Our world is illustrated

.. and full of sounds and smells … and less and less transparent

Boris Škandera
Informetal, interest association
VÚHŽ, 739 51 Dobrá, Czech Republic
informetal@vuhz.cz

INFORUM 2010: 16th annual conference on professional information resources
Prague, 25th – 27th May 2010

Abstract

Significance of non-textual information. Brief evaluation of development of databases from text-oriented to multi-media databases. New possibilities offered by the so called "deep indexing". Increasing importance of full-text information sources. Role of digitisation in daily life and professional life. Impact of internet on development of specialised databases. Problems created by the current Babylon. Reflections on trends of possible future developments of information processing and searching.

Importance of visual information in human life

Perception of richness of external stimuli of the world, in which we live, is limited by our senses, i.e primarily by eyesight, hearing, smell, touch and taste. Many opinions as to what was the decisive factor for principally different development of human being as a specie from development of other living creatures. Practically all animate beings are able of mutual communication at various degrees of development (and in most cases we have not yet been able to understand it), so this probably is not the decisive distinguishing feature. It appears that we can observe comparatively high level of intelligence in animals, and also that man is not the only being capable of creating necessary tools. Development of speech was undoubtedly of paramount importance, but in my humble opinion the key role of development of mankind was played not only by ability of mutual communication, it means in conveying various information, but first of all the ability put down this information, that it to conserve it in time for possible later and repeated use. Historians assume on the basis of existing level of knowledge that human beings developed speech approximately 200 000 years ago.

Brief history of illustrations – visual information

It seems that everything started with first drawings, from which only those have survived till present time, which were made on high quality durable materials, i.e. particularly petroglyphs and paintings made on rocks or stones. Historians estimate that first still preserved drawings of this kind started to appear approximately 30 000 years ago. Human communication thus gained fundamentally new dimension, as original information became independent not only on its originator, but also on time and later on also on the place of creation. These images became another “speech” of human beings and they were gradually used more and more often. It can be therefore said that primarily visual information represented the beginning of recording of historical memory of humankind.

Those images later on served at various places of the Earth as basis for various symbols, from which alphabets, which we use nowadays, were developed. Alphabets are therefore in fact an abstraction of prehistoric images, the original sense of which we do not perceive anymore. This abstraction lead to creation of what we call today text information. Origins of systems of writing date back to the period of approximately 7 000 years ago.

Separation of text from images brought several advantages, namely uniformity and possibility of easier processing of text information, i.e. its ordering, sorting, etc. than processing of visual information. However, role of visual information did not perish by this, on one hand we are surrounded by abundant visual information literally wherever you, on the other hand visual information continues to be very important part of text information. On the contrary – modern technologies helped further enormous development of visual information. Art objects in past were mostly created by talented people. Thanks to tremendous development of photographic technology, video-technology and computer graphics, accompanied by its reduction of its prices practically anyone can express himself or herself graphically (i.e. photo-graphically or filmo-graphically), he or she even need not have truly professional equipment, often an ordinary mobile phone is sufficient.

Historically very recently – in 19th century - recording of image and text was completed by recording and therefore processing of sound. This dent could be briefly summarised into the sentence „from Altamira, passing by historiated manuscripts and comics to logos, television and internet“.

Recent history of text processing

Incredibly steep development of text information occurred during in the past decades. Similarly as we in our childhood listened open-mouthed to our parents or grandparents when they recalled their writing with chalk on slates at school, we today also look out-of-date when we talk to our offsprings about hand maintained card indexes in libraries and about hand made excerpts from books. Development of electronics intensively accelerated simultaneously both processing and transfer of information.

As recently as only 25 years ago we had miraculous helpers at work – computers of the type Sinclair Spectrum or Atari with operating memory of 48 MB, the fastest communication medium was teletype and later fax, which manages to transfer almost instantly black and white images of rather peculiar final quality.

When looking back we shudder at the admiration of human inventiveness and braveness, when we realise that the first landing of man on the Moon in 1969, i.e. 40 years, was realised with use on-board computer with capacity similar to that of already mentioned Sinclair Spectrum. It makes no sense to describe the present situation – we all know it and moreover it will be quite different already by tomorrow.

