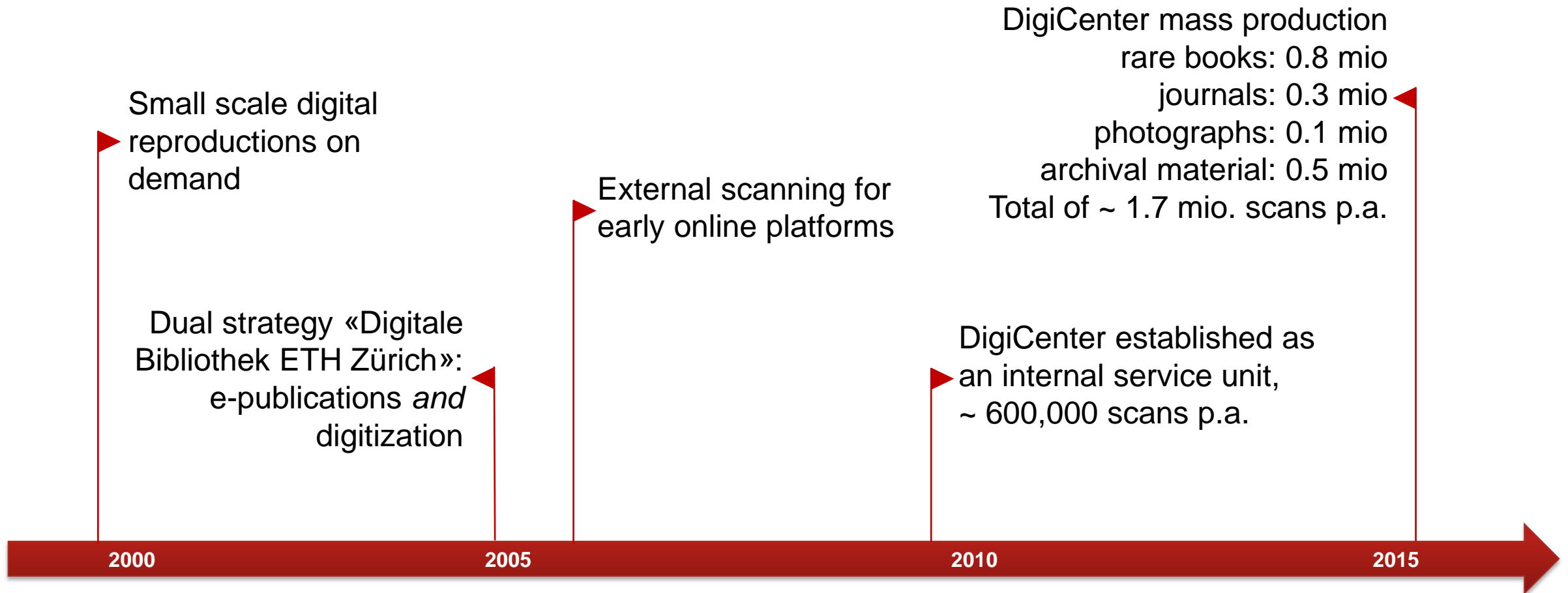IFLA Satellite Meeting 2017: Digital Humanities, Berlin, 15–17 August 2017

Research partnerships, user participation, extended outreach – some of ETH Library's steps beyond digitization
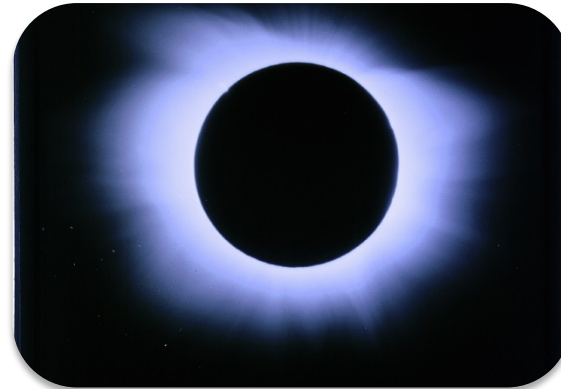
From «boutique» to mass digitization

DigiCenter mass production
rare books: 0.8 mio
journals: 0.3 mio
photographs: 0.1 mio
archival material: 0.5 mio
Total of ~ 1.7 mio. scans p.a.

