Vorlesung

E-Science, Wissenschaftsinformation, Szientometrie

Hans Uszkoreit & Brigitte Jörg

Hans Uszkoreit Vorlesung
Informationswissenschaft
und Informationssysteme

E-Science – Term

The term was created by John Taylor, the Director General of the United Kingdom's Office of Science and Technology in 1999

- **to describe a large funding initiative starting in November 2000**
- **examples of the kind of science include social simulations, particle physics, earth sciences and bio-informatics**
- **Particle physics has a particularly well developed e-Science infrastructure due to their need for adequate computing facilities for the analysis of results and storage of data originating from the CERN Large Hadron Collider**

(Wikipedia, July 2009)

Requires appropriate Infrastructure

E-Science – Definition

e-science is about inventing and exploiting new advanced computational methods to:

- **create a new approach to shared research between groups and facilities**
- **generate, curate and analyze data**
- **link publications to data**
- **develop and explore models and simulations at an unprecedented scale and to use simulations to run experiments**
- **help the set-up of distributed virtual organizations to ease collaboration and sharing of resources and information and the remote operation of facilities**

**(John Wood, ESFRI Chair 2007
at the euroCRIS Seminar, 2007)**

E-Science – Definition

e-Science – the invention and application of computer enabled methods to achieve new, better, faster or more efficient research, innovation, decision support or diagnosis in any discipline.

It draws on advances in computing science, computation and digital communications.

(e-Infrastructure Reflection Group Whitepaper 2008)

E-Science – e-Infrastructure

e-Science is

strongy related to

and dependent on

a working infrastructure

E-Science in the UK

- the UK e-Science Programme began in 2001 as a coordinated initiative involving all the Research Councils and the then Department of Trade and Industry

- the e-Science Core Programme
  - supported the development of generic technologies
  - such as the software known as middleware
  - to enable very different resources to work together seamlessly
  - across networks and create computing grids.

- Each Research Council has funded its own e-Science activities to develop techniques and demonstrate their use across a broad range of research and applications
  - Arts & Humanities Research Council (AHRC)
  - Biotechnology & Biological Sciences Research Council (BBSRC)
  - Engineering & Physical Sciences Research Council (EPSRC)
  - Economic & Social Research Council (ESRC)
  - Medical Research Council (MRC)
  - Natural Environment Research Council (NERC)
  - Science & Technology Facilities Council (STFC)

Source: http://www.rcuk.ac.uk/escience/

The early adopters: HEP

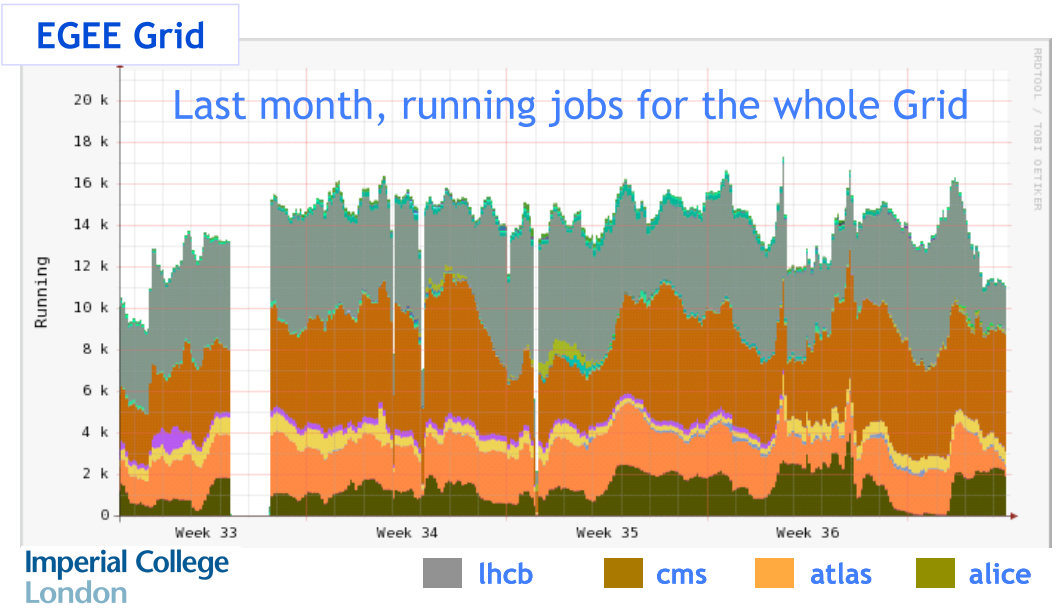
The **H**igh **E**nergy **P**hysics was the first research community to adopt globally the grid paradigm for data collection and analysis