Significance of databases and current trends of their development

Texts are today technically created, processed and spread in a much easier manner. There are so many databases of various kinds that it is impossible to find them without database of databases. Anything published on the internet is immediately available to anyone anywhere in the world. Internet is therefore flooded by unimaginable volume of information – in spite of this the biggest search engines manage to process them practically in real time in such a manner that they are able to provide information that suits in very precise manner to the query.

Recent novelty – deep indexing

Processing of text information still plays dominant role, however, processing of non-text information is continuously developing. Let us have a look at several examples, which can at the same time indicate the possibilities or direction of future development of information processing. The company CSA (nowadays named ProQuest) introduced in 2007 on the market a revolutionary novelty – database IIlustrata.

Until then the records in bibliographic databases contained apart from bibliographic data namely abstract and keywords = descriptors. Purpose of abstract was to express briefly the content of the abstracted article. Non-text information was in fact mentioned only in bibliographic part, usually in the form „6 tables, 4 figures“. This means that user got hold of valuable information contained in tables, graphs, figures, photos, etc. only when he obtained full text of the abstracted article.

Database Illustrata introduced so called „deep indexing“, it means that is focused specifically on indexing of information contained in those „non-text“ parts of publication. Statistics show that technical texts contain at the average 10 illustrations of the type mentioned above, and each such illustration can be described at the average by 7 keywords.

What does this mean in practice for the user:

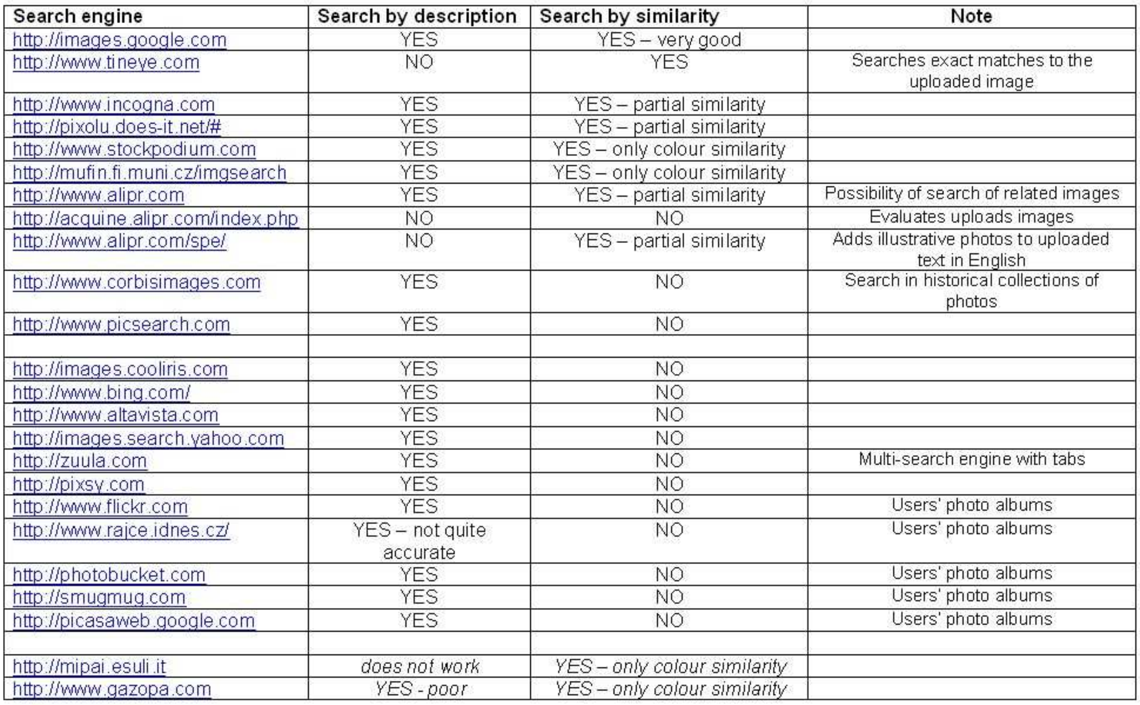
- number of keywords describing the abstracted article has substantially increased and therefore also precision of search has increased
- in this manner users gained access to information contained in the article, which were previously ignored by the processing focused on text only
- records show thumbnails of all illustrations, so that user can immediately judge their suitability for his needs
- internationally understandable illustrations made available in this way help to overcome in great extent language barriers

To summarise all this it can be said that in this manner the users obtain already at the level of secondary information all the richness of the given publication, including previously hidden content of its non-text parts.

