

Seminar: XML und Datenbanken

XML-Indexierung

Fabian Fichter
1.2.2008

Übersicht

- Einleitung
- Werteindex
- Volltextindex
- Strukturindex / hybrider Index
 - Index Fabric
 - BUS
 - SphinX
 - CAS Index
- Zusammenfassung

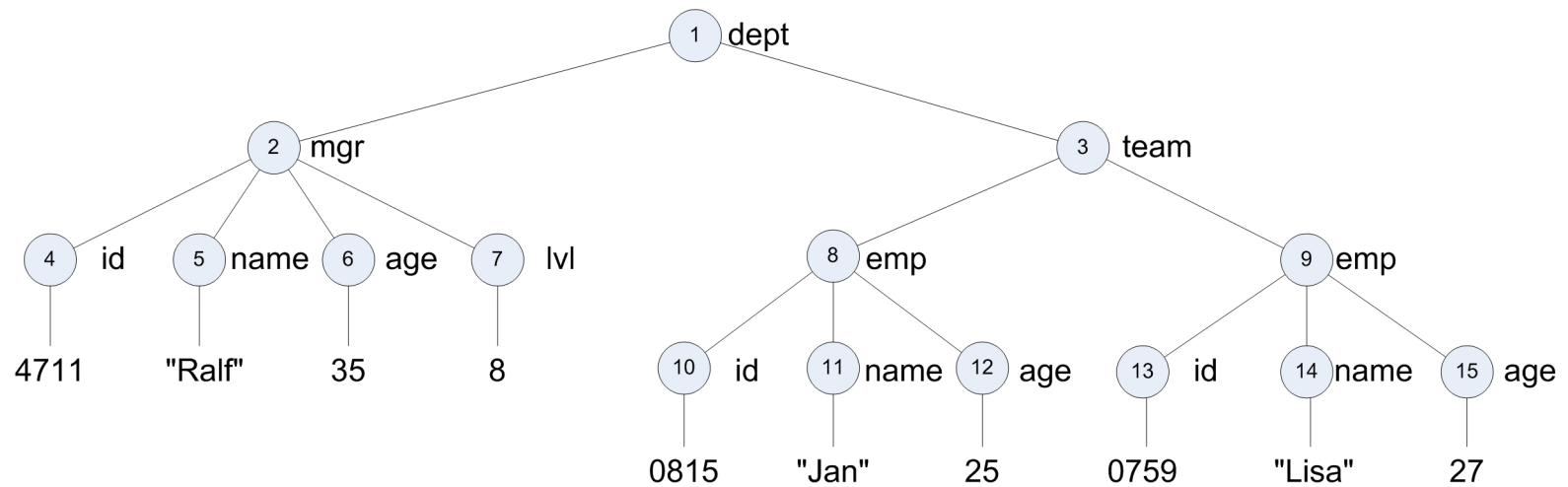
Einleitung – Index

- Zugriffsstruktur auf die Daten
- Ziel: schnelle Anfrageverarbeitung
- Grundlegende Datenstrukturen:
 - Bäume
 - Hashtabellen

Einleitung – Kriterien

- Anwendbarkeit
- Erweiterbarkeit
- Skalierbarkeit
- Auswertungszeit
- Speicherplatz
- Updates

Einleitung – XML-Graph



Wertindex

- Numerische Werte
- Elementnamen, Element Ids
- Attributwerte
- Geeignete Indexstrukturen
 - Dynamische Indexbäume wie B*-Baum
 - Hashverfahren wie das erweiterbare Hashing

Volltextindex

- Indexiert werden Wörter der natürlichen Sprache
- Kommt aus dem Information Retrieval
- Vorher findet Normalisierung statt
 - Erkennung von Interpunktionszeichen
 - Eliminierung von Wörtern mit geringer Bedeutung
 - Linguistische Normalisierung oder Grundformbildung

Volltextindex – Datenstrukturen

- Anfragetypen:
 - Stichwortsuche und Boolesches Retrieval
 - Mustersuche
 - Phrasensuche
 - Ranking

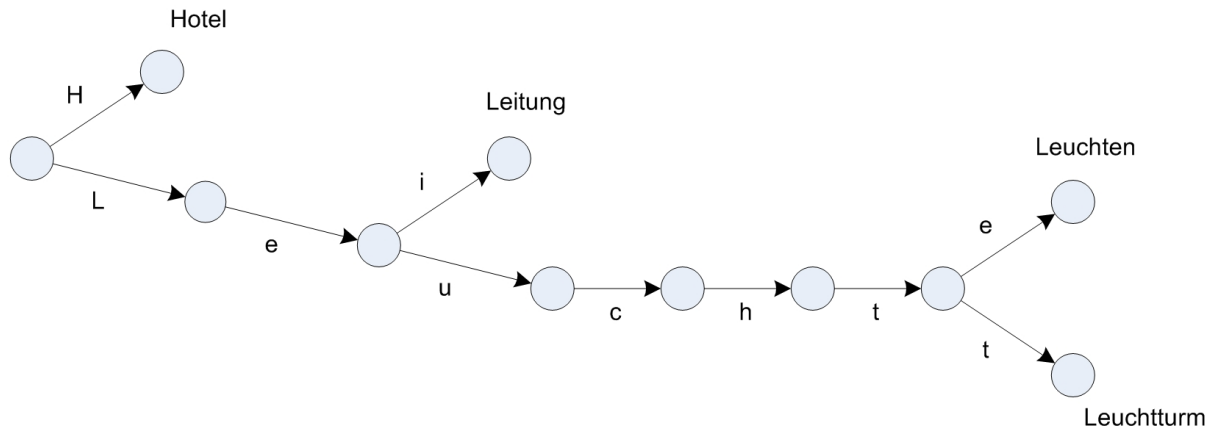
Volltextindex – Invertierte Datei

Worte	Wortposition
6	{ 11 }
150	{ 4 }
400	{ 7 }
200.000€	{ 14 }
haus	{ 1, 21 }
beschreibung	{ 16, 20 }
gesamtfläche	{ 6, 8 }
lage	{ 19 }
preis	{ 13, 15 }
quadratmeter	{ 2, 9 }
ruhige	{ 18 }
schön	{ 17 }
wohnfläche	{ 3, 5 }
zimmer	{ 10, 12 }

