Knowledge Retrieval in Complex Information Landscapes Related to Products and Production

Marcus Schichtel

Knowledge Retrieval in Complex Information Landscapes Related to Products and Production

M. Schichtel

BMW Group
markus.schichtel@bmw.de

Abstract. Developing and manufacturing complex vehicles today creates an increasing amount of information and knowledge that is by far not stored in a single data source. The challenge of knowledge retrieval in complex information landscapes has been approached at BMW with the help of semantic search technologies and a user oriented design of the "Knowledge Finder", in order to truly connect unstructured and structured data sources semantically.

Keywords: Information Landscapes, Semantic Search, Knowledge, PLM, Manufacturing

1 INTRODUCTION

Developing and manufacturing complex products such as vehicles today creates an increasing amount of information and knowledge that is by far not stored in a single data source like a PLM-system. Over time the number of isolated information islands has grown to a point where nobody oversees anymore how particular pieces of information are linked together that actually form a complex information landscape. In [1] a good example is given explaining what consequences could arise, if a specific dependency e.g. between a safety sensor and an environmental side condition remains hidden in a complex information landscape describing the functions of a car. That article points out in particular that the missing information that could have resolved the problem before product launch was simply located at an altogether different location.

To put this observation into another context it shall be pointed out (see fig. 1) that nobody can possibly know all available data sources in a large enterprise. Rather each of the people involved either has the needed piece of information in his memory or searches for it in those data sources that are personally known. Looking at the pie chart reveals that this way approximately only 10% of information and knowledge is tapped into. By far the largest portion is knowledge (approx. 90%), that an individual is not even aware of that it exists ("I don't even know that I don't know"). In fact making it possible for the individual to "sail this (to him) unknown sea of knowledge" is a major challenge today.

adfa, p. 1, 2011.

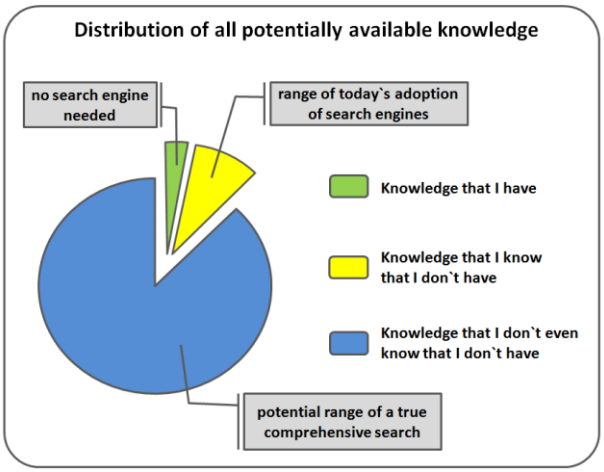

Fig. 1. - The Threefold way of categorizing potentially available knowledge in a large enterprise

An interesting and strong corroboration of this observation is the following quotation from [2]:

"Gelesene Bücher sind längst nicht so wertvoll wie ungelesene. Eine Bibliothek sollte so viel von dem enthalten, was man nicht weiß, wie der Besitzer angesichts seiner finanziellen Mittel … hineinstellen kann"[1]

The challenge of dealing with isolated information islands (or information silos) has drawn top management attention as pointed out in [3]:

"… in order to answer a question information silos are a hindrance. … How many different locations do I have to visit in order to put many single pieces together for meaningful information? Information must be accessible to all decision makers throughout the enterprise."

Figure 1 can also be interpreted as the vast offer of knowledge in any enterprise that is waiting to be uncovered again by the user that has a great potential of matching the pressing demand for information and knowledge from prior experiences. The following short list shall illustrate this demand showing a great bandwidth of complexity:

- What does "CFK" stand for?
- Who is the right contact to discuss an issue with the fender?
- What module does part "4711" belong to?
- What would be an alternative material for this alloy?
- What issues did we have with the headlight in former car projects?
- What would be the environmental impact of using this coolant instead of the one in use today?

---

[1] ***Translation by author:*** *"Books that have already been read are not as valuable as those unread. A library should always contain as much of those things not known to its owner as the owner can possibly shelve into his library according to his financial abilities"*

But digital information alone is not the whole story. Lots of information and knowledge is also buried in hardware prototypes! In former times the classical way of validating a product was putting hardware prototypes together. The process of putting together gave rise to insights into problems right away and naturally gave answers to questions not explicitly asked before, i.e. the physical prototype answered the question of "Does it work?" implicitly as a whole. The people carrying out the validation process gained a tremendous treasure of knowledge of experience with no particular need to store this knowledge digitally in data sources. However today, with the increasing pressure to reduce hardware prototypes more and more validation is carried out with virtual prototypes that are not put together like physical prototypes and therefore cannot reveal important issues right away.

