GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
ForMaT
UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative Neue Länder

UNIVERSITÄT LEIPZIG

Web Data Integration Lab: 2 Jahre praxisnahe Forschung

Prof. Dr. Erhard Rahm

21.12.2011

WDI Lab

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
ForMaT UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative Neue Länder

- Innovationslabor an der Univ. Leipzig zur semantischen **Web Daten Integration**
- Gefördert durch BMBF
  - 2009: Initiierungsphase
  - seit Jan. 2010 Vollausbau
- 10 Forscher in 3 Arbeitsgruppen + studentische Mitarbeiter

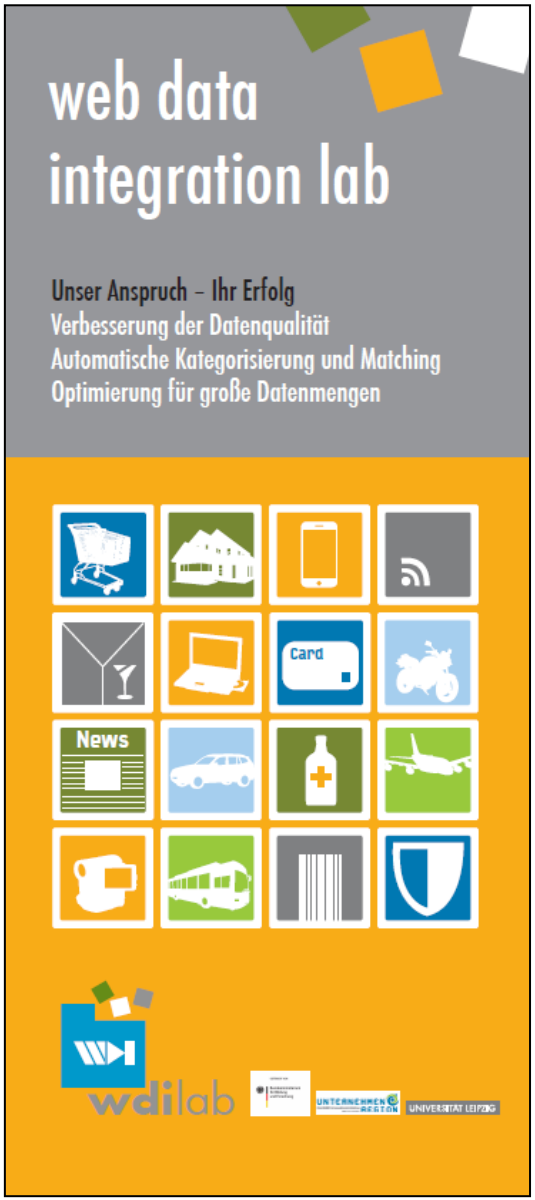
WDI-Lab: Zielsetzungen

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
ForMaT
UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative Neue Länder

- Semiautomatische Integration von heterogenen (Web-)Daten in hoher Datenqualität
- Schnellere Entwicklung von Datenintegrationslösungen im Vergleich zu konventionellen Ansätzen
- Weiterentwicklung vorliegender Forschungs-Prototypen für den Markteinsatz
- Anschub eines Startups

Beispiel-Anwendungen

- Online-Monitoring (z.B. Produkte, Flüge, Marken)
- Informationsanreicherung durch Web Content (z.B. Bewertungen, Herstellerinformationen)
- Automatische Integration großer Produktkataloge
- Automatische Kategorisierung von Produktangeboten
- Matching von Produkten und Angeboten



Leiter: Prof. Dr. Erhard Rahm

Teamstruktur

Projektunterstützung: Carina Röllig

AG 1: Workflowbasierte Datenintegration und Informationsextraktion
Leitung: Christian Wartner

AG 2: Schema- und Ontologie-Integration
Leitung: Sabine Maßmann

AG 3: Datenqualität und Objekt-Matching
Leitung: Hanna Köpcke

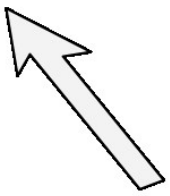
Workflowbasierte Datenintegration

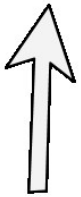
Anwendungen

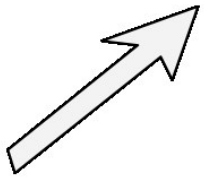
Mashups

Datenbeschaffung und Datenanreicherung

Web-Monitoring

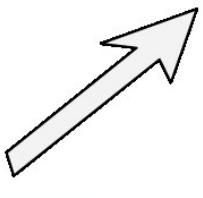






Ombat Tool-Suite

Datenquellen



Webseiten(Shops, Portale ...)