Present role of visual information

There is good reason for saying that an image speaks for hundreds of words. As consequence of development namely of television, internet and digital photo and video technology the number of images in our life hugely increased. If we take newspapers as an example, they were first published as purely text information, later illustrations were added replaced afterwards by photos – all this first in black and white and then in colour. Nowadays practically all the dailies and weeklies have their continuously updated web sites, completed with audio or video clips. Simultaneously we witness drop of sales of classical „paper newspapers“.

It is necessary to enable search in this abundant offer of visual information. Main search engines, such as Google, Yahoo, Bing, AltaVista, etc. offer search of images (though based on their text description). Apart from this we have search engines oriented specifically on images, such as Flickr, Picasa, Cooliris, Pixsy, Photobucket or Czech Rajče (Tomato) and others. Some of them have comparatively recently introduced also possibility of search based not only on text description, but on visual information itself – on the basis of similarity with the initial input image. The following search engines are examples of such possibilities:



Search engine	Search by description	Search by similarity	Note
http://images.google.com	YES	YES – very good	
http://www.tineye.com	NO	YES	Searches exact matches to the uploaded image
http://www.incogna.com	YES	YES – partial similarity	
http://pixolu.does-it.net/#	YES	YES – partial similarity	
http://www.stockpodium.com	YES	YES – only colour similarity	
http://mufin.fi.muni.cz/imgsearch	YES	YES – only colour similarity	
http://www.alipr.com	YES	YES – partial similarity	Possibility of search of related images
http://acquine.alipr.com/index.php	NO	NO	Evaluates uploads images
http://www.alipr.com/spe/	NO	YES – partial similarity	Adds illustrative photos to uploaded text in English
http://www.corbisimages.com	YES	NO	Search in historical collections of photos
http://www.picsearch.com	YES	NO	
http://images.cooliris.com	YES	NO	
http://www.bing.com/	YES	NO	
http://www.altavista.com	YES	NO	
http://images.search.yahoo.com	YES	NO	
http://zuula.com	YES	NO	Multi-search engine with tabs
http://pixsy.com	YES	NO	
http://www.flickr.com	YES	NO	Users' photo albums
http://www.rajce.idnes.cz/	YES – not quite accurate	NO	Users' photo albums
http://photobucket.com	YES	NO	Users' photo albums
http://smugmug.com	YES	NO	Users' photo albums
http://picasaweb.google.com	YES	NO	Users' photo albums
http://mipai.esuli.it	*does not work*	*YES – only colour similarity*	
http://www.gazopa.com	*YES - poor*	*YES – only colour similarity*	

Moving images - videos

Similar example is publication of videos and search in them, which was tremendously popularised particularly by YouTube, which has many followers and clones. Its popularity was undoubtedly helped by its ability to offer to the viewed video also a set of similar or related videos. This search is naturally based on text description, i.e. on indexing of each video recording, not on similarity of visual information of the video.

Sound information

Another important type of recorder information is sound (audio) information, broadcast nowadays not only by radio or compact discs (it means great-grandchildren of Edison's cylinder for his phonograph), but also in the form of audio files or so called podcasts, that is electronic recordings of speech or music. These new possibilities have changed and will further change the existing manners of perception of sound recordings, which has been lately most visibly manifested by massive use of the so called mp3 players, namely by young the ones.

However, what concerns search of audio files – it is again based predominantly on text descriptions, i.e. some kind of indexing – similarly as in case of search of video files on YouTube, etc. Website http://vozme.com (among others) offers quite interesting possibility of transfer of texts into audio mp3 file. This possibility exists not only for English, but also for other languages, namely Italian, Spanish, Catalan, Portuguese and even Hindi.

