TIB | German National Library of Science and Technology

Textual and non-textual objects: Seamless access for scientists

Uwe Rosemann

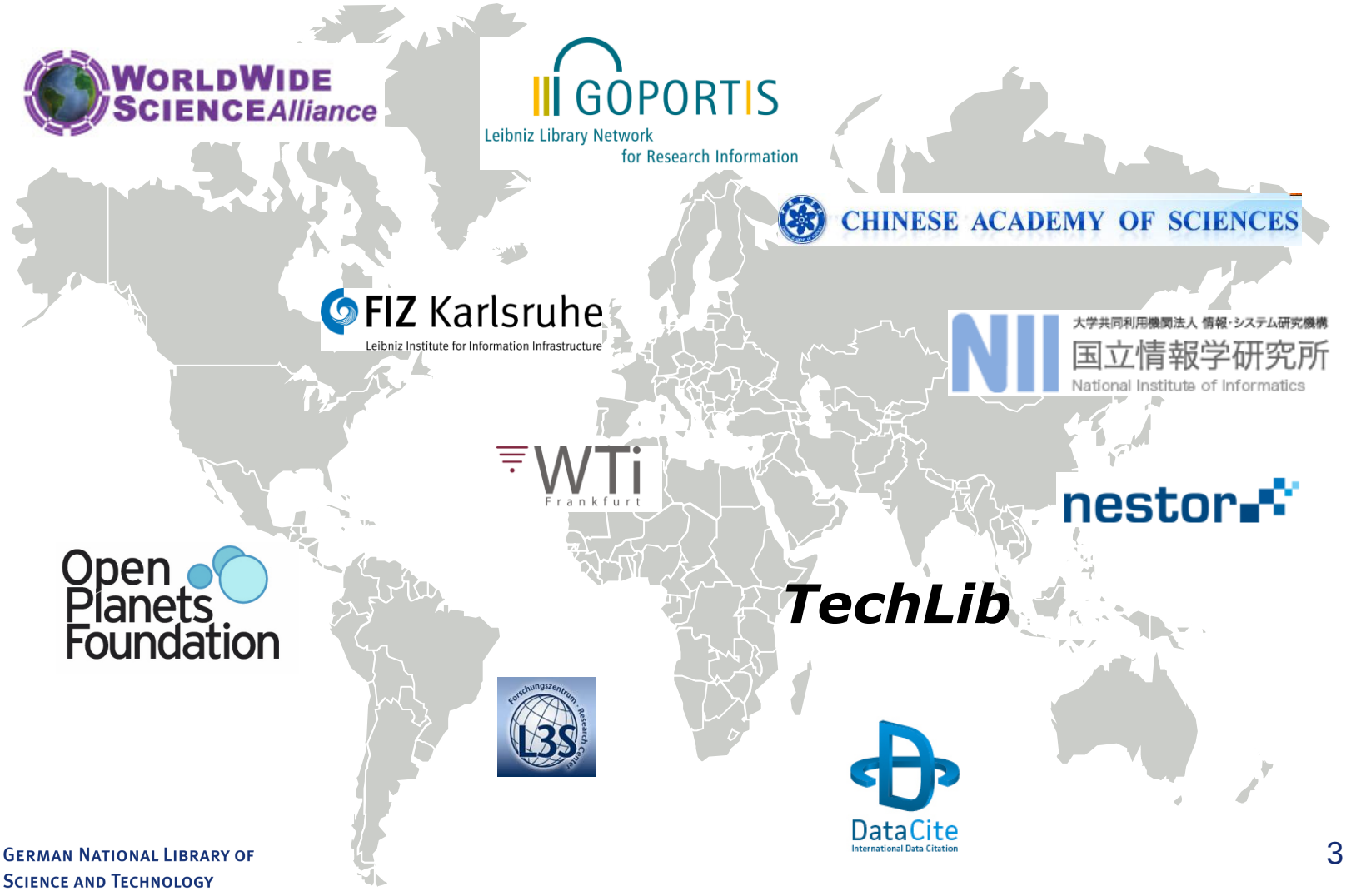
ICIC 2013

Vienna

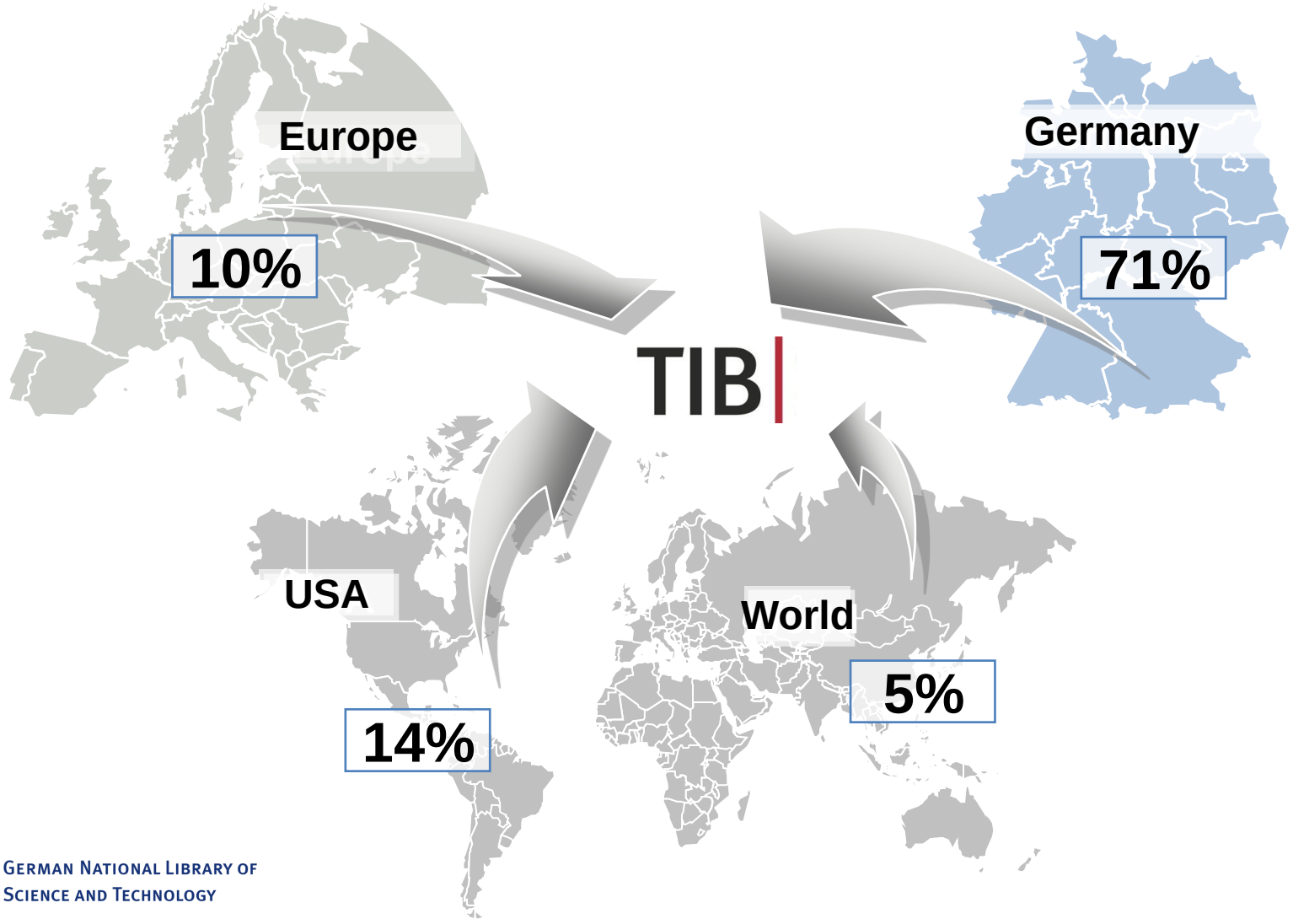
German National Library of Science and Technology (TIB)

- Specialized Library for Architecture, Chemistry, Computer Science, Mathematics, Physics, Engineering Technology
- Financed by Federal Government and all Federal States
- Member of the Leibniz Association
- Global supplier for scientific and technical information

Global Network



Customers



Main Services

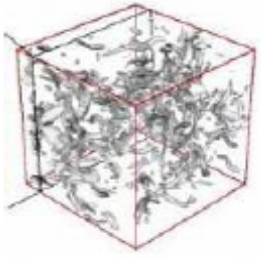
- Provision of scientific content
    - full texts, document delivery, interlibrary loan
- Scientific retrieval
    - portal GetInfo
- Long-term preservation
- DOI-Service for research data
- Research and development

Changes in the scientific process

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  unify theory, experiment, and simulation
  using data management and statistics
  - Data captured by instruments
    Or generated by simulator
  - Processed by software
  - Scientist analyzes database / files

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



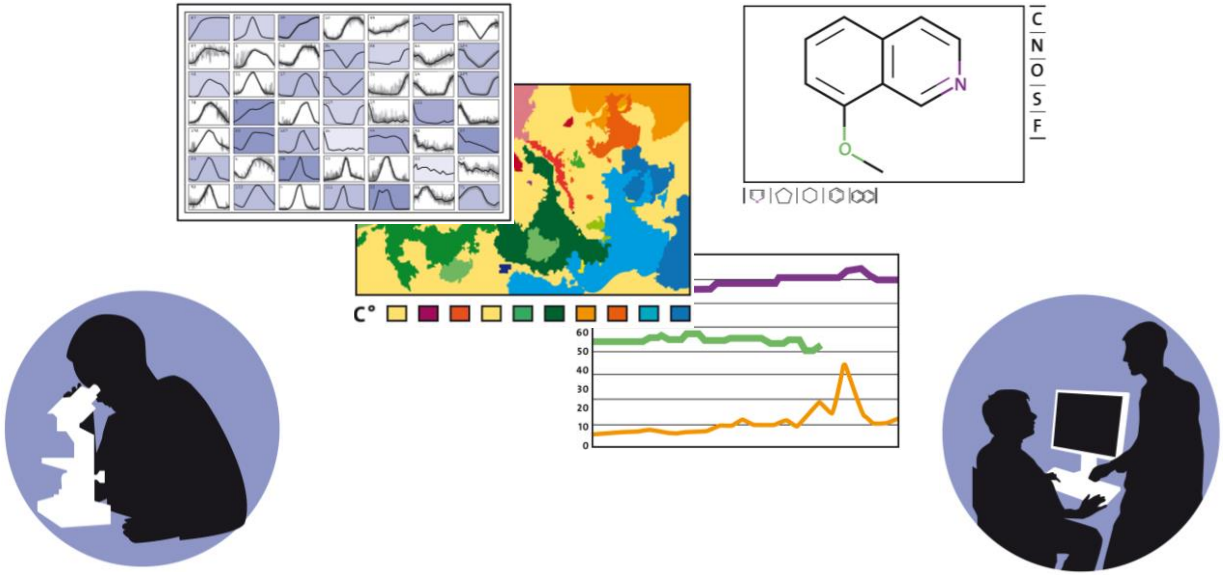
A gap

- A widening gap in the scientific record between published research in a text document and the data that underlies it
- As a result, datasets are
  - difficult to discover
  - difficult to access
- Scientific information gets lost

Requirements - Politics

Knowledge is power.

