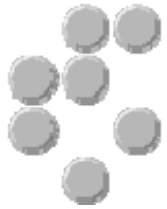


New Media and Knowledge Management

Part of
"New Media and e-Science"
MSc Programme 2006/07

Nada Lavrač

Jožef Stefan Institute

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Course participants

I. IPS students

- Jan Rupnik
- Matjaž Juršič
- Miha Grčar
- Vid Podpečan
- Xiaobin Li

II. Other students

- Ina Kovalna
- Katarina Pollak
- Nejc Trdin
- Anže Vavpetič
- Janez Kranjc

Course Schedule - 2007/08
Knowledge Management (KM)

- KM - Wednesday, 24 Oct. 07, 15-19 - Lavrač, lectures, MPS
- KM - Tuesday, 20 Nov. 07, 15-19 – Mladenić, Fortuna, lecture and practice - text mining, E8 Orange room
- KM - Thursday, 22 Nov. 07, 15-19 – Ljubič, Ferlež, lecture and practice – social network analysis, E8 Orange room
- KM - Wednesday, 27 Feb. 08, 15-19 – seminar results presentations, MPS

DM - Credits and coursework

- 6 credits (15 hours)
- Lectures
- Practice: Exercises and hands-on (Pajek and OntoGen)
- Group 1: working on abstracts of articles about questionnaire design, provided by prof. Vasja Vehovar from the Faculty of Social Sciences, Univ.Ljubljana:
  - Xiaobin Li (student MPS) + Anže Vavpetič (student FRI)
  - Jan Rupnik (student MPS) + Katarina Pollak (student FDV)
- Group 2: Working on abstracts of articles in the area of Inductive Logic Proghramming, provided by Springer
  - Vid Podpečan (student MPS) + Nejc Trdin (student FRI)
  - Matjaž Juršič (student MPS) + Janez Kranjc (student FRI)

- Contacts:
  - Nada Lavrač nada.lavrac@ijs.si
  - Jure Ferlež, jure.ferlez@ijs.si (social network analysis)
  - Blaž Fortuna, blaz.fortuna @ijs.si (text mining)

DM - Credits and coursework

- Wednesday, 27 Feb. 07, 15-19 – seminar results presentations, MPS
- For social network analysis:
  - Perform the analysis with Pajek, on one domain (Vehovar, or ILP)
  - Oral presentation of seminar results (max. 8 slides), each group member should present part of the results. Use slides template at Petra Kralj's web page
  - Deliver written report + electronic copy (4 pages, double column, possibly with appendices, in Information Society paper format, see instructions on Web pages of Petra Kralj)
- For text mining
  - Perform the analysis with OntoGen, on one domain (Vehovar, or ILP)
  - Oral presentation of seminar results (max. 8 slides), each group member should present part of the results. Use slides template at Petra Kralj's web page
  - Deliver written report + electronic copy (4 pages, double column, possibly with appendices, in Information Society paper format, see instructions on Web pages of Petra Kralj)

Knowledge Technologies for KM

JSI Department of Knowledge Technologies

- **Knowledge management - Knowledge technologies relationship:**
  - **Knowledge management**
    - Main topics: knowledge acquisition/ generation, storage/development, transfer, customization/use
    - Three aspects of KM: organizational, technological and sociological
  - **Knowledge technologies**
    - technological aspect of KM – methods, techniques and tools

Knowledge Technologies for KM

Department of Knowledge Technologies - main research areas

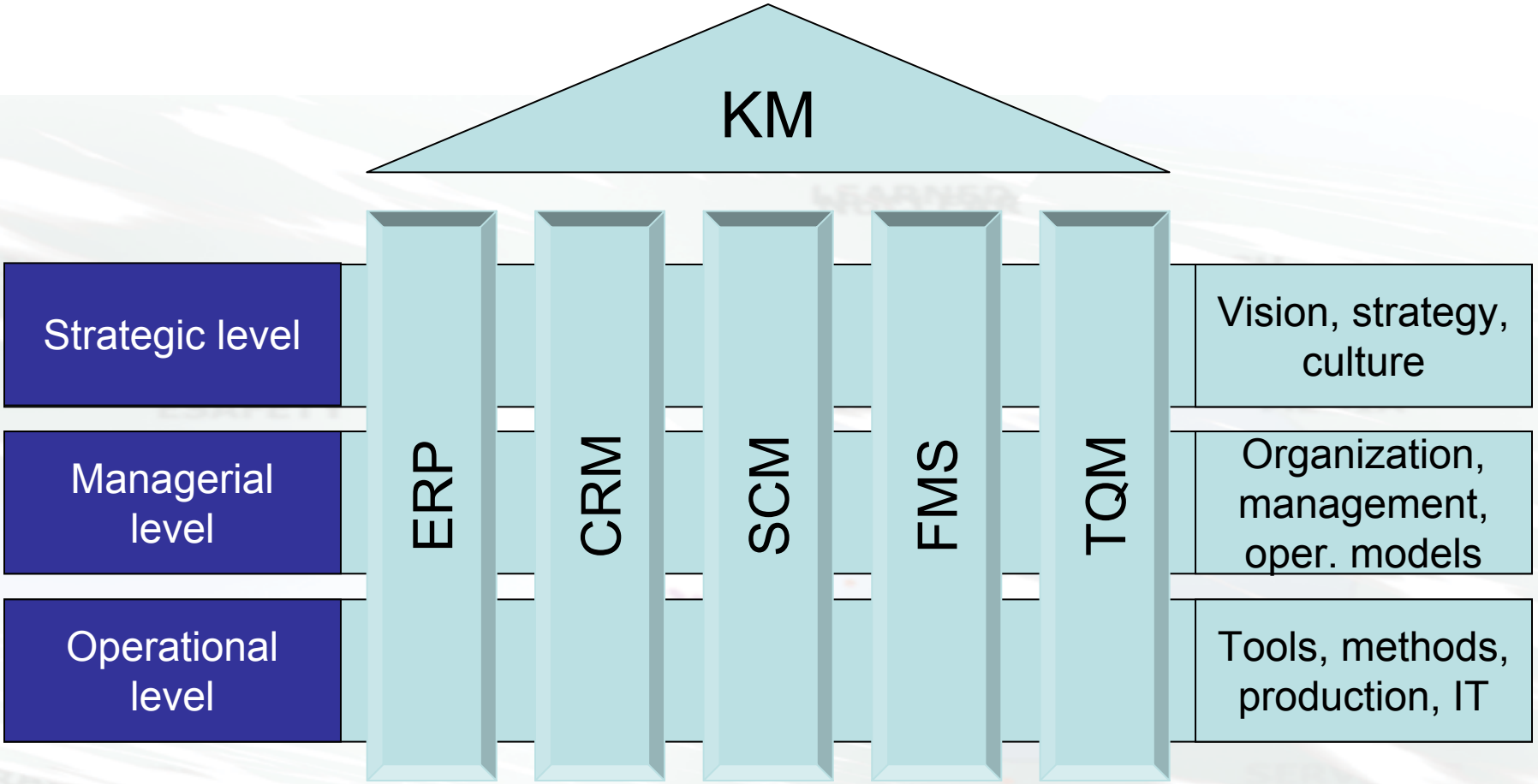
- data (text, web) mining and knowledge discovery
- decision support
- human language technologies
- semantic web
- knowledge representation, logical and probabilistic reasoning, expert systems, artificial intelligence
- Applications and eScience
- eLearning (Center of knowledge transfer in IT)

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Introduction to KM: Outline

- **What is KM: A traditional view**
- **KM in New economy: A Networked Organizations (NOs) perspective**
- **Selected knowledge technologies for KM in NOs**

Traditional KM

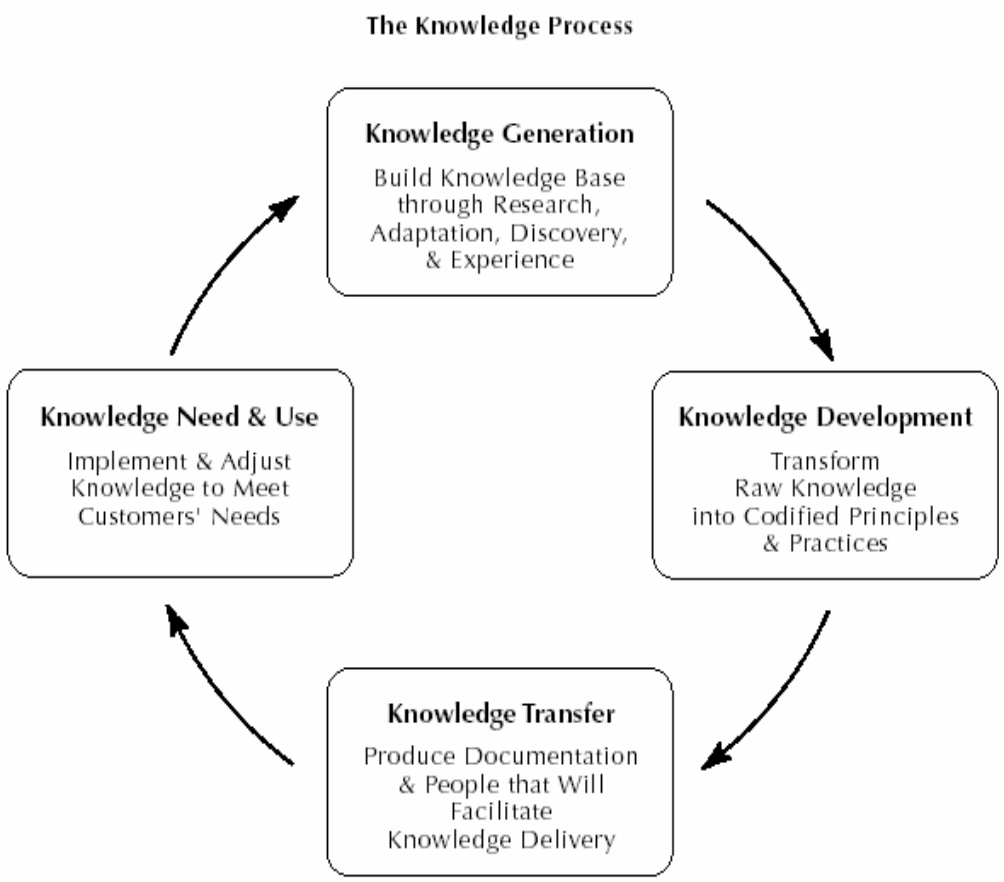


ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), SCM (Supply Chain Management) FMS (Flexible Manufacturing Systems), TQM (Total Quality Management), ...

Traditional KM

Managing knowledge

- generation (acquisition)
- storage and development
- transfer
- use and customization



The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation by Ikujiro Nonaka, Hirotaka Takeuchi, 1995

What is KM

Knowledge Management is a systematic approach to improve the way organizations, groups and individuals handle knowledge in all forms, in order to improve effectiveness, innovation and quality.

Knowledge Management aims to transform the intellectual capital of an organization –stored organizational knowledge and tacit knowledge of individuals - into a new corporate value resulting in increased productivity and improved competitiveness. KM teaches all members of an organization how to optimize existing knowledge and how to generate new knowledge as a **collective entity**.

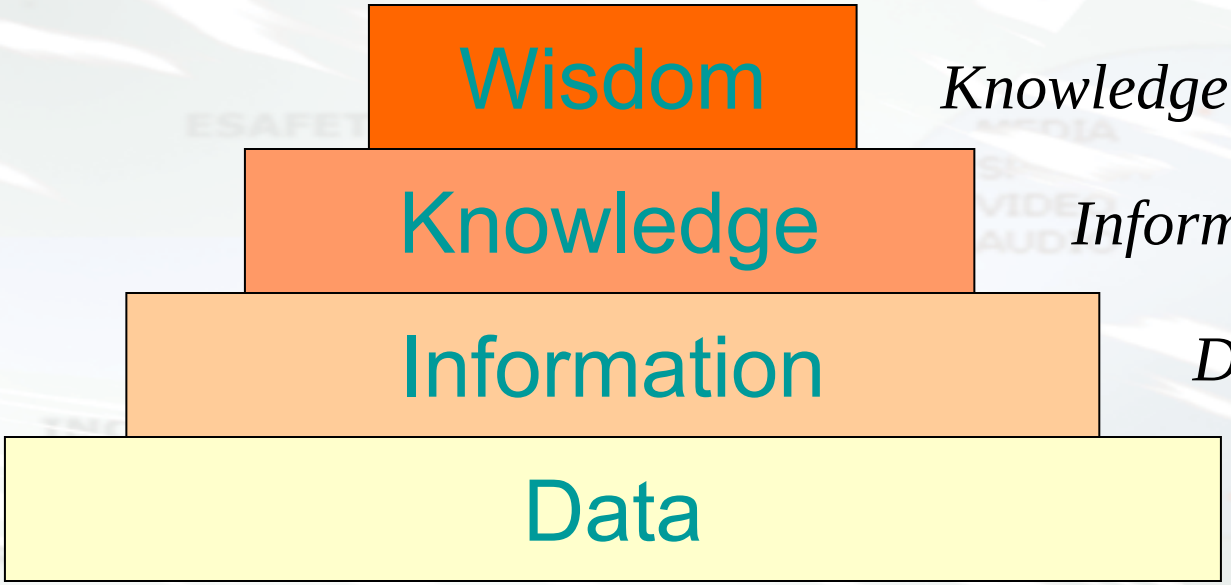
What is knowledge

- Knowledge is a model of (a part of) the reality as perceived by an agent
- Pragmatic definition: Knowledge is the information that confirms itself in use
  - Knowledge can not be uniquely defined, as the definition depends on the characteristics and goals of the organization
  - Knowledge is embedded in organizational processes, products and services

What is knowledge

- Principles
  - Knowledge is expensive to acquire, cheap to exploit
  - Property rights for knowledge are hard to define: IPR
  - Using knowledge does not mean wearing it out, knowledge grows and becomes richer through its use
  - Sharing knowledge with others does not imply losing it, knowledge evolves and multiplies through sharing

Data-Wisdom Pyramid



Knowledge plus experience

Information plus rules

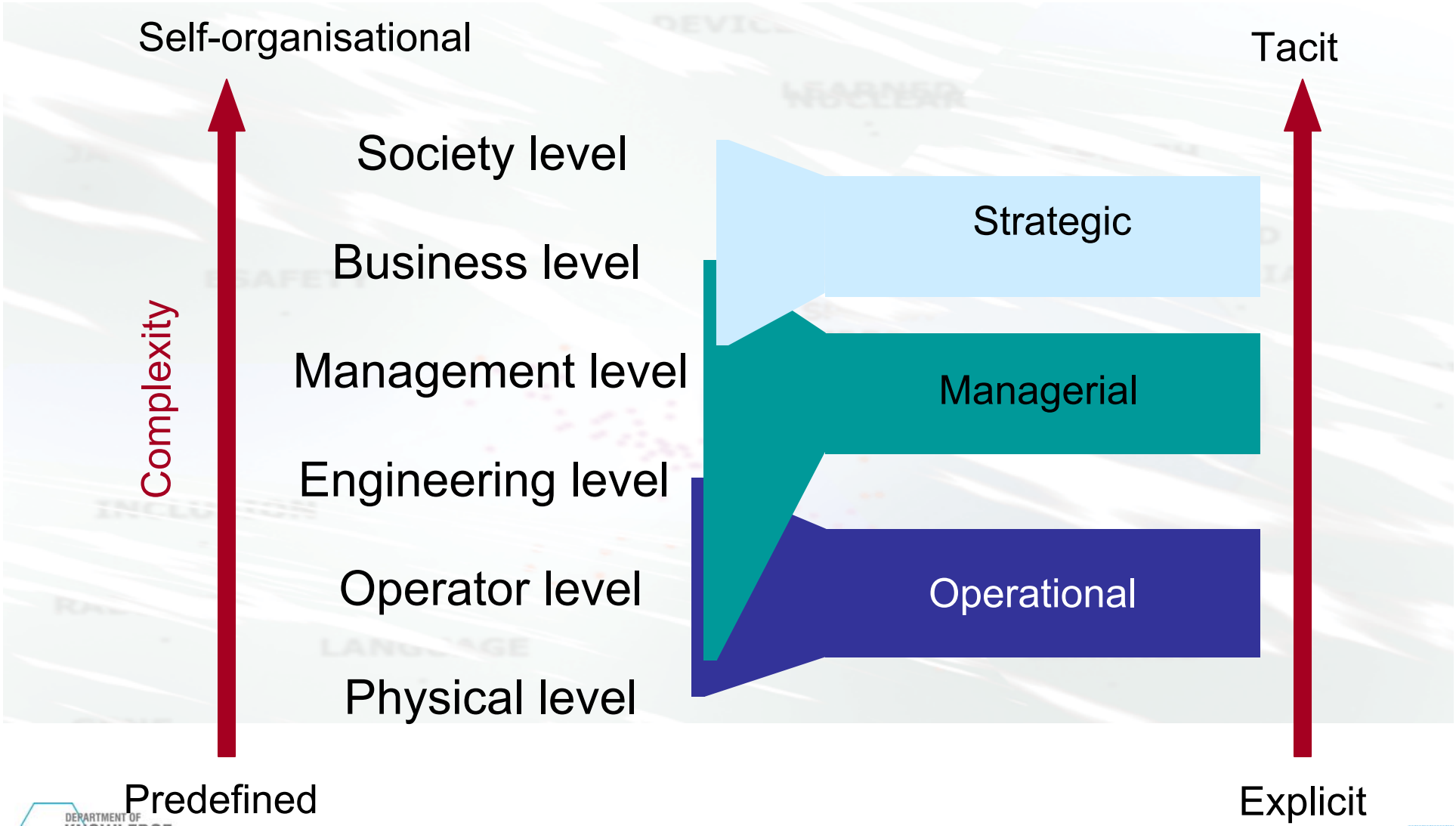
Data plus context

Observations, measurements

Data, information, knowledge

- Data is an individual observation or measurement, that yet needs to be interpreted
- Information is interpreted data – it is “the difference, which makes the difference”
- Knowledge is the structure from which the meaning of information can be derived (“why” and “what for”) - nothing can become information without pre-knowledge (background knowledge)

Knowledge and society



Tacit / implicit vs. Explicit / codified knowledge

- Tacit (silent, mute), Implicit (can not be explicitly articulated)
  - formed of experiences, values, judgments and skills, enabling autonomous triggering and performance of actions. Hard to verify and accept. Two strategies:
    - try making tacit knowledge explicit
    - enable free flow of tacit knowledge
- Explicit (can be explicitly articulated), Codified (explicit, articulated in a specific language)
  - Encoding enables knowledge transfer, provided that the recipient knows the tacit ingredients of encoding used by the encoder
- Knowledge continuum, with barriers to knowledge encoding
  - costs of acquisition of implicit knowledge, codification, learning, problems of misunderstanding and misinterpreting

Introduction to KM: Outline

- What is KM: A traditional view
- KM in New economy: A Networked Organizations (NOs) perspective
- Selected knowledge technologies for KM in NOs

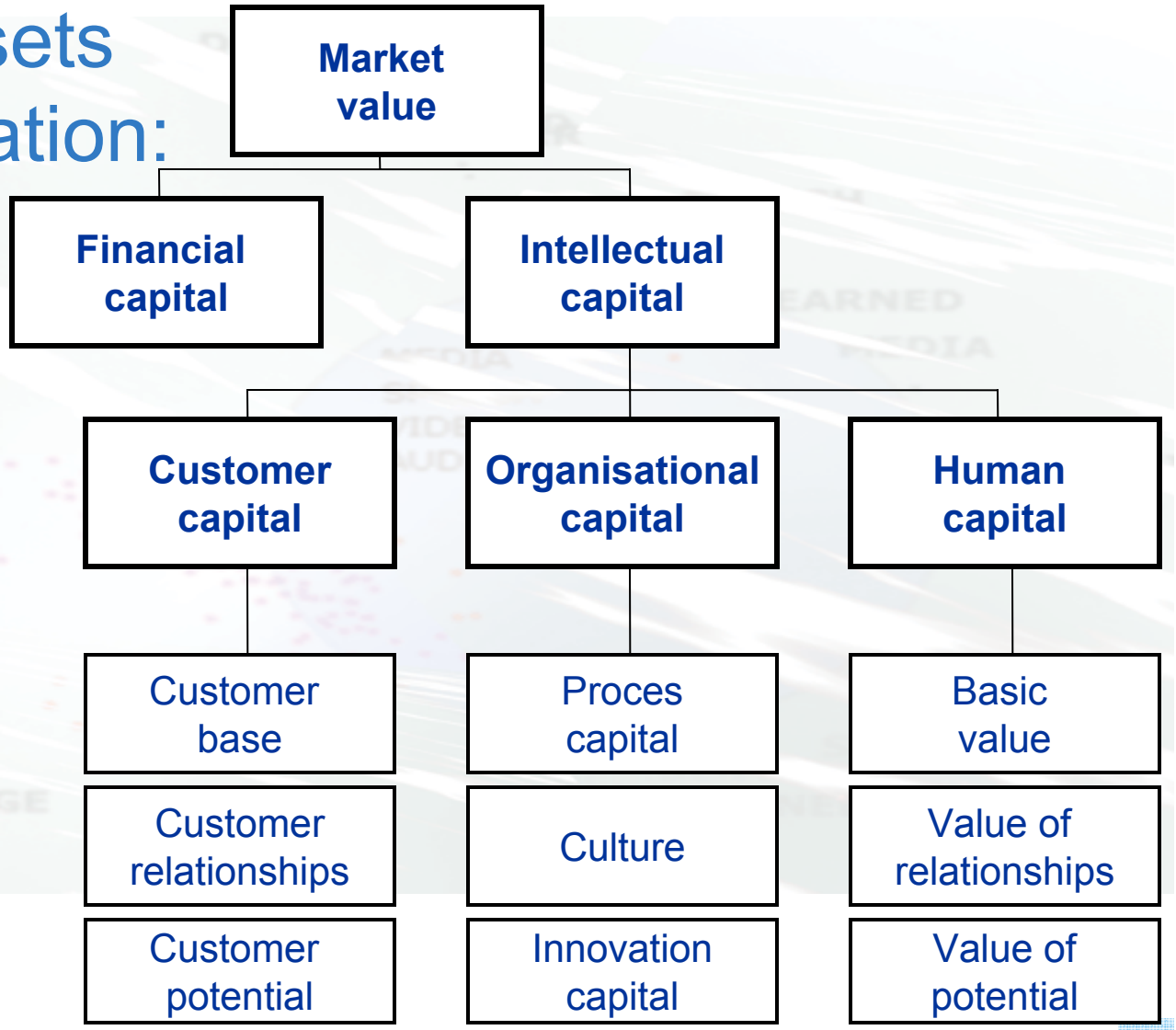
KM in New Economy

- KM: Traditional view
- KM: View shift
  - Information Society - Knowledge society
    - 10% of workforce produces all needed food and material goods, decreased dependence from natural resources (synthetic materials, decoding of human genome, ...), globalization and ease of accessing knowledge through new media, increased amount of people dealing with symbolic descriptions of things rather than things themselves (knowledge workers)
  - New economy - Knowledge economy
    - services rather than production, human networking, large corporations, virtual organizations, rapid changes, lifelong learning, knowledge as a source of intellectual capital

KM in New economy: Intellectual capital

Intangible assets of an organization:

- Internal (organizational structure)
- External (customer structure)
- Human (personnel competencies)



KM in New economy: A Networked Organizations Perspective

- **eBusiness, eScience, eMedicine, …**
  - doing business, science, medicine, … in a collaborative setting, supported by new media and computer networks
- **Networked organizations (NOs)**
  - non-static e-collaborative networks of individuals/organizations, enabled by information and communication technologies

NO infrastructures: New media

Infrastructures for KM: New media for eBusiness, eScience, ...

