Wikipedia / WikiCon 2013

Kategoriesystem 2.0

November 2013

Inhalt

- 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

Kategorien: Grundlagen

- 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

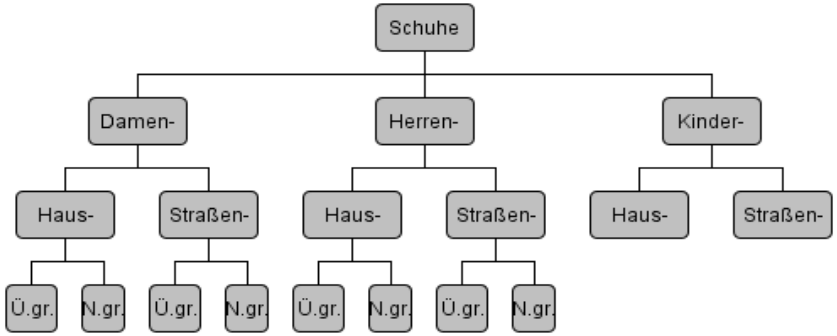
- Stärken
  - einfach
  - flexibel, vielseitig nutzbar
- Schwächen...
  - ...der MediaWiki-Implementierung
    - vielseitig nutzbar
    - verbesserungsbedürftige Benutzerschnittstelle
    - Kategorieänderungen sind aufwändig
  - ...der derzeitigen Systematik
    - inhomogen
    - arbeitsaufwändig und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

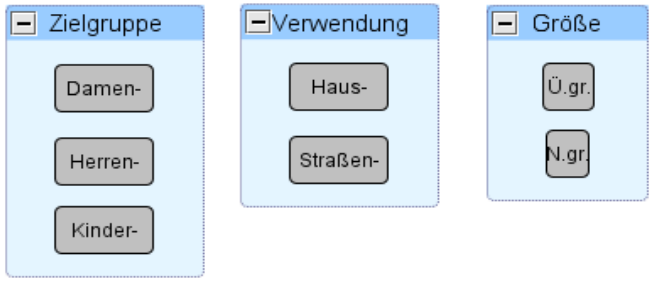
Wozu sind Kategorien gut?

- Abbildung von „Zuständigkeitsbereichen“ für Redaktionen und Projekte
- Ermöglichung von Auswertungen und Statistiken
- Steuerung von Massenbearbeitungen
- Recherche- und Orientierungshilfe
  - für Wikipedianer
  - für „normale“ Leser [wirklich?]
- …

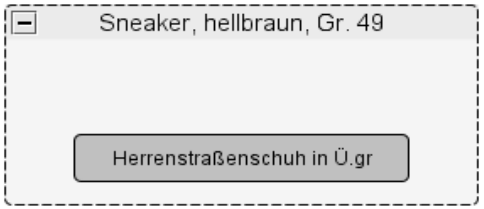
Präkombination vs. Postkombination

- WP:Kategorien: [WP]
  „Grundlegendes: ...Facettenklassifikation...“ [WP]





Präkombination



Postkombination

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

Präkombination vs. Postkombination

- Tatsächlich?
  - Beispiel: Kategorie:Geographisches Objekt [WP]

