ETH zürich

Wenn Algorithmen Zeitschriften lesen. Vom Mehrwert automatisierter Textanreicherung

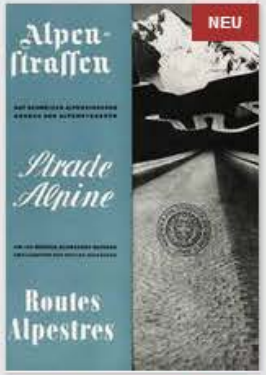
107. Deutscher Bibliothekartag: offen & vernetzt

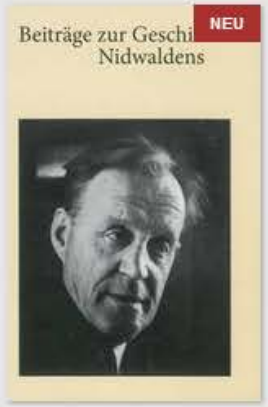
Berlin, 13. Juni 2018

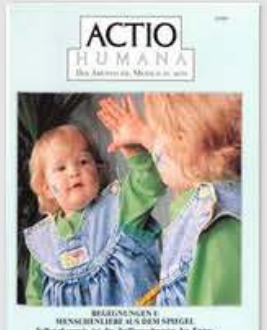
Regina Wanger, Michael Gasser

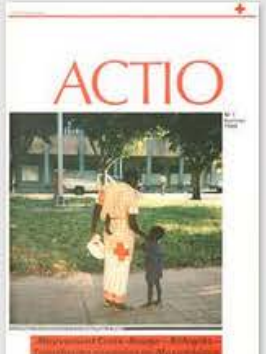
Schweizer Zeitschriften online. Revues suisses en ligne. Swiss journals online.

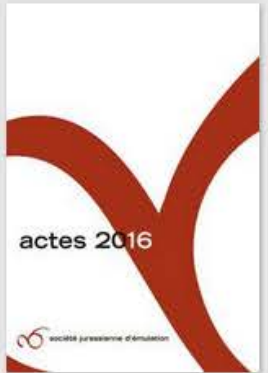
Ein Service der ETH-Bibliothek. Un service de ETH Library. A service by ETH Library.











E-Periodica

- Die Online-Plattform für digitalisierte Fachzeitschriften der Schweiz
  - Rund 300 Zeitschriften online
  - Rund 7 Mio. Seiten
- Hosting und Betrieb durch die ETH-Bibliothek
  - Laufender inhaltlicher Ausbau
  - Laufender funktionaler Ausbau
- Datenbestand
  - Master-TIFF, JPEG
  - Metadaten (XML)
  - Volltexte

Pilotprojekt zur automatisierten Textanreicherung

Dataset aus der Plattform E-Periodica

- Zwei Architektur-Zeitschriften
  - *Schweizerische Bauzeitung* (142 Jahrgänge, 1874–2016)
  - *Werk, Bauen und Wohnen* (102 Jahrgänge, 1914–2016)

} 380’000 Seiten

Forschungspartner

- Institut für Computerlinguistik der Universität Zürich: Prof. Martin Volk, Ismail Prada

Ziel: Mehrwert in drei Bereichen schaffen

1. Textkorpus automatisiert anreichern:
   Named Entity Recognition (NER) / Named Entity Linking (NEL)
2. Ausbau von Fachwissen an der ETH-Bibliothek
3. Zusatzfunktionen für Nutzerinnen und Nutzer von E-Periodica schaffen

NEUE AAREBRÜCKE IN THUN

01 Siegerprojekt: Das Beurteilungsgremium ist überzeugt, dass der Brückenentwurf von Bänziger Partner und Corinna Menn eine gute Eingliederung in den heterogenen Siedlungs- und Landschaftsraum erlaubt und eine sehr gute Grundlage für das künftige Bauprojekt bietet
(Visualisierung: Bänziger Partner, Chur)

Für die neue Aarebrücke zwischen dem kantonalen Entwicklungsschwerpunkt Thun Nord / Steffisburg und der Glättlimühli schrieb das kantonale Tiefbauamt Bern im Sommer 2008 einen Studienauftrag für vier Ingenieurbüros aus. Den Wettbewerb gewannen die Ingenieure von Bänziger Partner aus Chur zusammen mit der Architektin Corinna Menn.