Small scale digital reproductions on demand

External scanning for early online platforms

Dual strategy «Digitale Bibliothek ETH Zürich»: e-publications *and* digitization

DigiCenter established as an internal service unit, ~ 600,000 scans p.a.

| 2000 | 2005 | 2010 | 2015 |

(Cooperative) online platforms for digitized content

Rare books

e-rara

Images

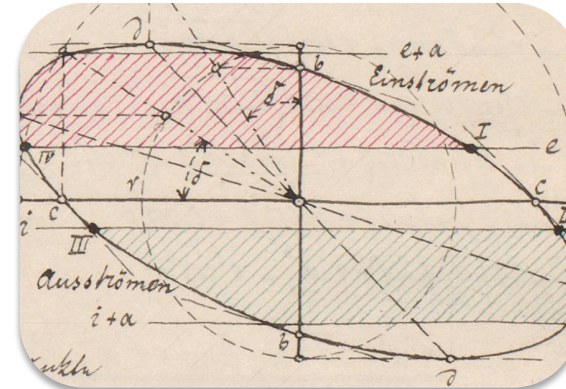e-pics Image Archive

Archival material

e-manuscripta

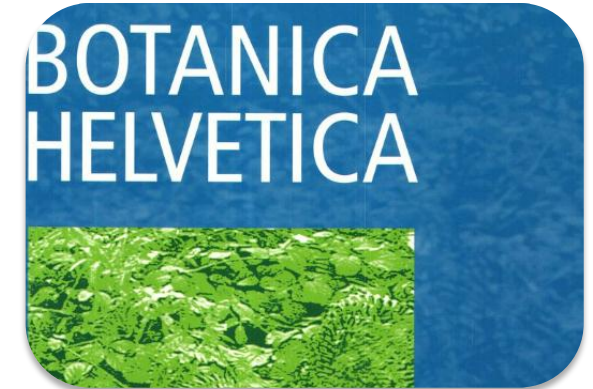**Journals**

e-periodica

~ 57,600 books
(share ETH Library:
~ 14,300 books)

Public Domain Mark

~ 385,000 images

Mostly Public Domain
Mark and CC BY-SA

~ 54,800 sources
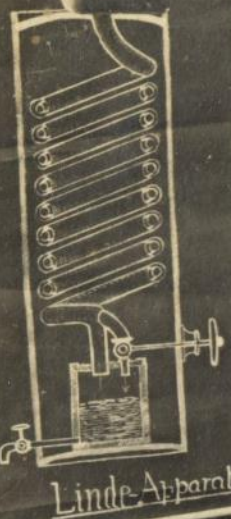(share ETH Library:
~ 13,400 sources)

Mostly Public Domain
Mark

~ 300 journals
(~ 600,000 articles)

How can we realise the full potential of digitized content?

Research partnerships

User participation

Outreach

Potential I
Research partnerships

Providing a machine readable corpus:
Enhancement of OCRed text

- **Available material: digitized journals as PDFs with OCR in e-periodica**
- **Additional service (aim)**
  - Provide access to the texts in a machine readable format
  - Text enrichment (named entity recognition)
- **Test case**
  - Two sample journals (~ 380,000 pages)
  - Tokenized text-files with OCR results
  - Aim I: Automatic correction of OCR errors
  - Aim II: Explore possibilities of automatic text enrichment (recognition and linking of the named entities [persons] and [place names])
- **Research partner**
  - Professor Martin Volk (University of Zurich, Institute of Computational Linguistics)

