eResearch

A Max Planck Perspective

Kurt Mehlhorn

Max Planck Society

Chairman, Steering Committee sInfo

Overview

- The MPG
- eResearch, a Definition
- Max Planck Digital Library (MPDL): our central unit for eResearch infrastructure and tools
  - Information Provision
  - eSciDoc
    - Information Management (Pubman)
    - Research Tools (Scholarly Workbench)
- An example for eResearch
  - Knowledge Extraction from the Web
- Summary and Outlook

The MPG: A Characterization

- We are a research performing organization, not a funding agency
- Our mission: basic research of highest quality for the advance of science and for the benefit of humanity
- Our ambition: at the or near the top in the fields in which we engage in
- Structure: 78 research institutes
- Intensive cooperation with universities

The MPG: Some Numbers

- 78 institutes
  - 265 directors, 4000 PhDs, 4000 PhD students
- 30 locations
- Annual budget about 1.4 Billion Euro per year
- 18 Nobel prizes in the last 50 years

Locations

we are a distributed organization

cooperation between different locations is our daily life

locations outside Germany are not shown (Nijmegen, Florence, Rome, …)

Fields

- We work in a diverse set of fields
  - Chemistry, Physics, Technology
  - Biology, Medicine, Brain Science
  - Law, Art, History, Cognition
- not a broad as most universities
- but more interdisciplinary

Our Scientific Standing

Chemistry

Number of "Top Papers" published by Top-Ranking Institutions, between January 1995 and October 2005

Top-ranked institutions within research field	top papers	total papers	% top papers
HARVARD UNIV	246	2.550	9,65
MIT	237	3.443	6,88
UNIV CALIF BERKELEY	278	4.972	5,59
ETH ZURICH	123	4.529	2,72
MAX PLANCK SOCIETY	**276**	**11.242**	**2,46**
UNIV TOKYO	147	8.073	1,82
KYOTO UNIV	143	8.869	1,61
CNRS	97	7.432	1,31
CHINESE ACAD SCI	87	20.622	0,42
RUSSIAN ACAD SCI	36	29.942	0,12

Auswertung:Informationsvermittlungsstelle / Information Retrieval Services for the institutes of the Bio.-Med. Section of the Max-Planck-Society http://www.biochem.mpg.de/iv/
Quelle: „ISI - Essential Science Indicators" (http://www.db-hosts.mpg.de/WoS/)

Infrastructure, Tools, Instruments

Our research infrastructure

must match our research ambitions

Instrument (tool) building is an essential

part of doing science

Infrastructure, Tools, Instruments

- Buildings and Laboratories
- Libraries and Access to Information
- Communication Infrastructure
- Telescopes, electron microscopes, computing power, sequencers, semi-conductor lab, …

- Development of new instruments is an essential part of science: electron microscope, patch-clamp technique, frequency comb, search engines, …
- It can be much much more than service

eResearch =

Use of information technology

for enhancing research

New infrastructure, new tools, new instruments

my personal definition

Max Planck Digital Library

- **Max Planck Digital Library** (MPDL) is our central unit for eResearch infrastructure and tools (and instruments)

- The axes:
  1. Information provision (journals and data bases)
  2. Information dissemination and open access
  3. Research tools

- considerable eResearch activities in the institutes (virtual observatory, intelligent search engines, computational XXX, machine learning)

MPDL and eSciDoc

- **Max Planck Digital Library** (MPDL) is our central unit for eResearch infrastructure and tools (and instruments)

