processed per day (studies, research, news, etc.).

The situation has changed completely since information volumes have begun to grow more quickly. Considering for example an organisation with incoming mail (paper, facsimile) of more than 100,000 items a day or electronic communication across several 10,000 mailboxes. Even at a first glance, environments like this clearly exceed feasible limits for productive solutions. And the quantities will continue to grow!

2.3 The quality of indexing defines the quality of retrieval

The quality of retrieval and access directly depends on the quality of indexing. “Quality” here means firstly, short filing time and quick access and secondly, context-related retrieval results which can be accessed immediately.

With regard to manual and semi-manual indexing, today's situation can be described succinctly: Document volumes and complexity of content are directly correlated with the set of attributes / features for building an index. This means, with fast growing volumes and increased complexity, attribute schemes are simplified more and more thus resulting in a loss of content, related context and information.

Depending on the specific needs of an organisation, different strategies for storage and retrieval can be distinguished – with varying requirements for the indexing task. In principle, all these approaches can be summarised in three different concepts:

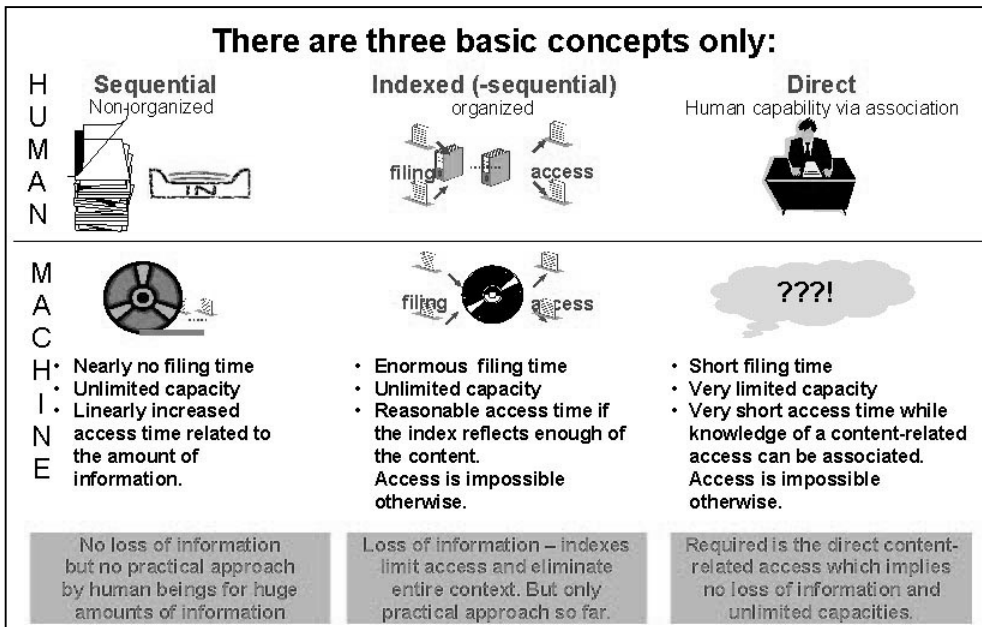


Figure 1: Basic concepts for storage and access

2.3.1 Sequential storage and access

People pile documents in a sequential manner, one on top of the other as they come in. No organisational aspects are considered. The comparable machine is a tape or a cassette, which records the data sequentially. This way of storing information takes almost no time. The storage capacities are only physically limited but they are logically unlimited. To access a document requires the same sequential approach. It is only a question of time until the required information is found. The sequential access approach leads to a linearly increased access time related to the amount of information which has been stored earlier. The conclusion is that no information gets lost but the effort to access the required information is not practical and thus not acceptable for individuals when handling a huge amount of information.

2.3.2 Indexed storage and access

Incoming information is indexed and categorized under these 'labels', i.e. the stored items are organised in an index. The manual approach is comparable with adding new documents to a defined folder. The matching electronic system can be e.g. a CD or a hard disk where the data are indexed as well. The manual indexing of all incoming information takes an enormous amount of time because for most of it, the index or indices are not readily evident. The storage capacities are only physically limited but they are logically unlimited.

Using the index-based access method appears to be reasonable and leads to the required information as long as the index reflects enough of the content. But as soon as the index does not reflect the content, the information access fails.

This leads to a loss of information. But nevertheless, this has been the only approach so far which is practical and acceptable for handling large amounts of information. A special version of indexing is the full text index which is commonly used in press systems, document management systems, the Internet, etc.. Other special index methods (most of them are proprietary) are delivered by all kind of DMS, IDM and eDM solutions.

2.3.3 Storage independency and direct access by association

This approach can be usually chosen by people only because a person's content-related access capabilities are driven by associations. The time, which is required to file new information, is that short that it is negligible. The required capacities are very limited by human nature. A comparable approach provided by a machine has not been generally introduced as of yet.

The access time for a human is very short because the knowledge of a content-related access can be associated, i.e. as long as the location can be associated direct access is possible. The flip side is that no access is possible when no content is associated and the question for the location does not even come up.

As a result, a machine is required that can access content and context related to any type and limitless quantity of information and knowledge, no matter where and how it is stored. This would prevent a loss of information through unlimited capacities and provide enormous time savings.

2.3.4 Which access method is the best?

Conventional access methods have been well known for a many years but the problem that remains is how to get direct, associative access to the required knowledge which is to be withdrawn from the vast amounts of stored information. Otherwise information will be lost because you cannot find it anymore or you do not attain access the best matching information. Conventional access methods simply fall short because they are not able to efficiently master the amount of information in time and quality anymore. These access methods are not scalable enough to fulfil people's needs.

New access methods are required which deliver an unlimited scale with fast and reliable direct access. The sort of direct access which can be compared with the human's way of accessing knowledge – by association, which combines content with the individual context.

The fulfilment of this requirement for direct access will not replace the sequential and indexed methods but it will become a complementary and strongly desired method. The indexation method of DM systems remains meaningful because such a cascading indexation allows the declaration of a logical context between a document and a folder or a folder inside another folder.

The combination of all access methods frees the way to the entire information and knowledge space that is required. The results which are produced by the context-related accessing of a dedicated knowledge domain inside the knowledge space will be more complete than if you look for the matching documents yourself because you would struggle not only with the amount of information but also with the information which is

available in thousands of copies of the original. In the end, content-related accessing is faster, more precise and higher in quality.

2.4 The role of metadata for indexing and information exchange

As already mentioned at the beginning of this section, indexes can look very different since diverse criteria have to be considered. Depending on the required and selected indexing methods (i.e. simple number series or extensive descriptors with formulas and thematic keywords), the question is to what extent it is possible to create separate metadata with the index information.

Document management and archiving systems are usually operated using separate index metadata, which is not only separate from the contents, but is also stored separately in the database. Most importantly this serves to distribute the search criteria and to keep them redundant whereas the documents themselves are stored centrally. In addition to the performance aspects, the security aspects also advocate metadata and documents being stored separately.

Organisationally relevant information is no longer just processed in document management systems. Content management systems, which also have storage and access methods, are used parallel in Internet and Intranet environments

In addition, the introduction of Internet systems have made it more and more apparent that most information is not maintained in DMS or CMS applications but instead is stored more or less unstructured on file servers. In order to make all this information internally and externally available, a method has to be introduced which is not limited to separate indexing within an individual application environment. The volume problems are thus intensified even more.

As already mentioned at the beginning, the smallest common denominator for information is to be used for storage and accessing. A “neutral key” based on natural language, which includes the entire document content when storing as well as when accessing, is fast and reliable because its full content metadata can be used separate from the original document.

2.5 The need for quality standards, costs and legal aspects

Quality standards

The definition of quality standards and the implementation of quality management procedures is an important issue and becomes more important at the moment when intelligent software takes over responsibility for capturing, categorization and indexing tasks – finally resulting in highly automated operations within an organisation often without human interaction.

Today organisations work more closely together than ever before – both interdepartmentally and internationally – and have to agree on storage and access standards which can guarantee that the quality level of document related information processing always remains consistent. Rules and regulations applied to a distinct business case should always lead to the same result, independent of any specific operating conditions an organisation is facing. This is true for manual processing as well as for semi-automatic and automatic document processing.

Cost aspects

Factors impacting costs arising from an intelligent software solution with auto-categorization, information extraction capabilities and direct access are listed below. Due to significant increases in efficiency, which are realistically achievable, fix costs (related to each document / case) can be reduced as well.

- Cost reduction drivers for storage / archiving
 - Reduced effort / time for document preparation
 - Reduced effort / time at document auto-categorization (only correction + completion)
 - Reduced effort / time for data capture, completion and perhaps correction.
- Cost reduction drivers for retrieval / access:
 - Reduced effort / time to locate and access the items desired.
 - Reduced effort / time to get access to context related items.
 - Almost no training efforts if natural language is used.

Legal aspects

During the past decade, most legal aspects have been resolved for paper documents put into digital storage systems. The issue has now come up again along with a wider use of electronic documents covering a broad range of business and legal relevant cases with the Internet as the driving force behind it. Electronic documents are easy to handle and simple to access. Creating an arbitrary number of copies is inexpensive and done in a straightforward manner. In general, each copy is an exact reproduction of the original document and it can be distributed very quickly. Modifications can also be made to digital documents using standard techniques however, and even major changes might not be easily recognised.

Applications which rely on electronic documents and binding communication transactions require new concepts for document authentication and enterprise-wide document security. Appropriate and reliable methods also have to be applied to digital documents. This also impacts categorization and indexing.

As part of business processes handled between a document “creator / owner” and a document “user / consumer” – also involving service providers and information distribution mechanisms of any kind - the approval of rights and authenticity is a key issue. This is true for document owners within organisations and users outside of organisations and vice versa. In both cases, proof of intellectual property rights for the document owner and proof of authenticity for the document user are necessary in many cases to come to a legally binding transaction.

The best protection mechanisms usually lie inside the documents and are most effective against misuse regarding the original content and format. Security settings avoid opening without password, changes, copy and print. Digital watermarks can additionally prove the genuineness of e.g. legal documents. All other cryptographic techniques can protect against distribution and storage but not against removal or change. Nevertheless, security and authentication aspects can have an influence on the set of attributes for building a “safe” index, because the representation of an electronic document may be changed / altered or additional information (digital watermarks) may be added to the original document.

3. Methods for indexing and auto-categorization

The majority of today's storage / archive systems are organised in an indexed-sequential manner. This means that they are able to provide arbitrary and fast access, however some information is always lost, at least context information. In general, index attributes must reflect all content-related aspects. Content, which is not represented within the index cannot be accessed at all.

An additional practical issue is that index information is usually logically structured (assorted tables, mapped onto hierarchies) and its reorganisation is extremely time-consuming and perhaps even difficult. Creating new “views” - meaning different ways to analyse and retrieve information related to business relevance - is not particularly feasible. The best approach for overcoming the dilemma can be to introduce additional content related indexing which is based on portions of textual representations or even more by using the complete document content. New technologies are currently available which overcome the limitations of conventional full-text indexing / retrieval systems – achieving stable operations, language independence and fault tolerant processing. These methods are not contradictory to conventional indexing approaches in organising storage and retrieval but create a logical extension, leading towards a multi-dimensional content-based architecture for storage and retrieval.

3.1 Types of indexing and categorization methods

Manual indexing

Manual indexing is realised by manually keying-in all index relevant attributes from the documents within a given case. The user enters the indexing terms into the designated fields on the indexing screen. To reduce the many individual variants and errors when entering terms, such fields can pre-filled with default values which can also be declared as overwriteable.

Data entry fields can be programmed so that manual entries are not allowable but instead a selection list is displayed. Such lists can be integrated coming out of any program or database and can be automatically accessed by clicking an entry field. The prerequisite for integrating such list is that they are accessible via ODBC interfaces. Selection lists can contain a defined vocabulary like e.g. a thesaurus with which the selection of certain terms is limited. A further aid when entering values can be achieved with hierarchically controlled entry fields. This means for example that a selection list is shown for the first entry field and depending upon what is selected, further selections are limited in the second entry field. This can be continued over several levels.

Manual entries, predefined values and values from the selection list can be checked for plausibility in several ways:

- regarding individual format and spelling, e.g. an item number format: xxxxx-xxx.xxx
- regarding the logical relationships of the field contents, e.g. customer names and customer numbers
- regarding logical field relations, e.g. checking to see if there is an order in the ERP system which matches an invoice that has been received.

This approach is clearly limited regarding performance (= throughput) as well as complexity of the information contained (= related content). An error-rate of several percent is typically experienced and the risk of unclear or ambiguous assignments grows as the number of attributes regarding categories is increased. Manual indexing is done using paper documents as well as digital images.

The fields on a storage screen for delivery notes are filled using database entries. The user simply selects a dataset from the supplier table. This prompts multiple fields on the screen to be simultaneously filled and ensures that only valid supplier data is used for indexing.

Semi-automatic indexing

Semi-automatic indexing gives certain software support by identifying well-defined document types (e.g. dedicated forms issued by an organisation) and capturing data from previously defined reading-zones on these forms. Semi-automatic indexing for paper documents is always based on digital images, which have to be scanned in advance. Such indexing has been realised in many cases since the 80's and is able to show a productivity increase for well-defined processes with a clearly defined and structured communication based on documents, especially forms. The importance of these methods has decreased over the last years, since the degree of unstructured documents and information is constantly growing and multi-channel communication has been introduced.

Automatic indexing and categorization

State-of-the-art automatic indexing and categorization methods are introducing a new quality into the entire process. They no longer concentrate on processing a distinct sub-set of documents only (for example: forms), but are designed to properly cover the complete document stream within a proper organisation.

The goals for the entire process of automatic indexing and categorization can be defined in the following way:

- Broad coverage = every document item can be assigned to a certain business process.
- Detailed analysis = every piece of information required can be extracted and captured.

Automatic indexing and auto-categorization do not make employees superfluous but instead lead to a redefinition of their roles within an organisation. With support from such a software solution, employees work on the tasks of correction, completion and confirmation of all those document items which could not be processed completely automatically, or which have not reached the defined quality levels for categorization and extraction. Practical experiences have shown that a combination of complementary methods for categorization and extraction lead to the best results according to the goals defined above: "broad coverage" plus "detailed analysis".

