Meta Data – Guide to Digital Information and Deep Web Word

Digital information is one of significant factor of modern word. From one hand they contribute to unbelievable data volume growth, but from the other they are guides to valuable information, which quite often is available only through the meta data. Current article deliver short introduction into digital information, information searching paradigms and meta data role, importance and errors during their creation.

Our world is going digital. For new generations of information society it is not a change they can observe, it is a normal, current state of the world they met. Going digital is a little bit more controversial for older generations. However, despite of how we, mature generations, judge this process, we cannot avoid digitization and we can't escape from living partially online. Naturally the changes affect the core of modern society – Information[1]. One may not forget that digitization is not only opportunity – it is complex process with full spectrum of occasions and threads. Information is a form independent entity. It can be stored and distributed by the mean of physical media, like a book, a CD, etc. as well as digital media – Internet-based, thus is extremely easy adaptable to digital world. The digitization covers, of course, not only storing and distribution of information. A substantial part of the process is also creation of digital information – digital-born information[2]. Despite of origin, digital information needs to be retrieved in order to have any value for users.

Opportunities and threads of information digitisation

There are opportunities and threads of digitization of information. Opportunities are, for example, low costs of production, storage and distribution of information, while one of the most dangerous threads is information flood. What is an opportunity for digital information in a narrow sense, it can cause extremal overproduction of information in broader view. Eric Schmidt, Google CEO, in 2003 estimated:

"Every two days now we create as much information as we did from the dawn of civilization up until 2003".

Additionally E. Schmidt estimated also there is about 5 Exabytes of information created every two days, and such amount of information our civilization has created till year 2003.[3] In fact, R. J. Moore detailed and explained the statement, we were already creating or multiplying about 6,8 Exabytes of information around year 2010, but it was much more from the beginning till 2003.[4]

As IFLScience assess, there are about:

- 456000 tweets send to Twitter,
- 4 6000 of new pictures in Instagram,
- nearly 74 000 of new Tumblr posts.

---

1 Here we understand by the term also a data, which in canonical view are only base factor of information, together with interpretation given to them.

2 As digital-born information we define any information that is created, processed, stored and distributed only in digital form.

3 Cited after Siegler MG., Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003 – post at TechCrunch, online: https://techcrunch.com/2010/08/04/schmidt-data/

4 Moore R. J., Eric Schmidt's "5 Exabytes" Quote is a Load of Crap, online: https://blog.rjmetrics.com/2011/02/07/eric-schmidts-5-exabytes-quote-is-a-load-of-crap/

Every Minute![5]

In case of information the more is not the better! Such overproduction of information raises problems of access to desired data and information and, maybe more substantial in terms of daily life facilitation[6], the problem of information quality. Using search engines one can easily find thousands or millions of documents estimated as relevant to given query. Even thousands are too much for human brains reception, and impossible to only read in reasonable time.

Indexing and searching paradigms

Since opening of the Internet to the wide public, we could indicate few information searching paradigms. First it was manual selection, evaluation and distribution of information about digital, online resources. The process was just adaptation of traditional library cataloguing routines to a new environment. In manual classification paradigm users queries are limited to given system categories or classes, furthermore users are expected to learn the structure of web cataloguing system. As examples of the paradigm we may recall Gopher protocol[7], as well as quite current web catalogues – and the most known one – The Open Directory Project (DMOZ). Manual classification of Internet resources by DMOZ was used among others by Google, in order to increase search results relevance. Unfortunately, as of 17 March, 2017 DMOZ is no longer available. Unquestionable value of manual cataloguing is its reliability – during the process digital information is analysed and its subject is adjusted to one of a priori set categories or classes. And its obvious drawback is time – the process of manual resources cataloguing is simply way too slow. Digital information is categorized too late, from users point of view.