des kantonalen Entwicklungsschwerpunkts Thun Nord / Steffisburg. Dabei muss das insgesamt 500 m lange Bauwerk bestehende Nutzungen und künftige Entwicklungen berücksichtigen.
Die Wahl der Jury fiel auf das Projekt von Bänziger Partner und Corinna Menn. Ihr Entwurf beruht auf konservativen, aber bewährten Prinzipien: Das Bauwerk ist in monolithischer Bauweise geplant, wodurch der Kräftefluss klar ablesbar ist. Die vorgespannte [illegible] vouteter Hohlkastenträger mit einer Hauptspannweite von 64 m. Wegen der variablen

13 in variablem Abstand angeordneten Pfeilern fort und lässt das Bauwerk als homogene Einheit erscheinen. Der Stützenquerschnitt verändert sich nur unmerklich je nach Belastung.
Die Brüstung, die zugleich als Leitschranke und Lärmschutzelement dient, schliesst die Brückenplatte links und rechts als einfaches, durchlaufendes Band ab. Die Jury kritisierte, dass die kandelaberartige Beleuchtung der Fussgängerzone [illegible] mittelt auf die [illegible] das Projekt bezüglich Unterhalt ausgezeichnete Noten: Die Zahl der Brückenlager ist auf

sbz-2009-135-999-09-z-0006.txt - Editor

Datei Bearbeiten Format Ansicht ?

```
dass 839,961,53,17
der 903,961,40,17
Brückenentwurf 955,960,196,18
von 1160,966,39,12
Bänziger 1212,960,105,23
Partner 1329,961,93,17
und 1433,961,41,17
Corinna 1486,960,93,18
Menn 1590,961,63,17
eine 1665,960,47,18
gute 1723,961,51,22
Eingliederung 1786,960,165,23
in 1961,959,20,19
den 112,997,40,16
heterogenen 164,997,149,22
Siedlungs- 325,996,130,23
und 467,997,41,16
Landschaftsraum 522,996,205,17
erlaubt 739,997,87,16
und 837,997,40,16
eine 889,996,47,17
sehr 947,997,54,16
gute 1011,997,51,22
Grundlage 1074,997,124,22
für 1208,996,36,17
das 1254,997,40,16
künftige 1306,996,94,23
Bauprojekt 1413,996,130,23
bietet 1554,996,68,17
(Visualisierung: 113,1026,185,26
Bänziger 312,1030,105,22
Partner, 429,1030,97,22
Chur) 539,1026,66,26
<EOS>
<EOP>
Für 114,1140,51,22
die 192,1140,43,22
neue 262,1146,72,16
Aarebrücke 359,1140,182,22
zwischen 568,1140,139,22
dem 113,1184,58,23
kantonalen 191,1184,164,23
Entwicklungsschwer- 374,1184,332,29
punkt 113,1229,83,29
Thun 208,1229,72,22
Nord/Steffisburg 298,1226,276,32
und 591,1229,52,22
der 651,1229,51,22
das 474,1273,51,22
kantonale 561,1273,147,22
```

