

The hidden potential of **unstructured content**

How to better leverage ECM
(Enterprise Content Management) investments

2009 / 07



Table of Contents

| | | |
|-------|---|----|
| 1 | Executive Summary | 3 |
| 2 | Initial Situation | 4 |
| 3 | An architectural approach for unstructured content | 5 |
| 3.1 | Content Repository Service..... | 6 |
| 3.2 | Meta Data..... | 8 |
| 3.2.1 | Organization-driven Meta Data | 8 |
| 3.2.2 | People-driven Meta Data..... | 8 |
| 3.3 | Business Logic | 9 |
| 3.4 | Presentation | 10 |
| 3.5 | Search | 10 |
| 4 | Consolidation as the first step towards an integrated architecture..... | 12 |
| 5 | Conclusion & recommendations | 13 |
| 6 | The Authors..... | 14 |
| 7 | The Company..... | 15 |

1 Executive Summary

The consolidation and standardization of IT applications always was and still is one of the main drivers for many enterprises. Today's CIOs are trying to achieve a standardized landscape of business applications. The consolidation and standardization of ERP or CRM systems for instance is one way to reduce complexity and to leverage investments. However, having many different content management solutions throughout the enterprise and not knowing what kind of content is stored in these repositories and how relevant they are for the business does not seem to be a big deal. Unfortunately this scenario is still common reality in many organizations. According to Gartner 75% of enterprises have 6 or more different content repositories and 34% are using three or more search engines.¹

Unstructured content is a general term describing every type of content that does not necessarily contain a structure which can be easily processed by computers. Typical examples are office documents, web content, videos, audio files etc. According to a study done by Merrill Lynch more than 80% of all potentially usable business information is stored in an unstructured form.² There is an enormous business value hidden in unstructured content which is readily available, but which is spread throughout the organization. Despite the fact that many other studies are showing the significance of unstructured content, most of the time these systems are not treated in a strategic way.

In addition the fragmented market of content management products and the lack of standards have led many organizations into using a variety of different content management solutions. The dawn of Enterprise 2.0 technologies such as corporate wikis and blogs brings even more complexity into the mix. In fact many companies are paying license and maintenance fees for the same features spread across multiple content management solutions. Having a consolidated landscape for unstructured content will reduce both complexity and costs. But it is not only consolidation and cost cutting that we describe in this opinion paper.

Leveraging the value hidden inside the unstructured content is not easy. In order to do that it is important to have a state-of-the-art content management architecture available that goes hand in hand with your business application platform. This architecture must also cover content-related components such as enterprise search and meta data in order to bring added value to the existing silos. In this document we describe a blueprint for such an architecture based on best practices. Implementing this architecture along with the consolidation of existing content management solutions can help to leverage the existing content in new ways, while achieving the necessary cost cutting and complexity reduction.

We strongly believe that a strategic and architectural approach is needed to get the most value out of unstructured content in an organization. In times where most business applications are built on top of standardized products off the shelf, maybe this is one of the few remaining areas in which to gain competitive advantage.

¹ Mark Gilbert, "Enterprise Content Management: Architecture and Governance", September 2008

² Christopher C. Shilakes and Julie Tylman, "Enterprise Information Portals", Merrill Lynch, 16 November, 1998

2 Initial Situation

Take a look at knowledge workers in modern organizations. The vast majority of the information used on a daily basis is so-called unstructured content (office documents, emails, web content etc.). The lower you go down the hierarchy, the more employees are dealing with this kind of information and less with structured data coming out of business applications. The amount of content produced by knowledge workers is vast and increasing at exponential growth rates. So it is obvious that unstructured content represents a company's knowledge, which is one of the core assets in modern organizations. However, a strategic awareness of solutions for managing that content and to leveraging its hidden value is often missing.