Modern approaches to information indexing and searching are known under common name Information Retrieval (IR). As we may read in [Maning, Raghavan & Schütze, 2008] web search is confronted to structural search (typical for data bases). The latter one requires good knowledge of data base structure and relations between data base fields. Thus for each different data base user need to learn its structure. IR offers full text searching allowing to bypass obstacles of manual cataloguing of web resources. As answer for slow process of cataloguing of web resources we can observe a second paradigm – instant, automatic indexing. This approach offers quite current information on Internet resources, indexation process covers full web documents text and allows full text searching. Each significant word from document becomes a search index. Users are free to give any words in their queries. By there appears question of search results quality. Users are free in terms of building queries. Mostly those are single keywords, given as a query. For ambiguity, synonymy and popularity of keywords user receive wide sets of documents a system answer. The problem of finding the most relevant documents is limited to a small fraction of all available documents, but there still remains a lot of filtering work to do.

Semantic searching, promoted as a third searching paradigm, is a promise of improving the searching process.[8] This approach applies linguistics solutions in order to retrieve documents matching the meaning instead of the word form. Semantic searching recognizes semantic relations between concepts and their word representations. However already the idea of semantic searching is well known, it still remains in planning phase. The structure of information and, especially of meta information, is not sufficient enough to open semantic web era.

Beyond full text searching

It appears clearly full text searching is not capable to index the whole Internet content. There is a lot of documents which are ignored or are not available for search engines. Most of web databases content is not searchable by web crawlers, either. Despite of easiness and simplicity of giving information in form of free texts there is many application areas which require properly structured or described information and data. All BigData application are triggered towards detection any meaningful order among available data, for grouping and comparing them.

From data analytical perspective we cannot escape from some way of describing data structure, order or relations. Except pure DeepWeb[9] resources, we may recall web repositories. A web repository is a structured base of documents stored and available through Internet web sites. AS the major content of such resource is stored in data bases, it is mostly not accessible for universal search engines, like Google. A prominent example of such repositories in Europe is Europeana.[10] The digital library of European Culture Heritage (CH) objects hosts currently nearly 52 million of records. CH objects are digital representations or reproductions of physical artefacts, but there are also digital born objects. All the core content of digital libraries, due different form and formats (e.g. pdf documents, jpg or png files, movies, mp3 music files, etc.) is not easily searchable nor indexable. In case of non-textual documents it is not possible to apply full text searching, simply there are too many different factors to index which are not intuitively searching keywords for human user.

Europeana and other digital libraries provide, except files of digital representations of objects, also added value in form of meta data. In the scope of meta data there are object identifying data, as name, title, author, etc. and content identifiers as e. g. description. As example of good, informative description of object content we may recall e. g. object entitled: *Grudziądz – monument to a lancer and a girl*. It is a picture taken from the sculpture linked from Pictures-Bank[11]. Except of the picture itself, we can read description of its content in Polish (as the sculpture is from Poland and a photographer is Polish, too) and in English:

---

5 Hale T., How Much Data Does The World Generate Every Minute? online: http://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/

6 Which, ironically, is one of digital world promises.

7 Gopher (protocol) online: https://en.wikipedia.org/wiki/Gopher_(protocol)

8 Which in current state, in deed is already accepted by users as quite convenient.

9 Described e.g. in Devine J., Egger-Sider F., Going beyond Google: the Invisible Web in learning and teaching, London: Facet Publishing 2009

10 Europeana Collections, online: https://www.europeana.eu/portal/en

11 Pictures-bank: http://www.pictures-bank.eu/index.php?action=przegladaj_zdjecie&id=37250

[EN]: The monument to a lancer and a girl is situated on brick walls at Spichrzowa St. in the Old Town in Grudziądz. Both figures are slightly cuddled against each other. They look as if they just ran up to each other. They hold in their hands a posy of flowers....[12]

All the meta data of Europeana objects were, for a long time, given according to Dublin Core standard (<dc: ….> tags in description files). There were also Qualified Dublin Core elements (<dcterms: …> tags) and Europeana Semantic Elements just to mention the most extensively used standards. Currently the resources are described according to Europeana Data Model[13] – a richer model that allows for domain-specific meta data incorporation into Europeana resources.[14] An example IR research on Europeana collection data set done in 2013 within *Cultural Heritage in CLEF* (CHIC) was described in [Akaresh, Malak & Pawłowski]. The results of including semantic element lead to decrease of IR accuracy (relevance) for all subject terms were, at that time, automatically translated into all European languages, which leaded to false-positive matches, as quite often a term in e. g. English was similarly written in other language, but the meaning of both was not always similar.

