

# PRÍPRAVA DÁT NA ANALÝZU A PERSONALIZOVANÚ PREZENTÁCIU NA WEBE V DOMÉNE VEDECKÝCH PUBLIKÁCIÍ

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

---

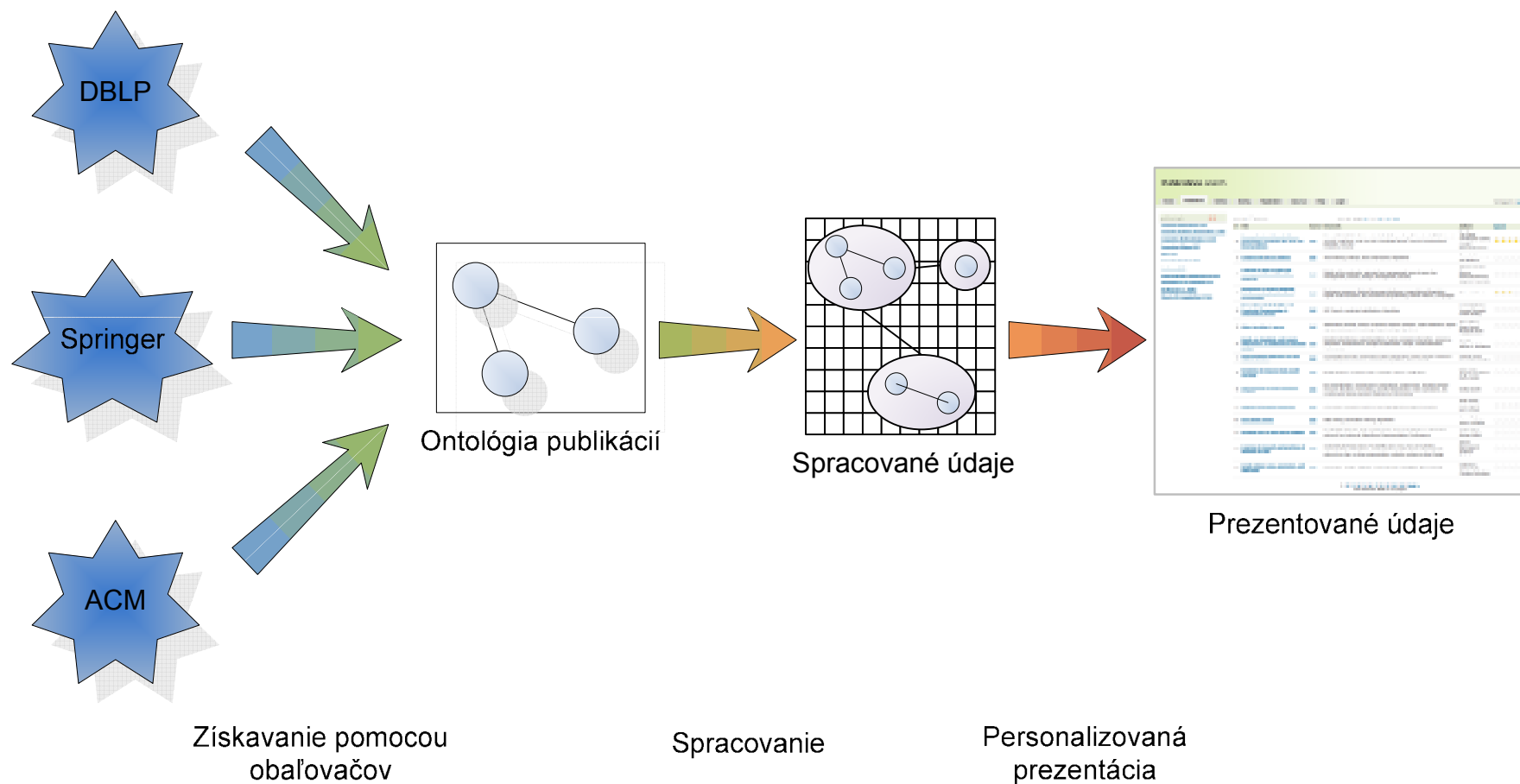
- V rámci výskumného projektu MAPEKUS (Modeling and Acquisition, Processing and Employing Knowledge about User Activities in the Internet Hyperspace)
- Personalizovaná navigácia vo veľkých informačných priestoroch
- Získavanie znalostí o používateľovi
- Prispôsobovanie sa používateľovi

