Open Corpus User Modeling

Michal Barla
Ontožúr 18.4.2010


HAVRANIA SKALA
554 m

Open Corpus User Modeling

It is not a brain surgery...

Open Corpus **User Modeling**

- collect and analyze data to produce User Model
  - knowledge
  - goals
  - preferences
  - interests
  - experience
  - background
  - ...

Open Corpus **User Modeling**

- collect and analyze data to produce User Model
  - knowledge
  - goals
  - preferences
  - interests
  - experience
  - background
  - ...

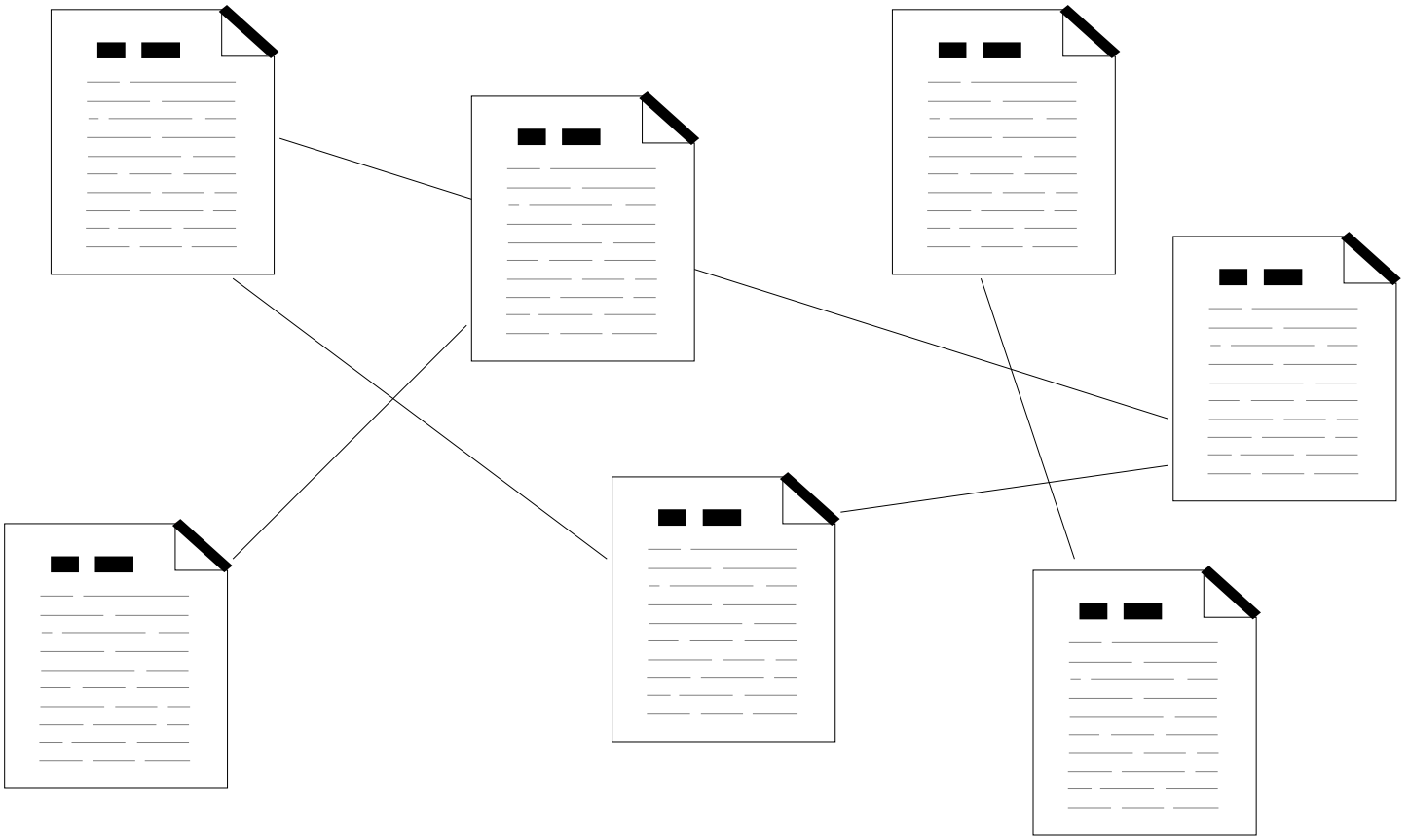
Characteristics are related to domain elements