Mehrwert 1: Automatisierte Textanreicherung

Vorgehen und Methode

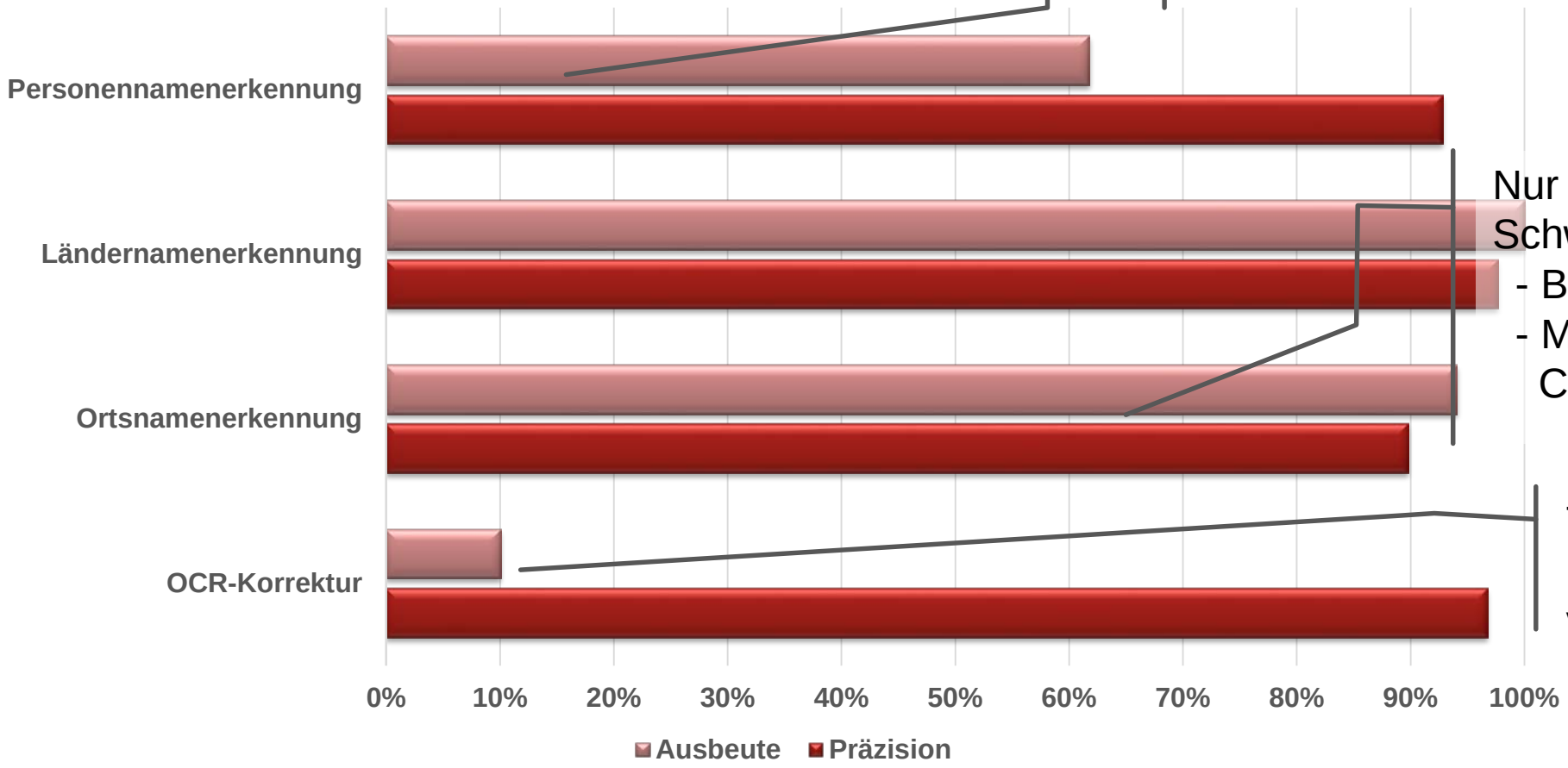
I. Automatische OCR-Korrektur
II. Orts- und Ländernamenerkennung
III. Personennamenerkennung
IV. Verlinkung mit GND (Gemeinsame Normdatei)

- Regel-basiertes Verfahren des Instituts für Computerlinguistik UZH (Python Skripte, u. a. Orts- und Namenlisten)
- Fokus auf Präzision (= sichere Erkennung)
- Goldstandard
  - Subset des Datenkorpus, bestehend aus 3 Jahrgängen (1895, 1940, 1990)
  - von Hand korrigiert und annotiert
- Vergleich mit einem statistischen System

Goldstandard Beispiel: Bauzeitung 1940
14’430 Zeilen (OCR-Textfile)
119 OCR-Fehler
269 Wörter in Personennamen
61 Ortsnamen (55 in der Schweiz)
12 Ländernamen

Automatische Erkennung I
(Bezug: Goldstandard)

Fokus auf Präzision zulasten der Ausbeute
Fehlerquellen z. B. Konstruktionen wie «Charles Aussage»



Nur Ortsnamen der Schweiz
Schwierigkeiten:
- Beznau: AKW oder Gemeinde?
- Montana: US-Bundesstaat oder CH-Gemeinde?