Enhancements by auto-categorization

Application of auto-categorization methods, which support and extend basic attribute schemata by utilising complementary analysis techniques for different types of documents. Among the set of methods used for "classification" are for example content-based similarity matching or phrase analysis.

Dedicated information is captured from documents by using “extraction” methods based on format analysis and zone reading. Again, content related information is taken into account to decide ‘what’ to capture and ‘how’ to validate extracted information.

Practical use and quality highly depends on ‘ease-of-set-up’ and ‘ease-of-use’ for the entire application (e.g. by “learning” new content and categories), the stability under real operating conditions within an organisation (quick adaptation to changing processes and different content) and the quality of the results themselves. All these issues precede fast growing document volumes.

The results delivered by an auto-categorization mechanism can be used for both the decision on how to route and further process a document within the organisation following the defined business processes and to collect a better and more complete set of attributes for storage or archiving.

Content-based “index” for context-related retrieval

Based on the results of the auto-categorization step, an initial set of attributes is supplied which determines the content and business dependence of each document much more precisely and with less errors as compared to manual indexing.

Taking this as an input, a “textual index” can be created through a second step, which allows content-related access to arbitrary documents within a given content store. Today’s methods allow context-related retrieval based on content samples - without the need for exact wording and phrases like full-text engines typically require. Additional features add more benefits to fault tolerance against misspelled words or wrong characters as well as language independence. The related index information is compact, quickly generated and updated and can be accessed very quickly.

In addition, a content based “index” is created, which allows the so-called ‘associative access’ to find related content within other documents by a similarity match. The combination of both approaches results will lead to best-quality results.

The supply of an enhanced indexing scheme based on auto-categorization methods combined with standard attribute patterns forms the basis for a high-quality index. In conjunction with additional content-related “indexing” methods, the retrieval quality can be improved – especially with regard to environments where a higher portion of unstructured documents and information are included.

The following methods (so-called “engines”) are part of today’s state-of-the-art products and have been successfully implemented in various productive environments. They target the complete range of structured and unstructured documents, which organisations currently handle and process.

3.2 Auto-categorization methods

3.2.1 Content-based similarity match (“learning” algorithms applied)

Based on natural language this method makes use of the capabilities of trainable ‘neural networks’, which are able to distinguish documents by their content and assign them to different (distinctive) categories. It is a very stable method due to its fault tolerance capabilities and language independent operation and can be applied to all text-based documents or fragments of text by choosing any representation or possible format.

The method is based on “learning” algorithms, which basically rely on a small set of documents (“learn-sets”, typically between 10 and 30) per category to learn how to distinguish previously unknown documents from each other based on their content. In fact, there is no additional effort or definitions required to create a categorization scheme.

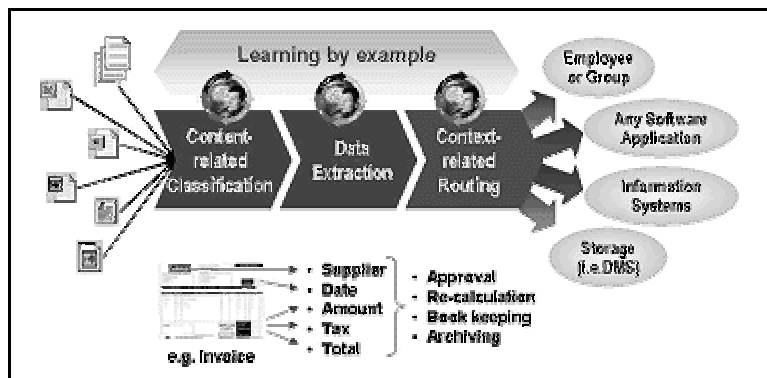


Figure 2: Data capture and collection after scanning/OCR

3.2.2 Template analysis (“learning” algorithms applied)

Templates are a specialisation of the content-based similarity match, tailored for typical business or formal documents. In such documents, the header and footer sections often include enough information and characteristic features to distinguish documents from one another and to assign them to the proper class. Template analysis is also a “learning” algorithm with which only one or two sample documents per class typically need to be defined. Traditional “pixel-based” methods - like logo analysis or label detection - are losing relevance because the portion of electronic documents is increasing steadily and the variety of document designs are changing more rapidly.

3.2.3 Phrase analysis

Phrase analysis allows either a precise or a fuzzy match of single words or multi-word phrases. It can be seen as an extension and enhancement to the two core analysis methods described above. Phrase analysis based on such approach is able to place emphasis on a dedicated aspect given by a certain document.

Phrase analysis is particularly meaningful if well defined terms or a set vocabulary is part of the regular communication within an organisation or between organisations. Such a specialised dictionary can be also used for categorization - to identify further information preceding the selection of a distinct category.

3.2.4 Forms and document size analysis

The analysis of forms and the inclusion of document size as additional criteria for categorization are well known from the area of forms processing. These methods may be still a good choice if an organisation has to deal with a major quantity of forms, which can be distinguished by detecting and reading a key-field (= forms ID) or even by simply

comparing the size of scanned images. Many organisations within the public sector still have a major portion of these types of documents, e.g. tax forms.

3.2.5 Usage of such methods

The set of categorization methods as outlined above can be combined in different ways to generate a result with the highest possible accuracy. Therefore, each result of the entire categorization process receives an individual “confidence” value between 0 and 100. If there is a categorization scheme built upon e.g. 10 classes, there are 10 confidence values per document calculated - one value for each class.

If more than one categorization method is applied, the overall confidence value for a document can be calculated in different ways, by choosing the maximum value from the individual results delivered by the different methods or by calculating the average or a weighted value – see figure 3.

Classes/Engines	Result	Templat...	Phrase Cla...	Brainware ...
Normale Fortsetzung	100,0	-	-	100,0
Leistung	85,7	-	0,0	85,7
Fortsetzung	75,1	-	-	75,1
Beschwerde	52,2	-	-	52,2
Rückabwicklung	36,8	-	20,0	36,8
Zahlung	0,0	-	0,0	0,0
Zahlung dringend	0,0	-	0,0	-
Zahlung Gerichtskosten	0,0	-	0,0	-
Zahlung Vorschuss	0,0	-	0,0	-
Vorstandsbeschwerde	0,0	-	0,0	-
Blitzbrief	No	9,0	0,0	-
Einzugsauftrag A4	No	1,9	0,0	-
Namensänderung	No	0,0	0,0	-
Antrag	No	-	0,0	2,1
Kündigung	No	-	0,0	54,5
Adressänderung	No	-	0,0	52,7
Adressänderungen-Postkarten	No	-	0,0	-
Inkassoänderung	No	-	0,0	54,8
Vertragsbeschwerde	No	-	0,0	-
Universelle Änderung	No	-	0,0	-
Sonstige	No	-	0,0	-
Bestand	No	-	No	-

Figure 3: Result built by using three different auto-categorization methods

3.2.6 Hierarchical Categorization

With an increasing number of different topics to be handled, the complexity of the categorization scheme can be reduced while quality is increased by introducing “hierarchical categorization”.

The semantics within the hierarchy can be modelled based on different approaches, like

- organisational structures: divisions, departments, teams, members, ...
- business processes such as inquiries, orders, invoicing, ...
- structure of a product range or services
- skill profiles, etc.

Based on such a hierarchical tree, the concerned entity can be directly matched with the document-related requirement. Two examples follow which help to explain this:

- If a letter is received with a precise request/question/response, it can be assigned to a highly specialised and dedicated category (= a “leaf” of the categorization tree) and directly forwarded to a dedicated group of employees who are qualified to handle these types of specific cases.
- If the content is more vague or of general nature, one of the “nodes” within the categorization tree may be chosen. In this case, the document will be at least assigned to the correct department.

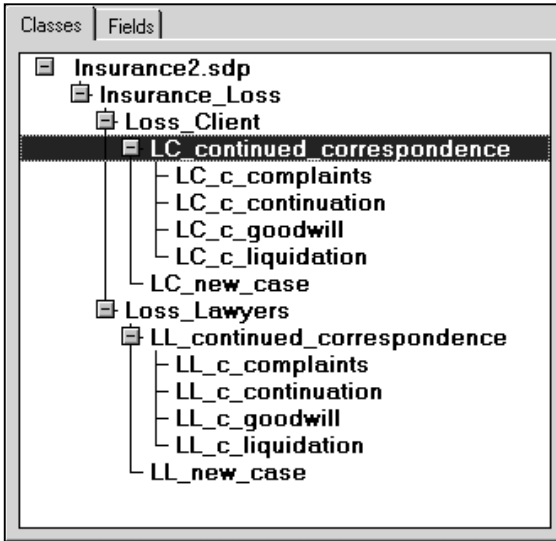


Figure 4: Customised class hierarchy

Based on a set of complementary methods, auto-categorization can be realised for organisations with a broad coverage of all documents to be processed. Content analysis and context-related matching techniques are key for productivity and quality – also under dynamically changing production conditions.

3.3 Extraction methods

Based on the categorization described above, results can be achieved using basic indexing schemes. By capturing additional data using intelligent extraction methods, the set of attributes can be completed and the necessary information for subsequent case treatment collected. Again a set of complementary methods can be applied to achieve the best results in terms of quality. And again “learning” capabilities will reduce the maintenance efforts and enable adaptation to changing content and varying context in short time.

3.3.1 Format analysis

Format analysis techniques are used to build a set of “candidates” for certain types of data fields to be extracted. Format analysis is based on the textual representation of an

electronic document. Geometric information (= locations) may be also used if available – in this case working with OCR/ICR results. Format analysis is typically applied to determine format characteristics for process relevant information like:

- Name / Identification number / Date / Amount values / Account number, etc.

3.3.2 Context-related extraction (“learning” algorithms applied)

Based on a set of ‘candidates’ delivered by the format analysis, context-related extraction is used to filter the correct candidate out of a list. This can be done for each document type by forming a specific class but may also be applied to more general categories (= “nodes” within the classification hierarchy) like applications, inquiries, complaints, ...

Context-related extraction performs an analysis for each candidate in the surrounding context information and creates a measure for performing a similarity match. Again, the decision is built by using a “confidence” value as a measure for quality - similar to categorization.

Through example documents, context-related extraction learns where to look and which candidate to select. Fuzzy similarity-matching technologies result in stable operations and high quality results even if the document structure and wording change.

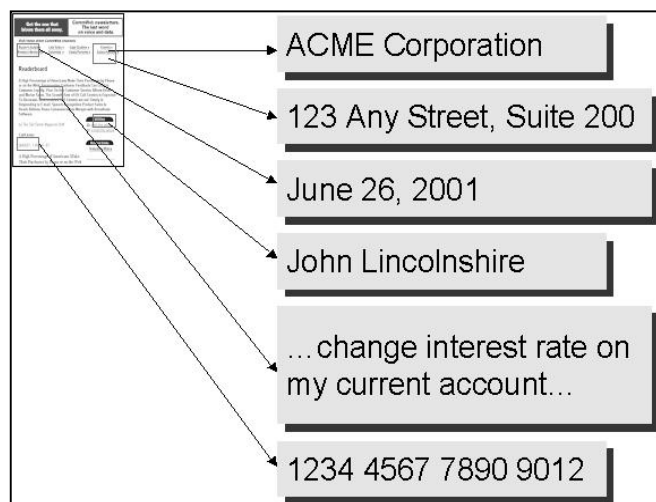


Figure 5: Content-related extraction of invoice dates by using a list of “candidates”

Depending on the application environment, additional routine checks can be applied to validate the information:

- Validation based on the document content only
This is restricted to just the document content by performing single-field and multi-field validations on the extracted data. These routines include a proof of checksums on specific ID-codes, plausibility checks among dates (ranges, sequence of dates) or a cross-calculation of amount values (totals, tax, ...).
- Validation based on the document content matched with external information
The document content is matched with additional information, which has been stored

by another application in a database, e.g. getting access to customer or client related information. It may also include access to all stored case-related information.

3.3.3 Address Analysis

Address analysis targets the specialised task of extracting and validating address information taken from documents by utilising public information sources as well as incorporating organisation-specific databases.

Due to a high degree of redundancy within address information, the extraction quality is always very high; also under poor conditions regarding document quality and incomplete captured information.

The screenshot shows a dialog box titled "Address Analysis" with a "General" tab. It is divided into three sections:

- Address Items and Min. Similarity:** Contains three rows of checkboxes and similarity percentage inputs. "Name" is unchecked with a 40% similarity. "Street + No" is checked with a 40% similarity. "ZIP Code + City" is checked with a 40% similarity. An "Overall Min. Similarity" field is set to 60%.
- Address Layout:** Contains four checkboxes: "Horizontal Alignment" (checked), "Vertical Alignment" (checked), "Inverse Street/City" (checked), and "Header Style" (unchecked).
- Included/Excluded Zones:** A table with the following data:

	Zone Name	Page Number	Included/Excluded
0	KUex1	1	Excluded
1	AdresseKU	1	Included

Figure 6: Parameter set for address analysis

3.3.4 Table Analysis

The table analysis has to deal with a large variety of various table layout techniques. Although a table basically always consists of a number of rows and columns with optional header and footer sections, the variety found under practical conditions is much higher.

It includes:

- hierarchies within tables, partial amounts, sub-totals
- split tables, comment inclusion
- single line items occupying more than one row
- tables without header / footer
- poor image quality and bad fonts.

Intelligent and content-based methods for table analysis do not rely on geometric information and graphic layout features anymore. They concentrate on the analysis of

content relations and based on this allow the correct extraction of information even if the internal table structure changes.

3.3.5 Zone analysis

The definition of “reading-zones” and the extraction of characters from these zones has become common practise since the introduction of automated forms processing. With the growing importance of dynamic or unstructured documents, the importance of zone reading is decreasing. Today’s zone reading techniques include the definition of flexible (relative) reference points as well as dynamic zones with varying dimensions.

3.4 Handling types of information and document representations

Organisations are currently challenged by the number and variety of electronic and paper-based documents. These can be letters, forms, fax, emails, data and information which comes from any software application and database or even from the Web/Internet. The more precise document contents can be analysed and categorized, the better the quality of the case specific treatment and the storage-indexing scheme will be.

The overall challenge is to successfully handle the variety of different sources and document representations. This requires transformation techniques, which generate a unified electronic representation of the textual content based on the different sources.