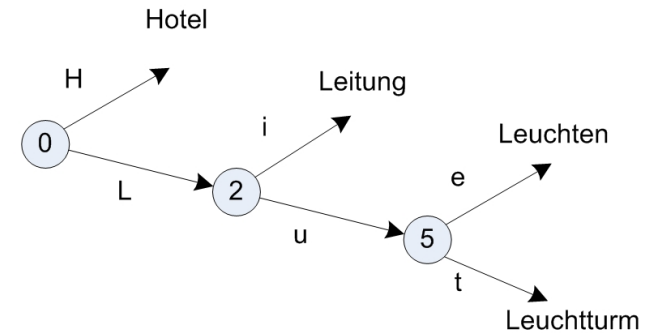
```
<haus>  
  <quadratmeter>  
    <wohnfläche>150</wohnfläche>  
    <gesamtfläche>400</gesamtfläche>  
  </quadratmeter>  
  <zimmer>6</zimmer>  
  <preis>200.000€</preis>  
  <beschreibung>schön, ruhige Lage</beschreibung>  
</haus>
```

Volltextindex – Tries, Patricia-Bäume, Suffixbäume

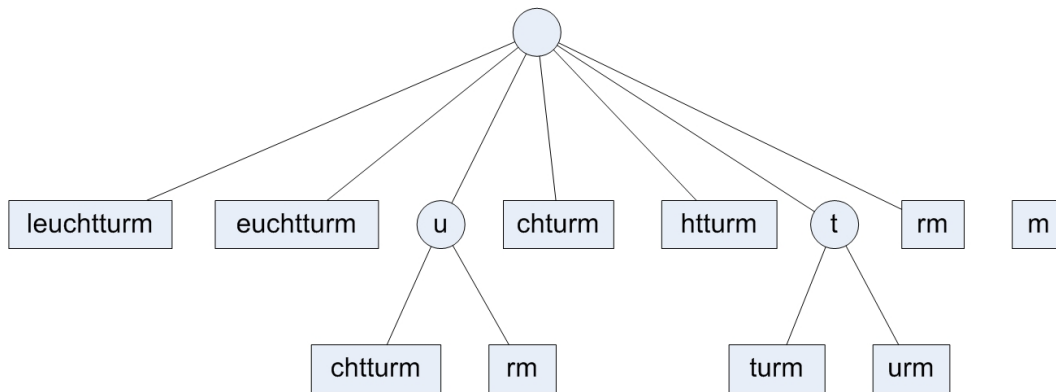
Trie



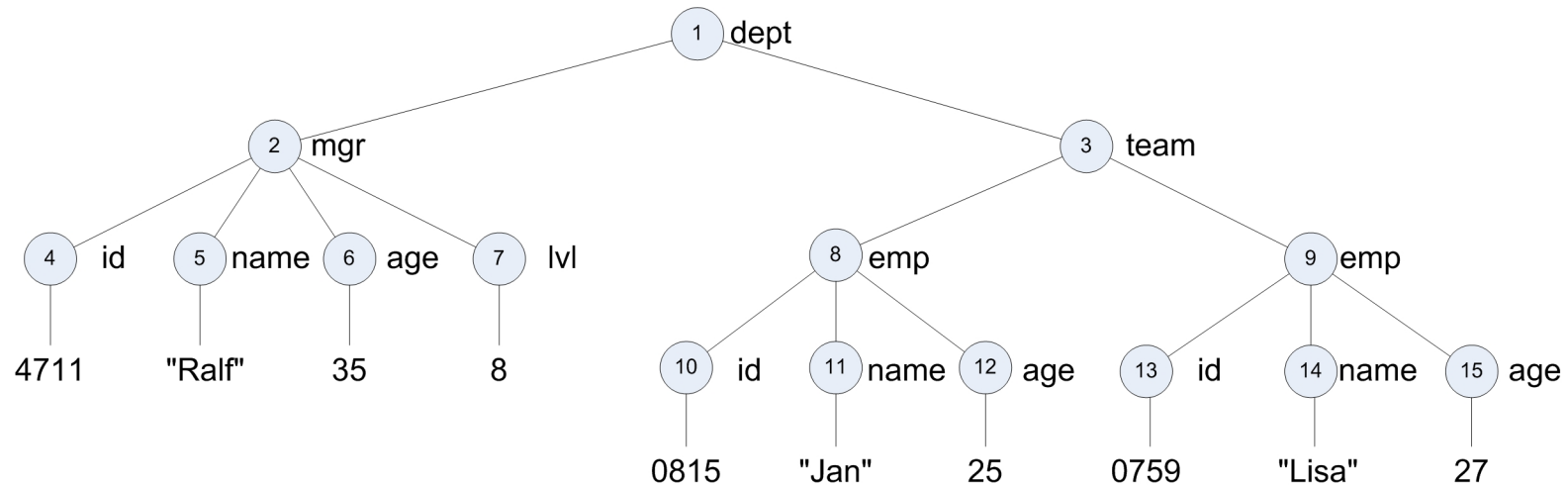
Patricia-Baum



Suffixbaum



Strukturindex / hybrider Index

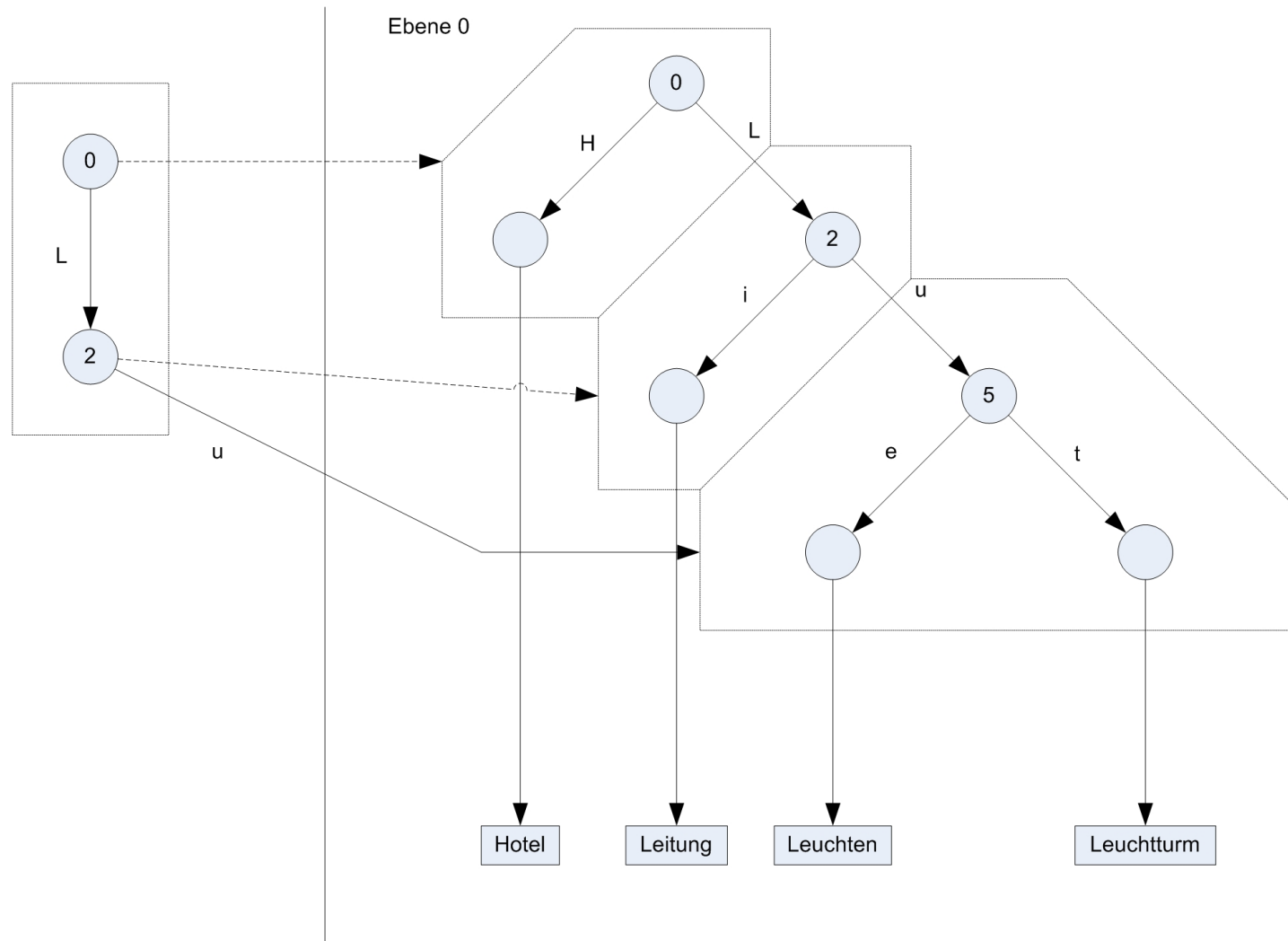


- /dept/mgr/id
- //age
- //emp[name="Jan"]

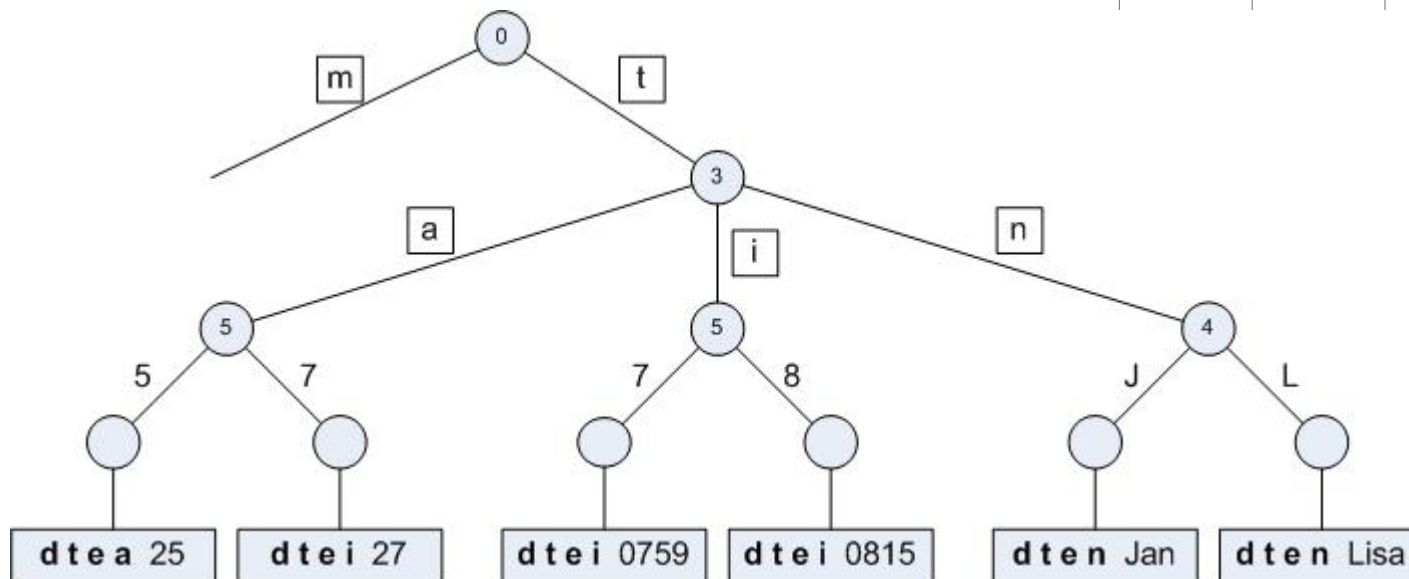
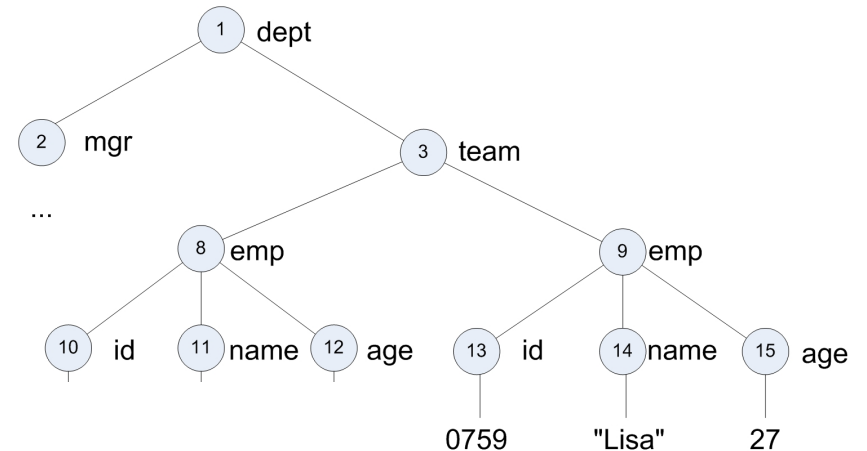
Hybride Indexe – Index Fabric