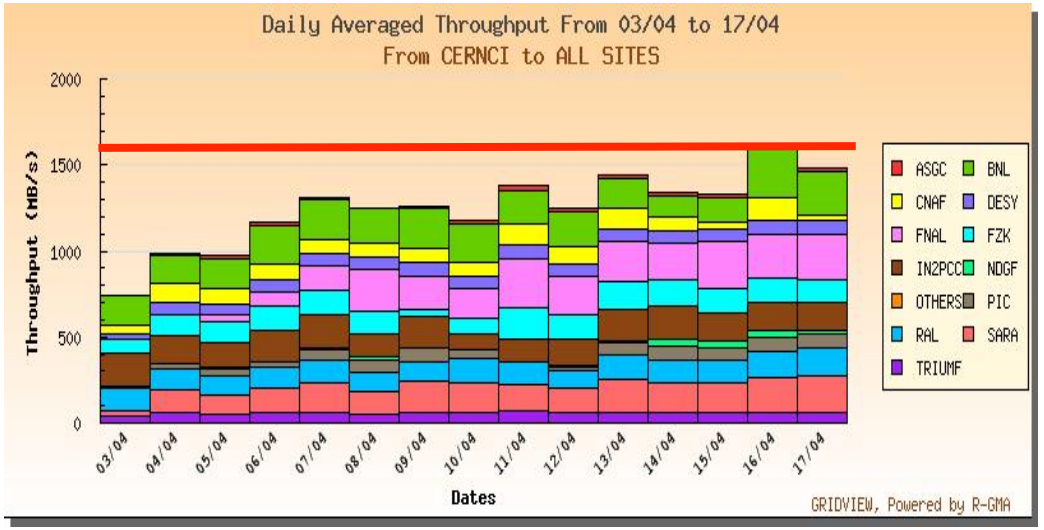
- High Energy Physics adopted grids for LHC to handle the unprecedented volume of data produced
- Highly structured community acting as “Guinea pig”
- High Energy Physics is the n°1 user of e-infrastructures around the world

Achievements in High Energy Physics

EGEE example
(Enabling Grids for E-sciencE)

- ~50K jobs/day
- > 10K *simultaneous* jobs during prolonged periods
- Reliable data distribution service demonstrated at
- 1.6 GB/sec from CERN to LHC Computing Grid national nodes





(by John Wood, ESFRI Chair 2007 at the euroCRIS Seminar, 2007)

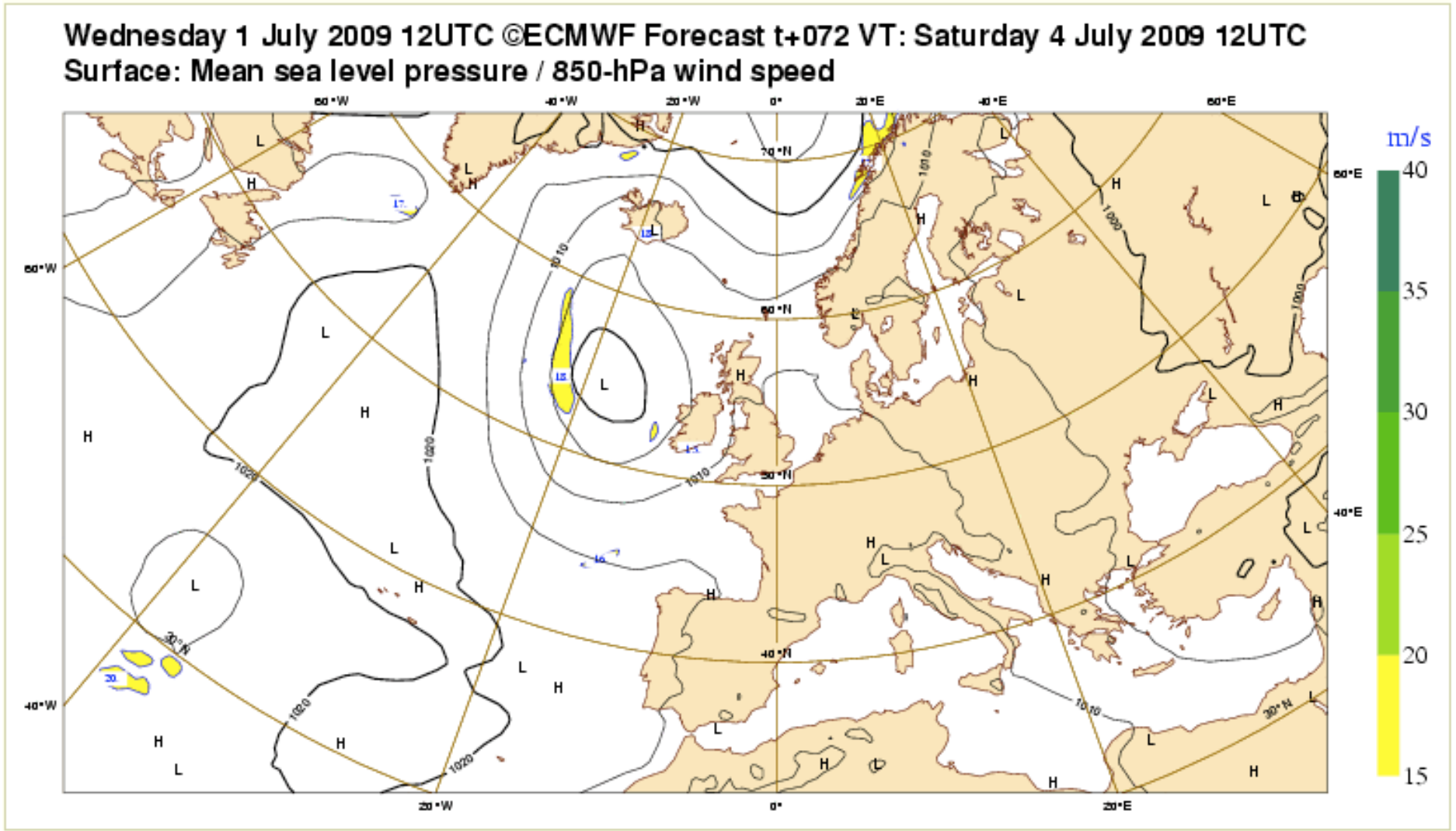
e-infrastructure supporting the physical sciences

Mean sea level pressure, wind speed at 850 hPa and geopotential 500 hPa

Step (-> valid time) | Forecast base time

72 (Sat 4 Jul 2009 12UTC) | Wed 1 Jul 2009 12UTC

Europe



Source: http://www.ecmwf.com/

GÉANT: global reach

Who are the users today?