Rather, validating a virtual prototype implies the need to know issues from the past so that the engineers validating a product can inspect the virtual prototype accordingly. Complex questions during validation such as "What implications on production processes would be incurred if we used a different blank thickness in car body design?" would resemble such a typical situation.

In order to make this new paradigm work we need to first digitize older knowledge and second find a way to make it accessible efficiently.

With the increasing pace of business (faster processes, increasing fluctuation among employees) less time is available to find missing vital information for the task at hand.

The classical idea of connecting IT-systems by interfaces that are hard and expensive to develop, to maintain and sometimes to even redesign may have worked in the past but cannot catch up with the increasing speed and complexity of today`s engineering processes. In fact, today it is not uncommon that business processes in general are expected to happen in real time. Hence new and more sophisticated IT-tools are needed guiding the user through the complex information landscape to find the right knowledge in acceptable time.

2 THE BMW-APPROACH: KNOWLEDGE FINDER

2.1 Business Context

The principle ideas of the Knowledge Finder originated from a close collaboration between business (department in charge of the car validation process) and IT while creating a new form of review method in order to detect issues early on in a car development project. During the first pilot of the new review method it became evident from the beginning that the method stimulated the open flow of knowledge among the participants across all disciplines. Though, as it has been noted evaluating the first reviews, the flow of knowledge couldn't be captured for future usage.

In other words knowledge was elusive and the immediate question was whether this knowledge could be easily captured by a knowledge database. Soon it became clear that creating a new knowledge database was not the right answer to the challenge at hand for a number of reasons:

- The creation of a new data source would have meant to implement a very challenging, if not impossible, change management task considering the vast amount of business processes that would have to be changed in order to store data into the new database rather than using the existing data sources
- Further it would have meant to create a new silo of information in addition to the existing ones that still continue to contain very valuable information
- It would have been totally unclear at what point in time the amount of "new information" would have reached a sufficient level so that the users would have reaped a benefit tapping into the new knowledge database compared to the existing data sources

Hence the fundamental design decision for the Knowledge Finder came to rest on three pillars (see fig.2):

- Reuse of existing data sources rather than creating an additional new knowledge database
- The idea of Knowledge Finder acting as an agent networking in the "community of data sources" similar to a human that networks among his coworkers
- Exploitation of "state of the art"- semantic search technologies in order to follow a user-centric approach and to create an intuitive and "fun to use" – tool to find knowledge rather than search for it

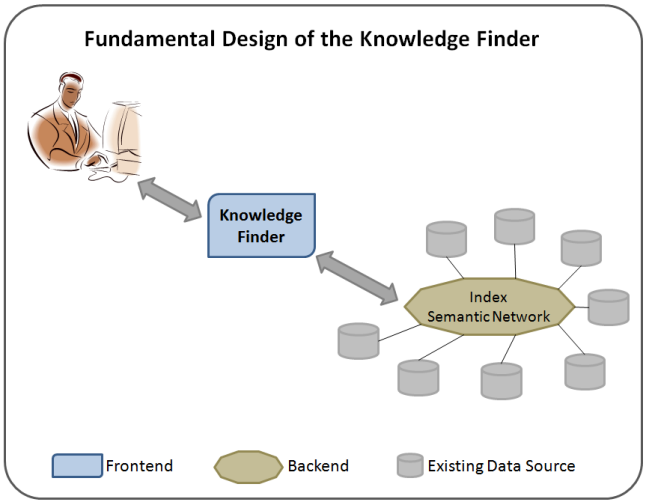

Fig. 2. - The fundamental design idea of the Knowledge Finder

Applying the fundamental design of the Knowledge Finder to the new review method mentioned before the following description of one of its main use cases could be described as follows

- The discussion in a review session stalls because a vital piece of information/knowledge is needed and none of the participants knows this piece of information
- The Knowledge Finder is triggered in order to answer the question "What (else) do we know about the issue being discussed?"
- The Knowledge Finder taps into all available data sources as a single access portal and presents its findings in a user friendly form thereby harnessing the flood of information
- The participants of the review either obtain the missing information at once or at least get a clue in which direction they should proceed with their discussion

2.2 Exploration of the idea

With this fundamental design idea in mind it was decided to apply a stage gate innovation management process to explore the potential of the Knowledge Finder while keeping risks at a minimum at the same time.