► Geographisches Objekt nach Staat (24 K)
*
► Glaziologisch geprägtes geographisches Objekt (3 K, 6 S)
► Humangeographisches Objekt (39 K, 10 S)
A
► Atoll (1 K, 286 S)
B
► Becken (3 K, 1 S)
► Berg (5 K, 2 S)
► Bergwerk (7 K, 4 S)
► Biotop (2 K, 10 S)
► Bucht (5 K, 10 S)
D
► Damm (34 S)
► Deich (11 S)
► Drumlin (19 S)
► Düne (1 K, 15 S)
E
► Einschlagkrater (Erde) (6 K, 3 S)
► Einzelpflanze (1 K, 2 S)
F
► Felsen (5 K, 62 S)
► Fluss (4 K, 29 S)
► Flussdelta (16 S)
► Geest (21 S)
► Gletscher (5 K)
H
► Halbinsel (3 K, 4 S)
► Höhle (4 K, 13 S)
I
▼ Insel (8 K, 31 S)
► Insel nach Eigenschaft (7 K)
► Insel nach Gewässertyp (3 K)
► Insel nach Inselgruppe (67 K)
▼ Insel nach Kontinent (8 K)
► Insel ohne Kontinentalbezug (2 K, 75 S)
► Insel (Afrika) (263 S)
► Insel (Antarktika) (100 S)
► Insel (Asien) (754 S)
► Insel (Australien und Ozeanien) (716 S)
► Insel (Europa) (1.627 S)
► Insel (Nordamerika) (1 K, 692 S)
► Insel (Südamerika) (130 S)
► Insel nach Region (4 K)
▼ Insel nach Staat (129 K)
► Insel nach abhängigem Gebiet (8 K)
► Insel (Ägypten) (7 S)
► Insel (Äquatorialguinea) (6 S)
► Insel (Albanien) (1 S)
► Insel (Antigua und Barbuda) (3 S)
► Insel (Argentinien) (3 S)
► Insel (Aserbaidschan) (1 S)
► Insel (Australien) (7 K, 2 S)
► Insel (Bahamas) (17 S)
► Insel (Bahrain) (6 S)
► Insel (Bangladesch) (3 S)
► Insel (Barbados) (2 S)
► Insel (Belize) (6 S)
► Insel (Bolivien) (6 S)
► Insel (Botswana) (1 S)
► Insel (Brasilien) (18 S)
► Insel (Bulgarien) (7 S)
► Insel (Chile) (24 S)
► Insel (Republik China) (1 K, 1 S)
► Insel (China) (1 K, 20 S)
► Insel (Costa Rica) (4 S)
► Insel (Dänemark) (111 S)
► Insel (Demokratische Republik Kongo) (1 S)
► Insel (Deutschland) (14 K, 2 S)
O
► Oase (7 K, 25 S)
P
► Pass (5 K, 2 S)
Q
► Quelle (6 K, 43 S)
R
► Region (20 K, 25 S)
► Rohstoffvorkommen (7 K, 39 S)
S
► Sandbank (1 K, 31 S)
▼ See (6 K, 8 S)
► See nach Eigenschaft (4 K)
▼ See nach Kontinent (7 K)
► Stausee nach Kontinent (6 K)
► See in Afrika (145 S)
► See in Amerika (2 K)
► See in Antarktika (13 S)
► See in Asien (211 S)
► See in Australien und Ozeanien (298 S)
► See in Europa (3.015 S)
▼ See nach Staat (125 K)
► Stausee nach Staat (27 K)
► See in Afghanistan (4 S)
► See in Ägypten (7 S)
► See in Albanien (13 S)
► See in Algerien (3 S)
► See in Angola (3 S)
► See in Argentinien (24 S)
► See in Armenien (2 S)
► See in Aserbaidschan (1 K, 1 S)
► See in Äthiopien (26 S)
► See in Australien (7 K, 5 S)
► See in Belgien (2 K)
► See in Bolivien (26 S)
► See in Bosnien und Herzegowina (9 S)
► See in Brasilien (1 K)
► See in Bulgarien (19 S)
► See in Burkina Faso (7 S)
► See in Burundi (3 S)
► See in Chile (21 S)
► See in China (1 K, 46 S)

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, Finnland

Intuitivere Oberfläche

- „Mehrwurzliger“ 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 Kategorieliste eher ungeeignet [...]
  - Einstieg über Facetten?

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]

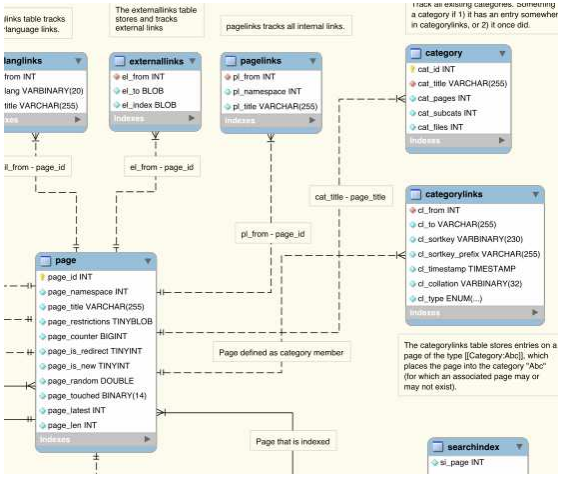
Facetten, Designüberlegungen

- 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 Kunst-schlüsseln (IDs) eigentlich angebracht

MediaWiki Datenmodell [Dok.]



categorydependents
cd_sup VARCHAR(255)
cd_dep VARCHAR(255)

Optimierung des Datenmodells

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)

Weitere Gedanken

- 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?

Resümee

- 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