Open Corpus **User Modeling**

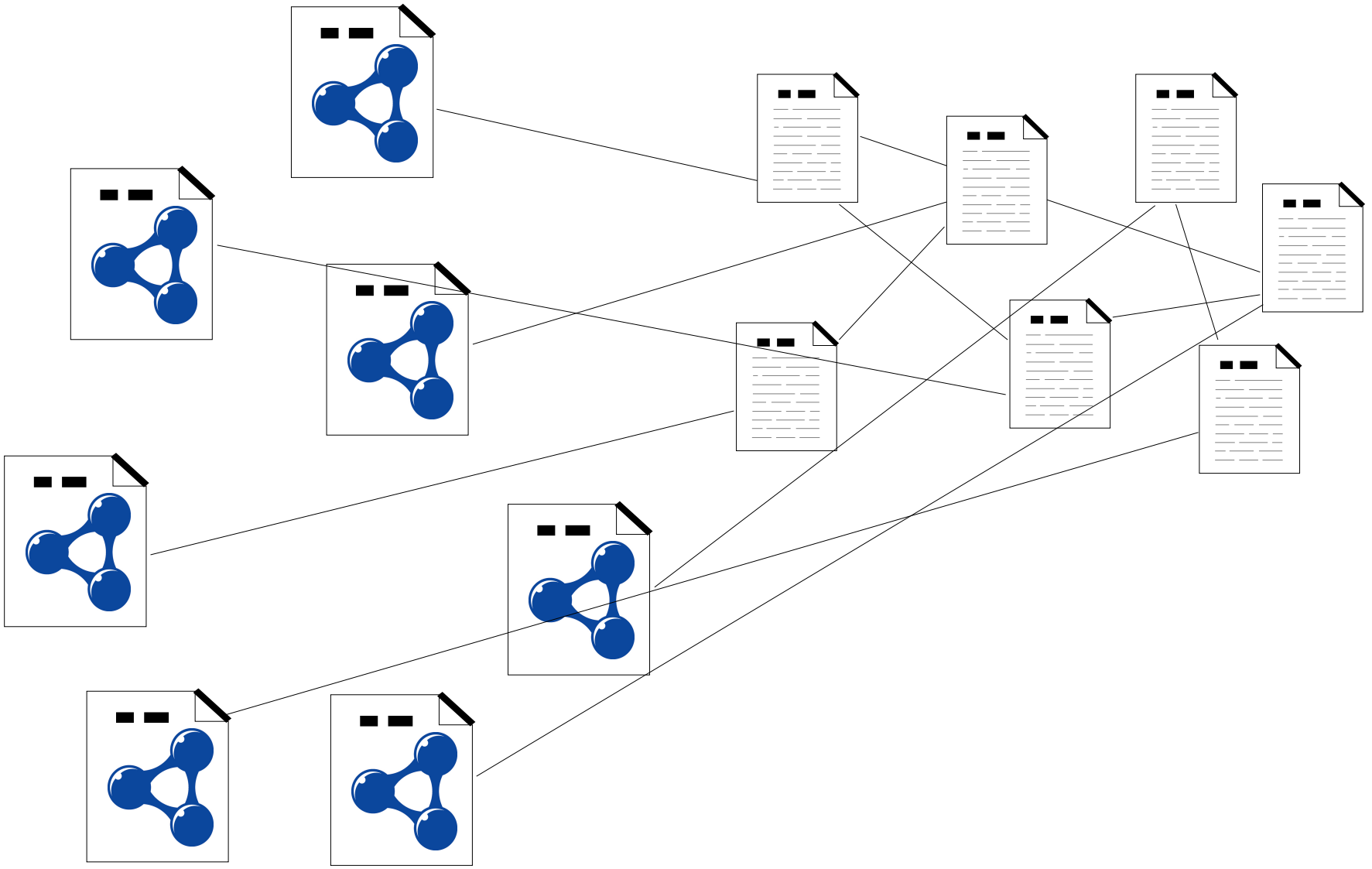
- collect and analyze data to produce User Model
  - knowledge
  - goals
  - preferences
  - interests
  - experience
  - background
  - ...

