eResearch

A Max Planck Perspective

Kurt Mehlhorn
Max-Planck-Institut für Informatik
Interim Head, Max Planck Digital Library
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
  (= TUM + LMU, < Harvard)
- 18 Nobel prizes in the last 50 years
we are a distributed organization

coopertion between locations is our daily life

locations outside Germany (Nijmegen, Florence, Rome, Florida)
We work in a diverse set of fields

- Chemistry, Physics, Technology
- Biology, Medicine, Brain Science
- Humanities, Social Science, Law, History, Art, Psychology

- not as 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

<table>
<thead>
<tr>
<th>Top-ranked institutions within research field</th>
<th>top papers</th>
<th>total papers</th>
<th>% top papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>HARVARD UNIV</td>
<td>246</td>
<td>2.550</td>
<td>9.65</td>
</tr>
<tr>
<td>MIT</td>
<td>237</td>
<td>3.443</td>
<td>6.88</td>
</tr>
<tr>
<td>UNIV CALIF BERKELEY</td>
<td>278</td>
<td>4.972</td>
<td>5.59</td>
</tr>
<tr>
<td>ETH ZURICH</td>
<td>123</td>
<td>4.529</td>
<td>2.72</td>
</tr>
<tr>
<td>MAX PLANCK SOCIETY</td>
<td>276</td>
<td>11.242</td>
<td>2.46</td>
</tr>
<tr>
<td>UNIV TOKYO</td>
<td>147</td>
<td>8.073</td>
<td>1.82</td>
</tr>
<tr>
<td>KYOTO UNIV</td>
<td>143</td>
<td>8.869</td>
<td>1.61</td>
</tr>
<tr>
<td>CNRS</td>
<td>97</td>
<td>7.432</td>
<td>1.31</td>
</tr>
<tr>
<td>CHINESE ACAD SCI</td>
<td>87</td>
<td>20.622</td>
<td>0.42</td>
</tr>
<tr>
<td>RUSSIAN ACAD SCI</td>
<td>36</td>
<td>29.942</td>
<td>0.12</td>
</tr>
</tbody>
</table>


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
- Telescopes, electron microscopes, DNA sequencers, semi-conductor lab, ...
- Computing Power, Communication Infrastructure

- Development of new instruments is an essential part of science: electron microscope, patch-clamp technique, frequency comb, ...
- It is much much more than service
eResearch =

Use of information technology for enhancing research

my personal definition
Max Planck Digital Library

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

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, bioinformatics, computational XXX, machine learning)

Other relevant central units: Garching Computing Center, German Climate Computing Center
MPDL

- Founded in 2006
- Interim Heads: Kurt Mehlhorn, Jürgen Renn, Bernard Schutz
- Departments
  - Information Provision (Ralf Schimmer)
  - eResearch Tools (Malte Dreyer)
- Complemented by activities in many institutes
- Close cooperation with MPDL --- institutes
- Link to eResearch activities outside MPG
  - FIZ, SUB Göttingen, Humboldt, DFG, DANS, NIMS, ...
Very large electronic journal collection – also on international scale

Vast growth rate; even accelerated since MPDL foundation

Access to content in 2007
- 3.5 million full-text downloads
  (cf. 3.0 million database searches)
- 350 per researcher and year

Journals with MPG access

Number of journal titles

- 17834
- 21287 (+19.4%)
- 30291 (+42.3%)

02/2006 02/2007 02/2008
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
  - Side effect: might reduce dependency on publishing houses

- MPS supports open access in many ways (see next slide)
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
- Intended impact
  - Strategic project for FIZ and MPG
  - Impact beyond our own organizations (Leibniz, MPG)
  - 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
- Roll-Out has started (with enthusiastic feedback)
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.

D1 MPI-INF Publications, generated 10:33, 29 September 2009

381. Kurt Mehlhorn and Michael Sagraloff, A Deterministic Descartes Algorithm for Real Polynomials, ISSAC 2009


.............

2. Kurt Mehlhorn, The 'almost all' theory of subrecursive degrees is decidable
   In: Automata, languages and programming : 2nd colloquium (ICALP-74), Saarbrücken, Germany, September, 22-23, 1974, 317-325, [PDF: Download: Mehlhorn_1974_d_m.pdf]

1. Kurt Mehlhorn, On the Size of Sets of Computable Functions
   In: 14th Annual Symposium on Switching & Automata Theory (SSAT-73), Iowa City, Iowa, USA, 1973, 190-196, [PDF: Download: Mehlhorn_1973_a_m.pdf]
Scholarly Workbench

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

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

- Collaboration with MPI Bildungsforschung
  - 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
- A step towards a generic image management solution
  - NIMS (Japan) uses it for images of diamond cuts
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).
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 X-ray 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 I

General-purpose thesauri and concept networks: WordNet family

cast into graph enhanced with weights for relation strengths, derived from co-occurrence statistics
Wikipedia and other lexical sources

List of Nobel laureates

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

1 Life and work
  1.1 Early Childhood
  1.2 Education
  1.3 Academic career
  1.4 Family
  1.5 Professor at Berlin University
  1.6 Black-body radiation
  1.7 Einstein and the Theory of Relativity
  1.8 World War and Weimar Republic
  1.9 Quantum mechanics
  1.10 Nazi dictatorship and Second World War
2 Honours and medals
3 See also
4 Publications
5 Bibliography
6 External links
  6.1 Biographies
  6.2 Articles
7 Notes

Life and work

Early Childhood

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
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:

Max_Planck bornIn Kiel
Kiel isInstanceOf City
Graph-based search on YAGO-style knowledge bases with built-in ranking based on confidence and informativeness.

**Discovery queries**
- Kiel bornIn $x$ isa scientist

**Connectedness queries**
- German novelist isa Thomas Mann

**Queries with regular expressions**
- Ling hasFirstName | hasLastName
- Beng Chin Ooi
- scientist
- worksFor
- locatedIn Zhejiang
**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 developed his 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 emerged later in cooperation with Albert Einstein and Niels Bohr.

<table>
<thead>
<tr>
<th>Person</th>
<th>BirthDate</th>
<th>BirthPlace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Planck</td>
<td>4/23, 1858</td>
<td>Kiel</td>
</tr>
<tr>
<td>Albert Einstein</td>
<td>3/14, 1879</td>
<td>Ulm</td>
</tr>
<tr>
<td>Mahatma Gandhi</td>
<td>10/2, 1869</td>
<td>Porbandar</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Person</th>
<th>ScientificResult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Planck</td>
<td>Quantum Theory</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Constant</th>
<th>Value</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planck’s constant</td>
<td>$6.226 \times 10^{23}$ Js</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Person</th>
<th>Collaborator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Planck</td>
<td>Albert Einstein</td>
</tr>
<tr>
<td>Max Planck</td>
<td>Niels Bohr</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Person</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Planck</td>
<td>KWG / MPG</td>
</tr>
</tbody>
</table>

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
Louvremuseum
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!
4 Challenges

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: Scalable and 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 how science is conducted
- Research organizations and universities must manage this change
  - The MPG is taking up this challenge
  - Improved cooperation MPDL – Computing Centers – Institute IST groups
  - Improved exchange: MPDL – research activities in institutes