Vizualizace sémantického webu na příkladu doménově vymezené digitální knihovny

Zuzana Nevěřilová, Petr Sojka
Masarykova univerzita v Brně, Fakulta informatiky
{xpopelk | sojka}@fi.muni.cz

18. dubna 2005

Motivace

„Graphics *reveal* data.“ (Tufte)

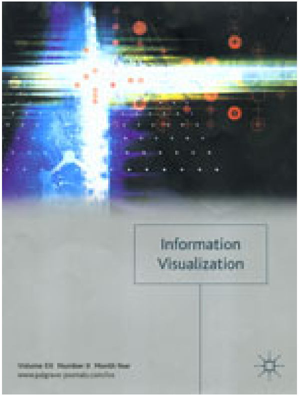
Vizualizace: komunikace informací (obsahu dat), která je *přesná, rychlá,* a přesto *srozumitelná.*

Samostatný vědní obor s konferencemi, časopisy.

Konference

- *HCI International*
  HCI International 2005, Las Vegas, NV, July 22–27, 2005.
  http://www.hci-international.org/
- *InfoVis*
  IEEE Symposium on Information Visualization 2005, October 23–25, Minneapolis, Minnesota, USA.
  http://www.infovis.org/infovis/2005/
- *KIV*
  I-Know 2005 Special Track on Knowledge and Information Visualisation 2005 (KIV 2005). June 29, 2005, Graz, Austria.
  http://i-know.at/kiv

Časopisy

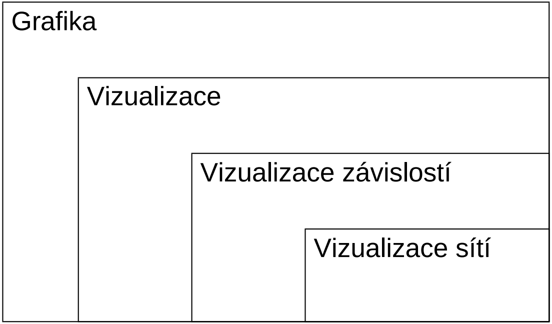


Information Visualization is a central forum for all aspects of information visualization and its applications. The journal is essential reading for researchers and practitioners of information visualization and is of interest to computer scientists and data analysts working on related specialisms.
Palgrave Macmillan Journals

Vymezení vizualizace

Vizualizace je zobrazení modelu struktury, která sama o sobě často nemá vizuální podobu.

Vizualizace statistických dat vs. vizualizace struktur.



