Universität Bielefeld
Institute for Interdisciplinary Studies of Science (I²SoS)

competence centre for bibliometrics

Institutionen- und Geokodierung als Instrumente zur Qualitätssicherung in der Bibliometrie

Christine Rimmert

Forum Bibliometrie, TUM

Bibliometrie

Grundlegende Fragestellungen

- **Anzahl Publikationen**
  Verlauf über Publikationsjahre, Vergleich von Ländern, Regionen, Institutionen, Autoren...
- **Anzahl Zitationen**
  → hochzitierte Publikationen in bestimmten Ländern, Regionen, Institutionen, von welchen Autoren?
- **Kooperationen**
  Länder, Institutionen, Autoren....

→ institutionelle & geografische Zuordnungen von Publikationen erforderlich!
→ Autorenadressen

Adressen

Beispiele aus dem Web of Science

Univ Cologne, Inst Phys, Zulpicher Str 77, D-50937 Cologne, Germany Univ Wurzburg, Inst Theoret Phys & Astrophys, D-97074 Wurzburg, Germany Commiss European Communities, Joint Res Ctr, Inst Transuranium Elements, D-76125 Karlsruhe, Germany Charite, Inst Radiol, D-10117 Berlin, Germany Karlsruhe Inst Technol, Inst Expt Kernphys, D-76131 Karlsruhe, Germany Max Planck Inst Kernphys, D-69117 Heidelberg, Germany Univ Munster, Dept Neurol, D-48149 Munster, Germany Univ Marburg, Inst Pharmazeut Biol & Biotechnol, Deutschhausstr 17A, D-35037 Marburg, Germany Thuringer Landessternwarte, D-07778 Tautenburg, Germany EMBL Hamburg DESY, Bldg 25A,Notkestr 85, D-22603 Hamburg, Germany **Ruhr Univ Bochum, Tech Univ Crete, Istanbul Tech Univ, Univ Hamburg, GeoForschungsZentrum Potsdam, Bochum, Germany** Univ Duisburg Essen, Dept Neurol, D-45122 Essen, Germany Tech Univ Munich, Chair Analyt Food Chem, Freising Weihenstephan, Germany Univ Hosp, Heart Ctr Bonn, Dept Med 2, Bonn, Germany Univ Saarland, Dept Pulmonol, D-66421 Homburg Saar, Germany Univ Vet Med Hannover, Dept Pharmacol Toxicol & Pharm, Buenteweg 17, D-30559 Hannover, Germany Bavarian Acad Sci, Commiss Geodesy & Glaciol, Alfons Goppel Str 2, D-80539 Munich, Germany Univ Bayreuth, Dept Anim Physiol, POB 101251, D-95440 Bayreuth, Germany DESY, Hamburg, Germany Max Delbruck Ctr MDC Mol Med, Berlin, Germany IFW Dresden, D-01069 Dresden, Germany