Different types of documents from different sources

OCR / ICR recognition techniques are used during document capturing and scanning processes and are directly applied to digital images. Documents are typically scanned and saved as TIF images or stored using similar formats. OCR / ICR engines and the collection of filter routines deliver pure text information which can be enhanced (only if available) by additional geometric information, mainly including the location of textual segments within a document by absolute / relative coordinates. This unified text result is taken as input for the categorization and extraction steps which follow as described in the previous section.

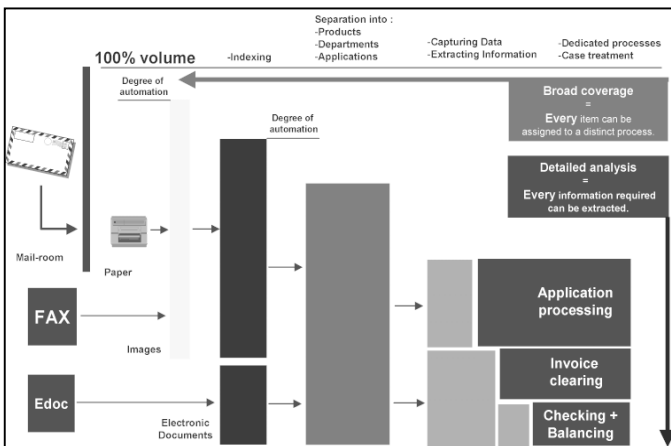


Figure 7: Multi-source document input

The relevance of OCR / ICR has changed during the past years due to changing boundary conditions. Some of today's paradigms are:

- better printer quality
- affordable ink-jet and laser printers are available
- the importance of facsimile is still growing.
- forms to be filled out by hand are restricted to their initial application cases, characterised by high volume, well-defined and established processes and a lower content complexity.
- forms based applications are moving more and more towards the Internet.

Recognition engines for typewritten characters (OCR) are common today and available in different variants. Handwritten character recognition (ICR) and constraint handwriting recognition technologies are more specific domains, this is a result of existing regional differences.

“Character recognition” as part of the document input process is supported by multiple recognition engines, mainly distinguished by characters (OCR / ICR), barcodes and optical marks (OMR).

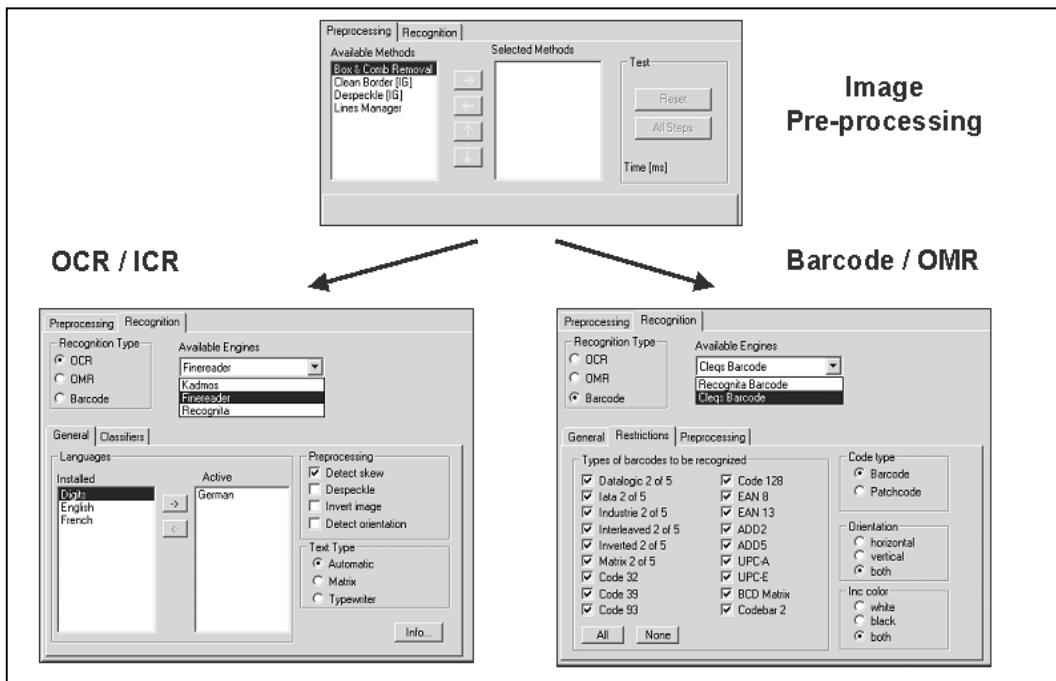


Figure 8: Set-up for multiple recognition engines for OCR / ICR / OMR and barcode

All these engines can be used in a very flexible manner and - to increase accuracy - the results delivered from each recognition engine can “compete” against each other on word level. In this case, the decision for a particular result is taken as ‘valid’ if all recognition engines participated have come to the same result.

Different document formats

Various document formats can be handled by electronic filters, which work on several hundred different input formats. Some of the most commonly used formats are HTML, PDF, DOC, XLS, PPT, TXT, etc. Such filters enable application viewers to display the content of most documents in the same manner and in one general style. Using these filters facilitates the quick access and reading of documents. It is not necessary to open five different applications to access five documents in five different formats. A viewer provides the content and then users can decide to open the application themselves.

Comprehensive content base

The overall target is to achieve a comprehensive information base which is able to contain the textual content of all relevant documents independent from their type, format, source, generator, owner and location. As soon as such an abstract of the overall content is available, the related metadata and the appropriate access methods can be applied.

Representation for storage

To collect and store all information including results delivered by categorization and extraction for building the storage index, document representations by using XML are preferred. A major benefit created with XML is the ability to manage dynamic sets of attributes – this also with regard to changing document content and structure. By using XML, the different sets of information / attributes - per document - can be easily managed, modified and updated over the complete life cycle of a document. These include:

textual content	: pure ASCII representation
geometric information	: for text segments and words
classification results	: including confidences
extraction results	: including confidences
document status	: during processing

4. The Role of Databases

This chapter will describe the different types of databases and search engines (Internet search engines, full text, relational) and their characteristics in regard to indexing and retrieval.

4.1 Database types and related indexing

Basically, a database can be described as a collection of information, which is organised for flexible searching and utilisation. There is a wide array of databases, from simple examples such as simple tabular collections to much more complex models such as relational database models.

The different types of databases are distinguished by many characteristics. One commonly used characteristic is the programming model associated with the database. Several models have been in wide use for some time.

4.1.1 Flat or table databases

The flat (or table) model is basically a two-dimensional array of data elements, where all members of a given column are assumed to be similar values, and all members of a row are assumed to be related to one another. For instance, columns for name and password might be used as a part of a system security database. Each row would have the dedicated password associated with a specific user. Columns of the table often have a type associated with them, defining them as character data, date or time information, integers or floating point numbers.

4.1.2 Networked databases

The network model flat model by adding multiple tables. A table column type can be defined as a reference to one or more entries of another table. Thus, the tables are related by references to each other, which can be seen as a network structure. A particular network model, the hierarchical model, limits the relationships to a tree structure instead of the more general directed referencing structure implied by the full network model.

4.1.3 Relational databases

Relational databases are similar to network databases, which can also contain several tables. Unlike network databases, however, the relations between tables are not explicitly encoded in the definition of the tables. Instead, the presence of columns whose name and type are identical implies the relationship between the tables. In such a relational database the data and relations between them are organised in tables. A table is a collection of records and each record in a table contains the same fields (formally known as attributes). The data type and name of the fields specify the record's format. Certain fields may be designated as keys, which means that searches for specific values of that field will use indexing to speed them up.

Where fields in two different tables take values from the same set, a join operation can be performed to select related records in the two tables by matching values in those fields. Often, but not always, the fields will have the same name in both tables. For example, an "orders" table might contain (customer ID, product code) pairs and a "products" table might contain (product code, price) pairs so to calculate a given customer's bill you would total the prices of all products ordered by that customer by joining on the product code fields of the two tables. This can be extended to joining multiple tables on multiple fields. Such relationships are only specified at their retrieval time.

A mathematical model implies a set of consistent operations that can be performed on such databases. As a result, relational databases can be flexibly reorganised and reused in ways their original designers did not foresee.

4.1.4 Object databases

In recent years, the object-oriented paradigm has been applied to databases as well, creating a new programming model known as object databases. Objects are very similar to table rows in the other models, and the columns of a table are very similar to the data definitions included as a part of an object class definition. In addition, the object database can include stored behaviours, which are triggered when an operation is performed on an object. But in this case, two sets of code have to be maintained, the application code and the database code.

Object databases and relational databases have meanwhile become very similar so that object systems are often implemented on top of a relational database foundation, so-called object-relational databases. Nevertheless, there is always a fundamental mismatch between the object and relational data models, so object-relational databases require object-relational mapping software which maps the relational model (row-oriented, ad-hoc querying) with the object-oriented model (polymorphic types, navigational traversal of objects). This mapping always introduces inefficiency but may be the only choice for many systems, which mandate a relational database back-end for e.g. legacy reasons.

An object database (more correctly referred to as ODBMS or OODBMS for object database management systems) is a DBMS that stores objects as opposed to rows/tuples in a RDBMS or relational database system. It is most often used in the case of C++ and Java programmers that do not wish to deal with the mismatch of going from an OO language to a database query language like SQL programming language required by RDBMS. Developers prefer to be able to persist an object without having to go through a paradigm shift. Also missing in RDBMS is the concept of polymorphism, which is central to OO design, thus causing headaches when mapping from OO code to an RDBMS as already mentioned above.

Of course this has advantages and disadvantages. The ability to stay with an OO paradigm enhances the productivity. However, the RDBMS model is a mature and proven one that has had decades of development and testing.

Certain benchmarks between ODBMS and RDBMS have shown that ODBMS can be clearly superior. One of the main reasons is that ODBMS do not use joins to associate objects but references, which are usually implemented as pointers. In the RDBMS model, a join would in most cases minimally require a search through a B-Tree index. In general,

navigation in an ODBMS is via traversing references, whereas in an RDBMS data is joined in an ad-hoc fashion (which is better for arbitrary queries).

The successful market segments for ODBMS seem to be in telecommunications, high-energy physics and subsets of financial services. The things that work against ODBMS seem to be the lack of interoperability with a great number of tools/features that are taken for granted in the RDBMS world including but not limited to industry standard connectivity, reporting tools, OLAP tools and backup and recovery standards. Additionally, ODBMS lacks a formal mathematical foundation, unlike the relational model, which rests upon the firm and well-understood mathematical basis of the relational calculus.

4.1.5 Database indexing

All the kinds of mono-/multi-table databases can take advantage of indexing to increase their speed. The index is here a sorted list of the contents of some particular table column, with pointers to the row associated with the value. An index is a data structure that helps to organise the data by mapping a key value onto one or more records containing the key value, thus providing a mechanism to efficiently identify the storage location of records. An index allows a set of table rows matching some criteria to be located quickly. Various methods of indexing are commonly used, e.g. balanced trees like B(+)-trees*, hashes** and linked lists are all common indexing techniques.

*) A B-tree is an ordered tree data structure in which each node has at most two children. Typically the child nodes are called left and right. One use of binary trees is as binary search trees where every node has a value, every node's left sub tree has values less than the node's value, and every right sub tree has values greater. A new node is added as a leaf.

***) The hash table refers to a data structure that implements an associative array. Like any associative array a hash table is used to store many key => value associations. A hash table maintains two arrays, one for keys, one for values or possibly one array of (key, value) pairs. When required to find the associated value for a given key, the key is fed through a hash function to yield an integer called the hash value. This integer is then the index to the associated value.

In an object-oriented database, classes are organised according to three types of relationship, namely generalisation, association and aggregation relationships. Balanced trees such as B+-trees are popular for indexing structure. The advent of the object-oriented database (OODB) has introduced new indexing issues due to the semantic richness of the object-oriented model. The three relationships mentioned above impact the query language according to different dimensions. Many researches focus on generalisation and association relationships when trying to find indexing structures for OODB. For the aggregation hierarchy, there is no deep research in this field.

In discussion are recent advances in indexing and access methods for particular database applications. These are issues such as external sorting, file structures for intervals, temporal access methods, spatial and spatio-temporal indexing, image and multimedia indexing, perfect external hashing methods, parallel access methods, concurrency issues in indexing and parallel external sorting. To discuss all these possible types of indexing here would go too much into technical details.

4.2 Indexing and Search methods

4.2.1 Full Text

When storing a document a set of keywords needs to be indicated reflecting the main facets of the content. The information retrieval method is the straightforward comparison of query strings and the documents to be searched. These methods only look for exact matches. Efficient algorithms have already been proposed in the seventies and incorporated into commercial products and solutions during the past fifteen years.

Agrep for example is one of the ancestors of these search tools, which even tolerates typing errors. But most of these techniques strictly insist on correct spelling. The main advantages of full text scanning methods are that they require no overhead space and minimal effort on insertions and updates.

The disadvantage is the bad response time due to the fact that the search string has to be sequentially compared with every single string in the document collection being searched. But another profound disadvantage is the dependency on the spelling of the keywords you are looking for. A very evident example is the word “standardisation”. If you spell it like in the preceding sentence, you will receive a result which refers to documents written in British English. If you spell it with a “z” (“standardization”) then you will get the results written in American English. And of course, this could be an important difference.

4.2.2 Full text retrieval with Boolean operators

The method described above can be improved by adding Boolean functions which link the given keywords with each other in a certain way: AND, OR, NOT, MUST, etc. If users have experience using such mathematical operators then perhaps they will be able to obtain a result list, which matches their expectations. As soon as an entire phrase is used as the matching example, this retrieval method becomes complicated (many operators) or does not work anymore (max. number of keyword is e.g. 10 words).

The most frustrating disadvantage of the full text retrieval is that you get a result list of sometimes several hundreds or thousands of document hits but you cannot be sure that the best matching one is at the top of the list. And you are certainly not able or willing to read them all.

However, this full text method is just not able to work any better because the relation between one or even twenty given keywords with the entire contents of thousands of documents is not comparable enough. If a query is not set up with the matching keywords then the documents possibly matching the required context will not be found.

4.2.3 Signature Files

Using this method each document gets a “signature”. A signature is a bit string that indicates whether or not certain terms occur in a document. As a consequence, it is necessary to pre-process all search documents in order to set up the index. A separate index file is generated and is called the signature file. These signature files are of course much smaller than the original files and can therefore be searched much faster.