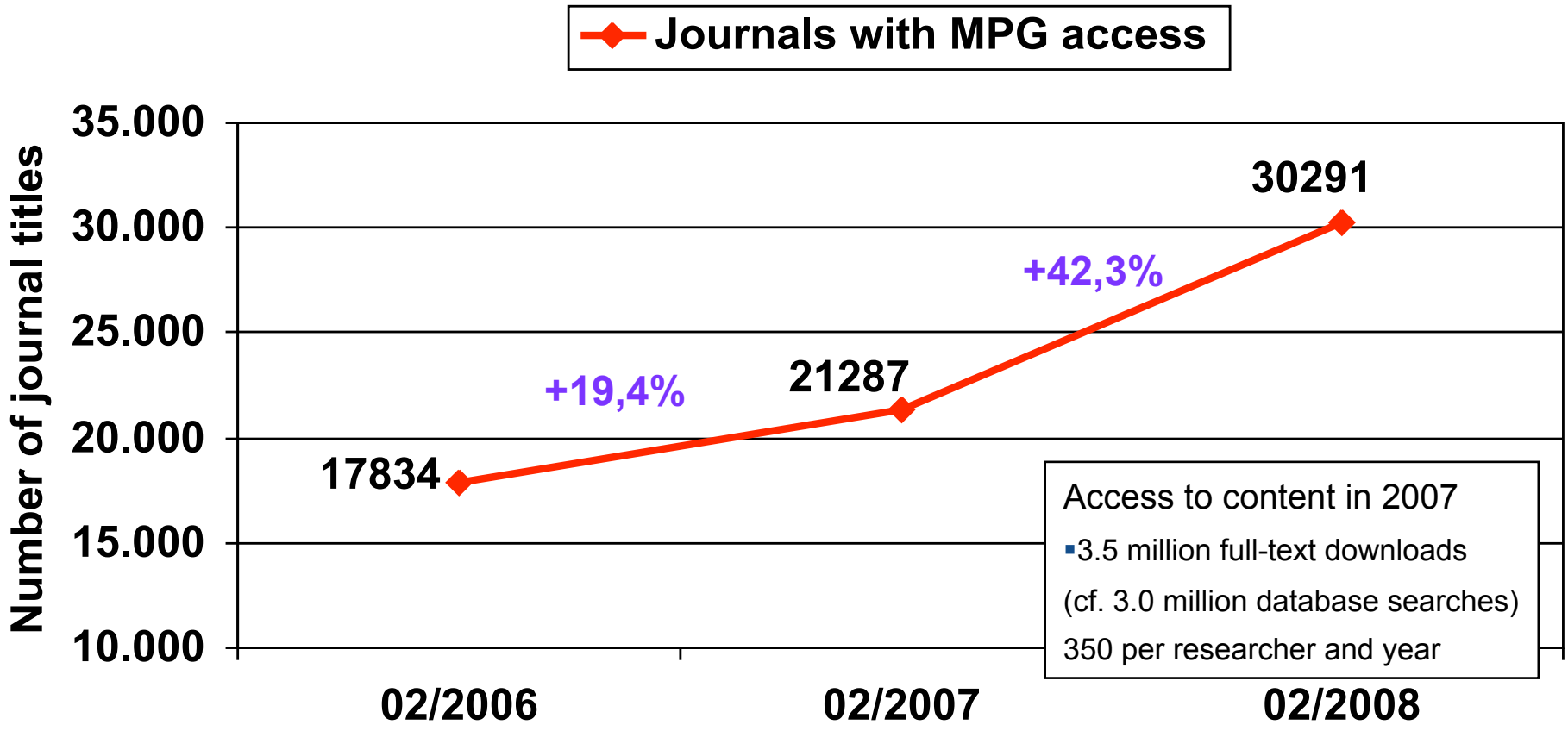
- The axes:
  1. Information provision (journals and data bases)
  2. Information dissemination and open access
  3. Research tools
- **eSciDoc** is our main project for
  - Axes 2: Pubman and
  - Axes 3: Scholarly Workbench

- considerable eResearch activities in the institutes (virtual observatory, intelligent search engines, …)

MPDL

- Founded in 2006
- Head: Laurent Romary
- Departments
  - Information provision (Ralf Schimmer)
  - eSciDoc (Malte Dreyer)
  - Open Access (Christoph Bruch)
- Complemented by activities in many institutes
- Close cooperation MPDL --- institutes
- Link to eResearch activities outside MPG
  - FIZ, Göttingen, Humboldt, DFG, DANS, NIMS, …

Information Provision: Journal coverage



Very large electronic journal collection – also on international scale

Vast growth rate; even accelerated since MPDL foundation

Open Access

- Free Access to Publicly Financed Scientific Results
  - Scientists want to be read (and cited)
  - Science will advance faster
  - The society has the moral right to freely accessing the science it already paid for
- MPS supports open access in many ways (see next slide)
  - but we do not fight a religious war

Support for Open Access

- Political action, e.g., Berlin declaration
- Framework agreements with
  - Open access publishers, e.g., Copernicus, New Journal of Physics, Biomed Central, PLoS, …
  - Traditional publishers, e.g., Springer
- Repositories (institute-level, MPS-level)
- Advice to our scientists about copyright agreements and how to change them
- Deposite mandate (under preparation)