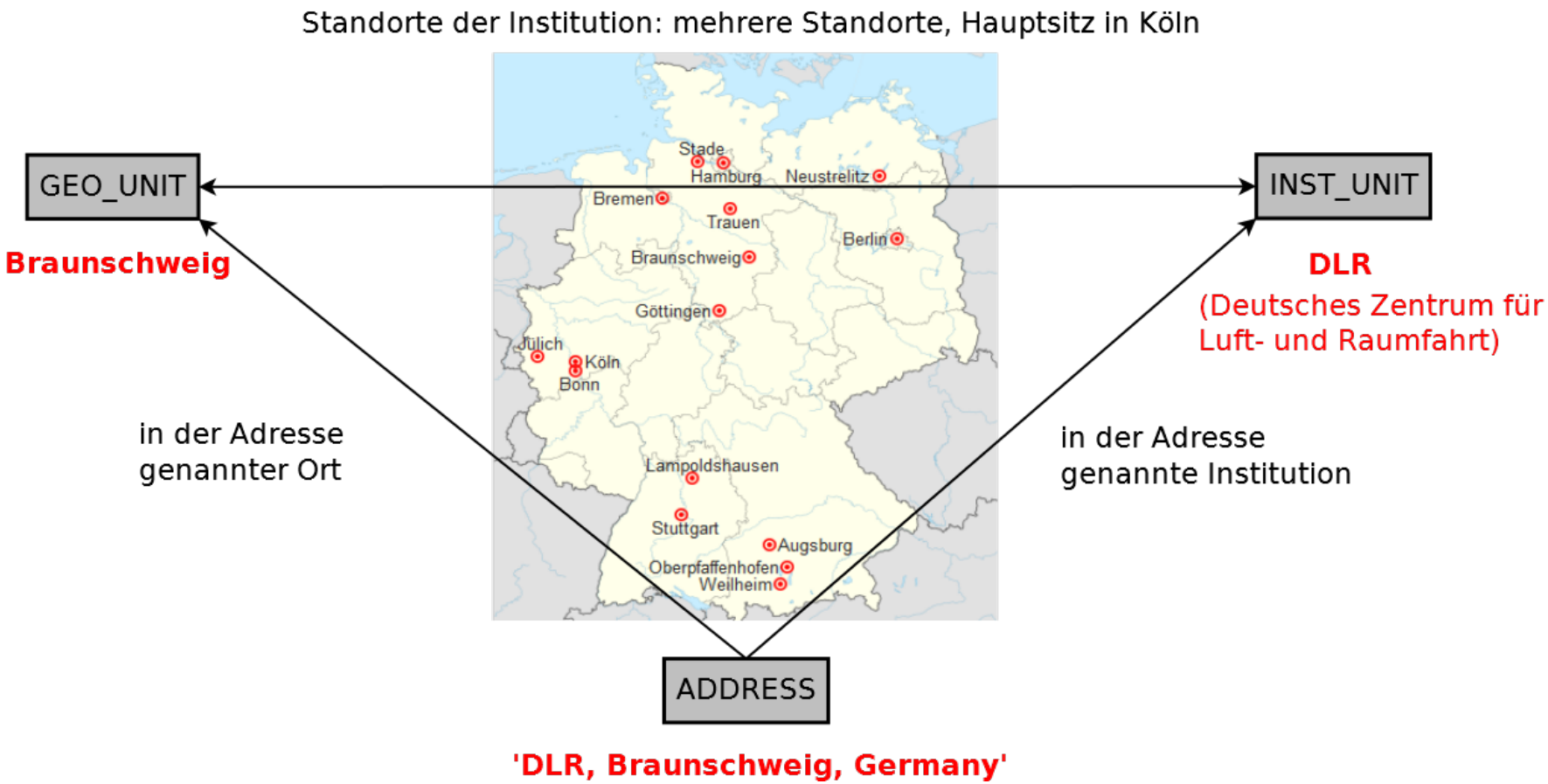
Adressen

Beispiele aus Scopus

Department of Palynology and Climate Dynamics, Albrecht-von-Haller Institute for Plant Sciences, University of Göttingen, Untere Karspüle 2, Göttingen, DEU Technische Universität München, Forschungsneutronenquelle Heinz Maier-Leibnitz (FRM-II), 85747 Garching, DEU Klinikum Kempten-Oberallgäu, Department of Surgery, Robert-Weixler-Strasse, D-87439 Kempten, DEU Technische Universität Dresden, Department of Psychiatry, Dresden, DEU Springer Verlag, Tiergartenstrae 17, 69121 Heidelberg, DEU Düsseldorf University, DEU Faculty of Medicine, University of Leipzig, Institute of Anatomy, Liebigstraße 13, 04103 Leipzig, DEU Pariser Platz A, D-81667 Miinchen, DEU **Université Paris-Sorbonne, Albert-Ludwigs-Universität Freiburg, DEU** Merz + Co., Dept. of Pharmacological Research, Eckenheimer Landstrasse 100-104, D-60318 Frankfurt/M, DEU Universitatsklin. Schleswig-Holstein, HNO-Klinik, Abt. fur Phoniatrie/Padaudiologie, Ratzeburger Allee 160, D-23562 Lübeck, DEU University of Stuttgart, Institute of Space Systems, Pfaffenwaldring 31, 70550 Stuttgart, DEU Military Hospital Ulm, Department of Surgery, 89070 Ulm, DEU Deutsches Krebsforschungszentrum, Abt. Med. Physik in der Radiologie, D-69120 Heidelberg, DEU Bavarian Acad Sci, Commiss Geodesy & Glaciol, Alfons Goppel Str 2, D-80539 Munich, Germany Univ Bayreuth, Dept Anim Physiol, POB 101251, D-95440 Bayreuth, Germany Friedelstr. 40, 12047 Berlin, DEU Univ. Potsdam, IPGP Paris, Alemania, DEU

Adressen

Institutionelle & geografische Information



Map: https://de.wikipedia.org/wiki/Deutsches_Zentrum_f%C3%BCr_Luft-_und_Raumfahrt

Adressen

Daten im Web of Science

Verfügbar im Web of Science:

- corporate addresses ('C1')
- separat angegebene reprint address ('RP')
- Vorstandardisierung der Adressen
- Autor – Adress – Linking ab 2008
- keine Adressdaten vor 1965

Namensvarianten *Schreibfehler* VERSCHIEDENE SPRACHEN Abkürzungen `unterschiedliche Hierarchieebenen` **semantische Fehler** *unvollständige Adressen* *Adressen, in denen keine Institution genannt ist* ***Adressen, in denen mehr als eine Adresse genannt ist***....

→ **viele Adress-Varianten je Institution und einige weitere Herausforderungen...**

Adressen

WoS – Vorstandardisierung

Adressen im Originaldokument:

Evaluating question answering over linked data ☆

Vanessa Lopez[a], Christina Unger[b], Philipp Cimiano[b], Enrico Motta[c]
[a] IBM Research, Smarter Cities Technology Centre, Mulhuddart, Dublin, Ireland
[b] Semantic Computing Group, CITEC, Universität Bielefeld, Bielefeld, Germany
[c] Knowledge Media Institute, The Open University, Milton Keynes, UK

→ **Adressen im WoS:**

Addresses:

[1] IBM Res, Smarter Cities Technol Ctr, Dublin, Ireland

[2] Univ Bielefeld, CITEC, Semant Comp Grp, D-33615 Bielefeld, Germany

[3] Open Univ, Knowledge Media Inst, Milton Keynes MK7 6AA, Bucks, England

→ Abkürzungen, Permutationen, Postleitzahl ergänzt, ...

Adressen

Datenfehler: Freiburg & Freiberg

Web of Science:

Addresses:

[1] Univ Nat Resources & Life Sci, Dept Appl Genet & Cell Biol, Muth

[2] Univ Freiberg, Inst Mol Med & Cell Res, Freiburg, Germany

accession number: WOS:000379886600013

Freiberg

Freiburg

Adressen

Datenfehler: Freiburg & Freiberg

ELSEVIER

Archives of Biochemistry and Biophysics

Volume 603, 1 August 2016, Pages 110–117

The death enzyme CP14 is a unique papain-like cysteine proteinase with a pronounced S2 subsite selectivity

Melanie Paireder[a], Ulrich Mehofer[a], Stefan Tholen[b], Andreas Porodko[a], Philipp Schähs[a], Daniel Maresch[c], Martin L. Biniossek[b], Renier A.L. van der Hoorn[d], Brigita Lenarcic[e], Marko Novinec[e], Oliver Schilling[b, f], Lukas Mach[a,]

[a] Department of Applied Genetics and Cell Biology, University of Natural Resources and Life Sciences, Vienna, Austria

[b] Institute of Molecular Medicine and Cell Research, University of Freiburg, Germany

[c] Department of Chemistry, University of Natural Resources and Life Sciences, Vienna, Austria

[d] The Plant Chemetics Laboratory, Department of Plant Sciences, University of Oxford, United Kingdom

Adressen

Datenfehler: Freiburg & Freiberg

Web of Science

Addresses:

- [1] Univ Freiberg, Inst Anat & Cell Biol, Dept Mol Embryol, Emmy Noether Grp Stem Cell Biol, Freiburg, Germany
- [2] Univ Freiburg, Spemann Grad Sch Biol & Med, Hugstetter Str 55, D-79106 Freiburg, Germany
- [3] Univ Freiburg, Fac Biol, Hugstetter Str 55, D-79106 Freiburg, Germany
- [4] Univ Freiburg, Ctr Biol Signaling Studies BIOSS, Freiburg, Germany
- [5] Univ Cambridge, Wellcome Trust Canc Res UK Gurdon Inst, Cambridge, England

Originaldokument

[1]Emmy Noether-Group for Stem Cell Biology, Department of Molecular Embryology, Institute of Anatomy and Cell Biology, University of Freiburg. [2]Spemann Graduate School of Biology and Medicine and Faculty of Biology, University of Freiburg, Freiburg, Germany. [3]Center for Biological Signaling Studies (BIOSS), University of Freiburg. *These authors contributed equally to this work. *Present address: Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, United Kingdom. Correspondence and requests for materials should be addressed to J.P. (email: jan.pruszak@uniklinik-freiburg.de)

→ **korrekte und fehlerhafte Adressen in einem Dokument?!**

accession number: WOS:000372049700001

Adressen

Datenfehler: weitere Beispiele

Stadt:

'Frankfurt am Main' vs. 'Mainz':

