

Vadász Pál

vadasz.pal@montana.hu

CASE STUDY FOR MEASURING THE FEASIBILITY OF A SEMANTIC SEARCH SYSTEM

Absztrakt/Abstract

Egy szemantikus keresőt alkalmazó tesztrendszer került kifejlesztésre a magyar bírósági rendszer számára. A bírósági rendszer munkafolyamatainak megvizsgálása során bebizonyosodott, hogy a mesterséges intelligencia a hatékonyságot nagymértékben megnöveli.

A test system was developed in the Hungarian judicial system using semantic search technologies. The working process of the judges was examined. It has been established that applying this kind of AI the efficiency of the judicial process would increase significantly.

Kulcsszavak/Keywords: *szemantikus keresés, mesterséges intelligencia, Magyarország, bírósági rendszer ~ semantic search, artificial intelligence, Hungary, judicial system*

BACKGROUND, HYPOTHESIS

In Hungary, the culture of semantic search is not only lagging behind that of the North-American and old EU countries by at least 15 years, but is also behind that of some neighbouring countries by 3-10 years. For the Hungarian multinationals, the hegemony of the foreign parent companies has had a profound effect on the domestic high-tech applications. Sensitive technologies are often installed and operated from outside Hungary. Large national companies often suffer from the absence of technological vision, whereas SME's lack the financial resources. In my view, the use of search technologies is the most prevalent in the public sector, though with enormous variations in the level of sophistication.

Indecisiveness characterises all sectors due to the fact that the decision-makers are not familiar with the return of their planned investment. The aim of this article is to present a simple financial model using an existing test system. The calculations are based on measurements and estimations gathered from the Hungarian legal system. This sample provides a basis for further modelling. Decision-makers can tailor their search system to meet their requirements by inserting parameters relevant to their organisations. The outcome of the equations and the resulting diagrams immediately show the justification of their future decisions. Although the model describes a specific environment, it can also be applied to any other, where the information needs to be extracted from a large volume of unstructured data.

In order to operate an organisation effectively, the appropriate management of the knowledge base accumulated is just as critical as the efficient operation of the work processes and the optimal cost structure. The private sector is measured on profitability whereas in the public sector the term effectiveness is more appropriate. An extensive amount of literature is available on this differentiation; however, discussing this would go beyond the scope of this article.

80% of the digitalised information in the world is in unstructured formⁱ. Word and PDF files, portals, databases, emails and many other formats of unstructured data store information. Sifting through unstructured data to find specific information is akin to searching for a needle in a haystack, so in order to make efficient use of one's time, the use of artificial intelligence as a magnet in the form of semantic search is not only increasingly necessary, but inevitable.

It is worth noting, that it is particularly difficult to handle the Hungarian language when employing statistical methods, due to the agglutinative nature of the language as opposed to Germanic, Latin or Slavonic languages. While the different expressions using prepositions in the latter languages show a linear growth, the variations in Hungarian using suffixes produce an exponential curve.

My hypothesis is as follows: applying semantic search systems can significantly increase the operating efficiency of any organisation. This is particularly the case in professional areas where the processing of vast quantities of unstructured text is vital.

While preparing this article we carried out a number of interviews with renowned judges of the Metropolitan Court, Second Instance. The estimates have been formulated on the basis of their professional views under the supervision of Dr. Zsuzsanna Mohácsy, member of the Metropolitan Court, Second Instance. Quoted data comes from the official statistics of the Hungarian court system.ⁱⁱ The tables underneath have been collated by myself, they have not been published so far.

PRESENTATION OF THE SYSTEM

Professional environment

The selected example concentrates on the judicial activity. This is a special field where a semantic knowledge management tool can increase operating efficiency to an especially high standard, which is hugely advantageous since judges work with thousands of legal documents in their everyday work.