- Research communities in urgent need for new advanced methods because they face unprecedented computational challenges
  -> **Example HEP**: LHC, Neutrino Mass, Gravitational Waves

- Research communities foreseeing the need for new advanced computational methods because of new major projects
  -> **Example:** fusion (ITER)

- Other research communities - a holistic approach
  - Geophysics
  - Condensed Matter
  - Meteorology
  - Energy
  - ..
  - Library – Digital Repository Community (**DRIVER**)
  - Language Technology (**CLARIN**)

(e)-Science is a global activity

eScience in Deutschland

D-Grid 1: 2005 – 2008

IT-Dienste für die Wissenschaftler, entwickelt und implementiert von erfahrenen Grid-Forschern und Anwendern. Sogenannte Grid-Communities testen diese globale Dienste-Infrastruktur inzwischen mit ihren rechen- und daten-intensiven Anwendungen aus den Gebieten der Hochenergiephysik, Astrophysik, alternative Energien, Medizin, Klimaforschung, Ingenieuranwendungen und Geisteswissenschaften.

D-Grid 2: 2007-2010

IT-Dienste für Wissenschaft und Industrie, die auf der D-Grid-Integrationsschicht aufbauen, wie zum Beispiel Bauindustrie, Finanzwirtschaft, Automobilindustrie, Luft- und Raumfahrt, Betriebsinformations- und Betriebsmittel-Systeme und geographische Datenverarbeitung.

Geplant sind weitere Schritte zur Erweiterung der D-Grid-Infrastruktur mit der Einführung professioneller Betriebskonzepte, Service-Level-Agreements für die Verhandlungen zwischen Nutzern und Betreibern von Ressourcen, einer Wissensschicht, dem Aufbau von virtuellen Kompetenzzentren, die Anbindung service-orientierter Architekturen der Industrie, und die Bereitstellung von Grid-Ressourcen zum Nutzen der ganzen Gesellschaft.

Source: http://www.d-grid.de/, gefördert vom BMBF.

eScience in Deutschland

- Die Rechner der Supercomputerzentren weisen unterschiedliche Architekturen auf, die recht spezifisch auf unterschiedliche Anwendungsschwerpunkte abgestimmt sind.
- In Deutschland stehen auf diese Weise Rechenkapazitäten für eine sehr breite Palette von Forschungsarbeiten in verschiedensten Disziplinen zur Verfügung.
- Für die internationale Wettbewerbsfähigkeit der Wissenschaft in Deutschland kommt es zunehmend auf eine **strategische Allianz der unterschiedlich ausgeprägten Höchstleistungsrechner** an.

Source: http://www.bmbf.de/de/298.php

eScience in Deutschland

- DFN (Deutsches Forschungsnetz)
- Jüngster Meilenstein: 26. Mai 2009 am FZ, Jülich
  -> schnellster Supercomputer Europas

"But in third place, a new contender has emerged-- a new IBM BlueGene/P system called JUGENE and installed at the Forschungszentrum Juelich (FZJ) in Germany. It achieved 825.5 teraflop/s (trillions of floating point operations per second) on the Linpack benchmarks and has a theoretical peak performance of just above 1 petaflop/s. FZJ is also home to the new No. 10 system. Called JUROPA, it is built from Bull Novascale and Sun SunBlade x6048 servers and achieved 274.8 Tflop/s. […]"

Source: http://www.top500.org/lists/2009/06/press-release

E-Science in Europa

Europäische Forschungsnetz-Organisationen

- **DANTE:** Delivery of Advanced Network Technology to Europe
  not-for-profit, established in 1993 in Cambridge, UK (the location of Cambridge was chosen as a result of an international competition; although based in the UK, it is a truly European company.)
  - plan, build and operate pan-European research networks
  - provides **data communications infrastructure** essential to the development of the global research community

- **TERENA:** Trans-European Research and Education Networking Association
  - offers a forum to collaborate, innovate and share knowledge in order
  - to foster the development of Internet technology, infrastructure and services
  - to be used by the research and education community

Source: http://www.dfn.de/globale-kooperation/europa/
See also: http://www.e-irg.eu/

Some European Projects

CLARIN (http://www.clarin.eu/):

Common Language Resources and Technology Infrastructure

The CLARIN project is **a large-scale** pan-European collaborative effort to create, coordinate and make **language resources and technology available and readily usable**. CLARIN offers scholars the tools to allow computer-aided language processing, addressing one or more of the multiple roles language plays (i.e. carrier of cultural content and knowledge, instrument of communication, component of identity and object of study) in the Humanities and Social Sciences.

Some European Projects

DRIVER (http://www.driver-repository.eu/):

Digital Repository Infrastructure Vision for European Research

DRIVER is a multi-phase effort whose vision and primary objective is to create a cohesive, robust and flexible, pan-European infrastructure for digital repositories, offering sophisticated services and functionalities for researchers, administrators and the general public.