- **New media:** A generic term for many different forms of electronic communications and services that are made possible through the use of Internet technologies
- The term is in relation to “old” media forms, such as newspapers, magazines, radio diffusion and TV

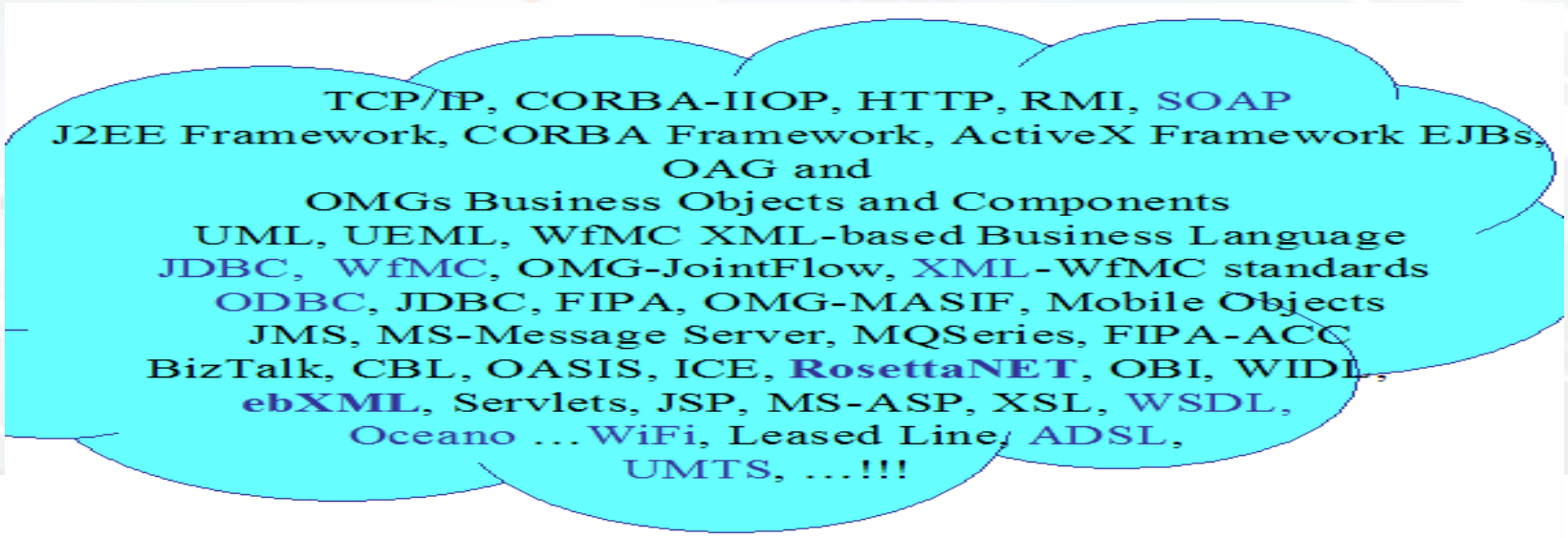
NO infrastructures: New media

- **Infrastructures:**
  - Networks (computer, satellites and telephone networks, cables, …)
  - Digital devices (DVD, CD-ROM, mobile telephones, wearable computers, …)
- **Services:**
  - WWW, internet, intranet, grid computing
  - streaming audio and video
  - chat rooms
  - e-mail
  - online communities
  - Web advertising
  - virtual reality environments
  - integration of digital data with the telephone, such as Internet telephony,
  - digital contents, digital libraries
  - mobile computing, wearable computing, ambient intelligence
  - …

NO infrastructures: Computer networks

Infrastructures for KM: Computer networks for eBusiness, eScience, ...

– ICT technologies, protocols and standards

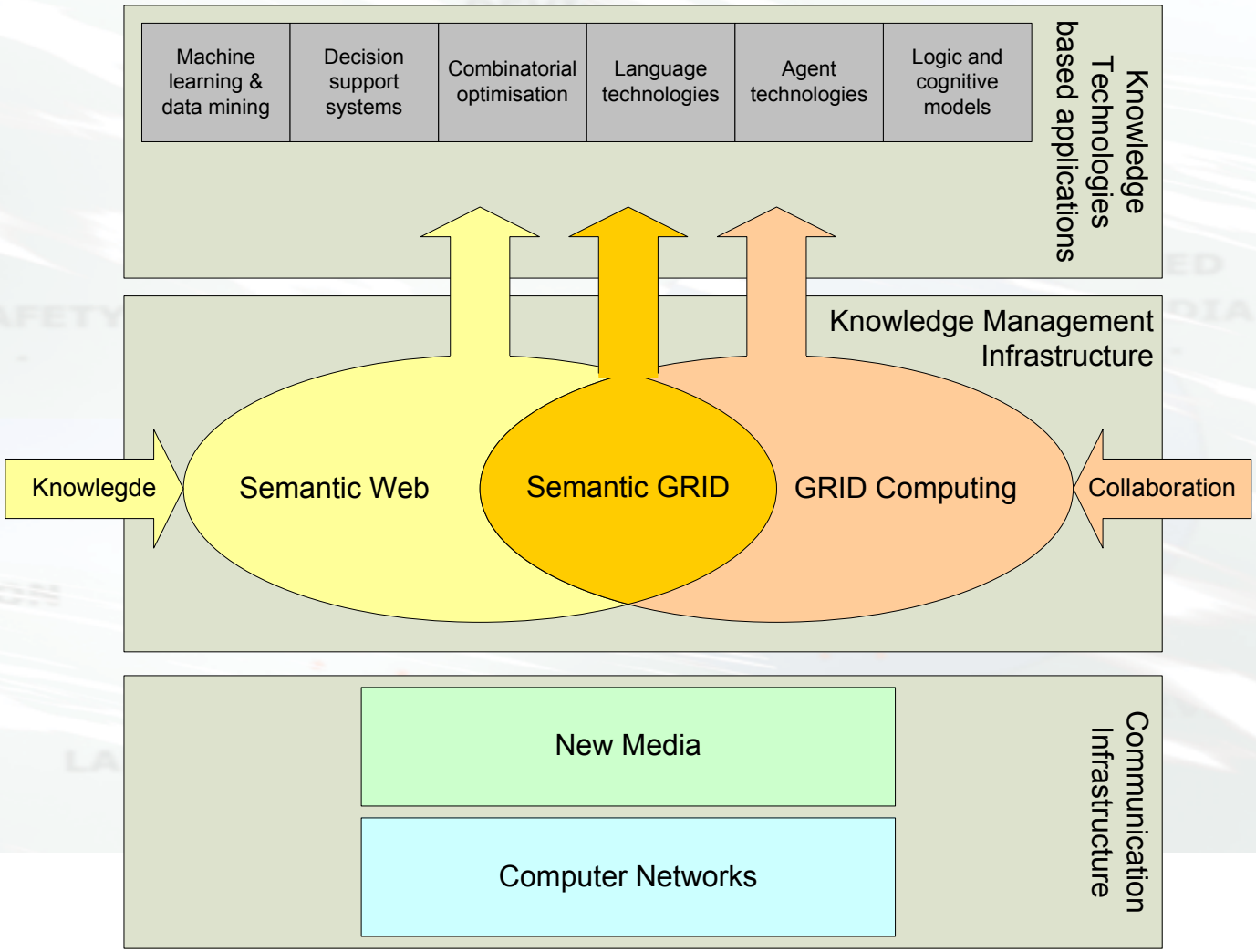


NO infrastructures: Towards the semantic grid

Infrastructures for KM: Semantic grid for eBusiness, eScience, ...

- **Grid computing:** coordinated resource sharing in dynamic, multi-institutional virtual organizations
- **Semantic Web:** extension of the current Web in which information is given a well-defined meaning, enabling data sharing and reasoning
- **Semantic grid:** extension of the current Grid in which information and services are given a well-defined meaning, enabling computers and people to work in collaboration

NO infrastructures and Knowledge technologies



Network economy

- **Network activities are facilitated by the use of shared infrastructure and standards, decreasing risk and costs**
- **Benefits of network membership increase by the number of other individuals and organizations in the network - the larger the network the better:**
  - a larger network is more competitive
  - has greater benefit of applications development
  - stimulates the speed and amount of learning and adapting of new technologies.
  - generates positive feedback where success generates success

Network economy

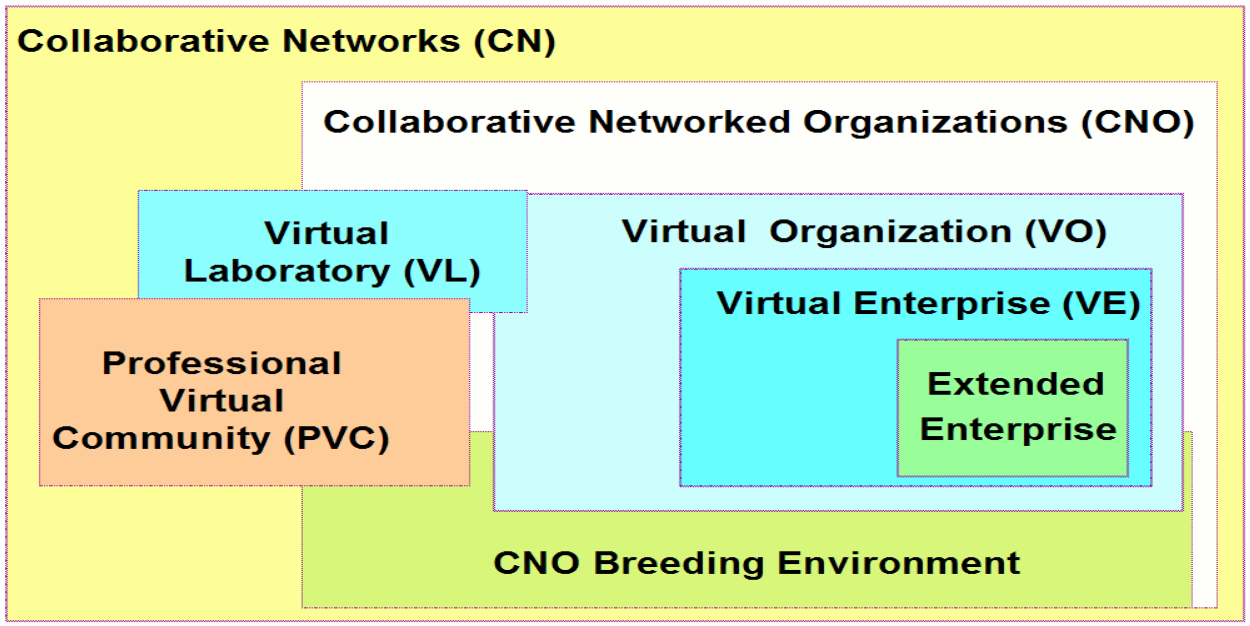
- **But: large networks are more complex to manage:**
  - increased complexity of the business environment and knowledge
  - managing processes instead of resources
  - agents as a source of knowledge
- **A partner in a NO can be viewed as an agent, capable of performing particular tasks**
- **The directing role is performed by an agent (net broker) acting as project leader in the process of:**
  - creating a virtual organization (VO) for a new project
  - planning, leading and controlling processes in a VO

Networked Organizations

- **Networked organizations (NO)** are non-static e-collaborative networks, enabled by information and communication technologies
- **Types of NO**
  - **Virtual organization (VO)**
  - **Virtual organization breeding environment (VBE)**
    - a cluster/association of organizations willing to collaborate
    - **VOs** are formed from **VBE** when a new business opportunity arises
  - **Professional virtual community (PVC)**

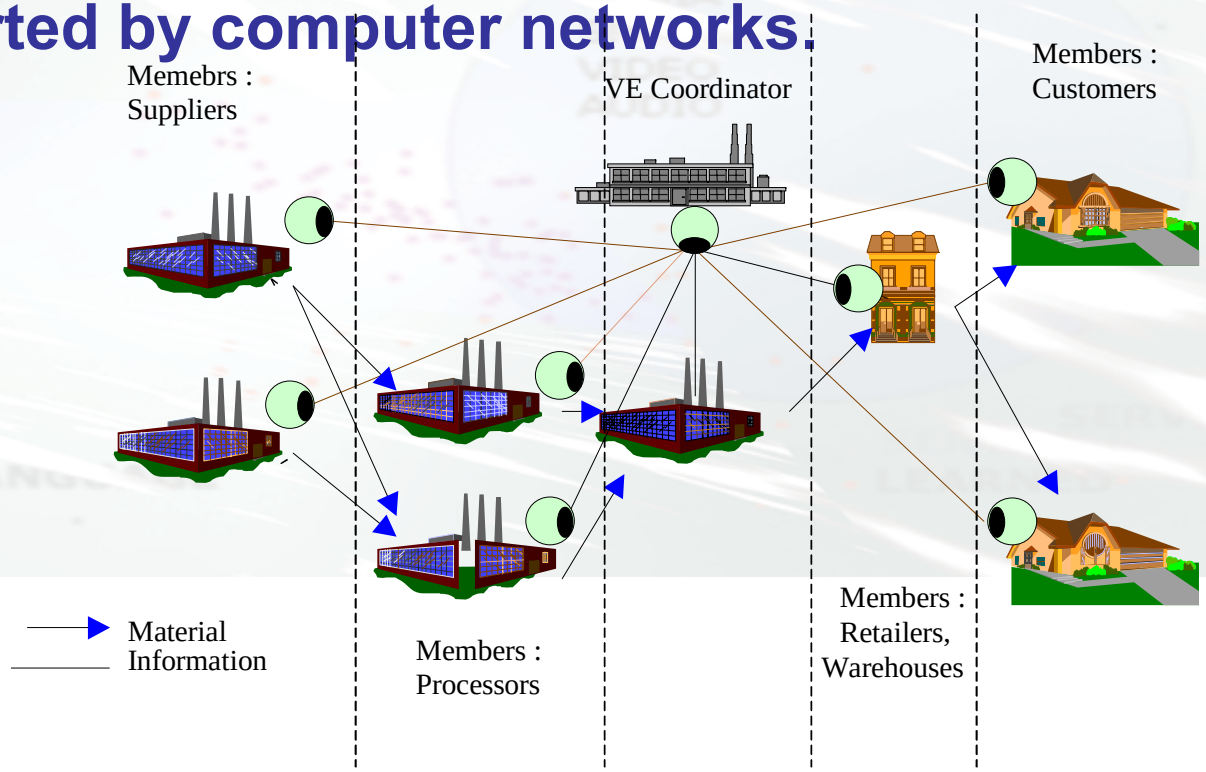
Networked Organizations

UNDERLYING CONCEPTS



Networked Organizations

Virtual organization (VO) is a temporary alliance of enterprises/organizations that come together to share skills or core competencies and resources in order to better respond to business opportunities, and whose cooperation is supported by computer networks.

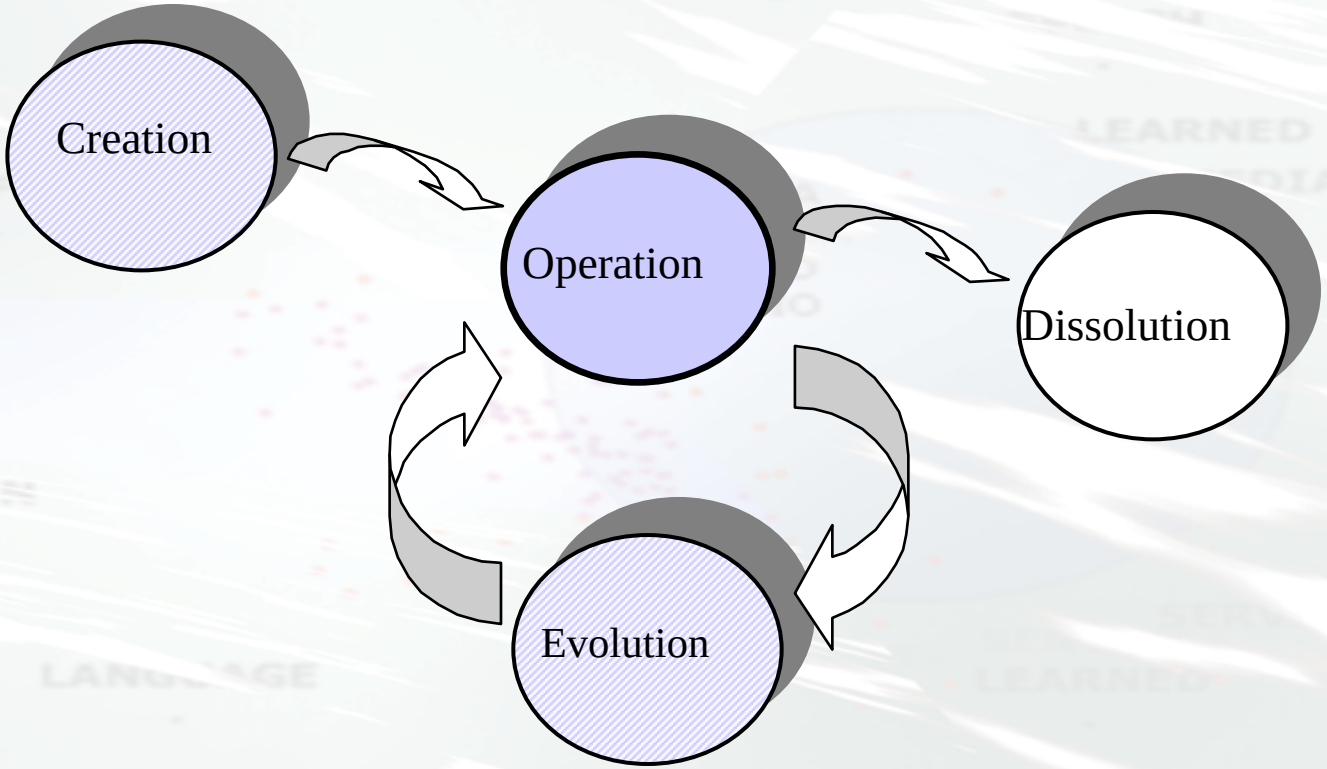


Networked Organizations

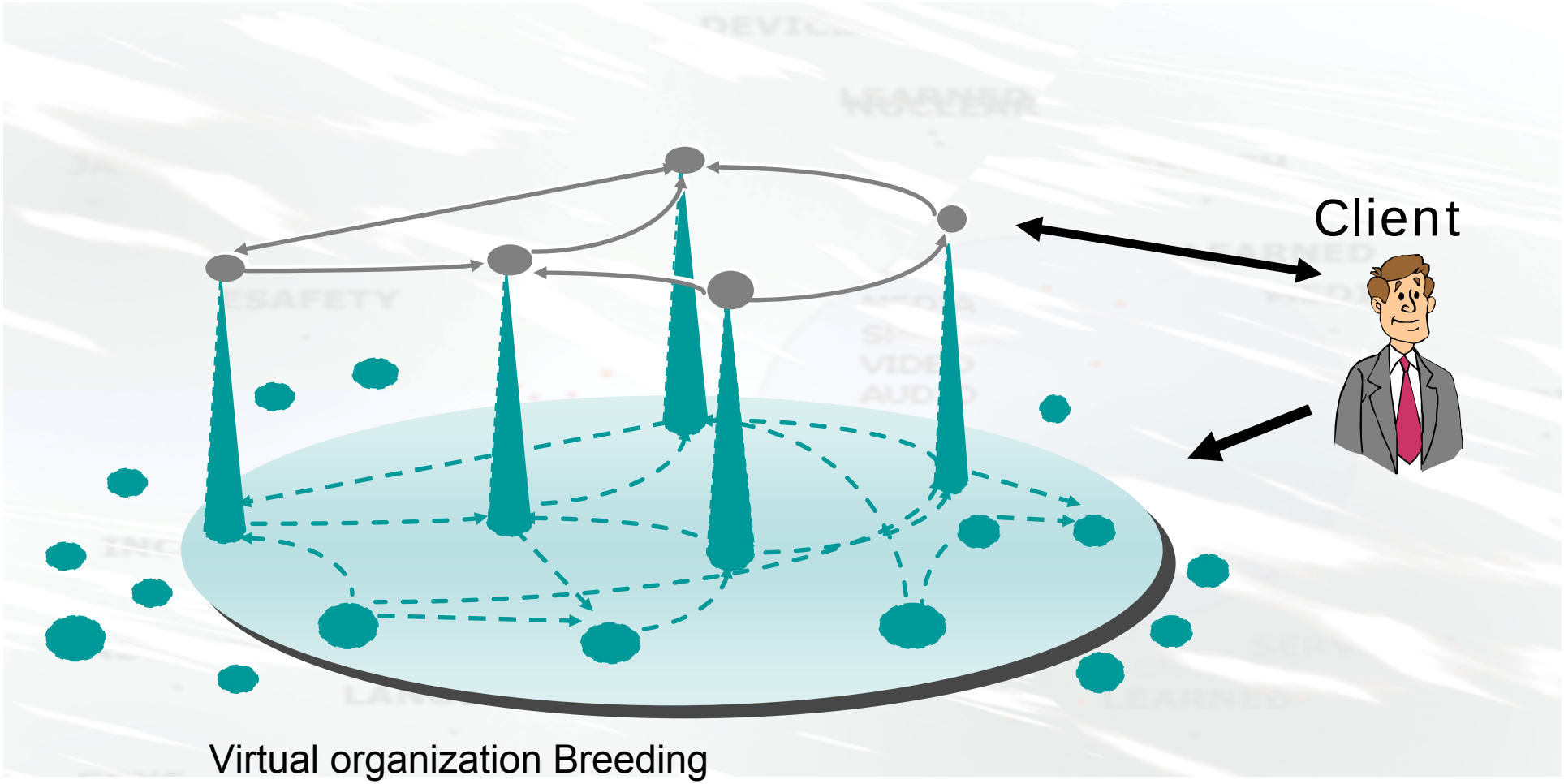
- **Virtual Organization Breeding Environment (VBE) represents an association or pool of agents -** organizations, supporting institutions, and individuals **- that have the potential and interest to cooperate.**
- **VBE is an establishment of a base long-term cooperation agreement**
- **When a business opportunity is identified by one member (acting as a broker), a subset of these organizations can be selected to form a VO**

Networked organizations

A typical networked organization lifecycle



Networked Organizations



Virtual organization Breeding Environment (VBE)

(Loss, 2005 – adapted from Bollhalter, 2004)

KM in NOs

- **Several problems occur:**
  - efficient storage of partners competencies
  - updating, sharing, promoting and transferring of these competencies
- Solved by adequate **knowledge management** using **knowledge technologies**
- **Knowledge map** - a knowledge resource repository is a necessity
  - each partner must have access
  - storing knowledge resources, process costs, resource availability

Knowledge technologies for knowledge mapping

- **Automatic gathering tools:**
  - Web crawling
  - Information and keyword extraction
  - Language technologies (lemmatisation, grammar, dictionary)
- **Data storage** - relational database technology
- **Data analysis and decision support**
  - Social Network Analysis
  - Data,Text and Web mining, clustering
  - Machine learning tools (classification trees,...)
  - Decision support systems and tools
- **Presentation**
  - Visualisation tools (text and data visualisation)
  - Social network visualisation and analysis tools

Introduction to KM: Outline

- What is KM: A traditional view
- KM in New economy: A Networked Organizations (NOs) perspective
- Selected knowledge technologies for KM in NOs:
  - Social Network Analysis:
    - A case study – ILPNet2, using Pajek
    - A case study in semi-automated trust modeling
  - Text mining:
    - A case study in semi-automated ontology construction – ILPNet2, using OntoGen
    - A case study in structuring of competencies of partners of the Virtuelle Fabrik Swiss industrial cluster, using gCLUTO

Goals of social network analysis

- **Coauthorship exploration through social network analysis (Pajek)**
  - Who are the most important authors in the area? Are there any closed groups of author, Is there any person in-between most of these groups? Is this same person also very important?