Overview of the system

The eCourt knowledge management system was prepared in co-operation between the Metropolitan Court and Montana Knowledge Management Ltd. in Budapest, Hungary. The Court provided the legal know-how, particularly in building the ontology, which includes the taxonomy and the rules that define the context related meaning based interrelations. Montana provided the whole search system, and the knowledge workers, who translated the human intelligence into machine-readable format for the artificial intelligence. Junior judges with a strong inclination towards IT were trained to operate the software with the view of transferring their know-how to more senior staff on a train-the-trainer basis.

6000 anonymised court decisions were indexed and then made available for semantic search carried out by the judges. The judges have been extremely enthusiastic about the meaning based operation as opposed to the character based system with which they have had to put up with for so long. It offered them far more flexibility in defining their queries, since no exact character strings had to be given as input. Synonyms, antinomies, and somewhat vaguer formulated questions were understood and responded to by the robot with a much higher hit rate and less irrelevant junk.

Operation of the system

The eCourt knowledge management system ranks the search results on the basis of content relevance, indicating the application of different set rules, and the content-based relevance of the results in relation to the query input. Furthermore, the system ensures that the assessment of any document (e.g. claims, lawsuit materials, decisions) can be performed by a competent employee of the court efficiently. The system automatically offers the previous decisions and lawsuit materials which are similar to the specific case without a further search query.

The system indexes the decisions, the digital lawsuit and legal materials, and then processes them electronically by the help of the developed specific knowledge trees (taxonomies) and context defining rules, i.e. the ontology.

In addition to the free word or similarity search, the system is able to search for a document or document group on the basis of the case number, the list number, the court, topic, date, legal regulations and other metadata.

The analysis of the decisions and lawsuit materials can also be used for making electronic summaries and excerpts enhancing quick results during search.

Data sources

The most important data source of the judicial system is the collection of court decisions in digital form. For the effective operation of the knowledge management system, other documents and lawsuit materials shall also be included.

The following text illustrates a court decision from the processing point of view.

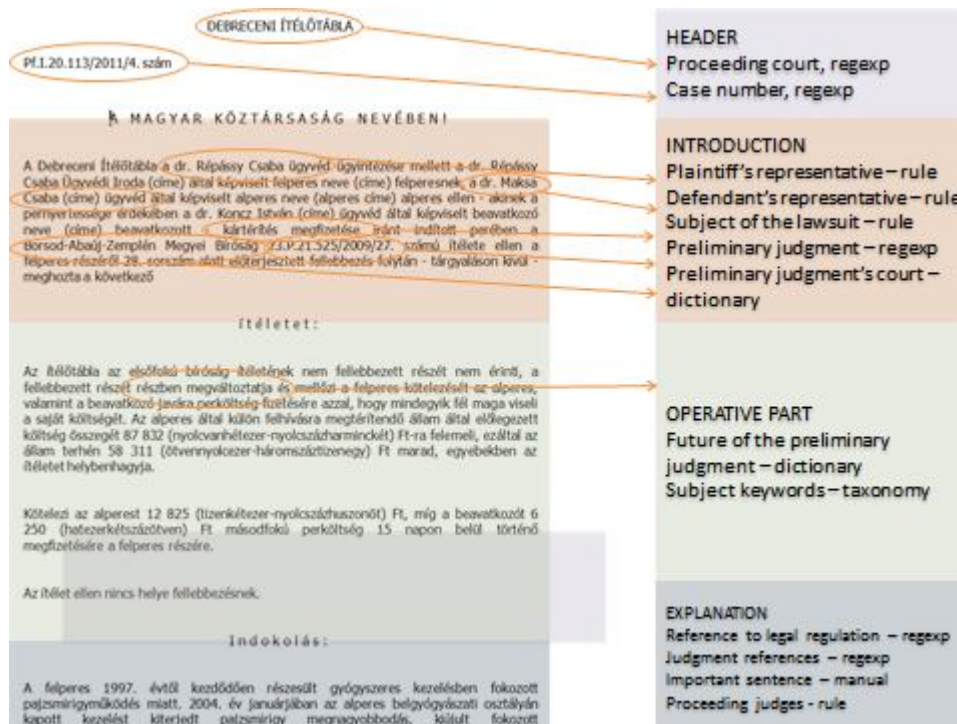


Figure 1. Court decision from the processing point of view.