DRIVER has established a network of relevant experts and Open Access repositories. DRIVER-II will consolidate these efforts and transform the initial testbed into a fully functional, state-of-the art service, extending the network to a larger confederation of repositories. DRIVER is integral to the suite of electronic infrastructures that have emerged in the worldwide GÉANT network and is hence funded under the e-Infrastructures call of the European Commission's 7th framework programme.

European Strategy Forum on RI (ESFRI)

European Strategy Forum on Research Infrastructures

ESFRI, the European Strategy Forum on Research Infrastructures, is a strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach. The competitive and open access to high quality Research Infrastructures supports and benchmarks the quality of the activities of European scientists, and attracts the best researchers from around the world.

The mission of ESFRI is to support a coherent and strategy-led approach to policy-making on research infrastructures in Europe, and to facilitate multilateral initiatives leading to the better use and development of research infrastructures, at EU and international level.

For more projects see: http://cordis.europa.eu/esfri/

ESFRI Roadmap (**Updated 2008**): ftp://ftp.cordis.europa.eu/pub/esfri/docs/esfri_roadmap_2008_update_20090123.pdf

eScience in the US

The term e-Science refers to large scale science that is carried out through distributed global Collaborations enabled by the Internet. Typically, such collaborative scientific enterprises require

- access to very large data sets
- very large scale computing resources
- high performance visualization

e-Science is a digital infrastructure that allows scientists to conduct research in new ways. Common terminology related to e-Science include **cyberinfrastructure, grids, grid computing, distributed networks,** and high performance computing.

- Projects often involve collaboration between large teams
- developed and managed by research laboratories
- large universities, and governments

In the United States these are **primarily funded by the National Science Foundation** (NSF) and the Department of Energy's Office of Science, which support the nation's supercomputing centers. Bioinformatics, earth sciences, and high-energy physics are examples of scientific disciplines with significant e-Science projects.

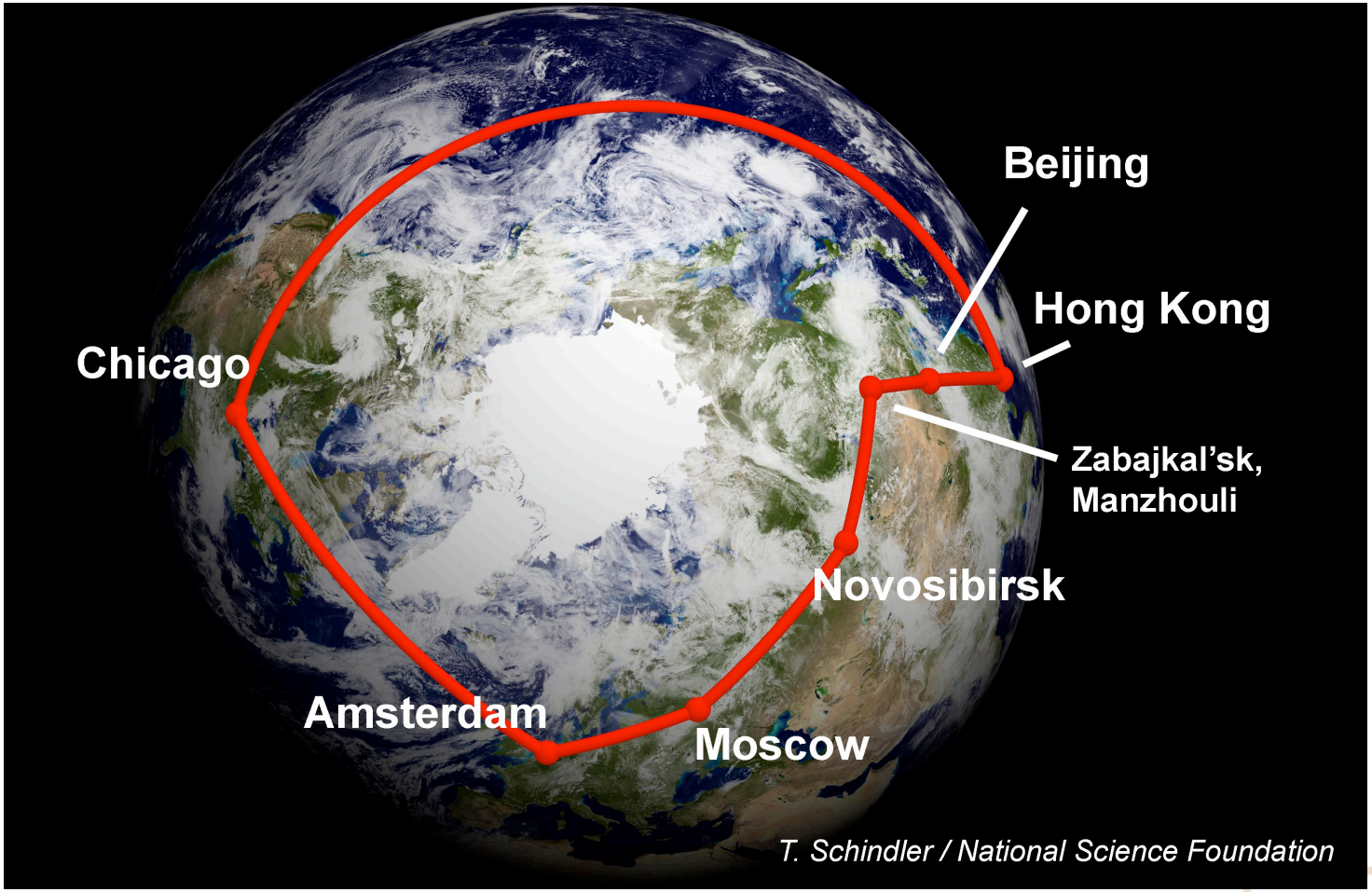
Source: http://www.loc.gov/rr/scitech/tracer-bullets/esciencetb.html

eScience in the US

The Teragrid is one of the most powerful computer infrastructures.