*J.W. Goethe-Universitaet, **Frankfurt am Main**, Germany*

→ *JW Goethe Univ, **Mainz**, Germany* (WOS:000248891500017)

'Homburg' vs. 'Hamburg':

*Unikliniken des Saarlandes, **Homburg**, Germany*

→ *Unikliniken Saarlandes, **Hamburg**, Germany* (WOS:000241747000003)

Land:

*Institut de Biologie du Développement de Marseille, CNRS, Université de la Méditeranée, Campus de Luminy-case 907, 13288 Marseille cedex 9, **France***

→ *Univ Mediteranee, CNRS, Inst Biol Dev Marseille, Campus Luminy Case 907, D-50829 Cologne 9, **Germany*** (WOS:000238688800016)

Adressen

Datenfehler: Originaldokument

Web of Science:

Addresses:

[1] Vienna Tech Univ, Inst Operat Res & Syst Theory, A-1040 Vienna, Austria

[2] Univ Bielefeld, Dept Econ, Vienna, Austria

[3] Tilburg Univ, Dept Econometr, Vienna, Austria

[4] Univ Vienna, Inst Management Sci, Vienna, Austria

Original document:

Alfred Greiner[a], Gustav Feichtinger[b], Josef L. Haunschmied[b], , Peter M. Kort[c], Richard F. Hartl[d]

[a] Department of Economics, University of Bielefeld, Vienna, Austria

[b] Institute of Operations Research and Systems Theory, Vienna University of Technology, Argentinierstrasse 8, 1040 Vienna, Austria

[c] Department of Econometrics and Center, Tilburg University, Vienna, Austria

[d] Institute of Management Science, University of Vienna, Vienna, Austria

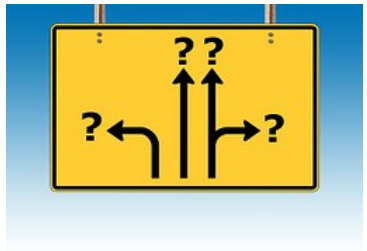
→ **Fehler im Originaldokument!**

accession number: WOS:000170961300010

Institutionenkodierung

Komplexe Institutionen-Realität

- **Hierarchieebenen**
  (Universität, Fakultät, AG, ...)
- **Definition** der 'Hauptinstitution' (Aggregationsebene für die Zuordnung)
  (Universitätskliniken, Lehrkrankenhäuser, An-Institute)
- verschiedene **Standorte**
- **Strukturveränderungen** über die Zeit
  (Teilungen, Fusionen, Eingliederungen, neu gegründete oder geschlossene Institutionen...)
- **Veränderungen der Attribute**
  (Namenswechsel, Sektorwechsel, Standortwechsel, ...)
- **Mehrfachzuordnungen**
  (z.B. bei Sektorzuordnungen oder Relationen zwischen Institutionen)



Institutionenkodierung

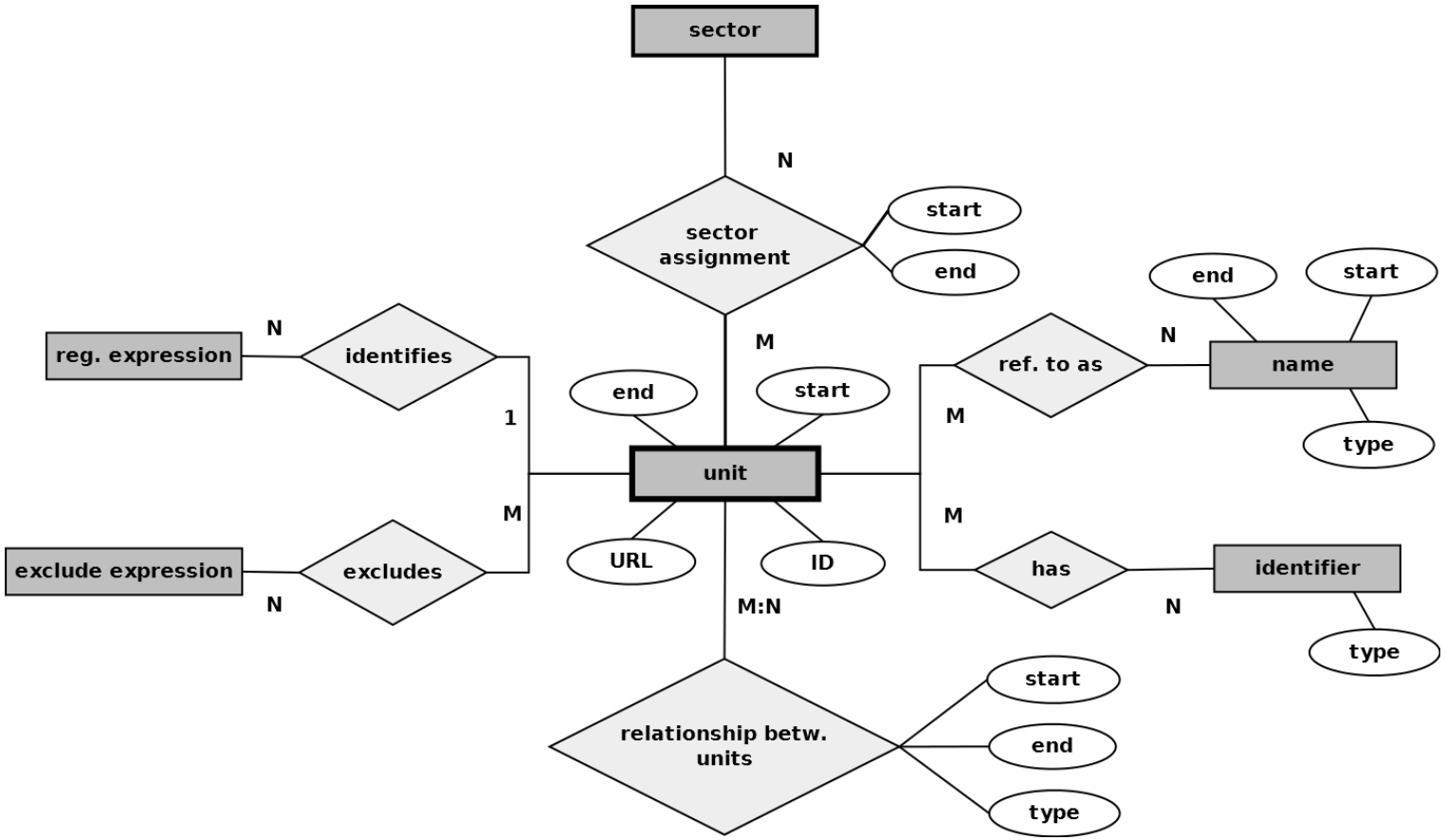
Projekt 'Institutionenkodierung' in Bielefeld