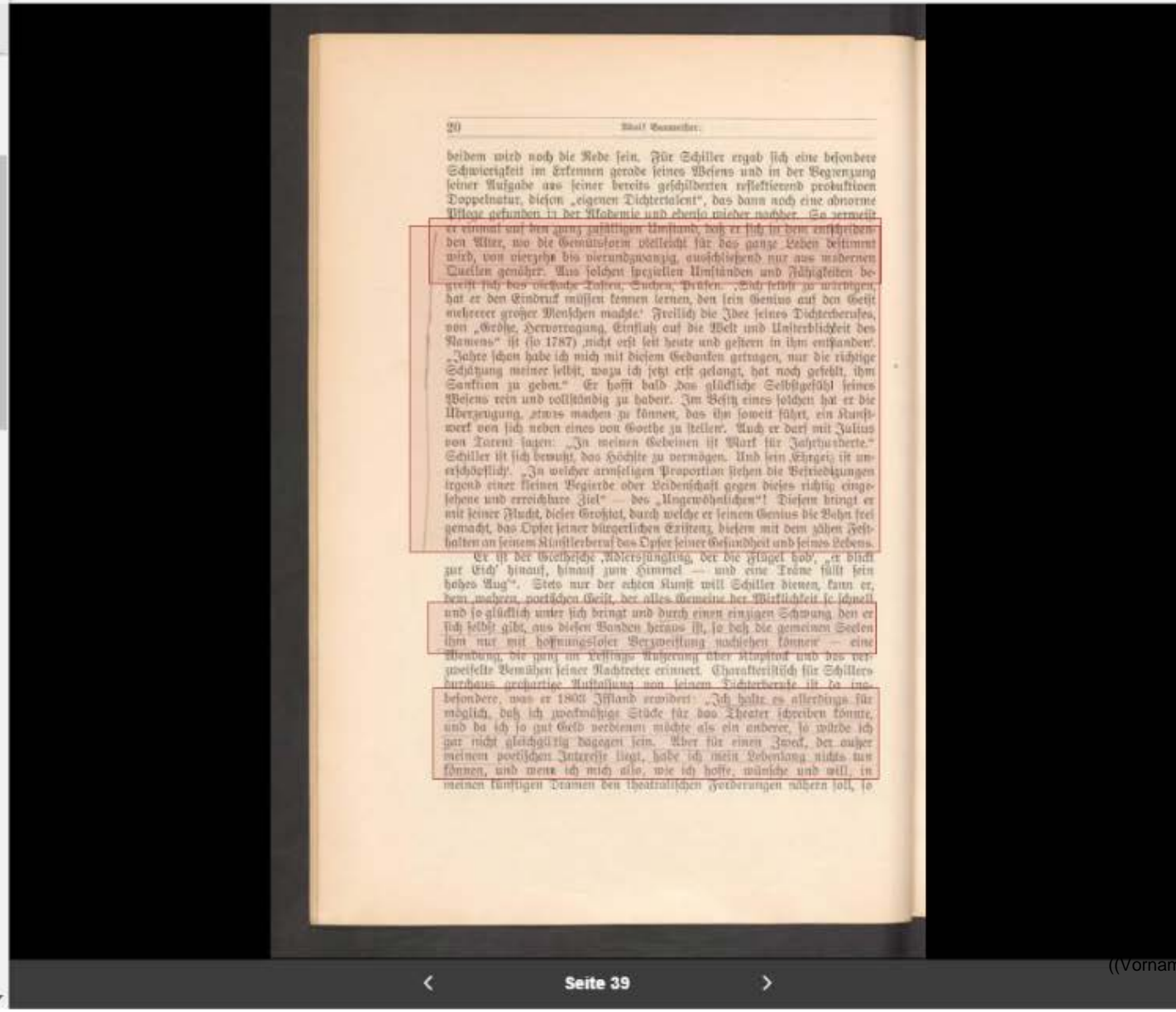# Providing a machine readable corpus: Enhancement of OCRed text

Building on existing solutions:
The "Thomas Mann personal library" project

- **Access to Thomas Mann's markings in his personal library**
  - Library is part of the Thomas Mann Archive, ETH Library
  - Integral digitization of 2,650 books containing handwritten traces
  - Transcription of the marginalia
  - Front-end for searching an displaying markings (building on e-periodica components)
  - Project duration: April 2016 to March 2019
- **Research partners**
  - Professor Andreas Kilcher (ETH Zurich, Chair of Literary and Cultural Studies)
  - Funded by Swiss National Science Foundation

Marks and annotations by Thomas Mann, dedication, owner mark, expropriation stamp. Compilation by Thomas Mann Archive

