



Wikipedia / WikiCon 2013

# Kategoriesystem 2.0

November 2013

- Kategorien:
  - Wie funktionieren sie?
  - Vor- und Nachteile des jetzigen Systems
  - Wozu sind sie gut?
- Gedanken und Anregungen für die Zukunft
  - Prä- vs. Postkombination, Facetten
  - Assoziationstypen
  - Weitere Aspekte



- Einführung
  - MediaWiki 1.3 (August 2004) [\[Release Notes\]](#)
  - Ersatz für Listen und Meta-Listen
- Prinzip
  - Ein Artikel (eine Seite) kann einer oder mehreren Kategorien zugeordnet werden
  - Eine Kategorie kann einer oder mehreren übergeordneten Kategorien zugeordnet werden (polyhierarchisch)
  - Eine Kategorie ist auch eine Seite
  - Parser extrahiert Kategoriezuordnungen aus Artikel-Quelltext

# Stärken und Schwächen

- Stärken
  - einfach
  - flexibel, vielseitig nutzbar
  
- Schwächen...
  - ...der MediaWiki-Implementierung
    - vielseitig nutzbar
    - verbesserungsbedürftige Benutzerschnittstelle
    - Kategorieänderungen sind aufwändig
  - ...der derzeitigen Systematik
    - inhomogen
    - arbeitsaufwändig und wartungsintensiv
    - teilweise schwer verständlich, kontraintuitiv

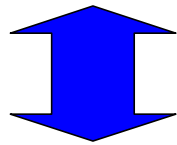
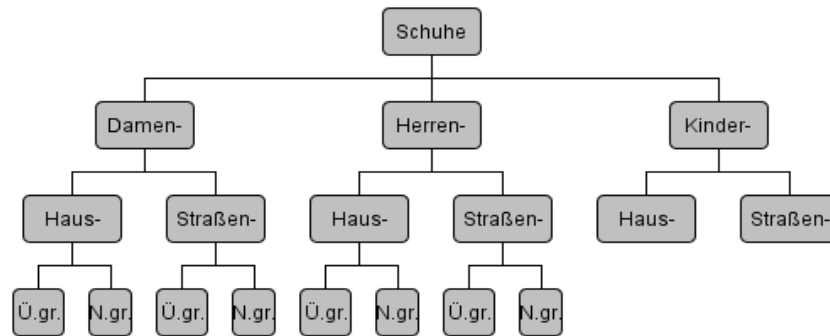


## Wozu sind Kategorien gut?

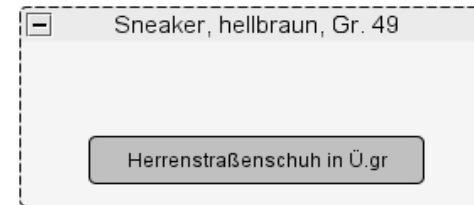
- Abbildung von „Zuständigkeitsbereichen“ für Redaktionen und Projekte
- Ermöglichung von Auswertungen und Statistiken
- Steuerung von Massенbearbeitungen
- Recherche- und Orientierungshilfe
  - für Wikipedianer
  - für „normale“ Leser [\[wirklich?\]](#)
- ...

# Präkombination vs. Postkombination

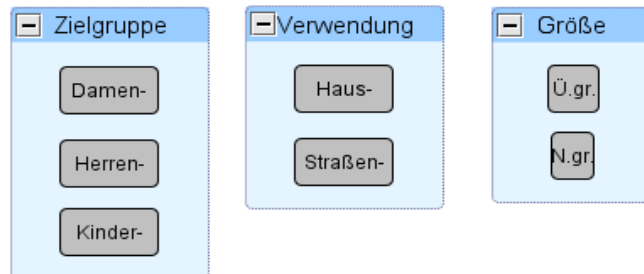
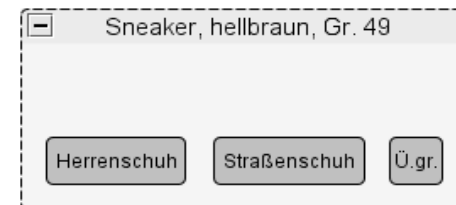
- WP:Kategorien: [\[WP\]](#)  
 „Grundlegendes: ...Facettenklassifikation...“ [\[WP\]](#)