The situation gets worse when taking a look beyond typical scenarios like inter-/intranet, file servers and document management systems. Having valuable high quality content can be a key differentiator in competitive environments. Providing rich media like video over the Internet, whether it is for marketing or training purposes, is not unusual anymore. A well-designed self-service portal including videos on how to solve common issues with certain products can decrease costs for customer support dramatically. It is not likely that most users will read long instructions on how to fix a problem. A short video is much more convenient to use and much more likely to be understood correctly. In addition to that companies are expected to have a periodically updated Blog, as well as department level Wikis and so on. Using separate and not integrated systems for each of these use-cases makes it hard to integrate the content into business processes and can lead to a situation where it is not possible to manage published content efficiently.

Taking a good look at the solutions and products available for managing content it becomes evident that the industry has not come up with any major innovation during the last ten years. The number of industry standards is still low and barely implemented by vendors. In addition to that many organizations have not put the necessary efforts into leveraging their investments in this area. As a result a large number of customers are dissatisfied with their current content management solution.

This document describes the challenges companies are facing, and the risks they are taking by not including content-related systems into their overall strategy. We describe how to put back this essential part of the business into the big IT picture and how to increase the value of existing content by bringing it into the context of the business.

We start to see an increasing interest in consolidation, especially in the area of unstructured content and content management software. Recent developments in the area of standardization and product innovation are backing up this trend. In the following chapters of this opinion paper we describe a guideline on how to leverage your content management investments.

3 An architectural approach for unstructured content

When business users end up evaluating a content management solution they actually start with a business requirement like "we need a way to manage contract documents of our customers" or "we need a place where we can share and discuss documents related to a new product". They go out and ask IT for a solution and they in turn end up buying or building a specific solution to an isolated requirement. The results of doing this over time are many small applications built on top of technologies such as Lotus Notes or MS-Access which are spread throughout the enterprise. Each one of these little solutions contains very valuable information for a certain limited number of people. However, their value is not transparent to anyone else outside this special interest group. Getting control over these applications and being able to evaluate their value for the enterprise as a whole is one major challenge many companies are facing now.

Even if there is an existing Enterprise Content Management (ECM) System available and part of the overall IT strategy, it seems not to fit when it is time to replace one of the vertical applications in which the content is generated. In order to provide a fast solution for a decent price, many companies go out and buy a technology that is wide-spread and easy to use, instead of a content management solution that is specified for managing corporate knowledge.

Content management products have been available for more than 20 years now and are considered to be mature. But why is it that they are not as ubiquitous as relational databases for instance? Maybe because there are no widely-adopted standards about what a content management solution should be able to do. Depending on whether you are talking about records management, document management or web content management you will get quite a variety of opinions about the set of features expected in such a product. The sheer variety of the market and its different types of vendors reflects this quite well. Uncountable numbers of companies are providing products of very different nature, starting from infrastructure components going all the way up to very specialized vertical applications, all claiming to do some kind of content management. This situation is maybe the biggest blocker of ubiquitous content technology.

A lack of standards and the complexity of proprietary products have led to a situation where most of the customers seem to be unsatisfied with their existing content management solution. Lack of business alignment and poor implementation seem to be the biggest pain points.

So how can ECM customers change this situation and be more effective in the way they manage and use their unstructured content? A blueprint as a result of best practices, which is described in this document, may give an answer to this question. This blueprint contains a repository service as the foundation and four functional building blocks for Meta Data Management, Business Logic, Presentation and Search. Content-centric vertical applications supporting the needs of the business are built on top of the functional building blocks. The following diagram shows an overview of this architecture followed by a detailed description of each part.