The domain: ILPnet2

- **Network of Excellence in Inductive Logic Programming (1998-2002), consisting of 37 universities and research institutes**
  http://www.cs.bris.ac.uk/~ILPnet2/
- **Successor of ILPnet (1993-1996)**
- **The ILPNet2 publications database:**
  - 589 authors, 1046 co-authorships, 1147 publications from 1971 to 2003

Social network analysis with Pajek

- Data extraction and preparation
  - Web data extraction
  - Data cleaning
  - Relational database construction
- Social network analysis, by exploring
  - Cohesion
  - Brokerage
  - Ranking

Data extraction and preparation

- Data in BibTeX format, one file for every year
  http://www.cs.bris.ac.uk/~ILPnet2/Tools/Reports/Bibtexs/2003, ...,
- Data acquired with the wget utility – a shell script that collects the data from the Web is as follows:

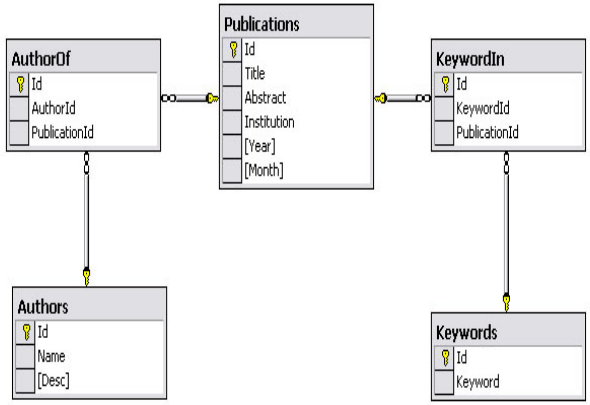
```
$ for((i=1971;i<2004;i++)); do
wget

http://www.cs.bris.ac.uk/~ILPnet2/Tools/Reports/Bibtexs/$;
done
```

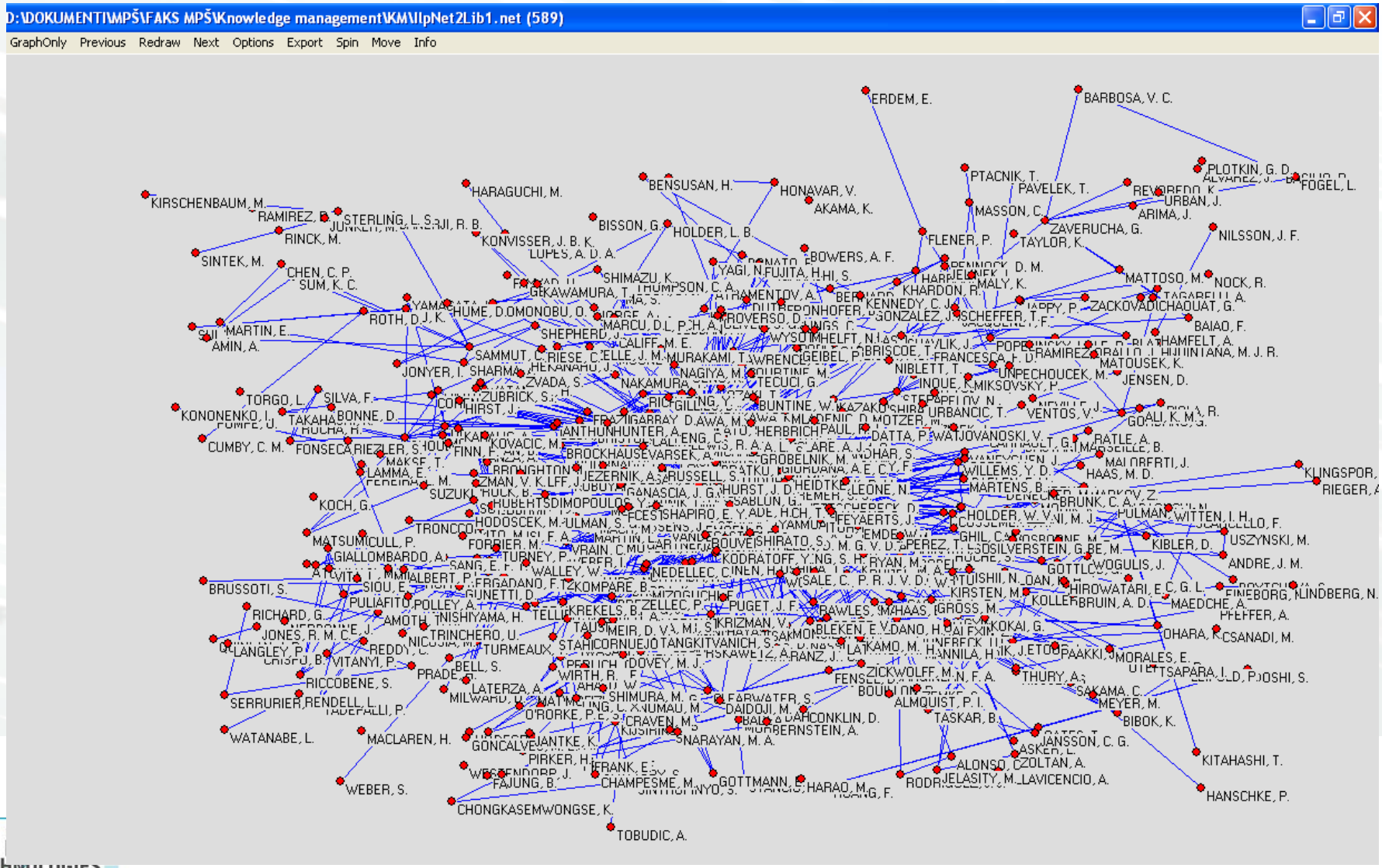
- Collected data converted into the XML format

Data cleaning and database construction

- Data cleaning
  - normalization of authors names
- Relational database construction
  - using Microsoft SQL Server
  - database schema
- Pajek input format
  - vertices:
    - author's ID and name
  - edges:
    - defined with two connected vertix IDs
    - weight correspond to the degree of collaboration (# of co-authorship) between the two authors.



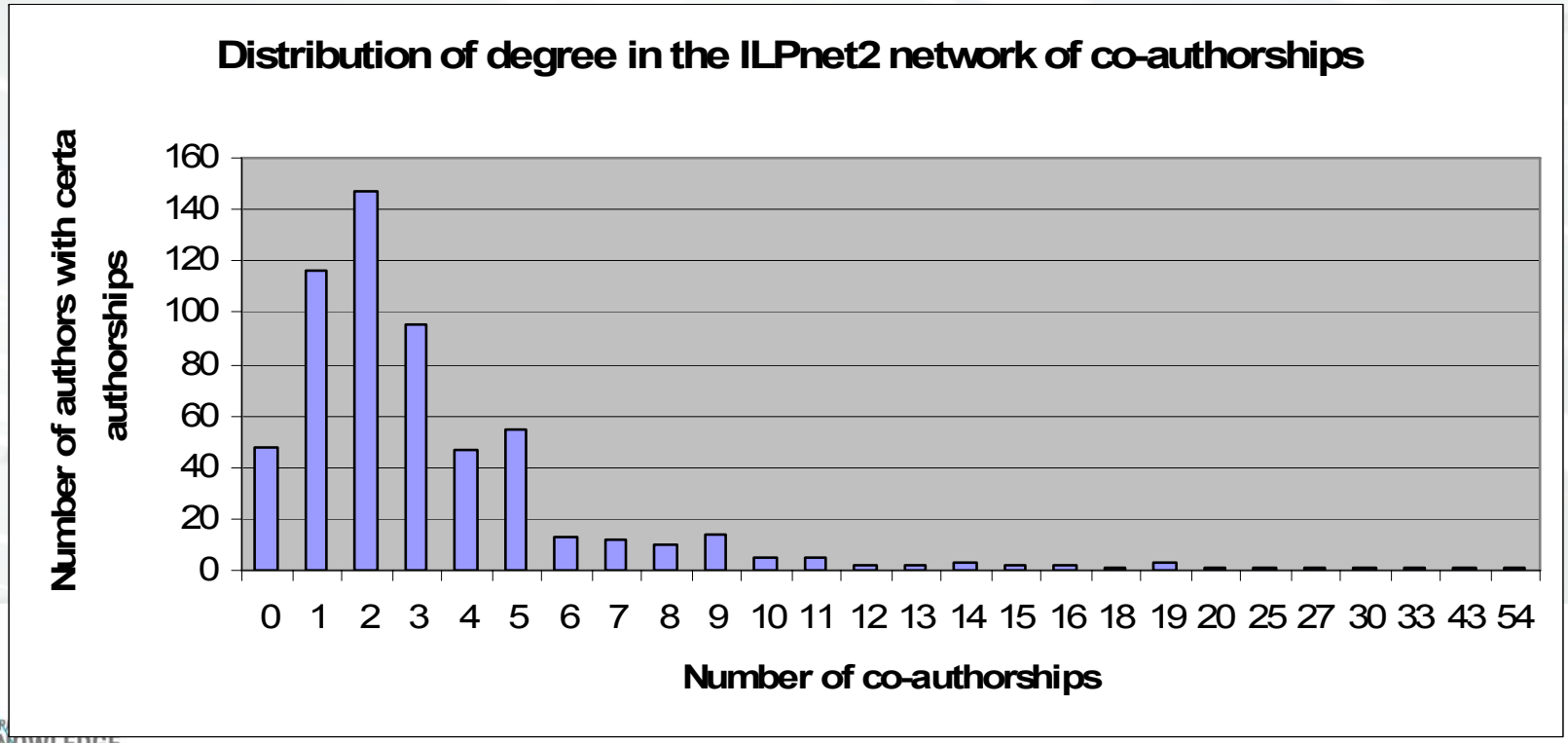
Social network of ILPNet2 authors



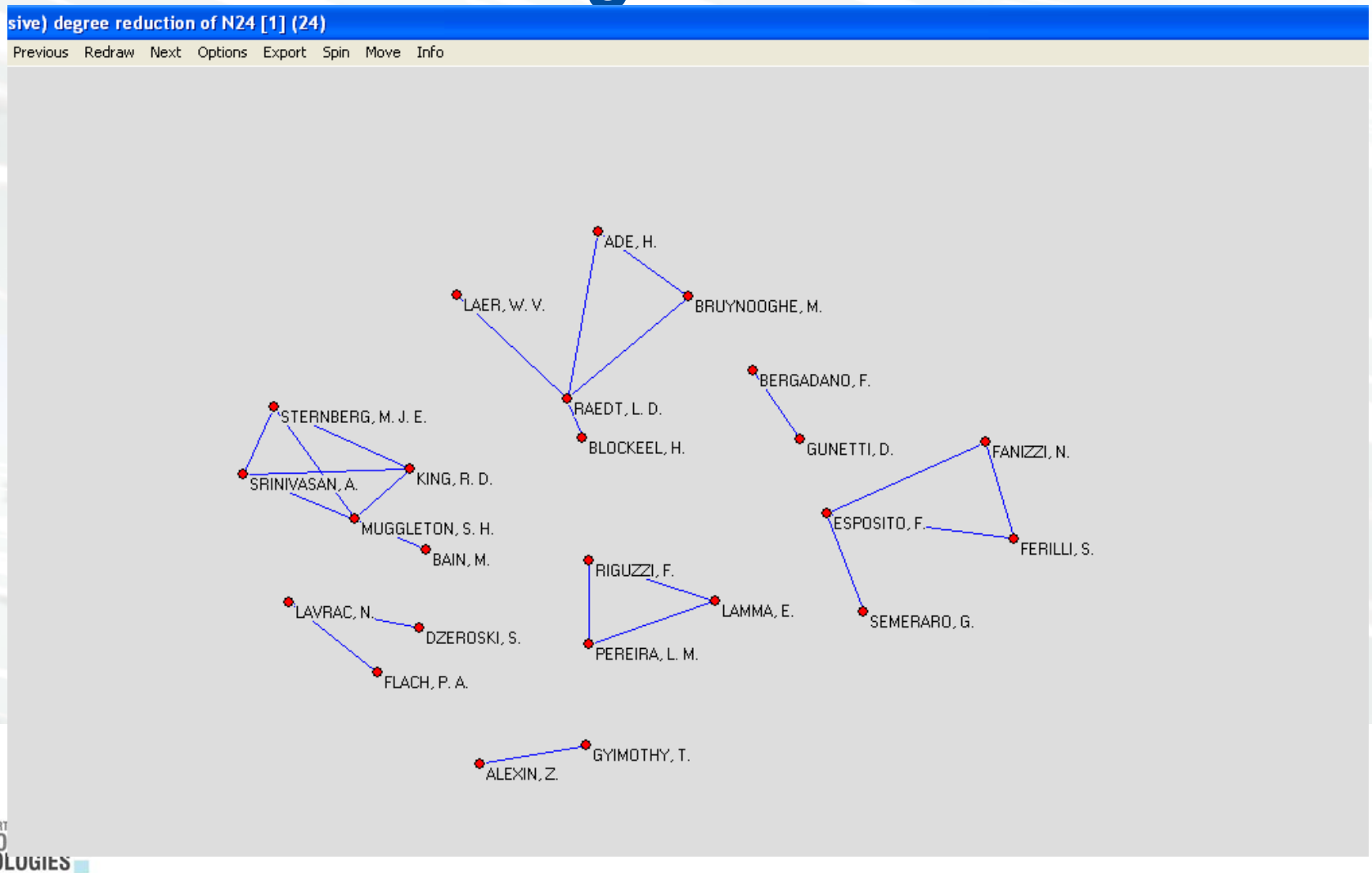
Vertex degree and density

Degree of a vertex = the number of lines incident with it.

ILPNet2 density = number of lines / maximum possible number of lines = 1046 / 173166 = 0.0060

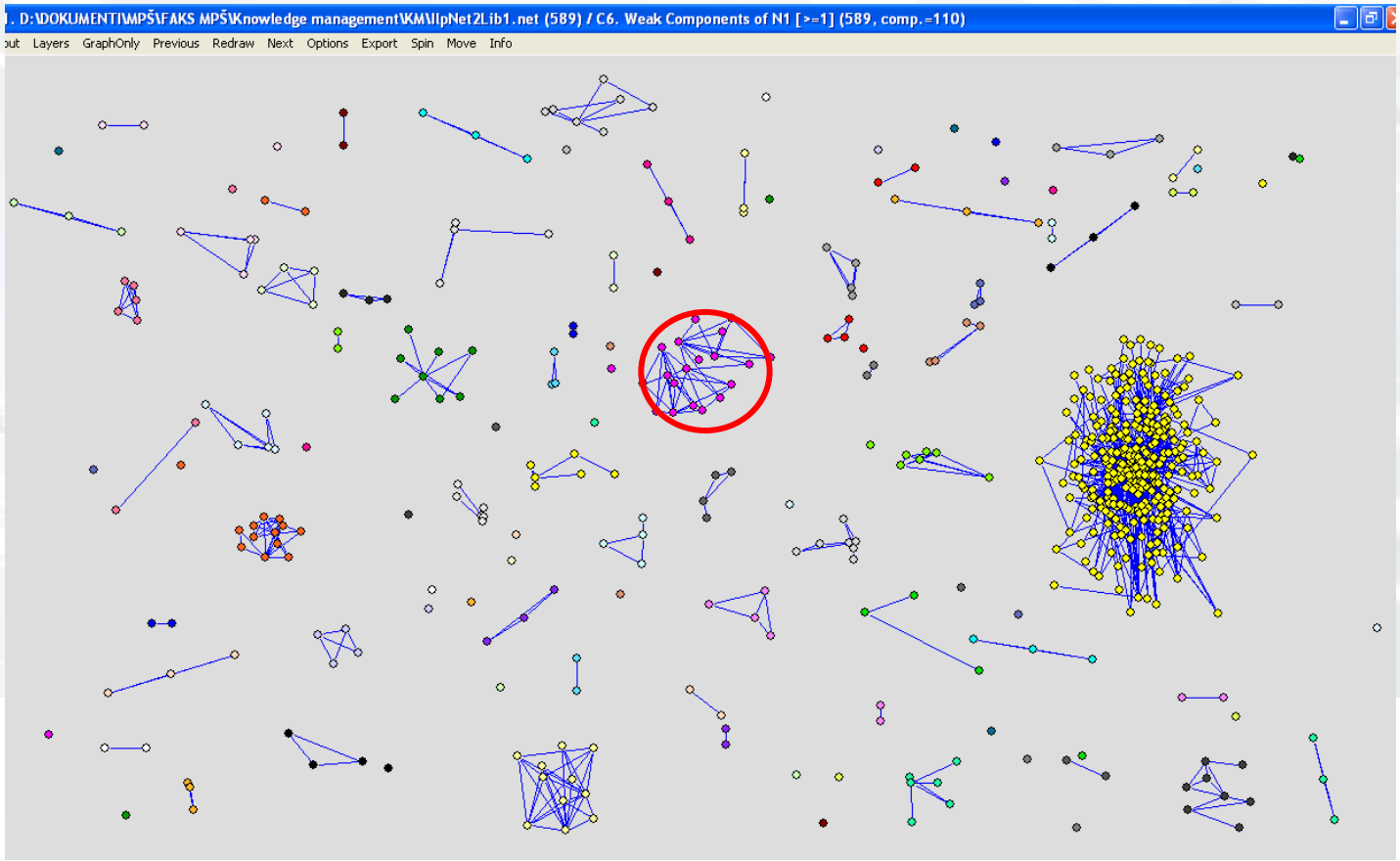


ILPnet2 social network – removed lines with value < 10 and vertices with degree < 1



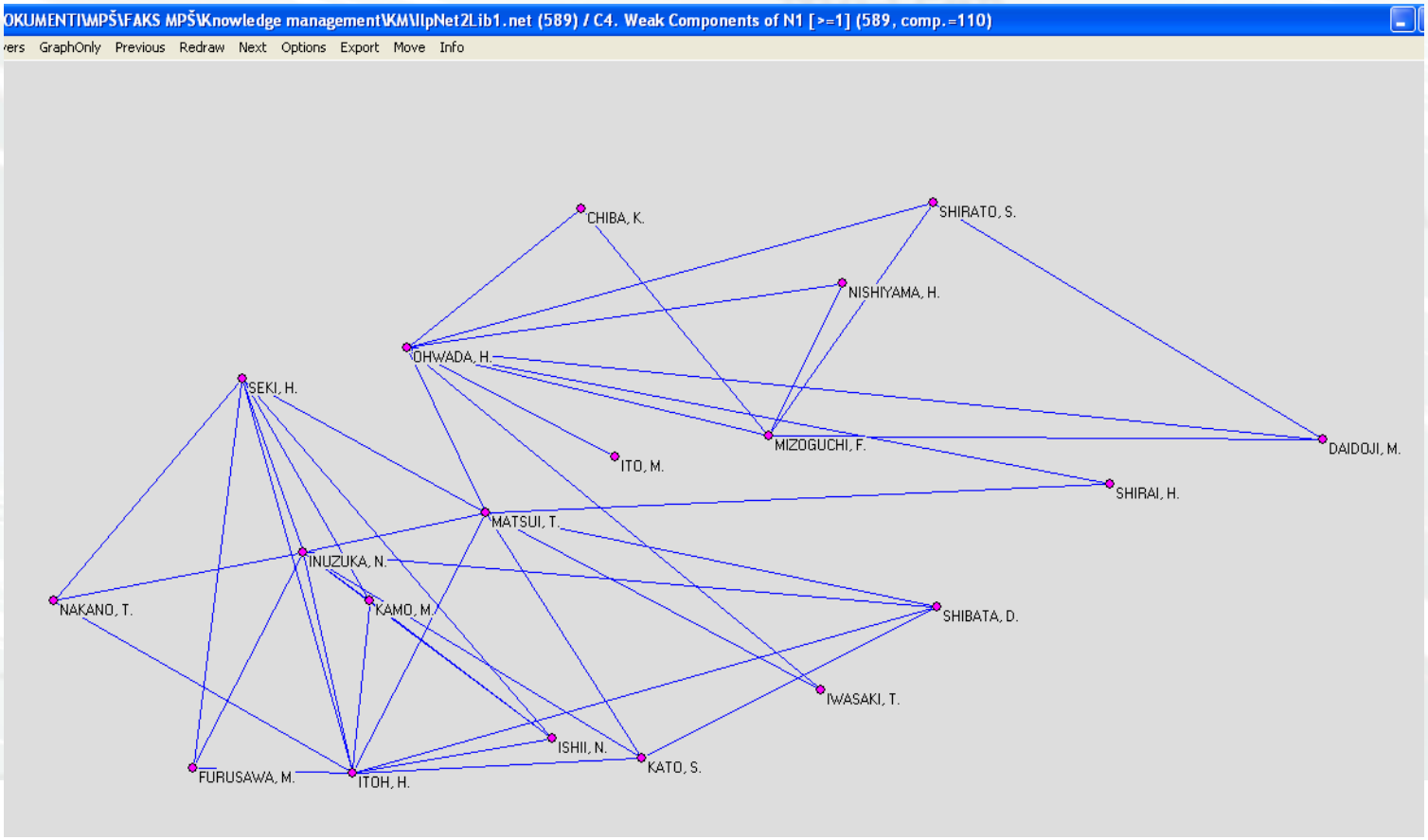
Components in the ILPnet2 network

Components identify cohesive subgroups – groups of vertices in a non-directed coauthorship network, connected by semipaths (with max 1 occurrence of every vertex)