A steering committee consisting of business and IT managers across development and production department alike was formed to decide at each stage whether to proceed in the process or not (see fig. 3).

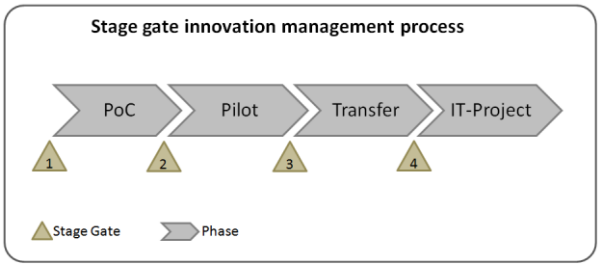

Fig. 3. - Stage gate innovation management process

The stage gate innovation management process was carried out as follows

- First a proof of concept was carried out with the intention to find out whether the idea of the Knowledge Finder would really create additional business value and would be accepted by the users. Note here that at this point a truly business-centric approach has been followed. No data sources were connected to the Knowledge Finder online in order to avoid efforts considered as too risky at this early stage.
- After an initial research for possible solutions a list of six IT-partners was compiled by a very small core team and proposed as participants in a first "Proof of Concept" (PoC) to the steering committee, which approved the selection for the PoC.
- Once the findings of the PoC in terms of additional business value and user acceptance turned out to be very positive, the steering committee approved the proposal to proceed with a pilot phase where eight different data sources (structured

data bases and unstructured data on a file server) should be truly connected to the Knowledge Finder.
- The core team then selected three IT-partners to be invited to the bidding process for the pilot solution. The winner of the bidding was then contracted to implement the Knowledge Finder as a pilot.
- Again the results of the pilot proved to be very positive and corroborated the earlier assessments of the Knowledge Finder. The steering committee consequently approved the proposal to proceed with the so called "transfer phase" which was intended to fulfill all the prerequisites for a formal proposal to initiate an IT-project to implement a productive solution.

2.3 IT-architectural considerations

While progressing along the stage gate process we also noted quite early on that the market for semantic search technologies is highly dynamic, i.e. the solution that seems appropriate today may not be the appropriate solution anymore in the future. Therefore we established two guiding principles:

- The solution we chose for the pilot would not necessarily be the natural choice for the prioritized IT-project, rather we would again look for the best suited solution available on the market by the time the IT-project would have been started
- The solution architecture must provide for the possibility to exchange the core semantic search technology while leaving the connectors to the various data sources unchanged (see fig. 4).

We designed a strict decoupling of the semantic engine together with its search index and the data sources with the corresponding crawler database views, such that at any future point in time the single investments in setting up the database views remain protected while the semantic engine is changed over to a new technology.

Furthermore we decided to decouple the **application** "Knowledge Finder" from the semantic engine such that other applications can reuse the semantic engine and the underlying search index for their purposes. This way we can both provide a semantic extension to a search engine inside a legacy system or provide other (new) applications (e.g. applications that have a focus on categorizing information) in need of semantic information processing with a common platform.

In fact, while being in the transfer phase another department (lab technology) requested to pilot the Knowledge Finder in order to assess possible improvements for their business process. And indeed, looking again at the market we found that a different solution provider could offer more added value to our Knowledge Finder (2.0) (see Table 1), which has been contracted for our second pilot.