# 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)

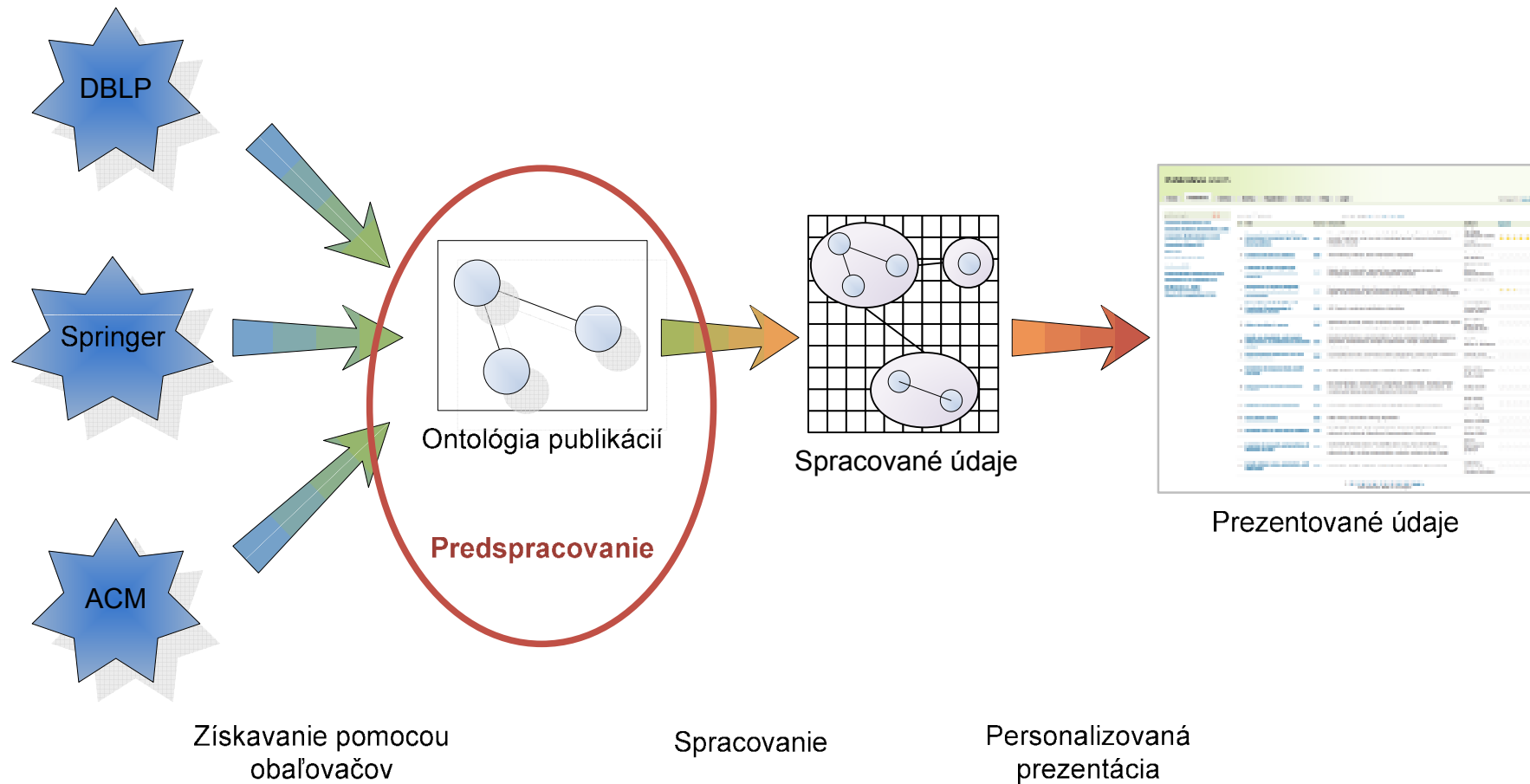
# Kontext a ciele prípravy dát (3)



