PRÍPRAVA DÁT NA ANALÝZU A PERSONALIZOVANÚ PREZENTÁCIU NA WEBE V DOMÉNE VEDECKÝCH PUBLIKÁCII

Ústav informatiky a softvérového inžinierstva
Fakulta informatiky a informačných technológií
Slovenská technická univerzita, Bratislava

Oto Vozár a Mária Bieliková

Obsah

- Motivácia - kontext a ciele prípravy dát
- Doména vedeckých publikácií a získané dáta
- Predspracovanie dát
- Overenie metód
- Záver a ďalšia práca

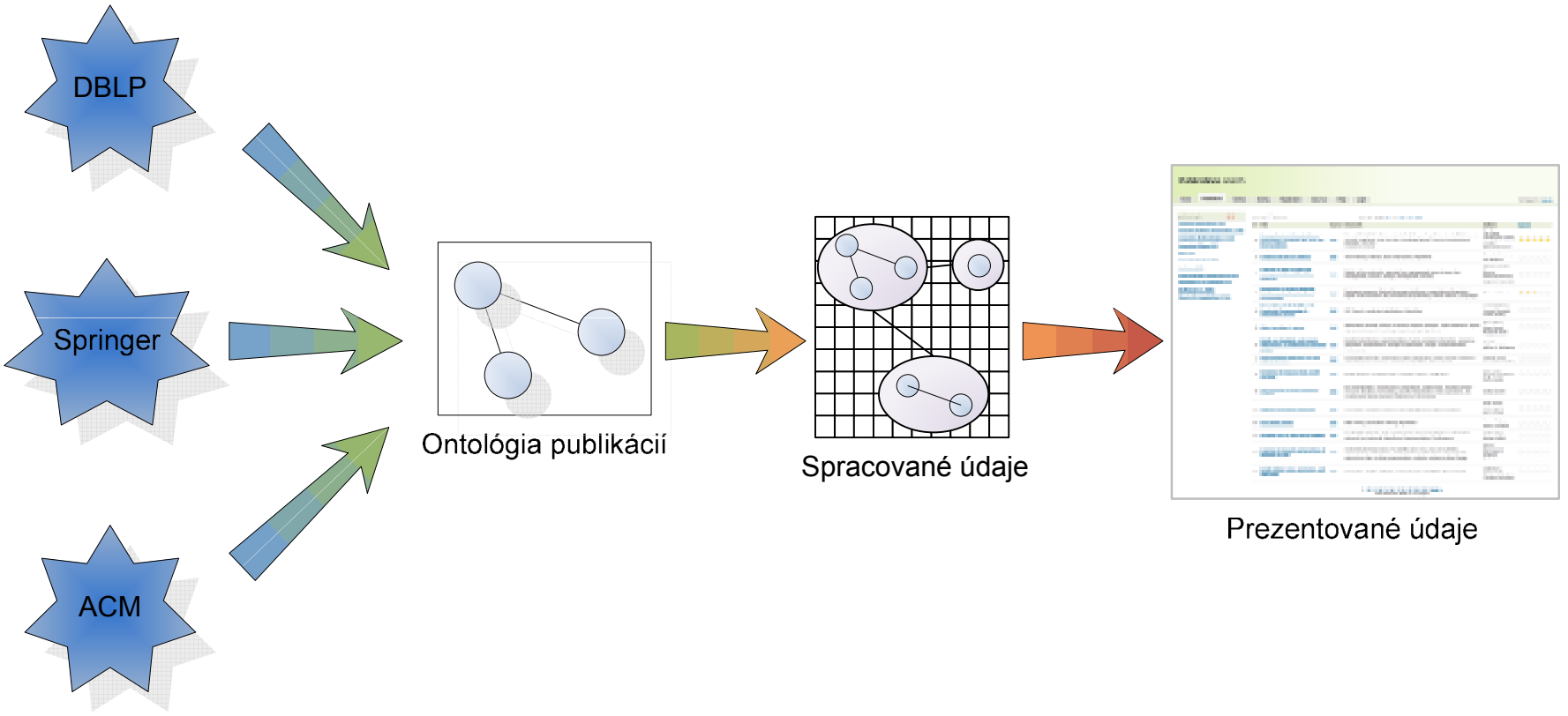
Kontext a ciele prípravy dát

- V rámci výskumného projektu MAPEKUS (Modeling and Acquisition, Processing and Employing Knowledge about User Activities in the Internet Hyperspace)
- Personalizovaná navigácia vo veľkých informačných priestoroch
- Získavanie znalostí o používateľovi
- Prispôsobovanie sa používateľovi

Kontext a ciele prípravy dát (2)

- Dáta z domény vedeckých publikácií
  - Dostatočne komplexné vzťahy
  - Dobrá prístupnosť dát (Springer, ACM...)
  - Nám blízka doména
- Údaje reprezentované pomocou ontológie (formát OWL-DL)

Kontext a ciele prípravy dát (3)

