know-how

innovation

Mögliche zukünftige Formen der Sacherschliessung

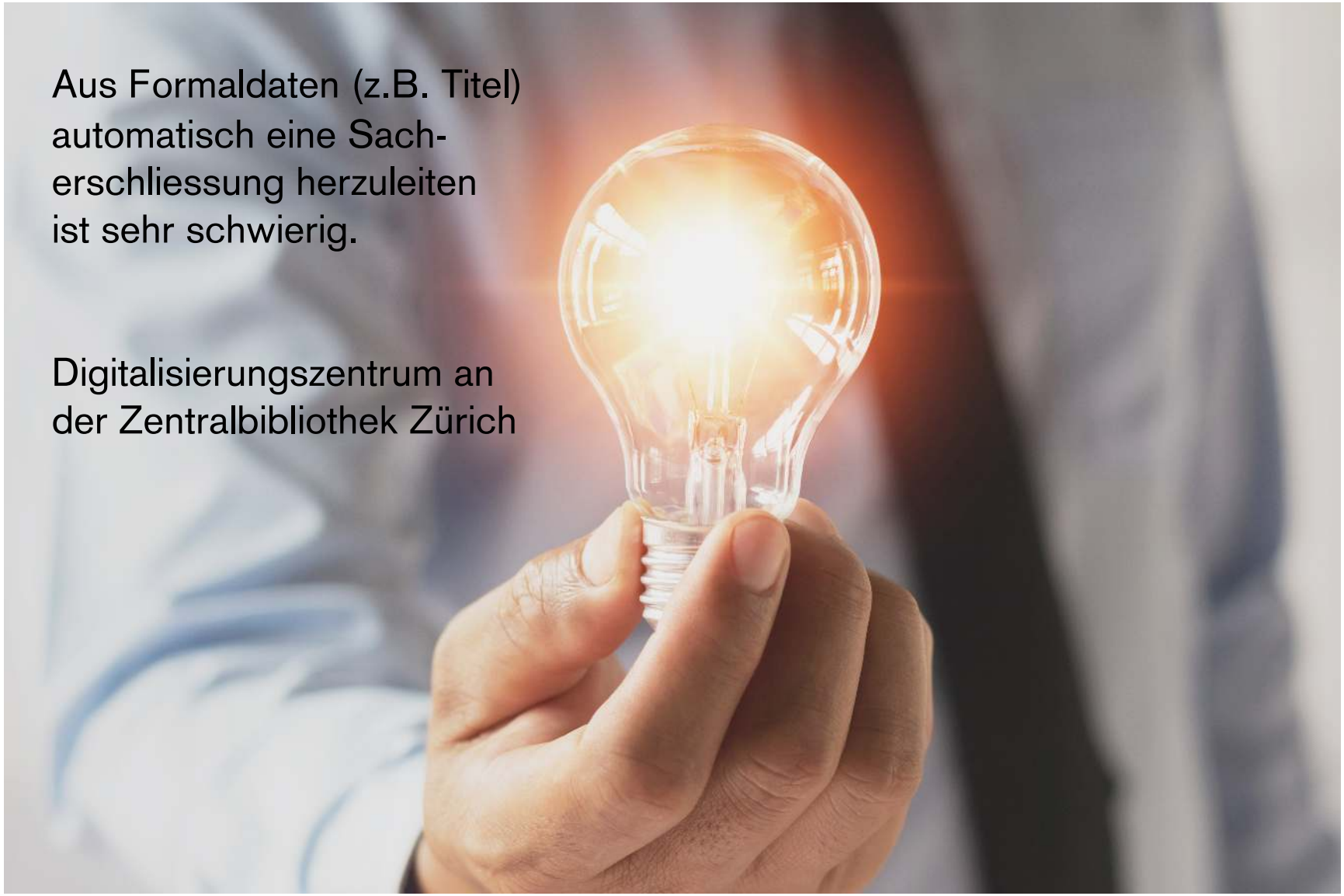
Dritter Stuttgarter Workshop Computerunterstützte Inhaltserschließung

7.11.2019, Peter Schäuble

eurospider
relevancy retrieval

solution

Idee der automatisierten Sacherschliessung vor 10 Jahren am 8.5.2009



Aus Formaldaten (z.B. Titel) automatisch eine Sach-erschliessung herzuleiten ist sehr schwierig.

Digitalisierungszentrum an der Zentralbibliothek Zürich

Vermutete Erfolgs-
faktoren, welche die
Sacherschliessung
weiterbringen
Erschliessungssystem
Trainingsdaten
Organisationsformen