# Prečo je potrebné predspracovanie ?

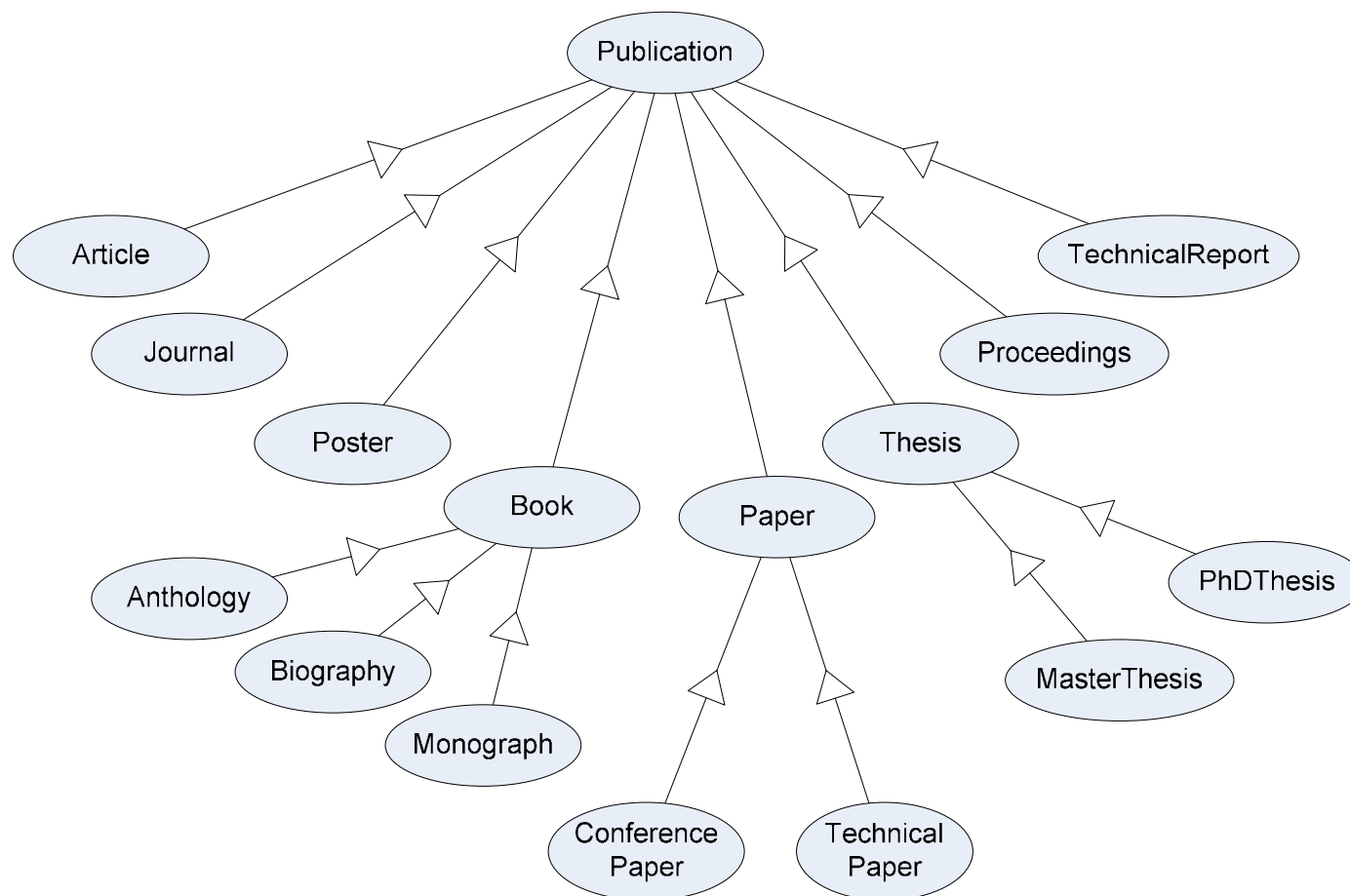
- Duplicity a nepresnosti v zdrojoch dát
- Nepresnosti vytvorené pri získavaní dát
  - ▣ Obaľovač z časových dôvodov nemôže „prezrieť“ všetky odkazy – napr. môže zle priradiť publikáciu v prípade autorov s rovnakým menom
- Duplicity spôsobené integráciou dát z viacerých zdrojov
  