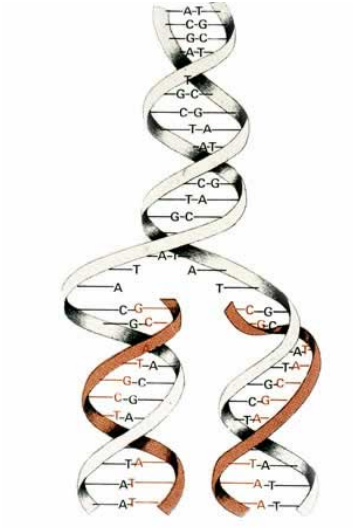
Grafika

Vizualizace

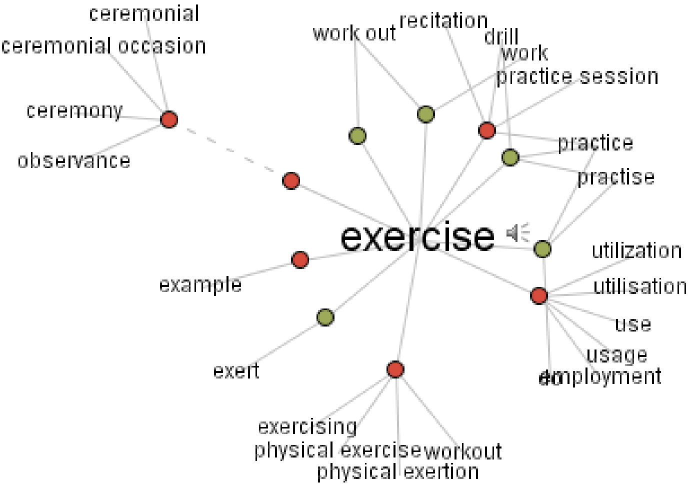
Vizualizace závislostí

Vizualizace sítí

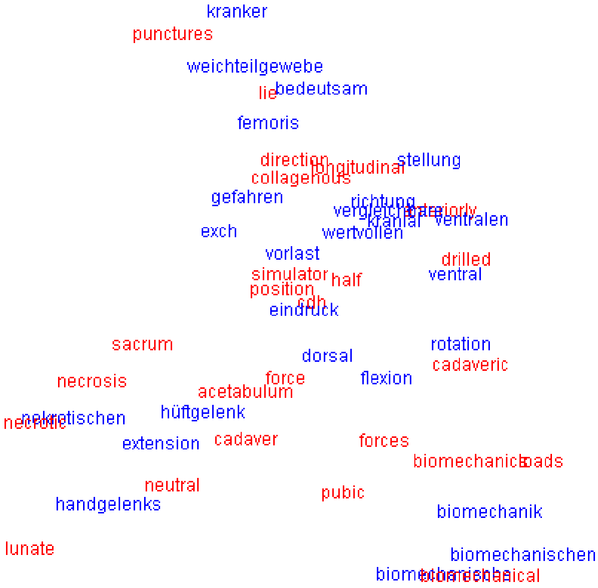
Příklady – obrázek za 1000 slov (DNA)



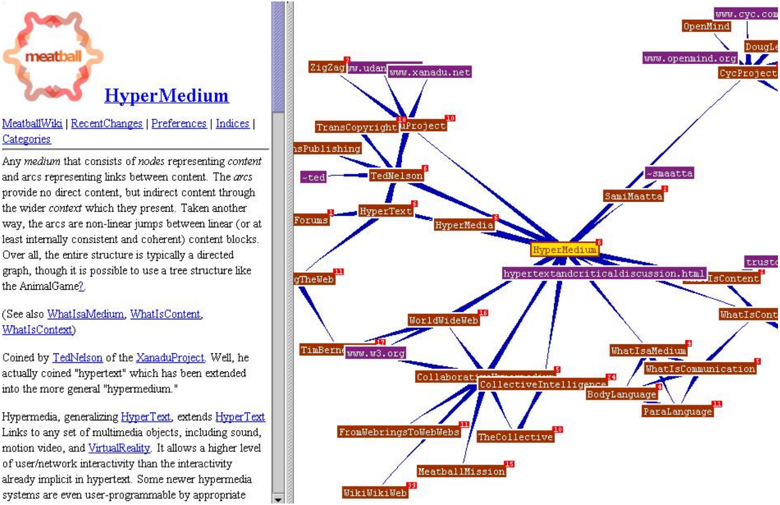
Příklady – obrázek za 1000 slov (Visual Thesaurus)



Příklady – obrázek za 1000 slov (Stanford Infomap)



Příklady – obrázek za 1000 slov (Visual WikiBrowser)



HyperMedium

MeatballWiki | RecentChanges | Preferences | Indices | Categories

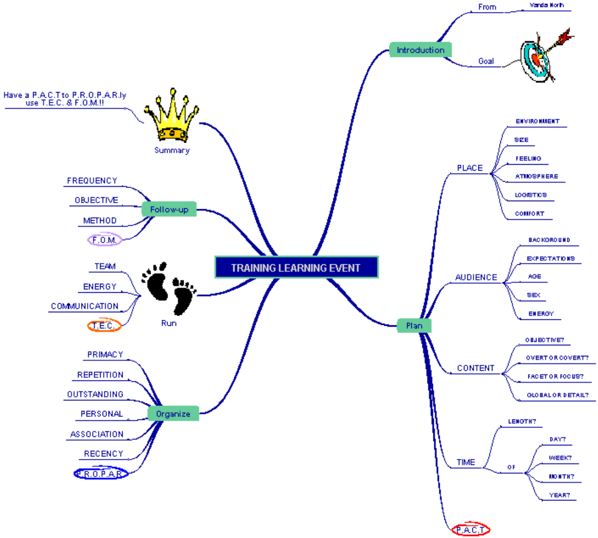
Any *medium* that consists of *nodes* representing *content* and arcs representing links between content. The *arcs* provide no direct content, but indirect content through the wider *context* which they present. Taken another way, the arcs are non-linear jumps between linear (or at least internally consistent and coherent) content blocks. Over all, the entire structure is typically a directed graph, though it is possible to use a tree structure like the AnimalGame?.