The disadvantage is again that a set of certain terms cannot reflect the whole content, and the individual context neither.

4.2.4 Inversion

Each document can be described with a list of keywords, which describe the content of the document for retrieval purposes. The terms allow text-based searching via string-matching. An alphabetically sorted list of all keywords in an index file with pointers to the documents that include the keywords allows fast retrieval as long as the number of documents does not exceed the number a human is able to process. The usage of such inverted index files in combination with full text databases usually requires a large memory capacity because the vocabulary and search structure is kept online for performance reasons.

Added acceleration can be achieved by using more sophisticated methods to organise the index files, e.g., B-trees, hashing or combinations of these. Advantages of this method are the easy implementation, speed, and an easy support of synonyms. The disadvantages are costly update and reorganisation of the index and the overhead storage.

4.2.5 Vector Model and Clustering

According to the cluster hypothesis, closely associated documents tend to be relevant to the same queries. As a consequence, clustering documents accelerates searching because only the identified cluster has to be searched. Documents are closely associated if they share the same terms (co-occurring terms). Document clustering methods involve the procedures of cluster generation and cluster search. For the cluster generation each document is represented as a t -dimensional vector.

The vector space is defined by keywords, which can be chosen automatically or manually. Automatic indexing procedures might be applied which use a stop word list to remove common words, a prefix and a suffix list for stemming and a thesaurus to assign each word-stem to a concept class. In this way, each document is represented by a t -dimensional vector, where t is the number of index terms (concepts). The presence of a term is indicated by a 1 and the absence by a 0. It is also possible to weight the presence of some terms (term weighting). Term weights are frequently used to reflect the importance of a term for a document. A possible weighting function could be the occurrence frequency of a term in a document.

The next step is to cluster documents into groups using either methods based on the document-document similarity matrix or iterative methods that are more efficient and proceed directly from the document vectors. If the similarity measure between two documents exceeds a predefined threshold then the documents are connected with an edge. The connected components of the resulting graph are the proposed clusters. Cluster generating methods generally require at least one empirically set constant: a threshold in the similarity measure or a desirable number of clusters. This constant greatly affects the final partitioning and therefore imposes a structure on the given data, instead of detecting any existing structure.

A simple and fast method is the single pass method. Each document is processed once and is either assigned to one (or more, if overlap is allowed) of the existing clusters, or it

creates a new cluster. To search a cluster is much easier than cluster generation. The input query is represented as a t -dimensional vector and it is compared with the cluster-centroids. The searching process takes place in the most similar clusters, i.e. those whose similarity with the query vector exceeds a threshold. A cluster-to-query similarity function has to be selected.

The vector representation of queries and documents allows relevance feedback, which increases the effectiveness of the search. The user pinpoints the relevant documents among the retrieved ones and the system reformulates the query vector and starts searching from the beginning. The usual way to carry out the query reformulation is by adding (vector addition) the vectors of the relevant documents to the query vector and by subtracting the non-relevant ones.

4.3 Indexing and retrieval methods using natural languages

The database type is not the only indicator of how information and knowledge can be organised today in order to keep it accessible for any individual and not only for DB experts. And the keyword based search methods described above and other extracted document data do not satisfy the human desire at best.

During the past few years, new methods have been more deeply explored. Meanwhile a status quo of marketable technologies based on a completely new natural language approach has been achieved.

There are two camps, which have developed their technologies based on natural language, both with fundamentally differing approaches: The linguistic and the statistics method.

4.3.1 Linguistic approach

This method rests on diverse language characteristics and their field of application. Semantic networks are set up in which a language's vocabulary (dictionary) and the typical characteristics of the language are retained within a sentence structure, e.g. subject, verb and object. The prevalent relationships of the sentence elements among one another are also classified. Such a formation can also be found in the Corpus Linguistics, such as the British National Corpus, which was created by several representative organisations. In order to get a feel for the scope of such a body of work, it is helpful to take a closer look at the parameters.

British National Corpus (BNC)

The British National Corpus is a very large (over 100 million words) corpus of modern English, both spoken and written. The BNC project was carried out and is managed by an industrial/academic consortium lead by publishers and academic research centres, and the British. The corpus was completed in 1994.

The Corpus is designed to represent as wide a range of modern British English as possible. The written part (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part (10%) includes a large amount of unscripted informal conversation, recorded by volunteers selected from

different age, region and social classes in a demographically balanced way, together with spoken language collected in all kinds of different contexts, ranging from formal business or government meetings to radio shows and phone-ins.

This generality, and the use of internationally agreed standards for its encoding, encourage us to believe that the Corpus will be useful for a very wide variety of research purposes, in fields as distinct as lexicography, artificial intelligence, speech recognition and synthesis, literary studies, and all varieties of linguistics.

The corpus comprises 100,106,008 words, and occupies about 1.5 gigabytes of disk space. To put these numbers into perspective, the average paperback book has about 250 pages per centimetre of thickness; assuming 400 words a page, we calculate that the whole corpus printed in small type on thin paper would take up about ten metres of shelf space. Reading the whole corpus aloud at a fairly rapid 150 words a minute, eight hours a day, 365 days a year, would take just over four years.

The corpus comprises 4,124 texts, of which 863 are transcribed from spoken conversations or monologues. Each text is segmented into orthographic sentence units, within which each word is automatically assigned a word class (part of speech) code. There are six and a quarter million sentence units in the whole corpus. Segmentation and word-classification was carried out automatically by the CLAWS stochastic part-of-speech tagger developed at the University of Lancaster. The classification scheme used for the corpus distinguishes some 65 parts of speech.

Every one of the 100 million words in the BNC carries a grammatical tag, that is, a label indicating its part of speech. In addition, each text is divided into sentence-like segments. This process was carried out at Lancaster University's Unit for Computer Research on the English Language (UCREL), using the CLAWS4 automatic tagger. The basic tag set distinguishes 61 categories found in most "traditional" grammars, such as adjectives, articles, adverbs, conjunctions, determiners, nouns, verbs etc. Tags are also attached to major punctuation marks, indicating their function.

The corpus is encoded according to the Guidelines of the Text Encoding Initiative (TEI), using ISO standard 8879 (SGML: Standard Generalized Markup Language) to represent both the output from CLAWS and a variety of other structural properties of texts (e.g. headings, paragraphs, lists etc.). Full classification, contextual and bibliographic information is also included with each text in the form of a TEI-conformant header.

The costs, the necessary resources and the time required for the completion of a single corpus are obviously considerable. Doubts regarding the feasibility and more importantly the availability arise when one considers that such a corpus would have to be completed for all languages – and consider for moment the number of languages spoken in Europe.

Semantic networks not only include the pure country specific aspects of language. Domain specific ontology is also set up which means the specialised vocabulary of e.g. the pharmaceuticals industry or legislation is also included.

Another linguistic building block is the thesaurus. The synonyms for individual words and their meanings are collected in such comparison tables. This is an immense help when searching with keywords because the search can be expanded in this manner. For example, the thesaurus entry for the word "legislation" could be law, code, order, ordinance, statute etc.

The linguistic approach offers many possibilities for collecting the content and context of language-based documents and also use as a search method. A simple semantic table in which an automobile is equivalent to car, for example, is already a great help. If a user queries an information retrieval system with a keyword, such as “intelligence”, the system should at least provide a list of documents in which the term “intelligence” occurs. More sophisticated systems would also provide documents, which contain terms like “cleverness”, “intellect”, “brains”, or even documents which talk about “intelligence” without using the term. Currently however, there are still some disadvantages to the linguistic approach, which have to be considered.

Languages are alive and continue to develop. This means that new words are constantly being created while at the same time other words fade into disuse. Some words are just short-lived buzzwords, but if they are not added to a corpus or thesaurus, then they won't be taken into account and lose their relevance in keyword searches. The same thing applies to e.g. a ten-year-old keyword index. Such an index would only be of partial use today.

A specific semantic network and the associated aids have to be set up for each language. Something, which has been created for the English language for example, cannot be reused for the French language.

The various industries (government, manufacturing, retail, finance etc.) all have their own specific vocabulary, sentence structure and ways of expressing something. In the literary and creative world, a situation is often described using analogies and metaphors. As such, the facts are not concretely mentioned, but implied. This of course inhibits determining the relationship to other documents. In comparison, an informational type language is used in the finance industry for relaying facts concerning amounts contained in a contract.

The linguistic approach is basically on the right path to being able to deliver contextually compatible document contents, but there are still a number of obstacles. The required data maintenance is considerable and the such systems are currently still relatively slow.

4.3.2 Statistical approach

The statistical approach to recognising the context and contents of documents first unleashed linguist scepticism. As mentioned in the previous section, linguistics is not as complex as it is for no reason. The statistical method was developed by linguists who weren't concerned with words, word stems, sentence structure, idioms and thesauri. They pragmatically broke down texts into their smallest parts so as to be able to recognise the textual contents via statistical calculations and multi-dimensional treatment. What was first met with doubt by experts lead to practical experiments and surprising results. This laid the cornerstone for further refining this method, which since then has reached a quality that is highly regarded for its effectiveness and speed.

In the following, a method is introduced which allows the usage of complete text passages as opposed to just keywords and synonyms. This “direct access” method does not even require a literal match between query and a search document. It uses pattern-matching algorithms to map similar information of the repository to the query. The representation of such a repository can also be seen as an enhancement to build a real “content-based index”.

Direct Access / Associative Access

Direct, associative access is based upon the principle of content and context matching of a sentence, a paragraph or a document, which returns suitable matches more reliably than any other form of keyword and current index-related search methods. For any document chosen by the user as an example text, the associative search detects the best matching documents similar to the given example and ranks them according to their relevance.

The main idea is that a paragraph-sized text provides more information than just keywords. This helps to determine not only a possible topic but also the correct topic since keywords may occur in different contexts.

Context based information thus helps to resolve the traditional problems associated with keyword searches. Resolving ambiguous meanings of a keyword is particularly important if the meanings differ from each other (e.g. for the word “bank” as the “edge of a river” or an “establishment for depositing money”).

The associative access does not create an ordinary index. Instead the statistically relevant patterns of the documents are captured, e.g. combinations of letters, words, fragments of sentences. Every single document in the knowledge base (representation of the document content collection being accessed by associations) is then represented in a metadata file, which is a coded definition for the existence or non-existence of certain patterns contained within all documents stored in the knowledge base. The internal representation of the knowledge base is created from the textual information of the documents in a single pass.

If a user performs a query using conversational speech, this query gives an example of what is to be associated and is mapped onto the patterns of the knowledge base using exactly the same process.

That is, either the pattern occurs or does not occur. Documents with no matching patterns are ignored and the remaining documents are compared with the query example and ranked by their compliancy in a result list.

Associative Access compensates for many of the disadvantages, which the previously described methods have. It simply uses a piece of natural language text as an example and compares it with the knowledge base in order to locate documents with matching content and context and presents these documents in a ranked result list (see figure 9).

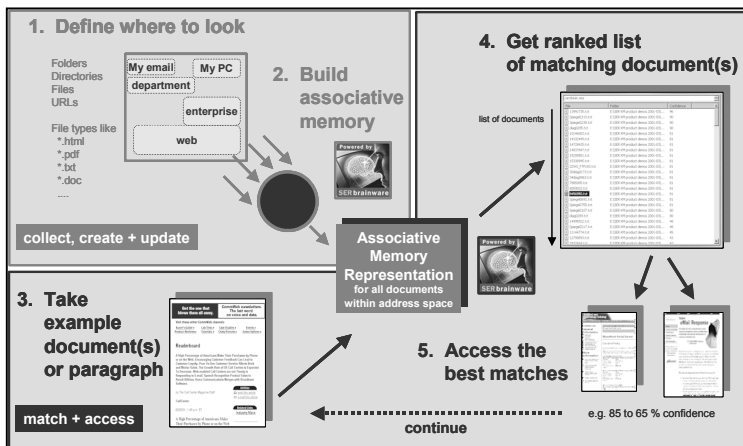


Figure 9: Direct Access by Associations

As compared to the other methods presented, the advantages are very evident:

- Automatic indexing of the complete content in a single pass. All manual intervention becomes obsolete which avoids insufficiency, errors as well as spending resources, time and costs.
- Automatic update of the document relationships. This avoids the manual update process of linking “old” documents with new documents, something which is not normally done in reality. So even in a computer system, old documents are usually forgotten if people no longer remember them. This leads to “forgetful” organisations which is unfortunate given that such document content may be very useful.
- Natural language queries which reflect a much more comprehensive expectation of what the user really wants to know. Such queries can be performed by anyone.
- No artificial, technical and user-unfriendly query languages, e.g. no Boolean operations, no SQL statements which can be only applied after having learned them, i.e. by experts.
- Language independence which saves a lot of resources which would otherwise have to be spent to set up semantic networks with language-specific corpus linguistics, thesauri, etc. Even if several languages are in use, the required maintenance efforts remain at the same level.
- Fault tolerant so that even users who do not have a good command of a foreign language or who are perhaps dyslexic can benefit from such associative accessing in which correct spelling does not mean success and failure.
- Document format independence which allows the use of any software applications. Old as well as quite new formats are supported.

Combination of indexed and associative search

The direct content access described above is being improved for an environment which scales up to several Terabytes of documents as well as for an environment where it is necessary to differentiate by certain expressions, by names or by the dedicated subject of the required content. This approach maps an indexed search with the associative access. An example which points out why it is worthwhile to combine these two is: If a lawyer has to make a decision regarding a given case then he will want to be sure to access just the case-related documents and not any others. Even if another case is quite similar to the one in question - he cannot allow himself to mix two different cases and risk making a decision based on the wrong information. In such a case, the lawyer needs to indicate the name of the client, the case ID or something similarly unique.

5. Standards for Indexing

5.1 Relevant standards for indexing and ordering methods

5.1.1 Metadata structuring through XML - eXtensible Markup Language

The Extensible Markup Language (XML) is the universal format for structured documents and data on the Web defined by the World Wide Web Consortium (W3C). XML is becoming an increasingly popular method of labelling and tagging digital material. Structured data includes things like spreadsheets, address books, configuration parameters, financial transactions, and technical drawings. XML is a set of rules (guidelines or conventions) for designing text formats that let you structure your data. The advantage of a text format is that it allows people view data without having to see program that produced it. You can read a text format with your favourite text editor.