- Basiert auf Patriciaabäumen
- Pfade als Zeichenkette kodiert
- Spezielle Bezeichner für Tags
- Zwei Varianten:
 - Raw Paths
 - Refined Paths

Hybrider Index – Index Fabric



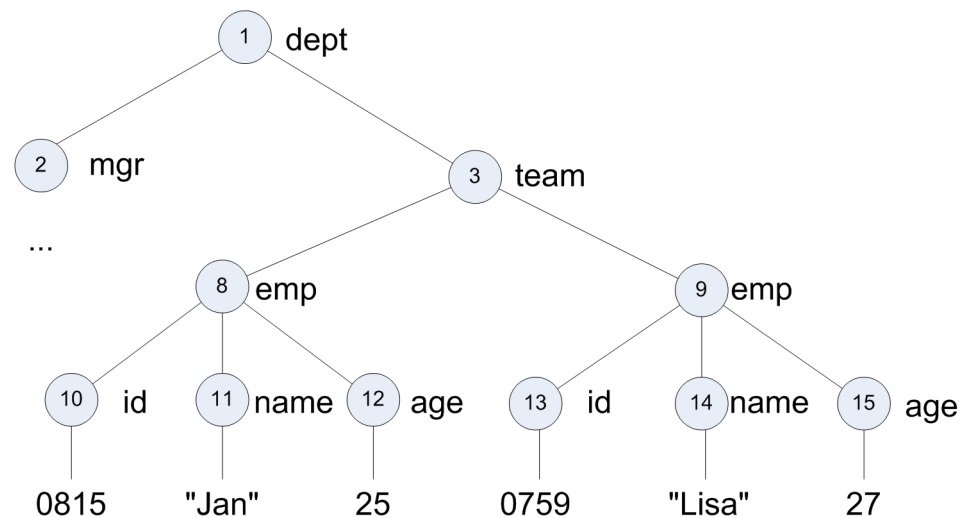
Hybrider Index – Index Fabric



- Tags werden durch Sonderzeichen kodiert

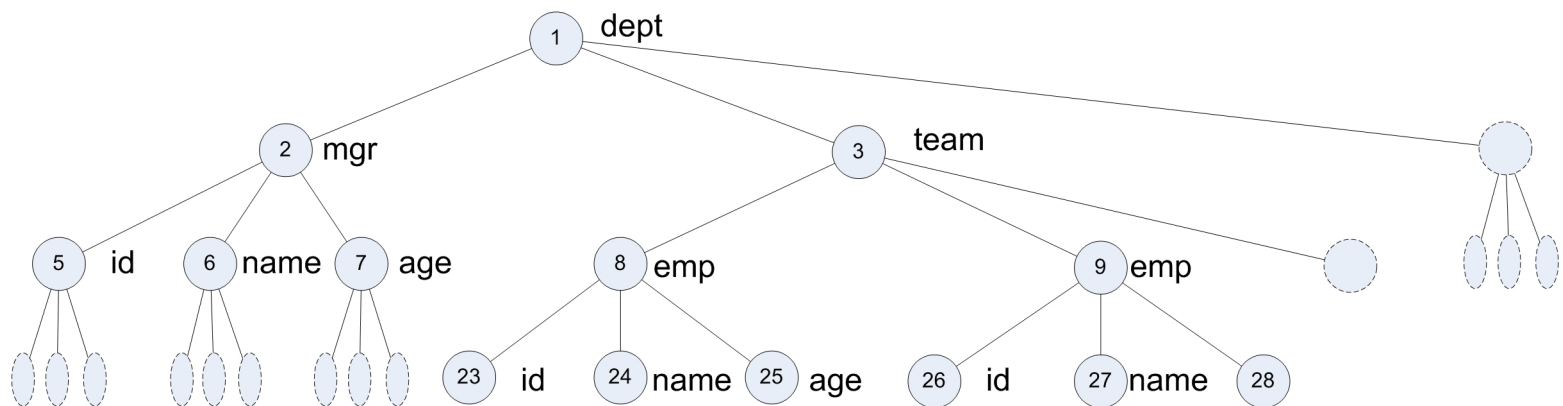
Hybrider Index – Index Fabric

- /dept/mgr/name
- Refined Paths ermöglichen Mustersuche
 - Muster erhält Sonderzeichen