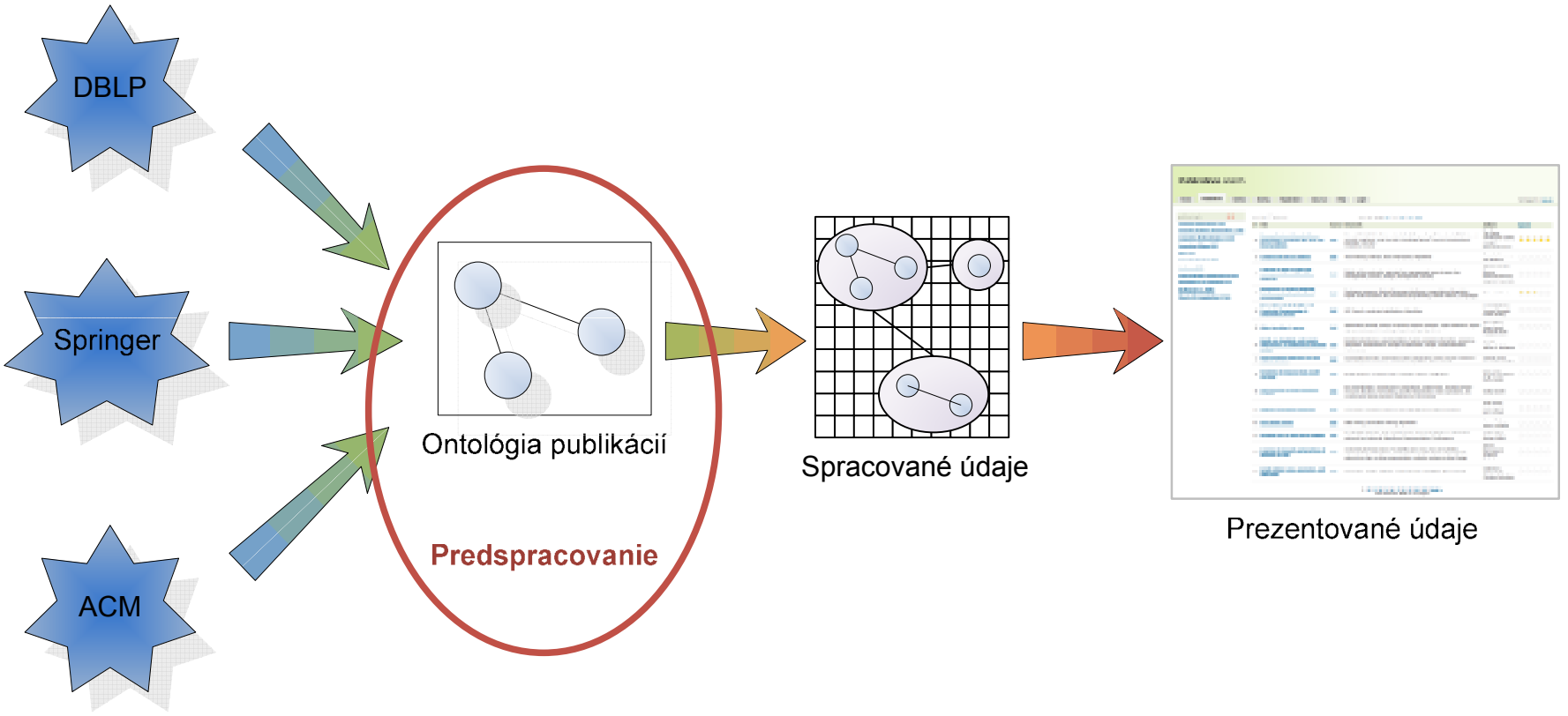


Prečo je potrebné predspracovanie ?

- Duplicity a nepresnosti v zdrojoch dát
- Nepresnosti vytvorené pri získavaní dát
  - Obaľovač z časových dôvodov nemôže „prezrieť“ všetky odkazy – napr. môže zle priradiť publikáciu v prípade autorov s rovnakým menom
- Duplicity spôsobené integráciou dát z viacerých zdrojov