Tiefe Ausbeute v. a. wegen Beschränkung auf Mindestlänge von sechs Zeichen

Automatische Erkennung II
(Bezug: Goldstandard)

Verlinkung mit der GND

- Suche nach Übereinstimmung zu jedem erkannten Personennamen
- «Sicherheitswert» 1-5 anhand der aggregierten Merkmale Name, Vorname, Geburtsjahr, Geschlecht, Beruf (Sicherheit 5: alle Merkmale stimmen überein)
- Ausbeute mit hohen Sicherheitswerten relativ gering (mehrheitlich Verlinkung von «VIPs»)

Verlinkungen Bauzeitung Jhrg. 1940		
Sicherheit	**korrekte Verlinkungen**	**Total Verlinkungen**
Stufe 5	**92.3%**	**12**
Stufe 4	**95.0%**	**94**
Stufe 3	**60.0%**	**123**
Stufe 2	**60.0%**	**207**
Stufe 1	**65.0%**	**251**

Vergleich mit statistischem Verfahren

- System der Zürcher Hochschule für Angewandte Wissenschaft (ZHAW)
- Test: Erkennung von Personennamen auf Basis des Goldstandards
- Fazit: Höhere Ausbeute bei geringerer Präzision (mehr *false positives*)

Ausgabedateien

- XML-Dateien
- Aggregation der Informationen pro Zeitschrift und Jahrgang
- Bsp. Personenerkennung
  - Identifizierende Merkmale
  - GND mit «Sicherheitswert»
  - Referenzen mit Positionsangaben

```
− <person id="1069" gnd="118730266" gnd_certainty="4">
    <address/>
    <titles/>
    <firstname>R.|Robert</firstname>
    <lastname>Maillart</lastname>
    <gender>M</gender>
    <profession>Ing.|Ingenieur</profession>
  − <references>
   − <reference>
     − <positions>
       − <position>
           sbz-1940-000-115-00-r-0003.txt:285:1703,2026,26,22:1590
         </position>
       − <position>
           sbz-1940-000-115-00-r-0003.txt:285:1743,2025,104,21:1592
         </position>
       </positions>
     − <positions>
       − <position>
           sbz-1940-000-115-02-b-0026.txt:128:1544,1464,31,20:1372
         </position>
       − <position>
           sbz-1940-000-115-02-b-0026.txt:128:1585,1463,126,21:1374
         </position>
```

```
def rankCandidate(uniq_id, cand_dict, entry, eval_dict):
    """Compares the last best candidate to the new one and exchanges them
        if appropriate."""
    old = cand_dict[entry]

    #~ print(old)
    #~ print(eval_dict)
    # calculate certainty:
    certainty = 5
    # firstname
    if not eval_dict["flagFirstName"] and (eval_dict["flagAltFirstName"] or eval_dict["flagShortFirstName"] == "Multi"):
        certainty -= 1
    elif not eval_dict["flagFirstName"] and eval_dict["flagShortFirstName"]:
        certainty -= 2
    # profession
    if 1 < eval_dict["profValue"] < 1000:
        certainty -= 1
    elif eval_dict["profValue"] <= 1:
        certainty -= 2
    if old is not None:
        eval_dict["OnlyChoice"] = True # This was the only candidate found
    else:
        eval_dict["OnlyChoice"] = False
    eval_dict["certainty"] = certainty
    if old is None or certainty > old[1]:
        cand_dict[entry] = (uniq_id, certainty)
    elif certainty == old[1]:
        cand_dict[entry] = (uniq_id+"|"+old[0], certainty)
```

Mehrwert 2: Ausbau von Fachwissen

Ausbau des bibliotheksinternen Fachwissens

- Rights Clearance für Text- und Datamining Projekte
- Technischer Know-how-Transfer
  - Installation der Skripte und Python-Programme
- Verifizierung der Resultate des Testlaufs anhand weiterer Zeitschriften
  - Überprüfung des Verhältnisses zwischen Präzision und Ausbeute
  - Übertragbarkeit der Resultate auf andere Zeitschriften bzw. Themengebiete
- Aktueller Fokus auf OCR Post-Processing
- Anwendung der Eigennamenerkennung auf das gesamte Angebot von E-Periodica