Hybrider Index – BUS

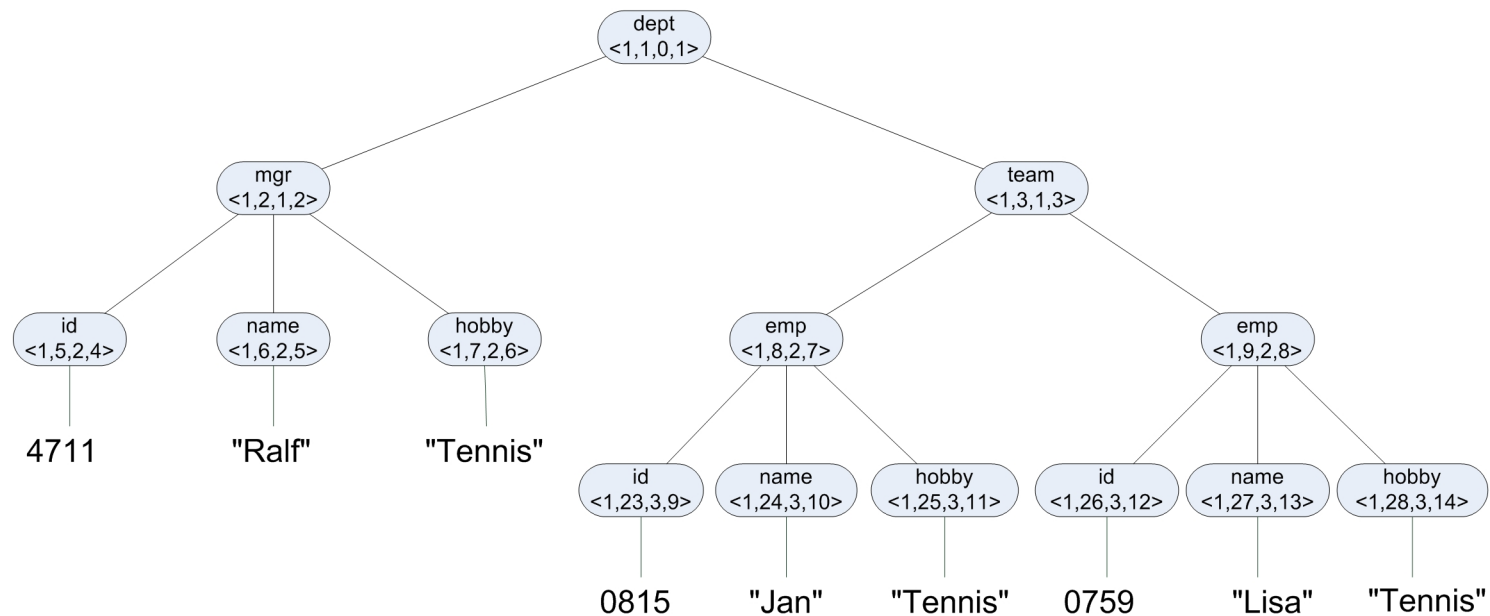
- Baumbasierte Indexierung
- Spezielle Nummerierung (UID)
- Nutzt invertierte Datei zum speichern
- Formel für Vaterknoten: $Parent(i) = \frac{i-2}{k} + 1$



Hybrider Index - BUS

- GID-Nummerierung
 - Dokumentnummer
 - UID des Elements
 - Elementlevel
 - Elementtyp

Hybrider Index – BUS



Invertierte Datei

0759	(1,<1,26,3,12>)
0815	(1,<1,23,3,9>)
4711	(1,<1,5,2,4>)
Lisa	(1,<1,27,3,13>)
Jan	(1,<1,24,3,10>)
Ralf	(1,<1,6,2,5>)
Tennis	(1,<1,7,2,6>), (1,<1,25,3,11>), (1,<28,3,14>)