Europe must manage the digital assets its researchers generate.

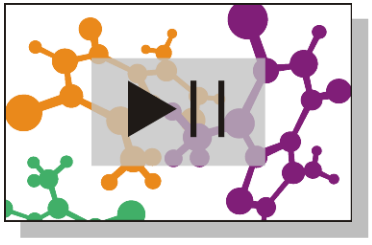


„Riding the wave“ – How Europe can gain access from the rising tide of scientific data

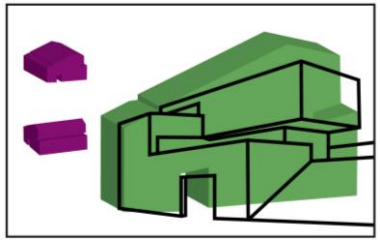
Final report of the High Level Expert Group on Scientific Data.

Strategy – Move beyond text

Software

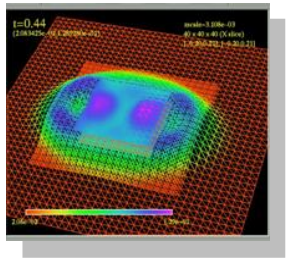


Scientific Films

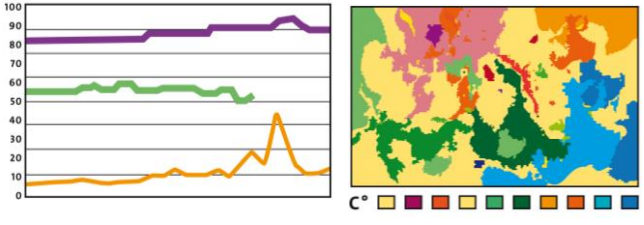


3D Objects

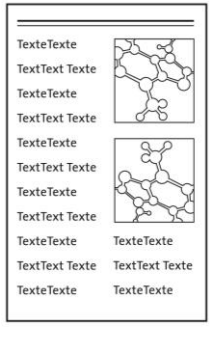
TIB



Simulation



Research Data



Text

Move beyond text – Consequences for TIB

- Research communities produce many types of scientific and technical information
- Each has its own unique characteristics and life cycle
- Must become capable of accepting and managing new media formats

Competence Center for Non-textual Materials I

- Develop a clear strategy for the use and integration of non-textual materials at the TIB
- Systematically collect non-textual materials from research and teaching
- Define, integrate and establish technical infrastructure
- Define and establish workflows for indexing, cataloguing, digital preservation, DOI names, licencing

Competence Center for Non-textual Materials II

- Develop innovative media-specific portals enabled by e.g. an automated video analysis with scene, speech, text and image recognition
- Linking non-textual materials to other research information such as full texts and research data via the specialist portal GetInfo
- Engage in communities, provide support and advice to media providers

→ **TIB will establish its own research capacity**

How have we been preparing ?

- ***Infrastructure for research data***
- Visual search tools for AV-media
- 3D-Objects
- chemOCR

Collaboration – Research Data

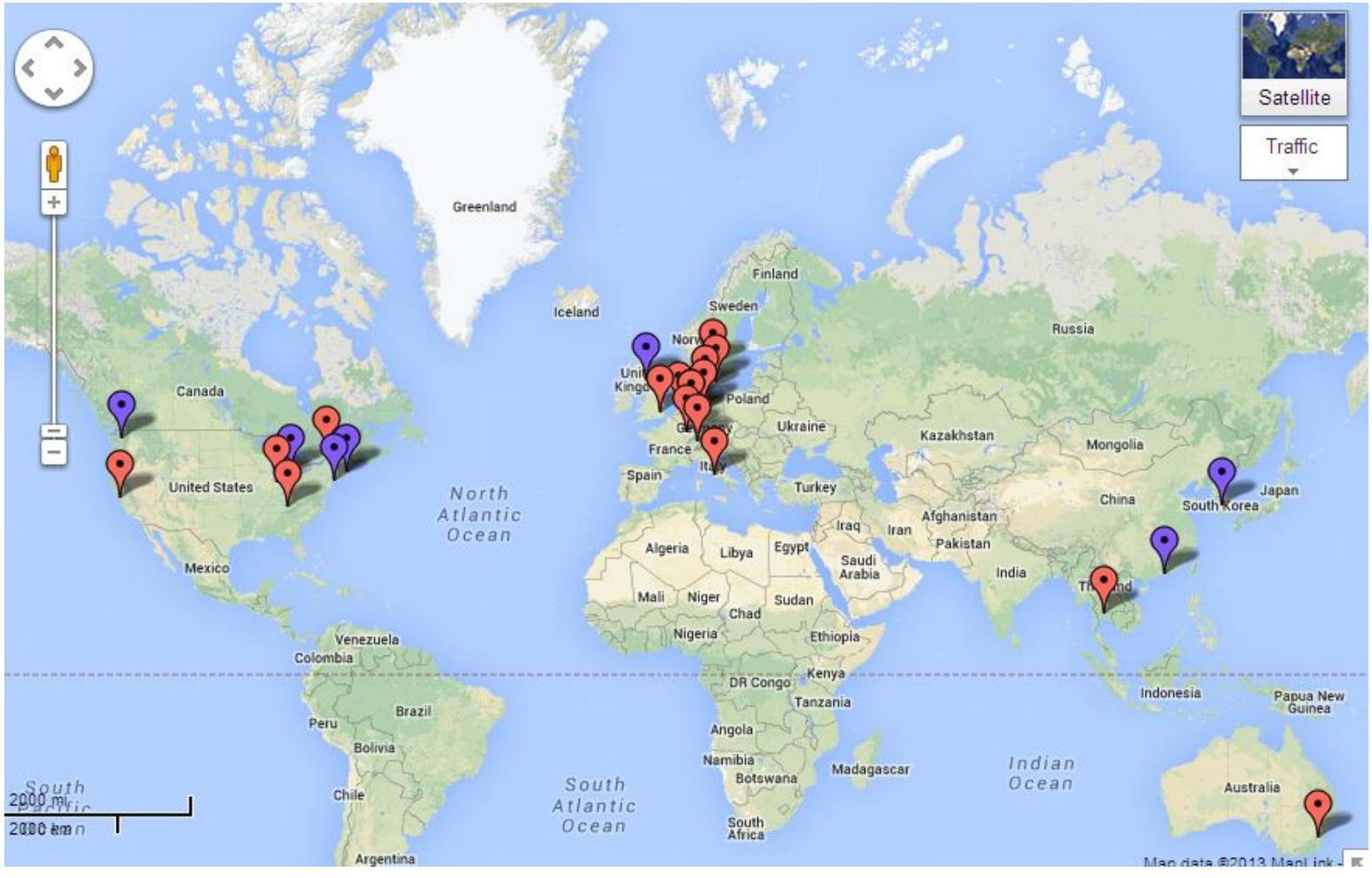
- In 2005, the TIB became a non-commercial DOI registration agency for research data

- In 2010, the TIB became co-founder of the international DataCite consortium to establish easier access to scientific research data on the Internet

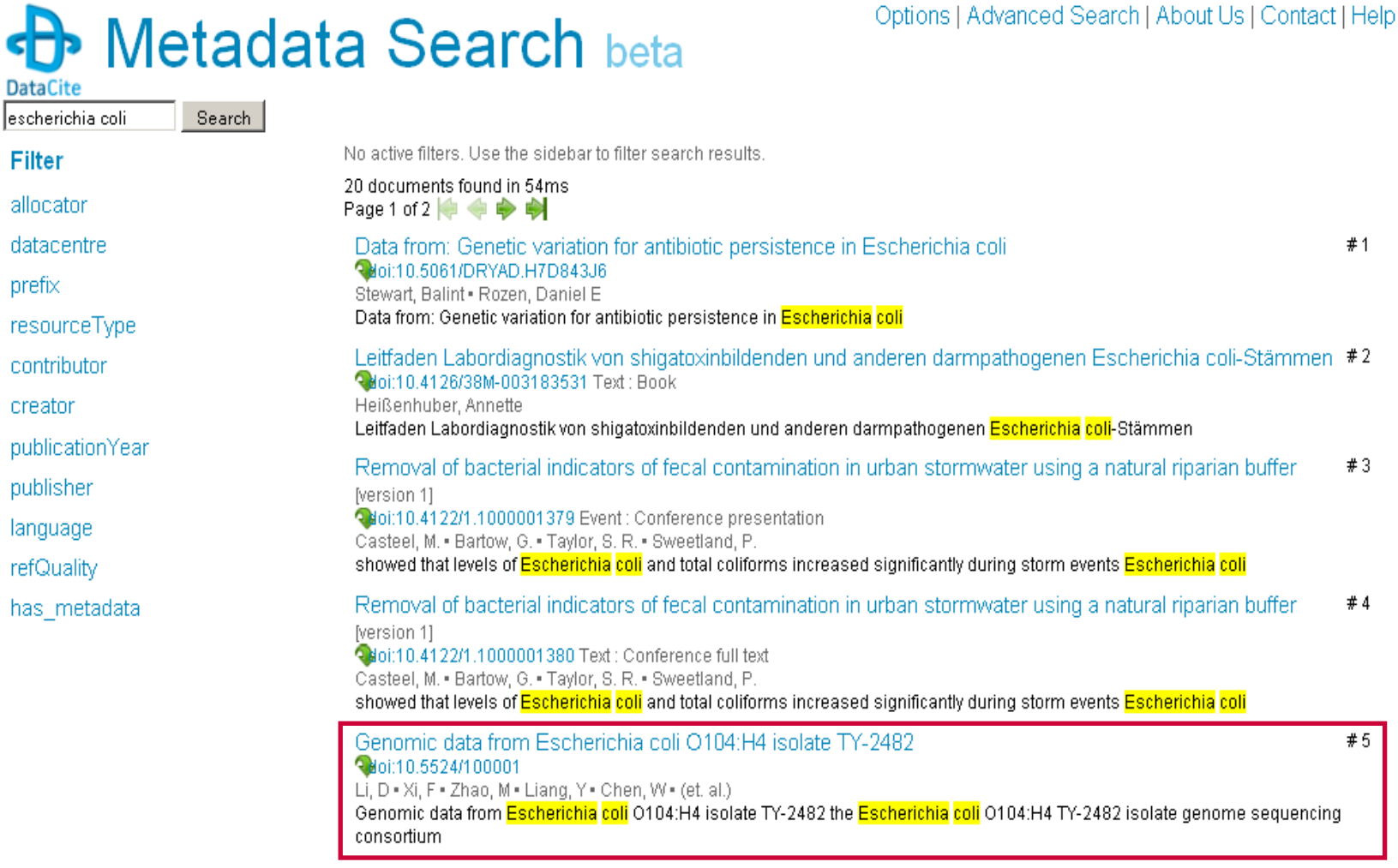
Mission

- Citability of research data
- High visibility of the data
- Easy re-use and verification of the data sets
- Increasing quality of published papers

DataCite Members



Example: EHEC virus



Example: EHEC virus

E. coli

The May 2011 outbreak of an *E. coli* infection in Europe resulted in serious concerns about the potential appearance of a new deadly strain of bacteria, *Escherichia coli* O104:H4 TY-2482. In response to this situation, and immediately after the reports of deaths, the University Medical Centre Hamburg-Eppendorf and BGI-Shenzhen worked together to sequence the bacterium and assess its human health risk.