Photo: National Science Foundation, Office of Cyberinfrastructure.

eScience in the US



In December 2003, the United States, Russia and China announced the start of operations for the first round-the-world computer network ring.

Photo: National Science Foundation, Office of Cyberinfrastructure.

So, what is the new (e)Science ?

- **data deluge:** ICT Infrastructure vs. Wet-Labs
- no longer one technique
- problems facing society demand a multiple technique approach
- **interoperability** between equipment and data sets becomes **imperative**
- virtual research communities
- cross-disciplinarity
- improved scientific process … role of simulation
- from data to publications
- from research to education
- …

E-Science – Verwandte Themen

- **Results: Publications (Grey Literature), Patents, Products**
- **Data, Metadata, Information, Knowledge**
- **Classifications, Subject-Headings, Indexes**
- **Standards**
- Middleware
- Sustainability
- Procurement
- (Open) Access
- Authentication
- Security
- Technical / Policy-Trust
- Ethical Issues (Data Protection; Privacy Rights)
- Policy Decisions (locations, priorities,…)
- Funding Opportunities / Priorities
- …

Wissenschaft / Wissenschaftsinformation

Man weiß heute: Forschung und Entwicklung führt zu mehr Wohlstand und zu verbesserter Lebensqualität:

- Forschung benötigt Förderung
- Forschung ist transnational (global)
- Wissenschaftssinformation ist relevant
  - für die Akteure im wissenschaftlichen Umfeld
  - für Entscheidungsträger
  - für strategische Planungen
  - für die Medien
- Wissenschaftsinformation als Mittler auch zwischen Wissenschaft und Gesellschaft

Wissenschaftsinformation

- konzentriert sich auf die **Information** im wissenschaftlichen Umfeld

- nicht die Daten als solches sind im Vordergrund sondern eher die **Information**, die man aus den Daten zieht, schlußfolgert, …

Wissenschaftsinformation - Beispiele

Informationen über wissenschaftliche Ergebnisse, Aktivitäten und Akteure:

- **Ergebnisse: Publikationen** (Peer-Review, Open-Access, Grey Literature, Governmental, …)
- **Ergebnisse: Patente** (International, Europaweit, National, …)
- **Personen** (Forscher, Projektleiter, Preisträger, …)
- **Organisationen** (Förderorganisationen, Forschungszentren, Universitäten, Forschungsbereiche, virtuelle Zentren, Communities, Forschungsgruppen, “Invisible Colleges”, Hosts, Publikationsorgane, Sammelstellen, …)
- **Projekte** (International, Europaweit, National, Lokal, …)
- **Förderung** (NSF, DFG, BMBF, EC, ESF, RCUK, …)
- **Förderprogramme** (Esprit, FP5, FP6, FP7 …)
- **Einrichtungen** (LHC, Supercomputer…)
- **Ressourcen** (Corpora, Tools, Infrastruktur)
- **Dienste** …

- **und deren Beziehungen (Kontext) !!**

Wissenschaftsinformation warum ?

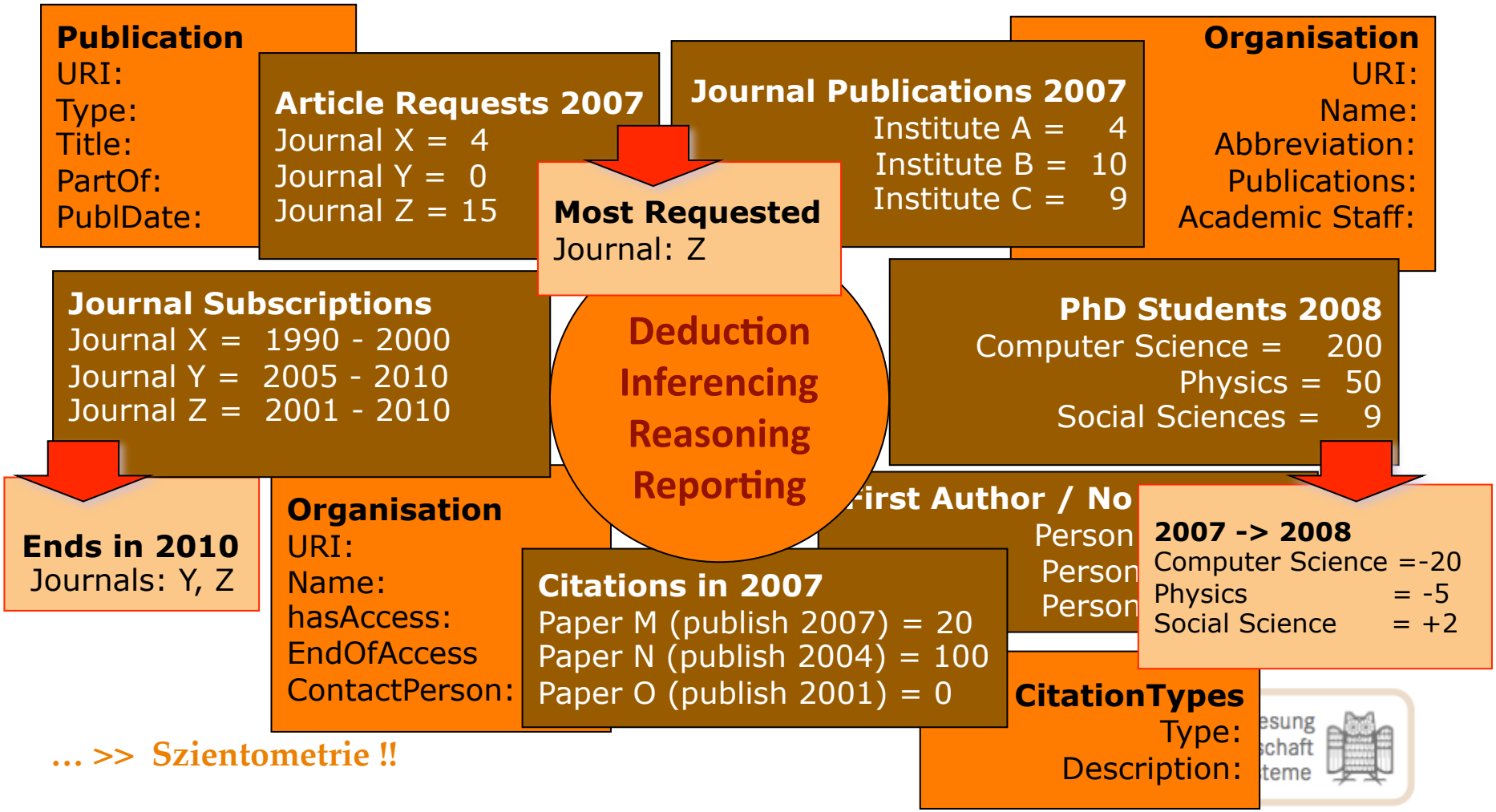
Welche Fragen wollen wir beantwortet haben