Anonymisation

Pursuant to current law, court decisions may only be disclosed to the public in anonymised form. In practice, this means that the judge must remove all data from the text from which the identity of the parties can be derived.

The document processor performs the actions that are the prerequisites for the anonymisation procedure by using the appropriate complete linguistic modules. It is important to convert the various types of documents into a standard format, and separate into parts of standard content. The words of the documents may be present in different language forms. Due to this and the complex nature of word search principles for establishing word distances, the documents are indexed in a way that is suitable for text mining search engines. In relation to the different content units, the entities indicated there can be handled as separate data. The document versions prepared this way are then anonymised.

The following workflow diagram illustrates the process of anonymisation.

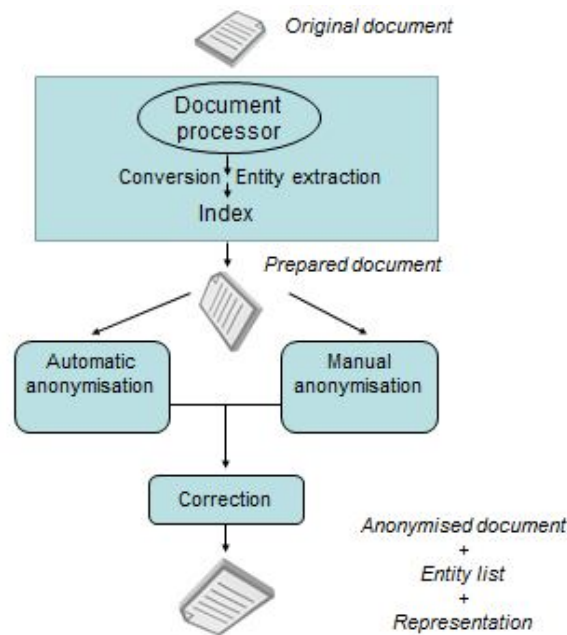


Figure 2. The process of anonymisation.

Ontology

The eCourt knowledge management system processes the decisions with the most up-to-date tools of text mining, by enforcing the legal professional aspects, partially in an automated way and partially by involving jurists. It also develops the taxonomies by the help of which the system supports the user in disclosing the correlations and judicial practice within the shortest possible time in the lawsuit materials and decisions.

The taxonomies allow for the automatic, thematic categorisation and pre-filtering of the lawsuit materials, and the monitoring of the changes and refining the results.

The essence of the professional system is provided by the included codes, thematic taxonomies and topic collections, as well as by the professional thesauruses and the related rules, that is, the whole ontology. The query language of taxonomies also works on the basis of similarity and regular expressions, and, navigating the tree, it also supports filtering and expansion type cross-queries.

The taxonomy is accessible by the user on a simple, user-friendly web interface.

The taxonomies of the effective Hungarian codes (Act on Business Entities, Civil Code, Criminal Code) were prepared in co-operation with the staff of the Faculty of Law of ELTE University, Budapest.

The following two pictures illustrate some of the Criminal Code taxonomy (left) and a set of search commands using Boolean operators (right). Underneath are the search results.