Tools and Instruments

- Projects at MPDL
  - eSciDoc (Pubman and Scholarly Workbench)
- Cooperation projects
  - MPDL + institutes
  - MPDL and outside partners
- Research projects in institutes

eSciDoc Project

- Partners
  - FIZ Karlsruhe (eSciDoc infrastructure)
  - MPG (eSciDoc solutions)
- Funded by BMBF, Nixdorf Foundation, and internal sources
- Key persons
  - FIZ: Mathias Razum and Leni Helmes
  - MPG: Malte Dreyer and Laurent Romary
- Intended impact
  - Strategic project for FIZ and MPG
  - Impact beyond our own organizations
  - Open source and community model

Pubman

- The repository solution
- Functionalities and user interfaces for the submission of publication data of multiple types and versions, such as article, conference-paper, poster, report, book, pictures, videos, primary data etc., along with the metadata needed for proper retrieval and long-term archiving.
- Advantage for our scientists: quality and completeness of data, export to local and institute home pages, versioning and persistence, export to search engines, long-term archiving
- Advantage for MPG: preservation of scientific output, open access, good scientific conduct

What hooked me

- Since 2006, my publication list is generated on the fly from the data in the institute's repository, complete, up-to-date, correct, with links to full-text in repository and at publisher

D1 MPI-INF Publications, generated: 10:58, 7 June 2008

371. Lutz Kettner, Kurt Mehlhorn, Sylvain Pion, Stefan Schirra, Chee Yap

Classroom Examples of Robustness Problems in Geometric Computations

Computational Geometry: Theory and Applications , 2008

370. Telikepalli Kavitha and Kurt Mehlhorn

Algorithms to compute minimum cycle basis in directed graphs

Theory of Computing Systems 40 (4): 485-505, 2007

Scholarly Workbench

- eSciDoc (Scholarly Workbench) is a framework for targeted eResearch solutions

- Targeted solutions for institutes (joint projects)
  - WALS Online (language description)
  - Faces (Images)
  - VIRR (primary textual sources)

Example: FACES

- Collaboration with MPI Bildungforschung
  - Experiments on the recognition of emotions
  - Collection of annotated photographs, see next slide
  - FACES will be the basis for future experiments
- Solution was built fast and with small effort
  - 2,5 FTEs over 3 months
  - Reuse of generic eSciDoc framework
- Strategy
  - Towards a generic image management solution

    E.g. Spectrographic images, Astronomical images

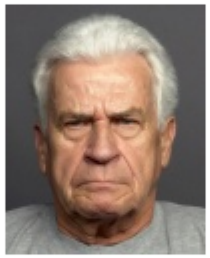
Welcome to the FACES Collection for MPI for Human Development. Version 0.9.

Not Logged in users can only view the picture sets of six persons (72 pictures). If you want to apply for an account, please fill out the application. Logged in users can see the picture sets of 171 persons (2052 pictures).

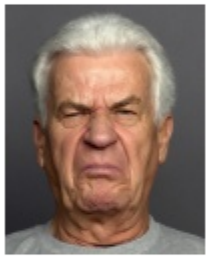
Show 12 hits of 72 |< < 1 2 3 4 5 > >| Number of pages : 6

Sort by identifier Order is ascending

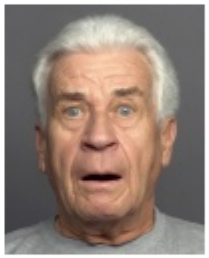

View Details


View Details


View Details


View Details


View Details


View Details


View Details


View Details


View Details


View Details


View Details


View Details

Show 12 hits of 72 |< < 1 2 3 4 5 > >| Number of pages : 6

Cooperation MPDL -- Institutes

- eSciDoc (Scholarly Workbench) forms basis
- Targeted solutions for institutes (joint projects)
  - WALS Online (language description)
  - Faces (Images)
  - VIRR (primary textual sources)

- **Competence centers**
  - **Identification of potential "MPDL subsidiaries »**
  - **MPIPL Nijmegen (Description and archival of multimedia information)**
  - **MPIWG Berlin (Cultural heritage data, XML processes)**