- How many articles has author X published in 2007 as a first author?
- How often have articles by author X been cited?
- Did author X publish with institutionally external authors?
- In how many FP7 projects does organisation Z participate?
- How many publications have resulted from project Y?
- How many people have been employed in the course of FP6 projects from the 1st call in the NMS?
- How many PhD students have participated in FP6 projects?
- How many women have been involved in FP6 projects?
- How often have articles from journal A been requested in 2007?
- How many articles have been published in the field of B?
- …

… >> Szientometrie !!

Wissenschaftsinformation

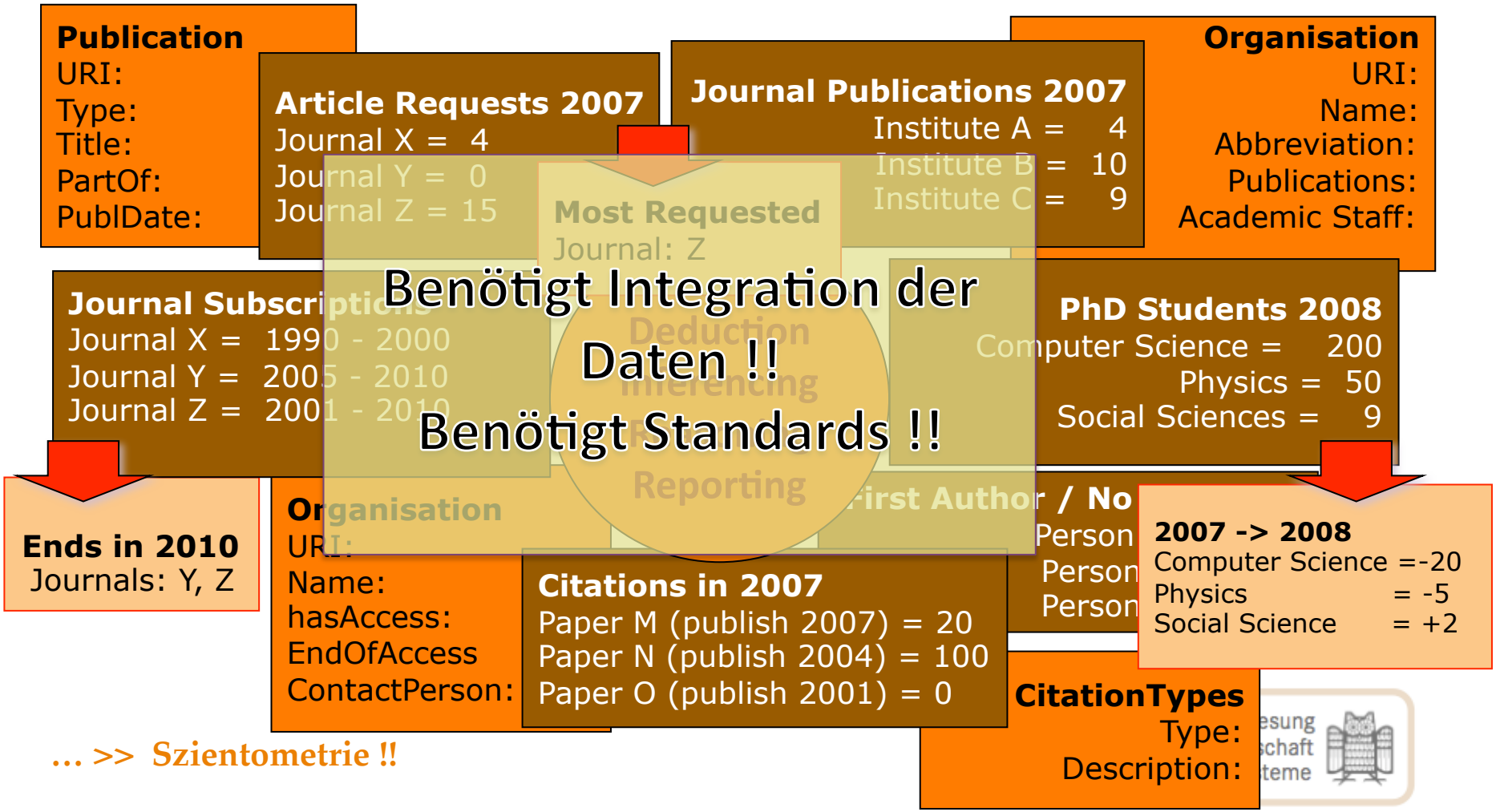
Welche Fragen wollen wir beantwortet haben



