TIB | German National Library of Science and Technology

Hochschule Hannover
University of Applied Sciences and Arts

How to discover reusable technology pictures

An joint grant proposal by TIB & HsH

Lambert Heller

36th Annual IATUL Conference 2015

Hannover, July 9, 2015

Pictures in technology & engineering related literature

More pictures, and more useful pictures, every day

- Today roughly 2-3 pictures (apart from graphs, formula, tables) in each journal article from technology & engineering areas
- Often key to quick understanding of research & its results
- Often **highly reusable**: education, journalism, other research…

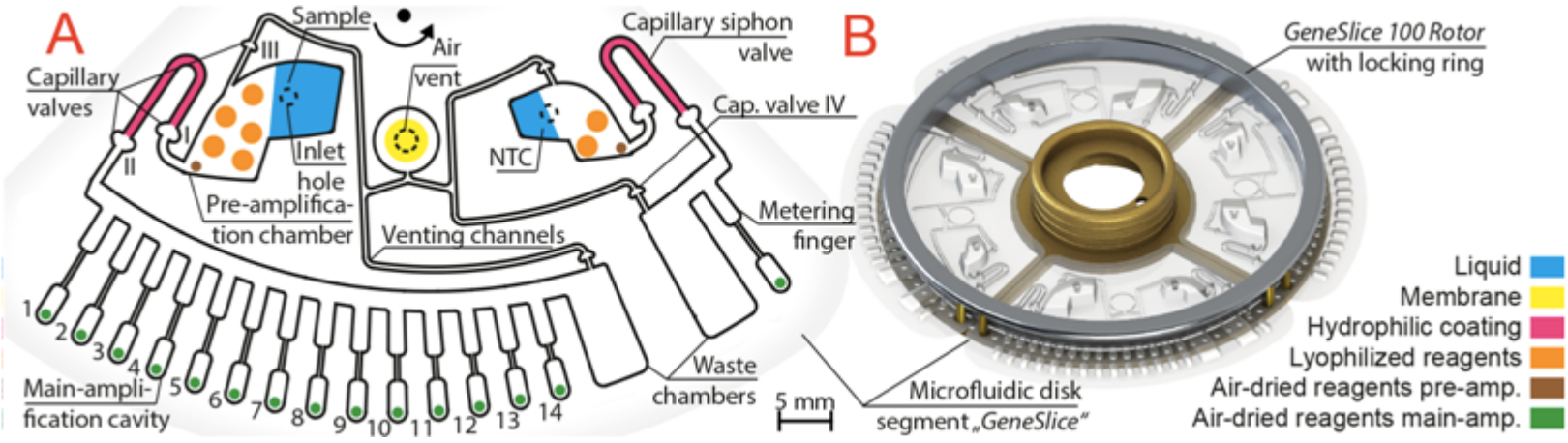


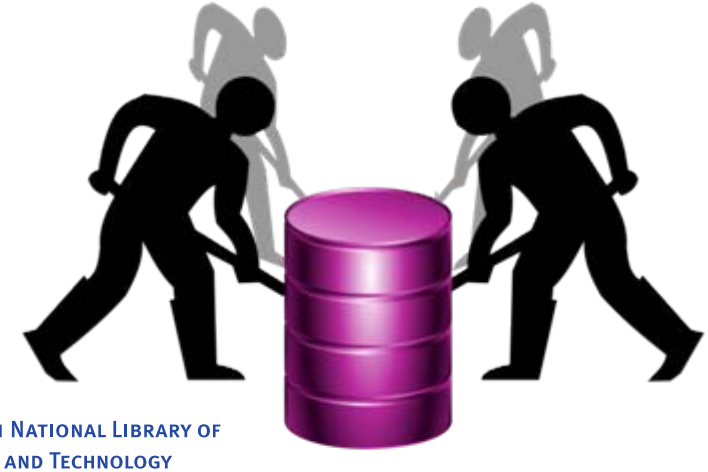
Image source: doi:10.1371/journal.pone.0131845.g001

How do you find reusable pictures, by topic?
Clearly depends heavily on expectations & field...

- Google image search (and similar) can be restricted to material licensed for reuse – but has low precision
- **Some disciplines** are highly centered on digital imagery, like biology & medicine, geo sciences incl. astronomy
- High precision & multi-lingual (that is, thesaurus-based) retrieval of pictures in broader fields, like technology & engineering? **Nothing we are aware of, so far…**

How to make a huge difference, quickly?
Content mining for pictures? How? Where?

- Large-scale thesaurus based content mining on tec+eng articles
- …focused on picture captions & picture references within the articles
- Selection criteria in regard to mining: all content in XML & CC-BY
- Only high volume journals worthwile (for a start we go with 90 journals from six large OA publishers, all of them OASPA members)
- We expect >100.000 pictures alone from last three years
- Last, crucial step: **Make result as far available as possible**



Thanks to Bastian Drees, graduate librarian trainee at TIB, for exploring and preparing a prospective article corpus

Where to store the pictures & metadata? **Wikipedia**

The all-open access, top reference website worldwide...