- Úloha predspracovania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

# Predspracovanie v kontexte prípravy dát



# Ontológia vedeckých publikácií

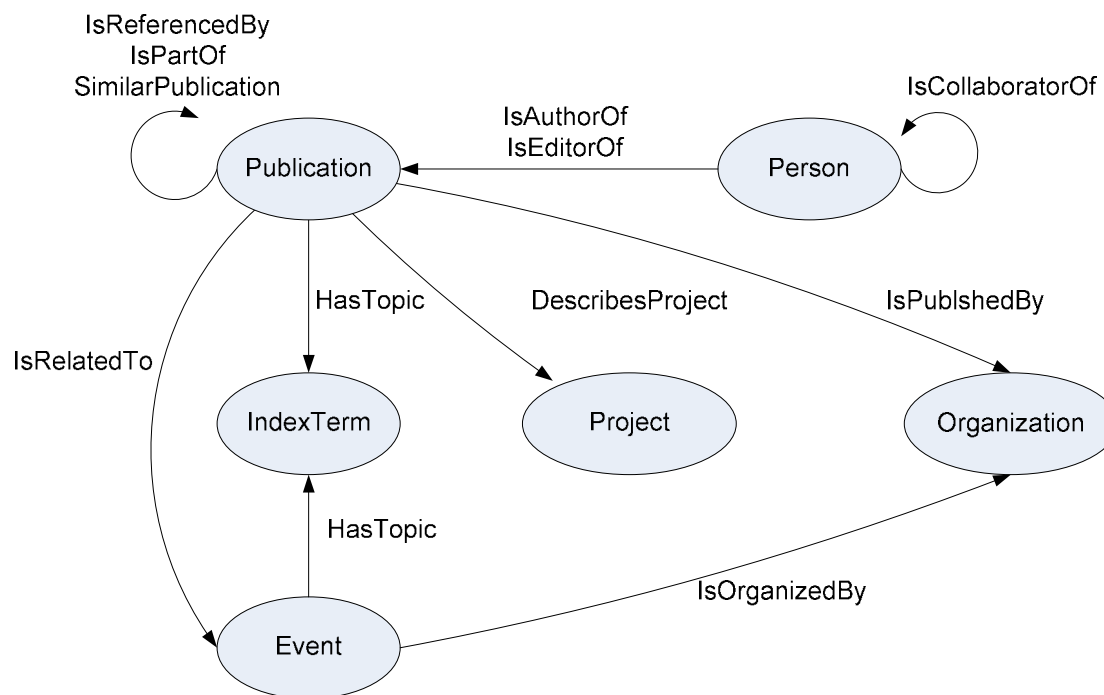
## □ Hierarchia publikácií





# 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	