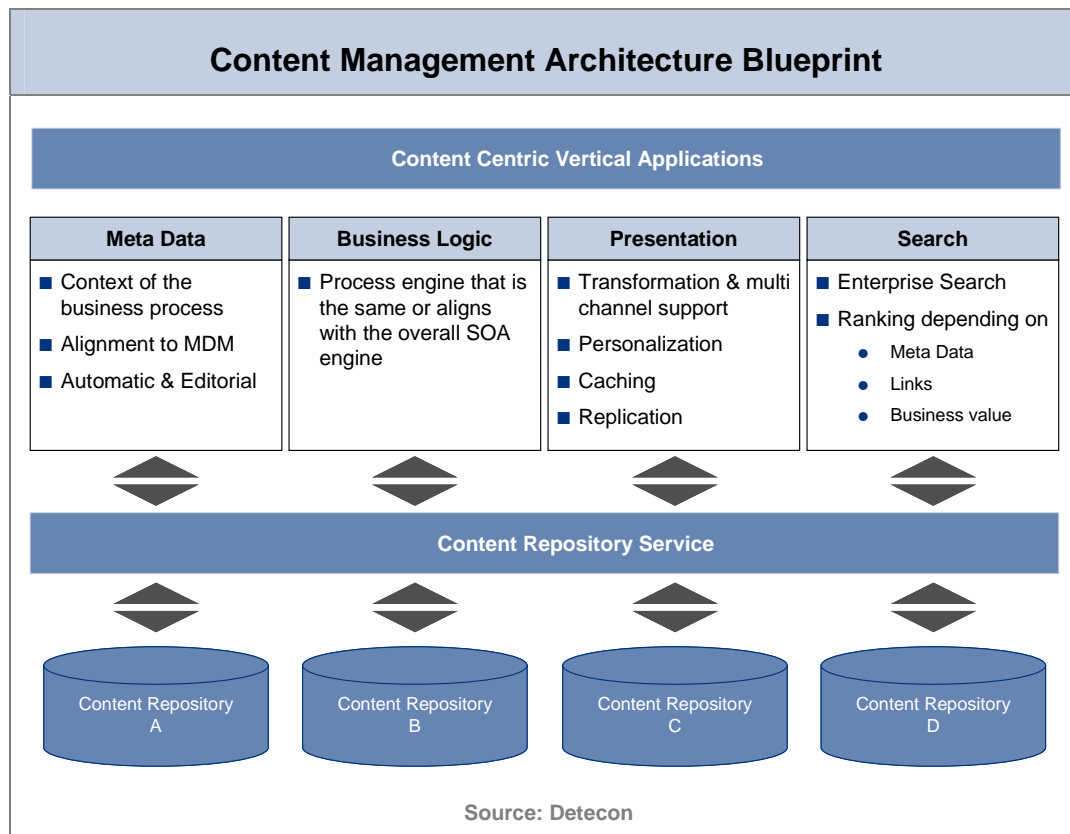


Figure 1: Content Management Architecture Blueprint

3.1 Content Repository Service

In order to meet current and future business needs, enterprises should look for a solid unstructured content infrastructure for building content-centric vertical applications. The role of the repository service is to provide services for basic object level access and storage. Having a unique way to access unstructured content is a huge step forward to enable vertical content-centric applications. If we compare ECM to relational databases we see that even if the differences between RDBMS (Relational Database Management Systems) from different vendors are still big, the existence of a standardized query language like SQL has boosted the penetration of this technology in enterprises. The lack of standardized access and query technology for unstructured content seems to be a crucial limiting factor.

A solid and modern content repository service should not require the replacement of existing content repositories. Unstructured content is hard to migrate, since it contains data, relationships and formatting information at the same time. This is the reason why many migration projects end up in manual copy & paste exercises instead of automatic processes. The repository service should act as a meta layer on top of existing content instead. This allows for re-using investments in existing content. In addition to that the repository service should allow the enrichment of existing content with meta data or with relationships to other repositories.

Until now only a few vendors have provided an implementation tool for a content repository service, based on either proprietary technology or on not widely-adopted standards. This is likely to change very soon. There are serious efforts under way to build fully-featured and standard-compliant repositories, and they are gaining momentum. This could have a huge impact on the costs for connecting existing content. All these products still need to be integrated in the enterprise environment, however. So the design of an appropriate repository service that complies with an existing IT architecture still remains a task which customers have to do themselves.

CMIS – The new Lingua Franca for content repositories?