Harvesting Knowledge from the WEB

Gerhard Weikum

MPI Informatik

In collaboration with Giorgiana Ifrim, Gjergji Kasneci, Josiane Parreira, Maya Ramanath, Ralf Schenkel, Fabian Suchanek, Martin Theobald

Gerhard's Goals

Opportunity: Web could be comprehensive knowledge base

Challenge: seize opportunity and turn vision into reality

Approach: combine and exploit synergies of

- hand-crafted, high-quality knowledge sources
- automatic knowledge extraction
- social networks and human computing

Why Google and Wikipedia Are Not Enough

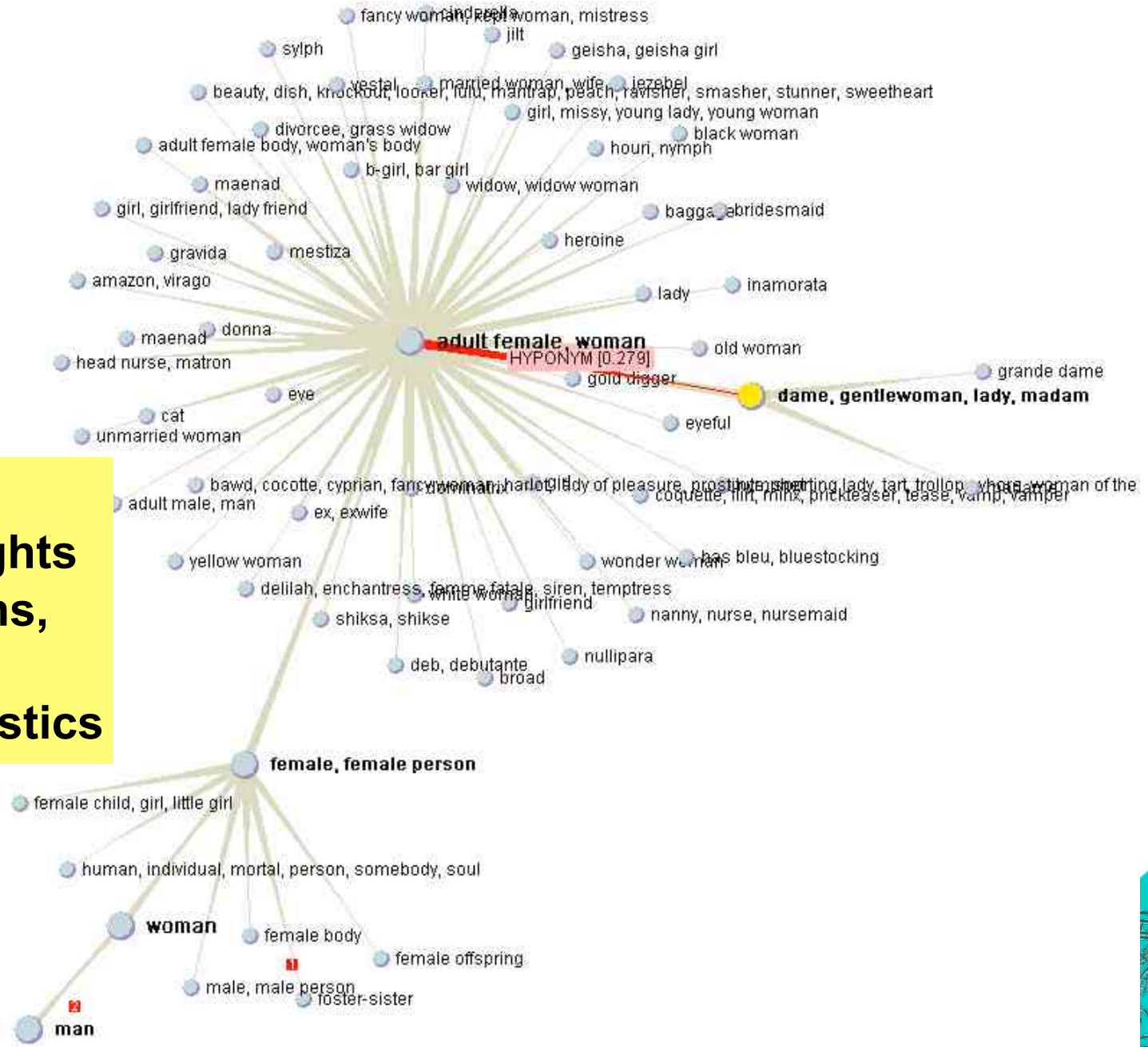
- neutron stars with Xray bursts > 10^{40} erg s^{-1} & black holes in 10“
- archaeological sites with both Roman and Celtic female clothes
- differences in Rembetiko music from Greece and from Turkey
- connection between Thomas Mann and Goethe
- Nobel laureate who survived both world wars and all his children
- drama with three women making a prophecy to a British nobleman that he will become king

The information to answer these questions is available in the web.