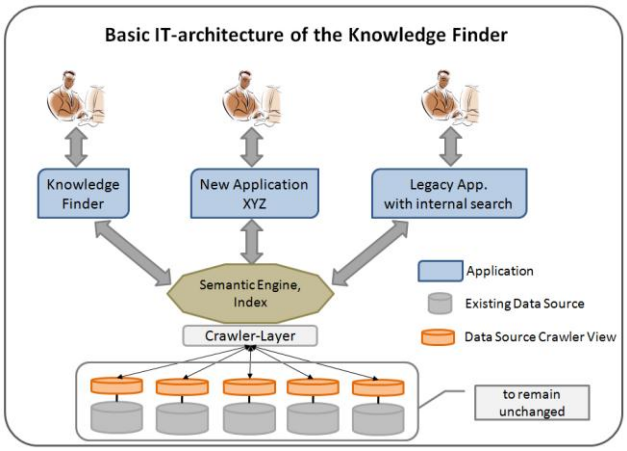

Fig. 4. - Basic IT- Architecture of the Knowledge Finder

Knowledge Finder 1.0	Knowledge Finder 2.0 (more added value)
Connecting structured and unstructured data sources	True semantic search
Transparent interactive and iterative search with backtracking	Bilingual search (German / English)
Fast localization of most relevant hits	Search for semantically similar phrases
Facetted search	Keyword in context search
First approaches to semantic search	Extraction of information objects from free text fields
Search for BMW-abbreviations and BMW-jargon	Suggestion – Autocomplete

Table 1. - The evolution of the Knowledge Finder

Figure 5 below gives an overview of what is possible with today's available semantic search technology. Given the search term "Plastik" Knowledge Finder 2.0 returns hits such as "Kunststoffanbauteile" and "plastic" demonstrating the capability of a semantic and bilingual search.

In addition Knowledge Finder 2.0 also extracts information objects such as "GS 97045-2" from free text, which denotes a so called "group standard" used at BMW that is mentioned within a lab technology report.

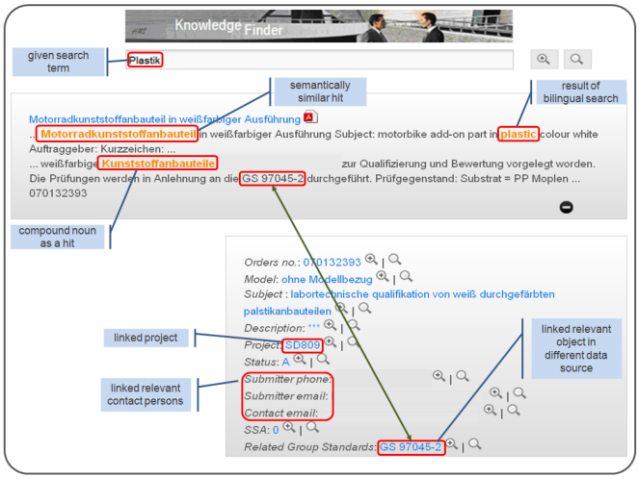

Fig. 5. - Features of semantic search

3 KEY FINDINGS

The two successful pilots resulted in a number of very interesting findings and observations which are outlined in the following sections.

3.1 User Experience

The possibility to ask a variety of different systems in a single consistent way rather than filling out different clumsy query interfaces obeying the internal data base schema was very well received by the users.

The Knowledge Finder's interactive and intuitive interface enables a search process where the human brain and the search engine work together in finding the best hits thereby also avoiding the "Google-Syndrome" of obtaining thousands of hits that a user is incapable of dealing with. In fact it has been found from a cognitive information processing point of view that it is much easier for the human brain to start with a single search phrase first and then to successively add more search phrases thereby reducing the number of hits very fast.

It has been found empirically that in the vast majority of cases the user could reach a list of less than 20-30 meaningful hits with no more than three steps (either adding another search term or selecting a filter value)

Apparently the list of found knowledge matching the first search term stimulates the brain to pick the next search term by association. There is no need for the user to come up with more or less complicated boolean expressions to define the search, which in many cases cannot be stated anyway in the first place since the user is not necessarily 100% clear about what he is really searching for in the beginning.

3.2 New Opportunities

Finding experts on specific issues has become much easier than before as the Knowledge Finder helps employees and organizational units alike in extending and strengthening their networks.

Particularly in light of the increasing fluctuation of employees this opportunity cannot be valued highly enough, because for a knowledge worker it is essential that the cost of obtaining missing information is minimized.

The Knowledge Finder allows to ask questions or at least to explore possible solution spaces which has not been an option before because of the high hurdles to retrieve information hidden in all the various data sources.

Another new opportunity that the Knowledge Finder may offer is the possibility to address the challenge to put all information islands together mentioned by Randy Mott, CIO of General Motors in [3] (see also introduction). By pulling information from various data sources together into one single search index, the Knowledge Finder might open up a new road to circumvent all the challenges and problems of classical tedious system interfaces, such that the search index could be viewed as a new form of information bus.