Zoomed ILPNet2 component

Smaller ILPNet2 components are country biased



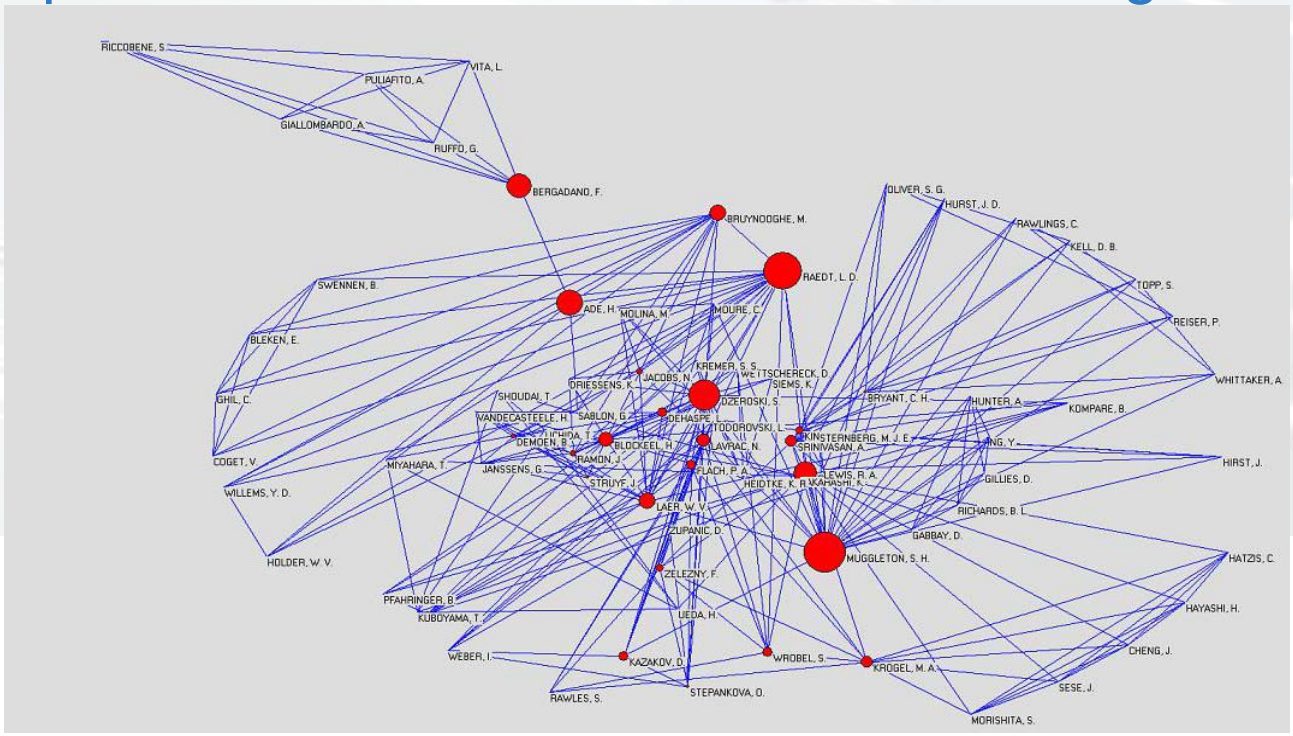
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Brokerage in the ILPNet2 network

Vertex degree of centrality = the number of lines incident with it

Closeness centrality = the number of other vertices divided by the sum of all distances between the vertex and all others

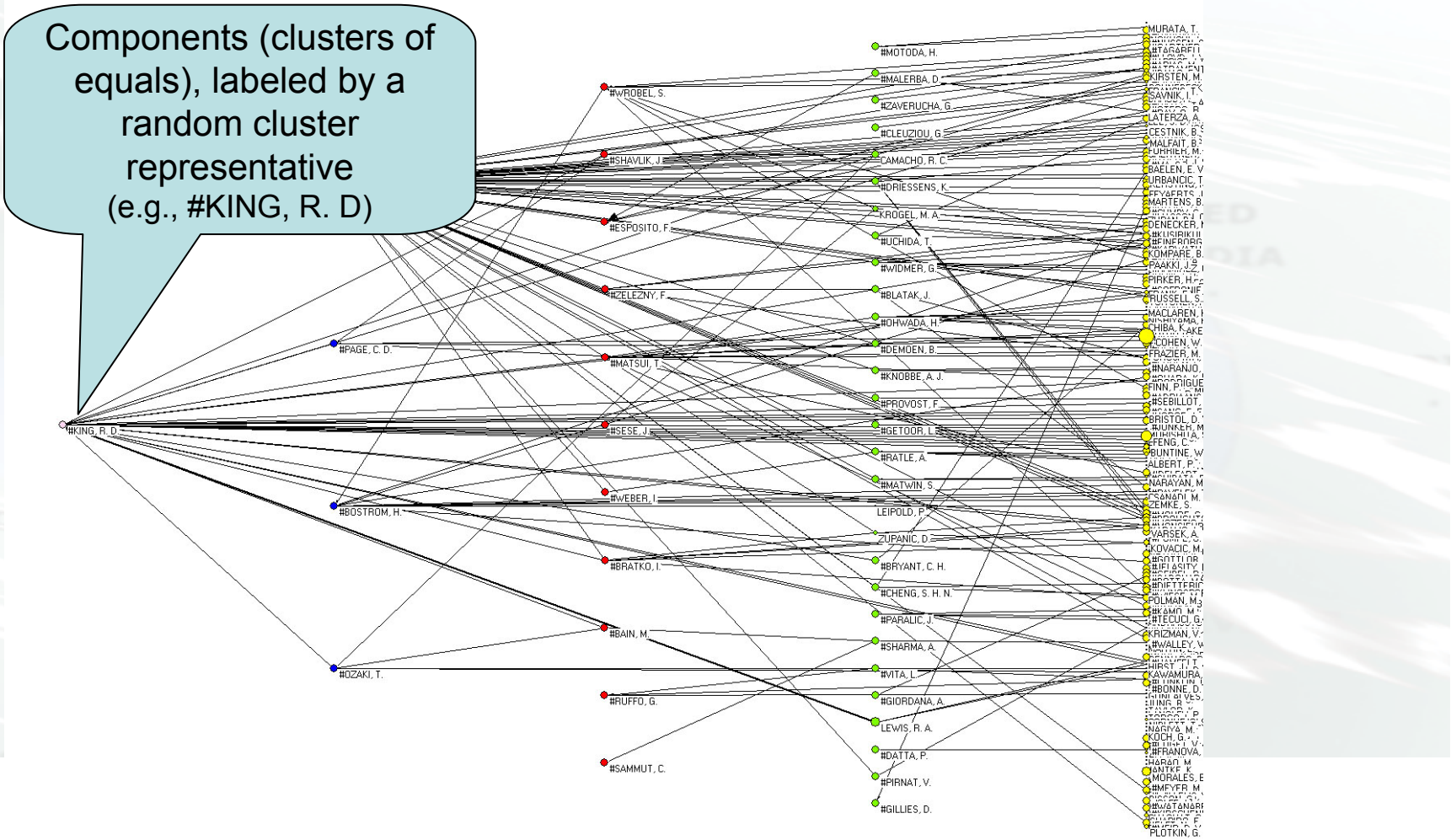
Betweeness centrality = the proportion of all shortest path between pairs of other vertices that include the given vertex



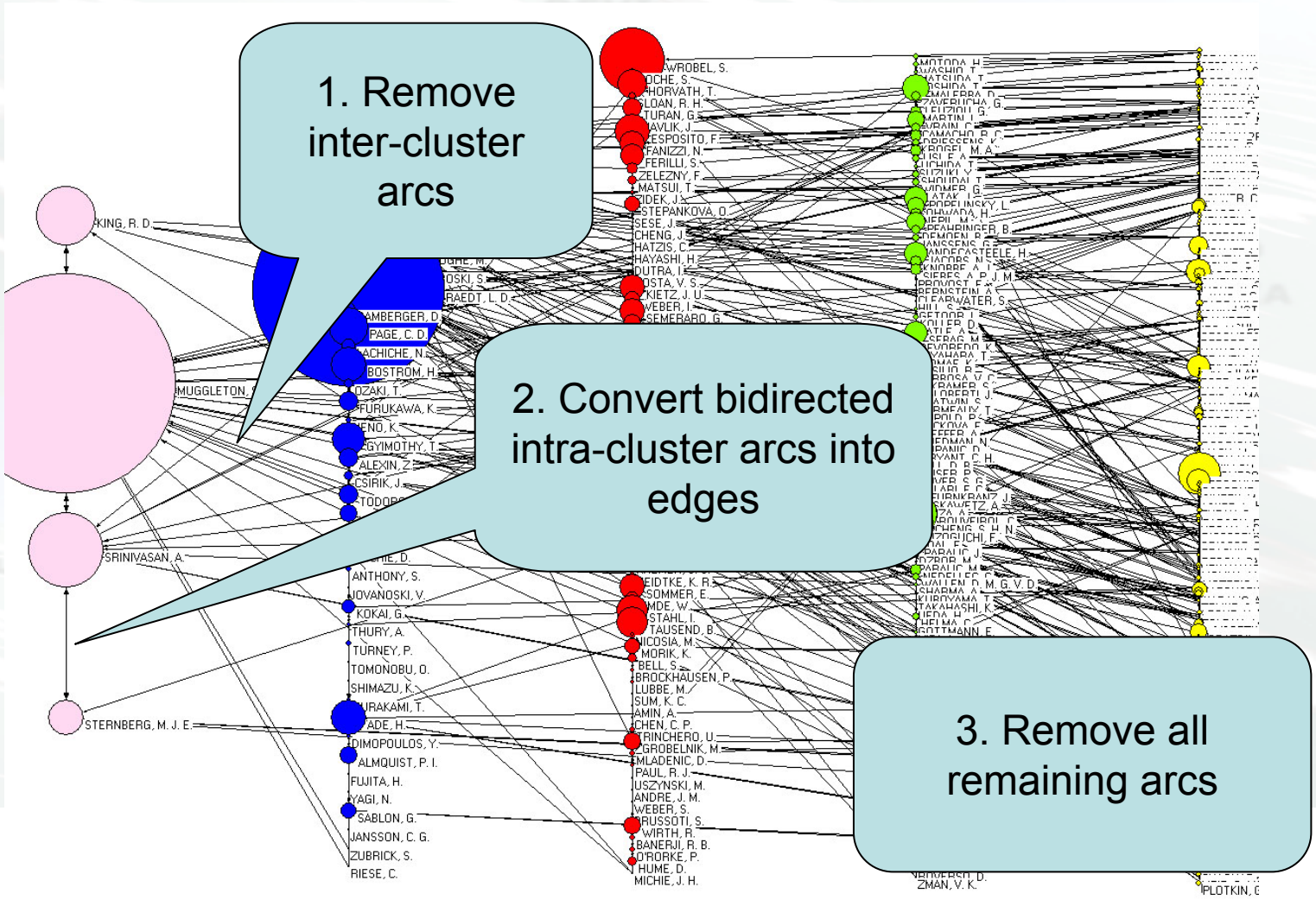
ILPNet ranking through structural prestige

Input degree		Unrestricted input domain size		Proximity prestige	
28	MUGGLETON, S. H.	152	LAMMA, E.	0.082030307	RAEDT, L. D.
21	RAEDT, L. D.	152	RIGUZZI, F.	0.077044151	DZEROSKI, S.
20	DZEROSKI, S.	152	PEREIRA, L. M.	0.068453862	LAVRAC, N.
17	LAVRAC, N.	152	RAMON, J.	0.066777042	MUGGLETON, S. H.
17	BLOCKEEL, H.	152	FLACH, P. A.	0.064946309	ADE, H.
12	FLACH, P. A.	152	LAVRAC, N.	0.06462585	BRUYNOOGHE, M.
12	SRINIVASAN, A.	152	STRUYF, J.	0.063683172	LAER, W. V.
11	GYIMOTHY, T.	152	BLOCKEEL, H.	0.060918631	TODOROVSKI, L.
10	JACOBS, N.	152	DEHASPE, L.	0.057783113	FLACH, P. A.
10	BERGADANO, F.	152	LAER, W. V.	0.054504505	SRINIVASAN, A.
9	WROBEL, S.	152	BRUYNOOGHE, M.	0.054346497	GAMBERGER, D.
9	STEPANKOVA, O.	152	DZEROSKI, S.	0.052812523	SABLON, G.
9	ITOH, H.	152	RAEDT, L. D.	0.051974229	DEHASPE, L.
9	ADE, H.	152	GAMBERGER, D.	0.051837094	BLOCKEEL, H.
8	KING, R. D.	152	LACHICHE, N.	0.048245614	KING, R. D.
8	OHWADA, H.	152	TODOROVSKI, L.	0.048015873	STERNBERG, M. J. E.
8	BRUYNOOGHE, M.	152	KAKAS, A. C.	0.047743034	KAKAS, A. C.
8	BOSTROM, H.	152	JOVANOSKI, V.	0.047283414	LACHICHE, N.
8	KRAMER, S.	152	TURNEY, P.	0.044957113	JOVANOSKI, V.
8	FURUKAWA, K.	152	ADE, H.	0.044957113	TURNEY, P.
8	CSIRIK, J.	152	DIMOPOULOS, Y.	0.043609897	RAMON, J.
7	HORVATH, T.	152	SABLON, G.	0.043226091	STRUYF, J.
7	ESPOSITO, F.	77	KING, R. D.	0.040507749	RIGUZZI, F.
7	SHOUDAI, T.	77	MUGGLETON, S. H.	0.040341393	DIMOPOULOS, Y.
7	DEHASPE, L.	77	SRINIVASAN, A.	0.035082604	LAMMA, E.

ILPNet2 ranking through acyclic decomposition

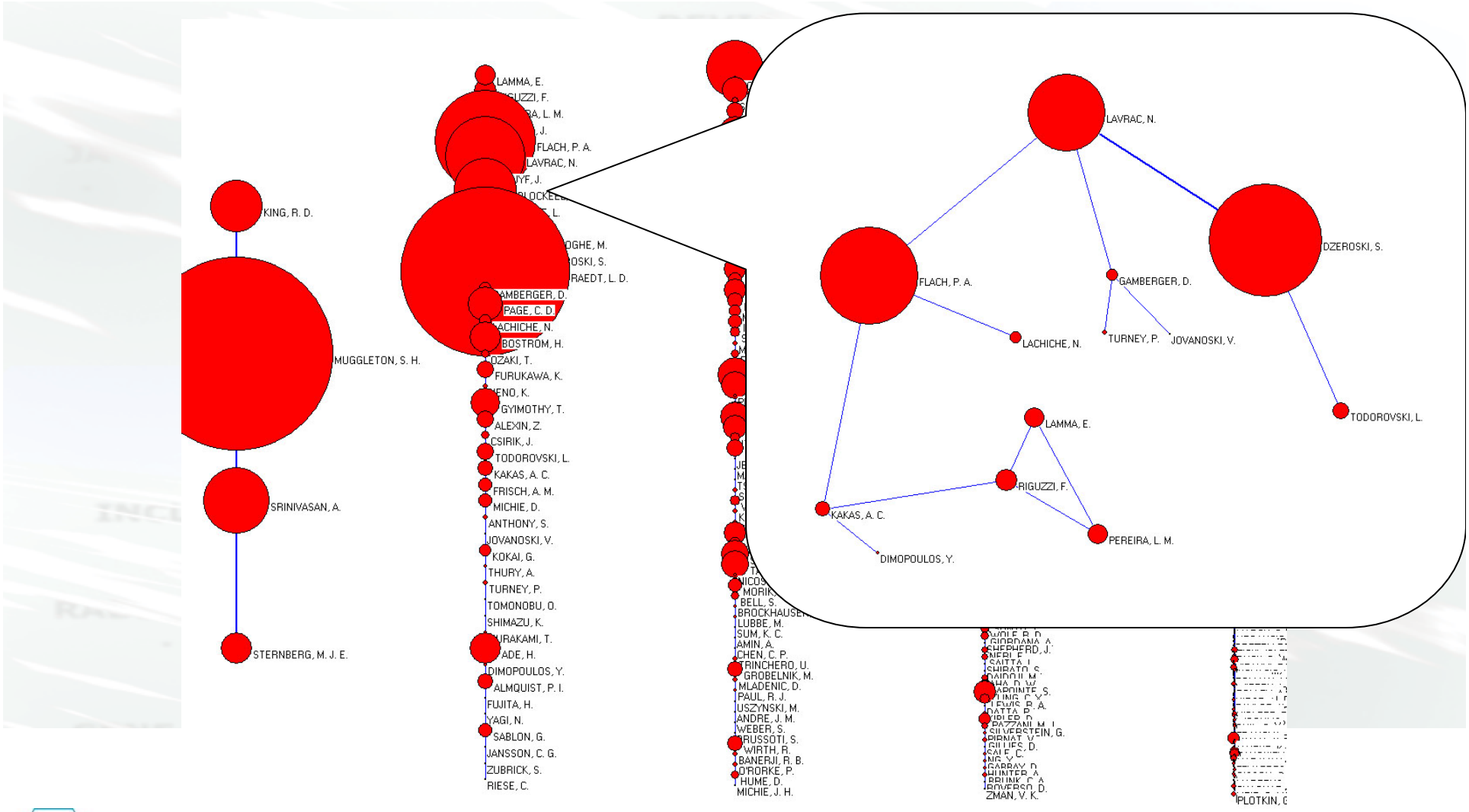


Acyclic decomposition
ILPnet2, hierarchical view (people)



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

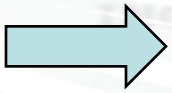
Acyclic decomposition
ILPnet2, hierarchical view (people)



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

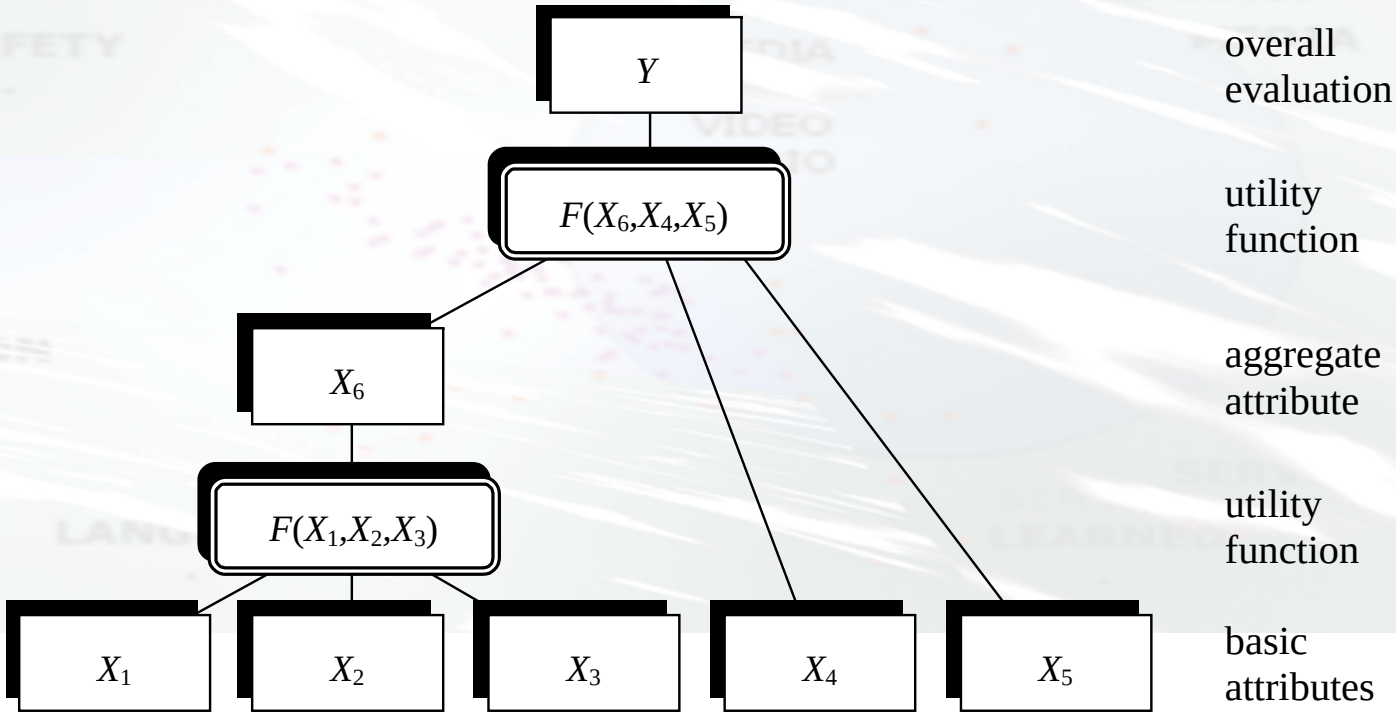
Introduction to KM: Outline

- What is KM: A traditional view
- KM in New economy: A Networked Organizations (NOs) perspective
- Selected knowledge technologies for KM in NOs:
  - Social Network Analysis:
    - A case study – ILPNet2, using Pajek
    - A case study in semi-automated trust modeling
  - Text mining:
    - A case study in semi-automated ontology construction – ILPNet2, using OntoGen



A questionnaire-based trust acquisition method

- **Modeling trust between partners (individuals, institutions) using multi-attribute decision support**



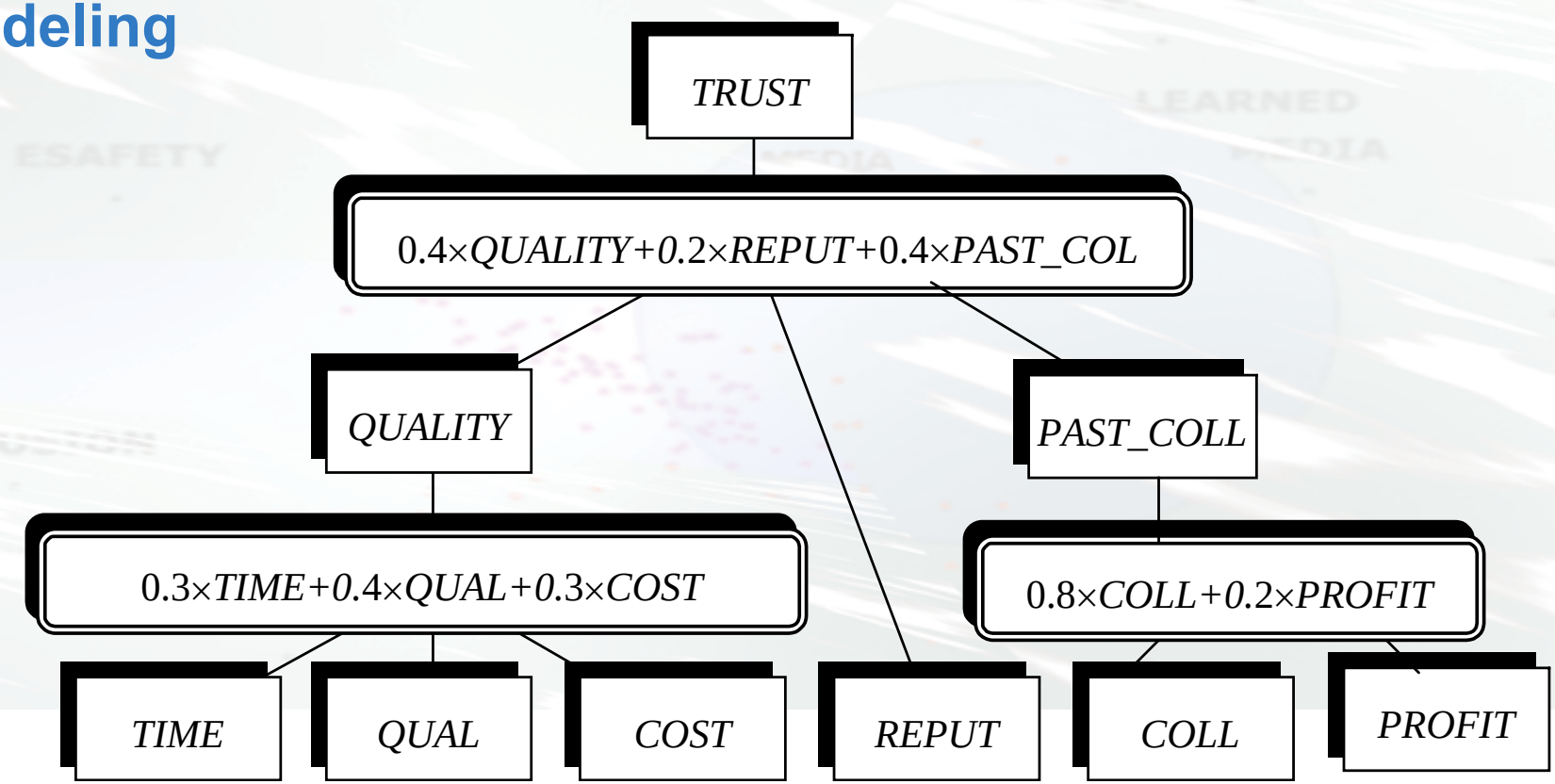
A questionnaire-based trust acquisition method