Characteristics
are related to
domain elements
=
Overlayed UM

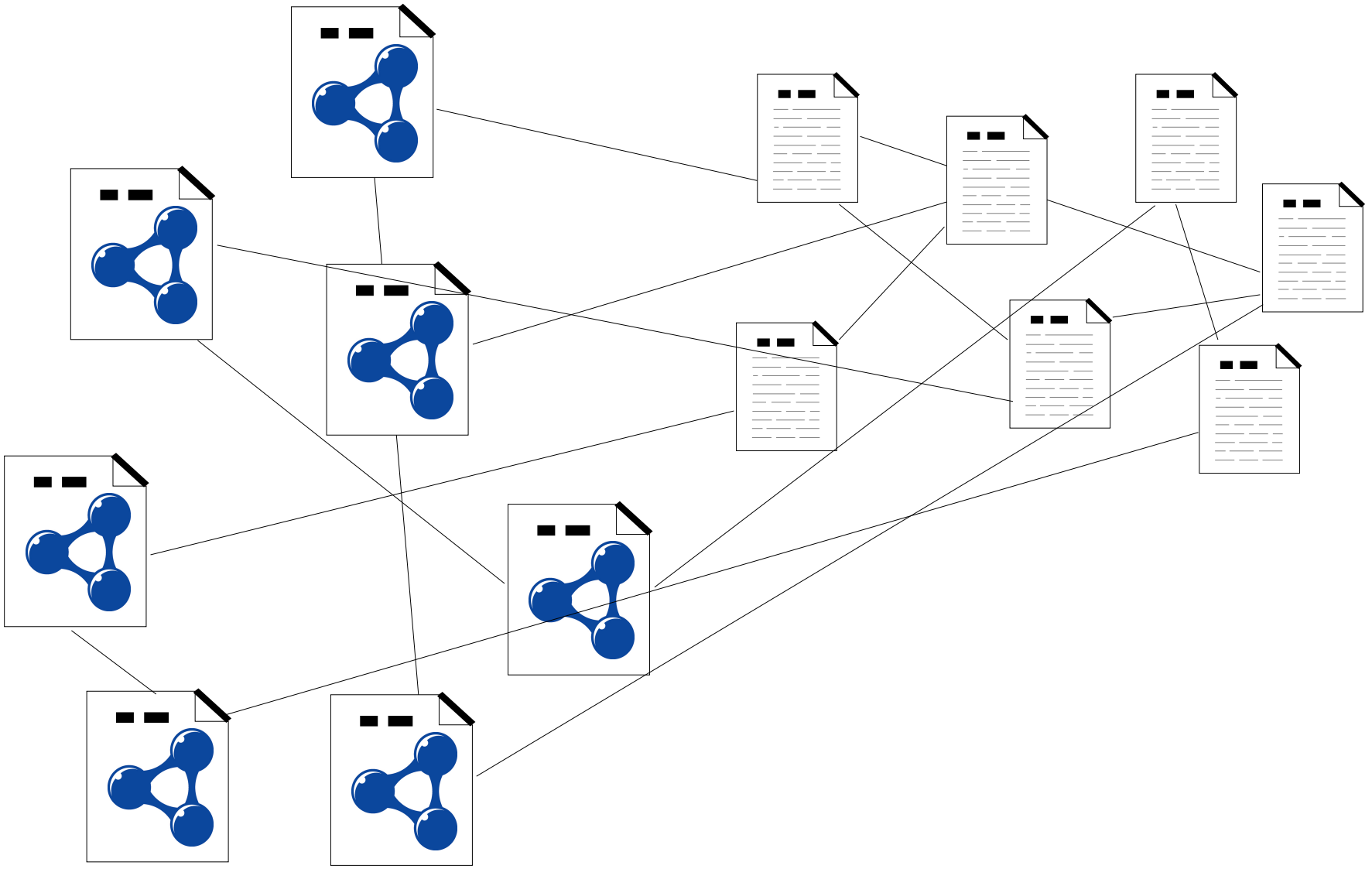
Closed Corpus User Modeling



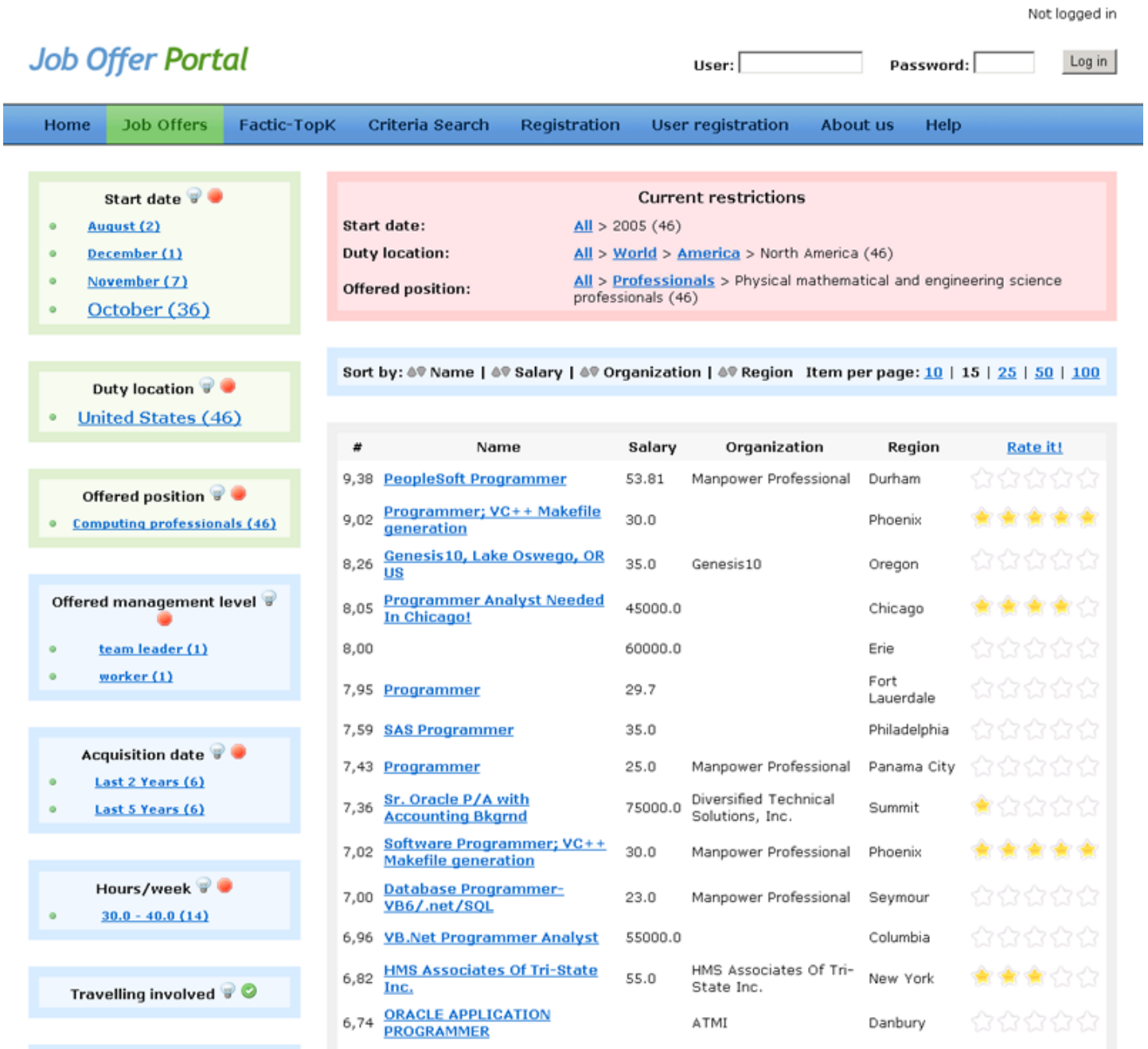
Open Corpus User Modeling



Open Corpus User Modeling



Example 1 – Job offers



Not logged in

Job Offer Portal

User: Password: Log in

Home | Job Offers | Factic-TopK | Criteria Search | Registration | User registration | About us | Help