- **Kontext:** Kompetenzzentrum Bibliometrie
- Zuordnung von Adressen zu real existierenden **deutschen Institutionen** (Adressen aus Dokumenten mit mindestens einer deutschen Adresse)

- **Textmuster** sind Hauptbestandteil des Verfahrens, aktuell $\sim$ 51.000 Muster
- Basisdaten und Ergebnistabellen werden in **relationalen Datenbanken** vorgehalten
- aktuell $\sim$ 4.600 'units', $\sim$ **2.300 Hauptinstitutionen**
- **institutionelle** Zuordnung wird von der **geografischen** Zuordnung getrennt
- **zwei verschiedene Varianten** (bezeichnet als 'A' (aktuell) and 'S' (synchron)) zur Behandlung der Historie von Institutionen
  (verschiedenen Anwendungskontexte)
  $\sim$ 'current potential' (A) und 'work done at' (S) im Fall von Autoren
- **regelmäßige Updates** für Basisdaten und Ergebnistabellen

Institutionenkodierung

Entity Relationship Model, Basisdaten I-Kodierung



Institutionenkodierung

Datenbankschema: Prinzipien/Eigenschaften

- **Flexibilität:** das Schema ist in der Lage, **alle** Typen von Einheiten (units), Veränderungen, Relationen... aufzunehmen, die in der Realität vorkommen (und in diesem Zusammenhang von Interesse sind)
- **Normalisierung:** für Namen, Identifier, Sektorzugehörigkeiten usw. existieren jeweils eigene/einzelne Tabellen
- **Einheiten** aller hierarchischen Ebenen können als Einheit aufgenommen werden
- **Relationen:** Beziehungen verschiedener Arten können erfasst werden (isPartOf, isPartOfNetwork, isAffiliatedInstituteOf, isAcademicHospitalOf, ...)
- **Vorgänger & Nachfolger:** Nachfolgen können abgebildet/aufgenommen werden
- Veränderungen über die Zeit können erfasst werden: **Start- und Enddaten** für alle Entitäten/ Attribute
- **Mehrfachzuordnungen** können abgebildet werden
- Fremdschlüssel und Trigger sorgen für **semantische Integrität**

Institutionenkodierung

Aggregation auf Hauptinstitutionsebene

- Universitätskliniken und Lehrkrankenhäuser: Teil der Universität?
- An-Institute: Teil der Universität?
- gemeinsame medizinische Fakultäten von Universitäten (z.B. *Charité*)
- ...

Prinzip: Definition über organisationelle und rechtliche Unabhängigkeit

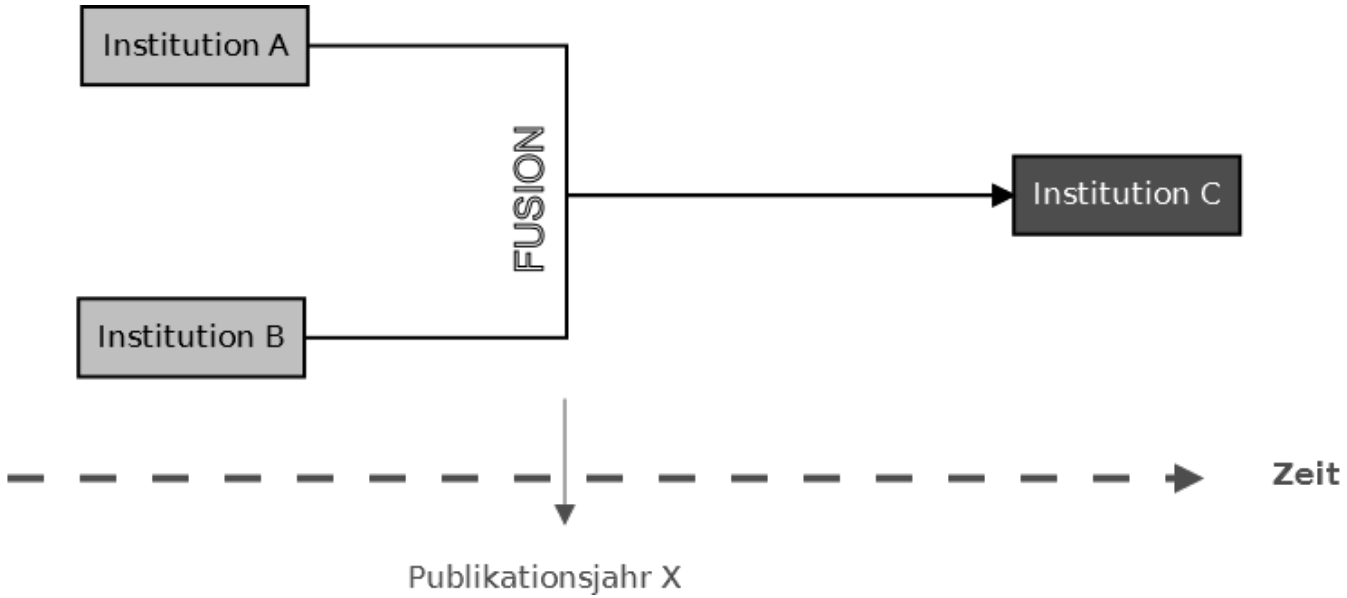
...in vielen Fällen einfach: Universitäten, Max Planck Institute, Kliniken (ausgenommen Universitätskliniken und Lehrkrankenhäuser), ...

Beispiele für Spezialfälle:

- **Universitätskliniken:** Teil der Universität
- **Lehrkrankenhäuser:** separate Hauptinstitution (LehrKH – Relation wird erfasst)
- **An-Institut:** separate Hauptinstitution (An-Instituts – Relation wird erfasst)
- *Charité* (and ähnliche Fälle): separate Hauptinstitution (in vielen Fällen ist hier die Zuordnung zu einer der zugehörigen Universitäten gar nicht möglich)

Institutionenkodierung

Strukturveränderungen über die Zeit, Beispiel Fusion



Modus S (synchron): Zuordnung unter Berücksichtigung der (historischen) Situation im Publikationsjahr (PY)
(bis PY=X werden die Adressen/Publikationen A oder B, danach C zugeordnet)

Modus A (aktuell): Zuordnung der aktuellen Situation
(alle Adressen/Publikationen werden C zugeordnet, denn C ist die einzige aktuell existierende Institution → C 'erbt' die Publikationen von A und B)