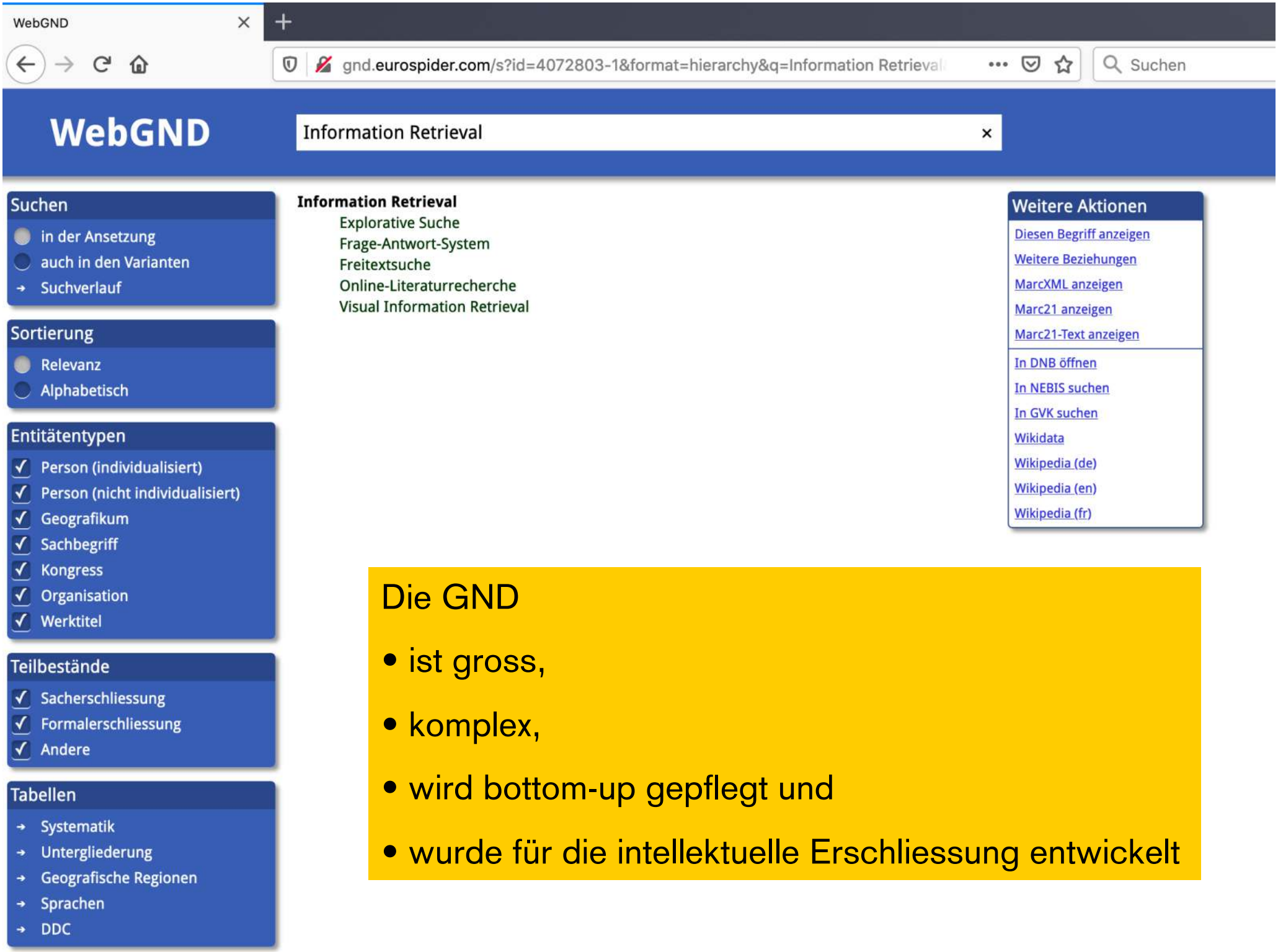
Ein neues Erschliessungssystem?

Geeignetes Erschließungssystem, welches

* sich für die automatische Erschliessung eignet und

* die Suchqualität von Discovery-Systemen verbessert.