Start date
- August (2)
- December (1)
- November (7)
- October (36)

Duty location
- United States (46)

Offered position
- Computing professionals (46)

Offered management level
- team leader (1)
- worker (1)

Acquisition date
- Last 2 Years (6)
- Last 5 Years (6)

Hours/week
- 30.0 - 40.0 (14)

Travelling involved

Current restrictions

Start date:	All > 2005 (46)
Duty location:	All > World > America > North America (46)
Offered position:	All > Professionals > Physical mathematical and engineering science professionals (46)

Sort by: Name | Salary | Organization | Region Item per page: 10 | 15 | 25 | 50 | 100

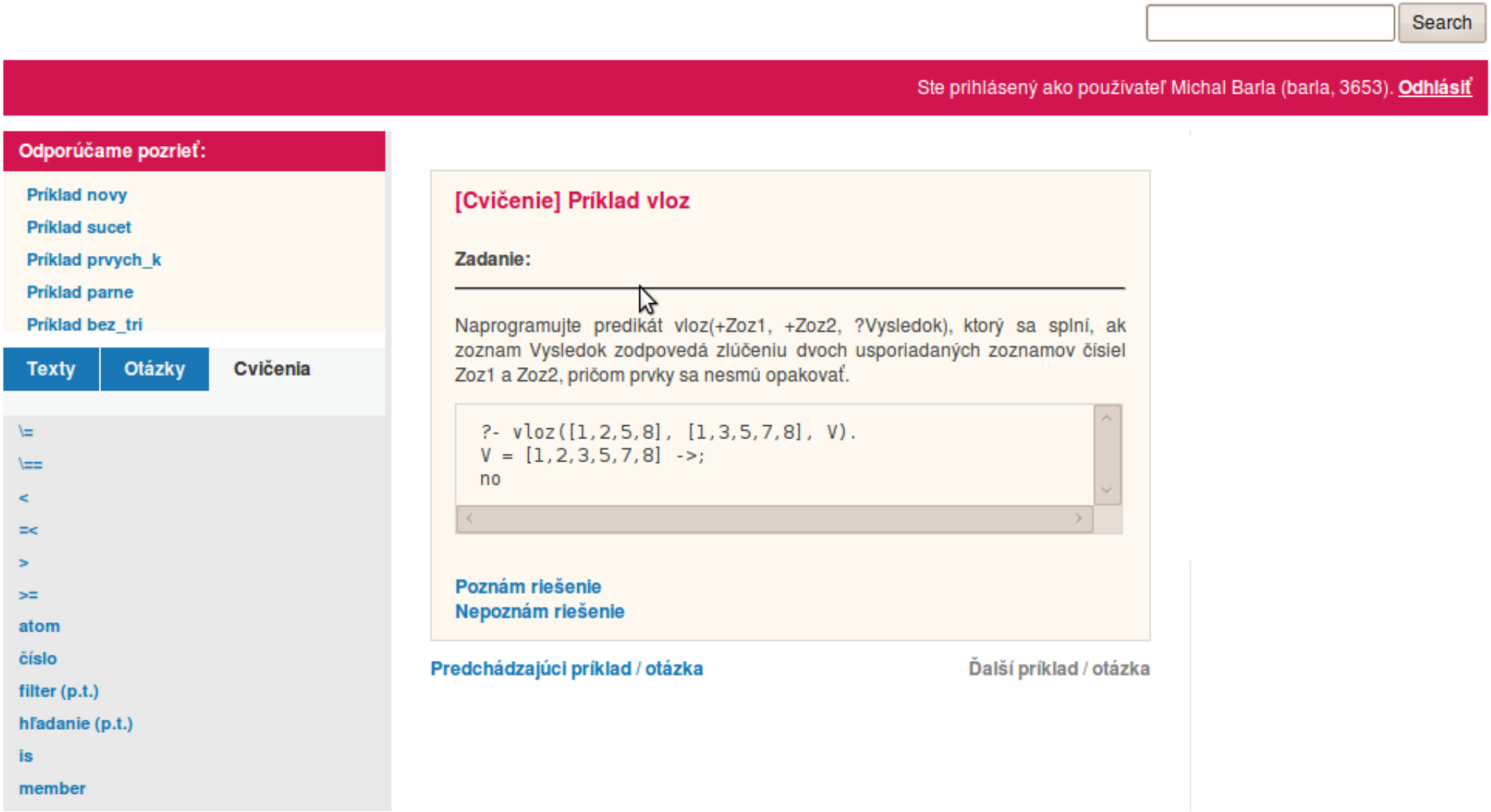
#	Name	Salary	Organization	Region	Rate it!
9,38	PeopleSoft Programmer	53.81	Manpower Professional	Durham	
9,02	Programmer; VC++ Makefile generation	30.0		Phoenix	
8,26	Genesis10, Lake Oswego, OR US	35.0	Genesis10	Oregon	
8,05	Programmer Analyst Needed In Chicago!	45000.0		Chicago	
8,00		60000.0		Erie	
7,95	Programmer	29.7		Fort Lauerdale	
7,59	SAS Programmer	35.0		Philadelphia	
7,43	Programmer	25.0	Manpower Professional	Panama City	
7,36	Sr. Oracle P/A with Accounting Bkgrnd	75000.0	Diversified Technical Solutions, Inc.	Summit	
7,02	Software Programmer; VC++ Makefile generation	30.0	Manpower Professional	Phoenix	
7,00	Database Programmer-VB6/.net/SQL	23.0	Manpower Professional	Seymour	
6,96	VB.Net Programmer Analyst	55000.0		Columbia	
6,82	HMS Associates Of Tri-State Inc.	55.0	HMS Associates Of Tri-State Inc.	New York	
6,74	ORACLE APPLICATION PROGRAMMER		ATMI	Danbury	