- E.g., Use user-defined features functions for trust modeling:
  - time
  - quality
  - cost
  - reputation
  - past collaborations
  - profit made in collaborations

$$NormalizedVal = \frac{ActualVal - MinVal}{MaxVal - MinVal}$$

A questionnaire-based trust acquisition method

- **User-defined features and utility functions for trust modeling**



Virtuelle Fabrik

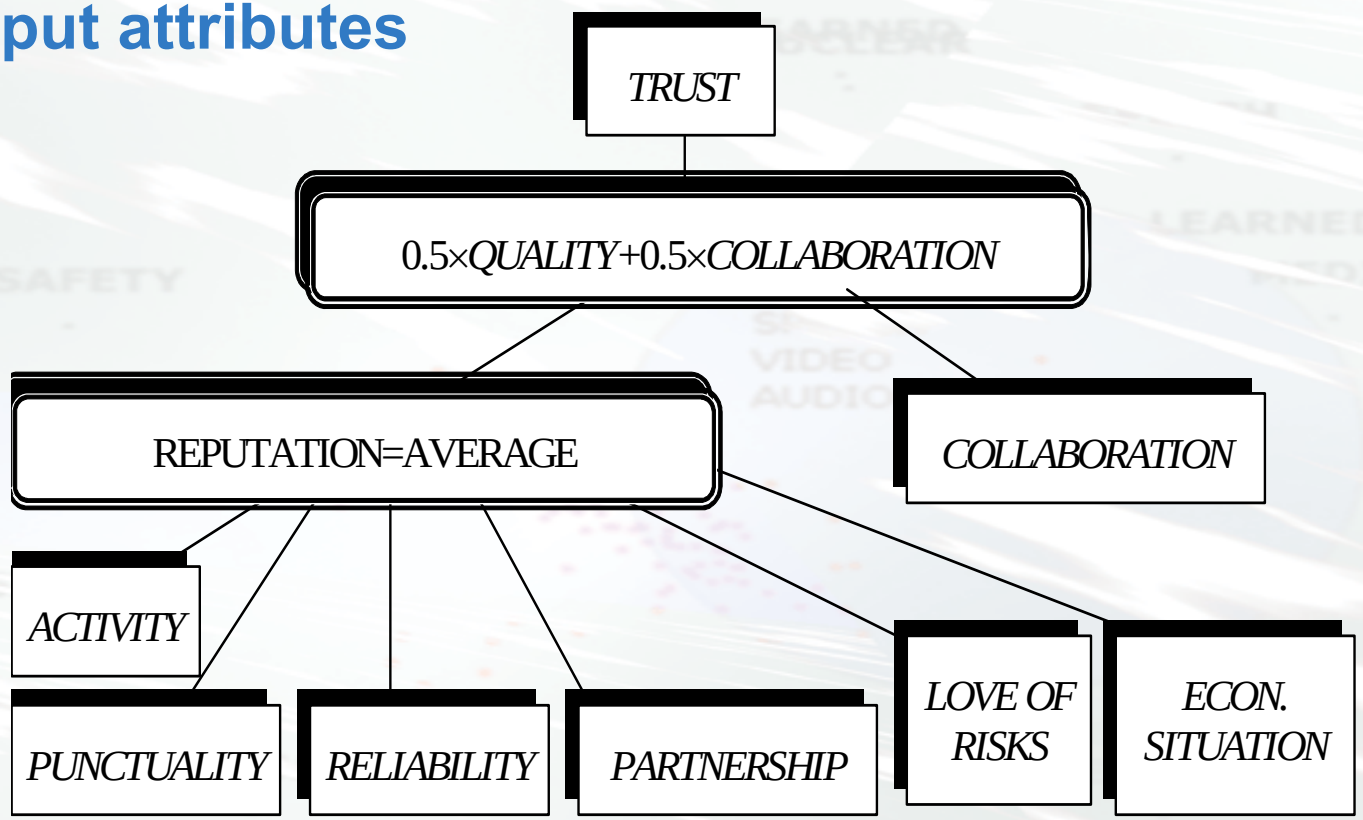
- a Swiss industrial cluster: Virtuelle Fabrik A.G., St. Gallen
- Cluster of partners from mechanical engineering industry
- http://www.virtuelle-fabrik.com
- collaborating expert: Stefan Bolhalter, a VF manager
- The goal of our project: Visualization of partners reputation and collaboration

Virtuelle Fabrik

- Reputation, each of properties has values from 1 to 6 (6 is very good, 1 is very bad)
  - activity
  - punctuality
  - reliability
  - partnership
  - love of risk
  - economical situation
- Collaboration:
  - matrix of collaboration, values from {1, 2, 3}

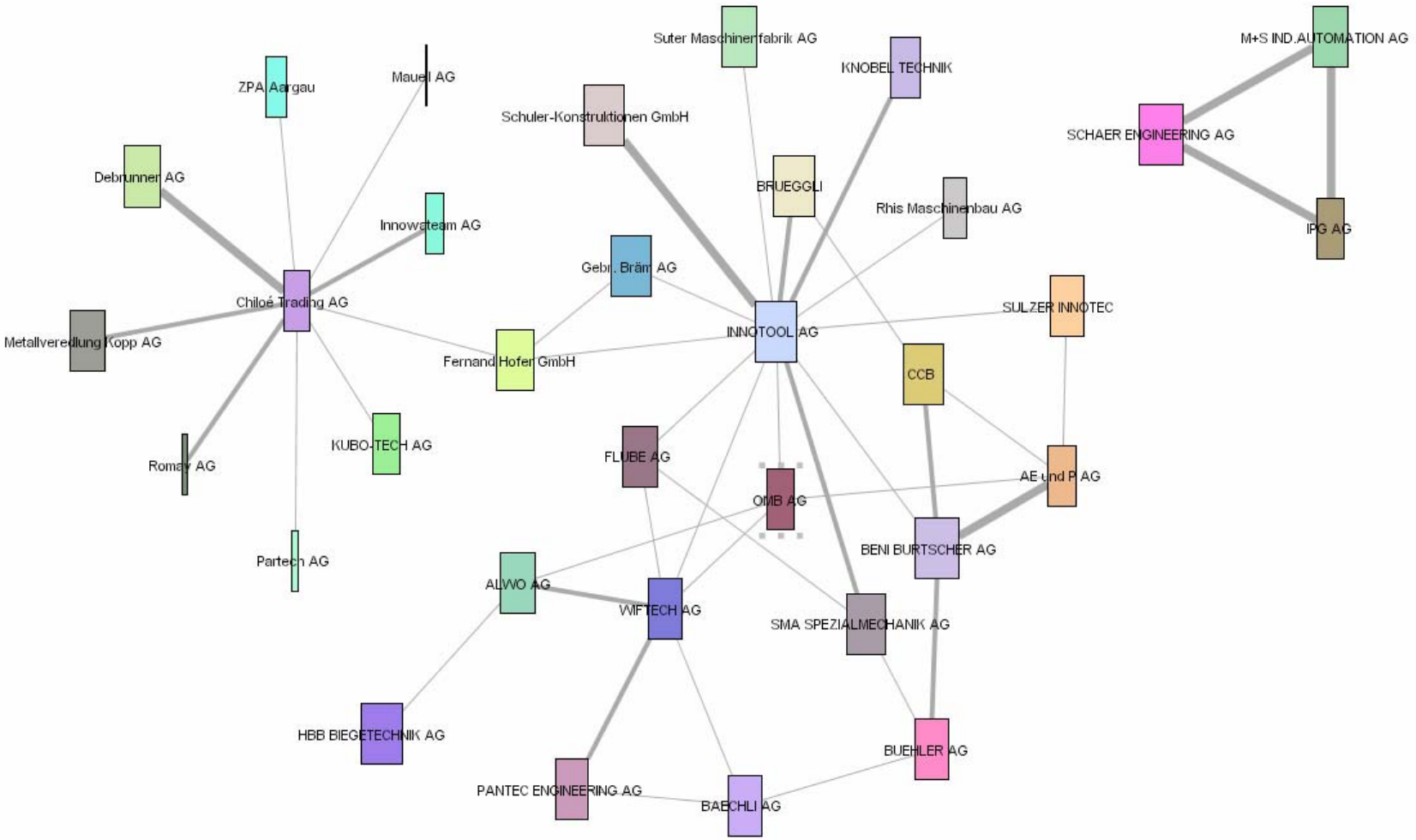
Virtuelle Fabrik

- **Reputation computed as the average of the basic input attributes**



- Other representations possible

Virtuelle Fabrik



Virtuelle Fabrik

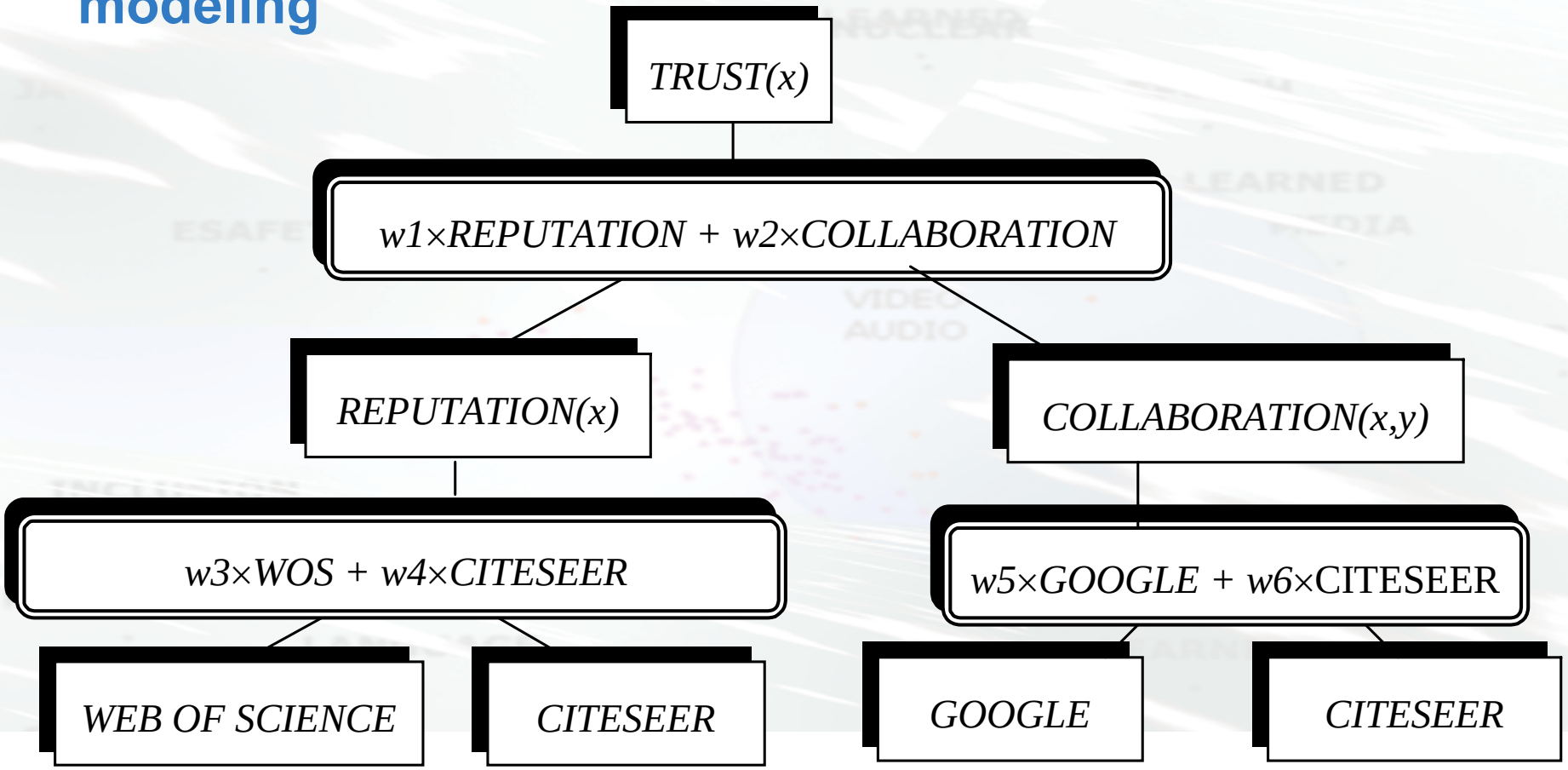
- The proposed decision support approach enables the evaluation and visualization of mutual trust between partners and can be used to find most trusted CNO partners in the process of creating a new VO
- The graph did not show new or surprising relationships to Stefan Bollhalter
- But the graph enabled him to visualize and confirm his knowledge about VF

Trust modeling through Web mining

- Analysis made for 102 individuals from 20 organizations participating in the ECOLEAD project
- Modeling trust between partners (individuals, institutions)
- Trust modeled from two components:
  - Reputation: measured by the # of papers published in SCI journals and # of SCI citations
  - Collaboration: measured by the # of joint papers and # of name co-occurrences on the web

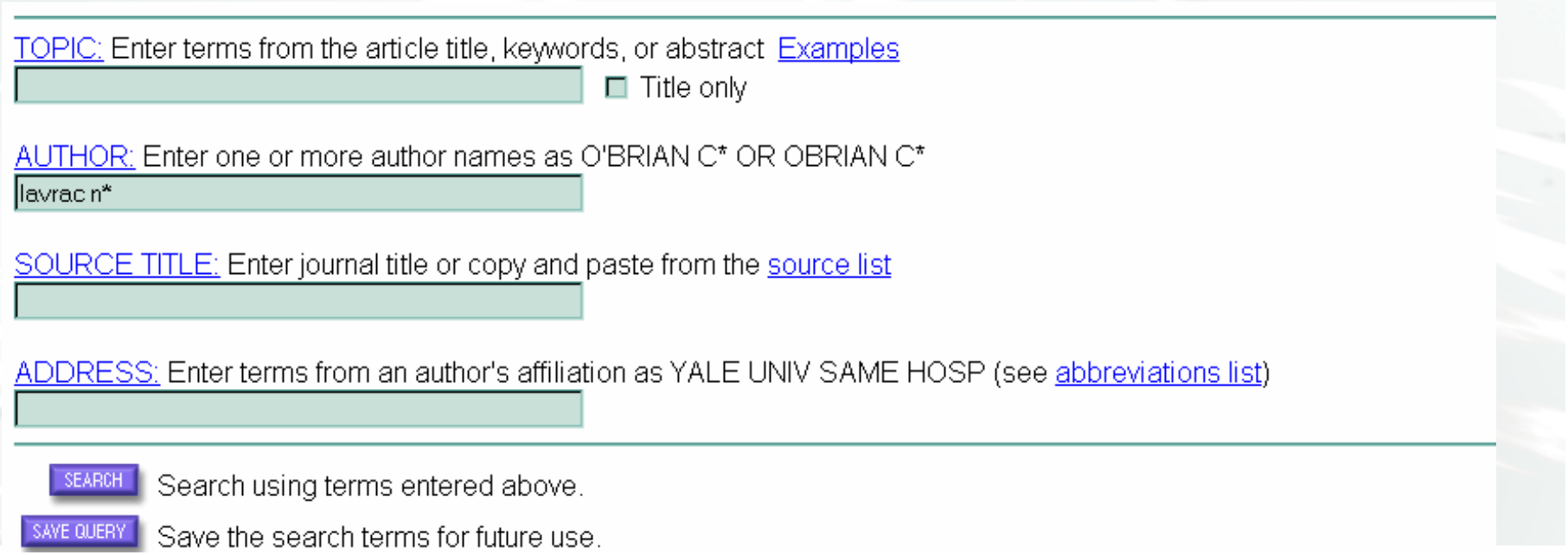
“Trust” computation

- **User-defined features and utility functions for trust modeling**



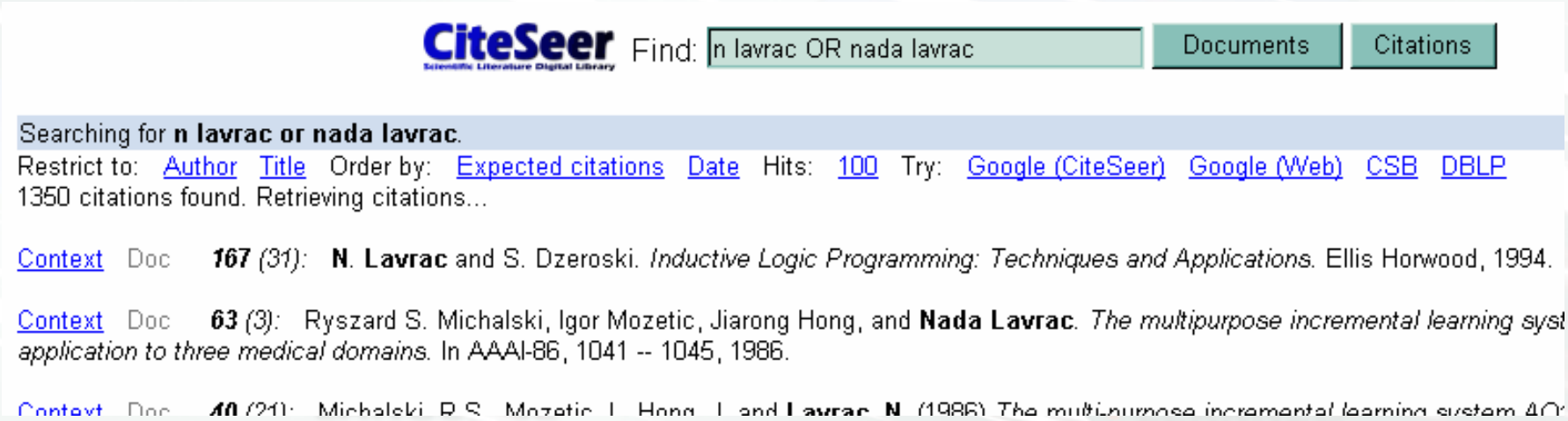
Reputation

- Citation index
- Taken from:
  - Web of Science, http://wos.izum.si
  - Citeseer, http://citeseer.ist.psu.edu



Reputation

- Citation index
- Taken from:
  - Web of Science, http://wos.izum.si
  - Citeseer, http://citeseer.ist.psu.edu

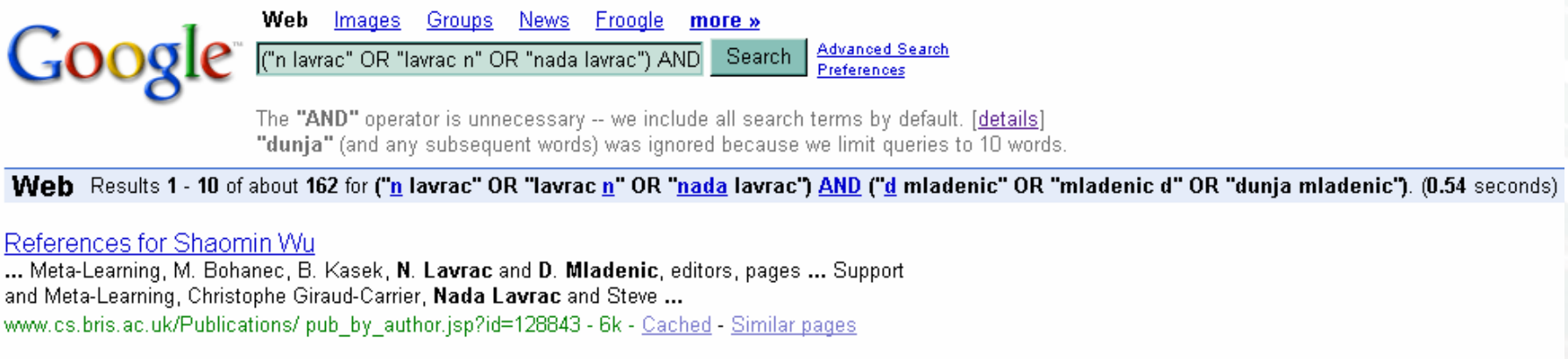


Collaboration

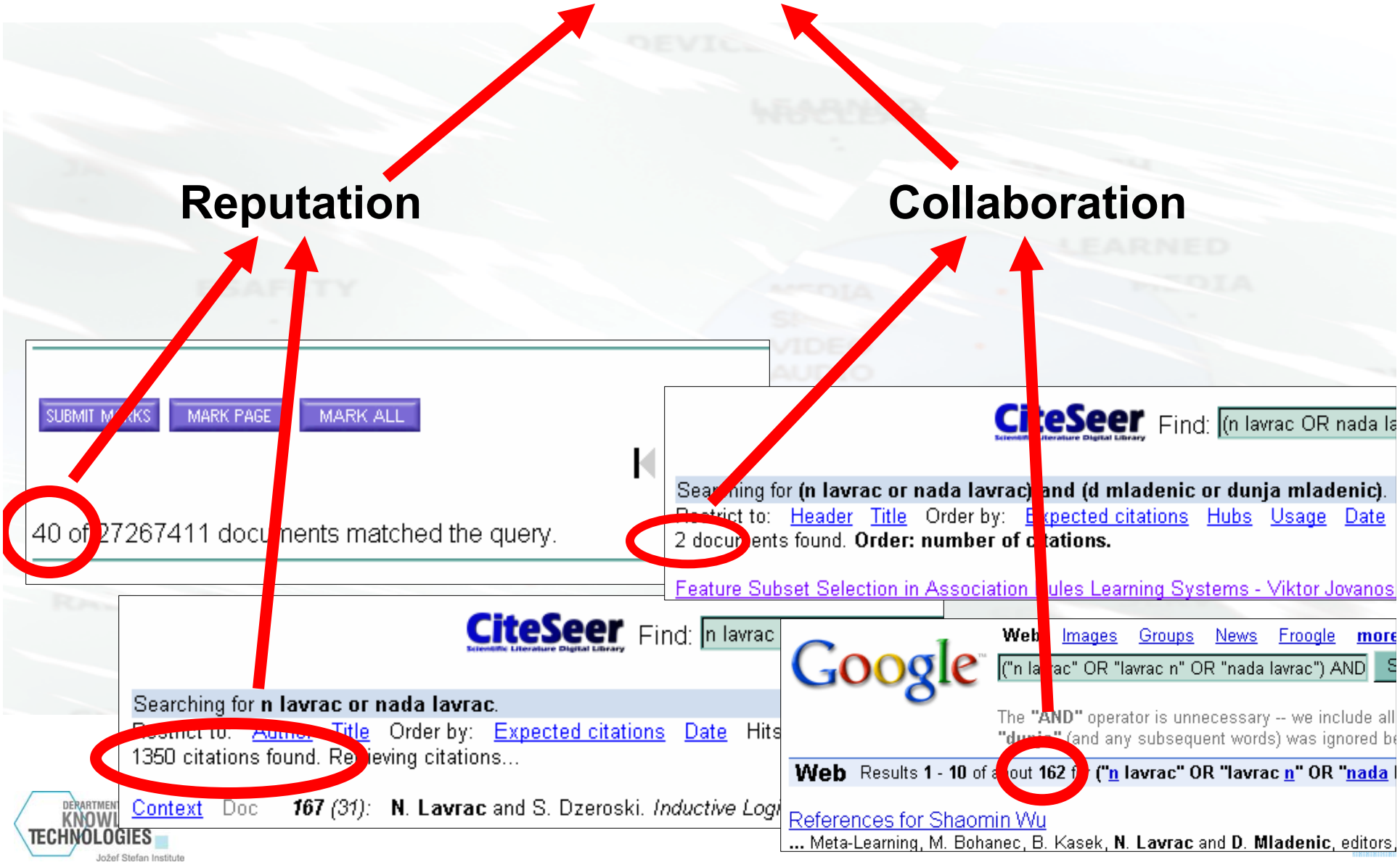
- Number of co-occurrences in:
  - Citeseer, http://citeseer.ist.psu.edu
  - Google, http://www.google.com