## Präkombination



## Postkombination



Beispiel nach: Jutta Betram: *Einführung in die inhaltliche Erschließung*. 2005

## ■ Tatsächlich?

### ■ Beispiel: Kategorie:Geographisches Objekt [WP]

<ul style="list-style-type: none"> <li>▶ Geographisches Objekt nach Staat (24 K)</li> <li>x</li> <li>▶ Glaziologisch geprägtes geographisches Objekt (3 K, 6 S)</li> <li>▶ Humangeographisches Objekt (39 K, 10 S)</li> <li>A</li> <li>▶ Atoll (1 K, 266 S)</li> <li>B</li> <li>▶ Becken (3 K, 1 S)</li> <li>▶ Berg (5 K, 2 S)</li> <li>▶ Bergwerk (7 K, 4 S)</li> <li>▶ Biotop (2 K, 10 S)</li> <li>▶ Bucht (5 K, 10 S)</li> <li>D</li> <li>▶ Damm (34 S)</li> <li>▶ Deich (11 S)</li> <li>▶ Drumlin (19 S)</li> <li>▶ Düne (1 K, 15 S)</li> <li>E</li> <li>▶ Einschlagkrater (Erde) (6 K, 3 S)</li> <li>▶ Einzelpflanze (1 K, 2 S)</li> <li>F</li> <li>▶ Felsen (5 K, 62 S)</li> <li>▶ Fluss (4 K, 29 S)</li> <li>▶ Flussdelta (16 S)</li> </ul>	<ul style="list-style-type: none"> <li>▶ Geist (21 S)</li> <li>▶ Gletscher (5 K)</li> <li>H</li> <li>▶ Halbinsel (3 K, 4 S)</li> <li>▶ Höhle (4 K, 13 S)</li> <li>I</li> <li>▼ Insel (8 K, 31 S) <ul style="list-style-type: none"> <li>▶ Insel nach Eigenschaft (7 K)</li> <li>▶ Insel nach Gewässertyp (3 K)</li> <li>▶ Insel nach Inselgruppe (67 K)</li> <li>▼ Insel nach Kontinent (8 K) <ul style="list-style-type: none"> <li>▶ Insel ohne Kontinentalbezug (2 K, 75 S) <ul style="list-style-type: none"> <li>▶ Insel (Afrika) (263 S)</li> <li>▶ Insel (Antarktika) (100 S)</li> <li>▶ Insel (Asien) (754 S)</li> <li>▶ Insel (Australien und Ozeanien) (716 S)</li> <li>▶ Insel (Europa) (1 627 S)</li> <li>▶ Insel (Nordamerika) (1 K, 692 S)</li> <li>▶ Insel (Südamerika) (130 S)</li> </ul> </li> <li>▶ Insel nach Region (4 K)</li> <li>▼ Insel nach Staat (129 K) <ul style="list-style-type: none"> <li>▶ Insel nach abhängigem Gebiet (8 K)</li> <li>▶ Insel (Ägypten) (7 S)</li> <li>▶ Insel (Äquatorialguinea) (6 S)</li> <li>▶ Insel (Albanien) (1 S)</li> <li>▶ Insel (Antigua und Barbuda) (3 S)</li> <li>▶ Insel (Argentinien) (3 S)</li> <li>▶ Insel (Aserbaidschan) (1 S)</li> <li>▶ Insel (Australien) (7 K, 2 S)</li> <li>▶ Insel (Bahamas) (17 S)</li> <li>▶ Insel (Bahrain) (6 S)</li> <li>▶ Insel (Bangladesch) (3 S)</li> <li>▶ Insel (Barbados) (2 S)</li> <li>▶ Insel (Belize) (6 S)</li> <li>▶ Insel (Bolivien) (5 S)</li> <li>▶ Insel (Botswana) (1 S)</li> <li>▶ Insel (Brasilien) (18 S)</li> <li>▶ Insel (Bulgarien) (7 S)</li> <li>▶ Insel (Chile) (24 S)</li> <li>▶ Insel (Republik China) (1 K, 1 S)</li> <li>▶ Insel (China) (1 K, 20 S)</li> <li>▶ Insel (Costa Rica) (4 S)</li> <li>▶ Insel (Dänemark) (11 S)</li> <li>▶ Insel (Demokratische Republik Kongo) (1 S)</li> <li>▶ Insel (Deutschland) (14 K, 2 S)</li> </ul> </li> </ul> </li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>O</li> <li>▶ Oase (7 K, 25 S)</li> <li>P</li> <li>▶ Pass (5 K, 2 S)</li> <li>Q</li> <li>▶ Quelle (6 K, 43 S)</li> <li>R</li> <li>▶ Region (20 K, 25 S)</li> <li>▶ Rohstoffvorkommen (7 K, 39 S)</li> <li>S</li> <li>▶ Sandbank (1 K, 31 S)</li> <li>▼ See (6 K, 8 S) <ul style="list-style-type: none"> <li>▶ See nach Eigenschaft (4 K)</li> <li>▼ See nach Kontinent (7 K) <ul style="list-style-type: none"> <li>▶ Stausee nach Kontinent (6 K) <ul style="list-style-type: none"> <li>▶ See in Afrika (145 S)</li> <li>▶ See in Amerika (2 K)</li> <li>▶ See in Antarktika (13 S)</li> <li>▶ See in Asien (211 S)</li> <li>▶ See in Australien und Ozeanien (298 S)</li> <li>▶ See in Europa (3.015 S)</li> </ul> </li> <li>▼ See nach Staat (125 K) <ul style="list-style-type: none"> <li>▶ Stausee nach Staat (27 K)</li> <li>▶ See in Afghanistan (4 S)</li> <li>▶ See in Ägypten (7 S)</li> <li>▶ See in Albanien (13 S)</li> <li>▶ See in Algerien (3 S)</li> <li>▶ See in Angola (3 S)</li> <li>▶ See in Argentinien (24 S)</li> <li>▶ See in Armenien (2 S)</li> <li>▶ See in Aserbaidschan (1 K, 1 S)</li> <li>▶ See in Äthiopien (26 S)</li> <li>▶ See in Australien (7 K, 5 S)</li> <li>▶ See in Belgien (2 K)</li> <li>▶ See in Bolivien (26 S)</li> <li>▶ See in Bosnien und Herzegowina (8 S)</li> <li>▶ See in Brasilien (1 K)</li> <li>▶ See in Bulgarien (19 S)</li> <li>▶ See in Burkina Faso (7 S)</li> <li>▶ See in Burundi (3 S)</li> <li>▶ See in Chile (21 S)</li> <li>▶ See in China (1 K, 46 S)</li> </ul> </li> </ul> </li> </ul> </li> </ul>
--	--	--