© 2019 Eurospider Information Technology AG

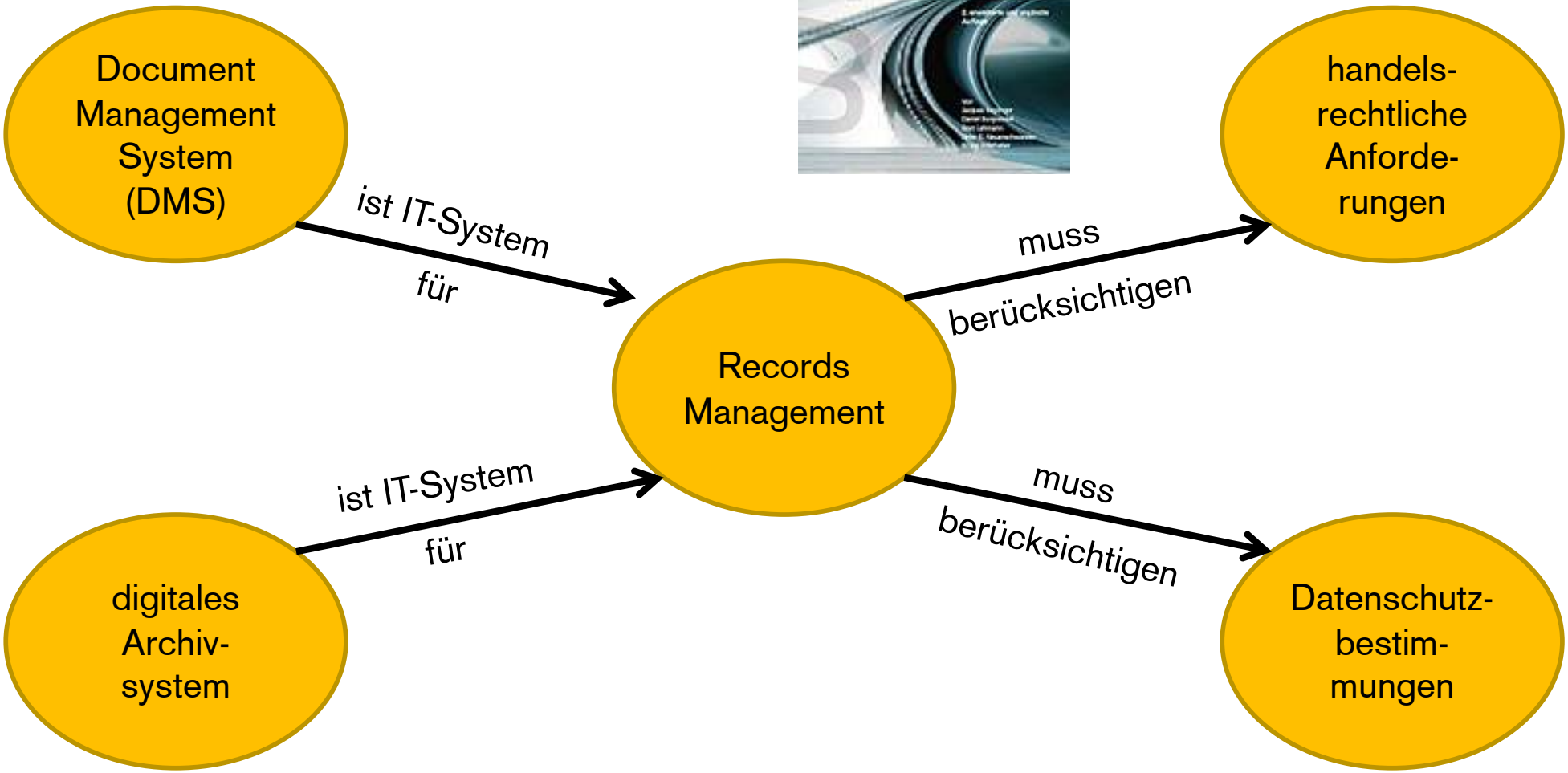


Die GND

- ist gross,
- komplex,
- wird bottom-up gepflegt und
- wurde für die intellektuelle Erschliessung entwickelt

Resource Description Framework (RDF) mit S-P-O

Die vielen Freiheitsgrade und die hohe Komplexität erinnern an das *5th Generation Project* in Japan (1982) und an das amerikanische Gegenprojekt CYC



https://en.wikipedia.org/wiki/Fifth_generation_computer

https://en.wikipedia.org/wiki/Cyc

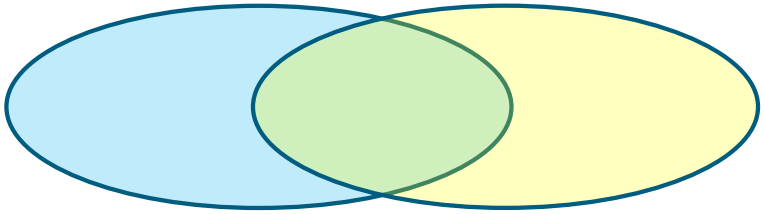
DDC Kurznotationen als neues Erschliessungssystem?

DDC Kurznotationen (DNB Jahresbericht 2018):

- 2018 sind fast 32.000 Monografien und 251.000 Artikel mit DDC Kurznotationen erschlossen worden
- Die klassifikatorische Erschließung wird jetzt um DDC Kurznotationen erweitert, also ausgewählte DDC-Notationen, die differenzierter sind als die DDC-Sachgruppen und gewisse statistische Eigenschaften haben (Dokumentenhäufigkeit)

Genügen ungeordnete Mengen mit Notationen ohne Operatoren?

- Dokumentenmanagementsysteme um handelsrechtliche Anforderungen zu erfüllen



Hohe Dokumentenhäufigkeiten alleine garantieren weder Erschliessungs- noch Retrieval-Qualität

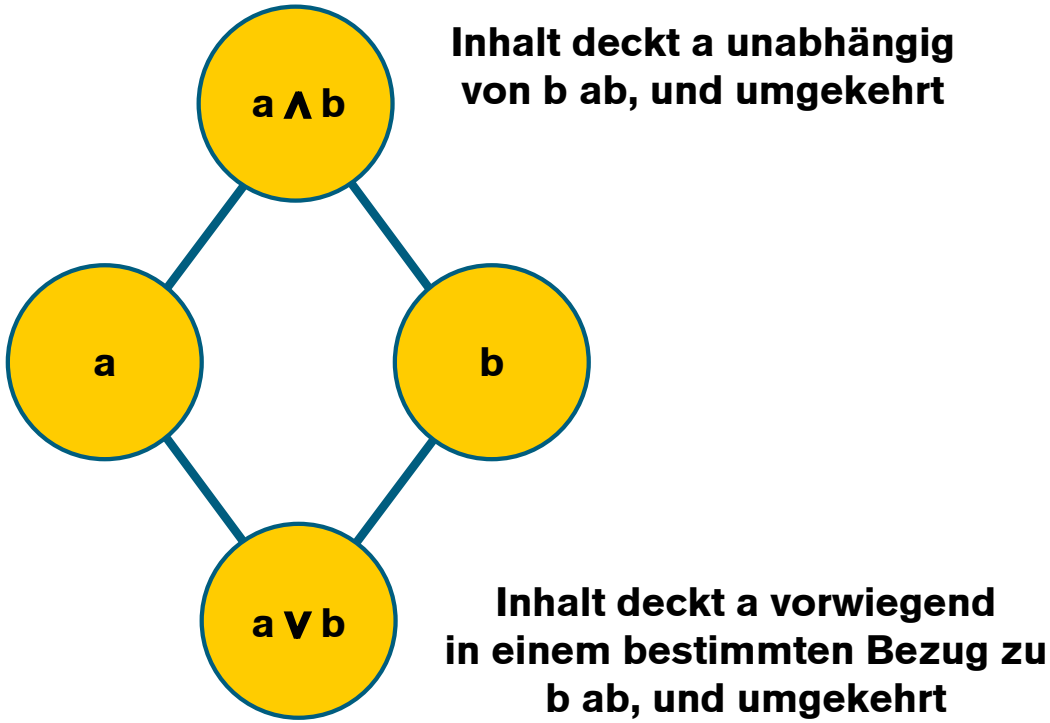
- Statistik und Algebra für Ingenieure

Verbände (Wikipedia)