-	-

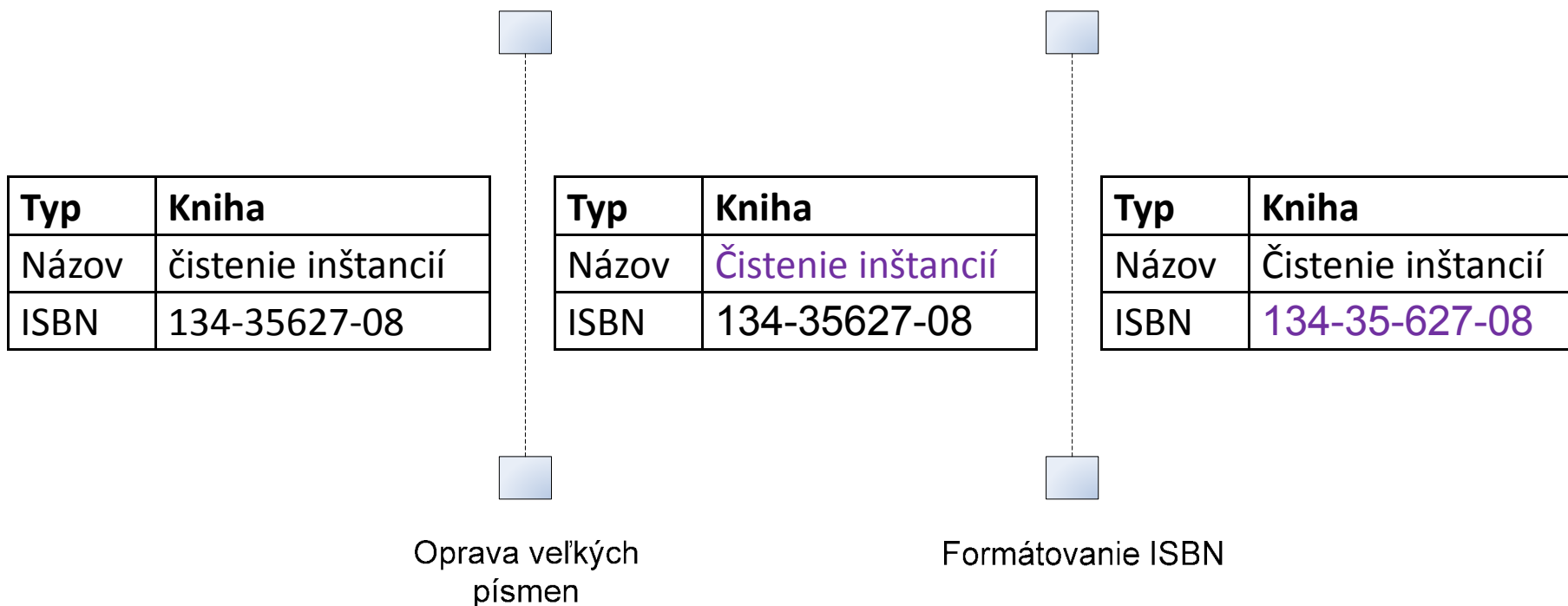
# Predspracovanie dát



- Jednoprechodové čistenie inštancií
- Odstraňovanie duplicit

# Čistenie inštancií

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



# Odstraňovanie duplicit

- 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 