Processing of sound information

New approach to processing of sound information offers for example the company Nexidia, which offers on commercial basis automatic indexing of audio recordings on the basis of their physical characteristics. In other words – text recorded in form of speech in sound recording is not transferred into its written form, the course of the diagram of audio record is indexed in special manner, which enables later search of requested words or phrases. It is then possible to transfer the searched expression into the form corresponding to the used sound indexing, and to find occurrence of the searched string in the database of all recorded and indexed sound recordings. It is in fact indexing of sound information freed from necessity of transferring the whole sound recording into text. These possibilities are at present used particularly by security services at monitoring and analysing of wiretaps, etc.

Nevertheless, internet already offers several services providing possibilities of audio search based on similar principle. Here are some examples:

Search engine	Search by description	Object of search	Other modes of search
http://www.midomi.com	YES	search of music	search of music by sung or hummed melody
http://www.owlmm.com	selection from catalogue	search of music	search of music by mp3 file
http://www.findsounds.com	YES	search of sound effects	
http://www.hibou-music.com	selection from catalogue	search of music recordings	
http://www.pdsounds.org	YES	encyclopaedia of sound recordings	

Possible trends of future development in search of information

Increasing role of the owners of full-text information = publishers

Users require the quickest possible and the most comfortable access to the answer to their request. This means in practice an increasing role of databases enabling direct access to full texts of documents, preferentially with highlighting of the searched text directly in the found primary source. Role of secondary information will not be extinguished by this, it will only change – they will continue to serve for finding the relevant primary sources, but they possible may not be even displayed in future as an "intermediate step" of search.

"Automatic" indexing will gradually prevail over "manual", i.e. "human" or "intellectual" indexing, before capacities of servers and memories will enable replacement of the existing databases based on secondary information by databases based on processing of full texts with outputs in the form of full-text information.

Due to the fact that full texts are mostly owned by publishers, they will gradually become operators of such full-text databases, dominating information market. Publishers will thus become database centres of higher type.

"Classical databases" will therefore be gradually replaced by full-text information.

Transformation of classical databases into multi-media databases

The above mentioned processing of full texts will not be limited to text only, but it will handle also other „non-text " information contained in publications. This will be further development of technology implemented for the first time in „Illustrata" databases. Indexing and search will therefore increasingly use also information contained in graphs, tables, captions of photos and figures, etc. Increasing role will be played also by specialised databases, enabling search of graphical or sound information, based not only on their text description, but also on similarity search of the given image or sound.

Increasing role of internal corporate information

Globalisation brings, however, also some other new elements, limiting exchange of information. We witness on one hand so far unseen possibilities of free access to huge amount of information wealth provided by internet, as well as rich offer of specialised commercial databases, the aim of which is to earn money by providing expert information of higher quality – compared to free information from „ordinary internet". On the other hand supra-national and global companies create their own proprietary world of information – internal corporate information system. Although this system uses apart from internal information also available external information sources, it contributes only insignificantly to expansion of this generally and publicly accessible information wealth.

The companies within their competitive struggle keep expanding the range of their so called "sensitive" and "confidential", which are classified. Technical articles of authors from these companies in magazines or their papers presented at conferences have thus very often character resembling more to publicity than to fact-based text. Information policy of supra-national corporations comprises even organising of their own intra-corporate conferences, the outputs from which are naturally not available to general public.

What concerns expert information we can see in practice three separated types of information systems:

- information available publicly and free of charge on the internet, which contain apart from quality information also big portion of noise and which do not guarantee access to the expert information of the top quality
- information available at some price from commercial databases, usually of high quality and enabling more precise searching
- publicly unavailable and therefore inaccessible expert information, circulating within intra-corporate information systems, especially in supra-national and global corporations

Globalisation requires local sources

In conformity with general trends databases also become more global. This brings about many new aspects and tasks, i.e. challenges. This concerns in the first place the issue of languages. In our area dominant role is so far played by international databases in English, which, however, will have to process more and more the sources from other languages. Apart from languages of the countries with very large population, such as for example China, Japan, Arab speaking world or countries of Latin America, it will be necessary for ensuring the most comprehensive coverage to process also the sources from languages spoken by not so large populations. Here to we see influence of internet, in which all important search engines very consistently localise their interfaces, as well as possibilities of searching.

This processing of local sources cannot obviously get along without collaboration with local partners, it will be thus globalisation with involvement of local potential. In this case too the increasing role will be played by the tools helping to overcome language barrier – be it already mentioned processing and accessing of „non-text " information or the ever improving "computer translators".