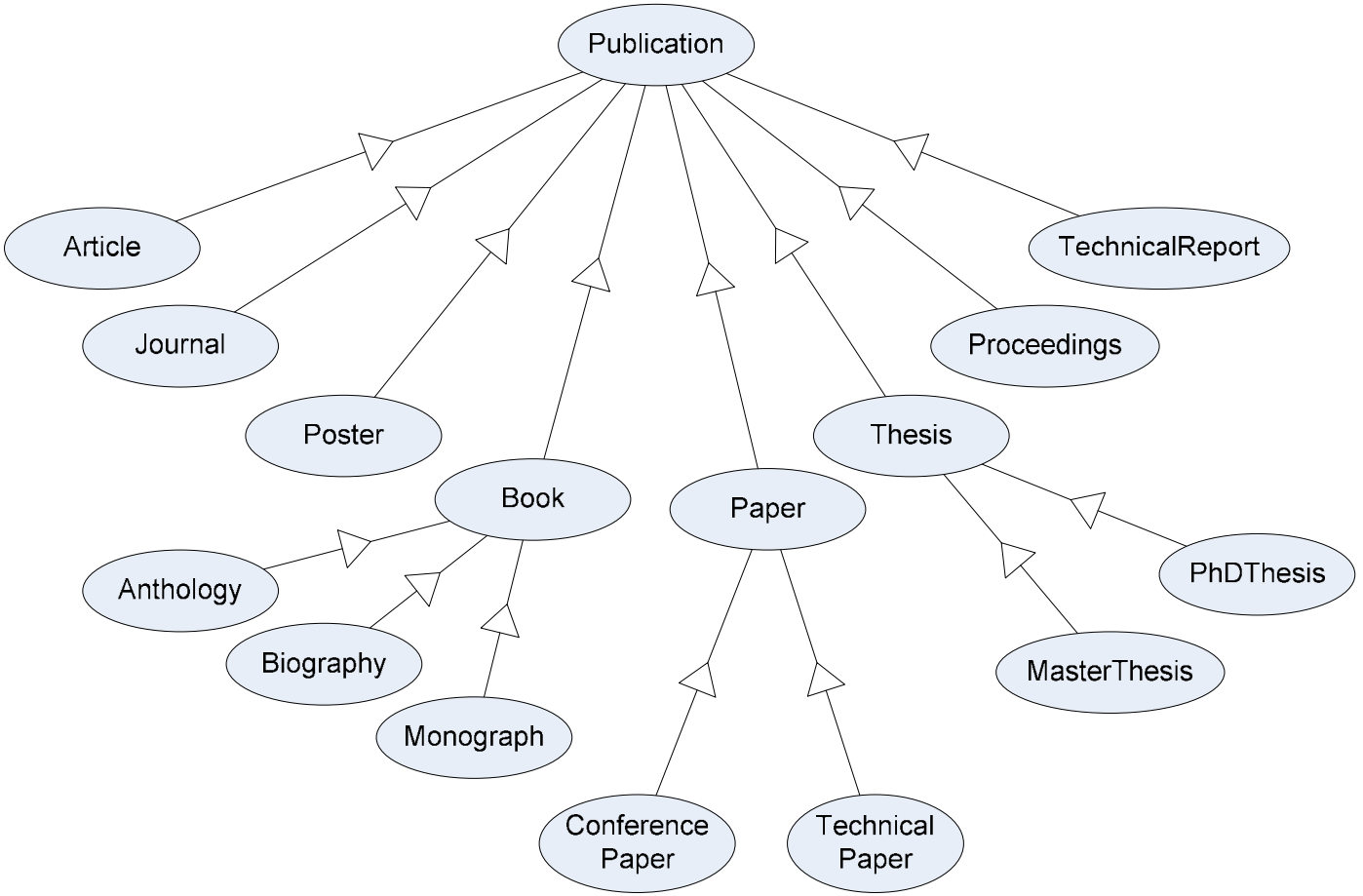
- Úloha predspracovania : riešiť tieto problémy

Predspracovanie v kontexte prípravy dát



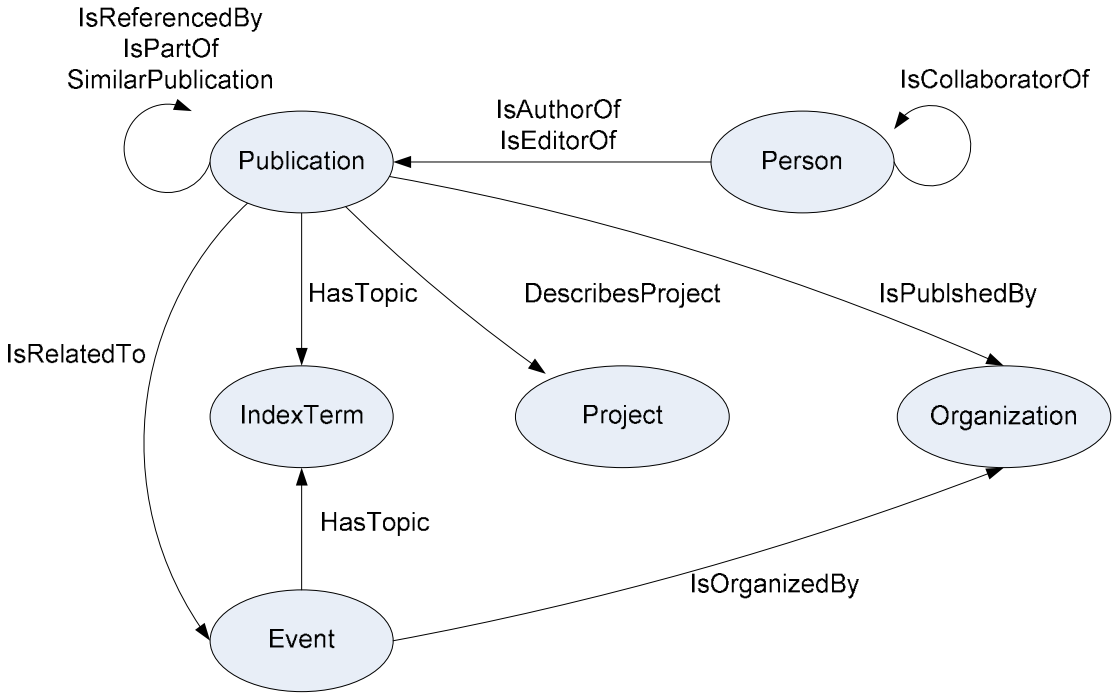
Ontológia vedeckých publikácií

- Hierarchia publikácií



Ontológia vedeckých publikácií (2)