(See also WhatIsaMedium, WhatIsContent, WhatIsContext)

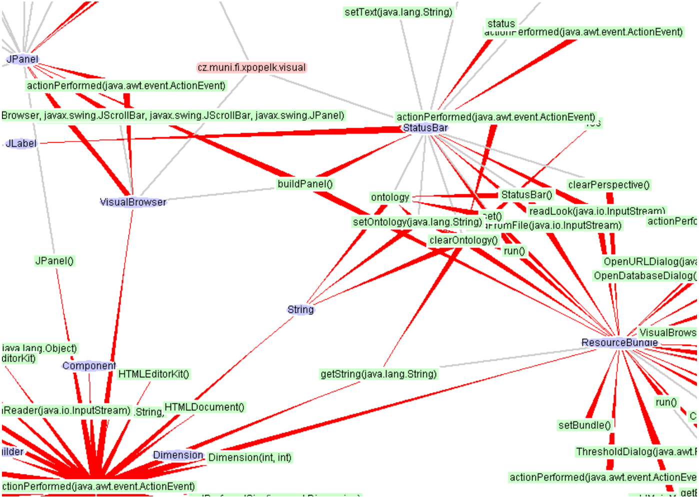
Coined by TedNelson of the XanaduProject. Well, he actually coined "hypertext" which has been extended into the more general "hypermedium."

Hypermedia, generalizing HyperText, extends HyperText Links to any set of multimedia objects, including sound, motion video, and VirtualReality. It allows a higher level of user/network interactivity than the interactivity already implicit in hypertext. Some newer hypermedia systems are even user-programmable by appropriate

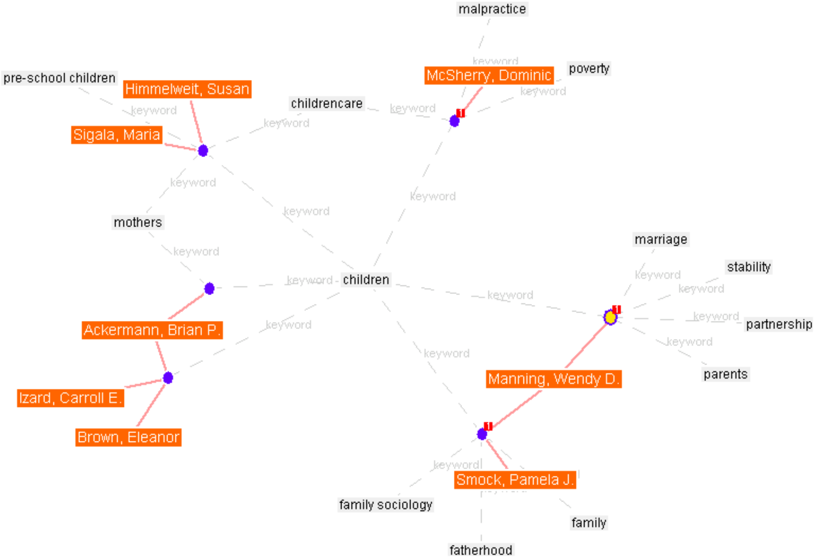
Příklady – obrázek za 1000 slov (Mind Maps)



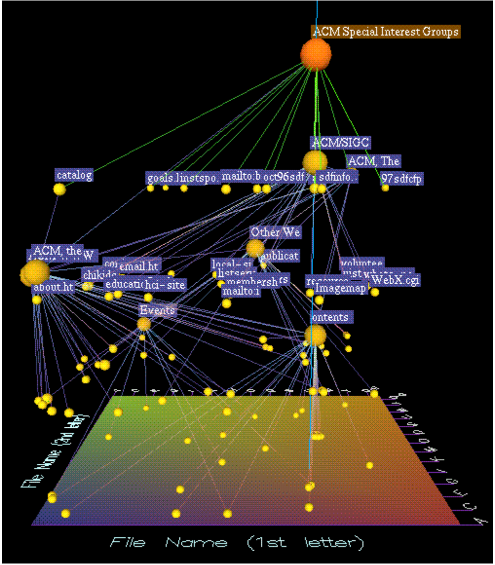
Příklady – obrázek za 1000 slov (Dependency Finder)


setText(java.lang.String)
status
actionPerformed(java.awt.event.ActionEvent)
JPanel
cz.muni.fi.xpopelk.visual
actionPerformed(java.awt.event.ActionEvent)
Browser, javax.swing.JScrollBar, javax.swing.JScrollBar, javax.swing.JPanel)
actionPerformed(java.awt.event.ActionEvent)
StatusBar
JLabel
buildPanel()
clearPerspective()
ontology
StatusBar()
VisualBrowser
readLook(java.io.InputStream)
setOntology(java.lang.String)
FromFile(java.io.InputStream)
actionPerfo
clearOntology()
run()
JPanel()
OpenURLDialog(java
OpenDatabaseDialog(
String
VisualBrows
ResourceBundle
java.lang.Object)
ditorKit)
Component
HTMLEditorKit()
getString(java.lang.String)
Reader(java.io.InputStream)
String,
HTMLDocument()
run()
setBundle()
ilder
Dimension
Dimension(int, int)
ThresholdDialog(java.awt.F
ctionPerformed(java.awt.event.ActionEvent)
actionPerformed(java.awt.event.ActionEve
getI