The bacterium's genome was first sequenced using Life Technologies; Ion Torrent sequencing platform. According to the results of the draft assembly, the estimated genome size of this new *E. coli* strain is about 5.2 Mb. Sequence analysis indicated this bacterium is an EHEC serotype O104 *E. coli* strain. Comparative analysis showed that this bacterium has 93% sequence similarity with the EAEC 55989 *E. coli* strain, which was isolated in the Central African Republic and known to cause serious diarrhea. This strain of *E. coli*, however, has also acquired specific sequences that appear to be similar to those involved in the pathogenicity of hemorrhagic colitis and hemolytic-uremic syndrome. The acquisition of these genes may have occurred through horizontal gene transfer.

To maximize its utility to the research community and aid those fighting the epidemic, this genomic data was released into the public domain under a CC0 license.

To the extent possible under law, BGI Shenzhen has waived all copyright and related or neighboring rights to genomic data from the 2011 *E. coli* outbreak. This work is published from: China.

Workflow and further information can be found at:

http://climb.genomics.cn/Ecoli_TY-2482

Citation

In further accordance with our terms of use, please cite this dataset as:

Li, D; Xi, F; Zhao, M; Chen, W; Cao, S; Xu, R; Wang, G; Wang, J; Zhang, Z; Li, Y; Cui, C; Chang, C; Cui, C; Luo, Y; Qin, J; Li, S; Li, J; Peng, Y; Pu, F; Sun, Y; Chen, Y; Zong, Y; Ma, X; Yang, X; Cen, Z; Song, Y; Zhao, X; Chen, F; Yin, X; Rohde, H; Liang, Y; Li, Y and the

DOI Services

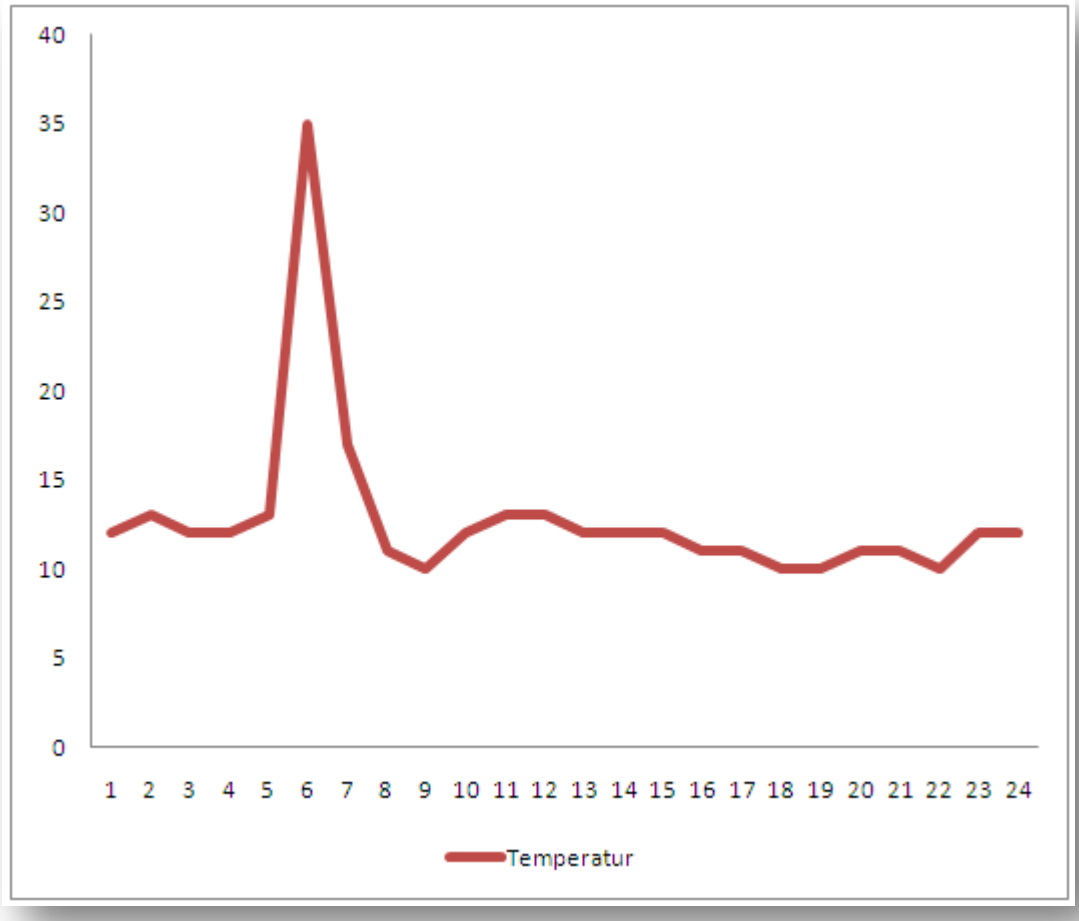
- Contracts with 60 data centres
  - Research Institutes
  - Universities
  - Libraries
  - Publisher
- 776.454 DOI registrations
  - 22.533 up to September 2013

Research data – Further developments

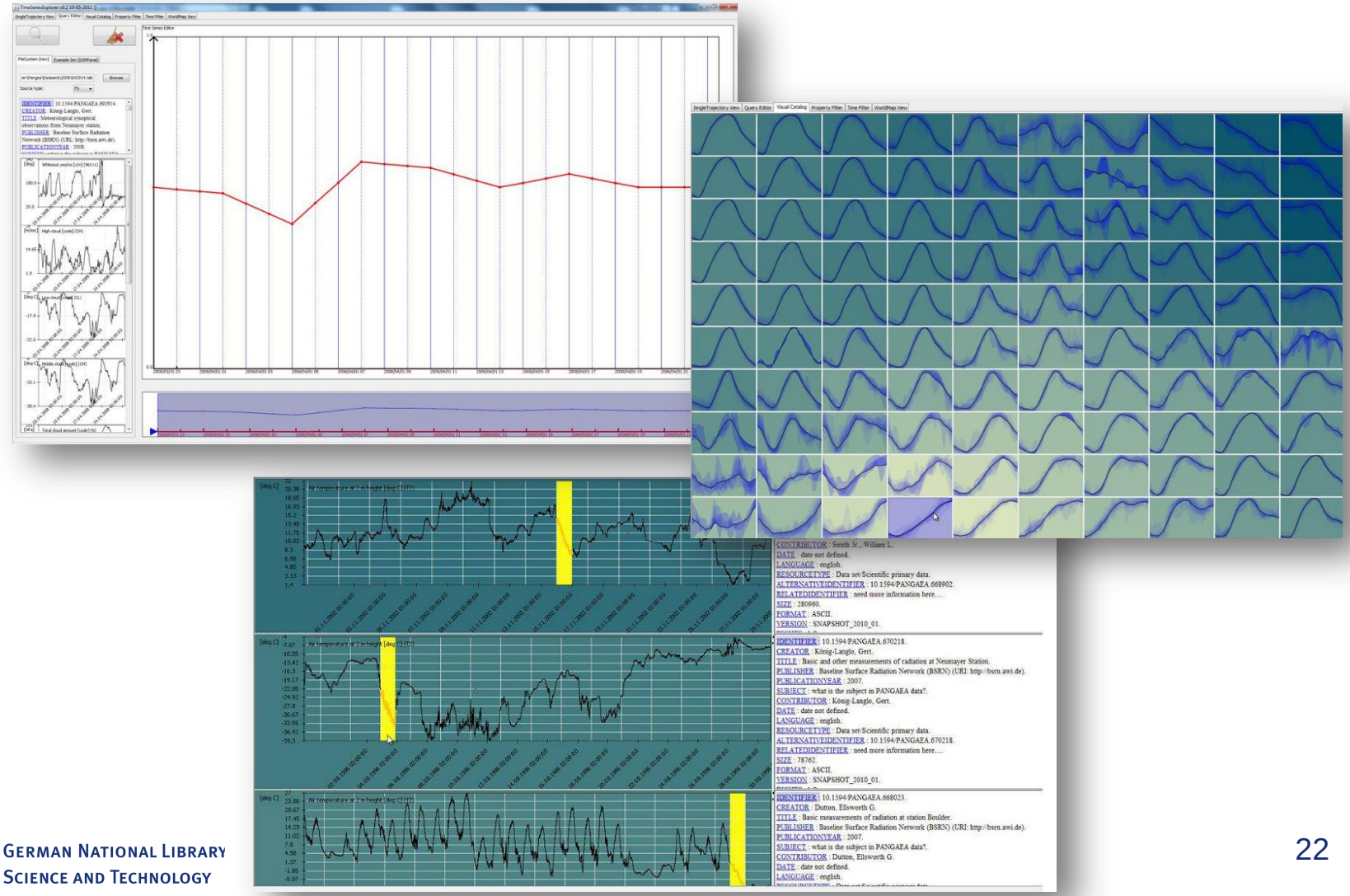
- KomFor
  - Centre of Expertise for Research Data from the „Earth and Environment“ project
- RADAR
  - RADAR - Research Data Repositorium
- Visual Analysis
    - VisInfo Methods

Numerical data

Zeit [h]	T [°C]
1	12
2	13
3	12
4	12
5	13
6	35
7	17
8	11
9	10
10	12
11	13
12	13
13	12
14	12
15	12
16	11
17	11
18	10
19	10
20	11
21	11
22	10
23	12
24	12



Visual access to research data

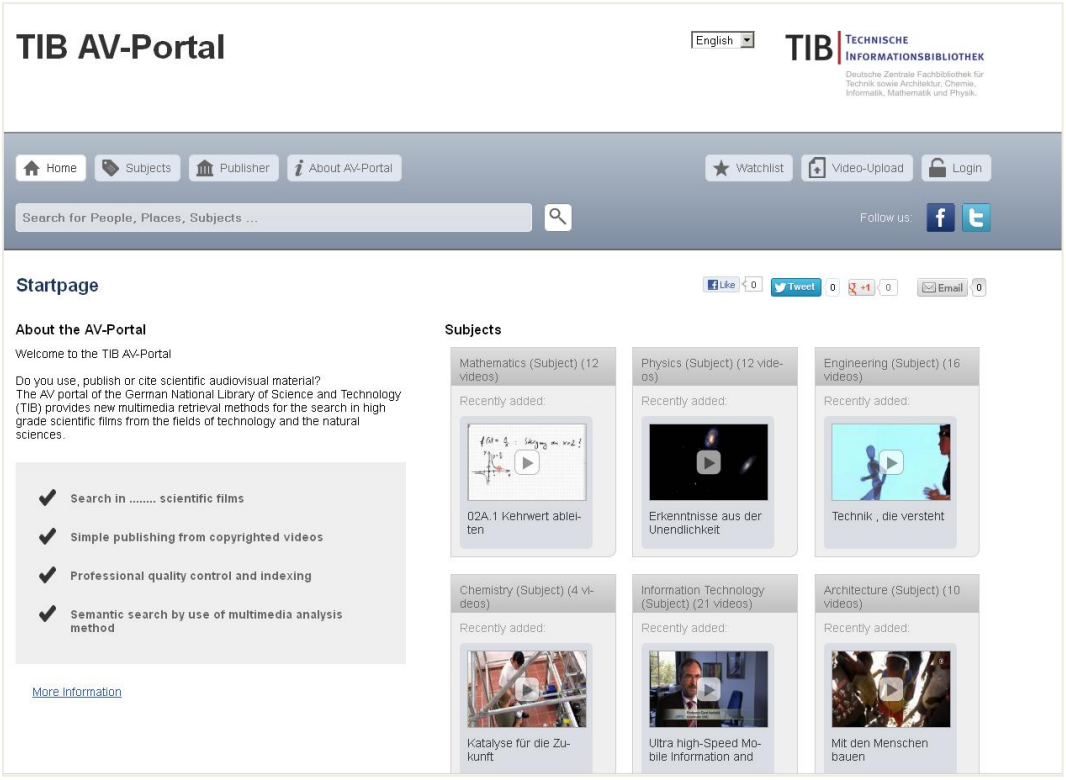


How have we been preparing ?