Collaboration

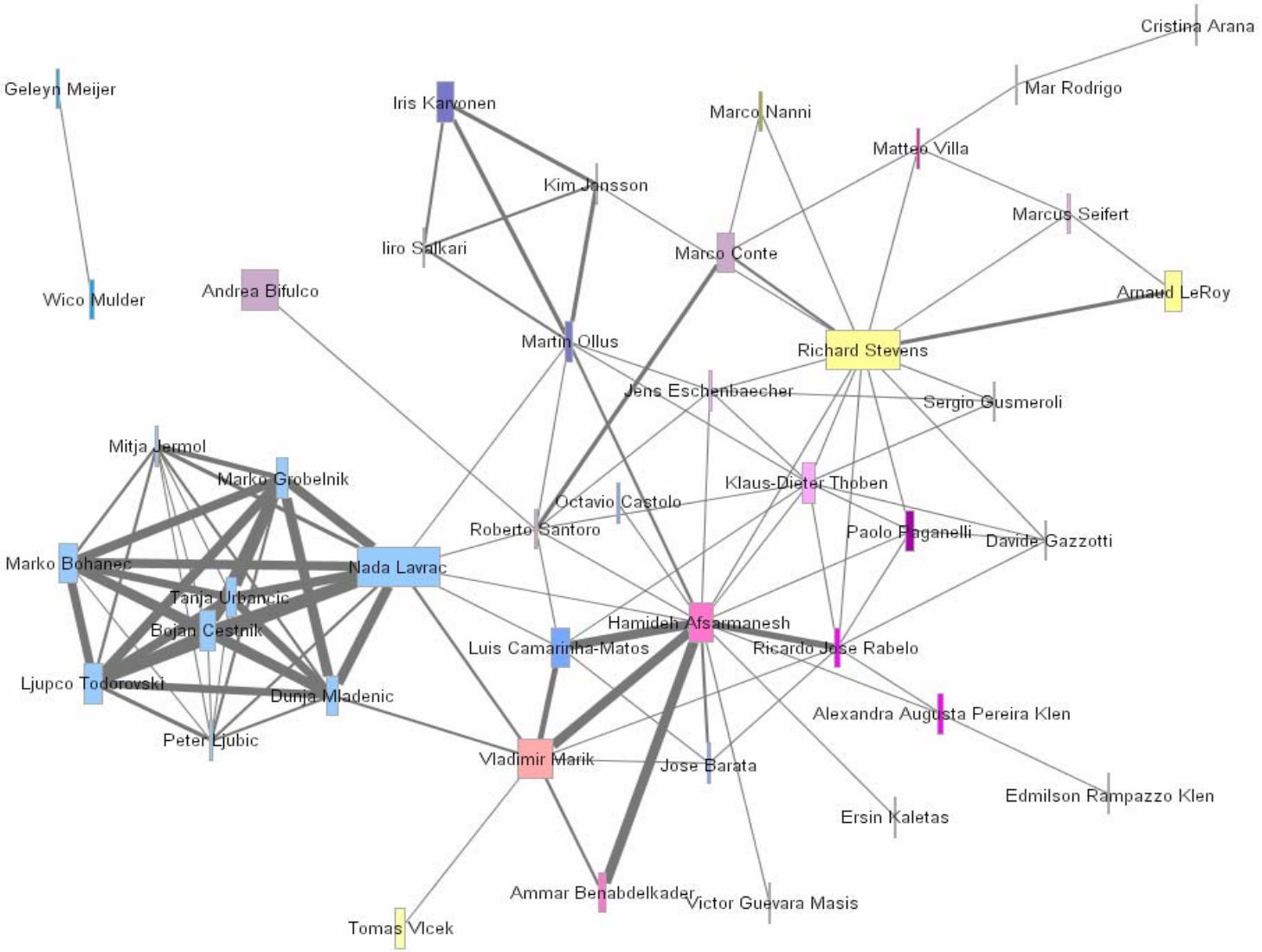
- Number of co-occurrences in:
  - Citeseer, http://citeseer.ist.psu.edu
  - Google, http://www.google.com



Trust

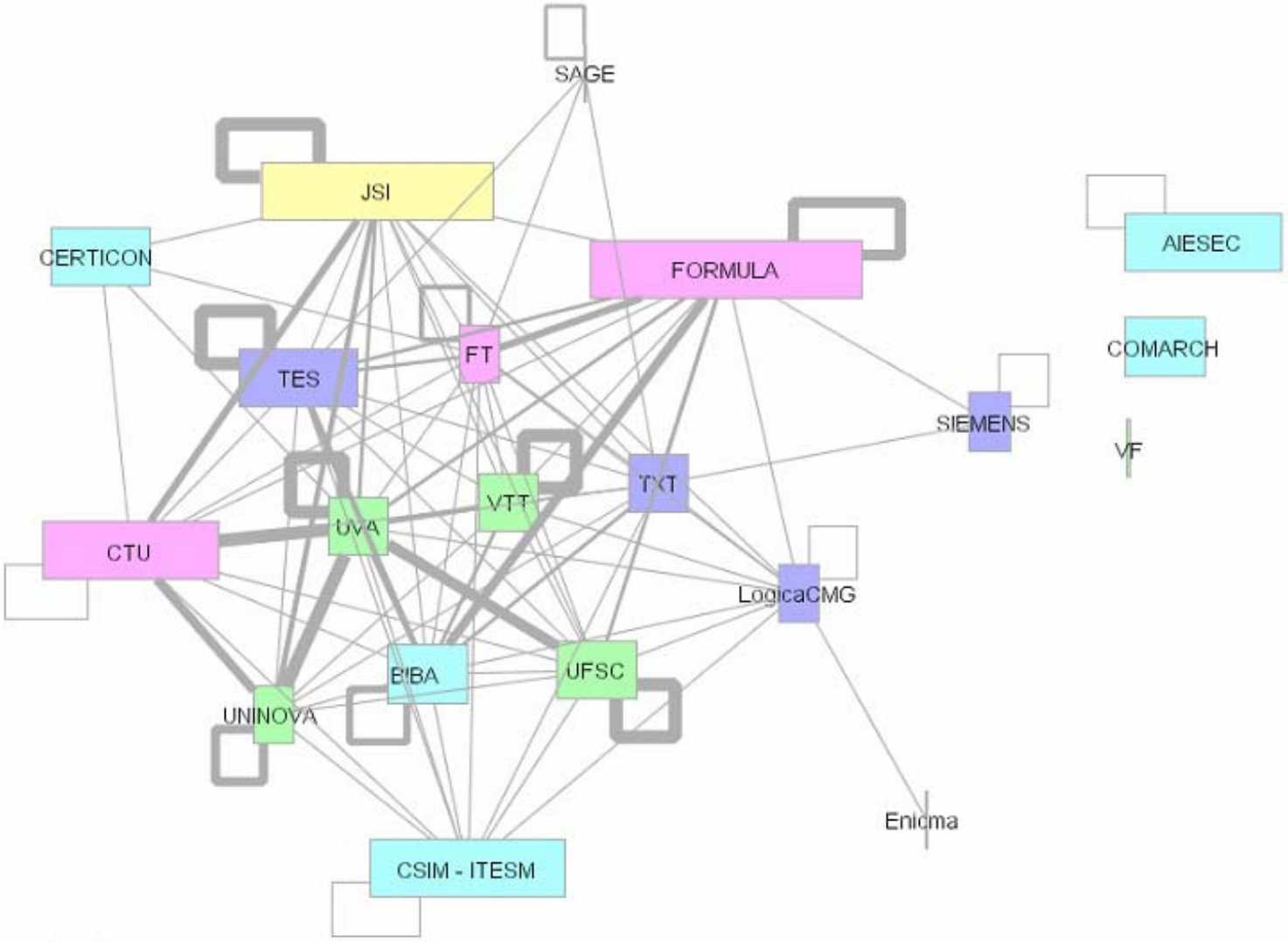

Reputation
Collaboration
SUBMIT MARKS
MARK PAGE
MARK ALL
40 of 27267411 documents matched the query.
CiteSeer
Find:
(n lavrac OR nada la
Searching for (n lavrac or nada lavrac) and (d mladenic or dunja mladenic).
Restrict to: Header Title Order by: Expected citations Hubs Usage Date
2 documents found. Order: number of citations.
Feature Subset Selection in Association Rules Learning Systems - Viktor Jovanos
CiteSeer
Find:
n lavrac
Searching for n lavrac or nada lavrac.
Restrict to: Author Title Order by: Expected citations Date Hits
1350 citations found. Retrieving citations...
Context Doc 167 (31): N. Lavrac and S. Dzeroski. Inductive Logi
Google
Web Images Groups News Froogle more
("n lavrac" OR "lavrac n" OR "nada lavrac") AND
The "AND" operator is unnecessary -- we include all
"dunja" (and any subsequent words) was ignored be
Web Results 1 - 10 of about 162 for ("n lavrac" OR "lavrac n" OR "nada
References for Shaomin Wu
... Meta-Learning, M. Bohanec, B. Kasek, N. Lavrac and D. Mladenic, editors,
DEPARTMENT KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

“Trust” between individuals



DEPARTMENT KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

“Trust” between institutions



Introduction to KM: Outline

- What is KM: A traditional view
- KM in New economy: A Networked Organizations (NOs) perspective
- Selected knowledge technologies for KM in NOs:
  - Social Network Analysis:
    - A case study – ILPNet2, using Pajek
    - A case study in semi-automated trust modeling
  - Text mining:
    - A case study in semi-automated ontology construction – ILPNet2, using OntoGen
    - A case study in structuring of competencies of partners of the Virtuelle Fabrik Swiss industrial cluster, using gCLUTO

Text Mining:
Levels of Text Processing

- **Word Level**
  - **Words Properties**
  - **Stop-Words**
  - **Stemming**
  - **Frequent N-Grams**
  - **Thesaurus (WordNet)**
- Sentence Level
- Document Level
- Document-Collection Level

Stemming and Lemmatization

- Different forms of the same word usually problematic for text data analysis
  - because they have different spelling and similar meaning (e.g. learns, learned, learning,…)
  - usually treated as completely unrelated words
- Stemming is a process of transforming a word into its stem
  - cutting off a suffix (eg., smejala -> smej)
- Lemmatization is a process of transforming a word into its normalized form
  - replacing the word, most often replacing a suffix (eg., smejala -> smejati)

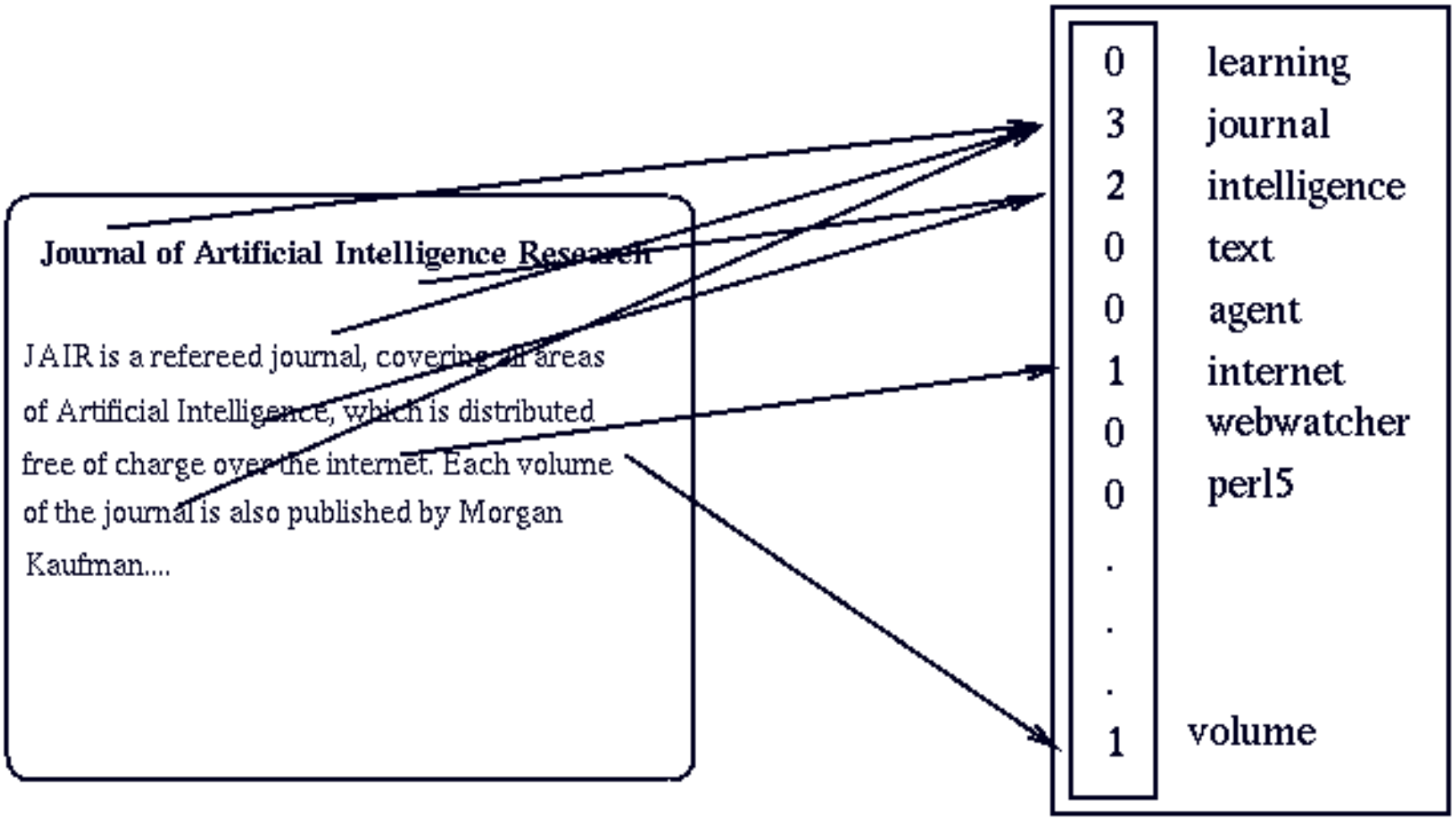
Stemming

- For English it is not a big problem - publicly available algorithms give good results
  - Most widely used is Porter stemmer at http://www.tartarus.org/~martin/PorterStemmer/
- In Slovenian language 10-20 different forms correspond to the same word:
  - ("to laugh" in Slovenian): smej, smejal, smejala, smejale, smejali, smejalo, smejati, smejejo, smejeta, smejete, smejeva, smeješ, smejemo, smejiš, smeje, smejoč, smejta, smejte, smejva

Text Mining: Levels of Text Processing

- Word Level
- Sentence Level
- Document Level
- Document-Collection Level
  - Representation
  - Feature Selection
  - Document Similarity
  - Categorization
  - Clustering
  - Summarization

Bag-of-words document representation



Word weighting

- In bag-of-words representation each word is represented as a separate variable having numeric weight.
- The most popular weighting schema is normalized word frequency TFIDF:

$$tfidf(w) = tf \cdot \log(\frac{N}{df(w)})$$

- Tf(w) – term frequency (number of word occurrences in a document)
- Df(w) – document frequency (number of documents containing the word)
- N – number of all documents
- Tfidf(w) – relative importance of the word in the document

The word is more important if it appears several times in a target document

The word is more important if it appears in less documents

Example document and its representation

- TRUMP MAKES BID FOR CONTROL OF RESORTS Casino owner and real estate Donald Trump has offered to acquire all Class B common shares of Resorts International Inc, a spokesman for Trump said. The estate of late Resorts chairman James M. Crosby owns 340,783 of the 752,297 Class B shares. Resorts also has about 6,432,000 Class A common shares outstanding. Each Class B share has 100 times the voting power of a Class A share, giving the Class B stock about 93 pct of Resorts' voting power.

- [RESORTS:0.624] [CLASS:0.487] [TRUMP:0.367] [VOTING:0.171] [ESTATE:0.166] [POWER:0.134] [CROSBY:0.134] [CASINO:0.119] [DEVELOPER:0.118] [SHARES:0.117] [OWNER:0.102] [DONALD:0.097] … [STOCK:0.035] [YORK:0.035] [PCT:0.022] [MARCH:0.011]

Cosine similarity between document vectors

- Each document is represented as a vector of weights D = <x>
- Similarity between two vectors is estimated by the similarity between their vector representations (cosine of the angle between the two vectors):

$$Similarity\ (D_1, D_2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2}\sqrt{\sum_k x_k^2}}$$

Document Clustering

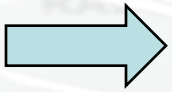
- Clustering is a process of finding natural groups in data in a unsupervised way (no class labels pre-assigned to documents)
- Document similarity is used
- Most popular clustering methods are:
  - K-Means clustering
  - Agglomerative hierarchical clustering
  - EM (Gaussian Mixture)
  - …

K-Means clustering

- Given:
  - set of documents (eg., word-vectors with TFIDF),
  - distance measure (eg., cosine similarity)
  - K - number of groups
- For each group initialize its centroid with a random document
- While not converging
  - each document is assigned to the nearest group (represented by its centroid)
  - for each group calculate new centroid (group mass point, average document in the group)

Introduction to KM: Outline

- What is KM: A traditional view
- KM in New economy: A Networked Organizations (NOs) perspective
- Selected knowledge technologies for KM in NOs:
  - Social Network Analysis:
    - A case study – ILPNet2, using Pajek
    - A case study in semi-automated trust modeling
  - Text mining:
    - A case study in semi-automated ontology construction – ILPNet2, using OntoGen
    - A case study in structuring of competencies of partners of the Virtuelle Fabrik Swiss industrial cluster, using gCLUTO



Goals of ILPnet2 analysis

- **Research contents analysis through ontology construction (OntoGen)**
  - Which are the main topics explored by ILP researchers? Can one reverse engineer the list of ILPNet2 keywords ? Can one classify the ILP papers into the suggested keyword categories ?
- **Improve ontology construction through term extraction and visualization**

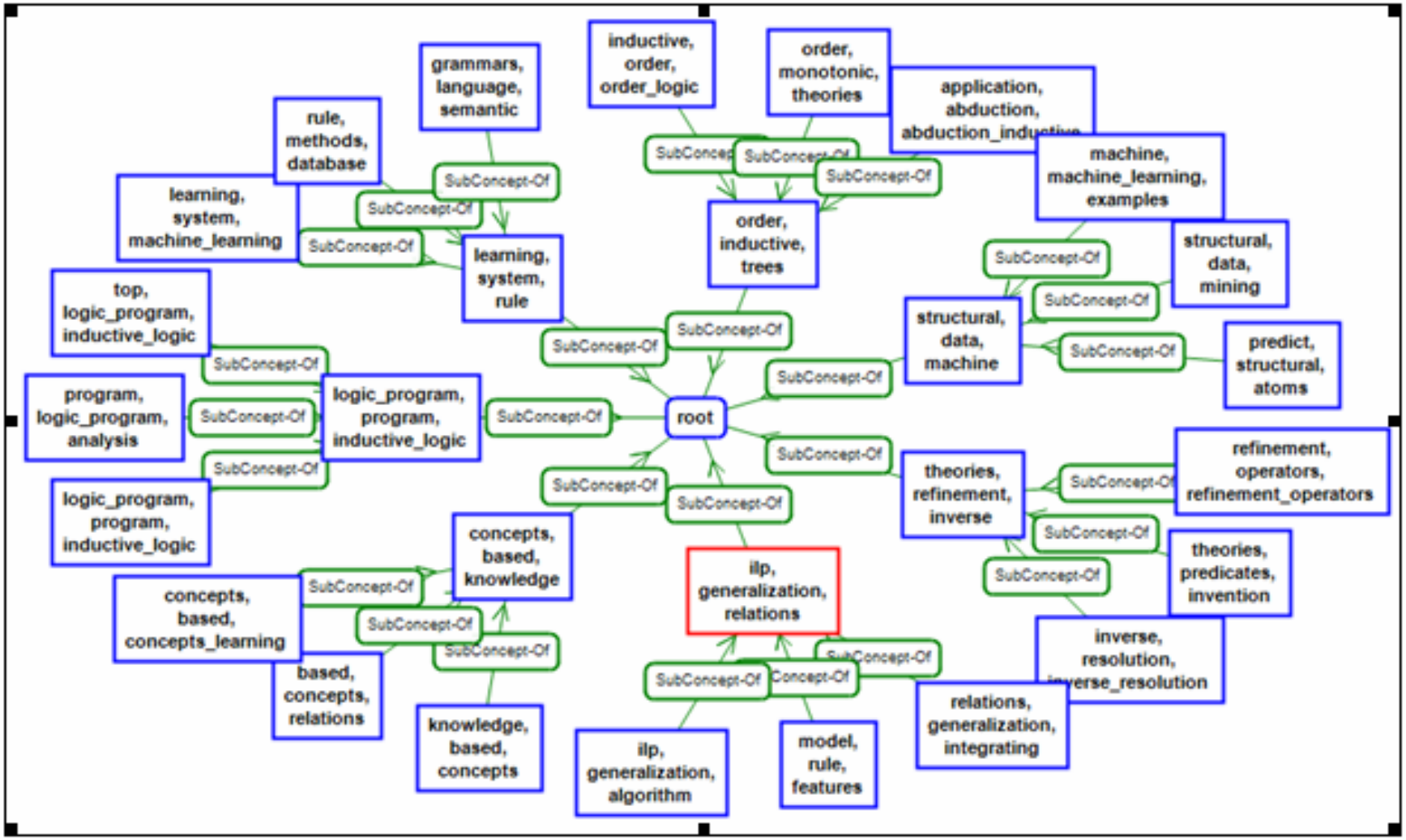
Ontology construction with OntoGen

- OntoGen: a system for data-driven semi-automated ontology construction
  - Semi-automatic: it is an interactive tool that aids the user
  - Data-driven: aid provided by the system is based on the data (text documents) provided by the user
- Freely available at http://ontogen.ijs.si

Data extraction and preparation

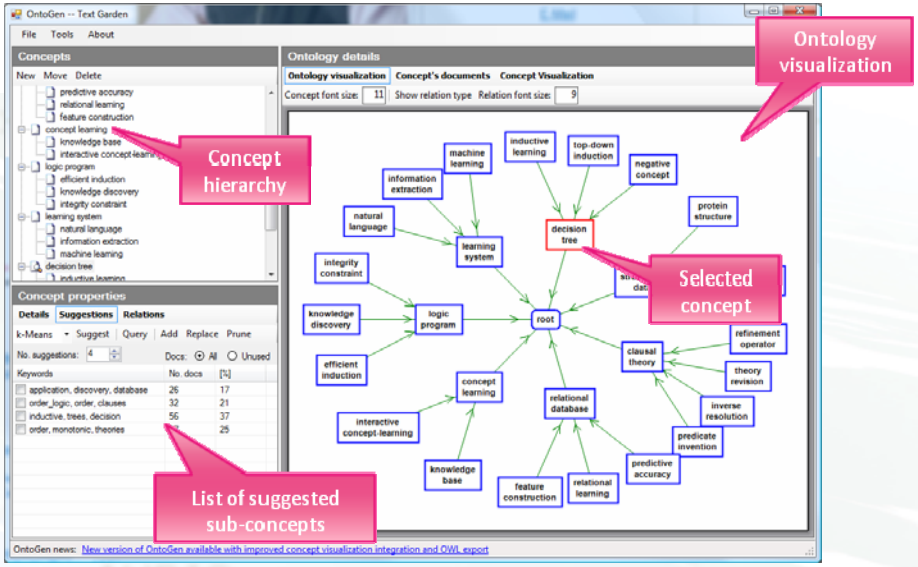
- Data in BibTeX format, one file for every year **http://www.cs.bris.ac.uk/~ILPnet2/Tools/Reports/Bibtexs/2003, ...,**
- Data acquired with the wget utility
- Collected data converted into the XML format
- Text data preprocessed using a predefined list of stop-words and the Porter stemmer.