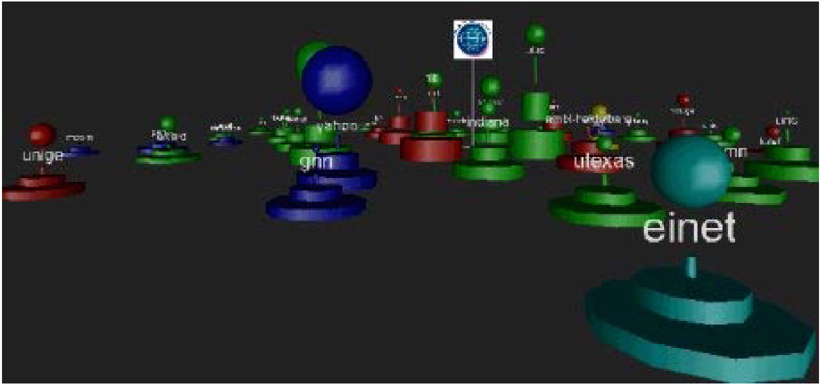
Příklady – obrázek za 1000 slov (Katalog knihovny)



Příklady – obrázek za 1000 slov (3D vizualizace webu I)

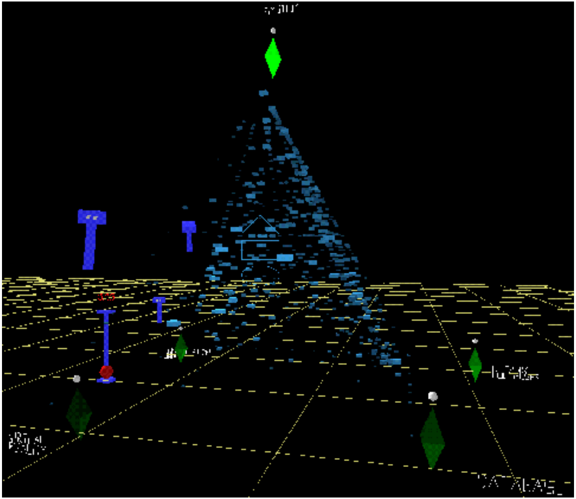


Příklady – obrázek za 1000 slov (3D vizualizace webu II)



Příklady – obrázek za 1000 slov (3D vizualizace webu III)

Příklady – obrázek za 1000 slov (3D vizualizace webu IV)



Od slov k informacím

Úrovně zpracování informací

slovní tvary → lemmata (morfologie) → významy (sémantika) → pragmatika, emoce, sebeuvědomění

’The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.’

– Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001 – W3C Definition of the Semantic Web

Resource Description Framework (RDF)

RDF je datový model pro popis metadat navržený W3C.

- Danou doménu popisuje pomocí *zdrojů* a *literálů*. Zdroje mají jedinečnou identifikaci (URI), literály jsou řetězce.
- RDF data tvoří *tvrzení* čili trojice (**S**ubject, **P**redicate, **O**bject).
- RDF data jsou uložena v XML, N-Triples (trojice) či relační databázi.

Příklad RDF trojic

<Mach>	<být spolužákem>	<Šebestová>
<Mach>	<být žákem třídy>	<3. B>
<Šebestová>	<být žákem třídy>	<3. B>
<Šebestová>	<nosit>	"červená mašle"

Velká vyjadřovací síla RDF

... díky *reifikaci*.

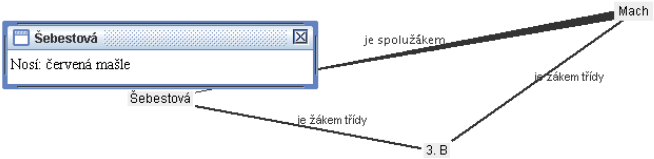
<Mach>	<být spolužákem>	<Šebestová>

Kropáček si myslí, že Mach je spolužákem Šebestové.

<Statement>	<subject>	<Mach>
<Statement>	<predicate>	<být spolužákem>
<Statement>	<object>	<Šebestová>
<Kropáček>	<myslet si, že>	<Statement>

Statement je tzv. *anonymní uzel*.

Vizualizace RDF trojic

- RDF lze vizualizovat jako orientovaný multigraf. Z trojice (**S**ubject, **P**redicate, **O**bject) jsou **S** a **O** uzly grafu, **P** je hrana grafu.
- 

- Pokud je **O** literál a nikoli zdroj, zobrazuje se trojice (**S**, **P**, **O**) jako *hint*.