- Infrastructure for research data
- ***Visual search tools for AV-media***
- 3D-Objects
- chemOCR

TIB‘s portal for audiovisual media

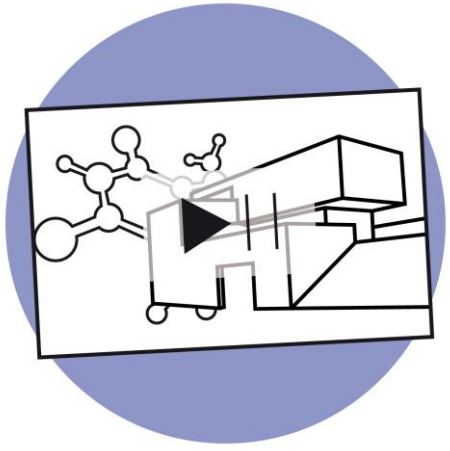
Project	Development of a portal for audiovisual media
Aim	Improve access to AV-Media
Time	July 2011 – December 2013
Partner	Hasso-Plattner Institut for Softwaresystemtechnology GmbH



TIB‘s portal for audiovisual media

How do I find what I‘m looking for in videos?

Today: Manual annotation of the whole video

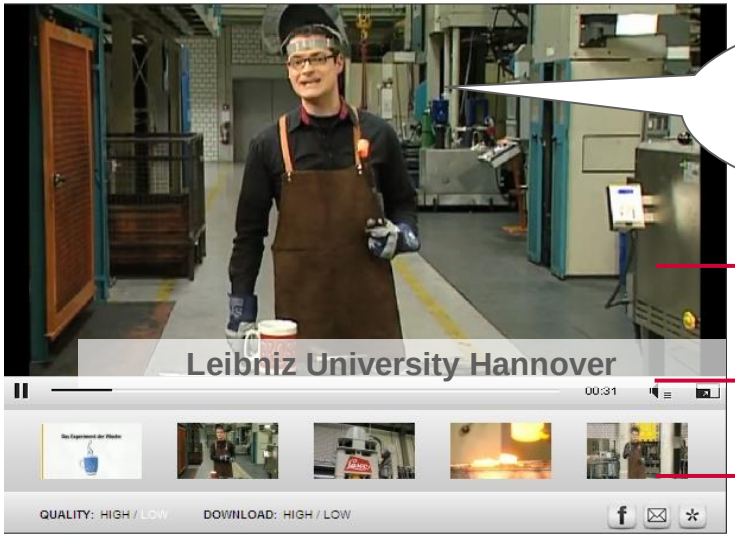


Metadata

- Titel
- Author
- Description
- Publisher
- Publication year
- Rightsholder
- …..

TIB‘s portal for audiovisual media

Future: Manual Annotation plus content-based information



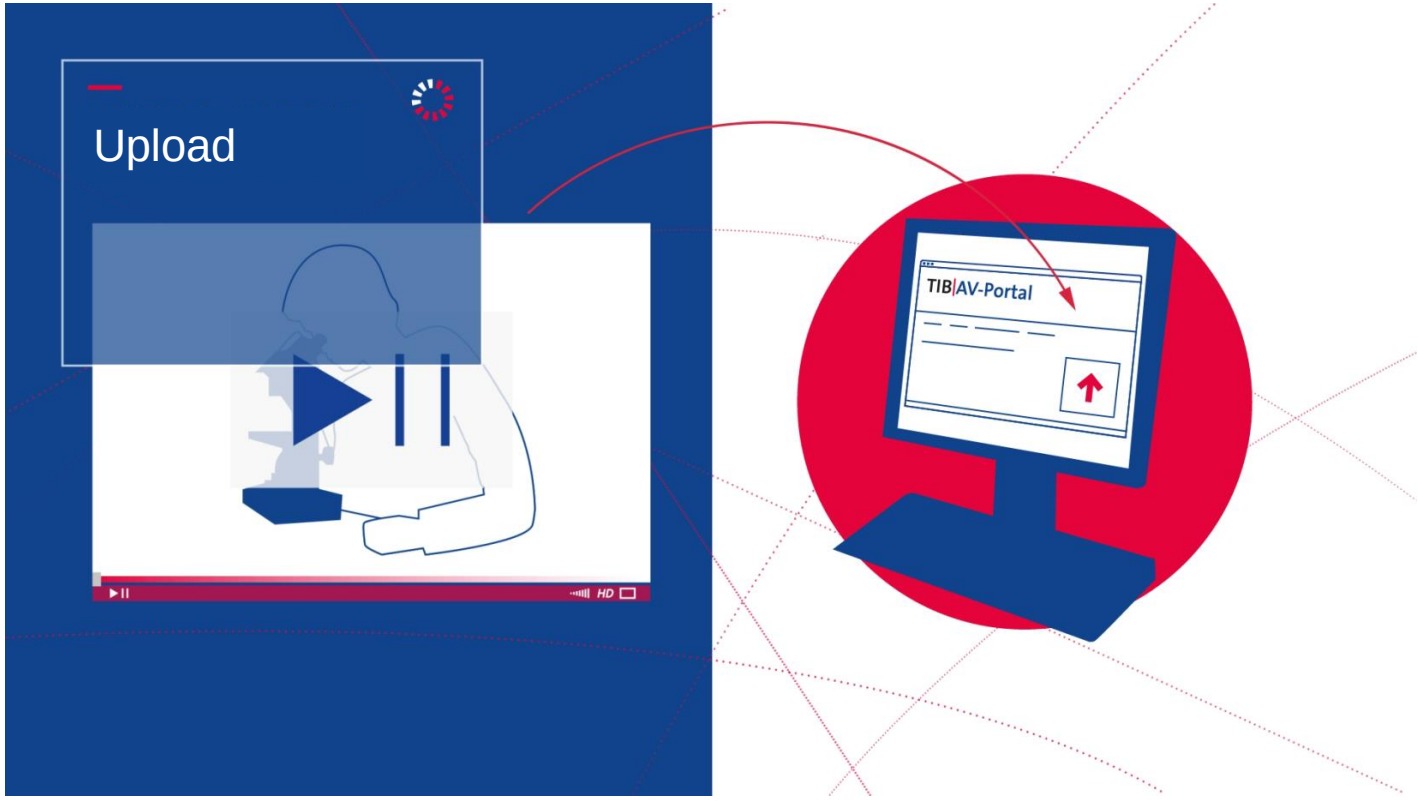
source: Scorupka, Sascha, Experiment der Woche, 2011

1. Speech

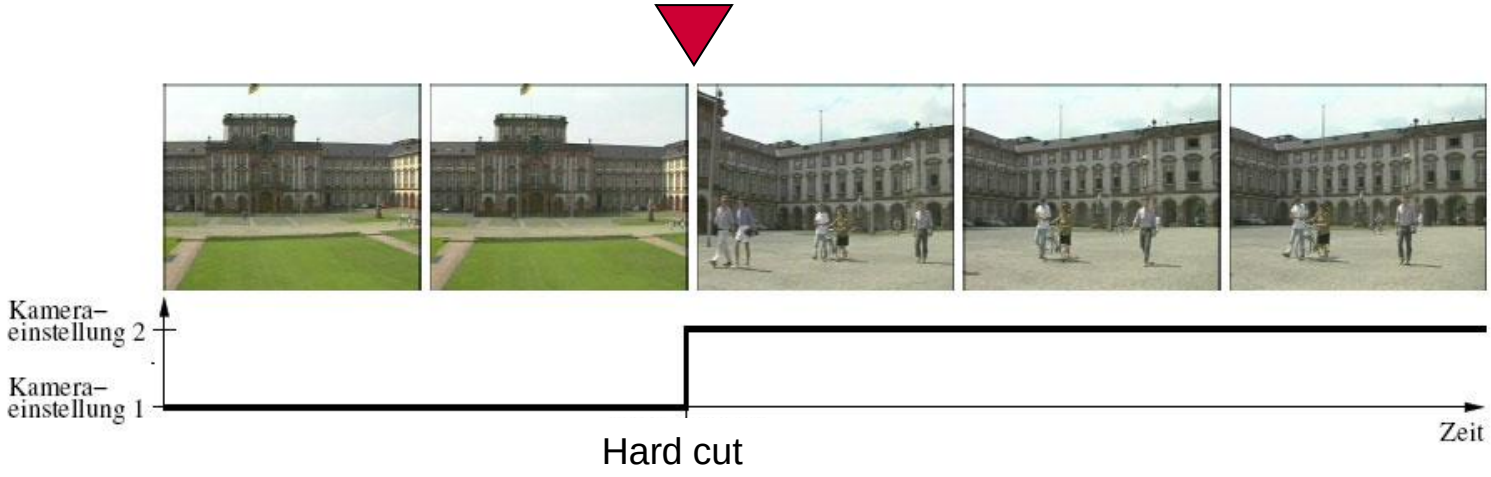
2. Visual features
   e.g. Indoor, Experiment, Technology
3. Textual information
4. Structural Information
   Scenes, Shots, Segments

TIB‘s portal for audiovisual media

Media analysis process



TIB‘s portal for audiovisual media
Scene recognition

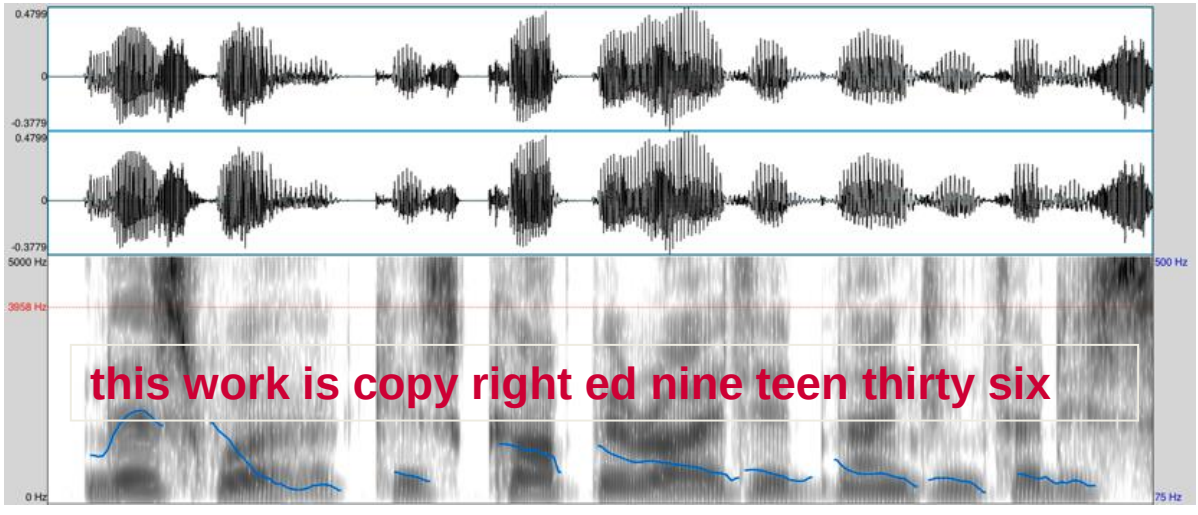


Automatic cut detection
→ luminance / contrast
→ colour distribution / colour histogramm
→ edges