porovnávanania

# 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 duplicit

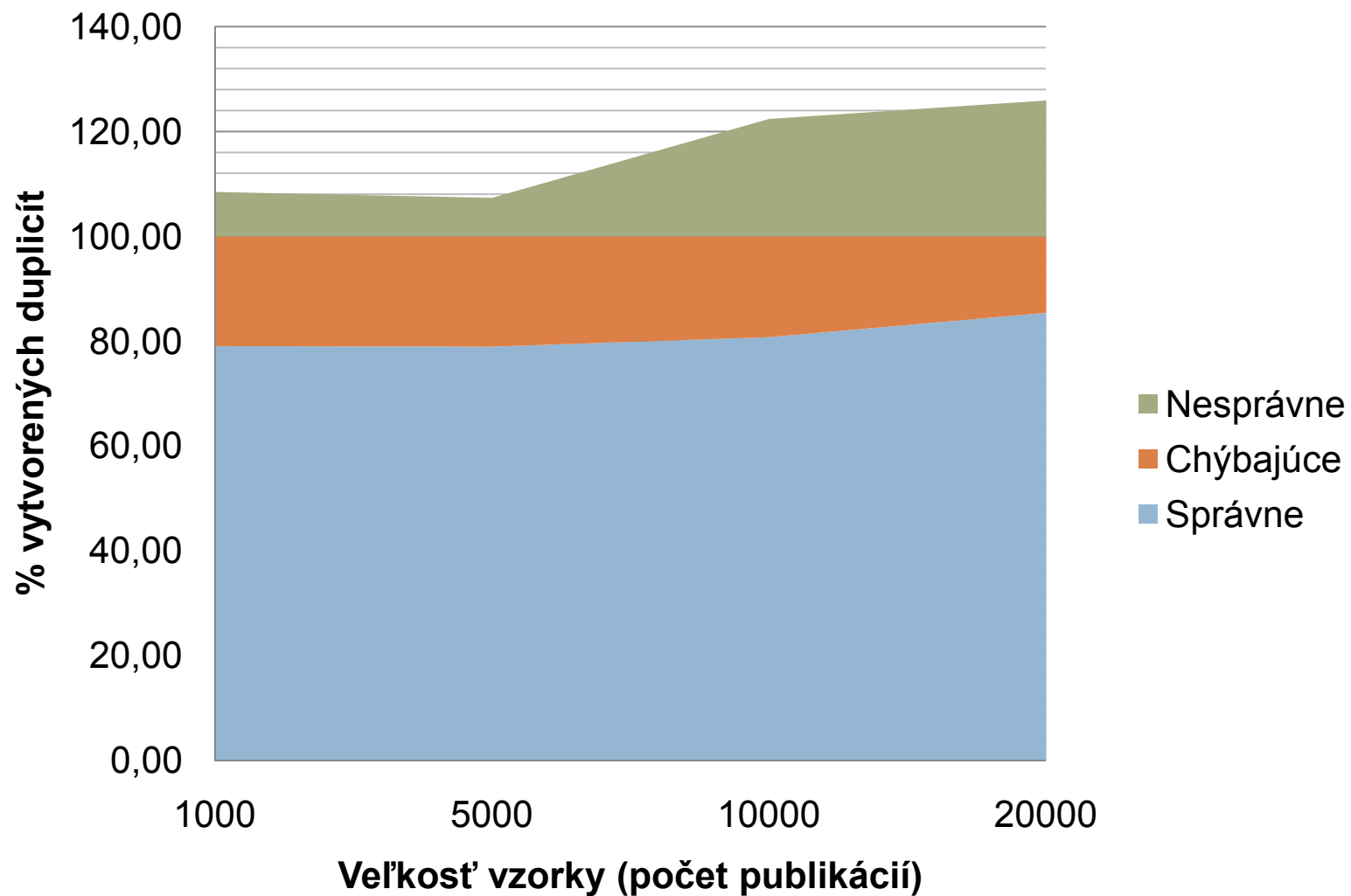


- Označenie duplicity špeciálnou reláciou
- Manuálne
- Vymazanie inštancie s menším množstvom informácií
- Zlúčenie inštancií