< Zurück zur Trefferliste < Zur Startseite
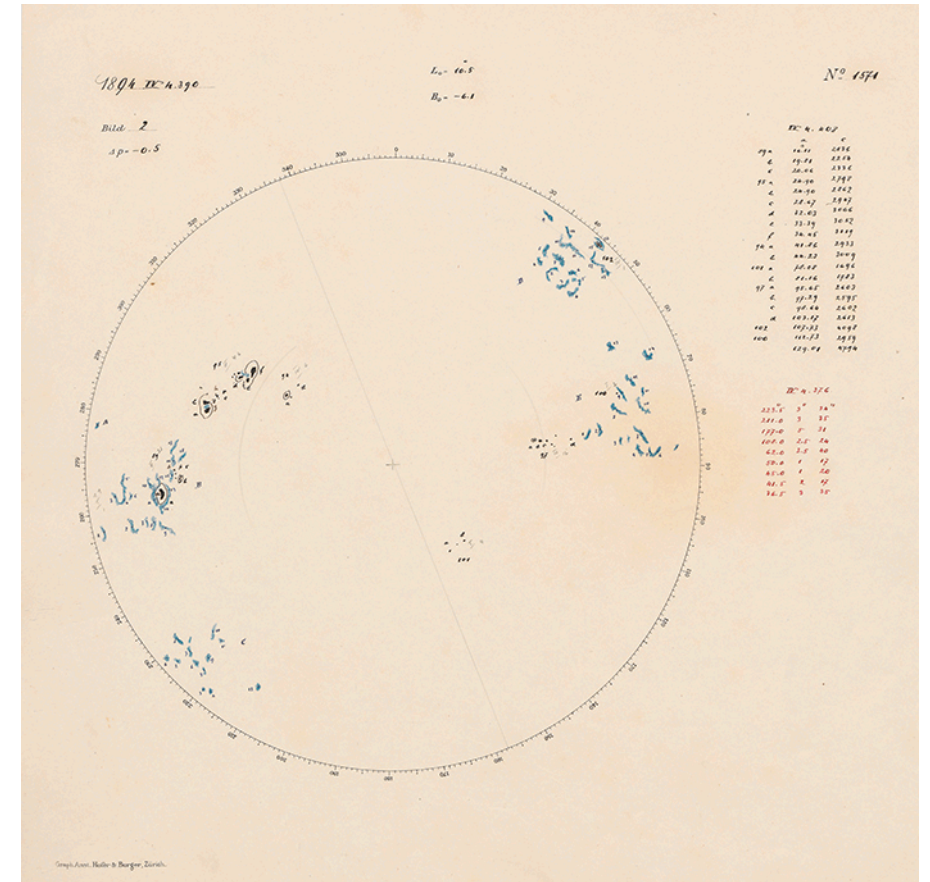
☑ Suchtreffer anzeigen ☑ Lesespuren anzeigen

Seitenminiaturen | Metadaten

i Volltext

Seite	34	35	36
37	38	39
40	41	42
43	44	45
46	47	48

ANNOTATIONTYPE **1**
Merkzeichen

SIGNTYPE
Unterstreichung

WRITINGUTENSILS
Bleistift

ANNOTATIONTYPE **2**
Merkzeichen

SIGNTYPE
Anstreichung

WRITINGUTENSILS
Bleistift

ANNOTATIONTYPE **3**
Merkzeichen

SIGNTYPE
Unterstreichung

WRITINGUTENSILS
Bleistift

ANNOTATIONTYPE **4**
Merkzeichen

SIGNTYPE
Unterstreichung

WRITINGUTENSILS
Bleistift

Providing primary data on demand:
Data extraction from analogue long-term measurement series

- **Seismograms for seismological research**
  - Series of paper seismograms from 1927 to 1955
  - Archival description, partial digitization and publication on e-manuscripa.ch
  - Partial data extraction by the Swiss Seismological Service at ETH Zurich
- **Sunspot drawings for solar research**
  - Sun observation at ETH Zurich from 1884 to 1995
  - Large series of sunspot drawings, data sheets, and sunspot statiscs at ETH Zurich University Archives
  - Archival description, digitization and publication on e-manuscripa.ch
  - Data extraction (years 1884–1949) by international solar researchers