Kopf, S. Computergestützte Inhaltsanalyse von digitalen Videoarchiven, Mannheim. 2006

TIB‘s portal for audiovisual media

Automatic speech recognition



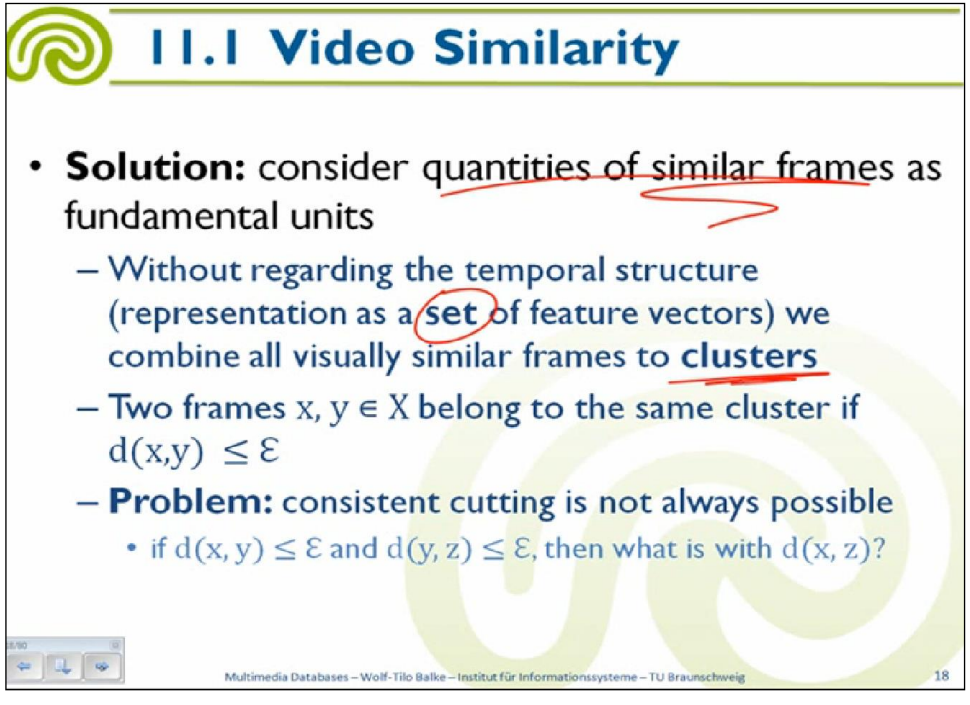
Quality of results is dependent upon

- quality of the speaker
- dialects
- background noises
- voice overlaps

Intelligent Character Recognition

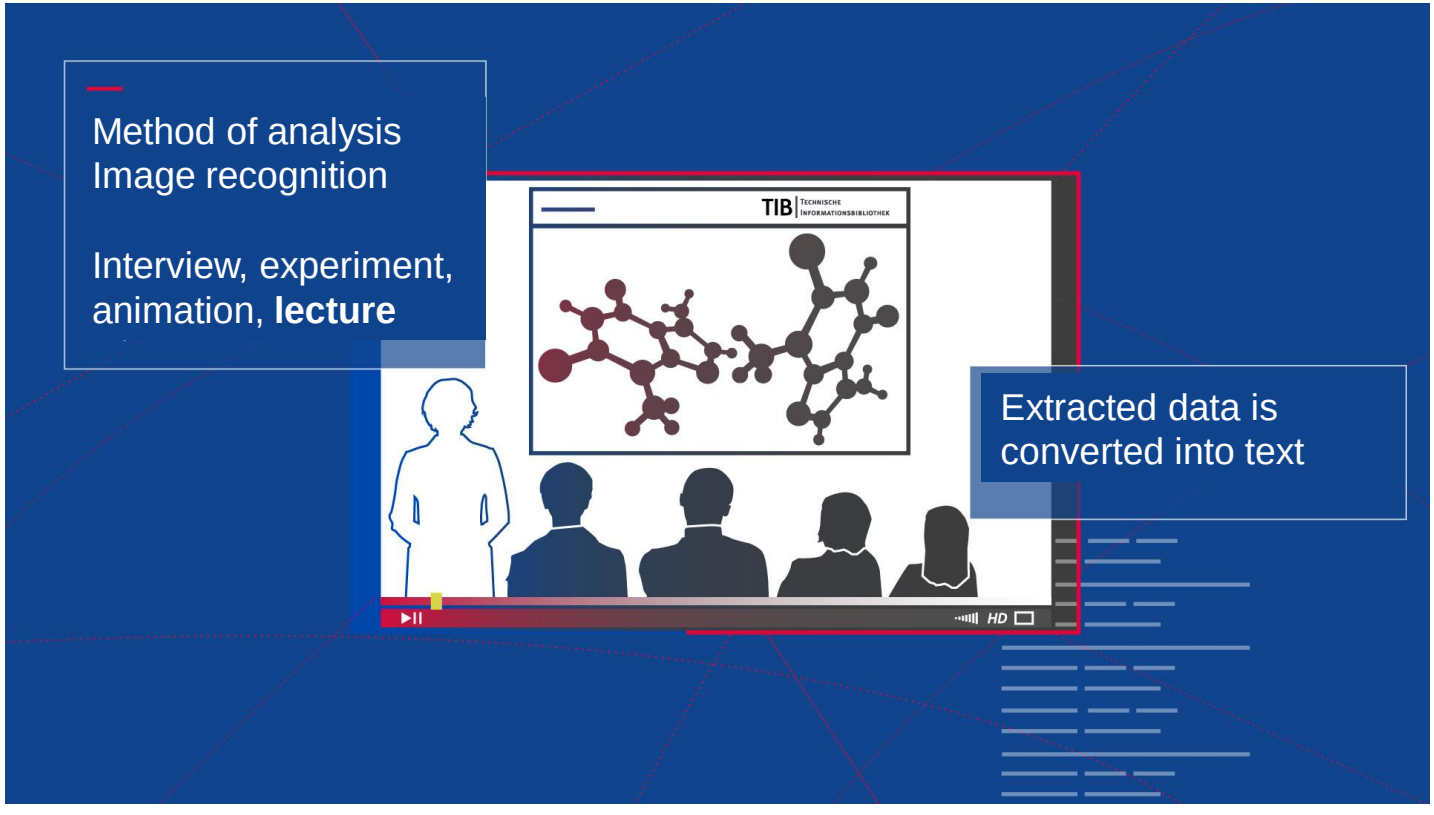
Intelligent Character Recognition (ICR)