Ein Verband ist in der Mathematik eine Struktur, die sowohl als Ordnungsstruktur als auch als algebraische Struktur vollständig beschrieben werden kann.

Als Ordnungsstruktur ist ein Verband dadurch gekennzeichnet, dass es zu je zwei Elementen a und b ein **Supremum a ∨ b** gibt, d. h. ein eindeutig bestimmtes kleinstes Element, das größer oder gleich a und b ist,

und umgekehrt ein **Infimum a ∧ b**, ein größtes Element, das kleiner oder gleich a und b ist.



Yang, Yiming (1997)

7 Conclusions

The following conclusions are reached from this study:

1. The performance of a classifier depends strongly on the choice of data used for evaluation. Using a seriously problematic collection[8], omparing categorization methods without analyzing collection differences[1], and drawing conclusion bas n the results of flawed experiments[2] raise questions about the validity of some published evaluation se problems need to be addressed to clarify of the confusions among researchers, and to prevent th on of similar mistakes. Providing information and analysis on these problems is a major effort in

aluations into a in
e baseline classi nce
baseline classifi nce
iers. Such an in ich
involving two o ogy
t to standardiz

top perform ally
had a relativ ods
outperformed WORD, the non-learning m are
not as large as previously claimed[1, 2 are
better than linear models, or that m sticated ents
on the strengths and weaknesses nferent models requires further research.

4. Scalability of a classifier when the problem size grows by several magnitudes, or when the category space becomes a hundred times denser, has been rarely examined in text categorization evaluations. KNN is the only learning method evaluated on the full set of the OHSUMED categories. Its robustness in scaling up and dealing with harder problems, and its computational efficiency make it the method of choice for approaching very large and noisy categorization problems.

The Performance of a classifier depends strongly on the choice of data used for evaluation.

Scalability of a classifier when the problem size grows by several magnitudes, or when the category space becomes hundred times denser, has been rarely examined in text catgeorization evaluation.