Sunspot drawing, 5 April 1894, ETH University Archives, Hs 1304:1572

Potential II
User participation

Image Archive: Crowdsourcing to fill patchy image descriptions

Original title: Uniform FA 1960 to 1970
New: Swissair hostesses in the uniform worn from 1970 to 1977

- **Online crowdsourcing via E-pics Image Archive Online**
- **Aim: fill the gaps in contemporary descriptions**
- **Step 1: Cooperation with Swissair retirees**
  - Project duration: end of 2009 until end of 2013
  - Closed group of experts
  - 135 interested volunteers, including:
    44 regular users and ½ dozen experts
  - 40,000 images processed
- **Lessons learned**
  - Importance of personal contacts
  - Effort involved heavily underestimated
  - Added value gained: priceless!

Image Archive: Crowdsourcing to fill patchy image descriptions

Original title: Untitled, 17.8.1932
New: Flawil-Gossau, reservoir in Glatt-Tobel near the ruins of Helfenberg Castle, 17.8.1932

- **Step 2: All volunteers welcome**
  - New comment function online since 9 Dec 2015
  - Huge reader and media response (newspapers, radio, TV)
- **Statistics** (1 July 2017)
  - 22,600 pointers received
  - 20,543 description of images improved
  - 798 volunteers (90% men), very active top ten
- **Ongoing community management**
  - Weblog ETH Bibliothek Crowdsourcing
  - Videos and events

ETH zürich

Archival sources: Crowdsourcing to transcribe manuscripts

- **Project "e-manuscripta full text"**
  - Add transcription module to e-manuscripta.ch
  - Project duration: September 2016 to November 2017
  - Cooperative project, lead: Zentralbibliothek Zürich
  - Funded by swissuniversities
- **Functionalities**
  - Extension of existing Visual Library online editor used by Franckesche Stiftungen
  - Workflow: registration, transcription, publication
  - Open to the public – simple transcription

Potential III Outreach

Multichannel distribution of metadata and content:
Go where the users are

- **Spreading metadata**
  - locally: ETH Library's Knowledge Portal also includes data from archival information systems (e.g. Image Archive, ETH Zurich Archives) and other sources
  - internationally: Finding aids of ETH Zurich Archives harvested by Archives Portal Europe (250 mio. archival units from serveral hundred archives)
- **Spreading content**
  - Ongoing upload of digitized images to Wikimedia Commons (~ 40,000 images so far)
  - Images used to illustrate Wikipedia articles (e.g. «Mount Kilimanjaro» article with ~ 100,000 visits per month)
  - Specialized metrics tool in development (Wikimedia CH)

Presenting content in a new way:
Explora – the new storytelling platform

- **Presenting content in story form**
  - Research combined with storytelling
  - Visually attractive, using multimedia content
  - Change of perspective
- **E.g. scientific plant images in "[Mallow, Iris, Orchid]"**
  - Link the past (historical holidings) and the present (modern research, teaching, scientific visualization)
  - From Renaissance herbal books to 3D-modelling
  - Combining material from several collections and archives and
  - existing scans with new video testimonials
- **New stories are published regularly**

Summary and outlook

Potential	Summary	Outlook
Research partnerships	Solid basis: • Existing platforms with their underlying processes (indexing, scanning, presentation, archiving) • Open data policy with clearly labelled rights statements	• Further standardization regarding bulk access (e.g. RDF, IIIF) • Growing demand for data driven research • Individual projects demand individual know-how • More active role of the library
User participation	Crowdsourcing works, but ongoing community management is key	Use «collective intelligence» for further and more complex tasks (e.g. to georeference images)
Outreach	• Targeted spreading of metadata and content increases visability • Visually attractive platforms reach out to new users	• Integration of data into link hubs (e.g. Wikidata) • Intensify cooperation with experts to produce content

Thank you very much for your attention!

Michael Gasser
Head Archives, ETH Library
Rämistrasse 101, CH-8092 Zürich
Tel. +41 44 632 21 82
michael.gasser@library.ethz.ch

You can find us on:
- www.library.ethz.ch/en
- Facebook
- Twitter
- Google Plus