- Character/Logo Detection
- Character Filtering
- Character Recognition



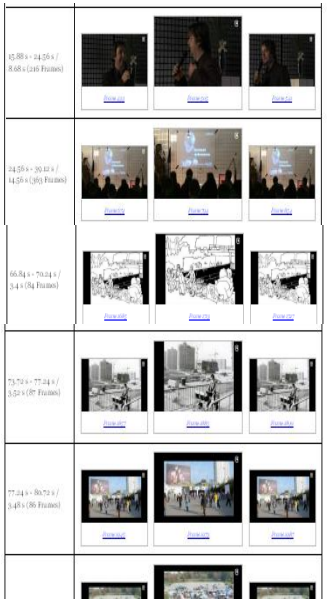
TIB‘s portal for audiovisual media

Automated analysis: Image recognition

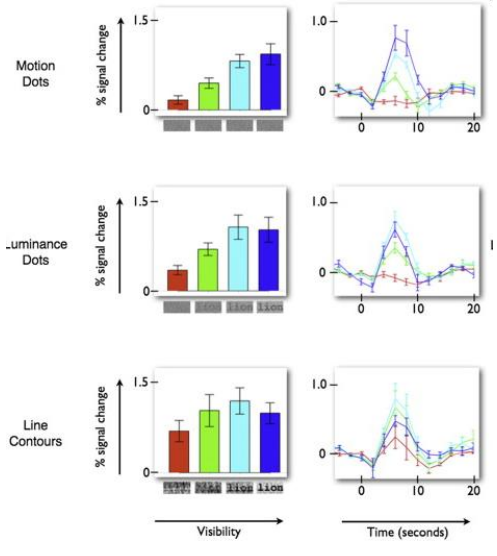


TIB‘s portal for audiovisual media

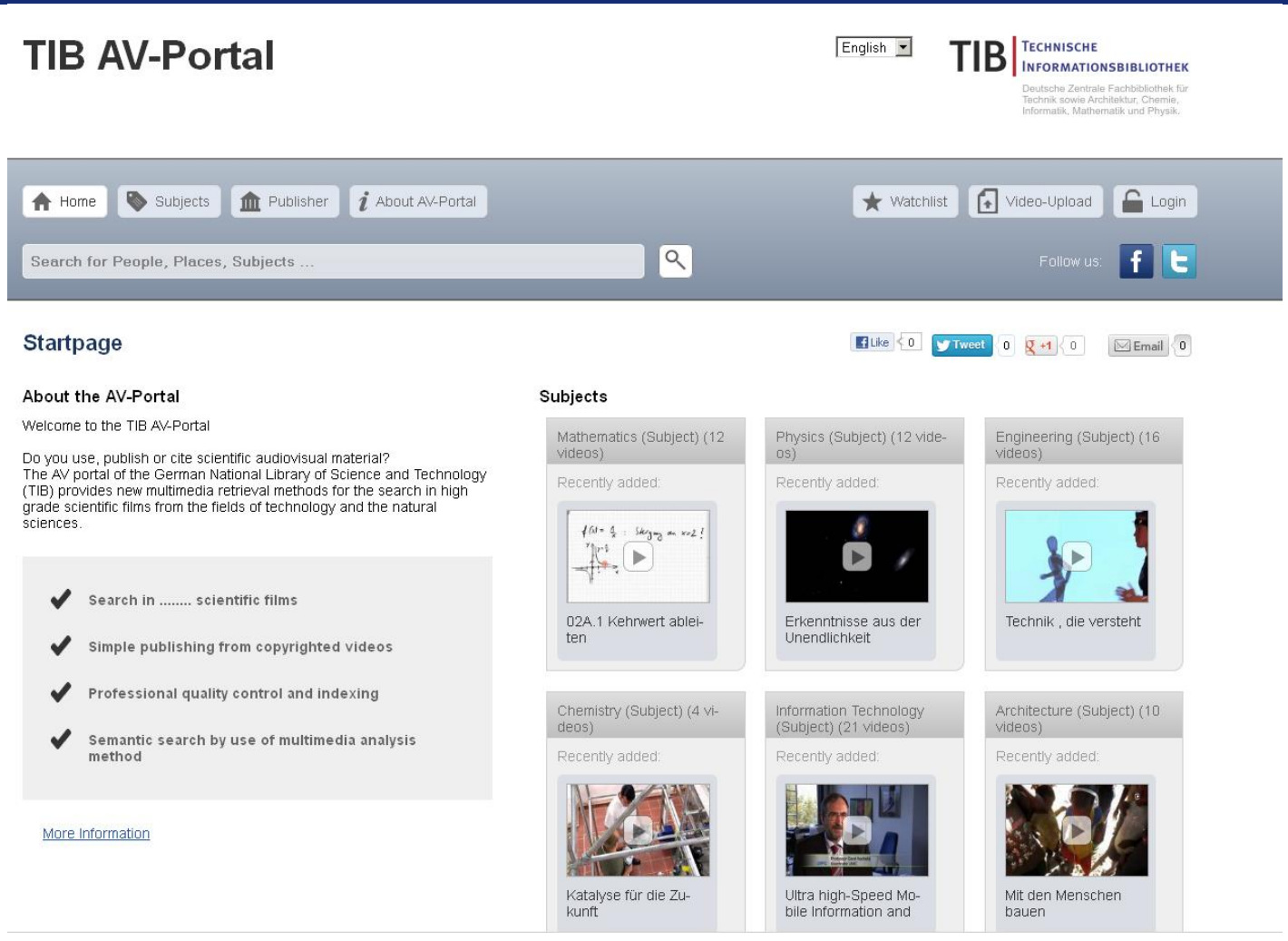
Keyframes → **Annotation** → **Machine learning using visual features**



Visual Concepts
Graphical : Animation
Graphical : Drawing
Graphical : Diagram
Real : Outdoor
Real : Indoor
Real : Lecture / Conference
Real : Interview
Real : Buildings
...



TIB‘s portal for audiovisual media

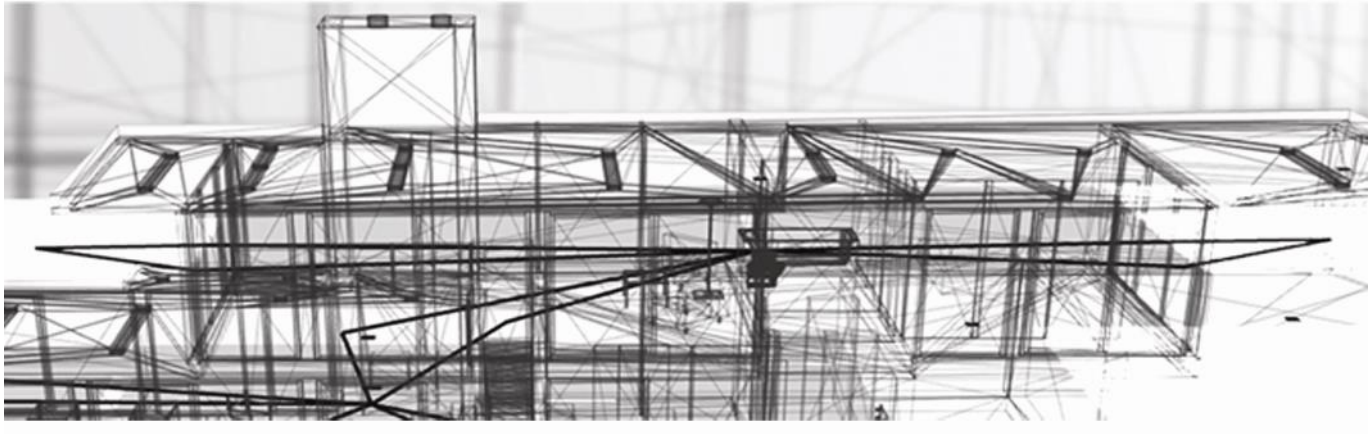


How have we been preparing?

- Infrastructure for research data
- Visual search tools for AV-media
- ***3D Objects***
- chemOCR

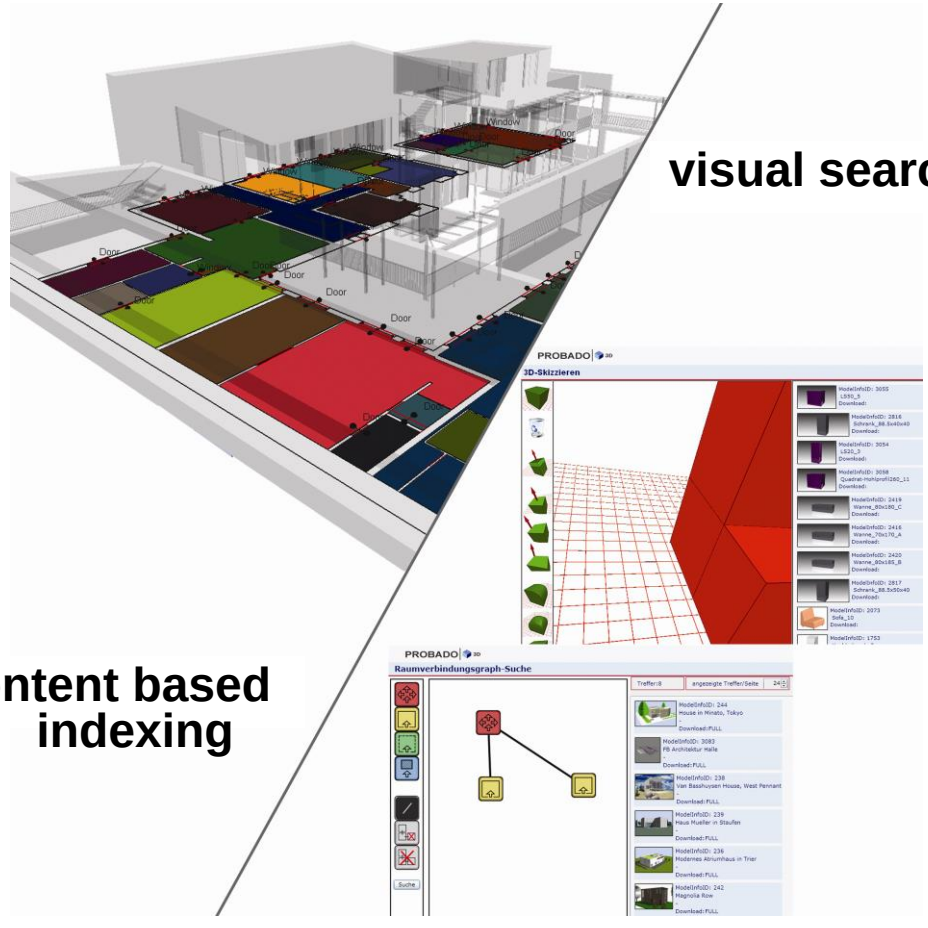
3D Objects – an excursion to Architecture

PROBADO | 3D

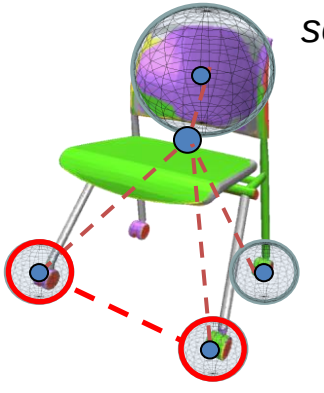


TECHNISCHE UNIVERSITÄT DARMSTADT | universitätbonn | TU Graz | TIB Technische Informationsbibliothek | DFG

Visual search tools

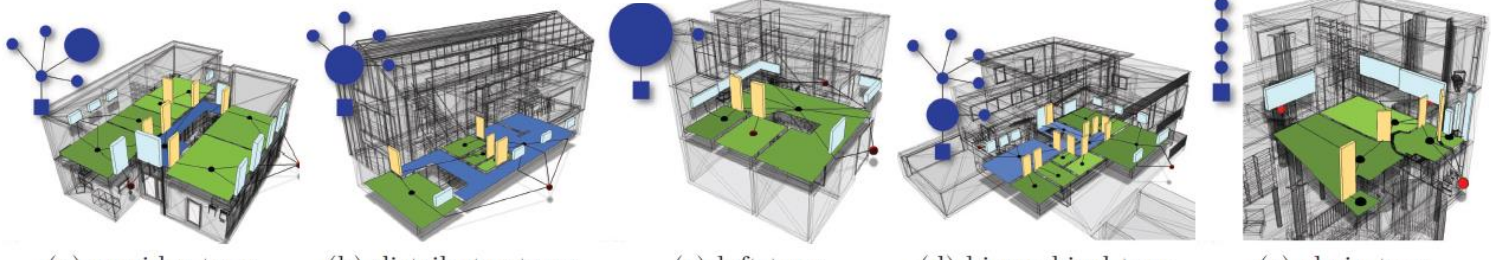


Content based indexing



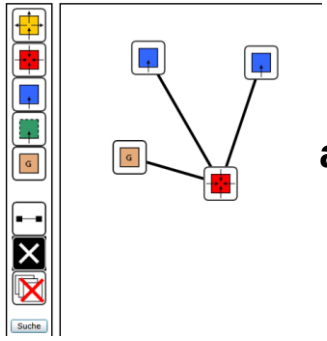
segmentation with
form-primitives

extraction of
room connectivity graphs

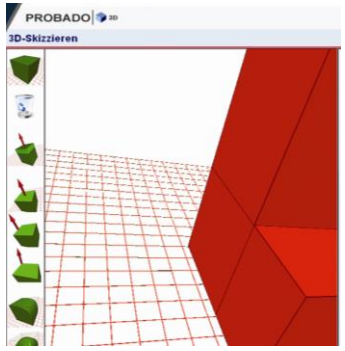


(a) corridor type (b) distributor type (c) loft type (d) hierarchical type (e) chain type