We must find ways to harvest it.

High-Quality Knowledge Sources

General-purpose **thesauri** and concept networks: **WordNet** family



cast into graph enhanced with weights for relation strengths, derived from co-occurrence statistics

High-Quality Knowledge Sources

Wikipedia and other lexical sources

List of Nobel laureates

From Wikipedia, the free encyclopedia

Max Karl Ernst Ludwig Planck (April 23, 1858 – October 4, 1947 in Göttingen, Germany) was a German physicist. He is considered to be the founder of quantum theory, and therefore one of the most important physicists of the twentieth century.

Contents [hide]

Life and work [edit]

Early Childhood [edit]

Planck came from a traditional, intellectual family. His paternal great-grandfather and grandfather were both theology professors in Göttingen, his father was a law professor in Kiel and Munich, and his paternal uncle was a judge.

Planck was born in Kiel to Johann Julius Wilhelm Planck and his second wife, Emma Patzig. He was the sixth child in the family, though two of his siblings were from his father's first marriage. Among his earliest memories was the marching of Prussian and

Max Planck

Max Karl Ernst Ludwig Planck

Born	April 23, 1858 Kiel, Germany
Died	October 4, 1947 Göttingen, Germany
Residence	Germany
Nationality	German
Field	Physicist
Institutions	University of Kiel Humboldt-Universität zu Berlin Georg-August-Universität Göttingen
Alma mater	Ludwig-Maximilians-Universität München
Academic advisor	Philipp von Jolly

From Wikipedia to YAGO

YAGO: Yet Another Great Ontology

[Suchanek/Kasneci/Weikum: WWW 2007]

- Turn Wikipedia into explicit knowledge base (semantic database)
- Exploit hand-crafted categories and templates
- Represent facts as explicit knowledge triples:

  relation (entity1, entity2)

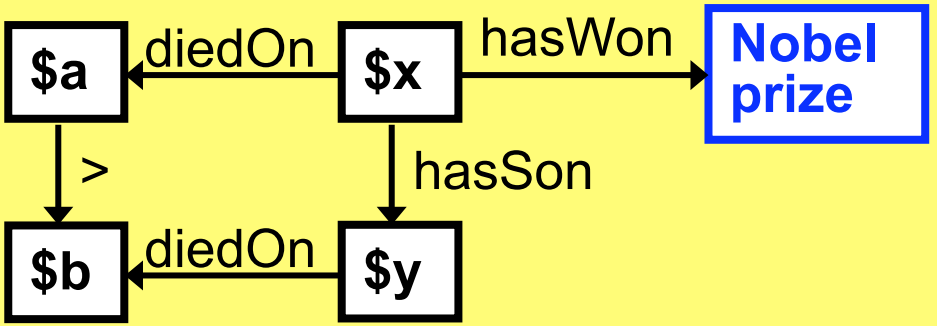
(in FOL, compatible with RDF, OWL-lite, XML, etc.)

Examples:

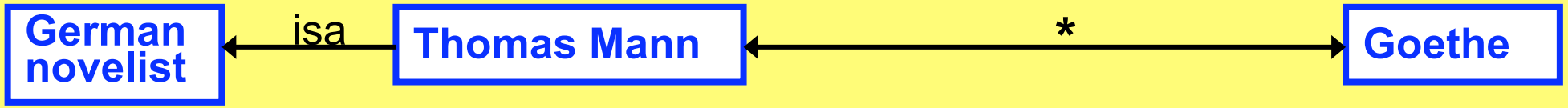
NAGA: Graph Search with Ranking

Graph-based search on YAGO-style knowledge bases
with built-in **ranking** based on **confidence** and **informativeness**