- Autor, editor, organizácia
- Udalosť, projekt
- Kľúčové slová – hierarchia prevzatá z ACM



Ontológia vedeckých publikácií (3)

- Získané údaje:

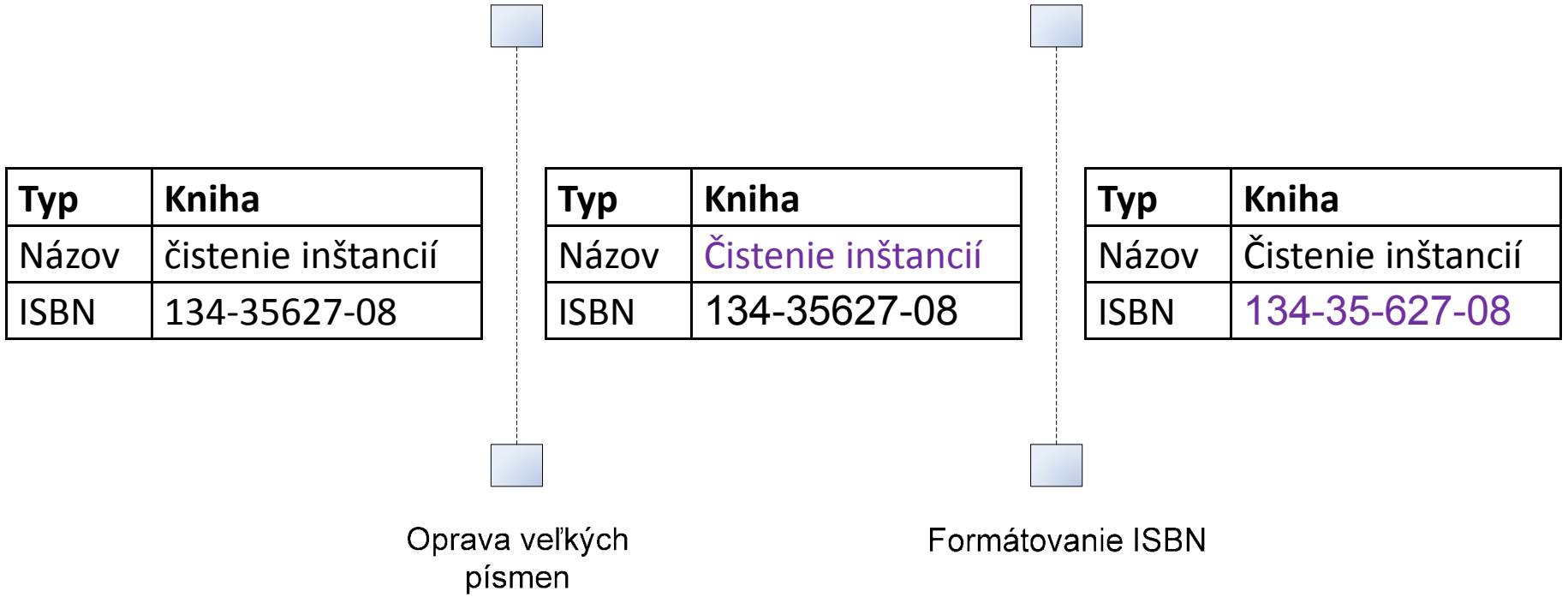
Typ inštancie	ACM	DBLP	Springer
Autor	126 589	69 996	57 504
Organizácia	17 161	-	6 232
Publikácia	48 854	47 854	35 442
Kľúčové slovo	49 182	-	-
Referencia	454 997	-	-

Predspracovanie dát

- Jednoprechodové čistenie inštancií
- Odstraňovanie duplicít

Čistenie inštancií

- Dátové relácie – formátovanie, oddeľovanie...
- Systémom filtrov



Odstraňovanie duplicít

- Potrebné vedieť odhaliť duplicitné inštancie
- Dve inštancie, ktoré popisujú to isté (autora, publikáciu...)
- Kombinácia dvoch metód porovnávania
  - Na základe dátových relácií
  - Na základe objektových relácií
- Ak je výsledná hodnota vyššia ako prahová - duplicita

Porovnávanie dátových relácií

- Porovnávajú sa inštancie rovnakých typov
- Zjednodušenie: predpokladá sa funkcionálny vzťah dátových relácií
- Využitie metrík pre určenie podobnosti reťazcov
- 15 štandardných metrík
  - Levenstein, Qgramy, Monge-Elkan...
- Vlastné spôsoby porovnávania

Vzdialenosť na klávesnici

- Pre nezhodné písmená v reťazci sa berie do úvahy vzdialenosť na klávesnici
- Berie sa do úvahy shift, prepnutie inej klávesnice