In September 2008 some of the biggest players in the ECM market announced Content Management Interoperability Services (CMIS) under the umbrella of OASIS (Organization for the Advancement of Structured Information Standards). This was not the first attempt to standardize access to content repositories, although there are signs that this specification will be more widely adopted by vendors than the previous ones (i.e. JCR a.k.a. JSR-170). Some reasons why this standard might be a success are:

- Three major ECM vendors (Microsoft, IBM & EMC) are throwing their weight behind this standard and have already been working with it for two years
- All other major ECM vendors (Oracle, Open Text, SAP etc.) are supporting it
- Even small and open source vendors are supporting it (i.e. Alfresco)
- The standard is not tied into a certain programming language (like JCR & Java)

Having such a standard with broad support by vendors is exactly what is needed to build an eco system around vertical content-centric applications. For tools vendors it would be easier to find skilled employees and the market would grow to a size that will allow them to invest in this area. Last but not least customers would not be handcuffed to certain proprietary repositories.

We strongly recommend to keep an eye on the CMIS standard and ask current ECM suppliers whether they intend to support it and if so, when.

Figure 2: CMIS – The new Lingua Franca for content repositories?

Right now there is a shift on the technology side from a document-centric way of managing content into object-oriented content management. The object level management of information allows much more re-usage of information than is possible with traditional document-based approaches. Object-oriented repositories store every piece of information as a single object that can have associations to other objects and that has its own set of properties. A document is an object tree with links to many other objects. The object level management of information has been pioneered by web content management vendors and is likely to be adopted by traditional ECM and DMS vendors soon. This is a key technology to enable a wider re-usage of content and reduce the costs to create new content by re-composition of existing content. We consider content re-usage and how it is supported by repositories as one of the key criteria for new investments in ECM technology.

It is impossible to set up one single physical repository for all content in the enterprise. When designing an infrastructure for content it is important that existing content in proprietary repositories can be connected to the service automatically.

It is also very important that this service can be scaled according to the amount of usage by the enterprise. Having sophisticated caching and replication functions built into it is a must.

The relationships between each piece of content are not limited to simple unidirectional links (like links in HTML). The repository service should be able to manage multidirectional associations between content objects. This can have a huge impact on how content can be found and ranked in query results.

The content repository service acts as a meta layer above the existing content and it provides connectivity to other existing content. Moreover it enables the storage of new content.

3.2 Meta Data

Metadata brings content into its business context. The quality of meta data has a massive impact on how fast relevant business information can be found. Creating and deploying high quality meta data throughout the enterprise is a very challenging task, though. Basically there are two approaches to solve this problem:

3.2.1 Organization-driven Meta Data

This approach requires a set of well-known meta data that align to the core business processes and are widely used in a consistent manner. Prior to implementing organization-driven meta data it is recommended to catch up with projects around master data management, if there are any under way. In this case it is strongly recommended that the meta data model aligns with the master data model as well. In a best-case scenario the meta data model for enterprise content is a subset of the master data in the Master Data Management system (MDM).

Once the meta data model is available, it has to be implemented throughout the enterprise. It is not possible to enforce consistent maintenance of meta data in a large organization. As a first step, systems must be identified that contain high quality and reviewed content. Such systems are likely to be updated easily by their users, if the number of users is limited and if a tight control process is in place.

For large systems with a large number of users changing things on a regular basis, it might not be a good idea to give everyone the right to make any changes at the meta data level. The usage of implicit meta data (for example depending on where an object is stored) or automatic classification with e-discovery tools might be a better solution.

3.2.2 People-driven Meta Data

A much more Web 2.0-like way is for users to tag content with any set of meta data. In this case there is no way to ensure a consistent set of meta data, but the amount of available meta data is likely to increase and to get better every day. IBM for example encourages its employees to tag useful intranet pages. The search engine uses the tagging for ranking its results. IBM estimates that this tagging saves each user 12 seconds a day. This would translate into an estimated annual productivity gain of 4.5 million US\$.³

³ Gary Matuszak, Enterprise 2.0 Tales from the Trenches, KPMG, 2008

However, implementing this kind of technology is not enough. In some cases cultural changes are necessary to get value out of this technology. Employees must be encouraged to share, review, rate and tag content as part of their daily job.

Both approaches can be used in conjunction. When it comes to relevance of the content (for example in the ranking of search results), meta data can be the key to achieving the next level of corporate knowledge. Some companies favor editorial and revised content from tightly controlled systems over tagged content from the intranet or from an internal bulletin board. The quality of editorial, automatically classified and tagged content must be examined and compared periodically. Accuracy of the meta data must be the driving factor behind any presentation of relevant data.