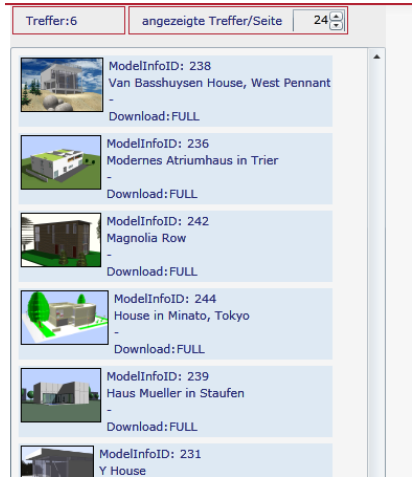
Visual search



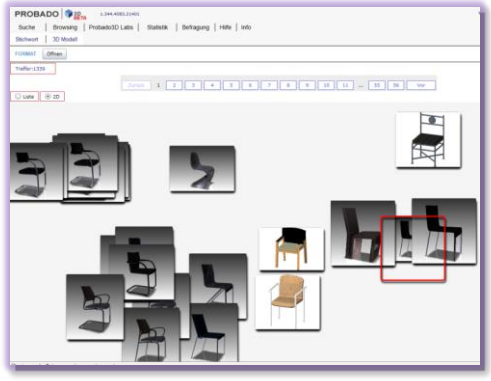
attributed graph



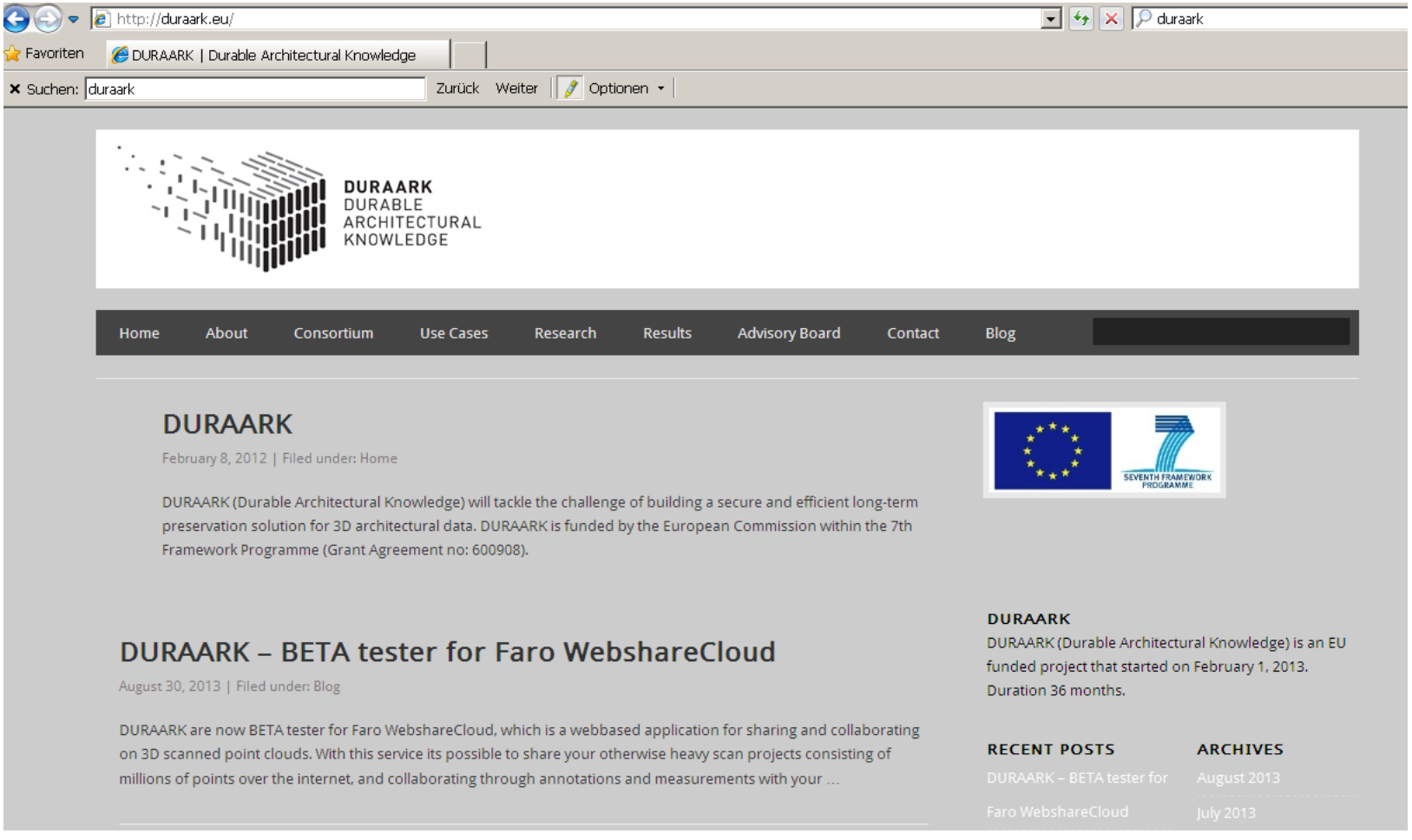
3D sketch



result visualization



Further developments



How have we been preparing ?

- Infrastructure for research data
- Visual search tools for AV-media
- 3D Objects
- ***chemOCR***

Information retrieval in Chemistry

Search for chemical structures – how?

Chemists are used to drawing

Textual and non-textual chemical information

Chemical Names

Linked entities from the table

Reaction scheme

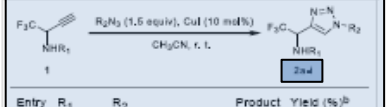
Table with reaction scheme

2a-i: Derivates from the reaction

Chemical structure

The copper(I)-catalyzed 1,3-dipolar cycloaddition [33-38] of organic azides and alkynes (also called "click chemistry") resulting in the formation of 1,2,3-triazoles has become an increasingly attractive area [39]. According to the literature [33-38], the Cu(I) species can be used directly (*e.g.* CuI), or generated by oxidation of a Cu(0) or reduction of a Cu(II) species. Catalysis by the CuI is known to yield exclusively the 1,4-disubstituted regioisomer [33,34]. First, the N-(*p*-methoxyphenyl)-1-(trifluoromethyl)propargylamine was reacted with benzyl azide in the presence of CuI (10 mol%) and showed good reactivity with completion of the reaction within 24 h, whereas the use of $CuSO_4$/Na ascorbate afforded the cycloadduct in low yield. The reaction was then carried out with different propargylamines (N-(p-methoxyphenyl) and N-benzyl) and various azides at room temperature in acetonitrile within 24 h which afforded the compounds 2a-i with good yields (63-92%) after purification by column chromatography. The results are summarized in Table 1.

Table 1: Copper(I)-catalyzed synthesis of 1,4-disubstituted triazoles



Entry	R_1	R_2	Product	Yield (%)[b]
1	-PMP[a]	-Bn	2a	82
2	-PMP	$-CH_2COPh$	2b	76
3	-PMP	$-CH_2OCOC(CH_3)_3$	2c	73
4	-PMP	$-CH_2CO_2CH_3$	2d	83
5	-PMP	$-CH_2CH_2OH$	2e	87
6	-Bn[a]	-Bn	2f	75
7	-Bn	$-CH_2COPh$	2g	63
8	-Bn	$-CH_2CO_2CH_3$	2h	92
9	-Bn	$-(CH_2)_2OH$	2i	73

[a]PMP: p-methoxyphenyl, Bn: benzyl. [b]Yield after flash purification.

As expected the new triazoles were formed in a fully regioselective manner affording the 1,4-regioisomer as highlighted from NOE experiments on compound 2c (Figure 1). A strong correlation was observed between the hydrogen H_a and H_b respectively. The structure of the other compounds 2a-i was assigned by analogy with 2c.

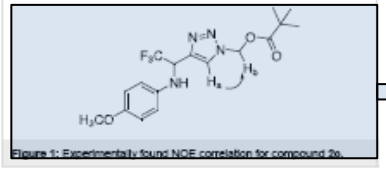
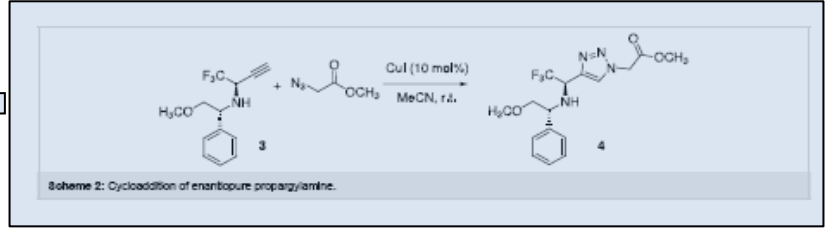


Figure 1: Experimentally found NOE correlation for compound 2c.

In our goal to study the influence of the CF_3 group on the conformation of peptidomimetics, we applied our strategy to the enantiopure trifluoromethyl-propargylamine 3 bearing the removable (*R*)-phenylglycinol chiral auxiliary (Scheme 2) [30-32].

The reaction was carried out under the same condition with azidoacetic acid methyl ester and afforded the cycloadduct 4 in good yield (79%) and as a single isomer without any racemization. This compound can easily afford the free amino ester which is a promising trifluoromethyl building block for the synthesis of new triazole-based trifluoromethyl oligomers.



Scheme 2: Cycloaddition of enantiopure propargylamine.