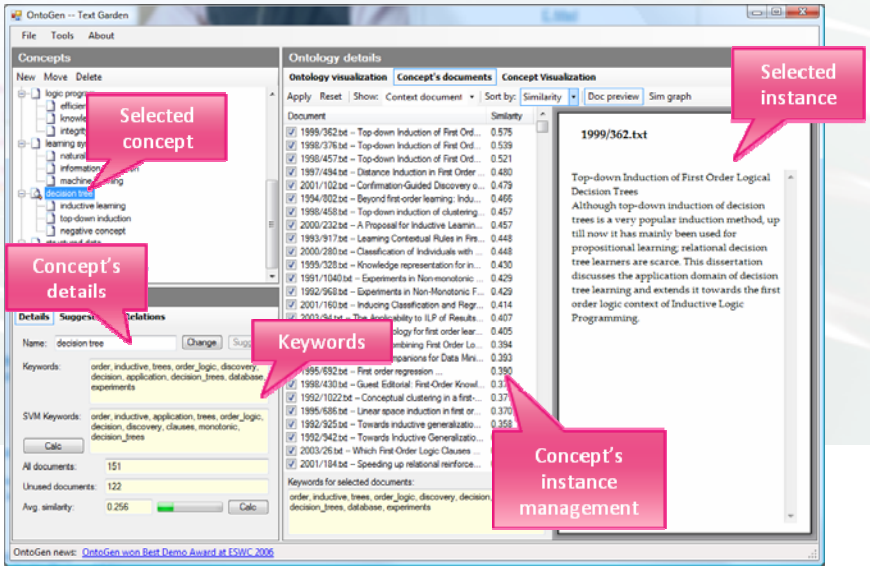
OntoGen ontology construction using k-means clustering



OntoGen

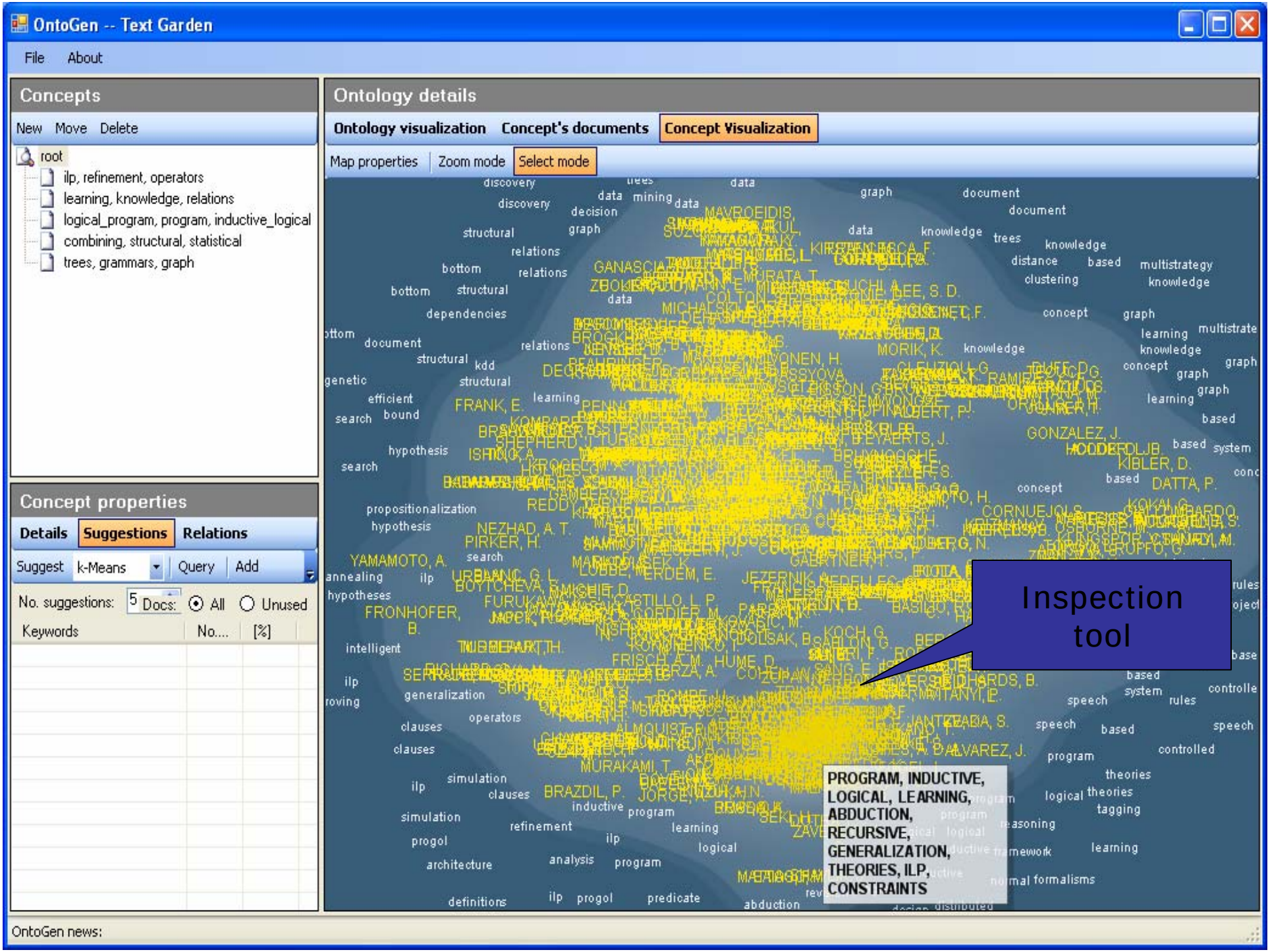
- Ontology construction and learning
- Semi-Automatic:
  - Text-mining methods provide suggestions and insights into the domain
  - The user can interact with parameters of text-mining methods
  - All the final decisions are taken by the user
- Data-Driven:
  - Most of the aid provided by the system is based on some underlying data provided by the system
  - Instances are described by features extracted from the data (e.g. bag-of-words vectors)

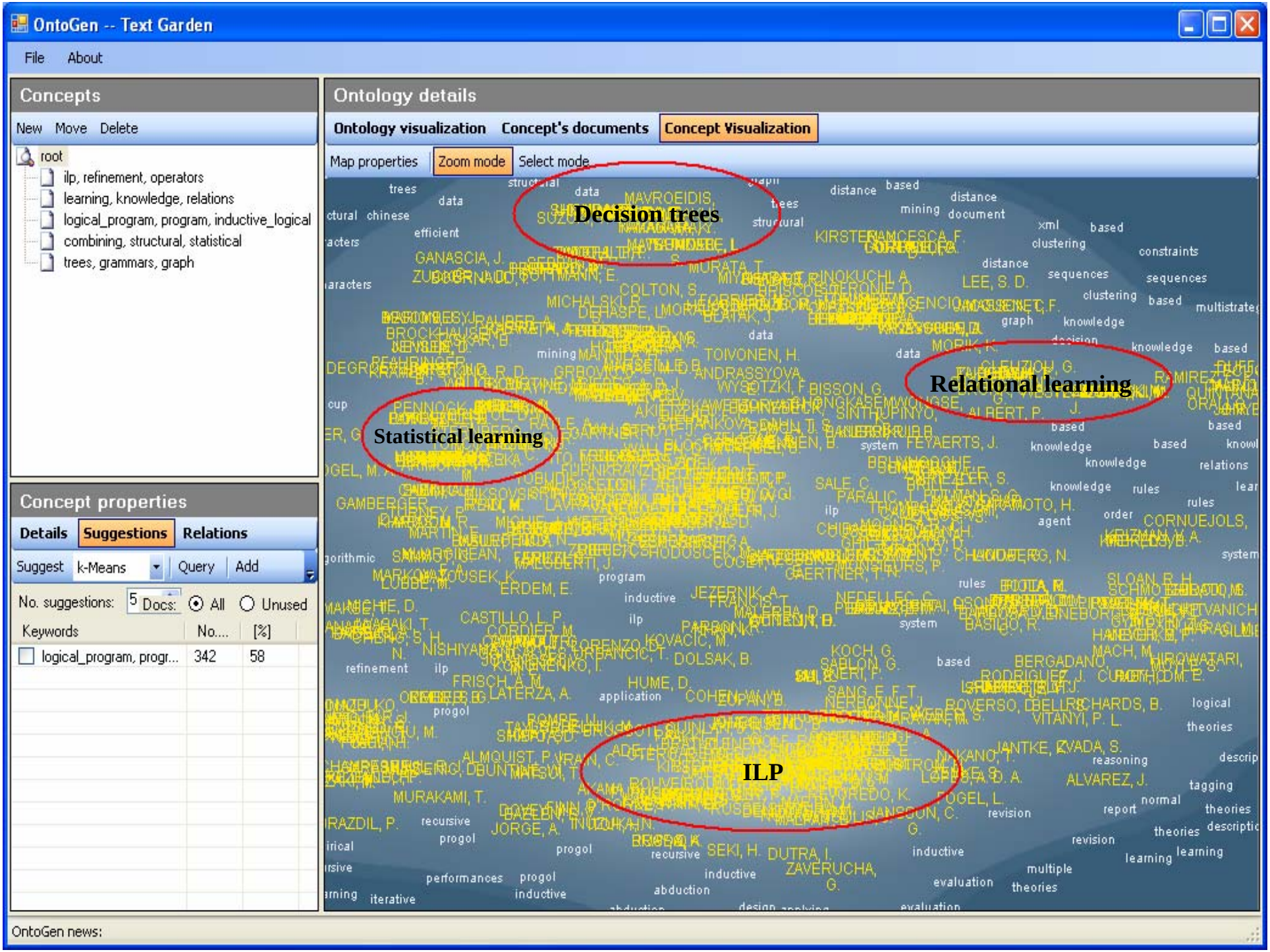




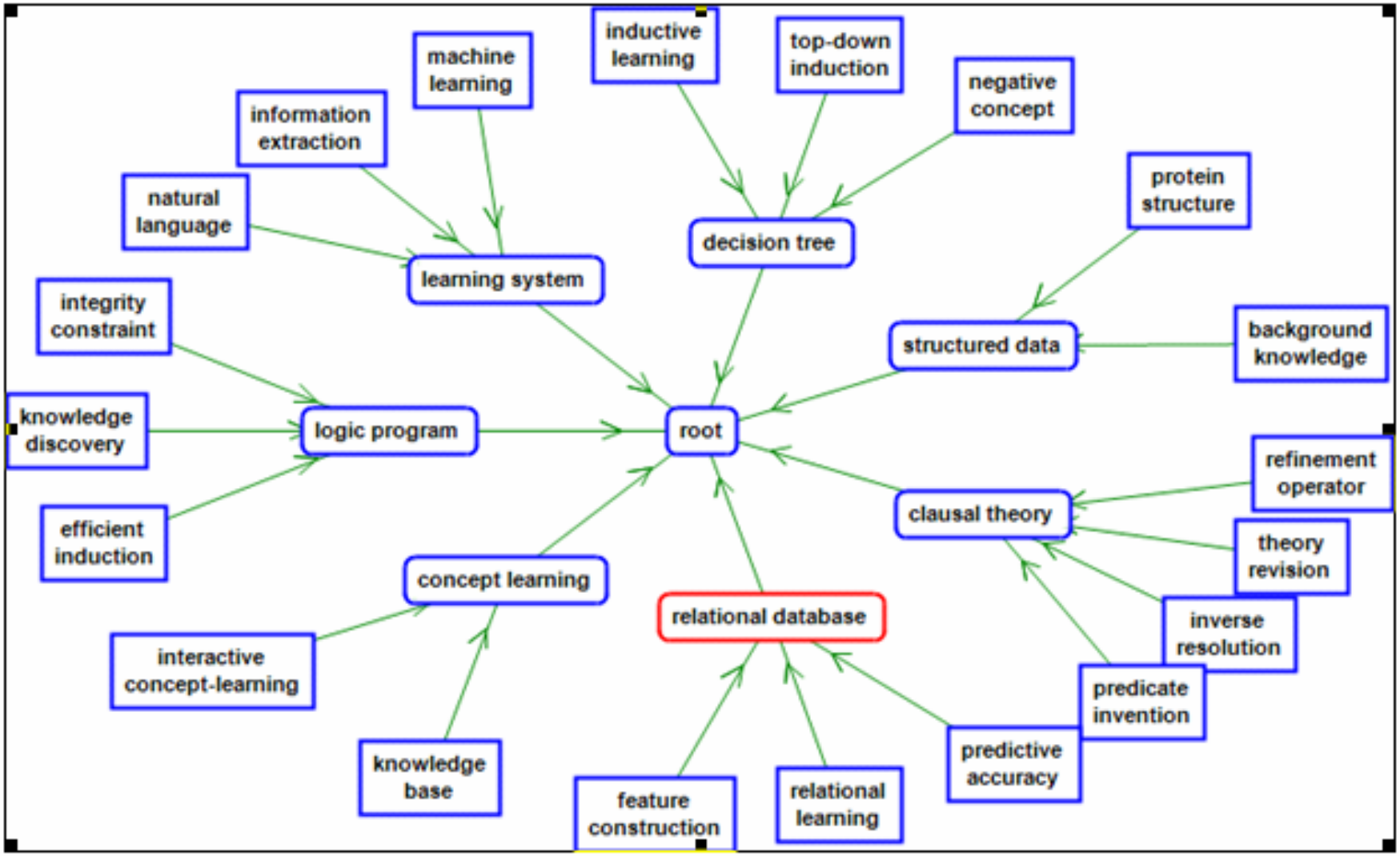
Recent advances in concept naming and visualization

- Visualization and sub-concept selection with DocumentAtlas
- Advanced concept naming with OntoTerm
  - Using TermExtractor
  - Populating the terms and keyword extraction


OntoGen -- Text Garden
File
About
Concepts
New
Move
Delete
root
ilp, refinement, operators
learning, knowledge, relations
logical_program, program, inductive_logical
combining, structural, statistical
trees, grammars, graph
Concept properties
Details
Suggestions
Relations
Suggest
k-Means
Query
Add
No. suggestions:
5
Docs:
All
Unused
Keywords
No....
[%]
Ontology details
Ontology visualization
Concept's documents
Concept Visualization
Map properties
Zoom mode
Select mode
Inspection tool
PROGRAM, INDUCTIVE, LOGICAL, LEARNING, ABDUCTION, RECURSIVE, GENERALIZATION, THEORIES, ILP, CONSTRAINTS
OntoGen news:


OntoGen -- Text Garden
File
About
Concepts
New
Move
Delete
root
ilp, refinement, operators
learning, knowledge, relations
logical_program, program, inductive_logical
combining, structural, statistical
trees, grammars, graph
Concept properties
Details
Suggestions
Relations
Suggest
k-Means
Query
Add
No. suggestions:
5
Docs:
All
Unused
Keywords
No....
[%]
logical_program, progr...
342
58
Ontology details
Ontology visualization
Concept's documents
Concept Visualization
Map properties
Zoom mode
Select mode
Decision trees
Relational learning
Statistical learning
ILP
OntoGen news:

Improved OntoGen ontology construction - advanced concept naming



Advanced concept naming method

OntoTermExtractor methodology:

- **Use document clustering to find the nodes in the topic ontology**
- **Perform term extraction from document clusters using the TexmExtractor tool, freely available at http://lcl2.uniroma1.it/termextractor,**
- **Populate the term vocabulary and repeatedly perform keyword extraction**
- **Choose sub-concept names by comparing the best ranked terms with the extracted keywords**

Best-ranked terms extracted from ILPNet2 publications by TermExtractor

Top-10 terms extracted from ILPNet2	Term Weight	Domain Relevance	Domain Consensus	Lexical Cohesion
inductive logic	0.928	1.000	0.968	0.557
logic programming	0.924	1.000	0.988	0.293
inductive logic programming	0.893	1.000	0.966	0.181
background knowledge	0.825	1.000	0.737	0.835
logic program	0.824	1.000	0.867	0.203
machine learning	0.785	1.000	0.777	0.221
data mining	0.776	1.000	0.691	0.672
refinement operator	0.757	1.000	0.572	1.000
decision tree	0.742	1.000	0.613	0.714
inverse resolution	0.722	1.000	0.557	0.894
experimental result	0.718	1.000	0.594	0.684

Populating the term vocabulary: Invoking Google search

Polulation of the term vocabulary, extracted by the TermExtractor, is performed as follows:

- Google web search was invoked by a query, generated from each term, by taking its words and attaching an extra keyword "ILP" to limit the search to ILP related web pages.
- The query was then sent to Google and snippets of the returned search results were used to populate the term.

Term Population - Details

- We want model for each term
  - Can tell if term is related to a given document
- For this we need instances
  - Retrieval approach
    - Each term was issued as a query over ILPNet2 documents
    - Documents weighted according to their TFIDF score were used as instances
  - Google search approach
    - Each term was issued as a query to Google
      - We added keyword “ILP” to limit results to ILP domain
      - Example: Term: “inductive logic programming, Query: “ILP inductive logic programming”
    - Snippets from results were used as instances
- In practice Google approach seams to give better results Web is richer and more diverse data source

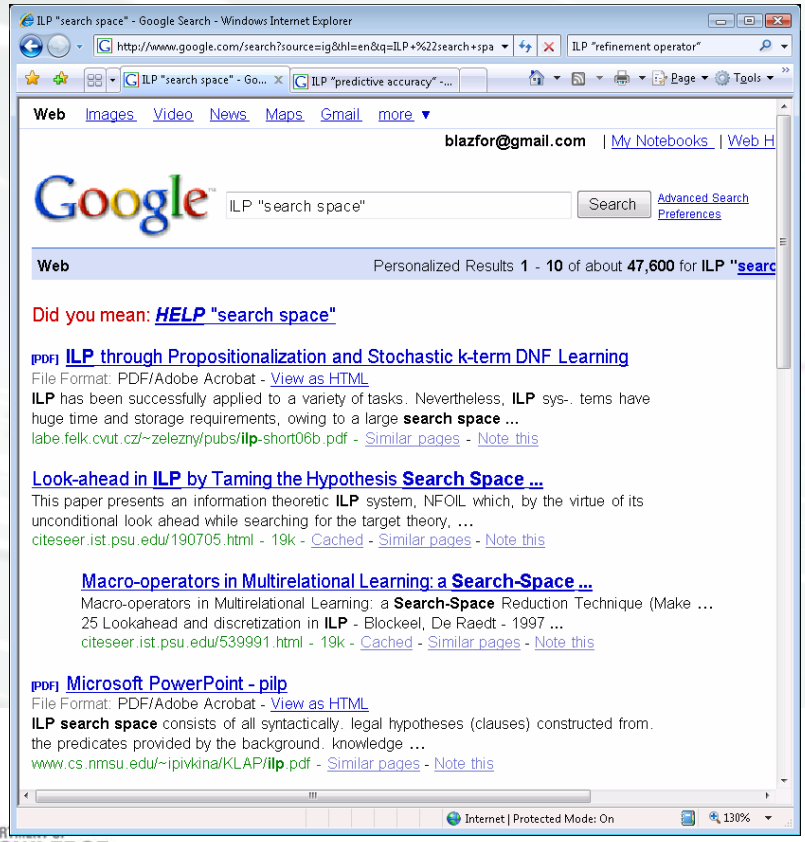
Populating the term vocabulary: Invoking snippets to populate the terms

Polulation of the term vocabulary, extracted by the TermExtractor, is continued as follows:

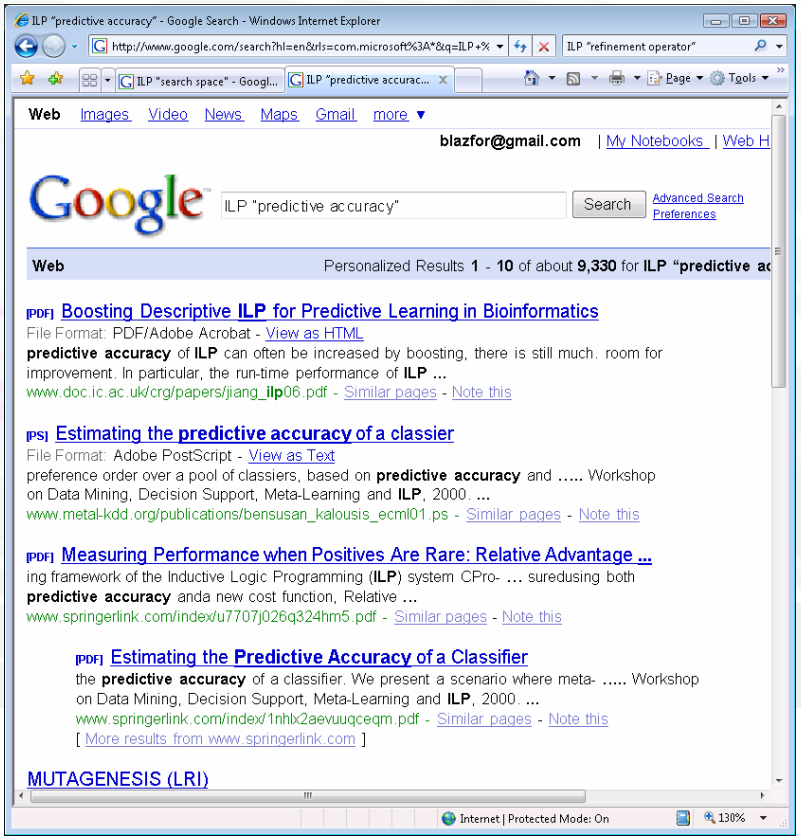
- For each query the Google snippets of the first 1000 results were used.
- The snippets served as input for term modeling.
- The models generated for each term, using this data, were then used for generating the concept suggestions and name suggestions in OntoGen.