Aplikace Visual Browser

- dvouvrstvá architektura
- interaktivita, animace, fokus
- implementace

Příklad vizualizace – CiteSeer

- 700 000 dokumentů, přes 2 500 000 stránek
- automatické budování citačních závislostí
- automatizovaný sběr metadat
- dostupnost přes web
- detekce duplicit
- doménově specifický – informatika (Computer Science)

http://citeseer.org

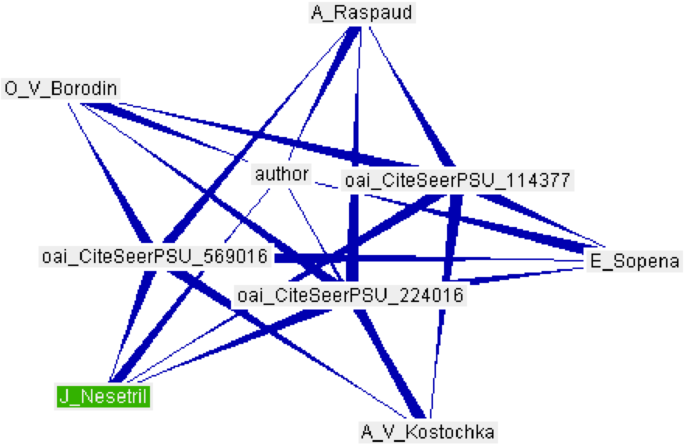
Dvouvrstvá architektura

Algorithms + Data Structures = Programs (Niklaus Wirth)
Data + Perspective = Visualization (ZN + PS :-)

Perspektiva pohledu je množina definic vzhledu uzlů, hran a hintů.

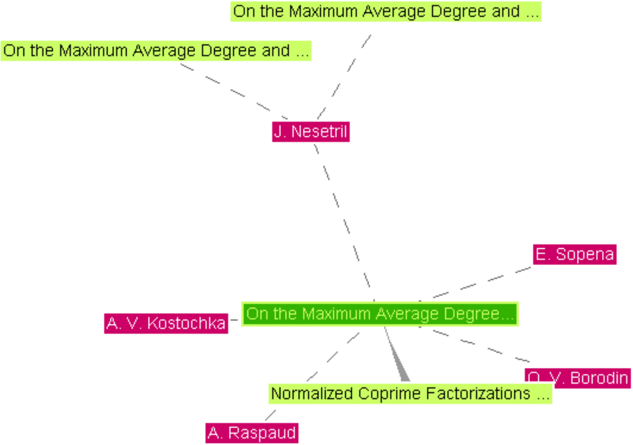
- Vzhled uzlu je definován pomocí tvaru, barvy pozadí, barvy textu, velikosti fontu.
  Vzhled hrany je definován pomocí tvaru, barvy a délky.
- Perspektiva umožňuje vzhledově odlišit různé třídy uzlů (např. autory a díla) nebo hran (různé druhy relací), ale také konkrétní uzly či hrany.
- Nad jedněmi daty lze definovat více perspektiv a pohlížet tak na data z více úhlů.

Stejná data zobrazená normálně a s perspektivou



každá trojice je zobrazena stejným způsobem

Stejná data zobrazená normálně a s perspektivou


On the Maximum Average Degree and ...
On the Maximum Average Degree and ...
J. Nesetril
E. Sopena
A. V. Kostochka
On the Maximum Average Degree...
O. V. Borodin
Normalized Coprime Factorizations ...
A. Raspaud

různé třídy uzlů i hran jsou zobrazeny různě

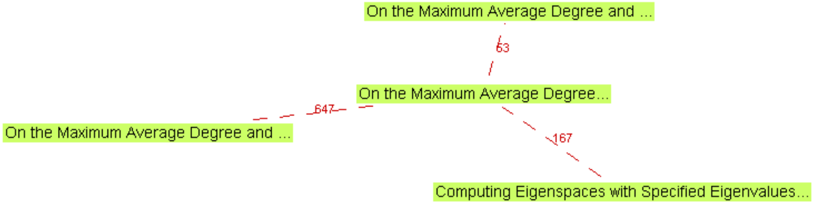
Příklad perspektivy

```
<?xml version="1.0"?>
<!DOCTYPE perspective SYSTEM "perspective.dtd">
<perspective>
   <name>Citeseer</name>
   <node subj="http://nlp.fi.muni.cz/citeseer#record"
      display="false"/>
   <node
pred="http://www.w3.org/1999/02/22-rdf-syntax-ns#type"
      display="true"
      obj="http://nlp.fi.muni.cz/citeseer#record"
      bgcolor="#ccff66" color="#000000"/>
   <edge
name="http://purl.org/dc/elements/1.1/isReferencedBy"
      shape="1" length="100" desc="is referenced by"
      label="is referenced by" color="#999999"/>
</perspective>
```

Vizualizace „spojitých“ dat

Trojici (**S**, **P**, **O**), kde **O** je literál, lze buď zobrazovat jako hint nebo interpretovat jako prvek vzhledu uzlu či hrany.