Like HTML, XML makes use of tags (words bracketed by '<' and '>') and attributes (e.g. name="value"). While HTML specifies what each tag and attribute means, and often how the text between them will look in a browser, XML just uses tags to delineate pieces of data, and leaves the interpretation of the data completely to the application that reads it. In other words, if you see "<p>" in an XML file, do not assume it is a paragraph like in HTML. Depending on the context, it may be a price, a parameter, a name, etc.

XML provides a standard set of descriptions for labelling digital information and is designed to automate the rapid identification of material and the seamless exchange of data between computers. As such, XML is designed to overcome the constraints of HTML as a single, inflexible document type and avoid the complexity of full SGML. By facilitating the mapping, transformation and routing of information based on a common set of "document type definitions" (DTD), the most common perceived application of XML is to provide metadata structuring for unstructured content, as well as the exchange of data between different applications and operating systems.

XML is likely to become prominent feature in the future development of online information sources. However, like all kinds of tagging schema, it suffers from a number of limitations. There are significant barriers to ensuring that XML decreases the costs and increases the efficiency of managing information. Insufficient awareness of such barriers and lack of understanding of how to automate the otherwise burdensome administrative processes upon which XML depends, can lead to high labour costs and descriptive inconsistency.

Manual process and specificity

The first limitation is the manual process of defining and correctly entering the tags. The well-known example of the effect of human behaviour and the inherent limitations of manually describing information can be illustrated by what happens e.g. when Intranet users are responsible for describing the contents of the documents they have written. It is a common and pragmatic procedure but a majority of documents are too often insufficiently/incompletely described and tagged as "general".

Whilst XML attempts to break away from such generalist terms, it remains dependant upon the same shortcomings of human behaviour that manifest themselves as

inconsistency. An individual's ability to describe information is dependant upon their personal experience, knowledge and opinions. Such intangibles vary from person to person and are also dependant upon circumstance dramatically reducing the effectiveness of the results.

Further imponderability arises when documents incorporate multiple themes. Should a document about "infrastructure development in Afghanistan as part of the new domestic policy" be classified as (1) Afghanistan infrastructure (2) Afghanistan domestic policy, or (3) Afghanistan reconstruction and development? The individual view onto such a question is different and the decision process is both complex and time consuming and introduces yet more inconsistency, particularly when the sheer number of hundreds of options available to a user is considered.

As soon as you want to be very specific in the retrieval and processing of XML based documents, you need to have a very high number of tags describing the content. For example, tag numbers in a company like the press agency Reuters run into the tens of thousands. It is easy to imagine that the effort and the likelihood of misclassification increase with the number of tags.

Interoperability

XML does not provide a set of standard tag definitions but instead is a set of definitions that guides you through defining tags, i.e. if two organisations intend to cooperate they have to explicitly agree on understandings and same meanings of the tags in question and they have to define them in advance. Here we can refer to that what we have learned from the database world: The technological basics are clearly defined but as soon as several parties want to cooperate, the first thing they realise is that they have their own understanding, structures and naming and are not able or willing to change and adapt quickly. For some industries, close cooperation has become so vital that they have not had any other alternative than to remain consistent. Just-in-time production and delivery are crucial points in manufacturing and transportation.

5.1.2 Metadata

Whatever the format, metadata is used for indexing the content of e.g. a document or a web page. This metadata is the attributes belonging to the content. They can be meaningfully used in document pre-processing. For instance, you can select all English documents identified by the attribute `<META NAME="Content-Language" CONTENT="en">` before an associated access is performed. This reduces the number of documents to be mapped with the query text.