3.3 Challenges

Of course, piloting the Knowledge Finder has revealed a number of challenges, some of which could very well be characterized as fundamental in terms of how do we manage information as a very valuable asset in a large enterprise in general. Within the scope of this paper three aspects of managing large quantities of information (quality, protection, and comprehension) shall be addressed and explained briefly why they are important.

3.3.1 Quality of Information.

It is common knowledge that for a communication to be successful it is necessary that sender and receiver have the same understanding of coding and interpreting information. If for example the German word for “expanding rivet” is written incorrectly in the data source (i.e. “Spreitzniet” instead of “Spreizniet”) then that piece of information is not found when entering the correct spelling even drawing upon semantic search technologies. In fact during the lab technology pilot we both found hits when searching for “Spreitzniet” (incorrect spelling) and searching for “Spreizniet” (correct spelling) but the sets of hits were disjoint!

In much the same way no semantic search technology today would be able to discern a variety of different ways to abbreviate (longer) words within free text (e.g. “homologation” = “hom.” = “homol.” = “homolog.”).

Most search engines today include a spell checker with the auto suggestion function when entering a search term but they cannot cope with misspelled words in the data sources. Even worse, a spell checker applied to “Spreitzniet” would change the search term to “Spreizniet” thus missing (possibly) the most valuable information hits associated with the misspelled word “Spreitzniet”.

Another type of lacking quality in information showed up in our first pilot. One of the test users turned very excited when he found a long missing Powerpoint-document only to turn disappointed when opening the document and discovering that its content didn`t match the search term at all. After a short investigation of this "false positive hit" it became clear that the meta information of the Powerpoint -document did match with the search term but by no means coincided with the content. What happened was, that (and that is typical) the Powerpoint-document originated from a template whose author and title had not been changed appropriately by the author of the found Powerpoint-document.

From a more general point of view the root cause of these observations can be explained by taking the approach that we are actually dealing with a flow of information.

With regard to a flow of material there exist a number of quality assurance techniques such as value stream analysis, (Ishikawa) fishbone diagrams or the application of Poka-Yoke principles in order to prevent wrong assembling of parts. For flows of information however, we do not possess corresponding quality assurance techniques nor have we transferred and adapted the well proven techniques for flows of material to flows of information.

One possible approach to do so could be to apply the so called DMAIC -cycle known from the six sigma toolset as indicated in [4] and [5]. However this is quite a generic practice which would lack specific tools in our context.

One such tool or best practice could be that the moment information is originally entered into any data source a spell checker automatically corrects input. This way we would also apply one of the lean production best practices known as "responsibility on site", meaning that any error that might occur should be corrected instantaneously right at the location where it occurred.

3.3.2 Protection of Information.

Naturally the idea of an all encompassing semantic search that links across various data sources triggers concerns with regard to intellectual property protection. In fact, in almost all data sources rules are in place to discern users that are allowed to read particular information from users that are not. Today it is common practice to grant (read) privileges on a "need-to-know-basis". However, we have found throughout the pilots that this principle has a number of disadvantages, of which some are listed below

- Once a read privilege is granted on a personal basis in most cases it will never be removed, i.e. the read privilege might still be in place even though the personal need is not given anymore
- Granting privileges on a need to know basis can be tedious, in particular if the number of information objects to be protected or the number of users is large. (we came across a situation where about 800 different folders on a file server needed to be overseen resulting in a full time job to do all the administration)
- Typically read privileges on a file server can only be granted at the level of folders. Hence if we grant read privileges to allow a user to read a particular file in that

folder we inadvertently give read privileges to other files residing in the same folder for which that particular user should not have read privileges.
- Last but not least the responsibility to grant any kind of privilege does not lie with the owner of the information object but rather with the (system) administrator, who even might not be employed by the enterprise

Therefore one of the key findings of the pilots was to propose a different principle to protect information which we call the "need-to-protect-principle". With this principle we strengthen the role of the information owner in the first place. The basic idea is that the information owner decides whether a certain piece of information can be read by anyone within the enterprise or shall only be visible to a restricted circle of users. In the latter case this information is not even presented to the Knowledge Finder, i.e. it is never been indexed and therefore cannot be found at all.

If, on the other hand, certain information can be read by anyone inside the enterprise then a fair balance of interests can be established between the urgent need to share information for better collaboration within the enterprise and the need to protect (strictly) confidential information.

As a consequence the unproductive efforts of granting read access on a need-to-know-basis are minimized if not eliminated.