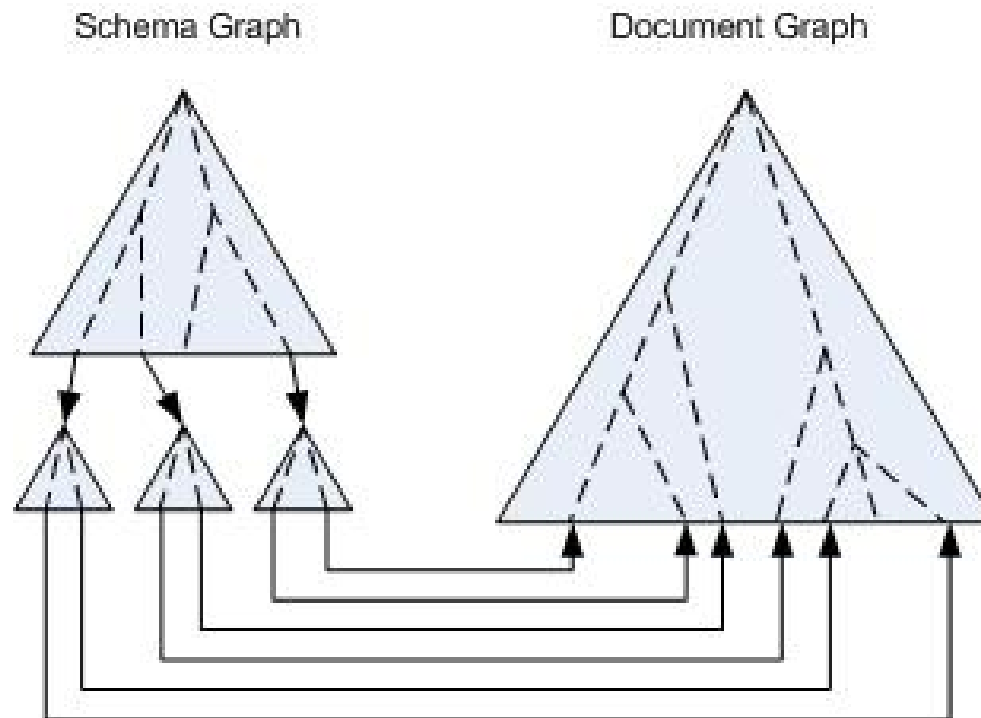
Akkumulator-Tabelle

(1,3)	2
(1,8)	1
(1,9)	1
(1,25)	1
(1,28)	1

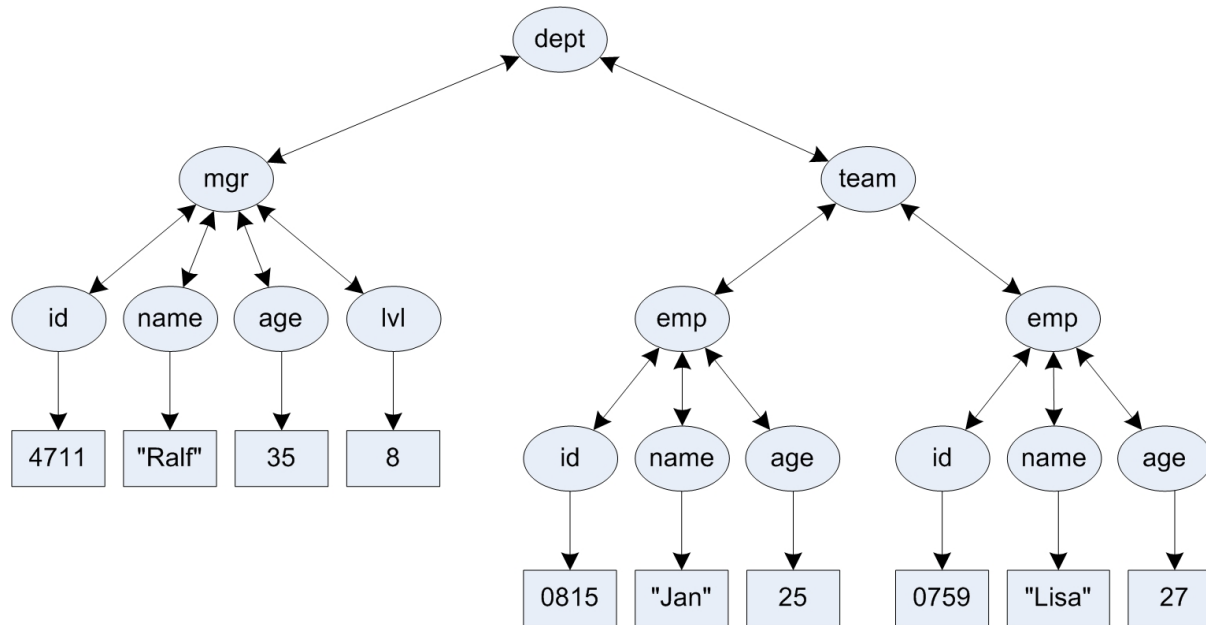
Anfrage: /dept/team[,„Tennis“]

Hybrider Index – SphinX

- Nutzt Schemainformation
- Unterteilt in Document Graph und Schema Graph



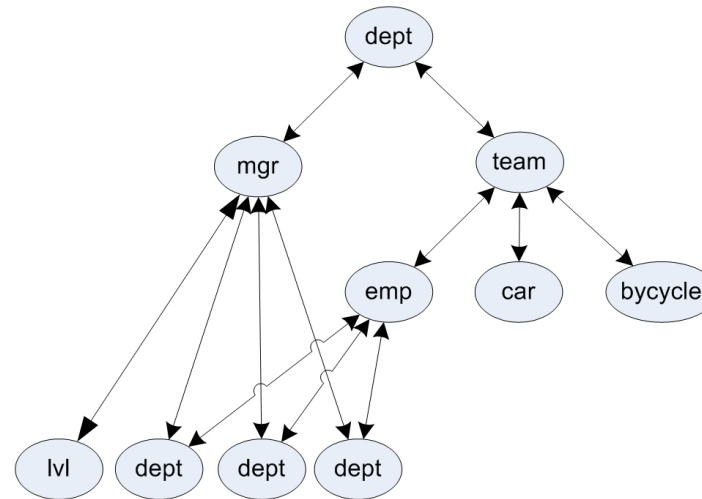
SphinX – Document Graph



- Gerichteter Graph
- Kann bidirektional navigiert werden