… >> Szientometrie !!

Wissenschaftsinformation

Welche Fragen wollen wir beantwortet haben



… >> Szientometrie !!

Wissenschaftsinformation

- **Entitäten**
  Person, Projekt, Organisation,
  Publikation, Patent, Ressource, …

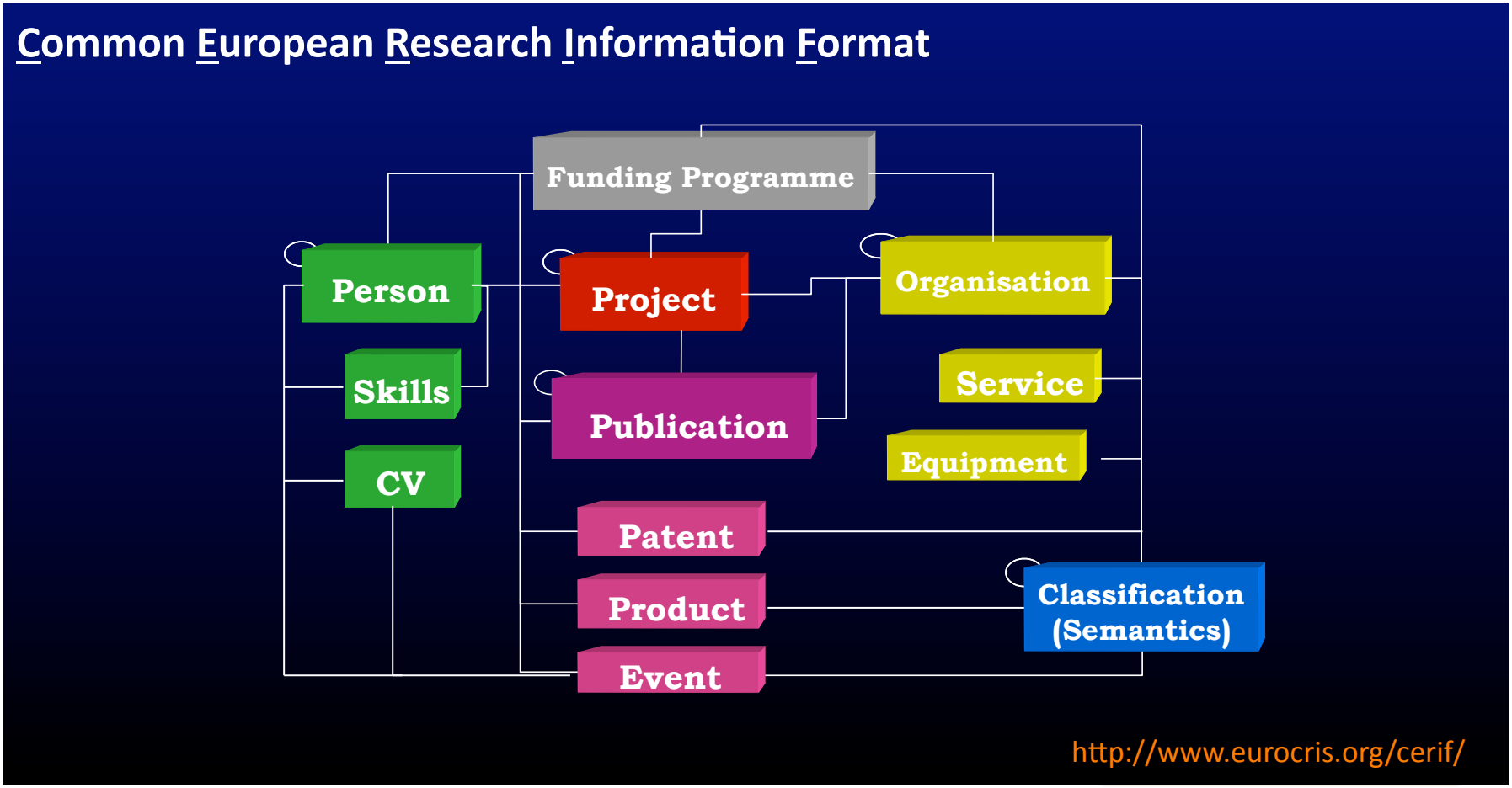
- **Attribute**
  Name, URI, Geschlecht, Startdatum,
  Enddatum, ISBN Nr., Publikationsdatum,
  Veröffentlichungsdatum, … Keywords