significance

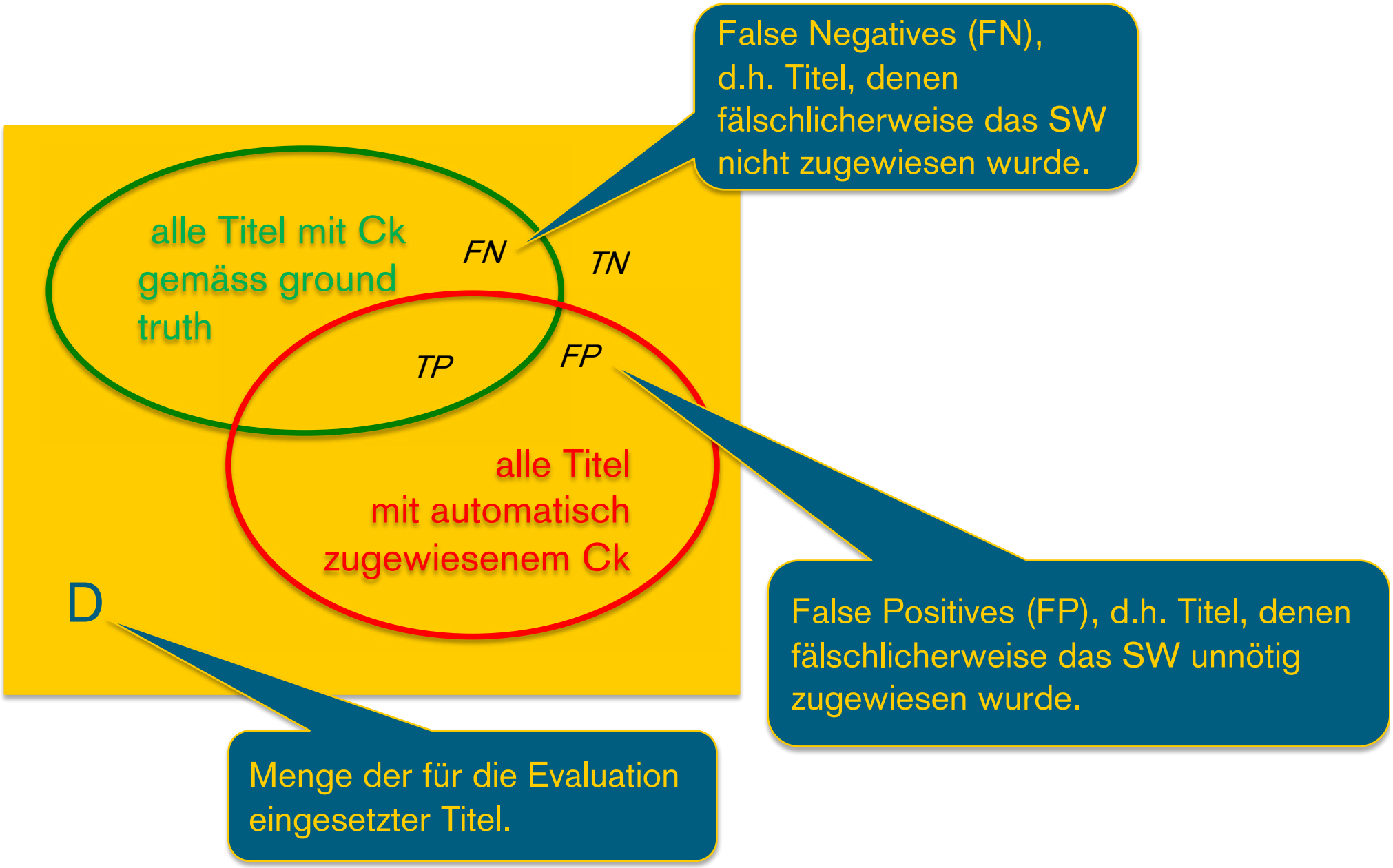
excellence

Trainings- und Testkollektion für automatische Kategorisierung

technology

eurospider
relevancy retrieval

Evaluation automatischer Zuweisungen von Ck (SW, Notation, etc.)