Porovnávanie mien

- „Obalenie“ Levensteinovej metriky
- Pre mená a priezviská
- Berú sa do úvahy skratky a chýbajúce časti
- Príklad.:
  - Michael von Schwarzwald
  - M. Schwarzwald
  - Vyhodnotí ako rovnaké

Zložená metrika

- Kombinácia viacerých metrík
- Každú metriku možno váhovať
- Priestor pre experimentovanie

- Napríklad: 0.7 Levenstein + 0.3 Monge-Elkan

Porovnávanie dátových relácií (2)

- Každú reláciu je možné porovnávať pomocou inej metriky
- Vplyv porovnania jednej relácie je možné váhovať

Porovnávanie objektových relácií

- Koľko inštancií v rámci nejakej relácie je podobných
- Napríklad koľko kníh majú porovnávaní autori podobných
- Využíva výsledky porovnávania dátových relácií
- Každú objektovú reláciu je možné váhovať

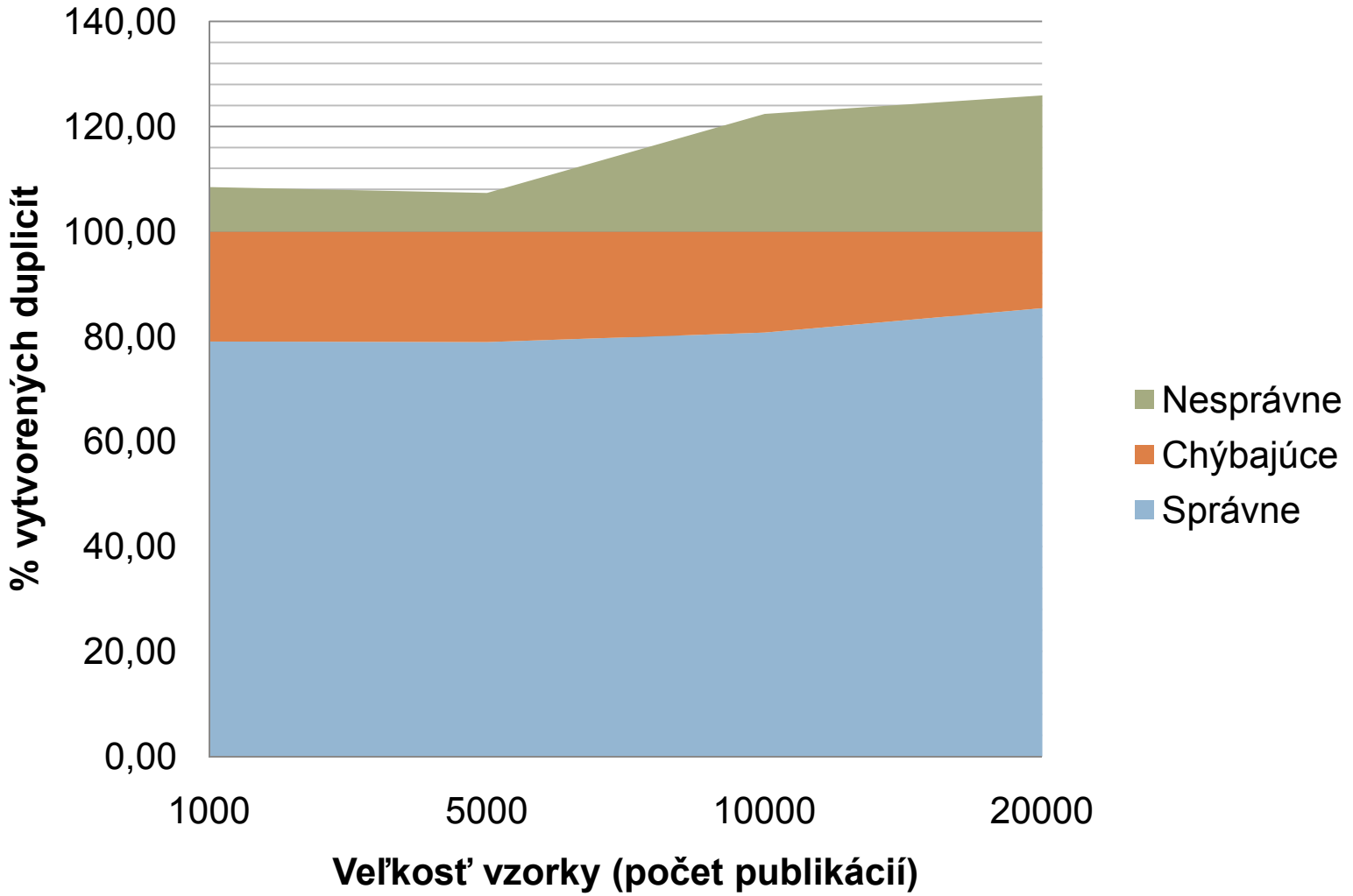
Metódy odstraňovania duplicít

- Označenie duplicity špeciálnou reláciou
- Manuálne
- Vymazanie inštancie s menším množstvom informácií
- Zlúčenie inštancií