Tímto způsobem lze do grafu zakomponovat „spojitou“ informaci, např. vzdálenosti vypočítané z četností v korpusu:


On the Maximum Average Degree and ...
63
On the Maximum Average Degree...
647
On the Maximum Average Degree and ...
167
Computing Eigenspaces with Specified Eigenvalues...

Příklad

```
<record rdf:ID="oai_CiteSeerPSU_224016">
   <related rdf:resource="#oai_CiteSeerPSU_569016"/>
      <proximity>53</proximity>
   <related rdf:resource="#oai_CiteSeerPSU_447389"/>
      <proximity>437</proximity>
         ...
</record>
```

Perspektiva:

```
<hint name="#proximity" eval="length" edge="#related"/>
<hint name="#proximity" eval="label" edge="#related"/>
```

Aplikace Visual Browser

- dvouvrstvá architektura
- interaktivita, animace, fokus
- implementace

Interaktivita, animace, fokus

- interaktivita umožňuje volbu zobrazovaných dat – skrývání uzlů či hran, „odsunutí” uzlu z cesty jinému uzlu, apod.
- animace rozmisťování uzlů
- fyzikální model grafu (řízené silou), převzato z (TouchGraph LLC)
- fokus – nezobrazuje se celá síť, vždy je vybraný jeden uzel a zobrazují se uzly uvnitř daného poloměru
- další vylepšení přehlednosti: automaticky generovaná legenda ke grafu, volitelné jmenovky hran. . .)
- lze nastavit práh pro maximální počet hran téhož typu vycházejících z jednoho uzlu

Aplikace Visual Browser

- dvouvrstvá architektura
- interaktivita, animace, fokus
- implementace

Implementace

- čistá Java aplikace (Java 1.5.0)
- k dispozici také jako Java Web Start na `http://nlp.fi.muni.cz/projekty/vizualni_lexikon`
- Apache License
- uživatelský komfort: snímek obrazovky v SVG, vícejazyčná podpora. . .

Další práce

- lepší prohledávání
- uživatelsky přítulné prostředí pro vytváření perspektiv
- . . .

Bottom-up way to WDML—DML-CZ

- Vision of WDML for years. Estimate of 50.000.000 pages in total only.
- Failure of global funding of DML-EU within FP6.
- Google Print project: massive digitization of Harvard, Stanford, Oxford, University of Michigan and New York Public libraries ($150.000.000).
- Niche “markets”, grey literature, mathematical literature published in CE not covered.
- Making WDML (bottom up)^2: with the help of the local goverment funding: DML-CZ, from scanned images to full text marked pages.

(W)DML Initiatives

NUMDAM Numérisation de documents anciens mathématiques.

ERAM The Jahrbuch Project Electronic Research Archive for Mathematics (1868–1942): „Jahrbuch über die Fortschritte der Mathematik“

JSTOR (AMS journals)

EMANI electronic mathematical archiving network (Cornell, SUB Göttingen, MathDoc, Tsinghua University Library)

RusDML Russian DML (2.000.000 pages of papers in Zbl refereed journals)

DML-CZ Digital Mathematical Library of mathematical literature published in Czech.

The Goal

- Czech Academy of Sciences grant (program Information Society) 2005–2009, *full* (retro)digitization of 50.000 pages of mathematical literature per year is planned.
- Design issues to discuss: gradual enhancement of the digital material by ‘knowledge enhancing’ filters on markup-rich XML data.
- We do not want to reinvent the wheel.
- New methods for (semantic) text processing tested on the available data.

More questions than answers at the present stage.

Specifics of Mathematical Publications

① review databases where entries are *classified* according to the Mathematics Subject Classification Scheme (MSC 2000).