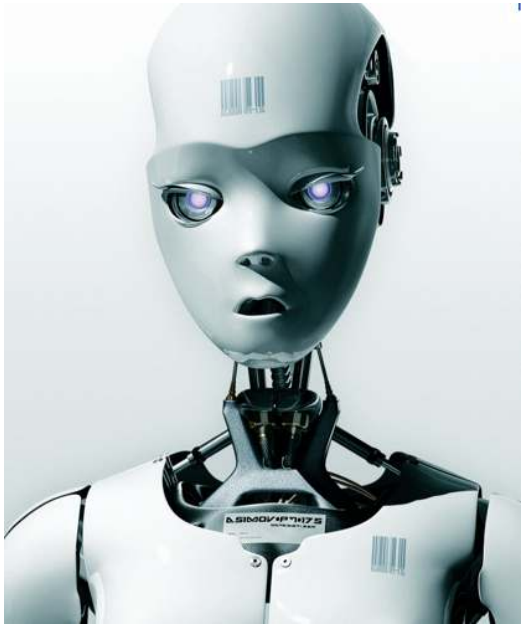


Maschinelles Lernen

Merkmalsvektor

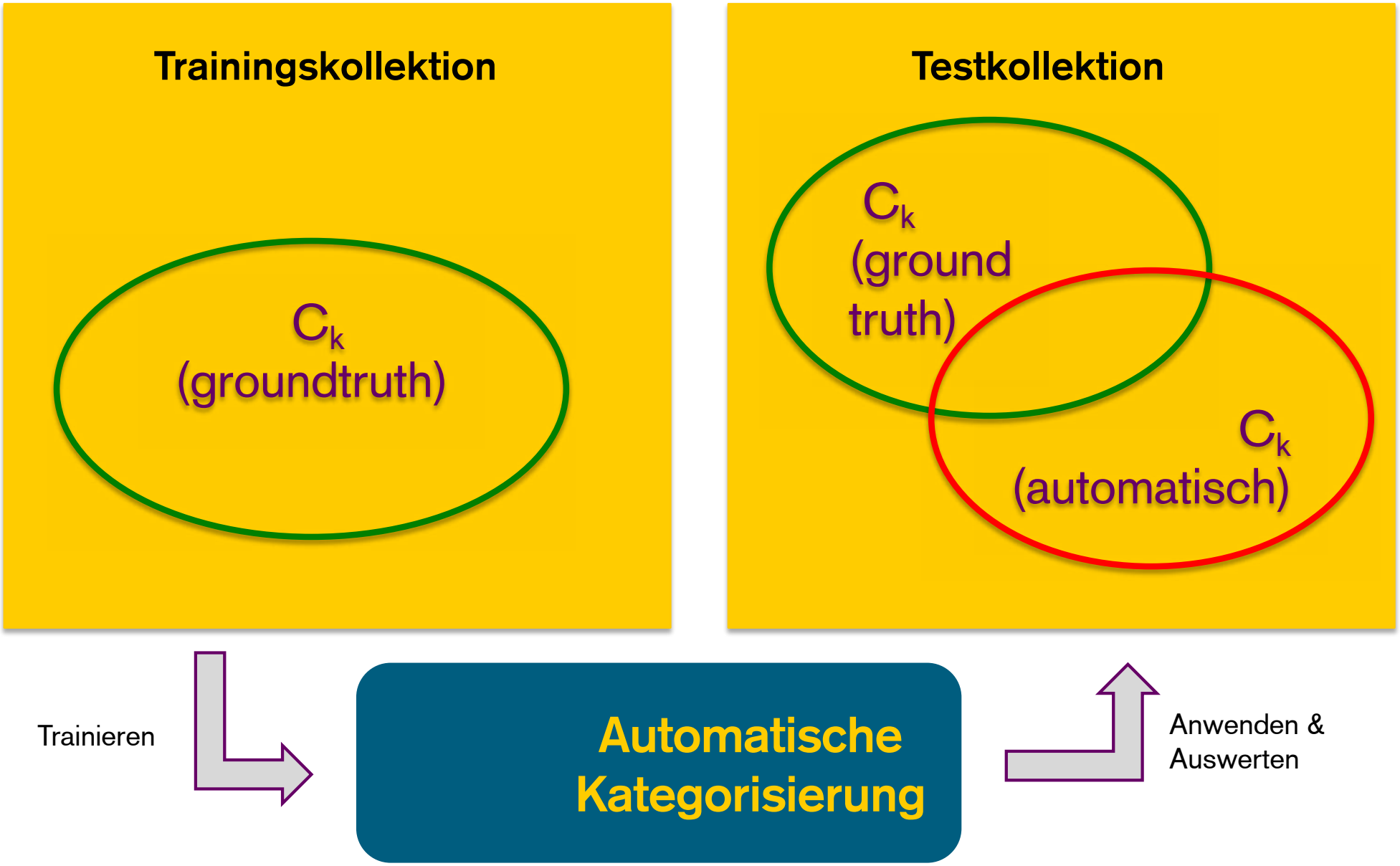
Notation

Merkmalsvektor



Notation

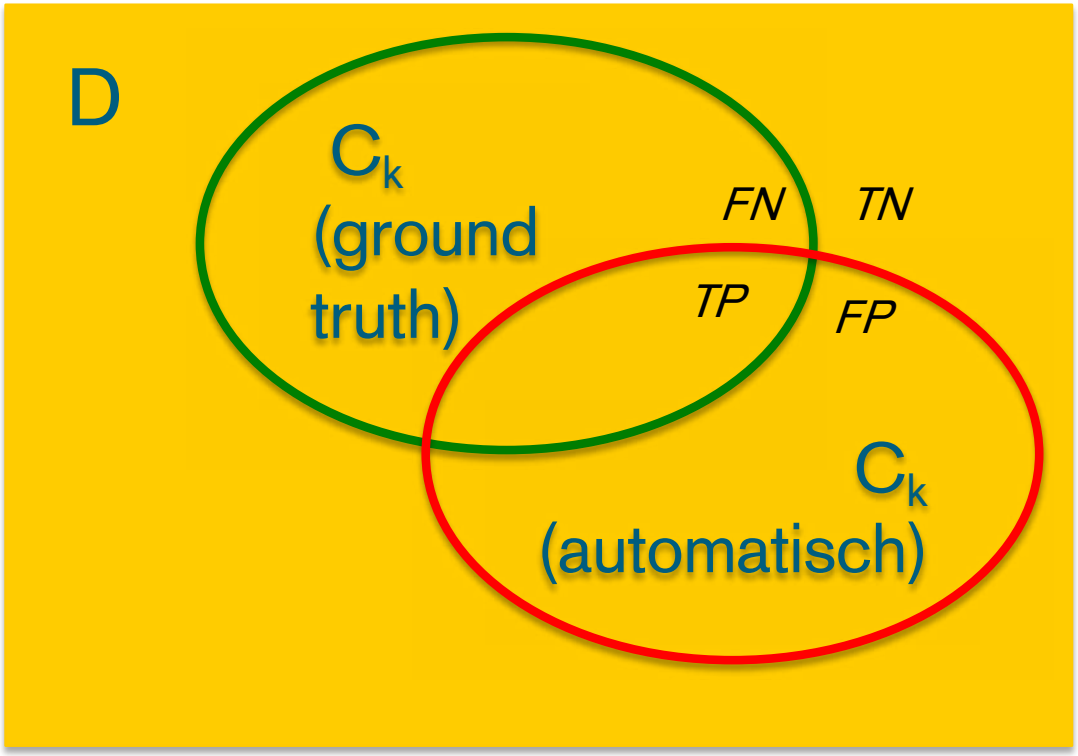
Test- und Trainingskollektion



Qualität der Kategorisierung ermitteln

Die Kategorisierungsqualität kann mit Ausbeute, Präzision, F1, etc. quantifiziert werden.

- Ausbeute (ρ= recall): Wieviele relevante wurden automatisch gefunden?
  $\rho = TP / (TP + FN)$
- Präzision (π = precision): Wieviele der automatisch gefundenen sind relevant?
  $\pi = TP / (TP + FP)$
- F1: Harmonisches Mittel
  $F1 = 2 * \rho * \pi / (\rho + \pi)$
  $F1 = 2 / (1/\rho + 1/\pi)$



Automatische Sachgruppenvergabe Ergebnisse Testfall II/3

DEUTSCHE NATIONAL BIBLIOTHEK

	Syst. A	System B	System C	System D
F-Measure$_1$*	0,75	0,68	0,71	0,65
Recall$_1$*	0,75	0,70	0,68	0,63
Precision$_1$*	0.77	0.70	0.76	0.86

	Syst. A	System B	System C	System D
F-Measure$_2$*	0,61	0,54	0,58	0,55
Recall$_2$*	0,88	0,83	0,85	0,79
Precision$_2$*	0,50	0,43	0,46	0,45

	Syst. A	System B	System C	System D
F-Measure$_3$*	0,54	0,48	0,53	0,51
Recall$_3$*	0,91	0,87	0,90	0,86
Precision$_3$*	0,43	0,37	0,41	0,40

*Gewichtetes Mittel über alle SG

know-how

innovation

Retrieval Effektivität

Wird relevante Information effektiv gefunden?

solution

Internationale Evaluationsforen: TREC, CLEF, NTCIR

Text REtrieval Conference (TREC) Home Page - Microso...