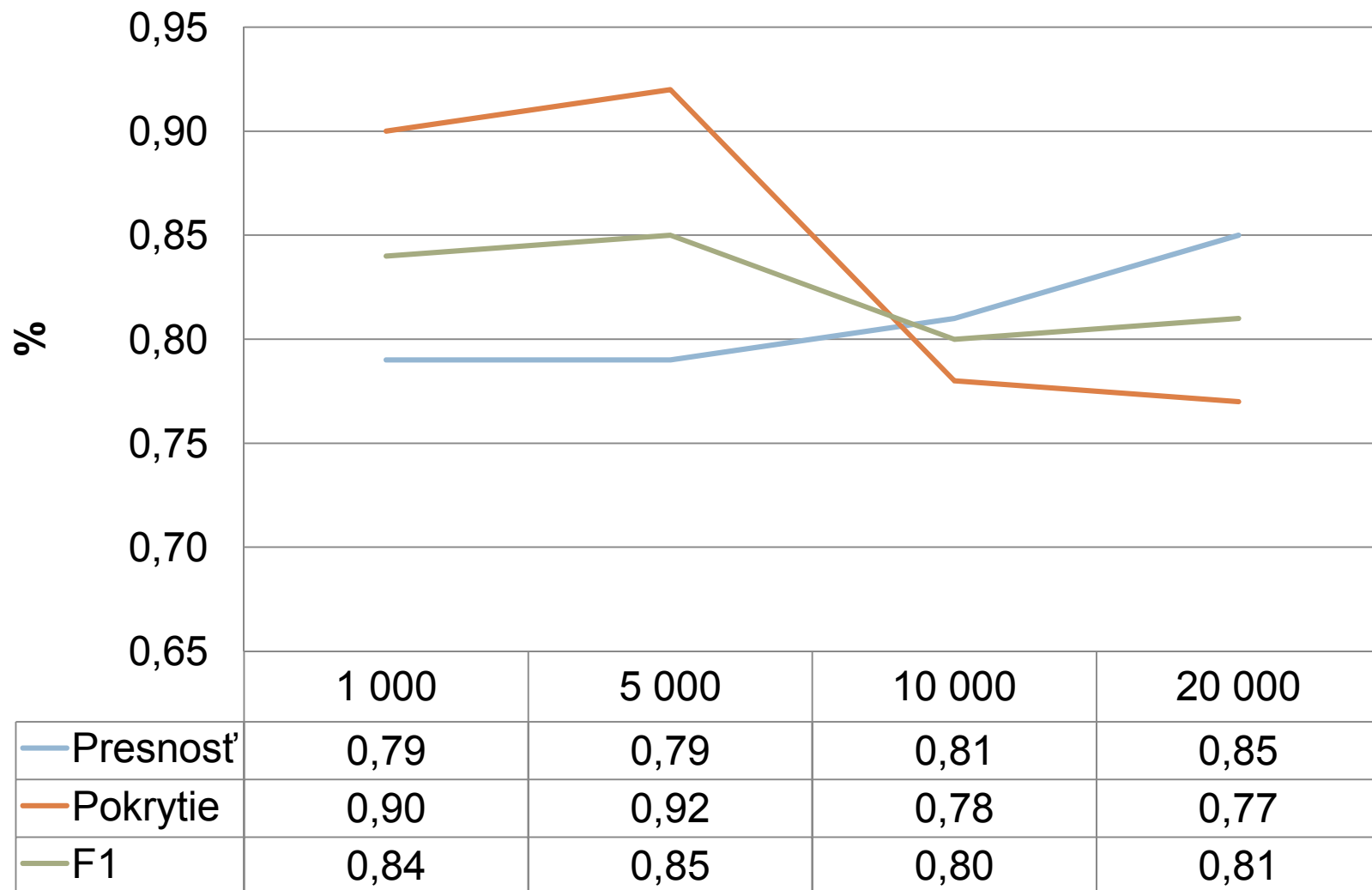
# Overenie riešenia – identifikácia duplicit