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html><head><title>SERintranet</title>
<META NAME="description" CONTENT="SERintranet">
<META NAME="author" CONTENT="SER Solutions Inc">
<META NAME="publisher" CONTENT="SER Solutions Inc">
<META NAME="copyright" CONTENT="SER Solutions Inc">
<META NAME="revisit-after" CONTENT="30 days">
<META name="generator" CONTENT="SER HTMLMAKER">
<META NAME="Content-Language" CONTENT="en">
```

Figure 10: Example of a set of metadata of an HTML web page. Important is the DTD reference.

Such metadata exists in XML and in HTML as well. If you want to use such attributes you have to first check which document type definitions (W3C//DTD) are referred to. Otherwise you would mix attributes, which for instance relate to “author” instead of “company”. This would lead to a useless or at least incomplete result.

5.1.3 Standard Generalised Markup Language (SGML)

SGML (Standard Generalised Markup Language) is an international standard for the definition of device-independent, system-independent methods of representing texts in electronic form. It emphasises descriptive rather than procedural mark-up; in other words, how a document is structured logically rather than how it should be displayed. It is used in association with Document Type Definitions (DTD), which allow the user to specify what type of document is being described.

5.1.4 Machine readable cataloguing for libraries (MARC)

MARC is an example of an index, which has been especially created for the library environment but as this segment has discovered the Internet as well, it has similar demands for indexing and accessing documents and related content in a given context.

Machine Readable Cataloguing (MARC) is an international standard for computer-readable bibliographic records. MARC records contain a description derived from AACR2R rules (for North American, British and Australian libraries), "access points" to the record (author(s), title), subject headings assigned from a controlled vocabulary such as the Library of Congress Subject Headings, and a "call number" to group together items in like subject areas. There are many MARC record variations, often country-specific. In the United States, the Library of Congress (LC) MARC record version has been used since the Library of Congress (the depository for U.S. publications) has produced standardised cataloguing records for libraries. Copying and modifying existing MARC records or buying the records constitutes a major source of cataloguing records for many North American libraries.

With the proliferation of Internet resources, especially those on the WWW, librarians have become concerned about providing access to resources not physically located in their libraries but used by their patrons. If a library does not subscribe to a particular periodical or a government publication and these are available through the Web, how should a library make them readily accessible to someone who wants to use them? What obligation, if any, does a library have to catalogue these resources? These questions plus the lack of standardised access to Internet resources has prompted a discussion of cataloguing for the Internet. Most North American librarians feel that cataloguing must be done but the "how" and is an ongoing debate. The hope is that some standardised form of header, whether the TEI header, the Dublin Core or another standard will be widely adopted to provide descriptive and subject information about an item which can then be used to create a MARC record manually or by mapping to the MARC format.

The risks, barriers and disadvantages of error-prone manual work and poor reflection of content and context by only using an index has been discussed in this document several times. Nevertheless, the clear demand for a broader standardisation of indexing and accessing methods also comes up here.

5.1.5 SQL programming language

SQL stands for Structured Query Language and as the name implies, SQL is designed for a specific, limited purpose - querying data contained in a relational database. As such, it is not a real programming language in the pure sense, such as e.g. C++, which is designed to solve a much broader set of problems.

SQL was originally created by IBM but many vendors directly developed dialects of it. It was then adopted as a standard by ANSI in 1986 and ISO in 1987. SQL was revised in 1992 as the known version SQL2. A new revision, SQL3, was released in 1998. SQL3 supports objects, which were not previously supported in other versions, but as of late 2001, few if any database management systems implement SQL3. SQL, although defined by both ANSI and ISO, has many variations and extensions, most of which are of a proprietary nature, such as Oracle Corporation's PL/SQL or Sybase's Transact SQL. Language extensions such as PL/SQL are designed to address this by increasing the scope of SQL while allowing the user to maintain SQL's advantages.

Regarding ergonomics and usability, SQL statements and multiple nested queries can now be generated by using user-friendly interfaces in which users can just point, click and link the items they want to connect with each other and which helps them to retain an overview. SQL statements, which are used inside user-oriented applications are hidden for the users and run in the background. Nevertheless, you need to be educated and experienced in this kind of technological logic.

5.1.6 Object-oriented query languages and views

The major issue arising during the development of object-oriented and object-relational databases concerns how to adapt SQL, the source of success of relational databases, to novel concepts, such as complex objects, relationships, nested collections, classes, inheritance, encapsulation and polymorphism. Two industrial standardisation bodies, ODMG and ANSI/ISO, have proposed competitive languages OQL and SQL3 (recently SQL2000) based on very different conceptual and semantic paradigms. Both OQL and SQL3 have received criticism: OQL for limited scope and imprecise description, SQL3 for eclectic design and too large of specification. On the other hand, the research community is seeking a formal model for object data structures and query languages. No query language can survive without automatic optimisation of queries, which requires simplicity, generality and a deep understanding of formal semantics.

Object-oriented concepts, new versions of OQL and SQL3 and new applications of query languages have shown that the idea of query languages must be relaxed from stereotypes having roots in formal paradigms of the relational model. There are discussions ongoing about a new approach to object query languages, having roots in semantics of programming languages. It employs the classical naming-scoping-binding paradigm and an environment stack for the definition of query operators. A query is a generalised programming expression, which can be used for various purposes: For retrieval a la OQL, for imperative statements a la SQL3, for the definition of views and procedures, etc. Example ideas are implemented in the Loqis prototype. (Author: Kazimierz Subieta, University of Hamburg Department of Computer Science)

5.2 Relevant standardisation bodies and initiatives

The main and relevant standardisation bodies for the Web / Internet environment are, in addition to the ones that have been mentioned, ISO (International Organisation for Standardisation) and ANSI (American National Standards Institute):

5.2.1 W3C – The world wide web consortium

The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential as a forum for information, commerce, communication, and collective understanding.

W3C defines the Web as the universe of network-accessible information (available through your computer, phone, television, or networked refrigerator...). Today this universe benefits society by enabling new forms of human communication and opportunities to share knowledge. One of W3C's primary goals is to make these benefits available to all people, whatever their hardware, software, network infrastructure, native language, culture, geographical location, or physical or mental ability. W3C's Internationalisation Activity, Device Independence Activity, Voice Browser Activity, and Web Accessibility Initiative all illustrate our commitment to universal access.

People currently share their knowledge on the Web in language intended for other people. On the Semantic Web ("semantic" here means "having to do with meaning"), we will be able to express ourselves in terms that our computers can interpret and exchange. By doing so, we will enable them to solve problems that we find tedious, to help us quickly find what we're looking for - medical information, a movie review, a book purchase order, etc. The W3C languages RDF, XML, XML Schema, and XML signatures are the building blocks of the Semantic Web.

The other subjects W3C focuses on is trust, interoperability, evolvability (with the principles of simplicity, modularity, compatibility, and extensibility guide all of designs), decentralisation, and last but not least cooler multimedia.

5.2.2 OASIS

OASIS is an international, non-profit consortium that designs and develops industry standard specifications for interoperability based on XML and SGML.

OASIS mission statement: The HumanMarkup TC is set forth to develop the HumanML and associated specifications. HumanML is designed to represent human characteristics through XML. The aim is to enhance the fidelity of human communication. HumanML is set forth to be an XML Schema and RDF Schema specification, containing sets of modules which frame and embed contextual human characteristics including physical, cultural, social, kinesic, psychological, and intentional features within conveyed information. Other efforts within the scope of the HumanMarkup TC include messaging, style, alternate schemas, constraint mechanisms, object models, and repository systems, which will address the overall concerns of both representing and amalgamating human information within data. Examples of human characteristics include emotions, physical descriptors, proxemics, kinesics, haptics, intentions, and attitude. Applications of

HumanML include agents of various types, AI systems, virtual reality, psychotherapy, online negotiations, facilitations, dialogue, and conflict resolution systems.

5.2.3 Dublin Core Metadata Initiative (DCMI)

The Dublin Core Metadata Initiative is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models as well as the development of specialised metadata vocabularies for describing resources that enable more intelligent information discovery systems. DCMI's activities include consensus-driven working groups, global workshops, conferences, standards liaison, and educational efforts to promote widespread acceptance of metadata standards and practices.

Conceived at an OCLC-sponsored workshop in 1995, DCMI has a growing global reach of about 45 nations around the world. The Dublin Core metadata element set has been translated into 25 languages, and has been formally adopted by 7 governments (with other governments currently in discussion phase).

The mission of the DCMI is to make it easier to find resources using the Internet through the following activities:

- Developing metadata standards for discovery across domains,
- Defining frameworks for the interoperation of metadata sets,
- Facilitating the development of community or disciplinary-specific metadata sets that are consistent with the first two items.

The participation base is global, and the management of DCMI is structured to reflect this. The DCMI Directorate is hosted at OCLC in the U.S. Stuart Weibel, Director of DCMI is employed by OCLC and works in Dublin. The Managing Director of the Initiative, Makx Dekkers, lives and works in Luxembourg. Other active members of the executive team live in the US, the UK, and Germany, and working group chairs include participants from the US, the UK, Sweden, Denmark, Germany, Portugal, and Australia.

One of the primary benefits of the Dublin Core is its usefulness as a cross-domain discovery tool. The general applicability of DCMI metadata makes it the strongest candidate to bridge disciplines and sectors to provide users with a common discovery model that will work throughout the Internet Commons. Active development of domain specific application profiles of Dublin Core metadata are ongoing in the library, museum, government, environmental science, publishing, agriculture, and corporate knowledge management domains.

DCMI's focus is the development of semantic metadata standards to support resource discovery. DCMI closely cooperate with application developers, including the W3C, and the RDF and XML developer communities.

6. Best Practice Applications

6.1 Automated distribution of incoming documents Project of the Statistical Office of the Free State of Saxony

The Free State of Saxony Statistics Office (Statistisches Landesamt des Freistaates Sachsen) is a sub-organisation of the Ministry of Interior of the German state of Saxony. The ministry is divided in departments for statistics, central administration and information technology. About 450 employees in 21 sub-departments located in 5 buildings create the official statistics for all Saxon districts that the state of Saxony has to provide. A modern data centre provides mainframe, server and PC and high-speed network support for all ministry tasks.

6.1.1 Description of the problem

The Statistics Office receives a vast number of documents every day. Documents arrive via mail, fax and email. About 2,000 documents arrive daily and about 80% of them are filled-in questionnaires covering various topics. The questionnaires have to be directed to the right department, the data has to be extracted and included in various statistical reports. 10% of the incoming documents were registered locally in small department-created solutions based on MS Access. A majority of the organisation's incoming documents first had to be transported to the right department. Documents like letters and faxes had to be transported physically to their destination, email messages were sent via email server.

This situation had a number of drawbacks that were to be solved by a modern solution:

- Document classification, registration and routing resulted in a labour-intensive, time consuming and error prone process.
- As there were a number of separate document databases, there was no possibility for a global search and retrieval mechanism.
- For the above reasons, the location and proper handling of misrouted documents generated significant overhead and delays.
- Maintaining several separate database solutions put an extra administrative burden on the departments, whose task is to deliver statistics.

This led to a number of requirements.

- All incoming documents - mail, fax, email - were to be handled in a coherent and transparent way.
- Classification of all documents was to be completed automatically.
- Logging of all incoming and outgoing documents in a database to provide reproducible results.
- Case relevant information was to be extracted automatically and attached to the documents as structured information in order to allow further decision support.

- A central document repository should enable users to store, search and retrieve documents as well as archiving.
- Reduction of overall time required processing a document.
- Provide a mechanism to generate document IDs that are consistent with the administrations' guidelines for outgoing documents.

6.1.2 Technical and organisational solution

In order to create a solution, which addressed the problems mentioned, a concept was created that covered both the organisational and technical structures that had to be adapted or created. The technical solution was built upon standard software components from the DOMEA® product suite. DOMEA® products follow the standards defined by the central IT Coordination and Consulting Agency of the German Government (KBSt, Koordinierungs und Beratungsstelle des Bundes). They have been successfully implemented in many government projects. DOMEA® also includes SERbrainware knowledge-enabling technology, which has proved to be crucial in this project.

- **Homogenous flow of information**
One central department receives and handles all incoming mail. There, all document capturing activities such as scanning, OCR reading, classification and distribution of incoming documents is completed. Every paper document that arrives at the Statistics Office, arrives at this department. With the exception of a few document types (newspapers, confidential mail, bids etc.) all documents are scanned and categorized here.
- **Fax and Email**
Email that is sent the central Statistics Office email address is handled in the office for incoming mail. Email that is sent to an individual address in the organisation is directly forwarded. A fax server handles all incoming faxes, which are electronically processed only.
- **Scanning and OCR**
All documents that have to be scanned are fed into a batch scanner, the document envelope is used as a separation page. Thus, all information which is provided by the sender of the document is preserved. Text from every document scanned is extracted by OCR and stored with the document for further processing. This is done using the application DOMEA® SERcapture.
- **Categorization and Knowledge Base**
Using DOMEA® SERdistiller and the underlying SERbrainware engine, the content of every scanned document is analysed, based on the OCR text. Each document is automatically associated with a document class (classification). The class information is necessary to decide, where a document has to be routed to and where it can be found in the document repository.
- **Information extraction**
In addition to document classification, DOMEA® SERdistiller reads the data located in a defined area of a document and extracts the values for further processing. Note that this feature is not restricted to a specific document format. Again, DOMEA®

SERdistiller can be trained to recognise specific sections of a document and extract data from these sections.

- Document repository

All scanned documents are put into a central document repository. Documents that have to be kept in paper format due to legal requirements are kept in the organisation for a certain time and are then transferred into a central archive. All other scanned paper documents are destroyed after scanning. The document repository is based on SERprocess Server. The repository provides a robust and safe space for all documents. It also provides advanced search and retrieval features.
- Document Routing

An SERprocess Server installation does not only maintain all documents but also controls the flow of documents to the various departments. Routing decisions are made automatically, based on the document category or other content. In some cases, manual routing decisions are made.
- User interface

End users can access the document repository and the processes using DOMEA® Outlook. DOMEA® Outlook provides features such as

 - Work list and routing
 - Document search and retrieval
 - Viewing and editing of documents.

DOMEA® Outlook is integrated into Microsoft Outlook, so Outlook users become familiar with the product very quickly.

Overall System Architecture

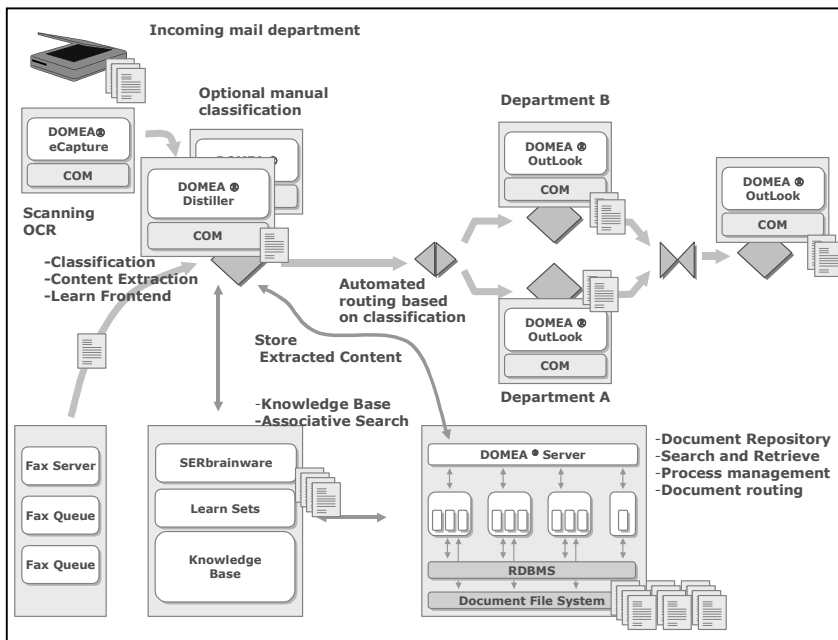


Figure 11: Overall System Architecture

6.1.3 Functionality

Incoming Post

When an incoming document arrives, an employee decides if the document will be handled electronically, depending on set organisational rules. If the document is to be handled electronically, it will be put in a scan batch and scanned. Scanning is done in batches and the image capturing software allows corrections to be made should this be necessary. After scanning, an OCR process extracts the text content. Scanned images and text are stored temporarily.

After scanning, DOMEA® SERdistiller starts automatically. DOMEA® SERdistiller identifies the document's category, which is connected to the target department to which the document is routed. DOMEA® SERdistiller extracts some basic information from the document, such as the return address creation date, document ID, subject and recipient.

Depending on the identified category, DOMEA® SERdistiller stores the document in the document repository and attaches document image, OCR extracted text and the extracted basic information for that document. Internally, document plus attachments are tied together by an SERprocess process instance. A workflow process is initiated that routes the document to the desired department for processing.

If a document is not classifiable with a reasonable confidence level (not classified), the system hands over control to the user for a manual decision..

Document Routing and Workflow

The SERprocess server handles the workflow process. The folders containing images and OCR text are placed in the work lists of individual users or work groups. A user can pick up any item in his work list and process it. Some routing steps might include process decisions, which are again based on the information extracted by DOMEA® SERdistiller.

User Interaction

Users in the departments see the routed documents in the in-boxes of DOMEA® Outlook. Users can now open and view the incoming documents and even create new documents using popular products such as MS Word or MS Excel from within MS Outlook. Newly created documents are also stored in the document repository maintained by SERprocess Server. Document access is regulated by an elaborated access rights system that allows to separate departments from each other, thus allowing maintaining one central repository for multiple departments.

6.1.4 Implementation

The implementation of the solution consists of standard DOMEA® and SERprocess components. In order to meet the projects goals, some customisation tasks had to be done.

Building the knowledge base

Based on organisational rules, every department defines a number of document classes, each fitting in an overall classification scheme called the Aktenplan. Initially, each department selects and provides a small number of documents (up to 20) for each class (about 30 classes) which makeup the learn set. These classes are then defined in