Non-textual data processing – chemOCR

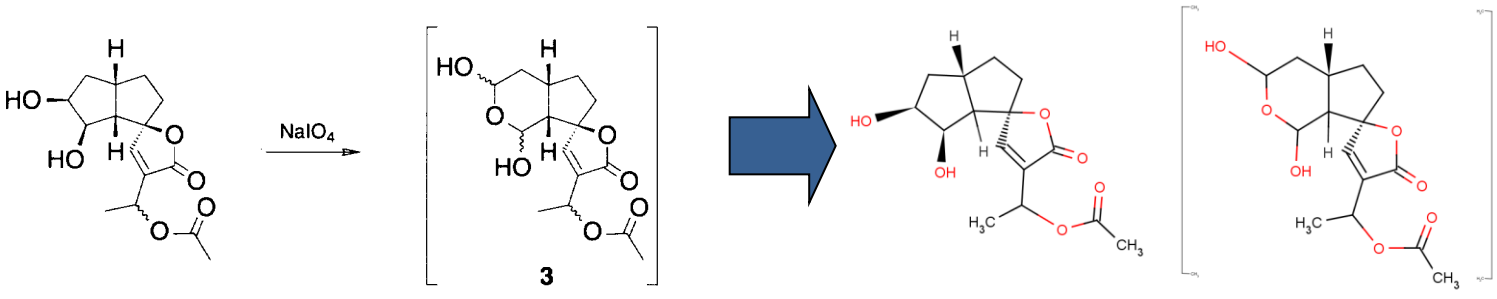


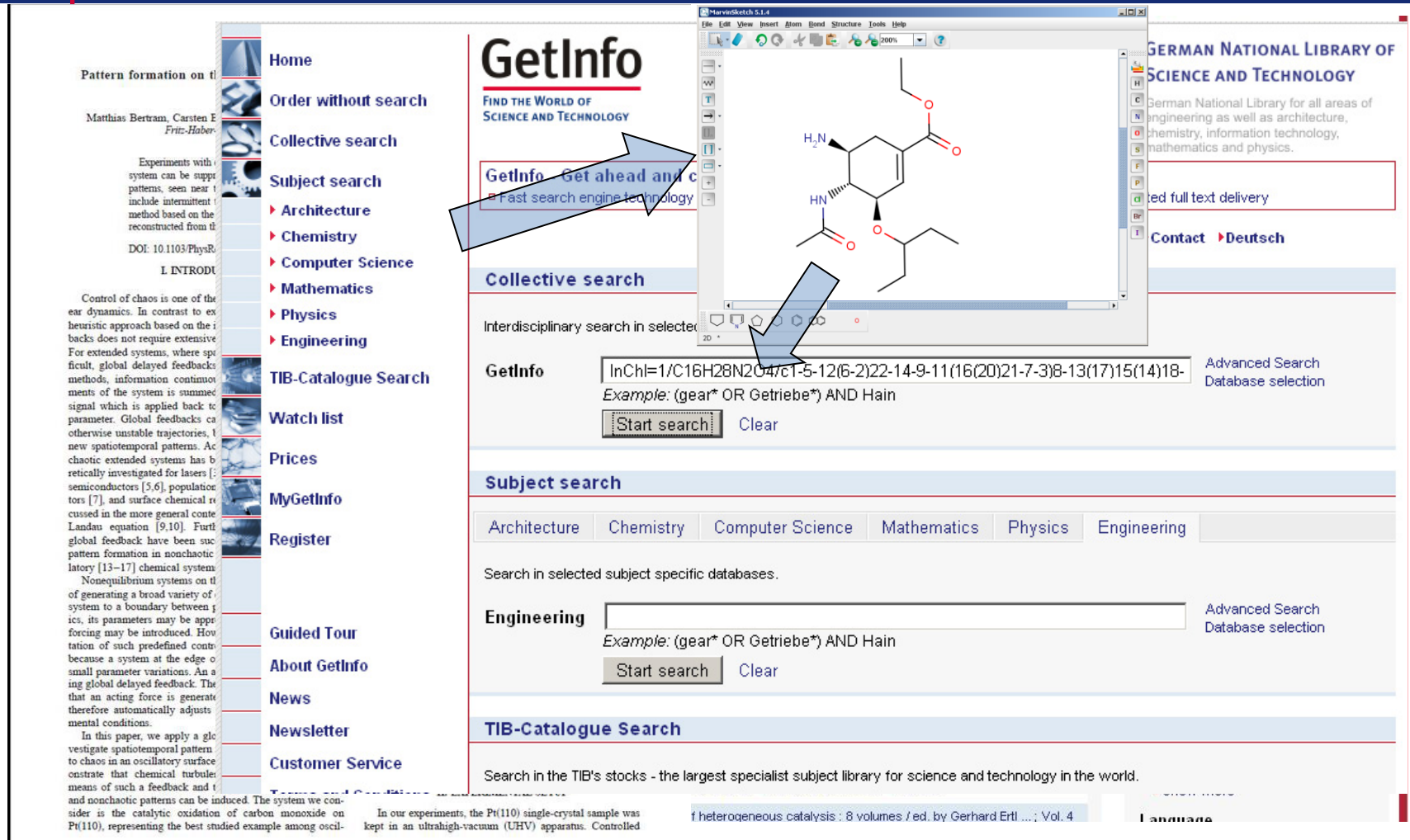
image data

chemical structure data

CLiDE

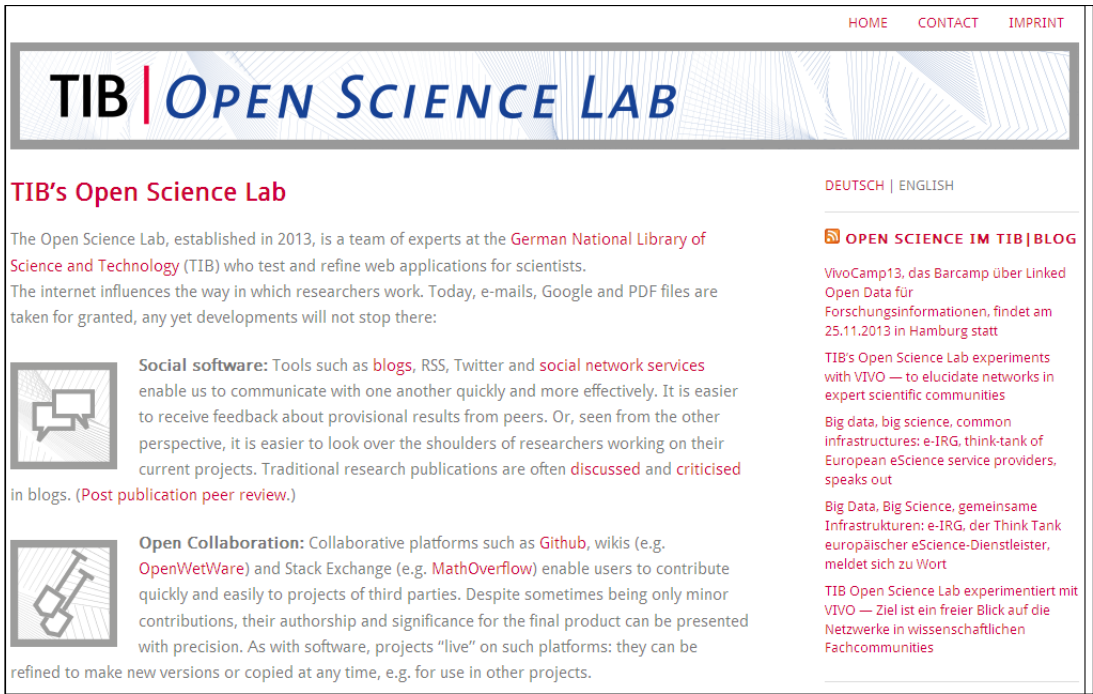
chemOCR

Information retrieval in chemistry
Text AND formulas



Further subjects

- Open Science Lab
- Ontology



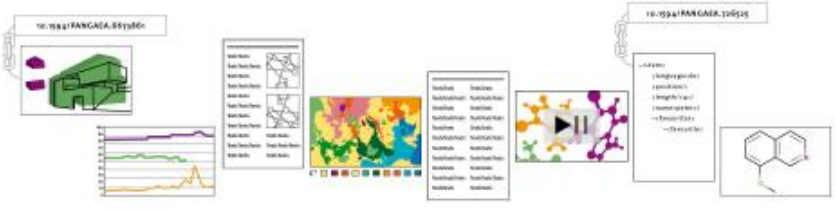
Conclusion

Dissemination of scientific and technical information has been a foundational mission.

The methods have completely changed, but the mission remains the same.

Conclusion

Ultimate Goal:



Interlinking and Search Across All Types of Digital Assets.

GetInfo – Portal for Science and Technology

- 58 m metadata in internal index
- 390 m metadata in external sources
- 900.000 pdf fulltexts
- Data, AV-Media, 3D Objects

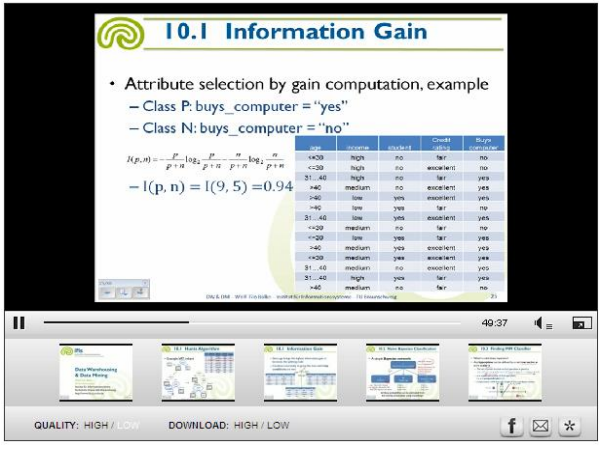
GetInfo

FIND THE WORLD OF SCIENCE AND TECHNOLOGY

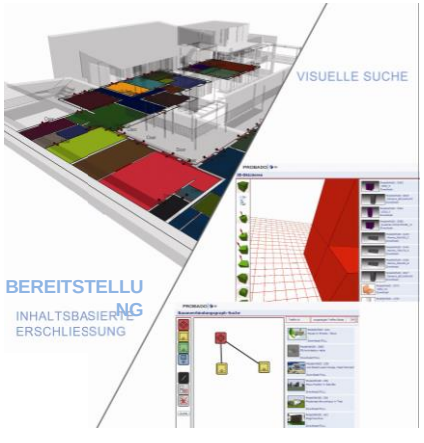
Development of media-specific portals

GetInfo

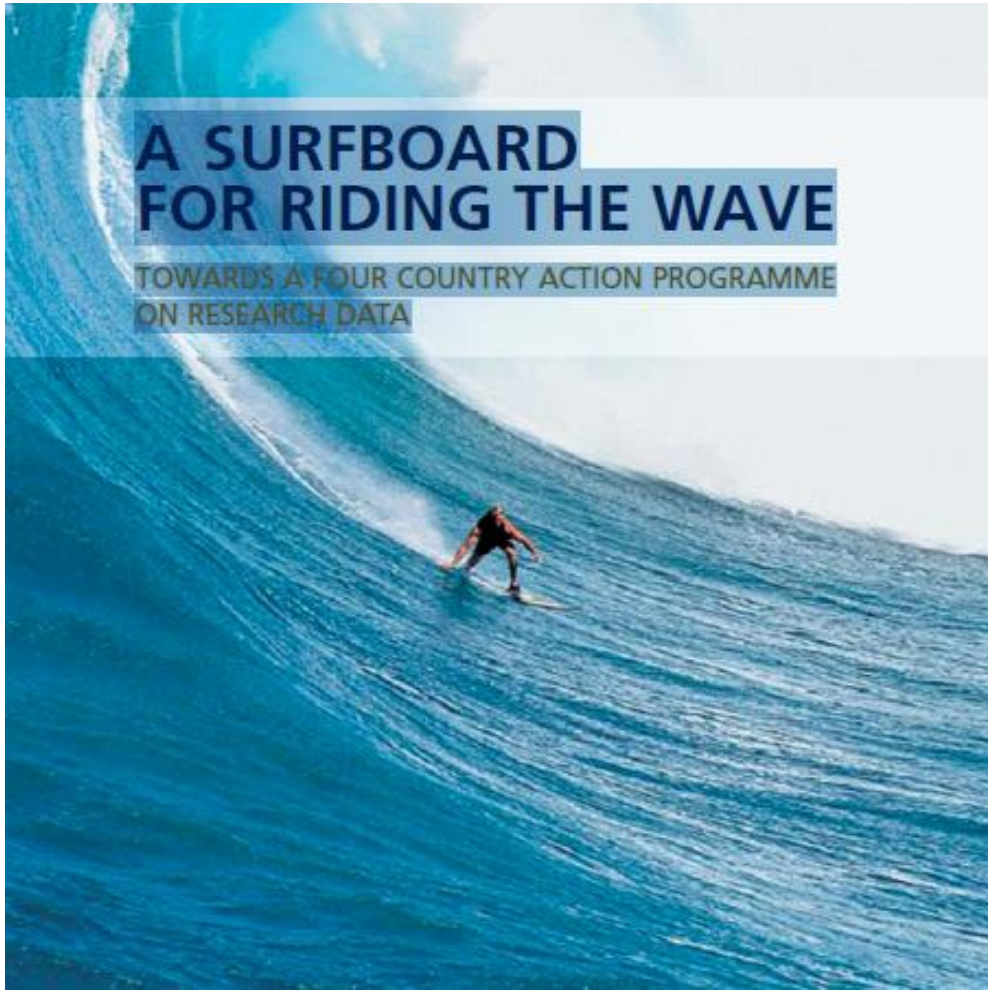
FIND THE WORLD OF SCIENCE AND TECHNOLOGY



Portal for audiovisual Media



Probado 3D


A SURFBOARD
FOR RIDING THE WAVE
TOWARDS A FOUR COUNTRY ACTION PROGRAMME
ON RESEARCH DATA

Questions?