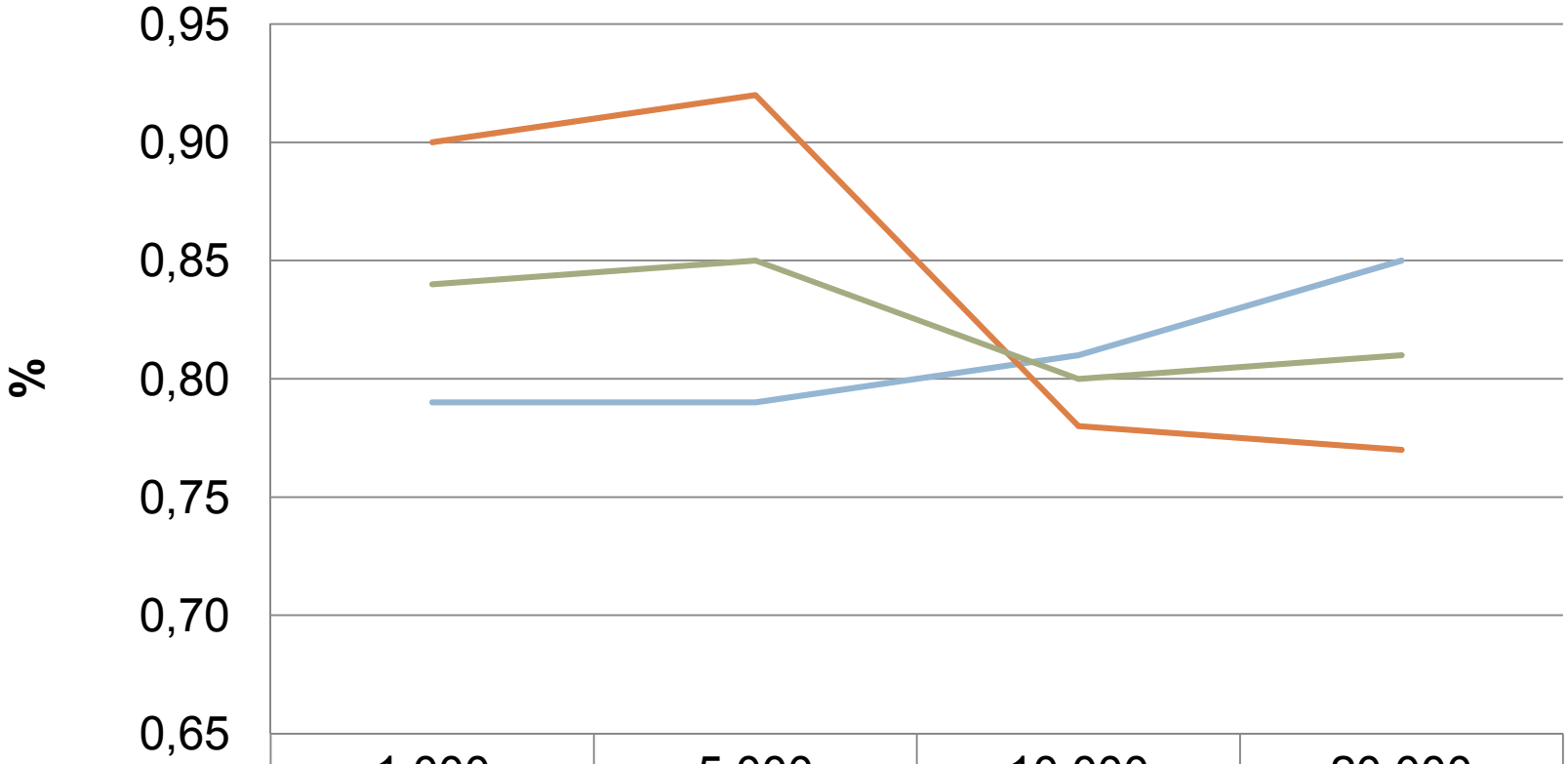
Overenie riešenia – identifikácia duplicít

- Veľkosť vzoriek (počet publikácií):
  - 1, 5, 10, 20 tisíc
- 10 meraní pre každú veľkosť
- Do každej vložených 100 umelo vytvorených duplicít
- Všetky dáta z DBLP
- Nezapočítavame duplicity, ktoré môžu byť v DBLP už prítomné

Overenie riešenia – identifikácia duplicít (2)