One should be aware Europeana itself does not create new digital data. It is meta data aggregator, one of the biggest. All the particular data are being posted to Europeana by digital libraries, museums, galleries or archives. All the local GLAM institution uses own meta data standards, but the leading one is Dublin Core.

There are not only standards which shape the meta data. The main factor are persons responsible for meta data creation. A pilot research on Polish digital libraries resources described in [Osińska, Malak] detected number of serious data errors, which excluded particular resources from search results even in resource mother digital libraries. As main categories of problems we may indicate:

- dc:type errors:
  - data provided as publication type,
  - title given as type,
  - original (physical) document dimensions given as type,
  - publication author,
  - publication date,
  - high fractionation of used document types
- dc:author errors:
  - spelling errors,
  - mismatching authors last name and first name
- dc:date errors:
  - inconsequence in data factor order (YYYY-MM-DD vs. DD-MM-YYYY) within data of the same digital library
- general errors:
  - spelling errors, like: photography, pohtography[15]

Conclusions

Resources of digital libraries are valuable source of data, that are searchable by the mean of meta data. But there is a problem of meta data quality, caused mostly by humans. Although using Dublin Core, digital libraries do not always follow an uniform standards and normalization in data. There are different terms used for the same type of digital documents or just spelling errors left uncorrected, also within meta data of one library. Persons responsible for meta data creation should pay more attention for the quality of data, as they are becoming more and more valuable source.

Acknowledgement

Research was partially granted by grant PL NCN 2013/11/B/HS2/03048

References

1. Akasereh Mitra, Malak Piotr, Pawłowski Adam, Evaluation of IR Strategies for Polish, [w:] Advances in Natural Language Processing. 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17–19, 2014. Proceedings, ed. by Adam Przepiórkowski, Maciej Ogrodniczuk, Heidelberg [et al.] 2014, s. 384–391 (Lecture Notes in Computer Science; vol. 8686).
2. Devine J., Egger-Sider F., Going beyond Google: the Invisible Web in learning and teaching, London: Facet Publishing 2009
3. Europeana Data Model Documentation, online: https://pro.europeana.eu/resources/standardization-tools/edm-documentation

[12] Europeana: Grudziądz - monument to a lancer and a girl, online: http://www.europeana.eu/portal/record/2023815/LinkedHeritage_Update_ProvidedCHO_Pictures_bank_eu__ICIMSS__37250.html

[13] Europeana Data Model Documentation, online: https://pro.europeana.eu/resources/standardization-tools/edm-documentation

[14] The Current Europeana Dataset, online: https://pro.europeana.eu/page/the-current-europeana-dataset

[15] Modern text editors do automatically correct most of spelling errors, but digital libraries software seems to be void of such features.

4. Hale T., How Much Data Does The World Generate Every Minute? online: http://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/
5. Manning Ch. D., Raghavan P. & Schütze H. (2008) *Introduction to information retrieval*, Cambridge University Press, retrieved October 21, 2017 from: https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html
6. Moore R. J., Eric Schmidt's "5 Exabytes" Quote is a Load of Crap, online: https://blog.rjmetrics.com/2011/02/07/eric-schmidts-5-exabytes-quote-is-a-load-of-crap/
7. Siegler MG. Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003 – post at TechCrunch, online: https://techcrunch.com/2010/08/04/schmidt-data/
8. Osińska V., Malak P., *Maps and Mapping in Scientometrics*. in: M. Góralska i A. Wandel (ed.) Metody i narzędzia badań piśmiennictwa cyfrowego i jego użytkowników, UWr 2016.

Piotr Malak

piotr.malak@uwr.edu.pl ■

(Faculty of Letters, University of Wroclaw, Wroclaw, Poland)