Institutionenkodierung

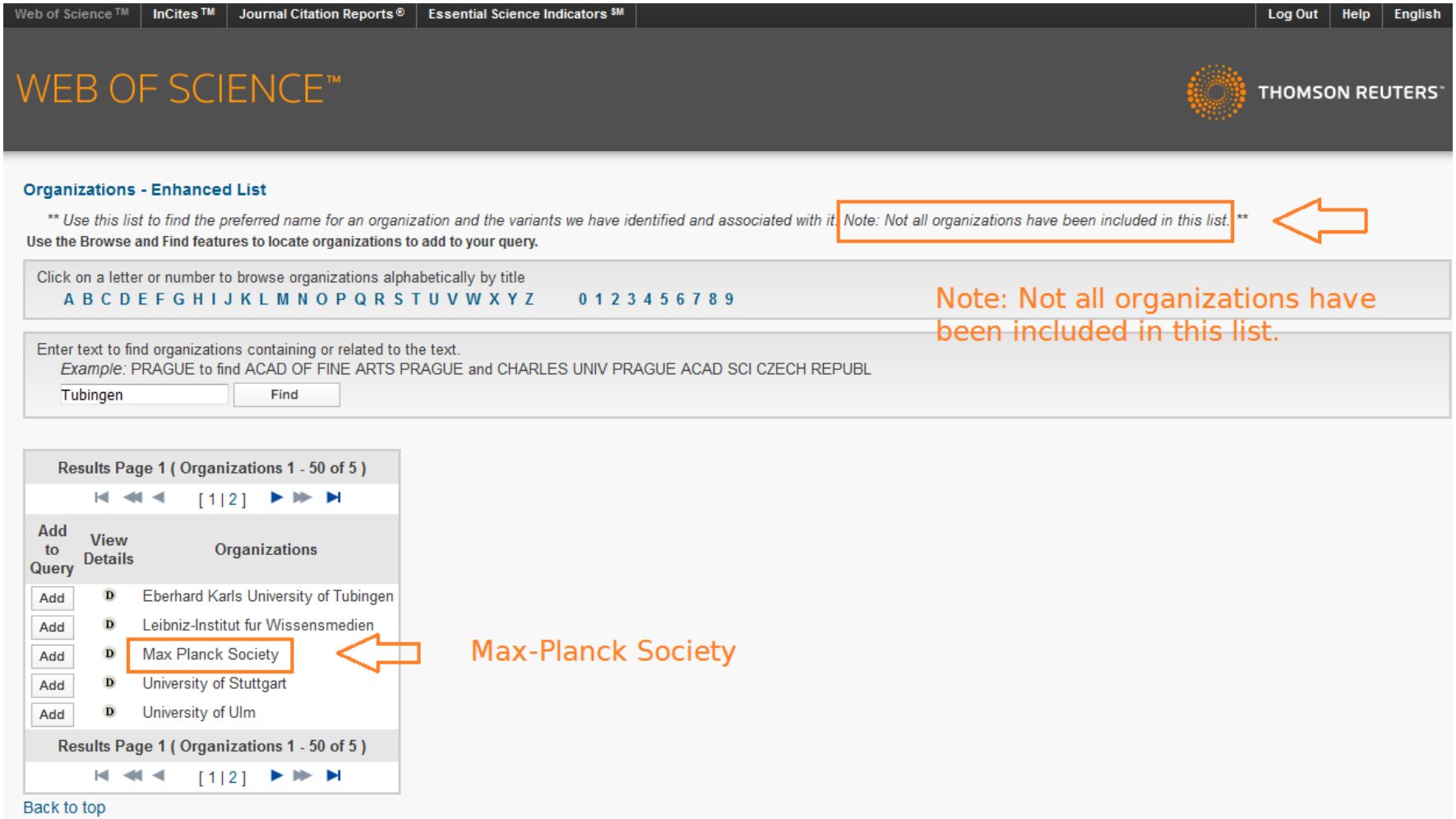
Effekte der Institutionenkodierung auf bibliometrische Indikatoren

Indikator	Anzahl von einbezogenen Institutionen	Werte in +/- 5% des korrekten Wertes (in %)	Median der Abweichung (in %)
# **Publikationen**	436	6.6	-50.4
# **Publikationen** fraktioniert	436	6.5	-51.9
# **Zitationen**	436	8.2	-51.6
# **Zitationen** fraktioniert	434	8.9	-48.9

→ Projekt 'Effects of institutional disambiguation on bibliometric indicators' mit Paul Donner, DZHW

Institutionenkodierung

Organization Enhanced



Institutionenkodierung

Organization Enhanced, Max Planck Gesellschaft

Organization Name:	Add	**MAX PLANCK SOCIETY**
Other Names:		MAX PLANCK SOCIETY
Address:		MUNICH, GERMANY
Website:		http://www.mpg.de/english/portal/index.html
Name Variants:	Add	3 MAX PLANCK INST MOL CELL BIOL GENET
	Add	3 MAX PLANCK INST PLASMAPHYS
	Add	3MAX PLANCK INST BIOL AGEING
	Add	4 MAX PLANCK INST QUANTENOPT
	Add	ABC MAX PLANCK INST
	Add	ABT STRUKTURELLE BIOL MAX PLANCK INST MOL PHYSIOL
	Add	ADM HEADQUARTERS MAX PLANCK SOC
	Add	AEI MAX PLANCK INST GRAVITAT PHYS
	Add	AG RIBOSOMEN MAX PLANCK INST MOL GENET
	Add	ALBERT EINSTEIN GRAVITAT PHYS
		...

- Max-Planck-Institut für Plasmaphysik (Garching bei München)
- Max-Planck-Institut für molekulare Physiologie (Dortmund)
- ...

→ Max Planck Gesellschaft mit all ihren Instituten ist als **EINE** Institution erfasst.

Institutionenkodierung

Organization Enhanced, Beispiel: Fehler

Geografische Angabe (hier Straßenname) führt zu einem Zuordnungsfehler in Organization Enhanced:

? Einzelfall ?

→ Web of Science Online Suche...

Institutionenkodierung

organization enhanced=Martin Luther University Halle Wittenberg

Reprint Address: Petersen, W (reprint author)

– Martin Luther Hosp, Dept Orthopaed & Trauma Surg, Caspar Theyss Str 27-31, D-14193 Berlin, Germany.

Organization-Enhanced Name(s)

Martin Luther University Halle Wittenberg

... Krankenhaus in Berlin ('MARTIN LUTHER HOSP')...

...gehört zur Universität Hamburg (Straße: 'MARTIN LUTHER KING PL') ...

Addresses:

[1] Prirodonaucen Muzej Makedonija, Bulevar Ilinden 86, Skopje, Macedonia

[2] Ellhornstr 21, D-28195 Bremen, Germany

– [3] Biozentrum Grindel, Martin Luther King Pl 3, D-20146 Hamburg, Germany

Organization-Enhanced Name(s)

Martin Luther University Halle Wittenberg

University of Hamburg

Institutionenkodierung

organization enhanced=Martin Luther University Halle Wittenberg

Author Information

Reprint Address: Hausotter, W (reprint author)

– Sozialmed Rehabil Wesen, Neurol & Psychiat, Martin Luther Str 8 D-87527 Sonthofen, Germany.