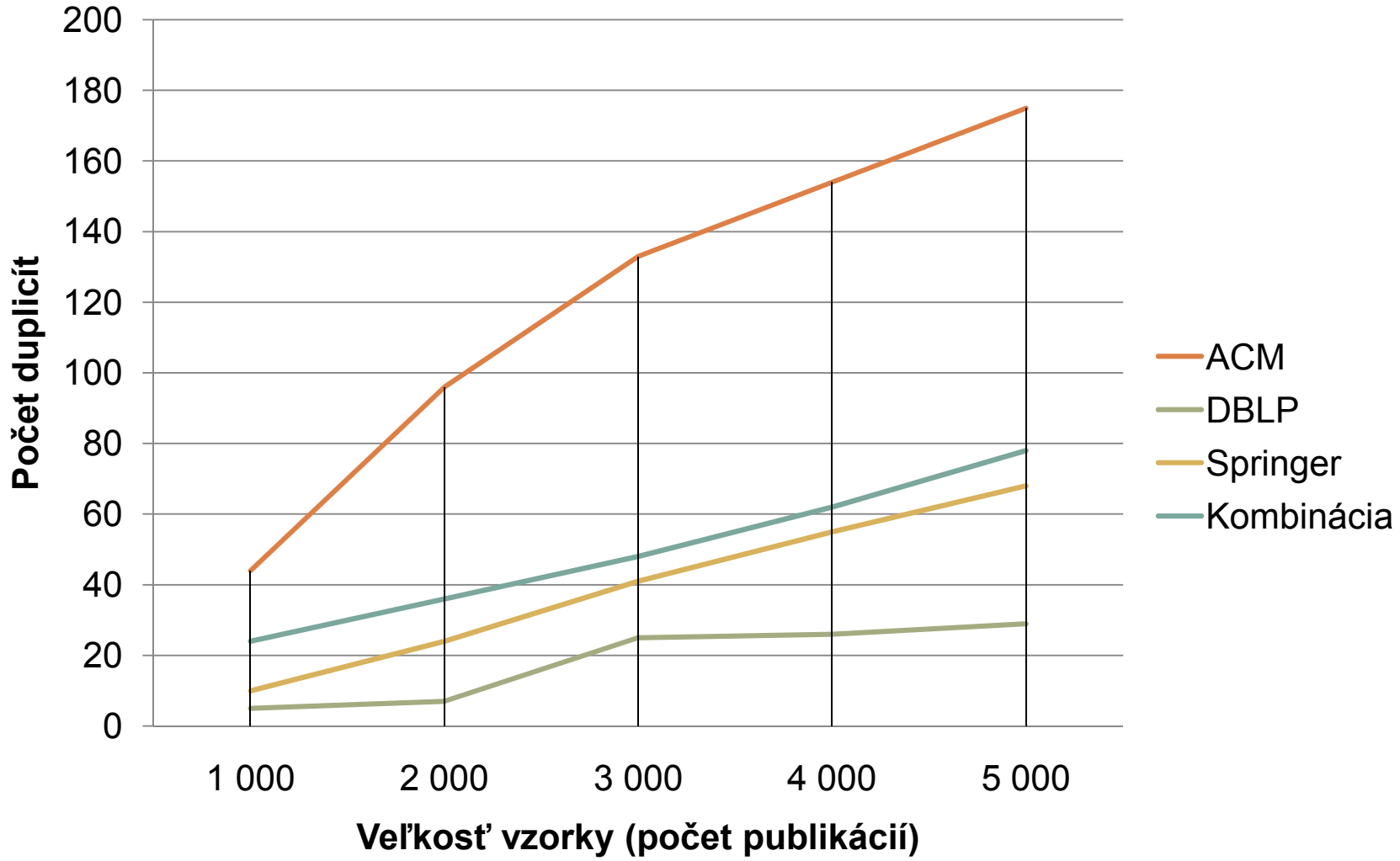


Overenie riešenia – identifikácia duplicít (3)



	1 000	5 000	10 000	20 000
Presnosť	0,79	0,79	0,81	0,85
Pokrytie	0,90	0,92	0,78	0,77
F1	0,84	0,85	0,80	0,81

Overenie riešenia – identifikácia duplicít (4)