Finally applying semantic technologies it is well conceivable to scan documents in order to help classify them into the right protection level.

3.3.3 Comprehension of Information.

Piloting the Knowledge Finder we soon discovered that extending the scope of data sources beyond those known to the user (compare fig. 1) it is likely that users might get overloaded with the sheer number of information objects and their interdependencies.

In other words the Knowledge Finder gives rise to unveiling the complex information landscape that we inherently create while progressing within our product development process today.

First of all, making interdependencies in a complex product (recall [1] in the introduction) visible is e.g. highly advantageous to an engineer who needs to validate a complex function of a complete vehicle and cannot do this anymore because he is lacking a physical prototype.

On the other hand (see fig. 6) it is questionable whether such a complex information landscape is always correctly mapped to the user`s brain so that the user is having the correct understanding of all the information and thus being able to draw the right conclusions.

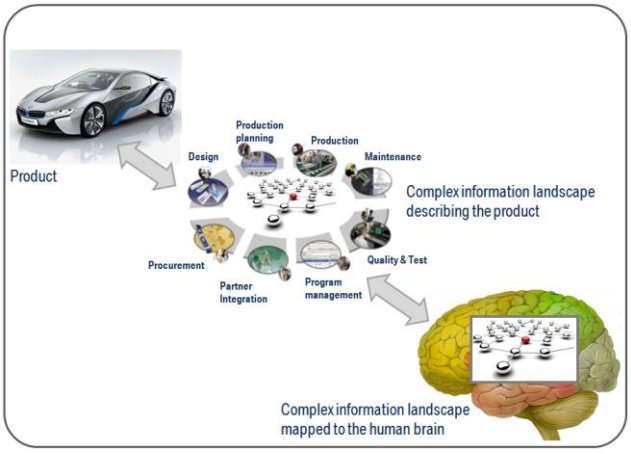

Fig. 6. - Complex information landscapes vs. processing capabilities of the human brain

4 CONCLUSIONS

In this paper it was shown how the increasing need to find buried information more efficiently again has led at BMW to the development of the Knowledge Finder and what road has been chosen to do so. During two pilot phases in two different disciplines (domains) many very valuable findings have been collected that influence the design of the next evolution of the Knowledge Finder for productive use.

In order to address the challenges mentioned before (quality, protection and comprehension of information) further research is necessary resulting for example in advancements with respect to semantic technologies.

As far as the aspect of comprehending complex information landscapes is concerned it is believed that new cross-discipline input from cognitive psychology and neuro sciences is needed in order to answer the following two questions.

- What is the appropriate design of new man-machine interfaces such that a minimum level of stress is imposed on the user when exploring complex information landscapes?
- What is the appropriate design of new user trainings so that the brain can comprehend complex information landscapes better?

These questions cannot be addressed by computer science alone. Rather an interdisciplinary approach together with cognitive psychologists and neuroscientists is needed that could even lead to a new branch of computer science that might be called “Psycho-Informatics” in the future much like “Bio-Informatics” today.

Such an interdisciplinary approach should look at new forms of interaction with information landscapes that go beyond the currently prevailing paradigm of interacting with information by screen, keyboard and mouse clicks.

It shall also be noted here that the further development of the Knowledge Finder will be a continuous ongoing process with adding more data sources and functional improvements and the design of the user interface step by step. We aim actually at offering the Knowledge Finder and its steady improvements as a service rather than rolling out software releases from time to time.

The idea is "Finding knowledge as a service" where suggestions from our users are added in an agile manner.

5 REFERENCES

1. Siemens PLM: „Herstelleralbtraum: Ein Produkt mit Funktionsfehlern"
   in: interface - Das Magazin für Product Lifecycle Management, 15. Jahrgang 1-2012
2. N. Taleb : "Der Schwarze Schwan", Seite 17 DTV 34596, 2. Auflage 2010
3. A. Bongard: Interview with Randy Mott, CIO General Motors
   in: automotiveIT, Ausgabe 10/2012
4. T. Schmidt: „Systematisches DQ-Management: den Daten bloß nicht immer hinterherlaufen"
   in: BI Spektrum, Ausgabe 5 / 2011
5. T. Schmidt: „Der DMAIC-Zyklus für ein systematisches Datenqualitätsmanagement"
   in: BI-Spektrum, Ausgabe 4 / 2012