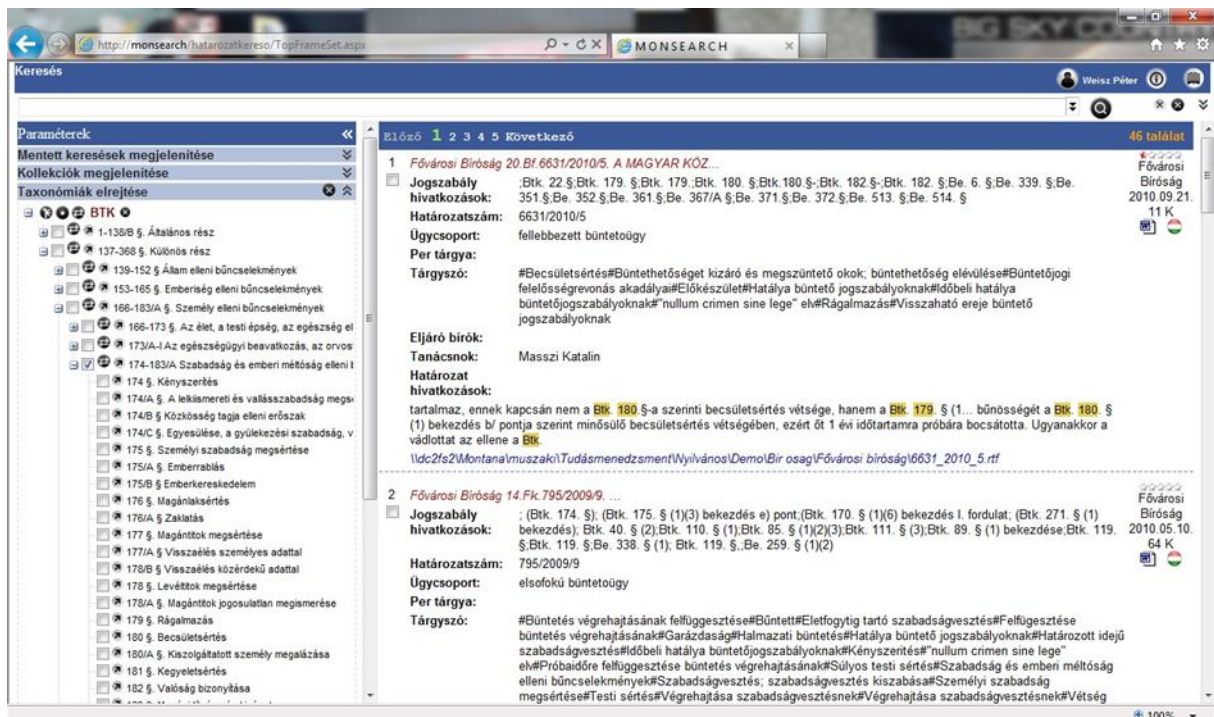
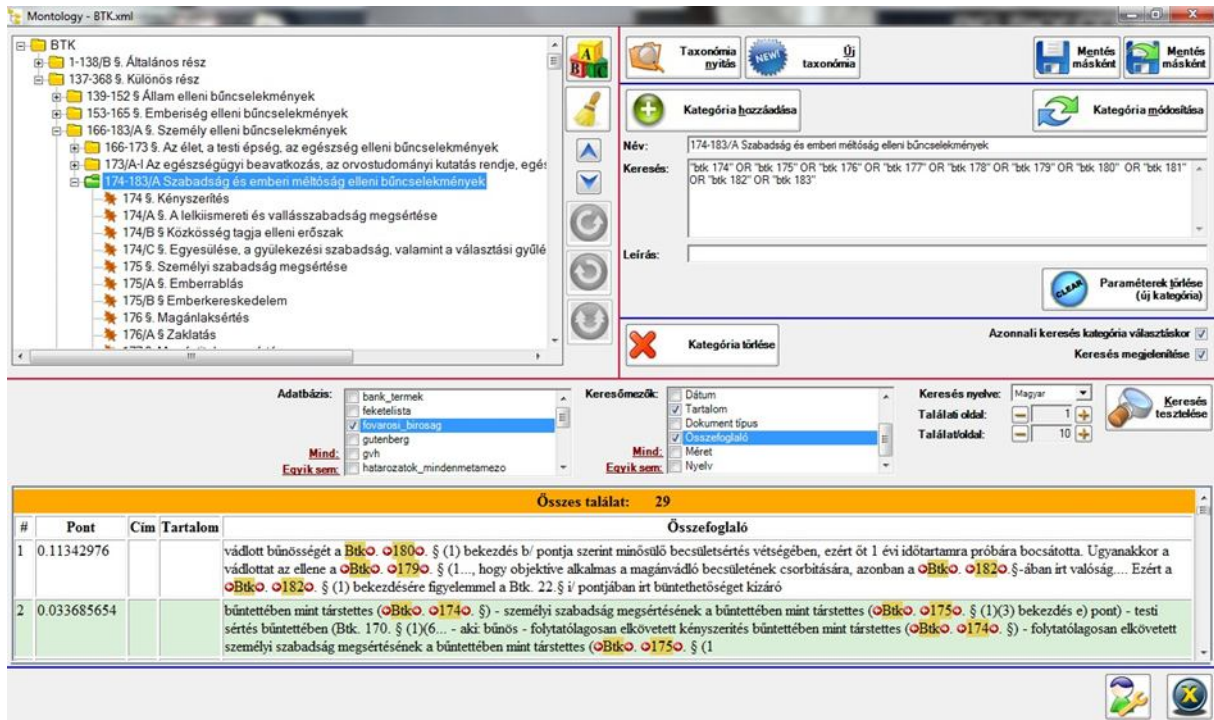