Example 1 – Job offers

- Each job offer has
  - duty location
  - salary
  - position
  - requirements

- Instead of modeling user's attitudes towards particular job offer, we capture user preferences of different types of attributes

Example 2 – ALEF



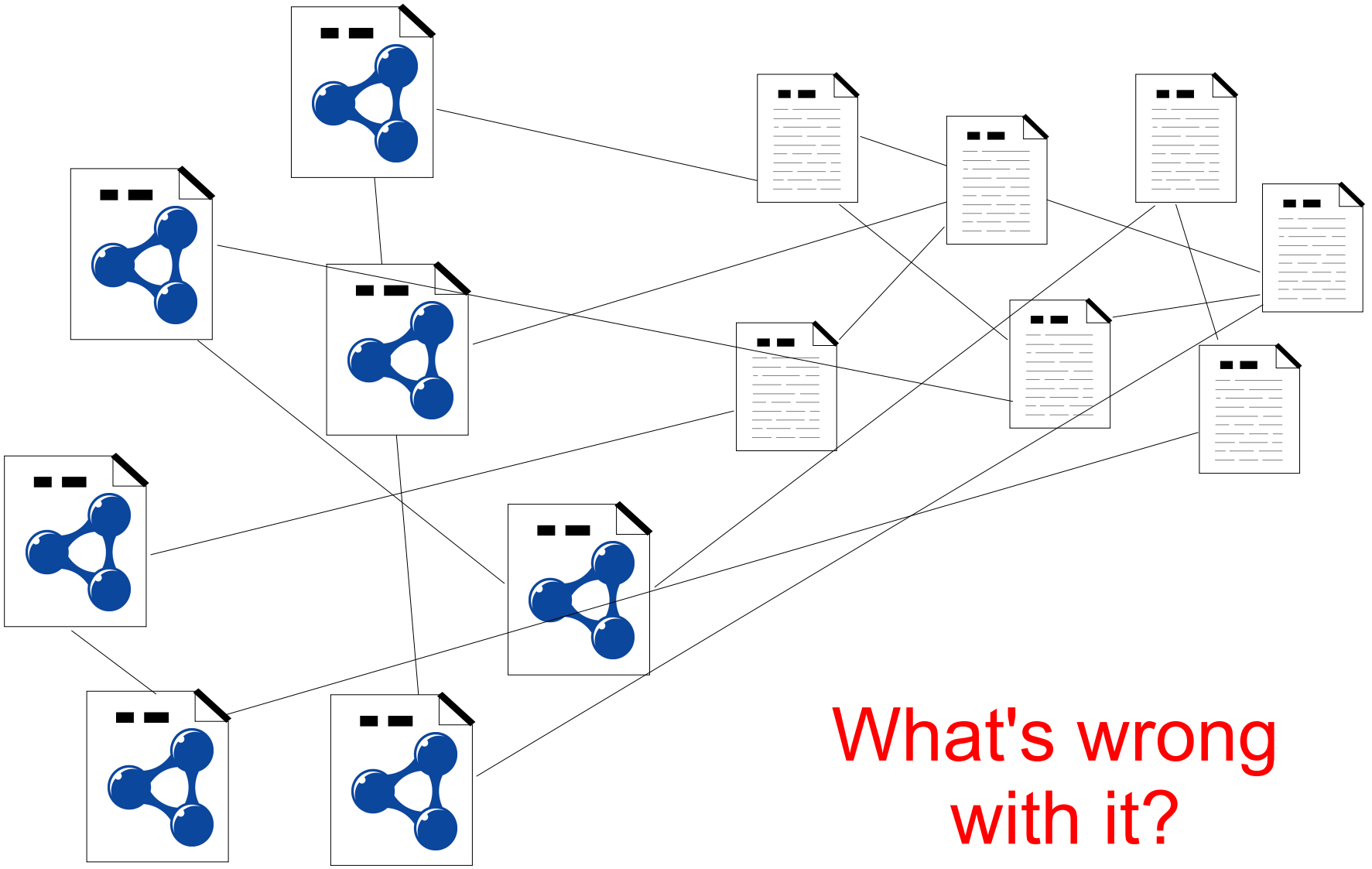
Example 2 - ALEF

- User is presented with learning objects (LO)
  - explanations
  - questions
  - exercices
- Learning objects are mapped to concepts
- Various concept-to-concept relationships
  - prerequisity
  - parent – child
- We model user knowledge of particular concepts