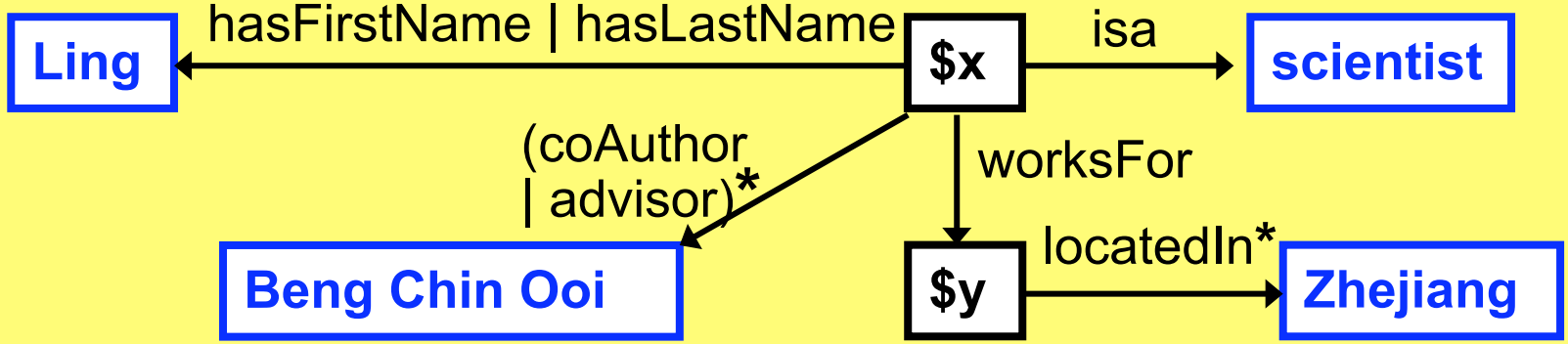
discovery queries



connectedness queries



queries with regular expressions



Information Extraction (IE): Text to Records

Max Planck

Max Karl Ernst Ludwig Planck (April 23, 1858 – October 4, 1947) was a German physicist who is considered to be the inventor of quantum theory.

Born in Kiel, Planck started his physics studies at Munich University in 1874, graduating in 1879 in Berlin. He returned to München in 1880 to teach at the university, and moved to Kiel in 1885. There he married Marie Merck in 1886. In 1889, he moved to Berlin, where from 1892 on he held the chair of theoretical physics.

In 1899, he discovered a new fundamental constant, which is named Planck's constant, and is, for example, used to calculate the energy of a photon. Also that year, he de... own set of units of measurement based on fundamental physical constants. One year later, he discovered the law of heat radiation, which is named Planck's Law of Radiation. This law became the basis of quantum theory, which emerg... later in cooperation with Albert Einstein and Niels Bohr.

Person	BirthDate	BirthPlace	...
Max Planck	4/23, 1858	Kiel	
Albert Einstein	3/14, 1879	Ulm	
Mahatma Gandhi	10/2, 1869	Porbandar	