Figure 3. The search results.

REPRESENTATION OF THE EFFICIENCY ENHANCEMENT

Efficiency aspects

The operating efficiency of the judicial system is evaluated on the basis of the following criteria by the profession:

- Duration of legal proceedings
- Practice of headcount management

- Quality of verdicts
- Ratio of backlogs
- Appropriate application of administrative tools

Judicial activity

On the basis of practical experiences and interviews conducted with judges the activities of judges can be grouped as follows:

- Preparation of court hearings:
 - Analysis of lawsuit materials
 - Revision of court decisions and practices similar to the specific case
 - Analysis of professional opinions
 - Searching and considering relevant legal regulations and acts related to the specific case
- Conducting hearings
- Meetings
- Participation in professional forums
- Documentation
- Preparation of statistics
- Checking anonymisation

Assumptions and comments

By analysing the work process of the judges and classifying their activities, it can be established that, without the semantic search tool the judges spend an estimated 23.50% of their working time on performing different searches, and an estimated 32.50% of their working time is left for analysis. The sheet *The distribution of the judges working time* shows the whole breakdown.

A fair assumption is that the use of a semantic search system reduces the search time to one tenth of the original duration of manual search. Therefore, the presented 23.50% search time decreases to 2.35%. See the *Overall cost analysis of the time spent by judges* table!

The development of the entire system requires one year, and the gradual national implementation and deployment would require an additional year. This latter is performed in three phases. The rate and scheduling of implementation can be seen in the last three lines of the *Return of the investment* table.

Only the decrease in time spent on searching is examined in terms of payback, but the increasing efficiency of analysis is disregarded which suggests that there is further hidden potential.

On the basis of the conducted interviews, the aforementioned time spent presents average, conservatively estimated values.

The 125 knowledge managers presented in the tables do not necessarily involve the hiring of new employees; this can be achieved by training the existing employees within the frameworks of the judicial training.

The distribution of a judge's working time

The distribution of a judge's working time							
Activities	Main activities in % of working time	Distribution of the main activities to detailed tasks	Detailed tasks broken down into subtasks	Total time spent on search	Total time spent on analysis	Total time spent on other tasks	
Preparation of hearings	20%	50%	analysis of lawsuit materials				
development of the judicial, evaluation criteria and searching for related entities			70%	7,0%			
discovering correlations (search)			15%	1,5%			
reading lawsuit materials, analysis of correlations			15%		1,5%		
Court of revision decisions, practices similar to the specific case		20%	20%	searching for cases	20%	0,8%	
analysing cases				80%		3,2%	
Analysis of expert opinions				20%	70%	searching for facts relevant for ordering experts, disclosing other comparable expert materials	2,8%
comparative analyses		30%				1,2%	
Searching and considering legal regulations, acts relevant to the specific case		10%	20%			search time	0,4%
evaluation				80%		1,6%	
Conducting hearings				30%			
Meetings		5%					5,0%
Participation in professional forums		5%					5,0%
Preparing and reviewing documentation	30%	20%	Preparation of facts, entities to be documented		6,0%		
Preparing documentation			70%		21,0%		
Checking documentation			10%		3,0%		
Preparing statistics	5%	50%	Searching information		2,5%		
Evaluating information			30%		1,5%		
Compiling statistics, statements			20%		1,0%		
Checking anonymisation	5%	50%	Searching for entities to be anonymised		2,5%		
Performing and checking anonymisation			50%		2,5%		
Total				23,5%	32,5%	44,0%	