Open Corpus User Modeling

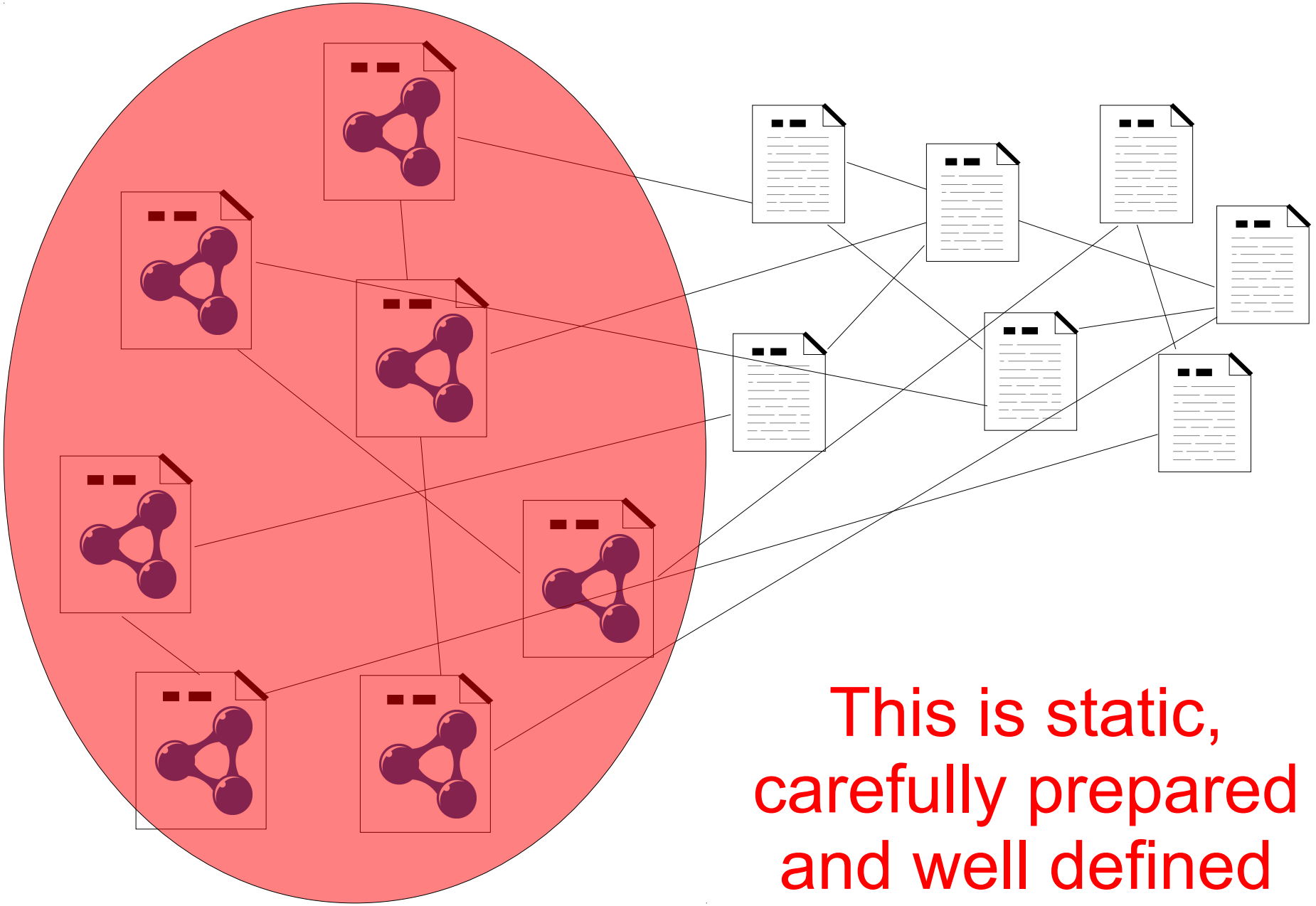
- Instead of overlaying the domain items, we put the user model layer on the top of domain items conceptualization
- Re-use of user model across similar domains
  - Loops in C# and loops in Java
- Support changes of underlying domain items without loosing user-related information
  - all we need to do is to map new content to our conceptualization

Open Corpus User Modeling



What's wrong with it?

Open Corpus User Modeling



This is static, carefully prepared and well defined

Static Conceptualization

- Limited to a particular domain
- Personalization is limited to isolated applications
  - Or isolated groups of applications sharing one UM
- Small islands within the whole web ocean

Adapt. & Person. of “the Web”