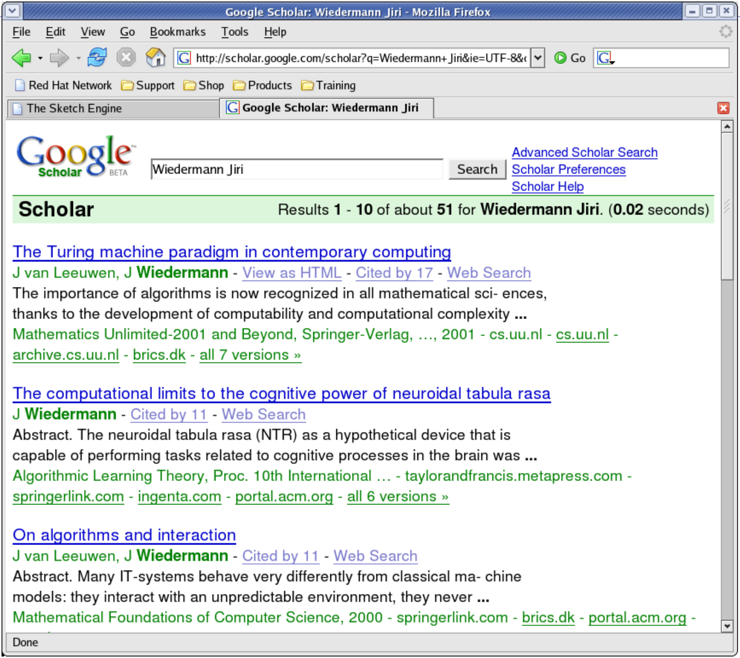
② *Zentralblatt MATH* (more than 2.000.000 entries drawn from more than 2300 serial and journals) Jahrbuch über die Fortschritte der Mathematik (JFM) covering the period 1868–1942 (200.000 entries digitized in ERAM).

③ *MathSciNet*: 24.157 items added in 2005; 1799 journals covered; links to 501.123 original articles; 11.304 active reviewers; 428.680 authors indexed. Since 1940.

④ 50 year old or even older papers are frequently cited.

Limited search in review databases, only things as collaboration distances.

On the ‘opposite’: CiteSeer, Google Scholar.



Google Scholar: Wiedermann Jiri - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://scholar.google.com/scholar?q=Wiedermann+Jiri&ie=UTF-8&
Go
Red Hat Network
Support
Shop
Products
Training
The Sketch Engine
Google Scholar: Wiedermann Jiri
Google Scholar BETA
Search
Advanced Scholar Search
Scholar Preferences
Scholar Help
Done

Wiedermann Jiri

Scholar Results **1** - **10** of about **51** for **Wiedermann Jiri**. (**0.02** seconds)

The Turing machine paradigm in contemporary computing
J van Leeuwen, J **Wiedermann** - View as HTML - Cited by 17 - Web Search
The importance of algorithms is now recognized in all mathematical sci- ences, thanks to the development of computability and computational complexity **...**
Mathematics Unlimited-2001 and Beyond, Springer-Verlag, ..., 2001 - cs.uu.nl - cs.uu.nl - archive.cs.uu.nl - brics.dk - all 7 versions »

The computational limits to the cognitive power of neuroidal tabula rasa
J **Wiedermann** - Cited by 11 - Web Search
Abstract. The neuroidal tabula rasa (NTR) as a hypothetical device that is capable of performing tasks related to cognitive processes in the brain was **...**
Algorithmic Learning Theory, Proc. 10th International ... - taylorandfrancis.metapress.com - springerlink.com - ingenta.com - portal.acm.org - all 6 versions »

On algorithms and interaction
J van Leeuwen, J **Wiedermann** - Cited by 11 - Web Search
Abstract. Many IT-systems behave very differently from classical ma- chine models: they interact with an unpredictable environment, they never **...**
Mathematical Foundations of Computer Science, 2000 - springerlink.com - brics.dk - portal.acm.org -

What to digitize?

Selection not yet finished: 5–8 journals, 100–200 conference proceedings, monographs and textbooks. In total 200–300.000 pages. First journals to start with:

① *Czechoslovak Mathematical Journal* (30.000 pages to scan, 7.000 are already born digital). Published by Academy of Sciences of CR, distributed partially by Springer. Founded as *Časopis pro pěstování matematiky* in 1872, under current name since 1951. 272 pages quarterly.

② *Applications of Mathematics* (20.000/5.000). Published by Academy of Sciences of CR. Founded in 1956 (as *Aplikace matematiky*). 80 pages bimonthly.

③ *Archivum Mathematicum* (2.000/4.000) Masaryk Uni in Brno.

Mathematica Bohemica already digitized in Göttingen,. . .

Who is in the project?

Four contractors (all from Czech Republic):

① **Czech Academy of Sciences, Prague** Jiří Rákosník, head of the project, responsibility for material selection, copyright negotiations.

② **Masaryk University in Brno** Petr Sojka (Faculty of Informatics) formats and tools, technical coordination.
Mirek Bartošek (Institute of Computer Science), content management system, metadata harvesting, long-term archiving.

③ **Charles University in Prague** Jiří Veselý, Oldřich Ulrych, selection and preparation of materials for digitization, metadata.

④ **Library of Academy of Sciences, Prague** Martin Lhoták, document scanning.