3.3 Business Logic

One limiting factor to a wide usage of content management technology in enterprises is hard-wired workflows and business logic built into these systems. Even if a content management system has a configurable business process engine built into it, in most cases the engine is tightly coupled with the repository and is only aware of the services built into the content management system.

In order to cost-effectively build and maintain content-centric vertical applications the business logic must be separated from the repository service. Business process engines are not a new technology anymore and they might already be in place for SOA⁴-related projects.

Here is an example: Companies doing e-business take special care about their online customer experience and invest in their CRM system, but this is not enough anymore. Customers are expecting richer shopping experiences, bigger bargains and specific offers for their needs and interests. Traditional CRM Systems are not optimized for personalization. Being able to combine the personalization engine of the web content management with the data from the CRM System would seem like the icing on the cake, but it does not exist at the moment! We might see this kind of integration as well as continuous improvements and tweaks by business users very soon, however. Their knowledge about their customers will increase by applying Knowledge Management methods and by using mature tools.

There is a long way to go yet towards the goal of consolidated vertical applications. The connectivity of content repository services and process engines is not even the hardest part of this challenge. First of all, as described before, most available content repositories allow only certain predefined workflows. Variations and new processes are poorly supported and require extensive customization by the customer. Another obstacle is the nature of unstructured content and content-centric vertical applications. They are not well supported by current process engines, which tend to focus more on processes relying on structured data.

⁴ Service-Oriented Architecture

3.4 Presentation

According to a study by Forrester for 2008 & 2009, 40 - 60% of enterprises in North America and Europe are going to invest in more mobility support for their employees, even in a down market.⁵ What this means is that it is unpredictable in which way individual end users will get in touch with the content they require. Dedicated and controlled desktop applications (i.e. Browsers like Internet Explorer) will not remain the dominant way of accessing information.

Multi-channel support is definitely not new for content management solutions. However, this support is also most likely to be tied tightly to the content management product and cannot be used universally. The ability to access information from the repository service and to transform the content into its appropriate format is not enough. Presentation services must also be aware of personalization and of the context in which users are accessing their information. In addition to that presentation services may take care of eliminating content replication and of synchronizing caches and proxies.

Setting up solid presentation services is the area where project and enterprise-specific solutions are most likely to be developed.

3.5 Search

The discovery of relevant information related to a given task and to the context of a user is maybe the biggest challenge that most companies fail to solve. Many off-the-shelf search engines provide very sophisticated algorithms to find related content based on the search terms a user has chosen, but more often than not the results remain poor and the ranking does not make any sense. The reason for this is that search engines are not aware of the meta data, the relationship between individual documents and how relevant they are to a specific user.

Once a content repository service is available as described above, there are many new ways to put some "intelligence" into the search engine and how its results are ranked. The data needed for this "intelligence" already exists and it is getting better every day.

Links and relationships between documents represent a very important set of implicit data that is available, even if it is not obvious at first sight. This is the reason why internet search engines like Google make wide usage of the links between sites to find out more about their relevance. In addition, different sources of links get treated differently, depending on their relevance. Doing a similar thing within the enterprise is not that easy, however. On the web everybody has agreed on the same standards (HTTP, HTML), which makes it quite easy to find relationships between content. In the enterprise world there is no such standard way to create relationships between content. Most systems have their own proprietary way of building associations between pieces of content and external systems. Therefore it is important to have a content repository service.

⁵ Michele Pelino, "Predictions 2009: What's In Store For Enterprise Mobility", December 2008

Each repository that is connected to the service can expose all relationships in a common way. This makes it much easier to make the search engine “aware” of these relationships.

Another valuable information is the meta data attached to the content. First of all meta data allow reasonable filtering of information depending on the specific context of each. As mentioned before, there are different kinds of meta data that can be attached to a document. For example, if a piece of content has been classified by a team of editors and/or experts and is explicitly assigned to certain meta data, it is likely that this kind of document is much more relevant than some intranet page where everybody can leave a comment or make changes.