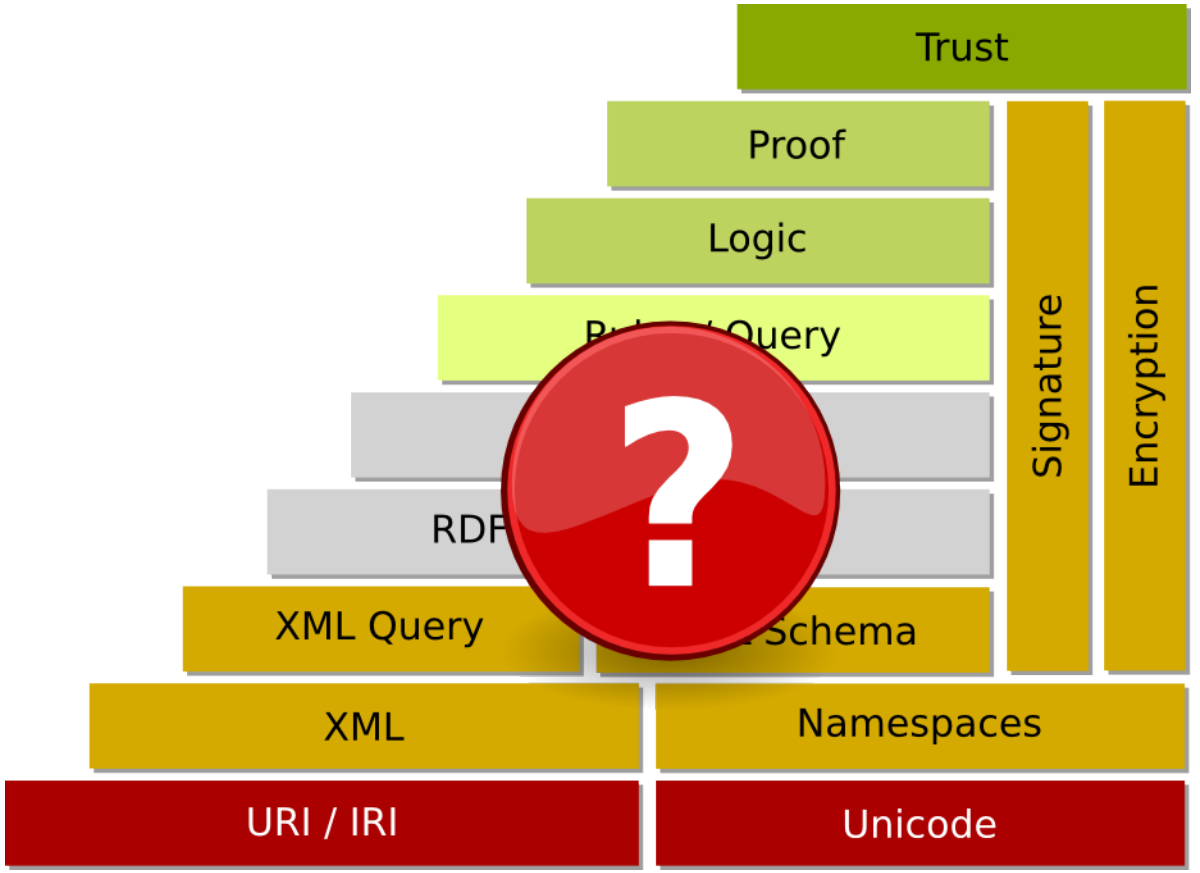
- Pure social-based approach
  - does not require any domain representation
  - System does not need to know the content, users will “tell” it whether it is worthy or not
  - Partitioning users into communities helps to provide more personalized recommendations, annotations etc. (clickstream analysis)

Adapt. & Person. of “the Web”

- Open corpus domain representation
  - dynamic and open conceptualization
  - Capability to retrieve and process metadata to any document (page) being viewed

- How can this be achieved?

Can we take advantage of semantic data?



Problems of the Semantic Web

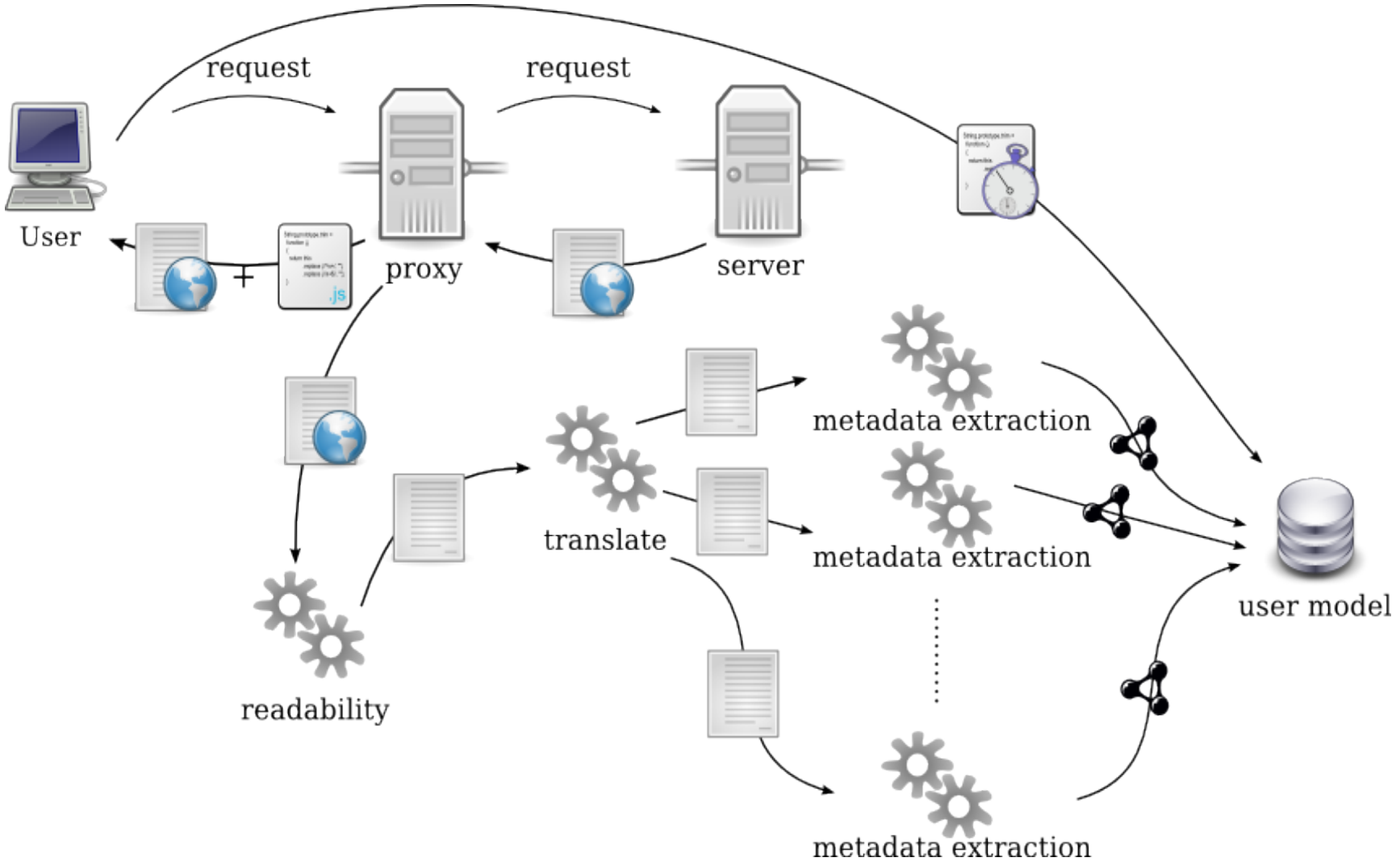
- Existing semantic systems restricted to a limited set of domains
  - a priori defined domain specific ontologies
  - no links to other ontologies
- The overall Semantic Web does not adequately cover specific terminology
- Many online ontologies have a weak internal structure
  - few online ontologies contain synonyms or non-taxonomic relations,