Organization-Enhanced Name(s)

Martin Luther University Halle Wittenberg

...medizinisches Zentrum in Sonthofen (Straße: 'MARTIN LUTHER STR')

...akademisch unabhängige Institution in Greifswald (Straße: 'MARTIN LUTHER STR')

Author Information

Reprint Address: Roesner, M (reprint author)

– Alfried Krupp Wissensch Kolleg Greifswald, Martin Luther Str 14 D-17489 Greifswald, Germany.

Organization-Enhanced Name(s)

Martin Luther University Halle Wittenberg

Institutionenkodierung

Organization Enhanced

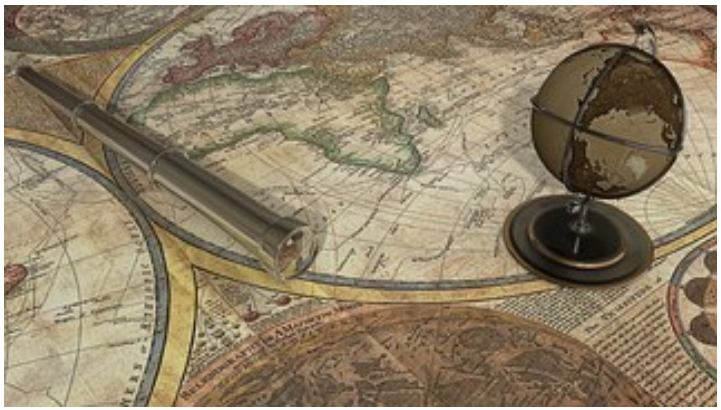
- 'not all organizations included' → viele deutsche Institutionen fehlen
- **Aggregationsebene/Definition der Hauptinstitution**
  (z.B. alle Institute der Max Planck Gesellschaft als eine Institution zusammengefasst)
- keine ausführliche **Nutzer-Dokumentation**
  (Aggregationsebene, Umgang mit Spezialfällen wie Lehrkrankenhäusern, An-Instituten, ... , Umgang mit Strukturveränderungen über die Zeit usw.)
- **bisher nicht in den Rohdaten enthalten**
  (Erweiterung der Rohdaten geplant → Auswertungen und Vergleiche möglich)

→ ...in der Entwicklung, aber (je nach Anwendungskontext) noch nicht ready-to-use ohne weitere Bearbeitung (mindestens für deutsche Adressen)

Geokodierung

Anwendungskontexte

- **Raumbezogene** bibliometrische Analysen
  (im Gegensatz zu organisationellen Analysen)
- **Verschiedene Aggregationsmöglichkeiten** für verschiedene Anwendungskontexte:
  z.B. Städte, Regionen, Bundesländer, ...
- Aufnahme von **Geokoordinaten** ermöglicht Kartendarstellungen
- **Input für die Institutionenkodierung**:
  Standardisierung der geografischen Bezeichnungen in Adressen
- ...



Geokodierung

Projekt 'Geokodierung' – Work in Progress

- **Kontext:** Kompetenzzentrum Bibliometrie
- Zuordnung von Adressen zu **geografischen Einheiten & Geokoordinaten**

Ziele

- **Statistische Auswertungen** zu Geoinformationen in den bibliometrischen Datenbanken WoS und Scopus
- Evaluation von **freien Quellen** zur Extraktion von Geoinformationen (Entitäten, Relationen, Namensvarianten, Geokoordinaten, ...)
- Entwicklung eines **Datenbankschemas für Basisdaten**
- Entwicklung eines **Verfahrens zur Zuordnung** von Adressen aus bibliometrischen Datenbanken zu geografischen Entitäten und zugehörigen Geokoordinaten
- Erstellung von **anwenderfreundlichen Ergebnistabellen** zur Integration in die Datenbanken des Kompetenzzentrums Bibliometrie

Geokodierung

Verfügbarkeit von Geoinformationen in WoS und Scopus

Garchhing Garchirg Garching Bayern GARCHING HOCHBRUCK GARCHING BEI MUNCH GARCHING GARCHING MUNCHEN Garchinq Garching Bei Mue

Strukturierte Informationen 'CITY' und 'POSTALCODE', aber:

- **unvollständig:** nicht für jede Adresse verfügbar
- **Fehler in der Struktur:** nicht immer ist im Feld 'CITY' auch tatsächlich ein Ortsname angegeben
  (sondern Straßennamen, Bundesländer, Namen von Institutionen, ...),
  'CITY' enthält zum Teil auch Postleitzahlen, die nicht gesondert in 'POSTALCODE' zu finden sind
  (Beispiel: CITY='33604 Bielefeld', POSTALCODE leer)
- **nicht eindeutig:** mehrere Ortsnamen in 'CITY'
  (Beispiel: CITY='Bielefeld und Bonn')

Garching Near Munich Garching B Munchen MUNICH GARCHING GARCHING MUENCHEN Garehing Garching Bei Muchen Garching BM Garchingen Garch

Geokodierung

Verfügbarkeit von Geoinformationen in WoS und Scopus

Garchingbei Munchen Gaerching Garching B Garching Bai Munchen Garching Bei Meunchen Garchning Garching Frm Munchen Garchine Garching D Garching

Strukturierte Informationen 'CITY' und 'POSTALCODE', aber:

- Namensvarianten *Schreibfehler* VERSCHIEDENE SPRACHEN Abkürzungen `unterschiedliche Hierarchieebenen (Hamburg ↔ Eppendorf)` **semantische Fehler** *unvollständige Angaben ('Frankfurt')*...

  → Probleme analog zur Institutionenkodierung

- **nicht nur 'CITY' und 'POSTALCODE'** können geografische Informationen enthalten
  (Beispiel (Scopus): Adresse='University of Hamburg, University Clinic Eppendorf, Institute for Sexual Research and Forensic Psychiatry, Berlin, DEU', CITY='Berlin')

D-8046-GARCHING Graching Garcing Garching By Munich GARCCHING Garchig Garshing Garching Bein Munchen GARSCHING MUNCHEN GARCHING G

Geokodierung

Verfügbarkeit von Geoinformationen in WoS und Scopus

Einfache Gruppierungen
über 'CITY' sind nicht möglich (Varianz)...