Examples

Q: *ILP* “search space”



Q: *ILP* “predictive accuracy”



Sample snippets for a given term

Top 5 snippets that were returned for the query "ILP predictive accuracy":

- Boosting Descriptive ILP for Predictive Learning in Bioinformatics -- general, this means that a higher predictive accuracy can be achieved. Thirdly,. although some predictive ILP systems may produce multiple classification ...
- Imperial College Computational Bioinformatics Laboratory (CBL) -- Results on scientific discovery applications of ILP are separated below ... Progol's predictive accuracy was equivalent to regression on the main set of 188 ...
- Evolving Logic Programs to Classify Chess-Endgame Positions -- indicate that in the cases where the ILP algorithm performs badly, the introduc-. tion of either union or crossover increases predictive accuracy. ...

Comparing best-ranked terms with OntoGen-generated concept keywords

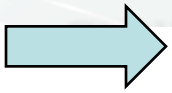
- For all the seven concepts the first-ranked term suggested from the vocabulary was selected.
- Sample lists of concepts with selected name, followed with the second suggested name, and the most important keywords as chosen by OntoGen:
  - Learning system (learning algorithm) -- learning, system, rule, language, methods, machine_learning, machine, approach, ilp, grammars
  - Decision tree (logical decision tree) -- order, inductive, trees, order_logic, discovery, decision, application, decision_trees, database, experiments
  - …
- OntoTerm method has, through term extraction and population, indeed succeeded to rank the terms and choose them for concept naming in a meaningful way.

ILPNet2 Summary

- Ontology construction with OntoGen was successfully used for research contents analysis in ILPNet2, but naming of concepts proved to be problematic
- A novel concept naming methodology was developed
- The developed OntoTerm method has, through term extraction and population, indeed succeeded to appropriately rank the terms, choosing them for concept naming in a meaningful way.
- Results of analysis were evaluated by domain expert (NL ☺)
- In further work we plan to implement this methodology as part of the OntoGen toolbox.

Introduction to KM: Outline

- What is KM: A traditional view
- KM in New economy: A Networked Organizations (NOs) perspective
- Selected knowledge technologies for KM in NOs:
  - Social Network Analysis:
    - A case study – ILPNet2, using Pajek
    - A case study in semi-automated trust modeling
  - Text mining:
    - A case study in semi-automated ontology construction – ILPNet2, using OntoGen
    - A case study in structuring of competencies of partners of the Virtuelle Fabrik Swiss industrial cluster, using gCLUTO



Ontology construction experiment: Structuring and visualization of NO competencies

- Approach: Applying knowledge mapping tools for competency visualization and structuring of competencies of partners of the Virtuelle Fabrik Swiss industrial cluster

Structuring and visualization of VF competencies

- **Structuring the expertise of companies: Analysis of VF partners business data (a subset of VF industrial cluster - 20 companies from the Bodensee sub-cluster)**

Table 2 - Company identifiers assigned to company names.

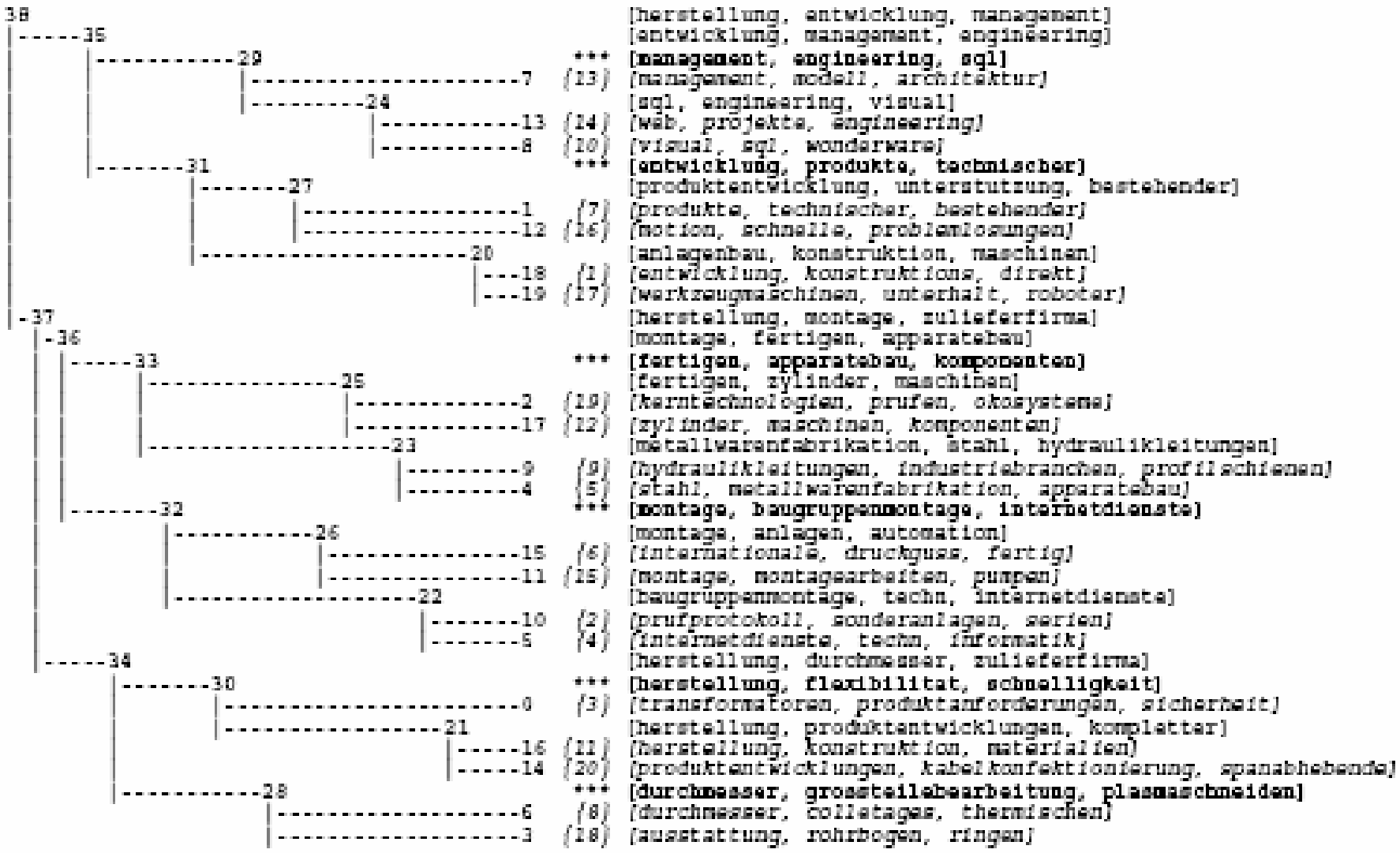
1 AE&P	2 ALWO	3 Bachli	4 Bruggli	5 Beni
6 Buchler	7 Ccb	8 Flube	9 KBB	10 Heese
11 Innotool	12 Knobel	13 IPG	14 M+S	15 OMB
16 Pantec	17 Schar	18 SMA	19 Sulzer	20 Wiftech

- **Our approach: Apply hierarchical k-means document clustering and visualization**

Descriptions of 20 VF partners

	A	B	C	D
3	Virtuelle Fabrik Euregio Bodensee (VFEB): Firmenprofile			vfeb.ch
4				
5				
6	Firma	Mitarbeiter	Produkte, Dienstleistungen	Kernkompetenzen bzw. Kerntechnologien
7				
8	AE&P AG, Aerne Engineering & Projectmanagement AG	6	Entwicklung und Konstruktion im Bereich Maschinen- und Anlagenbau, Lieferung von Komplettanlagen, Konstruktions-einsätze direkt beim Kunden, Projektmanagementmandate	Gesamtlösungen in der Automation, Entwicklung von Handgeräten, Breites CADKnowhow : Autocad2d/3d, Bravo, Catia, ME-10/30, Euklid
9	ALWO AG, Kreuzlingen	21	Zulieferfirma für Halbleiterindustrie; Werkzeugbau; Sonderanlagen; Baugruppenmontage	Serien bis 200 Stück; Kleinteilefertigung; drehen, fräsen; Montage mit Prüfprotokoll
10	Bächli AG, Kriens	48	Transformatoren, Drosseln, Speisegeräte	Flexibilität, Schnelligkeit, Sicherheit in den Normenerfüllung bei Produktanforderungen nach EN/IEC/CSA-Normen
11	Brüggli Produktion und Dienstleistung, Romanshorn	249	Druckerei, Informatik- u. Internetdienste, Fahrradanhänger, Techn. Textilprodukte wie Gurten, Taschen, Planen. Industrie- u. Kleingerätemontage, Mech. Bearbeitungen	Offsetdruck, Informatik- u. Internetdienste, Techn. Textilfertigung; Mechanik/ Montage; Profil-Rohrbiegen, Fräsen, Bohren, Baugruppenmontage
12	Beni Burtscher AG, Freidorf	38	Lohnarbeitsbetrieb und Zulieferer für die komplette Blechbearbeitung, Metallwarenfabrikation, Apparatebau und Metalldrückerei.	Stanzen und Laserschneiden von Blechteilen, Abkanten / Biegen / Pressen, Metallwarenfabrikation, Apparatebau und Metallkonstruktionen, Druckbehälterbau (SVTI-Zulassungen), Metalldrücken (spanlose Umformung), Schweissen von Stahl
13	Bühler AG, Uzwil	7.000	Systeme, Anlagen und Maschinen für die internationale Food-, Nonfood- und Druckguss-Industrie	Systemlieferant für Komplettherstellung vom Detail bis zur Fertig-maschine mit Kernkompetenzen in Montage, Automation/ Schaltschrank-bau, spanende/spanlose Fertigung, Logistik
14	ccb, Winterthur	2	Unterstützung bei der Entwicklung neuer und/oder bestehender technischer Produkte	Produktentwicklung
15	Flube AG, Lommiswil	9	Zulieferfirma für Präzisionsdrehteile bis 42mm Durchmesser und Décolletages bis 16mm Durchmesser.	CNC-Lang- und Kurzdrehen von Kleinst- Mittel- und Grosserien sowie die Herstellung komplett nach Kundenwunsch inkl. aller Thermischen- und Galvanischen Oberflächenbehandlungen.
16	HBB Biegetechnik AG, Walzenhausen	40	Schienenfahrzeugbau: Haltestangen,Hydraulikleitungen,Fensterrahmen,Türprofile,Eckstützen Automotiv: Ueberrollbügel,Spannbügel für Verdecke,Spaceframe,Hydraulikleitungen Fördertechnik: Profilschienen,Gehängestangen Aircraft: Strukturelle für Sitze	Als Zulieferer von Komponenten nammhafter Industriebranchen verfügen wir über ein grosses Know-how in den folgenden Techniken der Metallbearbeitung: 2-/3-D Biegen,Abkanten,Schweissen,CNC-Fräsen und Aluminiumwärmebehandlungen.
17	Ing.-Büro Heese	2	Engineering und Entwicklung von Software-Projekten für die Automatisierung in der Labormeß-, Prozeß- und Fertigungstechnik (PC-seitig)	Visualisierung / Steuerungs-, Regelungstechnik / Datenhandling: Wonderware: InTouch / InSQL; Intellution/GE Fanuc: iFix / iHistorian; Agilent: VEE Pro; Microsoft: Win 2000, Visual Basic, Visual C++, .NET; ODBC, ADO, SQL; Access, SQL Server, FoxPro, MySQL;
18	Innotool AG, Rothenhausen	8	Medizinaltechnik: Konstruktionsunterstützung und Beratung für Orthopädieimplantate und Instrumente, Herstellung von orthopädischen Implantaten und Instrumenten Präzisionsmechanik: Herstellung von präzisen Maschinenbauteilen in Klein- und Mittelserien Stanzwerkzeugbau: Konstruktion und Herstellung von Stanzwerkzeugen sowie kundenspezifischen hydraulischen Stanzmaschinen	Beherrschung der Schlüsseltechnologien für die Herstellung von medizinaltechnischen Produkten: Fräsen und Drehen von hochkomplexen Geometrien in hoher Genauigkeit und perfektem Finish aus rostfreien Materialien. Schnelligkeit und Flexibilität in Konstruktion und Fertigung.
19	Knobel Technik AG, Eschlikon	10	Wir konstruieren und fertigen seit über 10 Jahren Hydraulik-Komponenten: Patentierte Drehzylinder (Rotations-Zylinder), Linearzylinder, Mehrkolben-Zylinder, Ventil- und Steuerblöcke. Ferner entwickeln und fertigen wir Komponenten und Maschinen. Zulieferer für den Maschinen- und	Engineering, Produktions- und Operationsmanagement, CAM-CNC Fabrikation, Hydraulikkomponenten, bis hin zu kompletten Baugruppen und Maschinen.
	IPG AG	10	Wir entwickeln ganzheitliche Informatik-Innovationslösungen mit hohem Anwendungs-, Zuverlässigkeits- und Sicherheits-Niveau. Erarbeiten von projektorientierten und schlüsselfertigen Lösungen für den weltweiten Einsatz. Wir verfügen über Mitarbeiter mit betriebswirtschaftlichem und technologischen know how, bestückt mit gesundem Menschenverstand und	Technologiemanagement: Steuerungs- und Regeltechnik, Sicherheits- und Antriebstechnik, Informations-Technologie, Elektronik. Projekt-Management: Risiko-Management, Einkaufs-Management, Vertrags-Management, Konfigurations-Management, Change-Management, Prozess-Management (V-Modell, Hermes,

VF partners clustering



VF partners competency visualization

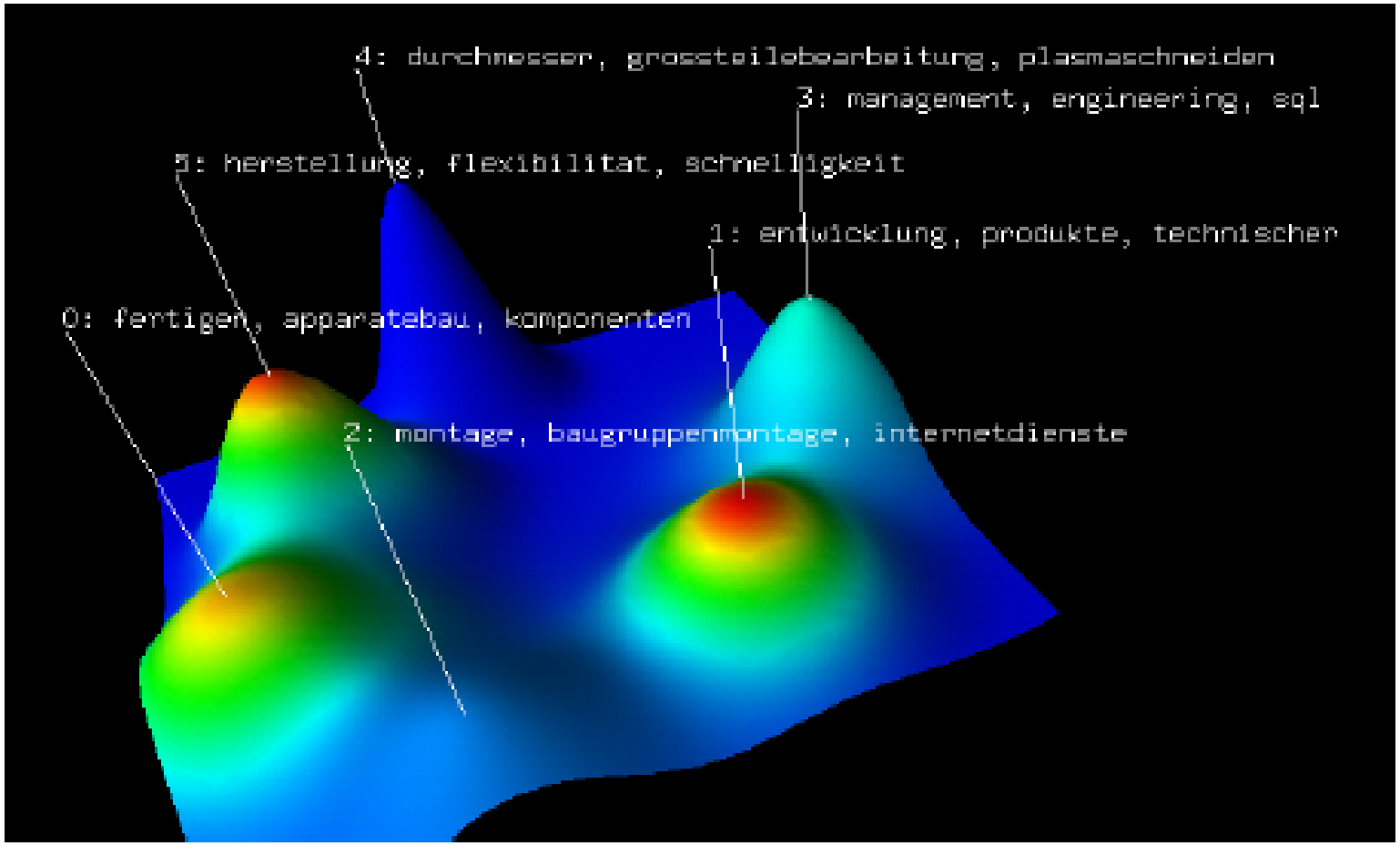
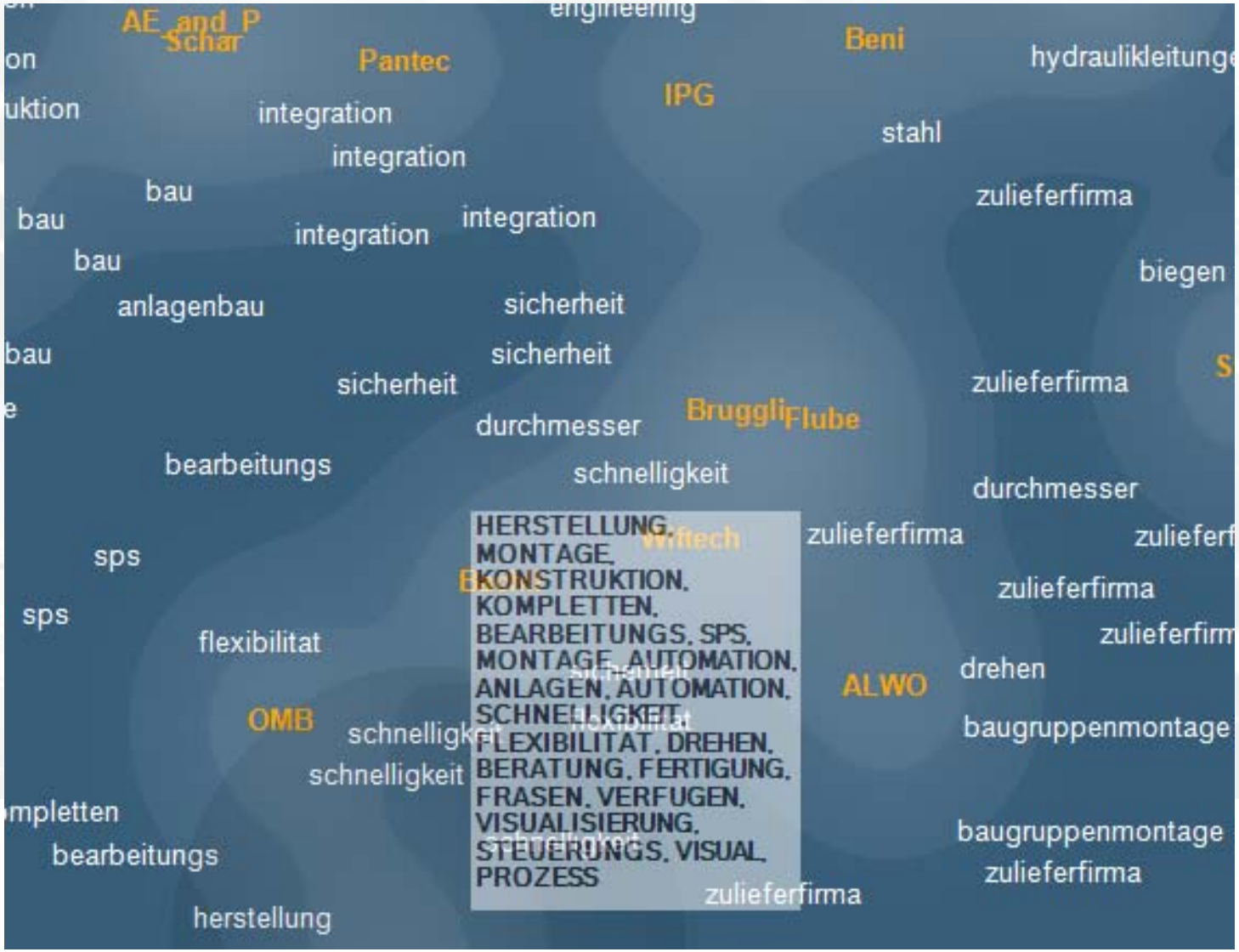


Figure 3 - Mountain visualization of six clusters, described with most descriptive words.

VF partners competency visualization



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Summary

- What is knowledge
- Traditional view of KM
- KM in the new economy: A networked organizations perspective
- Selected knowledge technologies for KM
- Using Web crawling and social network analysis for trust modeling and competency structuring through ontology construction