Waiting for Semantic Web killer app

Is there any other, reachable semantics?

- People got used to keywords
  - for searching
- People got used to tags and tagging
  - for future retrieval of information
  - but also for first-time retrieval in folksonomies
- People find enough semantics in keywords and tags
- We should give it a try as well...

Open Corpus **keyword-based** UM



Metadata extraction

- combining NLP, NER, Linked data and various text processing services

- JATR library
- OpenCalais
- TagTheNet
- AlchemyAPI

My tag-cloud


Web
Michal
University
Texas
Slovakia
Minister
web
Michael
Peter
search
Service
Barla
head
Bratislava
text
Semantic
Slovak
Social
Martin
script
html
user
Information
system
President
Vladimir
Tvaro
javascript
name
Prime
Joseph
project
page
systems
version
Web-Based
BRATISLAVA
application
Composition
United
Facebook
European
Bartalos
Marian
Holub
Robert
table
Ministry
Microsoft
Environment
title
Educational
Recycling
list
source
social
body
time
Republic
proxy
Shipman
Frank
Roman
server
States
HTML
Internet
Hungary
Technology
Anna
Dusan
Court
type
Content
electronic
Marshall
COURT
Tomas
Type
Setup
Interblue
WordNet
SME
Thomas
Prolog
Fund
Simko
Zacchi
Europe
New
Ivan
xml
Michelizzi
Bielikov
Kompan
Union
information
government
equipment
Intelligent
adaptive
web-based
Kramar
Jakub
EUR
TX
class
User
Moore
Group
href
Services
Iso
Framework
E-Learning
teaching
Anton
Czech
controller
California
processing
Andrew
Easter
repository
Server
Paul
document
site
Jason
date
content
year
ALEF
Upload
Technologies
e-learning
Miroslav
Press
Russia
deactivate
word
browser
president
John
Vojtek
Government
Konstantinos
configure
Schwandtner
Zelen k
overview
programming
Profiles
waste
receiver
model
Menu
speech
Fico
Resources
Exploration
Labaj
Office
Bulgarian
Kuric
recommendation
National
left
Product
exploration
Atom
Dutch
Scientific
machine
Paz rikov
development
Milan
git
channel
Meintanis
Haowei
Boris
Hsieh
subject
language
interpreter
Belarusian
Kaczynski
Problem
ODS
tools
region
Herlocker
Jan
Petit
Connect
interest
Search
state
Lisp
Windows
Badi
Rajiv
Author
System

Do you want your own tag-cloud? :)

- Currently 25 distinct IDs
  - max 22 real users
  - top user had 41.693 requests (at the time of preparing this presentation)
  - average was 7693.8
- peweproxy.fiit.stuba.sk
  - query expansion is in alpha testing and looks promising

Further processing of keywords

- Elimination of stopwords
- Identification of synonyms

- Clusterization
  - semantic-relatedness (Wordnet, Folksonomies, …)

  - this gives us different profiles of a user
    - researcher, photographer, programmer, hiker, ...

What we can do with it?

- Detecting virtual communities
  - “smarter” social-based recommendations
- search query disambiguation
- optimization of search results list
- ...

Conclusions

- We move from basic user models towards two (and more) layers of user models
  - domain items and their conceptualization
- Conceptualization does not need to be closed either
  - keywords could be good enough for performing cool stuff on the top of the “wild wild Web”
- Requires analysis of web traffic
  - proxy server
  - client-side agent as a part of huge multi-agent system