Overall cost analysis of the time spent by the judges

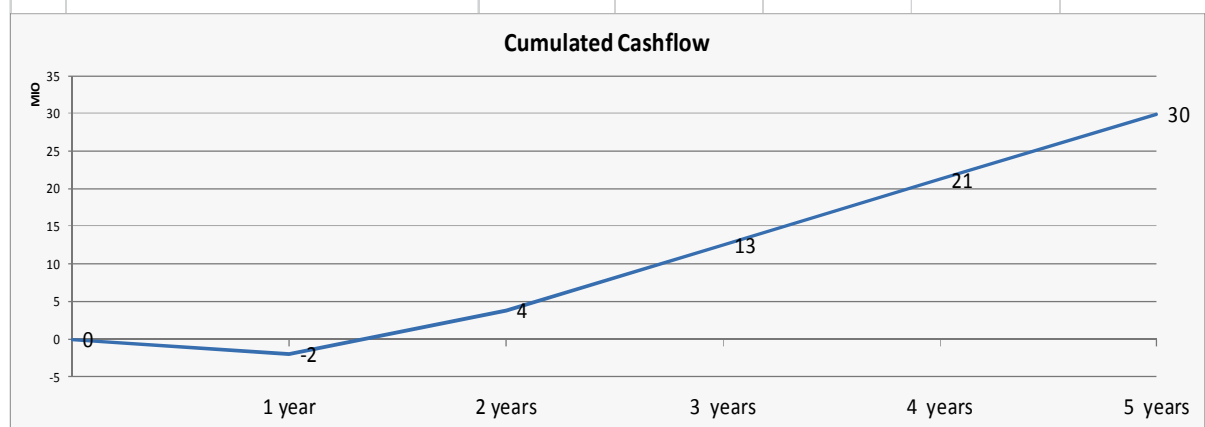
Overall cost analysis of the time spent by the judges		
	Manual search	Semantic search
Time spent on searching in % of working time	23,50%	2,35%
Judges' headcount (person)	2 936	
Monthly average gross salary/judge (EUR)	1 167	
Total gross annual personnel expenses of judges (EUR)	41 104 000	
Total cost spent on annual search (EUR)	9 659 440	965 944
Savings (EUR)		8 693 496
Savings in terms of judicial work (in man/woman years)		621
Rate of efficiency increase		21,15%

The court's course of work in numbers

The courts' course of work in numbers			
	Received	Completed	Still open
at local courts			
Local courts civil	161 335	164 702	64 807
Local courts economic	13 881	15 414	5 596
Non-litigious civil and economic lawsuits	64 328	66 087	2 820
Criminal lawsuit	77 980	82 676	48 752
Criminal lawsuit - public prosecution	61 510	66 337	42 894
Criminal lawsuit - private prosecution	15 569	15 472	5 196
Minor offences	107 276	106 910	16 366
county courts (of second instance)			
Civil	18 850	18 587	5 489
Economic	2 393	2 192	852
Non-litigious civil and economic lawsuits	13 697	13 212	2 245
Criminal lawsuit - public prosecution	12 472	12 033	5 592
Criminal lawsuit- private prosecution	676	566	288
Minor offences	1 159	1 143	46
Total	550 450	564 765	200 655
Annual national average/judge	187	192	68
Efficiency increase in case number		119 448	
Efficiency increase in case number %		21,15%	