http://trec.nist.gov/

Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

Overview
Other Evaluations
Publications
Information for Active Participants
Frequent Asked Questions
Tracks
Data
Past TREC Results
Contact Information

The TREC Conference series is co-sponsored by the NIST, Information Technology Laboratory's (ITL) Retrieval Group of the Information Access Division (IAD) and the Advanced Research and Development Activity (ARDA) of the U.S. Department of Defense

NIST
National Institute of Standards and Technology
is an agency of the U.S. Commerce Department's Technology Administration

Last updated: Friday, 10-Sep-04 11:45:25
Date created: Tuesday, 01-Aug-00
privacy policy / security notice / accessibility statement
disclaimer / FOIA
trec@nist.gov

Welcome to Cross Language Evaluation Forum - Microsoft Internet Explorer

http://clef.isti.cnr.it/

Cross Language Evaluation Forum

Home
Coordination
CLEF 2005
CLEF 2004
CLEF 2003
CLEF 2002
CLEF 2001
CLEF 2000
Publications
Links
Archives
Contact
Steering Committee
Deutsch

The Cross-Language Evaluation Forum (CLEF) supports global digital library applications by (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes.

Through the organisation of system evaluation campaigns, the aim is to create a community of researchers and developers studying the same problems and to facilitate future collaborative initiatives between groups with similar interests. CLEF will also establish strong links, exchanging ideas and sharing results, with similar cross-language evaluation initiatives in the US and Asia, working on other sets of languages. The final goal is to assist and stimulate the development of European cross-language retrieval systems in order to guarantee their competitiveness on the global marketplace.

The CLEF 2004 Evaluation Campaign has now concluded.
The results were presented at the CLEF 2004 Workshop
15-17 September, Bath, UK.
The Working Notes are online.

Information on the CLEF 2005 campaign will be made available shortly.

To be included on the CLEF mailing list and for further information, contact:
Carol Peters (carol.peters@isti.cnr.it)

CLEF 2004 is an activity of the DELOS Network of Excellence for Digital Libraries under the Sixth Framework Programme of the European Commission

Webdesign by Centromedia.com

http://clef.isti.cnr.it/2004/working_notes/CLEF2004WN-Contents.html

NTCIR Workshop - Microsoft Internet Explorer

http://research.nii.ac.jp/ntcir/workshop/

National Institute of Informatics

Workshop Home
The 4th NTCIR Workshop (2003/2004)
Evaluation of Information Access Technologies:
Information Retrieval, Question Answering, and Summarization
March 2003 - June 2004

The 4th NTCIR Workshop Meeting: June 2-4, 2004 NII Tokyo

[Japanese] [NTCIR Home]

The Online Registration for the 4th NTCIR Workshop has started!. Participation is invited from anyone interested in research on information access technologies, such as retrieval from large-scale test collections various genres, cross-lingual information retrieval of Asian languages, question answering and text summarization of Japanese texts. NTCIR Workshops are periodical event which are held once par one and half years.

- TASK DESCRIPTION
- TASK INFORMATION: CLIR - PATENT - QAC - TSC - WEB
- DATA (NTCIR-4 Test Collections)
- HOW TO PARTICIPATE
- IMPORTANT DATE NEW!
- How to prepare formal proceedings NEW!
- How to prepare working notes
- USER AGREEMENT FORMS
- NTCIR WORKSHOP 4 MEETING

CLEF

Nicht angemeldet Diskussionsseite Beiträge Benutzerkonto erstellen Anmelden

WIKIPEDIA
Die freie Enzyklopädie

Artikel Diskussion Lesen Bearbeiten Quelltext bearbeiten Mehr Wikipedia durchsuchen

Hauptseite
Themenportale
Zufälliger Artikel

Mitmachen

Artikel verbessern
Neuen Artikel anlegen
Autorenportal
Hilfe
Letzte Änderungen
Kontakt
Spenden