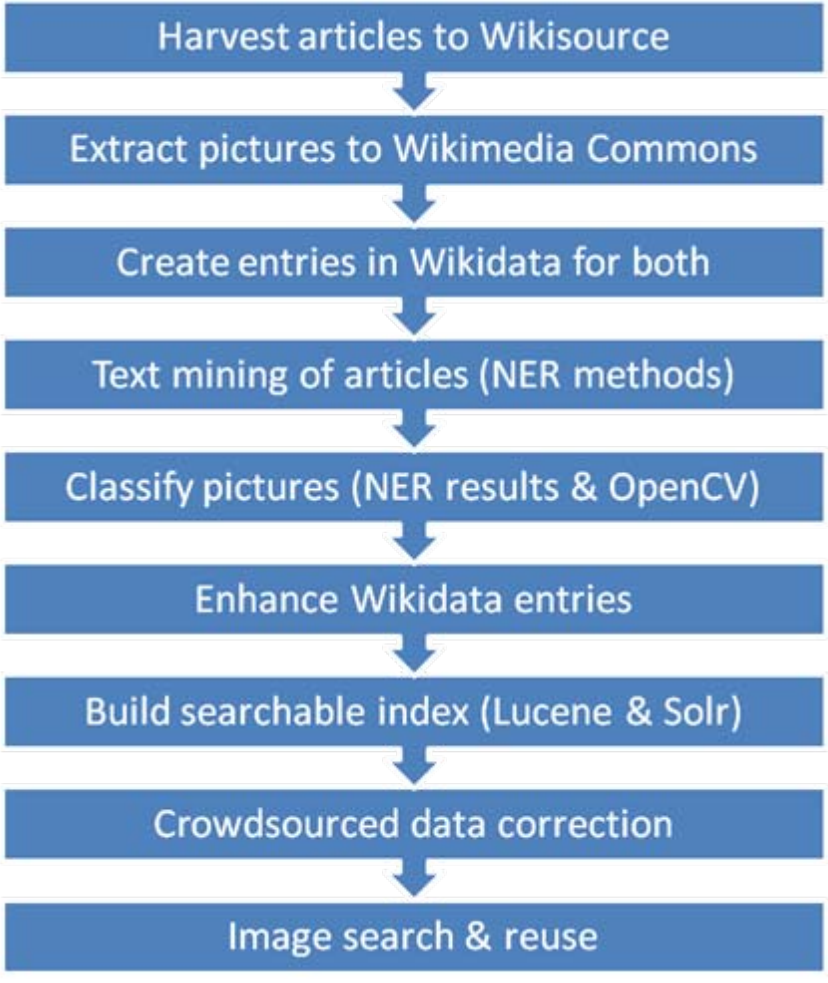
- Wikidata: Collaborative database, and de facto thesaurus
- Wikimedia Commons: Already the largest open picture archive
- Wikiproject Open Access: Already many science applications
- Wikipedia Zero: free mobile WP access in developing countries; Google knowledge graph: draws from Wikipedia & Wikidata…

Site	Domain	Alexa Traffic Rank (Mar 2015)[3]	SimilarWeb Top Websites (Mar 2015)[4]	Type
Google	google.com	1	2	Search engine
Facebook	facebook.com	2	1	Social network
YouTube	youtube.com	3	3	Video sharing
Baidu	baidu.com	4	36	Search engine
Yahoo!	yahoo.com	5	4	Portal
Wikipedia	wikipedia.org	6	7	Reference encyclopedias
Amazon	amazon.com	7	17	Shopping
Twitter	twitter.com	8	8	Social network

TIB

Projected articles & pictures processing pipeline

Built on top of Wiki* bots & other FOSS software



WIKISOURCE

WIKIMEDIA COMMONS

WIKIDATA

OpenCV

Lucene

Apache Solr

Project participants, status & perspectives

Hopefully to start in 2016. We‘re open for collaboration!

- Content mining: experience brought into the project by Prof. Christian Wartena, University of Applied Science Hannover
- Great outcomes from TIB & HsH joint projects with students (cf. Ina Blümel’s VIVO talk this afternoon & her VIVO15 keynote)
- Multimedia, content mining, digital archiving, ontologies, OA – project involves multiple TIB strategic development areas
- Cooperation & endorsements from Wikimedia & CrossRef
- First grant proposal to DFG was denied (& idea encouraged)
- Our open grant proposal: doi:10.5281/zenodo.12745 (german)
- We expect a successful second review & to start in 2016!
- **We are eager to share everything, including code and data**
- We want to enhance the project (cooperations on global scale, e.g. thesis papers?, encourage reuse of the picture index…)

TIB | German National Library of Science and Technology

Thank you for your attention!