Search engines are not likely to be able to utilize this kind of valuable information out of the box. It is very likely that additional customization is needed to make a search engine more “intelligent” in a way that makes sense. In the future we may see some improvements in this area, especially if standards like CMIS (Content Management Interoperability Services) get mature and will be widely adopted.

4 Consolidation as the first step towards an integrated architecture

Looking at the current situation most companies are far away from a well-defined architecture for their content management applications. The reasons for this are certainly manifold. We believe that the following two are the most relevant:

First, content management using information systems has been established for years. The first document management systems were installed in the 1980's. At that time nobody thought of IT-architecture management in the same way we do nowadays.

Second, nearly all content management systems are installed as isolated solutions. Although having interfaces to the "rest of the world" they remain niche applications. In general they support only subareas of end-to-end business processes and not the whole business process, the way e.g. ERP systems do.

Taking this into account it becomes quite clear that normally ECM-Systems are spread all over the business units of a company and are not always included in architecture management initiatives. Nevertheless it is of great importance not to neglect the applications dealing with unstructured content when architecture initiatives are undertaken.

Wherever several ECM systems are running, the best and most pragmatic way to achieve the target architecture is to combine an architecture initiative with a consolidation initiative. Using this combined approach the first step is to develop the target architecture as described above. This target architecture serves as a framework for the consolidation initiative. According to the principle "think big, start small" the second step comprises identifying appropriate consolidation candidates. The most obvious candidates are those who provide the same or nearly the same functionality. After identifying appropriate candidates the next task is to decide which of the redundant systems will be continued and which will be retired. For this decision it has to be taken into account that the big players like IBM, Oracle and Microsoft are very likely to capture the market with their solutions. We expect them to gradually increase the functionalities of their offering and thus attack established ECM vendors. So in case a redundant functionality is comparable to that of one of these offerings, a preference for one of these vendors would suggest itself under strategic considerations.

At this point it should be mentioned again that the consolidated system should match the Architecture Blueprint. It is recommended that all the needed functional building blocks are established right at the beginning. Having built all these components the whole system should – as a proof of concept – be tested and improved continuously concerning its scalability, performance, robustness, etc. Once such a nucleus has been set up, a roadmap to transform and integrate all the other ECM-systems can be defined.

Proceeding this way it is possible to achieve a well-defined architecture for content management, even if the company is already operating a lot of different ECM-systems. But this is not the only benefit. Consolidating applications – regardless of whether they are dealing with structured or unstructured content – helps to save IT costs. Besides license fees, savings can be realized in maintenance costs, platform operating costs, labor costs and indirectly by reducing the number of interfaces and thus reducing complexity. In times of economic downturn this is not a negligible aspect.

5 Conclusion & recommendations

The establishment of a consolidated architecture as described in this document is partly supported by existing and upcoming content management products. But there is no time to sit back and wait for these products to hit the market. According to a study done by AIIM (Association for Information and Image Management), organizations with large pools of content can save up to 30% on the costs of maintaining them, if they were able to leverage the re-use of content on a large scale. So the content re-use is clearly the killer application when it comes to ECM. Therefore, enterprises have plenty of homework to do before they are able to reap the returns on their investments.

First of all, the target architecture blueprint has to be developed, followed by a consolidation of the existing content repositories. The fewer repositories there are, the less complex it will be to implement a content repository service. A well-defined content migration strategy has to be the basis of any consolidation effort. The amount of content and the inability to extract the content in an automatic and standard-compliant way makes this step expensive. It is strongly recommended to consult experts in the area of content migration to make this step a success. It is also important to keep in mind that not every piece of content must be migrated. In fact, many companies use the migration process to separate valuable (i.e. re-usable) content from legacy that can be destroyed or archived.