Return of the investment

Return calculated with overall costs (incl. VAT)					
Definitions	Value	Unit	Ratio		
Required knowledge manager headcount	125	persons			
Gross additional wage of knowledge managers	750 000	EUR/year			
Costs of equipment (infrastructure)	973 667	EUR			
Service costs	1 247 608	EUR			
Total investment	2 221 275	EUR			
Development of the "Service provider Court" application	12	months			
National launch of the "Service provider Court" application	12	months			
Schedule of national launch expressed in % of the number of completed cases:					
phase 1	1	months	26%		
phase 2	5	months	35%		
phase 3	6	months	39%		
Costs and savings	1 year	2 years	3 years	4 years	5 years
Savings deriving from the rationalization of judicial works	0	3 846 872	8 693 496	8 693 496	8 693 496
Wage of knowledge managers	-750 000	-750 000	-750 000	-750 000	-750 000
Other administration costs	-300 000	-300 000	-300 000	-300 000	-300 000
Software monitoring	0	0	110 067	110 067	110 067
Operation, support	0	0	-203 200	-203 200	-203 200
Investment costs	-973 667	-1 247 608			
Depreciation		-740 425	-740 425	-740 425	
Marketing costs of launch		-125 000	-104 167		
Service provider Court – national savings per year	-2 023 667	683 839	6 705 771	6 809 938	7 550 363
Payback	1 year	2 years	3 years	4 years	5 years
Annual profit or loss	-2 023 667	683 839	6 705 771	6 809 938	7 550 363
Cumulative cashflow (CF)	-2 023 667	3 846 872	12 540 368	21 233 864	29 927 360



CONCLUSIONS, SUMMARY

Payback

Although the investment consultants accept the above “hockey stick”-type payback curves with strong scepticism, I have not found any factor that would contradict the above calculations. It would be the greatest recognition of this article if the experts and politicians considered it as material for discussion, added their own corrections and comments, enabling the legal IT to advance utilising this constructive change of ideas.

Knowledge transfer

By the help of the system, the professional knowledge database containing the historic decisions and comments would be available to the other active or younger judges, and thus knowledge transfer would be ensured. Additionally, it is important to support the management supervisory activities to further ensure efficiency.

Political importance

Increasing the efficiency of judicial work would not only mean enhancement in narrow professional circles, but it may also have political importance in two areas. On the one hand, faster decision-making would enhance investor enthusiasm and confidence. On the other hand, the consistent judicial practice would increase the trust of the whole population in the legal system.

If the time spared by the help of the eCourt knowledge management system were spent on professional work, the creative working time could be increased from 32.50% to 53.65% expressed in the percentage of total working time, which, among other benefits, would result in the decreased duration of lawsuits.

In terms of the decreased duration of lawsuits, the efficiency increased by the eCourt knowledge management system would be the same as if further 621 judges were involved at the national level.

Based on the statistical figures of 2011, the national average of finalised lawsuits by the local and county (second instance) courts is 192 cases/judge. Considering this, the involvement of 621 further judges would result in the theoretical completion of further 119,448 cases per year (see *The courts' course of works in numbers* table!). This would mean an approximate increase of 21.15% in finalised lawsuits.

At the courts, a specialised knowledge management system based on semantic search:

- would increase quality significantly by freeing up 21.15% more working time left which can be spent on effective work,
- would increase the number of closed lawsuits up to 21.15%,
- would see a return on investment *within 8 months* of the national launch,
- would ensure knowledge is transferred,
- would effectively support the creation of a standard judicial practice,
- would support the public administrative management of courts.

References:

1. Christopher C. Shilakes, Julie Tylman: Enterprise Information Portals, Merrill Lynch, 16. November 1998
2. www.birosag.hu, cumulated numbers from the 2011. statistics

Figures:

All figures created by the author of this article.

ⁱ Christopher C. Shilakes, Julie Tylman: Enterprise Information Portals, Merrill Lynch, 16. November 1998

ⁱⁱ www.birosag.hu, cumulated numbers from the 2011. statistics