Werkzeuge

Links auf diese Seite
Änderungen an verlinkten Seiten
Spezialseiten

Cross-Language Evaluation Forum

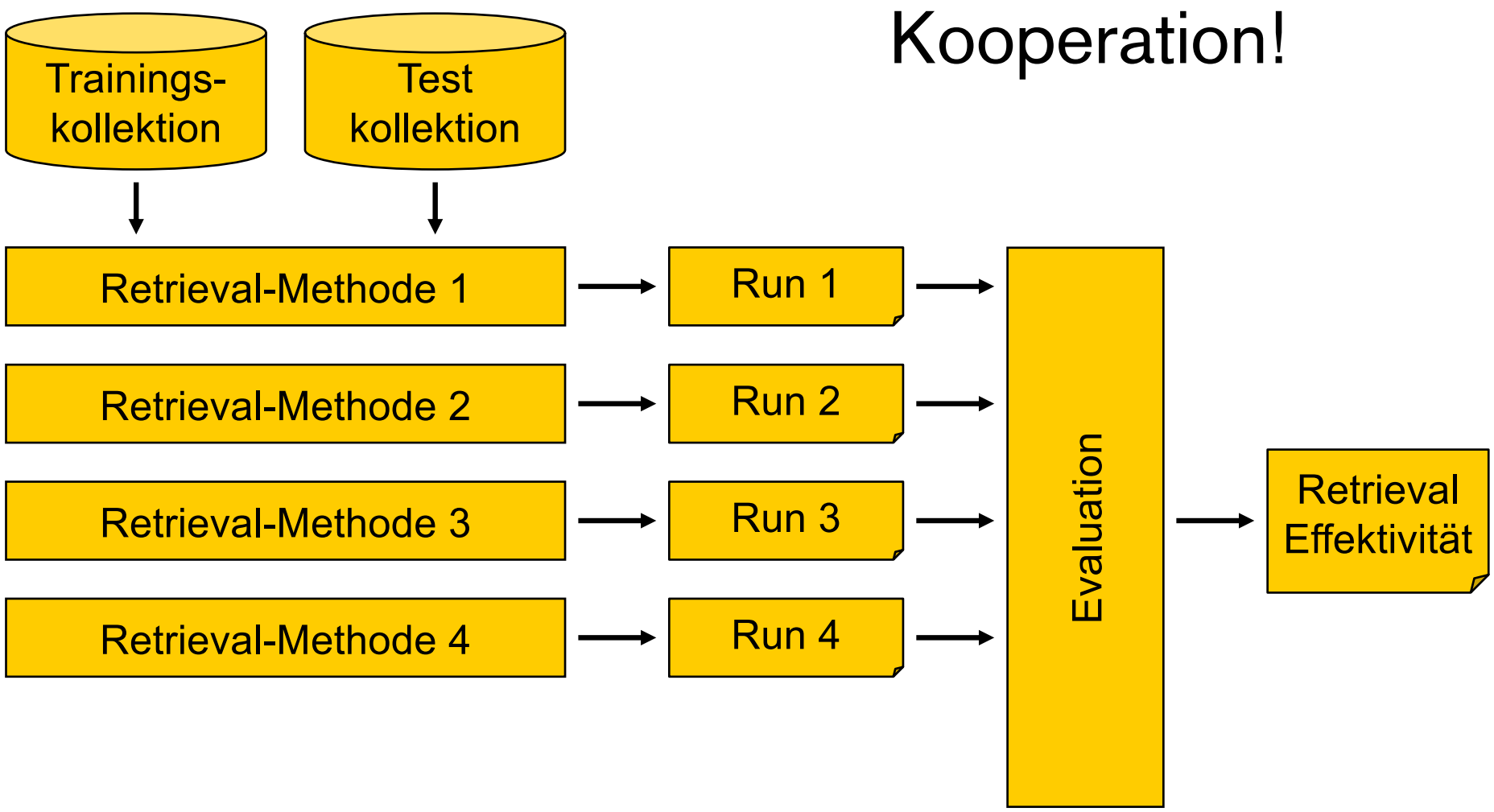
Das **Cross-Language Evaluation Forum** (*kurz*: **CLEF**) ist aus der TREC-Aufgabe Cross-Language Information Retrieval (CLIR) entstanden, welches sich hauptsächlich mit dem Cross-Language Information Retrieval europäischer Sprachen befasste.

CLEF ist mittlerweile ein eigenständiges EU-Projekt und bietet eine Plattform zur Evaluierung und Verbesserung von Information-Retrieval-Systemen für europäische Sprachen.

Die seit 2000 von CLEF jährlich organisierten System-Evaluations-Kampagnen sollen die Zusammenarbeit von Forschern und Entwicklern fördern und somit zukünftige Initiativen zur Zusammenarbeit von Gruppen mit ähnlichen Interessen vereinfachen und fördern. Es geht hierbei darum, Nutzeranfragen, die in einer beliebigen europäischen Sprache gestellt werden, in beliebigsprachigen Dokumentmengen abzuarbeiten und eine nach Relevanz geordnete Ergebnismenge zu erhalten, die auf diese Frage eine Antwort darstellt. Auch einsprachiges Information Retrieval stellt einen Schwerpunkt der Evaluierung dar, ist jedoch vor allem für Teams vorgesehen, die das erste Mal an der Kampagne teilnehmen. Es bestehen auch Kooperationen mit ähnlichen, anderssprachigen Initiativen aus den USA und Asien.

Das eigentliche Ziel ist, die Entwicklung der europäischen Cross-Language Retrieval Systeme zu unterstützen und anzuregen, damit ihre Wettbewerbsfähigkeit auf dem Weltmarkt gesichert ist.

Wie funktioniert ein Evaluationsforum?



significance

excellence

Zusammenfassung

technology

eurospider
relevancy retrieval

Sacherschliessung in 10 Jahren

Aufgrund der gemachten Erfahrungen bei der Entwicklung des Digitalen Assistenten und den sich stets weiter entwickelnden Anforderungswünschen an das System haben wir in einigen Gesprächen überlegt, wie wir insgesamt die Situation, um bessere Ergebnisse bei der automatisierten Inhaltserschliessung zu bekommen, verbessern können.

Wir haben dabei zunächst drei Erfolgsfaktoren identifiziert.

1. Innovative Organisationsformen, um Skaleneffekte zu erreichen (DINI Vortrag)
2. geeignetes Erschließungssystem, welches
   - für die automatische Erschliessung geeignet ist (weniger komplex als die GND) und
   - die Suchqualität von Discovery-Systemen signifikant verbessert (also detaillierter als die DNB Sachgruppen ist)
3. qualitativ und quantitativ geeignete Trainingsdaten, um die automatische Erschliessung und die Suchqualität von Discovery-Systemen zu evaluieren.

Tradition ermöglicht erst und hemmt dann wieder die Entwickelung der Gedanken

A. Einstein 1924.

Über die spezielle und die allgemeine Relativitätstheorie

(Gemeinverständlich)

Von

A. EINSTEIN