DOMEA® SERdistiller. DOMEA® SERdistiller is trained to recognise the content of the documents for each class. By just feeding the documents belonging to one learn set for one specific category into DOMEA® SERdistiller, the underlying SERbrainware engine builds the knowledge base for just that class. This process is repeated for every class.

Extraction rules

For proper extraction of document data, some rules had to be defined in DOMEA® SERdistiller by parameterisation and training, telling SERdistiller how to look for document areas such as address fields or document subjects.

Integration scripts

To integrate DOMEA® SERdistiller and the document repository of SERprocess server, a few Visual Basic scripts are used. Those scripts tell the SERprocess interfaces to store various values into the custom attributes of SERprocess.

Organisational definitions

Those definitions include the organisation model and the process definition. The organisation with users, departments and access rights is entered into the database using the admin tool. The sequence of routing steps is entered via SERprocess Designer. At the very beginning, the factory-supplied default process model can be applied. More complex models are interactively defined in cooperation with the users, resulting in process definitions which are in line with the users' requirements.

6.1.5 Project Experiences

DOMEA® SERdistiller proved to be a powerful and reliable engine for the categorization of documents and extraction of data. Automatic document distribution proves to be an efficient way to optimise the organisation. Initially, having all incoming mail processed by a single department was perceived as an overhead but the advantages of off-loading scanning and classification tasks and having a more reliable mail distribution outweighed that by far.

The solution is very extensible. When DOMEA® SERdistiller comes across a document that does not fulfil the matching criteria, then an operator is notified and he can perform the necessary steps manually. If documents that cannot be automatically categorized arrive frequently then this is an indication that a new class of documents has to be created so that these documents are processed automatically. Over time, while the knowledge base is growing, the ability of DOMEA® SERdistiller is increased to operate automatically. At the same time, routing process models can be modified over time without disrupting the system. It is even a good practise to start with a very simplified process model and refine the model based on user feedback. This approach makes sure that changing requirements are solved quickly and users are loyal because they are aware that the system has been built to support them.

6.1.6 Benefits, Costs and ROI

The first indicator of the efficiency of an automated solution is the grade of reliability that such a solution can achieve. The rate of documents that are categorized correctly is higher than 80%. This means that 80% of all documents are categorized with a level of

confidence that is regarded high enough for automated processing. Most of the remaining documents are correctly categorized as well but with a lower confidence level, so that a short user interaction is necessary to confirm the pre-decision of DOMEA® SERdistiller. Only a small minority of documents have to be handled manually.

The rate of successful data extraction is higher than 70%. For reasons similar to the above, the remaining documents must be completed or confirmed manually.

After the categorization is finished, the workflow processes handles document distribution automatically, so no interaction for distribution itself is necessary. This means that a significant amount of work is saved at the categorization and in the distribution process, while at the same time the overall process is much more reliable.

The availability of a central document repository proved to be a major benefit over the multiplicity of proprietary solutions. The overall costs for maintaining proprietary solutions, training users, keeping backups, etc. could be drastically reduced.

6.1.7 Adaptability of the example to similar problems

Basically, the example can be adapted to a large variety of problems. In each organisation that deals with external incoming documents, exact routing and exact content extraction is an important issue. Examples are service organisations, where service requests must be routed to the right department, depending on mostly unstructured requests.

6.2 Knowledge-Enabled Content Management Project of CHIP Online International GmbH

The InterRed GmbH, one of SER's partners in Germany, assisted CHIP Online International GmbH (the website of one of the most read IT magazines, <http://www.chip.de>) in quintupling the number of their online visits within twelve months. How were they able to achieve this?

In March 2001, CHIP Online had approximately 15 million page impressions each month. They were able to increase this number up to 45 million by December 2001, and expect to reach more than 70 million in March 2002. In addition, there is an above average number of page impressions per visit. With 7.2 page impressions per visit, this number is significantly above the national German average (these figures were calculated using the IVW method, <http://www.ivw.de>). The number of page impressions and page impressions per visit are the two most important indicators for determining the range and market significance of online offers. The consumer acceptance of an online offer rises and falls with the quality of the contents as well as with the technology in the background.

In other words, CHIP Online readers widely use the service because of the over 60,000 professional articles and 1 million information items, which the 50 editors continually create and update.

In addition, the portal technology that CHIP purchased last year before its online re-launch has also been an important factor in this success. As is readily evident in the CHIP Online example, InterRed's innovative content management system supports the performance factors which

- prompt the portal visitors to remain longer on the page and to return more often.
- eases and simplifies the work of the editorial staff.

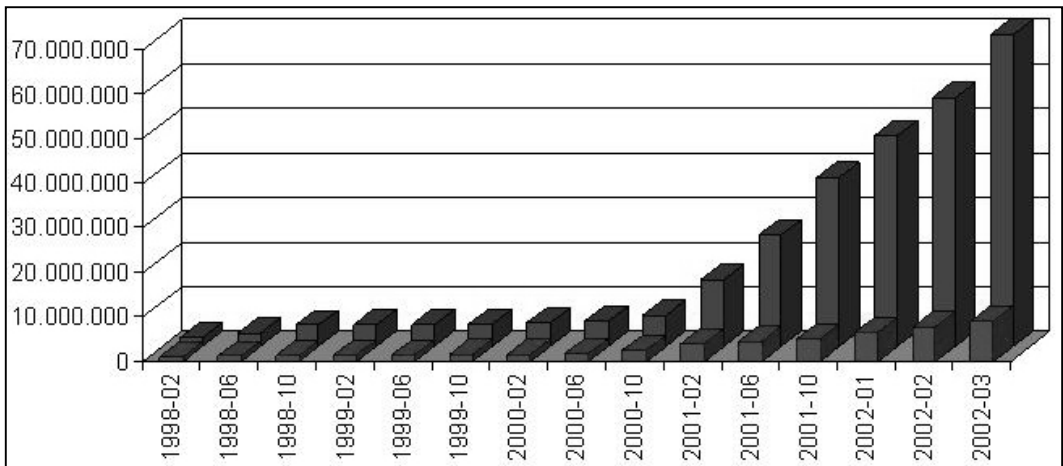


Figure 12: Page Impressions and Visits, the PI numbers have been quintupled within 12 months.

6.2.1 How did things work at CHIP Online before?

In order to provide an optimal user experience, CHIP Online created and maintained a taxonomy with more than 500 categories. Five employees had the task of keeping the taxonomy up-to-date. Their main task was to read each new article, then to classify it and relate it to the categories of existing articles. When the CHIP Online editors wrote new articles, they used this taxonomy to manually identify related articles and to set the appropriate links. Such links were listed as additional information beside the main article on a web page.

- Handling 60,000 articles in 500 categories in this manner lead to system immanent problems, which had to be solved:
- Different people tend to categorize differently which leads to inconsistent taxonomies. Everyone is most familiar with their own environments and are able to handle them well, but other subjects can only be handled less optimally.
- The links that editors offered as supplementary information to their articles were always older than the article itself. The related links were never be updated once an article had been published. For instance, if a bug report for a graphic card was released after an editor's initial hardware test article, then this new link was usually never added to the test article.
- A static taxonomy is not flexible enough to quickly adapt to new conditions and circumstances. Topics change, new topics occur – in a worst case scenario, this situation would have lead to a permanent update and re-categorization of all 60,000 articles and all 500 categories. As a result, it was not permanently updated and this prevented the best service with last news.

6.2.2 Content management, control and personalisation

In addition to the professional functions of the Internet based editing system and the impressive online page reproduction speed, the InterRed content management system has especially distinguished itself through its revolutionary software agent technology. The software agents offer a completely new way of automating content control and for personalising the CHIP Online website.

Such intelligent agents show users articles, which are contextually related to the ones they are currently reading. At the touch of a button, these agents can also support the article editors by showing a list of suggested links to other articles, which are thematically related to the main article.

The SERbrainware technology lends the software agents their capabilities. They understand the contents of an article based on natural language. Using these contents as a sample, they can compare them with the contents of other articles (content related content) and indicate close or distant relationships. As a result, the editor simply needs to select the desired links per mouse click from among the list of most closely related articles or the list is just taken automatically without any additional selection.

The software agents are also capable of getting to know each CHIP Online visitor personally and independently creating a profile of their favourite topics. The agents know the contents of the selected article and can offer visitors other related articles based on their individual profiles (profile related content). Visitors are recognised when they revisit the site and they quickly learn to appreciate the personalised service.

With a minimum of resources, the maximum use of all thematically related contents is achieved. Editors as well as visitors actively receive all information surrounding a desired topic regardless of where it is located in the web file structure.

The intelligent software agents

1. Context agents (content-related content)

When users enter the Chip Online website, the context agent automatically provides them with a list of similar articles and downloads related to the content of the web pages they are just viewing. The context agent knows the entire website content base and automatically recognises all related contents. In addition to factual information, it also presents links to ads or eCommerce pages. The context agent also provides a list of related links, articles or downloads etc. to someone who is reading an article about graphic card, for instance. The user also sees the article, which has just been released reporting that the graphic card has a manufacturing error.

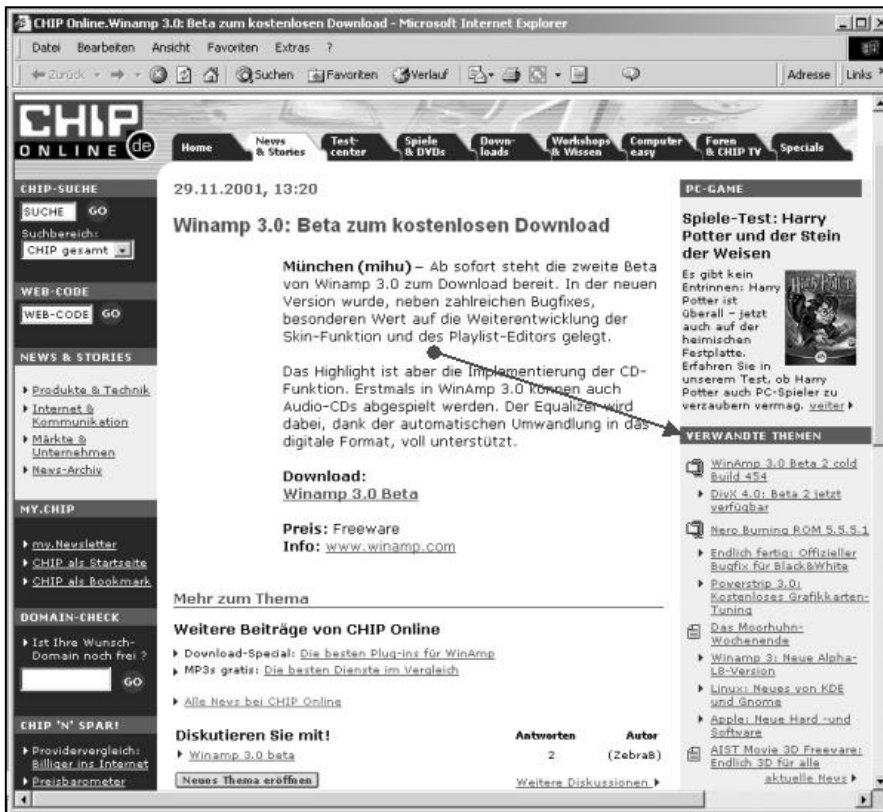


Figure 13: Fully automatic location and presentation of context related content (source: <http://www.chip.de>)

2. Profile Agents (profile-related content)

The profile agent provides the user with additional thematically relevant articles based on the user's previous behaviour. As a user "surfs" through CHIP Online's website, the profile agent makes note of their interests and adds them to the user's profile. As opposed to the conventional check box method, where users have to select keywords from a given list, the profile agent uses the entire article as an example for the selected topic so that the selection results are much more precise and comprehensive. Users can also manually add articles to their profiles. As soon as a CHIP Online user logs in he receives a personalised list of information regarding his areas of interest.



Figure 14: Personal interest profile and presentation of profile related content (source: <http://www.chip.de>)

InterRed offers two additional agents which round off its content management system. In their rather static statements, these agents deliver information regarding the most or longest read articles on the website.

HotSpot Agents recognise the most often read articles in the entire content pools. This not only allows the calculation of advertisement hits on the site, but also shows other visitors which articles are read the most. Natural human curiosity doesn't just increase the page impression rate but could also result in the editors adding more articles surrounding this topic. Based on the duration of the average visit, the Quality Agents know which article or news are especially worth reading. Stray bullets however are recognised and suppressed in the overall rating. These two agents measure the acceptance of the site. The results are indicators where the content could be possibly enhanced so that the overall content can be focussed on the mainly interesting topics.

6.2.3 Realisation of the CHIP Online solution

- The project was initiated by CHIP Online who analysed each software separately. The integration demanded by the CHIP required to create a team consisting of CHIP, InterRed and SER staff. The existence of clear interfaces allowed an easy and smooth integration of SERbrainware into InterRed's CMS. A stepwise implementation on the

user's site permitted an exact quality control.

The implementation at the customer site began in March 2001 when the core CMS system was successfully implemented. In parallel SERbrainware was tested by CHIP and implemented by SER and InterRed.

- Integration of the natural language based technology SERbrainware into InterRed's (www.interred.de) Content Management System (CMS). SERbrainware enables the CMS to identify dynamically and time-invariant content-related content as well as profile-related content. That is, because SERbrainware compares each new article's content with every other article inside the CMS and derives updated cross-links at any time. SERbrainware is also used to support the editor of a new article to automatically access articles and other items that he wants to use and link to his article.

The integration of the knowledge-enabling technology began in autumn 2001 and was completed in the same year. Each step was thoroughly tested and evaluated by October 2001. For each new piece of software the user's response (number of page impressions and page visits) was analysed. It turned out that expectations were exceeded. The software agent technology was finally implemented as the last important milestone in December 2001. CHIP Online appreciated the high quality of results tested with their own contents as well as the performance, speed and transparency of the technology.

- CHIP chose both SER's and InterRed's technology after a thorough test phase lasting several months. The knowledge-enabled CMS with its intelligent software agents was able to convince CHIP Online. The combination of the CMS and the agents provides the best possible content for the user and improves the user experience immensely. This had been proven by analysing the increase of page impressions on the website.

6.2.4 Benefits, costs and ROI

The benefits have been already discussed in the previous sections. A higher rate of page impressions resulting from better user experiences on the CHIP website has led to improved sales of the paper-based magazine. The advertisements in the magazine can also be sold for a higher price.

The editors' work can be streamlined and improved using knowledge enabled technology because it allows them easily to identify relevant information. Their output becomes more valuable by adding dynamic, time-invariant links.

"During the knowledge technology selection process, we were able to directly test SERbrainware using our own data. What won us over in the end was the high quality results which SERbrainware delivered and the speed and transparency with which it delivered them," said Helmut Scholpp, CHIP Online's Technology Director.

6.2.5 Adaptability of the example to similar problems in the public administration

Even though the application was developed for a very dedicated scenario it is possible to easily apply it to other scenarios in government as well.

The most promising way to improve the user's experience is to provide a variety of vertical and horizontal content and allow best possible information navigation. Vertical

content provides in-depth information for the user, e.g., a detailed hardware test on the latest graphics card. Additionally, if the site provides enough horizontal content then the visitor does not have to change the site to find the desired information somewhere else. Consider the example if the visitor wants to take a look on the latest tax regulations.

The best possible information navigation is when the user does not have to do anything, i.e. the information is made available proactively. This can be achieved using a push service based on the user's history on the website or by applying the user's profile of interest to present the latest news on the desired topics. The push service then either sends a message to the user (e.g. using email) or the information is supplied on the website the next time when the user is logged on. Furthermore, the user should have the chance to receive more information based on an already chosen content, for example a particular paragraph about a new law.

Some examples where the CHIP Online solution can be adapted will be presented in the following:

Scenario 1: Intranet or Information Portal

Today every public organisation is facing the problem of “finding” information related to content criteria on their local file servers – without having directed processes in place, which allow indexing and categorization by general means. “Intelligent Software Agents” like the content agents and the profile agents may be used to automatically present the relevant information and knowledge to the user.

If someone is interested in the latest regulations regarding smoking in public places, then this topic could be identified as relevant for this particular user and all new information regarding this topic would automatically forwarded to him. If another person is seeking specific information – perhaps on tax regulations – again related information can be directly made available by content-related links.

Content Agents additionally allow the user to interactively find content-related content. For example if someone is working on a particular case, the latest court decision for this case could be presented.

Scenario 2: Internet

Today most modern public administrations offer rich content on their websites. However, it is still a challenge for citizens to navigate directly to just that piece of information that they are looking for. Knowledge-enabling technologies as described here allow citizens to easily navigate through the website using their natural language and to access what they are looking for. Additionally, a push service can also be applied which forwards the information that matches a citizen's profile.

If public administrations were to offer such an easy-to-use service to the public, citizens would surely appreciate the improved service. Such a service would also increase information circulation as well as reduce the number of phone calls and queries.

7. Outlook

7.1 Citizen Portals

Indexing and retrieval within closed organisations could still be completed using traditional methods because the users who structure, store and retrieve documents are clearly defined. The structure could be uniformly set up organisation-wide and all users would be shown how to use the system. Although this would result in certain expenses, it could be realistically done. For large organisations, which undergo fast changes however, the risk of individual deviations is great and difficult to manage despite any regularly completed introduction measures. As long as we are working with dedicated electronic document archiving, meaning the secure, indestructible retention of facts, then document archiving systems still have their proper place. But that what is challenging us today is the information, which we receive daily and are expected to process as a part of our jobs. We are expected to quickly filter the information we need for our activities from among the volumes of information found on our desktop PCs, the organisation's Intranet and file servers as well as the Internet. To do this properly we need state-of-the-art support, even within closed organisations.

As soon as we leave closed organisations but still require individual and quick access to information then we can't continue to persist on using traditional indexing methods and analogue search methods. All members of the public must given access to services and information (e.g. tax returns), which they can readily and easily use. In doing so, acceptance for electronic processing via Internet will increase. As a result, the public would be able to take over many activities, like updating their addresses, and thus free up public institution resources for other activities.

In many cities, the first citizen portals have been opened by city administration offices. This approach is very good for providing citizens with general information. Such portals are currently almost always one way however, which means that a direct and open citizen dialogue is not yet possible. Consequently, traditional post or increasingly even email communication is used aside from the portal.

Let us imagine citizen portals of the future in this way: A citizen logs onto the desired web page in the Internet in order to see if there are any Wagner performances scheduled for the Semperoper opera festival in Dresden this year. Today this would still mean having to click through many pages on the Dresden city site in order to meander through the list of scheduled performances. Everything, which you find along the way would of course be very interesting, but the only thing you really want to know is if Wagner is included in this year's music festival. It would be so much easier if you could simply log onto the city portal and enter your request using one of two alternatives:

- In the dialogue window on the portal page, you would simply enter one or more sentences in your native language, like "Will any Wagner works be performed at this year's Dresden music festival in the Semperoper? For which days are they scheduled?" What you receive are the scheduled performances for the 2002 opera festival.

- Another variation would be that you wouldn't enter text in the dialogue window, but instead you would simply verbally pose your question. You would ask the same question in the website's integrated voice portal and would perhaps be asked for details in return just to make sure that your question was clearly understood. Then you would verbally receive the dates, which would also simultaneously appear on the screen in case you wanted to print them.

7.2 Natural language based portals

The first alternative has already been technologically realised, and the first cases have been implemented. As the best current practical CHIP Online example (chapter 6.2) shows, such innovative methods can be used with great success.

The second alternative is currently still in an experimental stage. Portal providers and their customers, e.g. telecommunication providers, still have not extensively gotten involved in natural language based technologies. Both currently have other projects, which do not allow them enough time to deal with language-based communication. Speech recognition results are already amazingly performant in terms of the recognition result quality and the speed at which the results are delivered. This type of technology is already often used in call centres for connecting calling partners.

Of course the background of such natural language based portals also has to be expanded so that not only is any sort of personalised query possible, but that extensive and current material is always made available for the answers. Manual capturing and indexing also has to be increasingly replaced in these steps with content related auto-categorization and data extraction including its validation. This is not just the case in portal application environments but also e.g. when handling the daily correspondence at a building authority. Building descriptions, which are normally included on the paper drawings and turned in for approval, can be quickly captured, classified and compared with the general building codes and laws in order to generate the respective basis for a decision. In which form the applicant is finally informed of the decision (per post, fax or email) is a separate matter; (see also chapter 6.1 about the Statistical Office of Saxony/Germany).

The technological possibilities for citizen friendly communication on the part of public administrations and authorities have not been nearly exhausted. Paper and archives are already very widespread, but now case relevant communication with companies and the public at large has to be tackled so that the volumes of paper, and the processing of the information contained within them, can be managed with at least the same resources if not even less. The applied index probably accepted by everyone within and outside of an organisation is the natural language they learned as young children. Why should institutions not draw upon the benefits and integrate it in their software applications or have their suppliers (software developers, systems integrators, etc.) integrate it for them? The future is not the question in this outlook, but the way we will master information and knowledge in future, which has already begun.

Glossary

ADL (Advanced Distributed Learning)	ADL is an initiative by the U.S. Department of Defence to achieve interoperability across computer and Internet-based learning courseware through the development of a common technical framework, which contains content in the form of reusable learning objects.
Associative Access	Knowledge retrieval based on pattern matching between an unstructured query (text paragraph) and a document content store.
Authoring tools	Tools/SW to create and adapt content to the web for use in an online course. They assist in creating e-learning solutions and provide a “do-it-yourself” option for placing content and materials online.
Categorization / Category	Assigning documents to different groups by performing content-related analysis - so called categories. Categorization schemes are typically built upon business processes and business rules or rely on knowledge domains within an organisation.
CD-ROM assessment	An assessment or survey that can be accessed and completed by using a CD-ROM launched through a company’s intranet. CD-ROM based assessments also can be used on a desktop stand-alone computer if the assessment is a self-assessment for the benefit of the trainee only. Alternatively, a CD-ROM-based survey can be printed (if the CD-ROM has a print capability) and used as a paper-based survey.
Computer-based training	A term used to describe any computer-delivered training, including CD-ROM, the Internet and Intranets. Sometimes referred to as Computer-assisted instruction (CAI), CBT is asynchronous learning.
Classification / Class	Collection of methods applied to categorize documents by analysing their content. In many cases, categories and classes are identical. Categories incorporate the semantics of the application, whereas classes may also be of formal nature.
Classify	Classification is a method of assigning retention/disposition rules to records. Similar to the Declare function, this can be a completely manual process or process-driven, depending on the particular implementation. As a minimum, the user can be presented with a list of allowable file codes from a drop-down list (manual classification). Ideally, the desktop process/application can automate classification by triggering a file code selection from a property or characteristic of the process/application.
Content Search	Information retrieval based on pattern matching between a query (text paragraph) and a document repository.
Distance learning/ Interactive Distance Learning (IDL)	Traditionally refers to a broadcast of a lecture to distant locations, usually through video presentations. IDL is a real-time learning session where people in different locations can communicate with each other. Videoconferencing, audio conferencing or any live computer conferencing (e.g., chat rooms) are all examples of IDL.
Document	A document (any form or format), an email message or attachment, a document created within a desktop application such as MS Word, regardless of format. There are two forms of document: Electronic Document: Body (text) of the document is stored in electronic format and can be read. If declared as a record, an electronic document becomes a managed record (i.e. a document may or may not be a (declared) record) Non-Electronic Document (Ndoc): A physical document of any form (maps, paper, VHS video tapes, etc.). Body is not recorded in electronic form, but descriptive metadata is stored and tracked within CM (profile). If declared as a record, an Ndoc becomes a managed record (i.e. an Ndoc may or may not be a (declared) record).
Document Life Cycle Management	The records life cycle is the life span of a record from its creation or receipt to its final disposition. It is usually described in three stages: creation, maintenance and use, and final disposition. e-Records applies management to all three stages. With e-Records, the records manager can create and maintain the official rules that will dictate when to destroy (or permanently keep) electronic records, as well as record

	and enforce any conditions that apply to destruction (e.g. destroy 2 years following contract completion). Finally, the records manager can carry out the physical destruction of electronic records, maintaining a legal audit file.
Document Security Control	Access control to documents (non-declared records) Note: Document security control is different from Records Security Control.
Electronic Recordkeeping	The practice of applying formal corporate recordkeeping practices and methods to electronic documents (records).
Electronic Signature	A signature is a bit string that indicates whether or not certain terms occur in a document.
Enterprise Content Management	Manage all content (i.e. unstructured information) relevant to the organisation. It embraces three historically separate technologies: web content management, document management, and digital media asset management. While outwardly dissimilar, all of these forms of enterprise content share similar needs for mass storage, search and access, personalisation, integration with legacy applications, access and version control, and rapid delivery over the internet.
EPSS (electronic program support system)	An electronic system that provides integrated, on-demand access to information, advice, learning experiences and tools. In essence, the computer is providing coaching support (i.e. the principal of technology based knowledge management).
File	A disk "file", something stored on electronic media, of any file. Does not necessarily denote a record. For example, "image files are stored on a server" simply refers to the electronic images, and implies nothing about the records status. Will be used in the context of describing the storage of documents and related information to electronic media.
File Plan Administration	Design and administration of the corporate file plan. The records manager can design file plan components. With Tarian's file plan designer, the records manager can design classes of file plan objects (files, records, folders, etc), then define the attributes of these classes. Relationships between classes are then defined (i.e. files can contain files, records and folders). Various views of the file plan may be defined. For instance, a warehouse view might present a view of the physical folders in the organisation, whereas a numeric view might present the sorted numeric structure for maintenance purposes. The records manager can create pick-lists enforcing consistency within the file plan, component profiles that define the characteristics of the file plan, and default values to simplify daily file creation tasks. Policies, Permissions, and Suspensions may be assigned to file plan objects.
Information mining	Linguistic services to find hidden information in text documents on content servers
Information Retrieval	An information retrieval (IR) system informs on the existence (or non-existence) and origins of documents relating to the user's query. It does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. This specifically excludes Question.
Keyword Search	Information retrieval method based on literal match of words.
Learning Resource interchange (LRN)	LRN is the Microsoft implementation of the IMS Content Packaging Specification. It consists of an XML-based schema and an LRN toolkit. It enables a standard method of description of content, making it easier to create, reuse and customise content objects with an XML editor, whether initially developed from scratch or bought under license from vendors.
Neural Networks	In information technology, a neural network is a system of programs and data structures that approximates the operation of the human brain. Typically, a neural network is initially "trained" or fed large amounts of data. A program can then tell the network how to behave in response to an external stimulus (for example, to classify a document based on its content).
Pattern Matching/Recognition	Matching/Recognition of objects based on features. Pattern Matching with regard to text documents means to identify and match words and phrases from different documents under the assumption that the more features match, the more similar the contents are.
Personalisation	The ability to provide the user with the right content both from the user's and Web

	site owner's perspective. A personalisation algorithm determines whether content is presented to the user, and if so, in what order of priority.
Portal	A single integrated point of comprehensive, ubiquitous, and useful access to information (data), applications, and people.
Record	Any form of recorded information that is under records management control. Records are either Physical or Electronic. Records may take any of the following four forms: Document: A document (see above) that has been declared as a record. Once declared as a record, the document is under records management control Folder: A folder of (paper) documents. Individual documents within the folder may or may not be treated as records (declared Ndocs). The physical handling of the folder is managed by Tarian's Physical Records Module Box: A box of (typically) paper documents. Usually contains folders (see above), which are individually managed as records, but may alternatively contain records other than folders such as loose documents of a given subject. The physical handling of the box is managed by Tarian's Physical Records Module Ndoc: A declared Ndoc (See above for definition of Ndoc) Important: A document (electronic or Ndoc) will not be considered to be a record until has been declared.
Record, Electronic	Electronic Records (e-Records). Any information (document) recorded in electronic form, on any digital media, that has been Declared to be a record. Characteristics of an e-Record: Document is in electronic form Metadata is associated with the document Document has been classified against a file plan Only the authorised Records manager has the means by which to apply retention/disposition to the document.
Record, Physical	Folders, Boxes, Ndocs to which records management control has been applied. A document (electronic or Ndoc) becomes an e-Record only once it has been declared.
Records Administration	The administrative infrastructure represents the tasks that the records manager carries out on the entire organisation's collection of declared records. Conducted within Tarian's Records Administration Client, a browser-based web application. End users never see this process. Consists of the following four broad activities; File Plan Administration, Records Security Control, LifeCycle Management, and Reporting.
Records Manager	Conducts one or more records administrative functions.
Records Security Control	Access control to declared records. Users and Groups of users may be created, and assigned roles and policies that will interact to determine the records users are able to access. Note: Records security control is different from Document Security Control.
Reporting	The process of generating reports from data managed by eRecords solution. It is a two-step process. Reports are first designed, and the design is saved for later reuse. Second, reports are generated by running the report design against the data.
Repository	Physical storage are for documents and/or electronic records.
Retention Rules	(Retention Schedule). The set of rules which specify how long to keep (retention) records, and what to do with them at the end of their lifecycle (disposition).
Syntactical Analysis	Syntactical analysis derives the syntactic category of words or phrases based on (language dependent) dictionaries and grammars. Example: house – noun.
Thesaurus	A book that lists words in groups of synonyms and related concepts.
Volume	Folder. A Volume will be referred to as a folder (common US terminology).
Virtual Reality (VR)	Virtual Reality simulations (usually involving wearing headgear and electronic gloves) that immerse users in a simulated reality that gives the sensation of being in a three-dimensional world.

Abbreviations

ASP	Application Service Provider
AVI	Audio Video Interleaving
BCR	Bar Coding
BPM	Business Process Management
CBT	Computer Based Training
CCD	Charge Couple Devices
CM	Content Management
COLD	Computer Output to Laser Disk
COM	Component Object Model
COOL	Computer Output On Line
DBMS	Database Management System
DMS	Document Management System
DRT	Document Related Technologies
ECM	Enterprise Content Management
E-Learning	Education, training and structured information delivered electronically
ERM	Enterprise Report Management
ERP	Enterprise Resource Planning
E-Term	European programme for Training in Electronic Records Management
FDDI	Fibre Distributed Data Interface
GIF	Graphic Interchange Format
HTML	Hypertext Mark-up Language
ICR	Intelligent Character Recognition
ICT	Information and Communication Technology
IDM	Integrated Document Management
ISDN	Integrated Services Digital Network
ISO	International Standards Organisation
JPEG	Joint Photographic Experts Group
KM	Knowledge Management
LAN	Local Area Network
LDAP	Lightweight Directory Access Protocol
MoReq	Model Requirements for the management of electronic records
MPEG	Moving Pictures Expert Group
NAS	Network Attached Storage
OCR	Optical Character Recognition
ODCB	Open Database Connectivity
OLE	Object Linking & Embedding
OMR	Optical Mark Recognition
PDF	Portable Document Format
PPP	Point-to-Point Protocol
RMS	Records Management System
RTF	Rich Text Format
SAN	Storage Area Networks
SQL	Structured Query Language
TCP/IP	Transmission Control Protocol/Internet Protocol
TIFF	Tag Image File Format
WAN	Wide Area Network
WAV	Audio Format File
WCM	Web Content Management
WebDAV	Web-based Distributed Authoring & Versioning
WORM	Right once read many times
XML	eXtensible Mark-up Language

Industry White Paper Series on Records, Document and Enterprise Content Management for the Public Sector

Capture, Indexing & Auto-Categorization

Intelligent methods for the acquisition and retrieval of information stored in digital archives

ISBN 3-936534-01-2

Hewlett-Packard GmbH

Conversion & Document Formats

Backfile conversion and format issues for information stored in digital archives

ISBN 3-936534-02-0

FileNET Corporation

Content Management

Managing the Lifecycle of Information

ISBN 3-936534-03-9

IBM

Access & Protection

Managing Open Access & Information Protection

ISBN 3-936534-04-7

Kodak

Availability & Preservation

Long-term Availability & Preservation of digital information

ISBN 3-936534-05-5

TRW Systems Europe / UCL - University College London / comunicando spa

Education, Training & Operation

From the Traditional Archivist to the Information Manager

ISBN 3-936534-07-1

Publishing Information

The series of six Industry White Papers is published to address the needs of public administration and archives at the European, national, federal and local level and to educate the public sector throughout Europe about available solutions for archival problems on relevant topics about acquisition, management, long term storage, multilingual access, indexing and training issues.

DLM-Forum

The current DLM acronym stands for *Données Lisibles par Machine* (Machine Readable Data). It is proposed that after the DLM-Forum 2002 in Barcelona this definition be broadened to embrace the complete "**Document Lifecycle Management**". The DLM-Forum is based on the conclusions of the Council of the European Union, concerning greater co-operation in the field of archives (17 June 1994). The DLM-Forum 2002 in Barcelona will be the third multidisciplinary European DLM-Forum on electronic records to be organised. It will build on the challenge that the second DLM-Forum in 1999 issued to the ICT (Information, Communications & Technology) industry to identify and provide practical solutions for electronic document and records management. The task of safeguarding and ensuring the continued accessibility of the European archival heritage in the context of the Information Society is the primary concern of the DLM-Forum on Electronic Records. The DLM-Forum asks industry to actively participate in the multidisciplinary effort aimed at safeguarding and rendering accessible archives as the memory of the Information Society and to improve and develop products to this end in collaboration with the users.

European Commission SG.B.3

Office JECL 3/36, Rue de la Loi 200, B-1049 Brussels, Belgium

A/e: dlm-forum@cec.eu.int

AIIM International - The Enterprise Content Management Association

AIIM International is the leading global industry association that connects the communities of users and suppliers of Enterprise Content Management. A neutral and unbiased source of information, AIIM International produces educational, solution-oriented events and conferences, provides up-to-the-minute industry information through publications and its industry web portal, and is an ANSI/ISO-accredited standards developer.

AIIM Europe is member of the DLM-Monitoring Committee and co-ordinates the activities of the DLM/ICT-Working Group.

AIIM International, Europe

Chappell House, The Green, Datchet, Berkshire SL3 9EH, UK

<http://www.aiim.org>

Industry White Paper Series on Records, Document and Enterprise Content Management for the Public Sector

The Industry White Papers are published by the DLM-Forum of the European Commission and AIIM International Europe to address the needs of public administration and archives at the European, national, federal and local level and to educate the public sector throughout Europe about available solutions for archival problems on relevant topics about acquisition, management, long term storage, multilingual access, indexing and training issues. The leading suppliers of Enterprise Content Management technologies participate in this series and focus on electronic archival, document management and records management for the public sector in the European Community.

Capture, Indexing & Auto-Categorization

This White Paper addresses the ever-increasing overload of information. An individual can read approximately 100 pages per day, but at the same time 15 million new pages are added to the Internet daily. Our limited human capabilities can no longer filter out the information that is relevant to us.

We therefore need the support of a machine which facilitates the exchange of knowledge by storing information and enabling personal, associative access to it through the lowest common denominator in human communication: The common human index is natural written and spoken language. All other types of indexing are limited aids which humans must first learn to use before they can employ them. To sum it up, the standard has already been set and recognised as natural language, but where are the systems which have adapted this natural standard?

ISBN 3-936534-00-4 (Series)

ISBN 3-936534-01-2