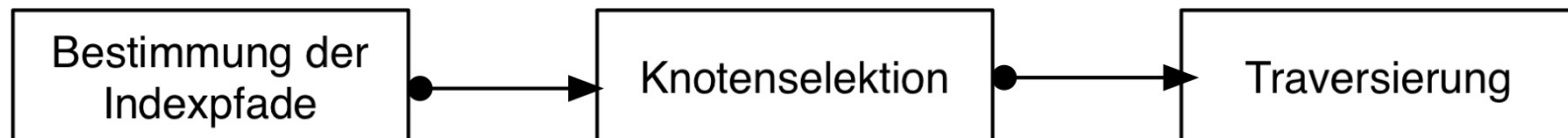
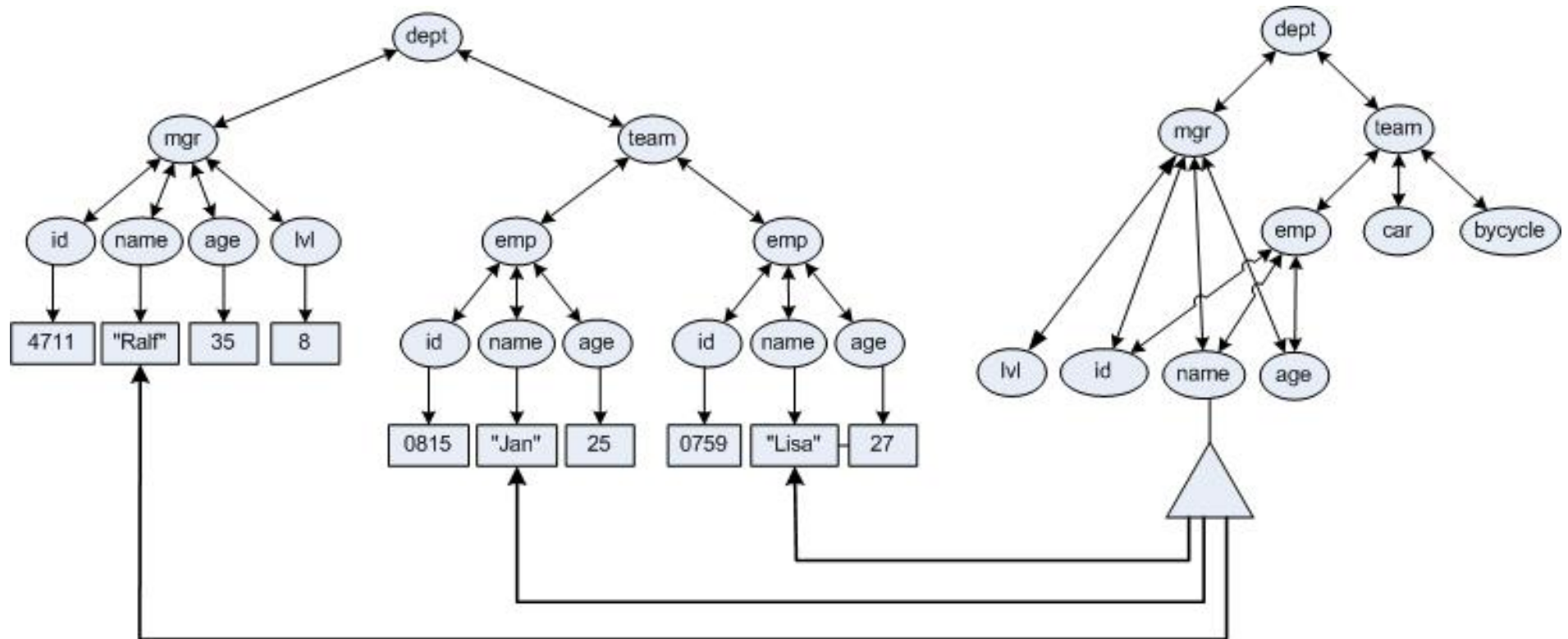
SphinX – Schema Graph

```
<!ELEMENT dept (mgr, team)>  
<!ELEMENT mgr (id, name, age, lvl)>  
<!ELEMENT team (emp*, (car | bycycle)?)>  
<!ELEMENT emp (id, name, age)>  
<!ELEMENT car (#PCDATA)>  
<!ELEMENT bycycle (#PCDATA)>  
<!ELEMENT id (#PCDATA)>  
<!ELEMENT name (#PCDATA)>  
<!ELEMENT age (#PCDATA)>  
<!ELEMENT lvl (#PCDATA)>
```