We can expect in future also a possibility of new kind of so called "cross search", i.e. not only search in several more or less related databases in one language, but even simultaneous search of several databases regardless of language of those databases or even type of their contents (text, sound, image, …).

Internet sets trends of searching

Specialised database already had to adapt to the mode of operation on the internet, oriented primarily on user friendly environment and simple search with possibility of "advanced search". The basic idea is to make for the user the route to the final target as simple as possible. It means not to bother him or her by complicated interface, not to congest him or her with unnecessary accompanying information or complicated possibilities of searching, nonetheless to offer him or her also a possibility of more precise, although more complicated queries – if needed.

Internet generation of users

Internet has been here already for sufficiently long time and it penetrated sufficiently broadly and deeply (i.e. practically into all countries, but also into majority of institutions, offices, schools and households), which means that it has already brought up new generation of users, which is not encumbered by previous possibilities and procedures of information searching.

This generation views "paper based files" as museum exhibit, similarly as databases, the output of which consists of records containing only abstract and bibliographic data. Internet user is „spoiled" by possibilities of internet search engines and he or she requires the simplest possible search and immediate access to original sources. Reading of abstracts followed by ordering of primary sources is regarded as useless loss of time. This again stresses the need to provide to the users in the easiest possible way full-text information, as mentioned above.

Time aspect

The present is characterised by ever increasing speed. Fight for higher efficiency of companies leads to reduction of number of employees and to increased requirements to the remaining ones by accumulation of functions and shortening of time for individual operations. Everything is required immediately. That's why very often when solution of some issue is searched the preference goes to internet providing an "immediate" answer instead of making a thorough analysis based on larger amount of information from various sources. In this connection I recall a quotation from famous writer John le Carré:

"What is important is only rarely urgent. Urgent is identical to ephemeral and ephemeral is almost a synonym for unimportant."

John le Carré, "Murder of quality"

Free internet versus paid specialised databases

Specialised thematically focused databases built for decades and enabling comparatively very precise searching of sources on the given topic feel more and more challenge from internet, which for many users creates an illusion that it enables finding of anything, immediately and free of charge. We more and more often encounter an opinion „I do not need specialised database or service of information centre since I have access to internet and I can find there everything myself".

Nevertheless, the estimates say that so called „invisible internet", i.e. that part of internet, which is accessible to the users only upon entering appropriate entry passwords, and which is therefore not covered by usual search engines, is approximately 500 times larger than generally publicly accessible part of information stored on the internet. This naturally comprises also the specialised databases (see for example http://websearch.about.com/od/invisibleweb/a/invisible_web.htm)

Database operators and providers must therefore exert much greater efforts on „public education of potential users" in order to explain them that role of specialised databases is irreplaceable even in the age of internet. They should also investigate possibilities of making access to their databases cheaper, learning from inspiring examples exactly of the big internet players, who provide "free" services thanks to clever use of possibilities of advertising offered by the internet.

Information and wisdom

Modern tools of communication together with mass media and internet on one hand "shrank the world" – since we have possibility not only to learn immediately what happens at the other side of the globe, but also see it "live". On the other hand they made the world somewhat "less transparent", since man can orientate worse in ever increasing heaps of information surrounding him or her. This is perhaps the best expressed by one of many witty drawings of Dan Perjovschi, which decorate the walls of the new building of National Technical Library in Prague.

258 CHANNELS
1 LIFE

I would like to conclude my paper by couple of quotations of the Russian author Victor Kornetskiy:

"We believed for millennia that sooner or later Wisdom – which is an ideal – will teach us how to manage the humankind. And we thus quite inconspicuously acquired another formula: Knowledge governs humankind. Knowledge has absolutely and irrefutably taken the power, while outshining the intellect, which apparently did not pass the test of maturity...

...Facts – this is information – At every crossroad you can hear: Give us information! We lack information!

God only knows why we do not hear: Give us sage advice! We lack wise ideas! What shall we do with abundance of wise ideas?!

Information more and more successfully replaces wisdom. Whatever you take, idiots, who have information, beat the smart ones by hundred points. And this has entranced the idiots ..."

Victor Kornetskiy "Sailor's dreams" (Morskiie sny)