Typ x Kontinent

# Präkombination vs. Postkombination

- „Klassische“ Kriterien
  - Sucheffizienz vs. Flexibilität
  - Vermeidung/Ermöglichung sinnloser Kombinationen (Einschlagkrater, Jupiter)
  - Pflegeaufwand des Katalogs
- Für Wikipedia zudem/besonders relevant
  - Präkombination verursacht extreme Redundanz im Kategoriesystem
  - Kategorieanpassungen sehr ineffizient
  - Inhomogenität kaum zu vermeiden



## Warum bisher Präkombination?

- Derzeitige Tools zur Schnittbildung kontraintuitiv
  - nicht für Endbenutzer („normale“ Leser) geeignet
  - (fast) keine Integration in die Endbenutzer-Oberfläche
- Ad hoc Schnittbildungen technisch sehr aufwändig (Ressourcenbedarf)
- Führung und Kontrolle bei Kategoriezuordnung
- Sammelleidenschaft? Gartenzaun-Mentalität?

# Gedanken/Anregungen

- Es gibt derzeit fast 190.000 Kategorien für 1,650,000 Artikel.
- Braucht es wirklich so viele?
- Gibt es sonstige Verbesserungsmöglichkeiten?



CC-BY-2.0 Micky Aldridge, Finland

- „Mehrwurzlicher“ Einstieg sehr anspruchsvoll
- Gedanken:
  - „Kategoriefilter“
    - als vorausgewählte Einschränkung bei Durchhangeln des Kategoriebaums
    - zusätzlich zu Suchfunktion [\[Parameter „incategory“?\] \[Bsp.\]...](#)
  - Angebot üblicher Schnittbildungen auf Kategorieseite (oder/und auf Portalseiten?)
- Auswahl einer Kategorie
  - Reine Kategorielliste eher ungeeignet [\[...\]](#)
  - Einstieg über Facetten?

- „Untergliederungsgesichtspunkt“
  - Facette („Kategoriefamilie“): „Sportart“
    - Foci (Kategorie): „Fußball“, „Radsport“
- Derzeitige Abbildung
  - teilweise, über Konventionen
    - über implizite Restriktion bei Unterkategorien
      - durchgehalten: [\[Kategorie:Geographisches Objekt\]](#)
      - uneinheitlich: [\[Kategorie:Biologie\]](#)
    - über Zwischenkategorien: [\[Kategorie:Person\]](#)