Mehrwert 3: Zusatzfunktionen für Nutzerinnen und Nutzer

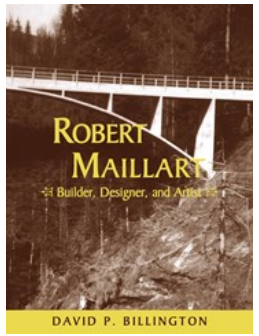
Zusatzfunktionen in E-Periodica

Wünsche von Nutzerinnen und Nutzern

- Kontinuierliche OCR-Optimierung
- Named Entity Recognition von Länder- und Ortsnamen als Basis für georeferenzierte Darstellungen
- Named Entity Linking von Personen als Basis für die Verlinkung …
  - … auf weitere Treffer innerhalb von E-Periodica
  - … auf weitere Ressourcen der ETH-Bibliothek und externe Datenquellen

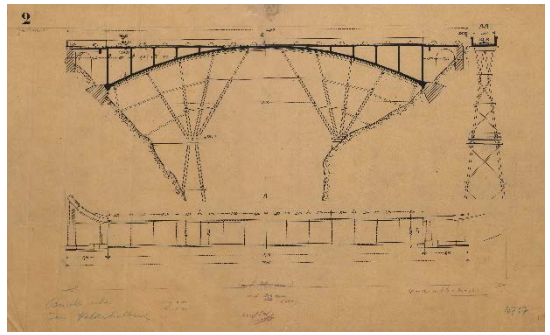
Robert Maillart (1872 – 1940)
Schweizer Bauingenieur

- Weitere Artikel in E-Periodica
- Weitere Ressourcen der ETH-Bibliothek



Bücher


Bilder


Archivalien

- Weitere Informationen
  - Wikipedia
  - Historisches Lexikon der Schweiz
  - Deutsche Biographie
  - Deutsche Digitale Bibliothek
  - Wikimedia Commons

Zusatzfunktionen in E-Periodica

Wünsche von Nutzerinnen und Nutzern

- Kontinuierliche OCR-Optimierung
- Named Entity Linking von Länder- und Ortsnamen als Basis für georeferenzierte Darstellungen
- Named Entity Linking von Personen als Basis für die Verlinkung …
  - … auf weitere Treffer innerhalb von E-Periodica
  - … auf weitere Ressourcen der ETH-Bibliothek und externe Datenquellen

Mögliche Forschungsdesiderate

- Erweiterung des Systems auf Texte in anderen Sprachen (F, I, gemischt)
- Erkennung und Verlinkungen weiterer Entitäten, z. B. Gebäude, geografische Bezeichnungen (Berge, Täler, Seen) oder Firmennamen, Organisationen etc.

Zusammenfassung

- Online Plattform E-Periodica mit 7 Mio. Seiten
- Pilotprojekt zur automatisierten Textanreicherung
  - Eigennamenerkennung in zwei Architektur-Zeitschriften
    - Präzision vor Ausbeute (zentrale Herausforderung für den praktischen Einsatz)
    - Low hanging fruits: Erkennung von Ländernamen und Ortsnamen der Schweiz
    - Gute Resultate: Personennamenerkennung
    - Klare Grenzen der sicheren Identifizierung und GND-Verlinkung: Named Entity Linking
- Know-how-Transfer und laufende Optimierungen
  - OCR Post-Processing als wichtige Zusatzaufgabe im Betrieb
  - Ausweitung des Systems auf weitere Themen / Zeitschriften in E-Periodica
  - Entwicklung von Zusatzfunktionen (als Prototypen) im direkten Austausch mit Usern
- Weitere Forschungsfragen

Fragen ?

Vielen Dank für Ihr Interesse!

ETH-Bibliothek, Regina Wanger, Leitung DigiCenter
Tel. +41 44 632 69 10, regina.wanger@library.ethz.ch

ETH-Bibliothek, Michael Gasser, Leitung Archive
Tel. +41 44 632 21 82, michael.gasser@library.ethz.ch
Twitter handle: @M_Gasser