In the consolidation step analyzing the existing content repositories will result in a map of functional capabilities they offer and how relevant they are for business processes. This can be used as a solid basis to design a content repository service, defining how each capability is mapped to existing systems. Functional analysis followed by matching the capabilities of each existing repository is not enough, however. Strategic factors should also be considered in the selection of the repositories that should continue to exist. Relying on small niche vendors as providers for such a strategic piece of infrastructure brings some risks that must be addressed. If content is to be re-used, users must be able to find it quickly and easily. One critical success factor here is proper management of meta data. Having a consistent set of meta data that supports all business needs is a key enabler for leveraging content. For all the meta data there must be a clear understanding about the information it provides for business processes and how it can be used to define the value of content. It is quite likely that the amount of information in the enterprise will double each year. The quality of the meta data therefore is essential for enabling each user to separate relevant from less relevant content in their particular context.

Last but not least is the ability to use meta data and the relations between individual items of content to rank these items as the result of search queries. We strongly believe that there is enormous potential in the area of enterprise search once the meta data is done right and once most of the relevant content is available through a central service.

There are quite a few mature technologies and methodologies available for unstructured content and they are being improved continuously. The key success factor in the future is to truly integrate these technologies into business processes. So the question is: Do you want to sit back to watch your competitors collect the first experiences with these kinds of approaches or do you want to be in the driver's seat and shape your own architecture as a first mover? We believe in the latter. By doing this, your organization will be ready by the time vendors and the content management market makes their next evolutionary step. Make sure you will be part of it.

6 The Authors

Ali Saffari has been a Managing Consultant for Detecon in the Competence Group Application Management since 2008. Following his studies of electro-engineering at the University of Applied Sciences in Cologne his career started as a consultant at one of Europe's leading system integrators. Through numerous projects he has gained substantial knowledge of document and records management as well as business processes in the public sector. He continued his career at a leading web content management vendor, where he was responsible for the program and product management of several content management solutions.

He can be reached at: +49 228 700 1914 or

Ali.Saffari@detecon.com

Dr. Norbert Hövelmanns is a Partner in Detecon's Information Technology Practice, where he is the head of the Application Management Group. Following his studies of computer science at the Technical University in Aachen he started his career as a consultant for technical IT at an automotive supplier. In 1992 he was promoted to the CIO of this company. Since joining Detecon in 1996 he has run a lot of IT-Management projects and through this gained broad experience in current topics of managing complex application landscapes. His area of expertise is the strategic aspects of application management.

Phone: +49 6196 903 286 or

Norbert.Hoevermanns@detecon.com

7 The Company

Detecon International GmbH

Detecon International is a leading worldwide company for integrated management and technology consulting founded in 2002 from the merger of consulting firms DETECON and Diebold. Based on its comprehensive expertise in information and communication technology (ICT), Detecon provides consulting services to customers from all key industries. The company's focus is on the development of new business models, optimization of existing strategies and increase of corporate efficiency through strategy, organization and process improvements. This combined with Detecon's exceptional technological expertise enables us to provide consulting services along our customers' entire value-added chain. The industry know-how of our consultants and the knowledge we have gained from successful management and ICT projects in over 100 countries forms the foundation of our services. Detecon is a subsidiary of T-Systems, the business customers brand of Deutsche Telekom.

Integrated Management and Technology Competence

We possess an excellent capability to translate our technological expertise and comprehensive industry and procedural knowledge into concrete strategies and solutions. From analysis to design and implementation, we use integrated, systematic and customer-oriented consulting approaches. These entail, among other things, the evaluation of core competencies, modular design of services, value-oriented client management and the development of efficient structures in order to be able to distinguish oneself on the market with innovative products. All of this makes companies in the global era more flexible and faster – at lower costs.

Detecon offers both horizontal services that are oriented towards all industries and can entail architecture, marketing or purchasing strategies, for example, as well as vertical consulting services that presuppose extensive industry knowledge. Detecon's particular strength in the ICT industry is documented by numerous domestic and international projects for telecommunications providers, mobile operators and regulatory authorities that focused on the development of networks and markets, evaluation of technologies and standards or support during the merger and acquisition process.

Detecon International GmbH
Oberkasselerstr. 2
53227 Bonn
Telefon: +49 228 700 0
E-Mail: info@detecon.com
Internet: www.detecon.com