...und 'CITY' und 'POSTALCODE bieten
auch keine vollständige Informationsquelle
(geografische Informationen in anderen
Teilen der Adresse gehen verloren)

→ **Geoinformationen liegen vor, müssen aber extrahiert, bearbeitet, zugeordnet und aggregiert werden...**

Geokodierung

Aggregationsebene

Sinnvolle Aggregationsebene/Zuordnungsebene für Geoinformationen?

→ Städte/Gemeinden

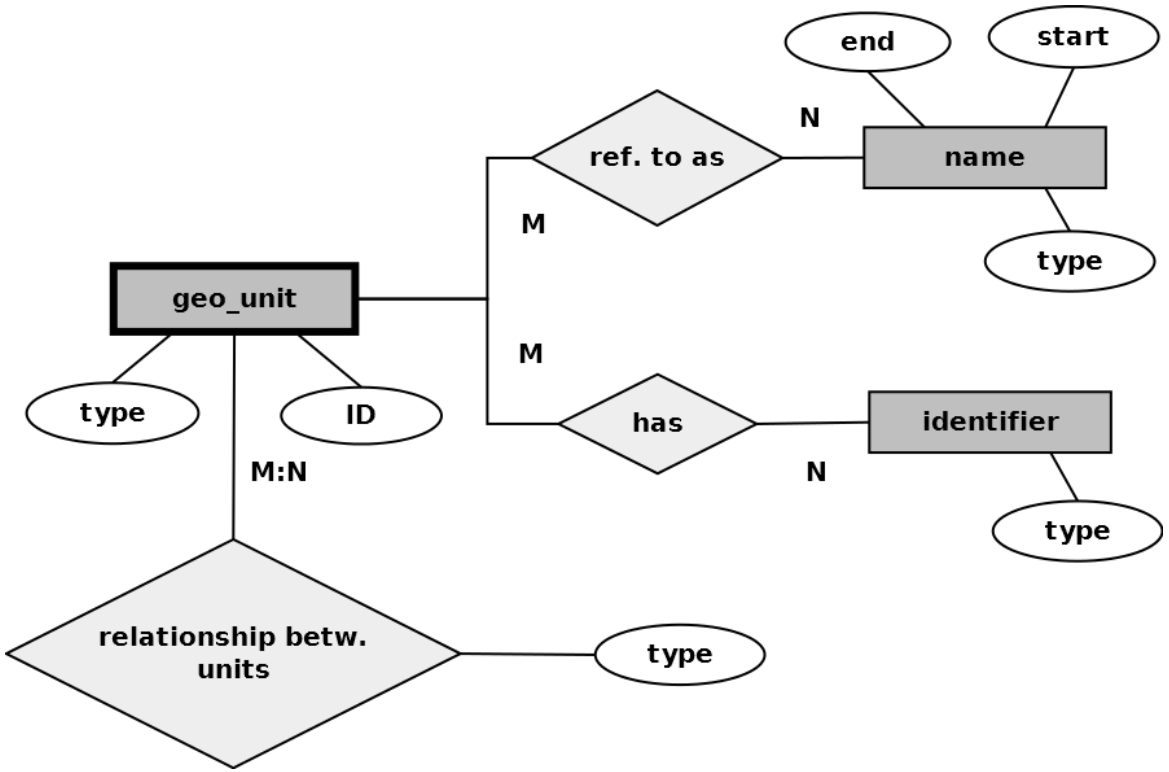
??? Identifikation über ...

- **...Ortsnamen?** ...nicht eindeutig: 'Frankfurt'? 'Neustadt' (Name von 21 Gemeinden/Städten in Deutschland lt. Gemeindeliste Statistisches Bundesamt)?
- **...PLZ-Ortnamens-Kombinationen?** ...mehrere Postleitzahlen je Stadt/Gemeinde → Aggregation nicht ausreichend.

!!! Lösung: amtlicher Gemeindeschlüssel. *Der Amtliche Gemeindeschlüssel (AGS) [...] ist eine Ziffernfolge zur Identifizierung politisch selbständiger Gemeinden oder gemeindefreier Gebiete.* [Wikipedia]