- **Relationen (Kontext)**
  Person-Publikation [Autor, Co-Autor, Editor, …]
  Person-Projekt [Investigator, Koordinator, Manager, …]
  Person-Organisation [Affilliierung, Role="CEO, ClO, CFO, …"]

- …

Wissenschaftsinformation

Ein europäisches Modell zur standardisierten Repräsentation von Wissenschaftsinformation: CERIF



Current Situation

Multiple proprietery Schemas / Formats / Applications / Services :

Publication Records:
- Dublin Core
- Marc Code
- Digital Item Declaration Language (DIDL)
- Metadata Object Description Schema (MODS)
- …

- DSpace
- Eprints
- Open Repositories
- …
- GoogleScholar
- CiteSeer
- …

For Person Records:
- FOAF

- LinkedIn
- Facebook,
- …

For Audio/Video Files:
- Metadata Encoding and Transmission Standard (METS)
- …

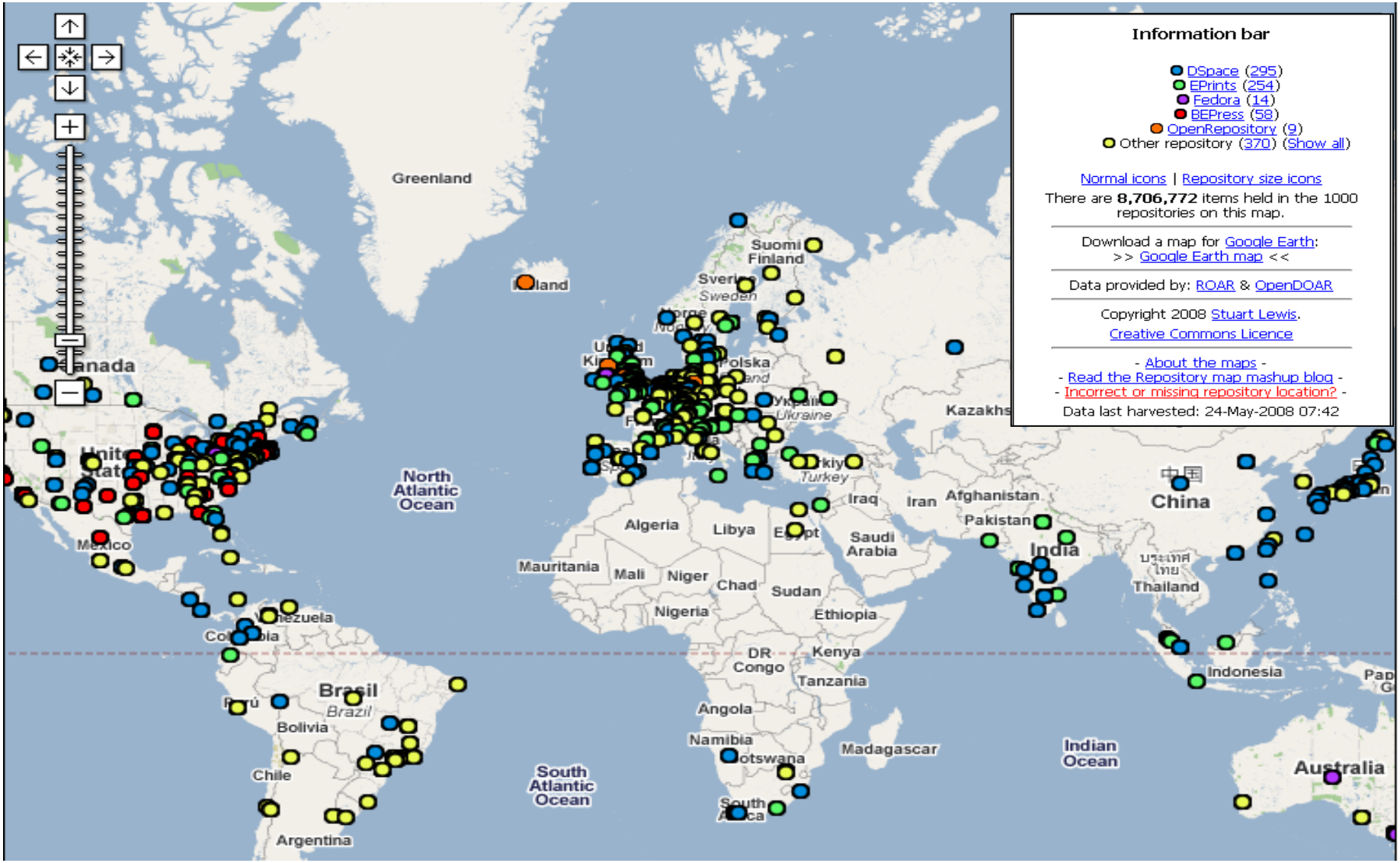
- MySpace
- YouTube
- …

Subject Headings:
- Ortelius Thesaurus
- MESH (Medical Subject Heading)
- …

- Wikipedia
- del.icio.us
- …

E-Science, Wissenschaftsinformation, Szientometrie

Landscape of Publication Repositories



Szientometrie

- enstanden aus der Bibliometrie, Informetrie
- Messung wissenschaftlicher Produktivität / Aktivität
- Analyse abhängig von verfügbaren Daten
- Ergebnisse sind von der Qualität der Daten abhängig
- Zur Messung: Integration verschiedener Ressourcen (Daten)

- …

- Szientometrische Berechnung von Ergebnissen auf Basis von vorhandenen Systemen (momentan hauptsächlich über ISI)

- …