Person	ScientificResult
Max Planck	Quantum Theory

Constant	Value	Dimension
Planck's constant	6.226×10^{23}	Js

Person	Collaborator
Max Planck	Albert Einstein
Max Planck	Niels Bohr

Person	
Max Planck	KWG / MPG

combine NLP, pattern matching, lexicons, statistical learning

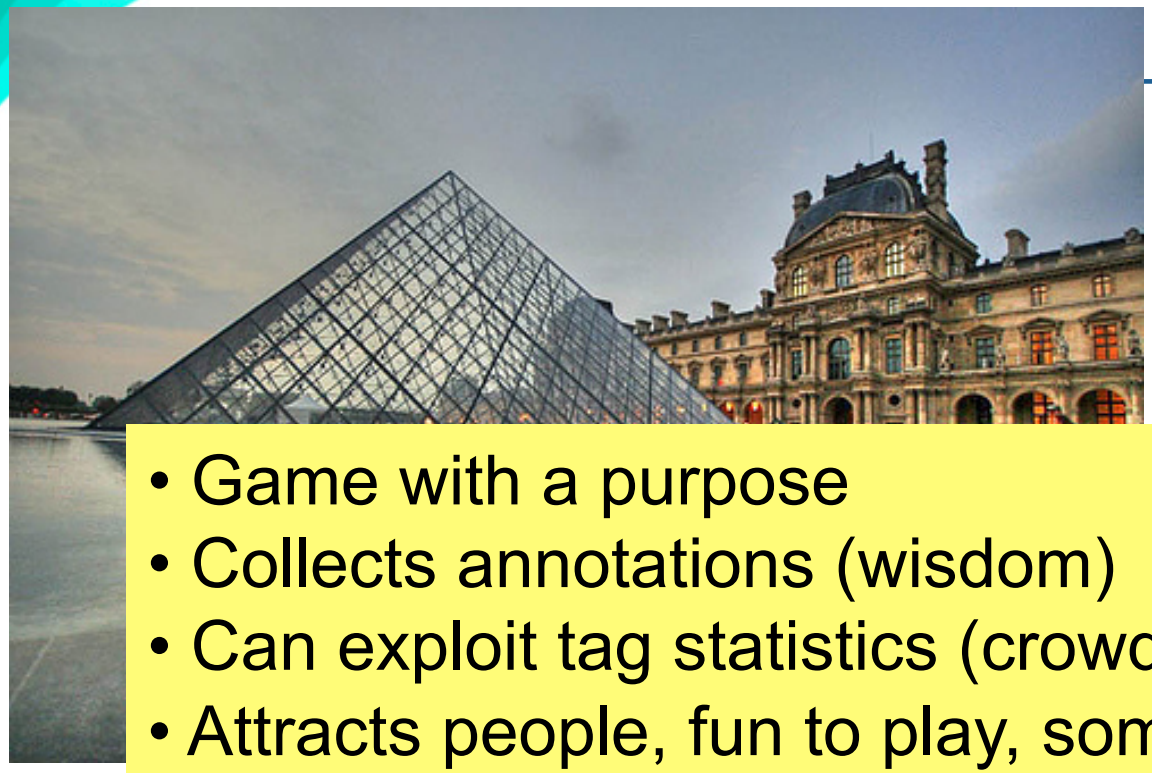
„Wisdom of Crowds“ at Work on Web 2.0

Information enrichment & knowledge extraction **by humans**:

- **Collaborative Recommendations**:
  - Google Page Rank
  - Amazon (product ratings & reviews, recommended products)
- **Social Tagging and Folksonomies**
  - del.icio.us: Web bookmarks and tags
  - flickr: photo annotation, categorization, rating
  - YouTube: same for video
- **Human Computing in Game Form**
  - ESP and Google Image Labeler: image tagging
  - Peekaboom: image segmenting and tagging
  - Verbosity: facts from natural-language sentences
- **Community Portals**
  - dblife.cs.wisc.edu for database research
  - www.lt-world.org for language technology

ESP Game [Luis von Ahn et al. 2004]

played against random, anonymous partner on Internet



taboo:
pyramid
Louvre
museum
Paris

- Game with a purpose
- Collects annotations (wisdom)
- Can exploit tag statistics (crowds)
- Attracts people, fun to play, some play hours
- ESP game collected > 10 Mio. tags from > 20000 users
- 5000 people could tag all photos on the Web in 4 weeks (human computing)

Congratulations!
You scored 1 point!

Mitterand
Mona Lisa
metro lignes 7, 14
Da Vinci code

Summary of Social Approach

Social tagging and social networks (Web 2.0) are potentially valuable knowledge sources

Games (human computing) are an interesting way of enticing „knowledge input“ and collecting statistics

Spectral analysis is a highly versatile tool for **rating & ranking** that can be extended and scaled by **decentralized** algorithms

Challenges:

- Design a game that intrigues serious scientists to „semantically“ annotate their scholarly work
- Develop an analysis method that identifies the „best“ facts, resilient to egoistic and malicious behaviors (incl. coalitions)

3 Tenets and 4 Challenges

T1: need both common knowledge and domain knowledge, and integration

T2: text (and speech) will remain major-value source

T3: hand-crafted ontologies, info extraction, social tagging go a long way, but no single approach will suffice

C1: Methods for automatically (and continuously) linking, matching, integrating ontologies & high-quality sources

C2: Scalable and robust IE methods for knowledge harvesting, with precision/recall tuning & minimum human supervision

C3: Intriguing, scalable, robust methods for social wisdom

C4: Combining the three methodological pillars with synergies

Summary and Outlook

- We have experienced substantial change over the last decade and this is going to continue
  - Recall that the first browser came in ’95
- Information technology has and will affect science how science is conducted
- Research organizations and universities must manage this change
  - The MPG is taking up this challenge