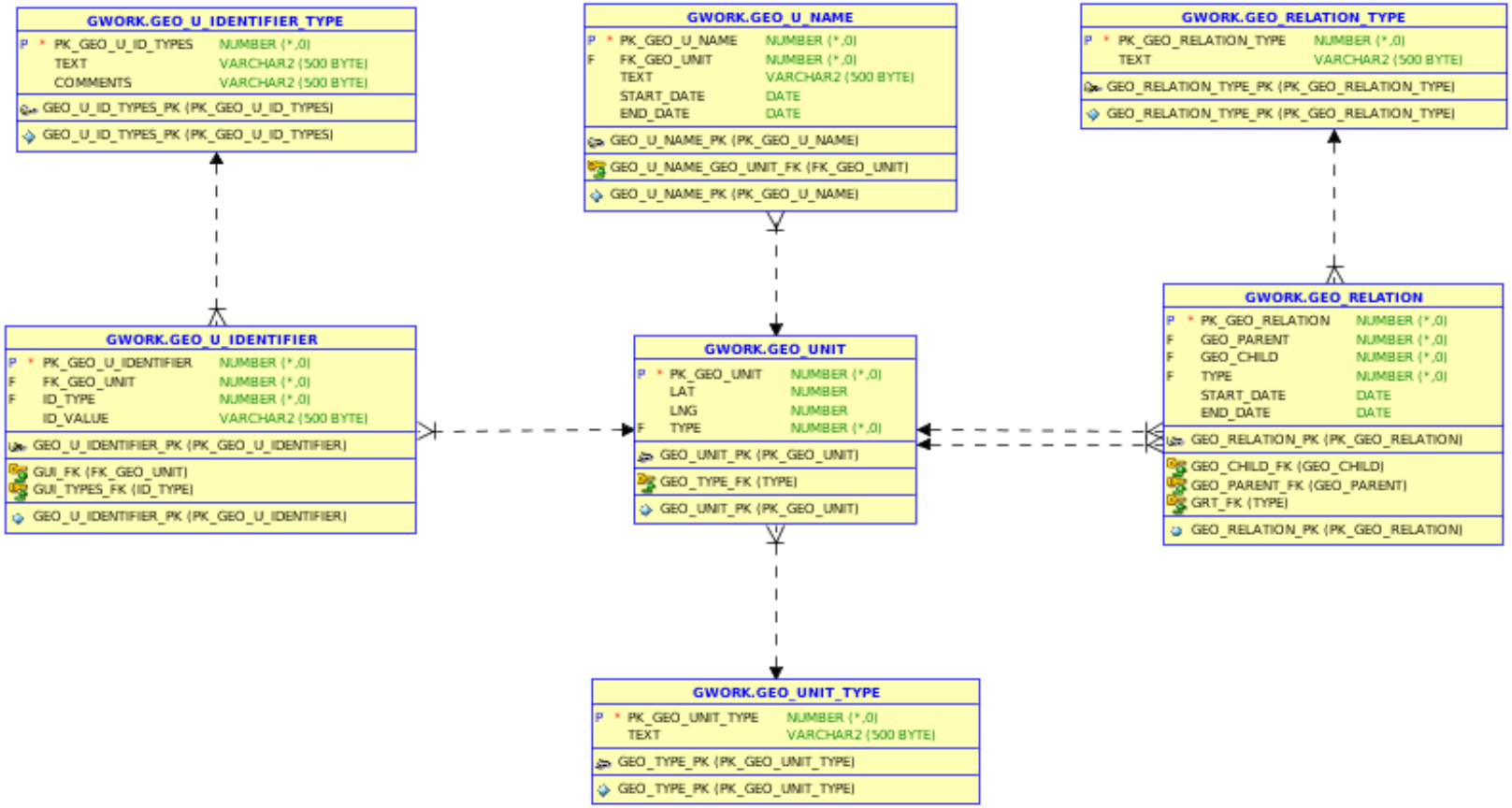
Geokodierung

Entity Relationship Model: Basisdaten G-Kodierung



Geokodierung

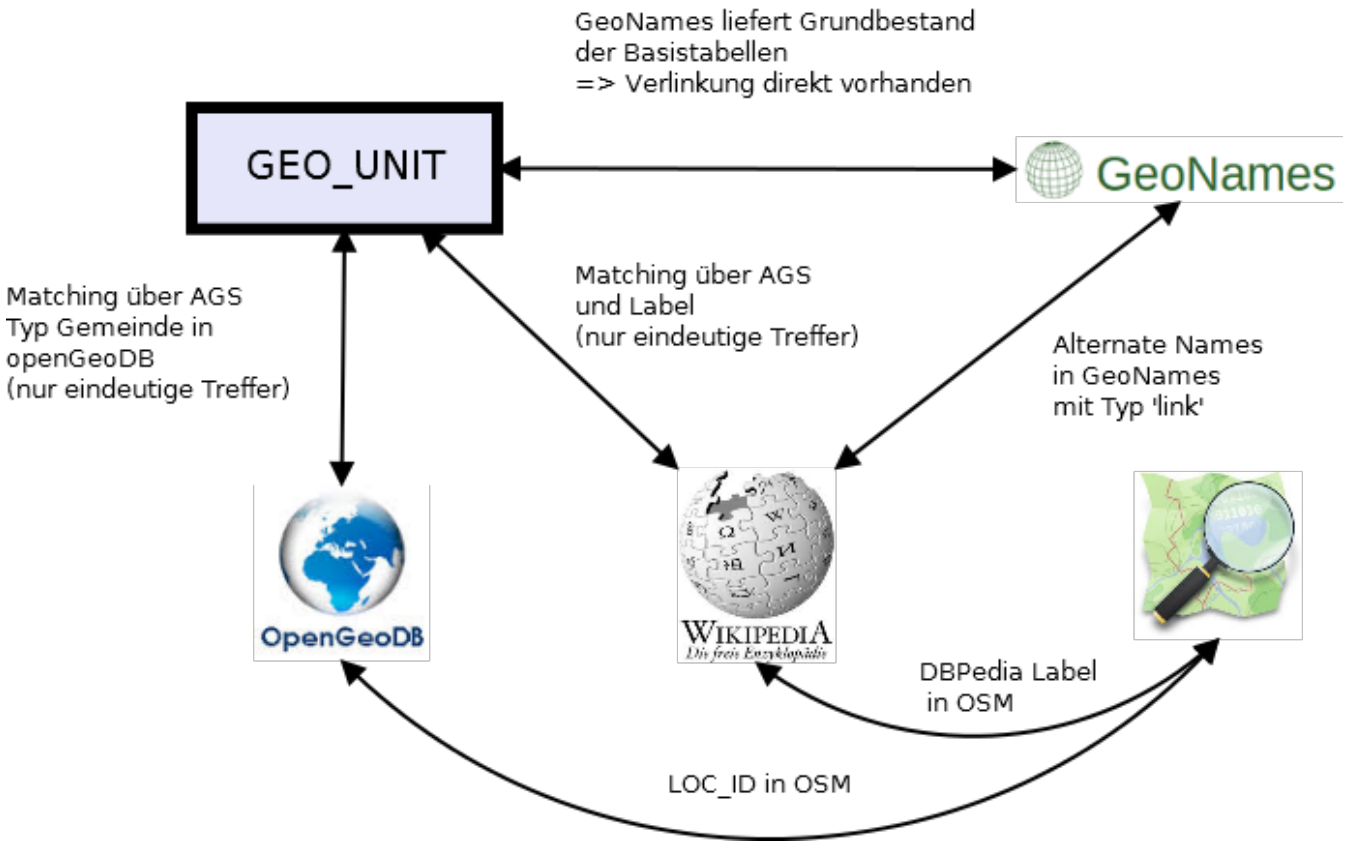
Datenbankschema



Geokodierung

Datenquellen

Datengrundlage für Basistabellen und Zuordnungsverfahren: Kombination aus freien Datenquellen (GeoNames, openGeoDB, OpenStreetMap, Wikipedia)



Geokodierung

Verfahren: Work in Progress

- **Basistabellen** werden aus den gewählten Quellen gespeist (Matching der Entitäten aus den verschiedenen Quellen erforderlich)
- alle verfügbaren hilfreichen **Informationen** (z.B. Namensvarianten, Postleitzahlen, Identifier, Relationen, Namen von Ortsteilen usw.) aus den Quellen werden in weiteren Tabellen erfasst und für das Matching genutzt ('CITY', 'PLZ', weitere Adressbestandteile)

- diverse **Stringmatchingverfahren** werden hierarchisch angewendet, Mehrfachzuordnungen möglich
- die Zuordnung erfolgt auf der **niedrigstmöglichen Hierarchieebene** (Stadt/Gemeinde) – ist eine solche Zuordnung nicht möglich, wird auf höheren Ebenen zugeordnet
- **Ergebnistabellen** mit der Zuordnung von Adressen zu einer oder mehreren Geo-Einheiten und zugehörigen Geokoordinaten und Aggregationsmöglichkeiten werden erstellt

Geokodierung

Zukunftspläne

- Weiterentwicklung des Verfahrens
- Auswertung weiterer Informationen aus den Datenquellen
- Auswertung und Hinzunahme weiterer Quellen
- Tests mit Adressdatensätzen anderer Länder
- Nutzung des Verfahren zur Aufdeckung von Fehlern in der Länderzuordnung
- .....

The end

...entweder heute, jetzt & hier... oder später an christine.rimmert@uni-bielefeld.de