# Facetten, Designüberlegungen

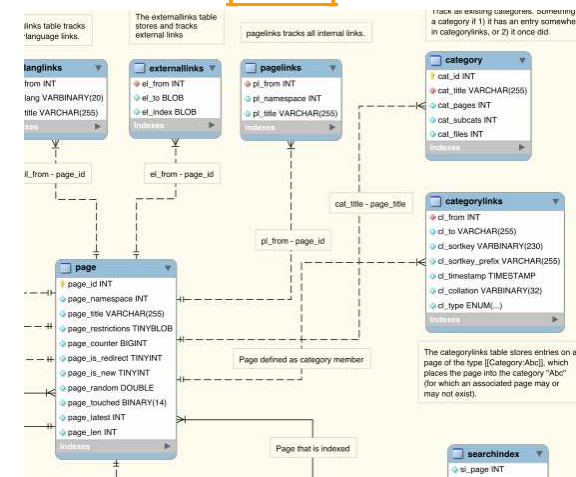
- Mehrstufig?
  - Unterkategorien müssen Spezialisierungen sein („ist ein“)
    - Beispiel: Sportart, Ballsportart, ...
- Überschneidungsfrei?
  - Beispiel: Europa, Asien, Eurasien, ...
- Vollständig?
  - Restekategorie „Sonstiges“?
- Abbildung
  - explizit?
  - über Konventionen?

# Optimierung des Datenmodells

- **Prinzip:**
  - „Vorabauswertung“ beim Anlegen/Ändern von Kategorien
- **Zusätzliche Hilfstabelle**
  - alle untergeordneten Kategorien zu einer Kategorie
  - auch die indirekt (transitiv) abhängen
  - Technische Hinweise:
    - auch `cd_sup=cd_dep` ist aufzunehmen
    - Verwendung von Kunstschlüsseln (IDs) eigentlich angebracht

## MediaWiki Datenmodell

[Dok.]



categorydependents
<code>cd_sup</code> VARCHAR(255)
<code>cd_dep</code> VARCHAR(255)

## Beispiel: Lebewesenartikeln mit Wartungsbaustein

- CatScan2
  - „2646 Kategorien durchsucht“ [\[Ausführung\]](#)
- SQL-Abfrage mit Verwendung der Hilfstabelle

```
select p.page_id, p.page_namespace, p.page_title
from categorylinks as cl
inner join page as p on p.page_id=cl.cl_from and p.page_namespace<>14
inner join categorydependents as cdep1
    on cl.cl_to=cdep1.cd_dep and cdep1.cd_sup='Lebewesen'
inner join categorydependents as cdep2
    on cl.cl_to=cdep2.cd_dep and cdep2.cd_sup='Wikipedia:Wartungskategorie-Bausteine'
```

## Explizite Assoziationen?

- Artikel zu Kategorie („ist ein“, „gehört thematisch zu“, ...)
  - Motivation:
    - Mögliche Alternative zu Objekt- und Themenkategorien?
    - Mehrfachzuordnungen zu Kategorien: „geboren in“, „gestorben in“, „lebte in“
    - Klarstellung des Bezugs [\[Kategorie:Person \(Berlin\)\]](#)
- Kategorie zu Kategorie
  - „ist ein“: Atoll → Geographisches Objekt
  - „ist Teil von“: Flusssystem Inn → Flusssystem Donau
  - „gehört thematisch zu“: Motorpresse → Kraftfahrzeug
  - Spezialfälle
    - Facetten! („Literaturgattung“)
    - vorweggenommene Artikel-Assoziationen! („gegründet 2004“)
    - sonstige Strukturierung (Beispiel: Begriffsklärung)



- Eigentliche Definition in Datenbank statt im Artikel-Quelltext
  - ähnlich Interwiki-Links über Wikidata
  - Kategorieänderungen deutlich problemloser
- Kategorie-Constraints
  - Beispiel: Artikel mit Fluss-Kategorie muss immer auch Kontinent- und Länderkategorie aufweisen
- Einfluss auf die Reihenfolge der Kategorieanzeige im Artikel (nach Assoziationen?)
- Namensraumbezug, Verwendungsbezug explizit (z.B. Wartungskategorien)?
- Zuordnung der Haupt- bzw. Übersichtsartikeln explizit?

- Flexibles System mit vielen Freiheitsgraden zu Beginn  
sicher kein Fehler...
  - aber: sollten aktuelle technische Möglichkeiten das Konzept  
dauerhaft prägen?
- Evolution von auf Konventionen beruhenden  
Systematiken zu explizit abgebildeten...
  - aber: nicht übertreiben!
- „Weniger ist mehr“ und Crowd-Sourcing?
- Zeitliche oder räumliche Schnittbildungen gezielt  
auflösen?
- Benutzergruppe zur Kategoriepflege?

# Danke...

- ...für's Interesse
- ... und Zuhören