Phases planned

acquisition preparation, document acquisition, copyright issues handling;

scanning document scanning (1/5 of the budget) main metadata entering, scanning checks;

image processing main OCR, image enhancements.

semantic processing document markup enhancement, semantic processing, document classification, citation linking, document clustering, indexing;

presentation visualization techniques of document repository, digital library web portal, interfaces to other services and search engines for the semantic based document processing/delivery.

Preparation

document selection criteria?, grey literature too?

preparation acquisition of documents for scanning.

copyright negotiation with publishers (or even authors?)

In what order? What is important when signing digitization contract?

Scanning

Floods in Bohemia three years ago. Many manuscripts were under water, and frozen (put into the refrigerator). Workflow for proces of defrosing includes scanning (Library of Academy of Sciences, Jenštejn near Prague, capacity of 40.000 pages per month or more!).

parameters 600 dpi bitonal according to BPS.

scanning facilities Digibook RGB 10000, A1 color book scanner; two book scanners Zeutschel OS 7000, A2 B/W.

software Book Restorer to make the scanned pages uniform (white space around text body,. . .); system Sirius for archival storage of scanned materials (they are put on CDs in TIFF G4); system Kramerius (open source, created under contract) for scanned documents delivery.

Metadata

OCR ABBYY FineReader? XDOC? Several OCR layers? storage of references to unresolved images (math) for future processing (AutoTag)?

metadata choice of, retyping or OCR tagging?

image enhancements multiple format, PDF, DjVU conversions, software?

semantic processing document markup enhancement, semantic processing, document classification, citation linking, document clustering, indexing;

Dublin Core, miniDML or ZentralBlatt+MR? Or all? BibTeX or XML? software for digital repository? (DSpace?) bibitem handling, addition of ZBL, MR, JHR hypertext links in miniDML? Technology for doing the linking?

Presentation, Visualization

visualization techniques ‘lost in hyperspace fear’, vizualization of document clustering, Visual Browser.

web portal unique and persistent URLs (DOI? URN? PURL?,. . .)

interfaces to other services OAI-PMH harvesting, bibitem export

indexing, search relevance EDBM-2?, mirroring? Google Scholar?

Storage, Indexing

space	multiple OCR layers, multiple attribute layers (lemmas, reviewer comments, semantic classifications, etc.) no problems to store and index all of that for *all* mathematics literature so far.
software	client/server architecture, Bonito and Manatee developed at NLPLAB FI MU, used by OUP dictionary development (Oxford Thesaurus of English, 2004) based on corpora of 100.000.000 word positions, superior scaling qualities.

Document Markup Enhancement Methods

① context dependent mapping from visual to logical markup

② algorithms of language identification (bi-gram, tri-gram based, par or even sentence level)

③ document classification, metrics, ontology construction, comparison with AMS 2000 classification

④ semiautomatic bibliography markup and metrics, *global mathematics* citation index, "MathRank"

⑤ document clustering (for visualization, . . .), identification of near duplicates

Conclusions

> We should experiment; we should try out new things; we should tinker with technology and find better ways to communicate. *John Ewing (2002)*

We are at the start—many problems are unresolved. Preliminary project web pages are at http://dml.muni.cz/. Will Google Print and Google Scholar projects take over before (W)DML is finished (90:10% rule)?

Real data are needed to explore methods further.

Properly designed *visualization* may help to *reveal* enormous amounts of (textual) *data.*

„Graphics reveal data." (Tufte)

G. Michailidis and J. de Leeuw.
Data visualization through graph drawing, 2001.
Available from: `citeseer.ist.psu.edu/michailidis01data.html`.

RDF Vocabulary Description Language 1.0: RDF Schema, 2004.
Available from: `http://www.w3.org/TR/rdf-schema/`.

RDF/XML Syntax Specification, 2004.
Available from: `http://www.w3.org/TR/rdf-syntax-grammar/`.

V. Geroimenko and C. Chen.
Visualizing the Semantic Web: XML-Based Internet and Information Visualization.
Springer Verlag, 2003.

A. Shapiro.
TouchGraph LLC at SourceForge, 2004.
Available from: `http://touchgraph.sourceforge.net/`.

E. Tufte.
Envisioning Information.
Graphics Press, 1990.

M. L. Huang, P. Eades, and R. F. Cohen.
WebOFDAV: Navigating and Visualizing the Web On-line with Animated Context Swapping.
In *Proceedings of the 7th International WWW Conference*, pages 638–642, 1998.

Steve Benford, Ian Taylor, David Brailsford, Boriana Koleva, Mike Craven, Mike Fraser, Gail Reynard, and Chris Greenhalgh.
Three dimensional visualization of the world wide web.
ACM Comput. Surv., 31(4es):25, 1999.