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

# Overenie riešenia – identifikácia duplicit (2)

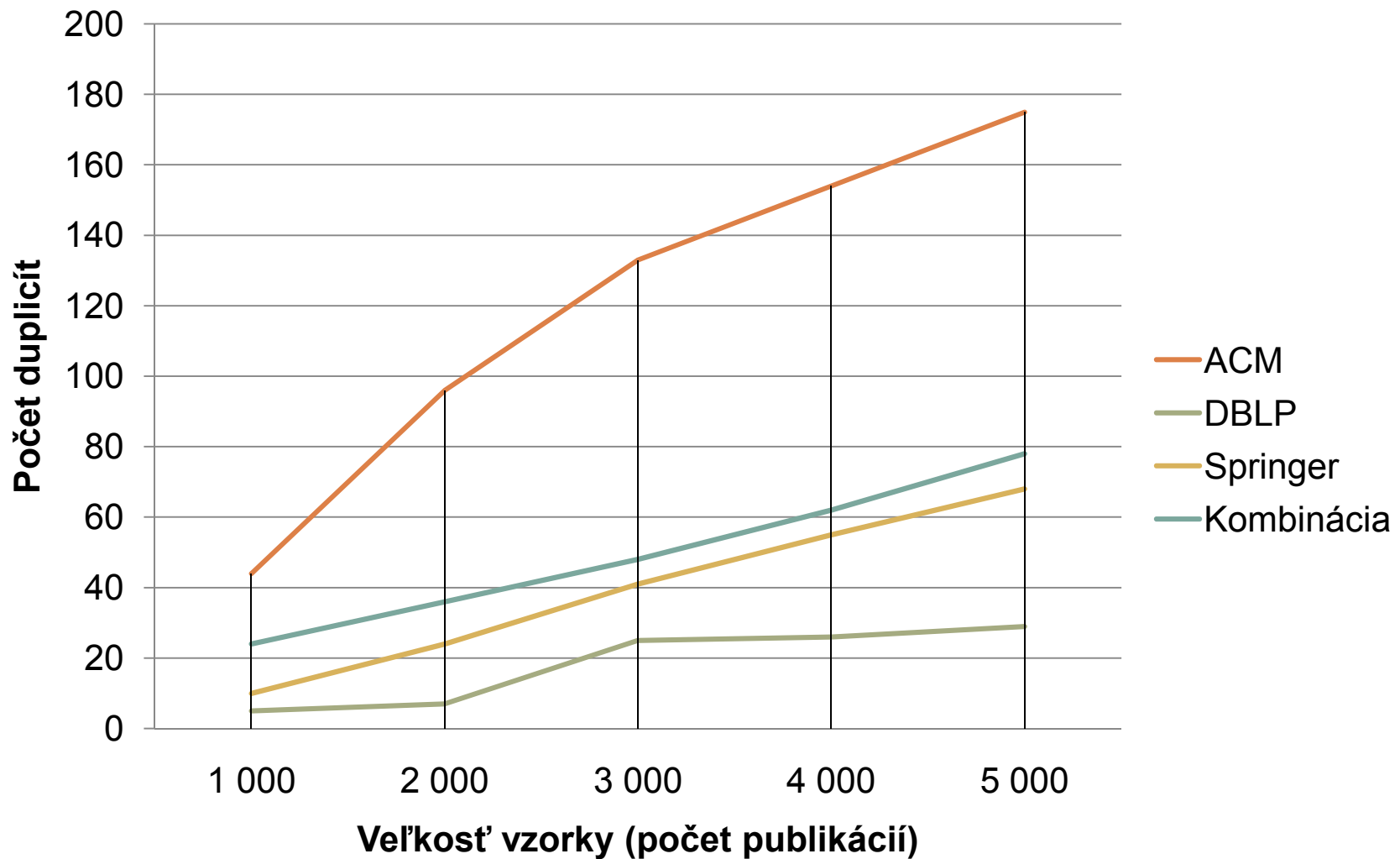


# Overenie riešenia – identifikácia duplicit (3)



# Overenie riešenia – identifikácia duplicit

## (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 rozoznávajú, 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>

**MAPEKUS**

**Data Sets - ontologies**

**Domain model**

Ontologies used in publication domain (schemata):

- **Region ontology (OWL)**  
The region ontology defines basic geographical regions, such as countries, states, cities, streets, currencies and languages.
- **Party ontology (OWL)**  
The party ontology defines a party which can be in relation to other concepts.
- **Publication ontology (OWL)**  
The publication ontology conceptualizes a publication.
- **Cluster ontology (OWL)**  
The cluster ontology describes hierarchically organized clusters of publications from publication ontology.

Ontologies used in publication domain (instances):

We created three ontologies populated by instances containing metadata information gathered from three different sources:

- **DBLP (OWL (160 MB, 7zipped))**
- **ACM (OWL (55 MB, 7zipped))**
- **SpringerLink (OWL (12 MB, 7zipped))**

To extract the archives get 7-zip here

**User model**

Ontology-based user model defines concepts representing user characteristics and identifies relationships between individual characteristics connected to domain ontology. Such a model is (after its population) used by presentation tools to provide personalized navigation and content. Model can be employed also in content organizing tools (e.g., perform sorting of items based on user's preferences).

User ontology in project MAPEKUS is composed of two standalone ontologies, which separate domain-dependent and general characteristics:

- **Generic user ontology (OWL)**  
Defines general user characteristics.
- **Publication user ontology (OWL)**  
Defines characteristics bound to the domain of publications represented by domain ontology

[User model details \(PDF, 60KB\)](#)

Copyright © 2006 - 2007 FIIT STU. All rights reserved.

```

RDF xmlns="http://mapekus.fiit.stuba.sk/mapekus/ontologies/v0
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>

```