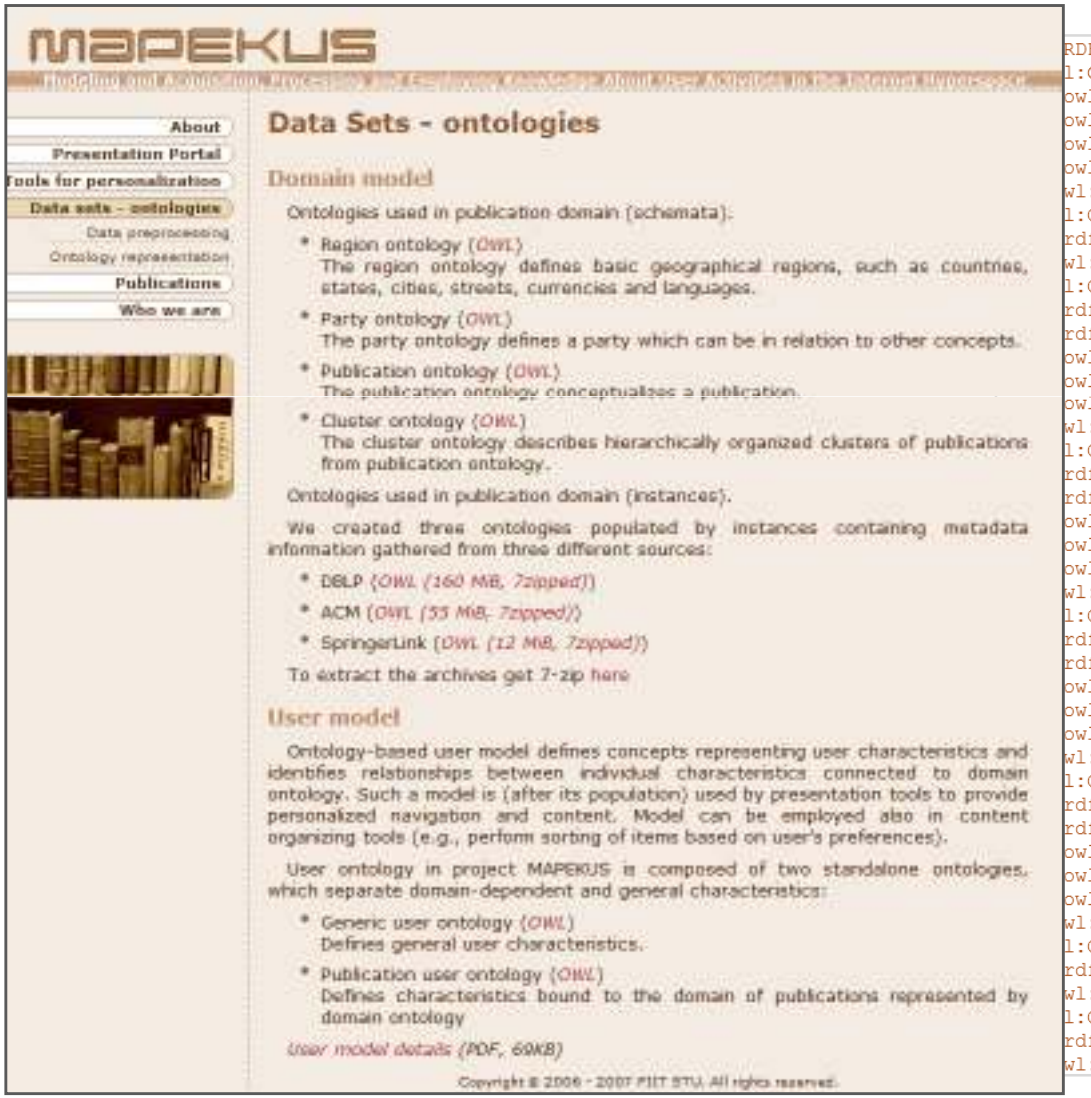
Záver a ďalšia práca

- Zistené skutočnosti
  - Veľa duplicít kvôli rôznym veľkým písmenám
  - Čínske mená a priezviská sa zle porovnávajú – veľmi krátke
  - Rôzne edície a vydania kníh sa ťažko rozpoznávajú, pretože sa ich názvy často líšia len jedným číslom alebo slovom
  - V niektorých prípadoch ani človek nevie bez ďalších informácií zistiť, či sa jedná o duplicitu

Záver a ďalšia práca (2)

- Porovnávanie je časovo veľmi náročné
  - Optimalizovať porovnávanie
- Zlepšovať metódu porovnávania
  - Ďalšie metriky
  - Váhovanie
- Experimentovať na rôznych iných vzorkách

http://mapekus.fiit.stuba.sk/?page=ontologies




```
RDF xmlns="http://mapekus.fiit.stuba.sk/mapekus/ontologies/v(
l:Ontology rdf:about="">
owl:imports rdf:resource="&dc;" />
owl:imports rdf:resource="&r;" />
owl:imports rdf:resource="&p;" />
owl:imports rdf:resource="&c;" />
wl:Ontology>
l:Class rdf:ID="Event">
rdfs:label xml:lang="en">Event</rdfs:label>
wl:Class>
l:Class rdf:ID="Activity">
rdfs:label xml:lang="en">Activity</rdfs:label>
rdfs:subClassOf rdf:resource="#Event" />
owl:disjointWith rdf:resource="#Conference" />
owl:disjointWith rdf:resource="#Meeting" />
owl:disjointWith rdf:resource="#Workshop" />
wl:Class>
l:Class rdf:ID="Conference">
rdfs:label xml:lang="en">Conference</rdfs:label>
rdfs:subClassOf rdf:resource="#Event" />
owl:disjointWith rdf:resource="#Activity" />
owl:disjointWith rdf:resource="#Meeting" />
owl:disjointWith rdf:resource="#Workshop" />
wl:Class>
l:Class rdf:ID="Meeting">
rdfs:label xml:lang="en">Meeting</rdfs:label>
rdfs:subClassOf rdf:resource="#Event" />
owl:disjointWith rdf:resource="#Activity" />
owl:disjointWith rdf:resource="#Conference" />
owl:disjointWith rdf:resource="#Workshop" />
wl:Class>
l:Class rdf:ID="Workshop">
rdfs:label xml:lang="en">Workshop</rdfs:label>
rdfs:subClassOf rdf:resource="#Event" />
owl:disjointWith rdf:resource="#Activity" />
owl:disjointWith rdf:resource="#Conference" />
owl:disjointWith rdf:resource="#Meeting" />
wl:Class>
l:Class rdf:ID="IndexTerm">
rdfs:label xml:lang="en">Index term</rdfs:label>
wl:Class>
l:Class rdf:ID="Keyword">
rdfs:label xml:lang="en">Keyword</rdfs:label>
wl:Class>
```