- Beschreibt die Schemainformation
- Quantifikatoren (a^* , a^+ , $a^?$) werden ignoriert
- Alternativen werden in Sequenzen umgewandelt

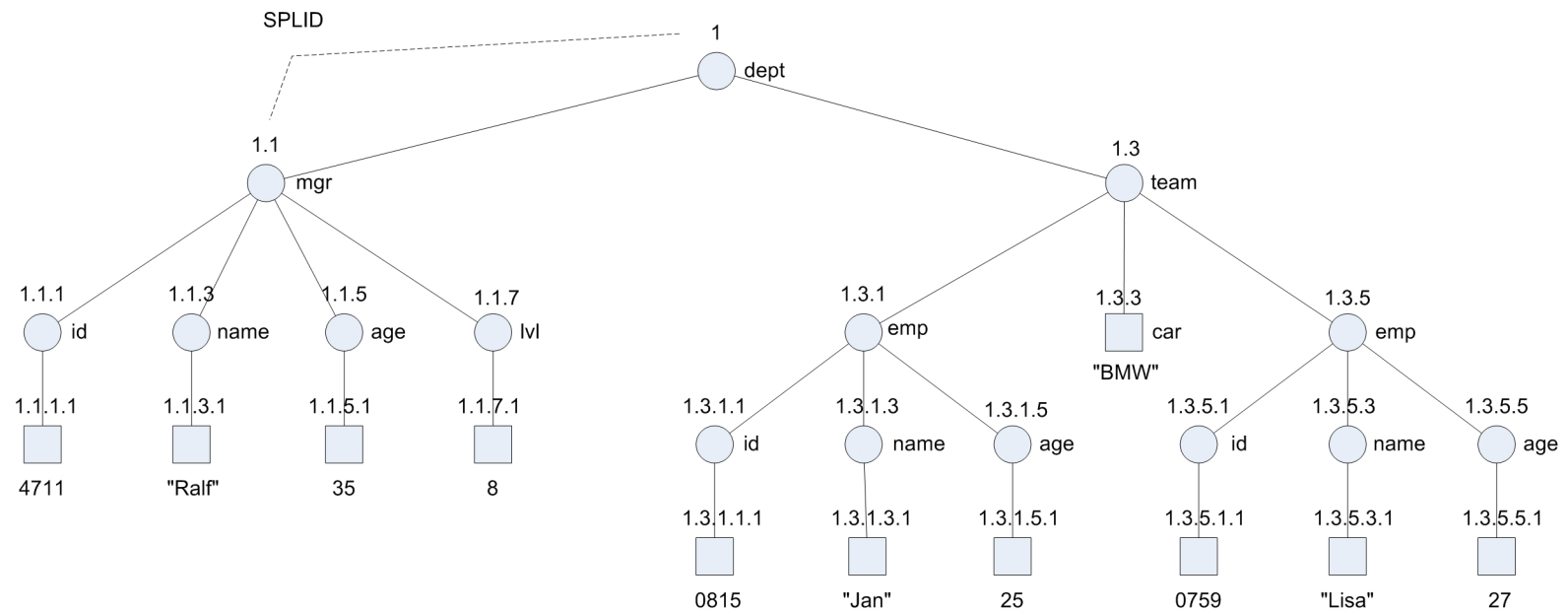
Sphinx – Suche



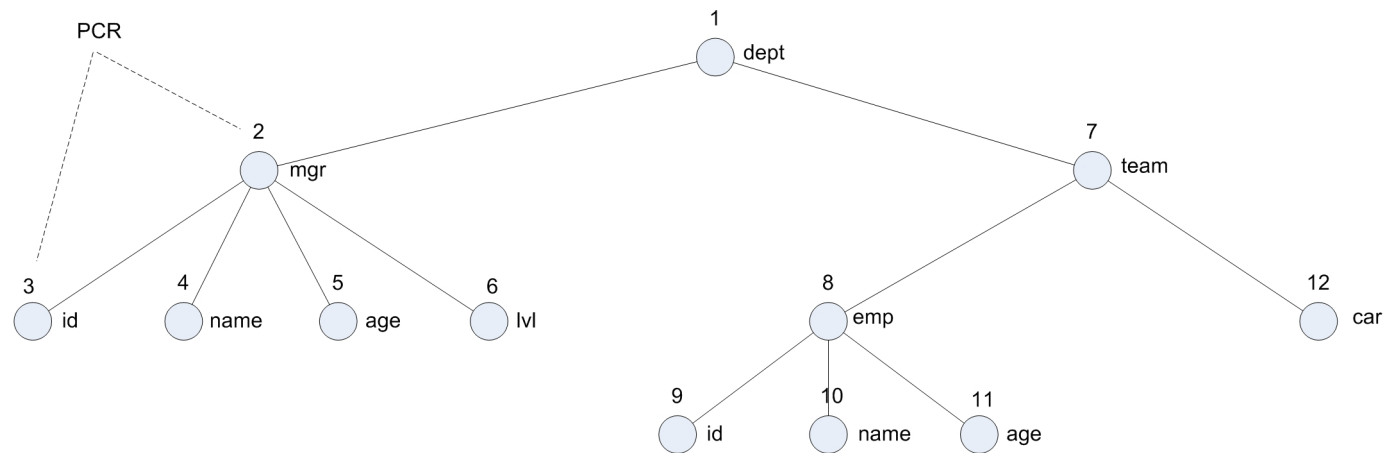
Hybrider Index – CAS-Index

- Spezielle Nummerierung
- Pfadübersicht
- Indexe für Mustersuche

CAS-Index – SPLID



CAS-Index - Pfadübersicht



CAS-Index – Konstruktion

- $I_D(p, T, C)$
- p : //age, //name
- T : String, Integer, ...
- C : Clustering (SPLID, PCR)

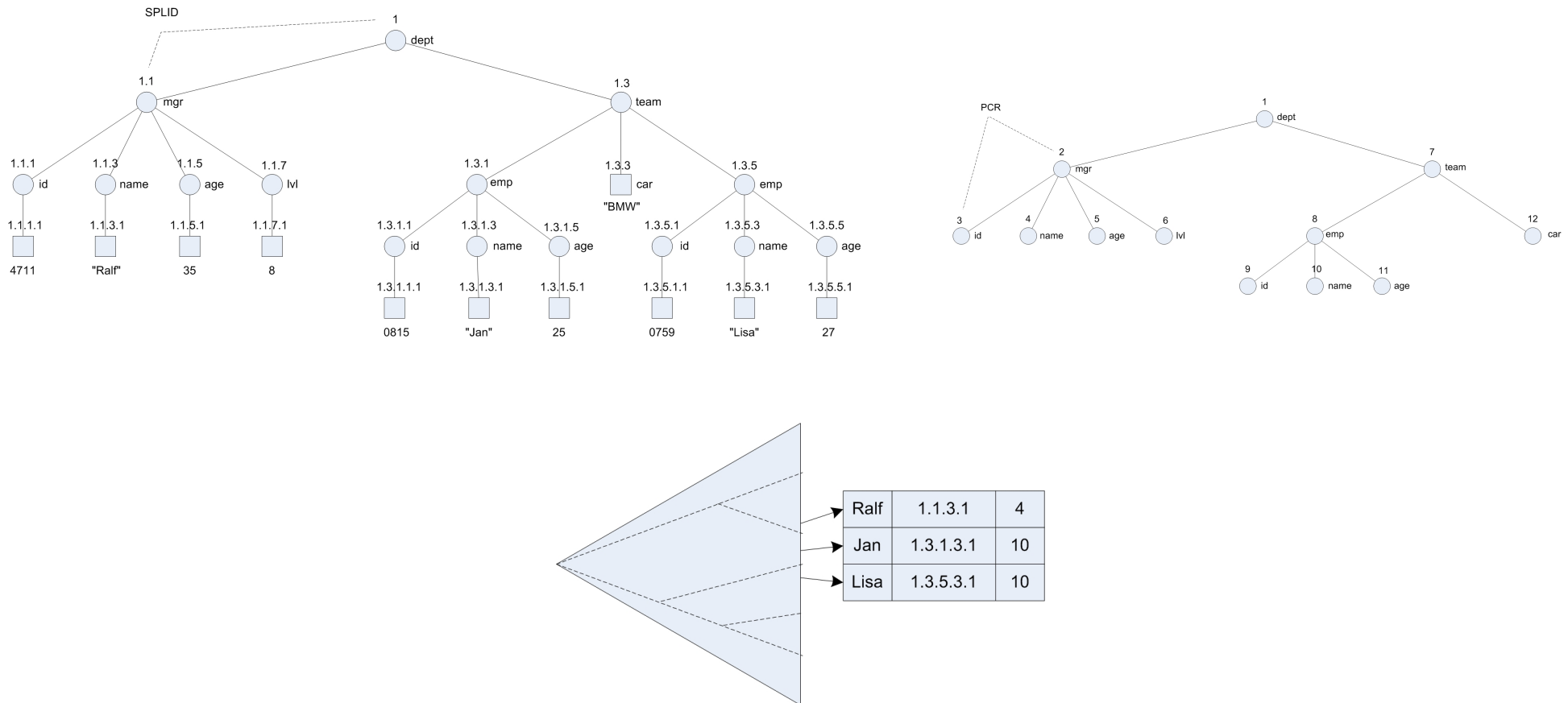
SPLID Clustering

Ralf	SPLID	4
------	-------	---

PCR Clustering

Ralf	PCR	4
------	-----	---

CAS-Index – Konstruktion



- $p = //name$, $T = \text{String}$, $C = \text{SPLID}$

Zusammenfassung

- Werteindex, Volltextindex, Strukturindex
- Index Fabric, BUS, SphinX, CAS-Index
- Aktuelles Forschungsgebiet
- Kommerzielle Implementierungen, z. B. von IBM und Oracle

ENDE