Web-Services

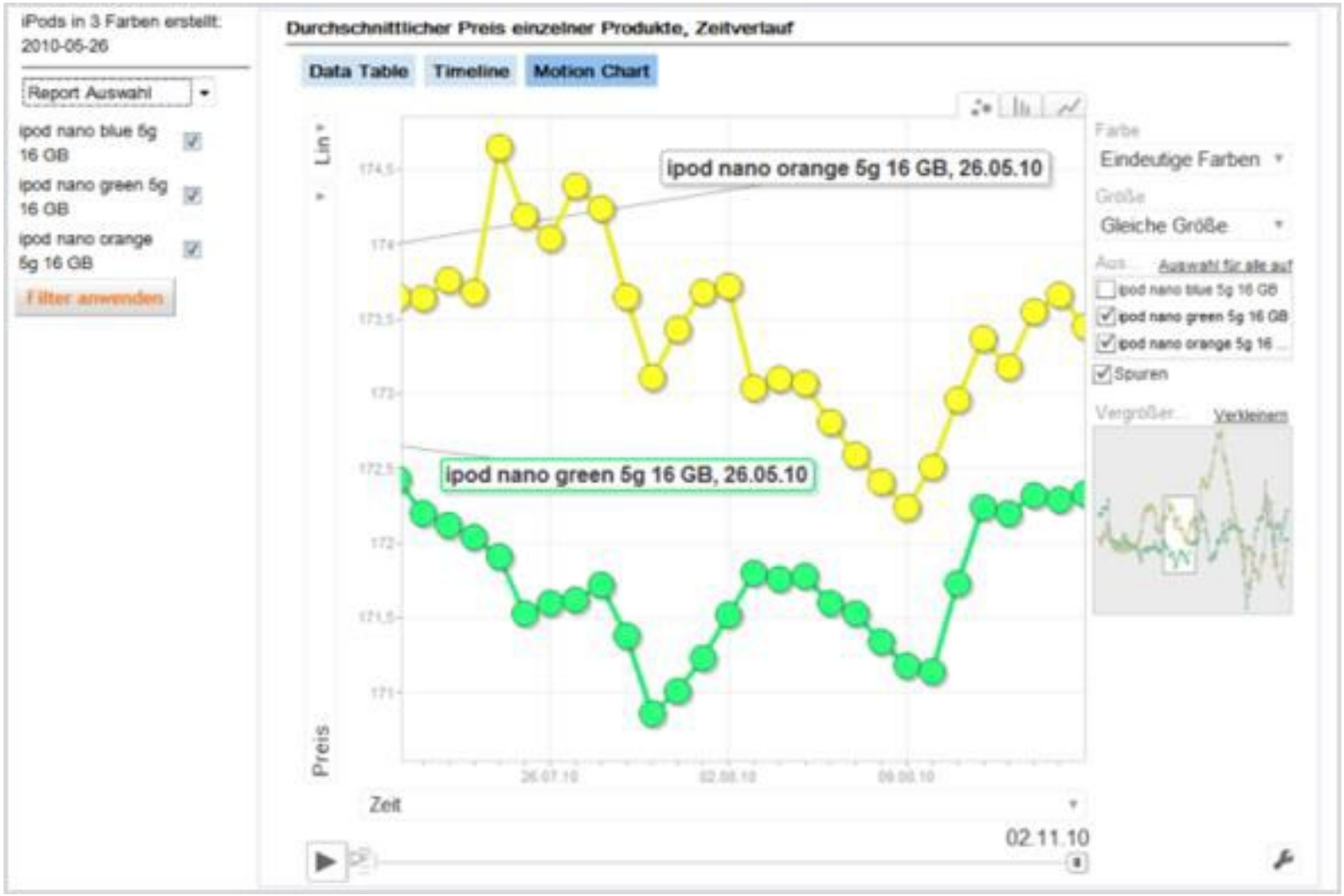
Unternehmenseigene Daten

Integrations-Workflows

- Mashup-ähnliche Workflows (Skripte) zur Integration von Daten unterschiedlichster Quellen
  - Daten aus Unternehmen und aus dem Web
  - Umfassende Unterstützung zur Webdatenextraktion
  - Optimierte Query-Techniken zur dynamischen Informationsbeschaffung „on-the-fly“
  - Nachbearbeitung der Ergebnisse, u.a. Matching

- Beispielanwendung: Produktmonitoring im Web

PROOF – Produktmonitoring im Web



Schema- und Ontologie-Integration

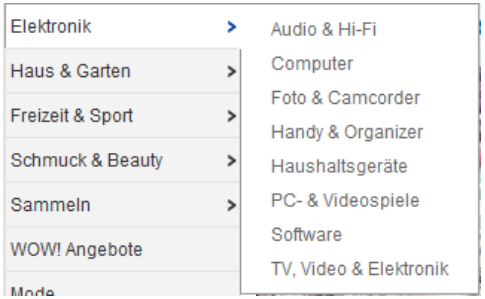
- Informationsstrukturierung durch Schemas und Ontologien
  - Datenbankschemas, z.B. Adressdaten

Vorname	Nachname	Geburtstag	Telefonnr	Emailadresse

  - XML-Schemas, z.B. Web-Service z.B. Amazon

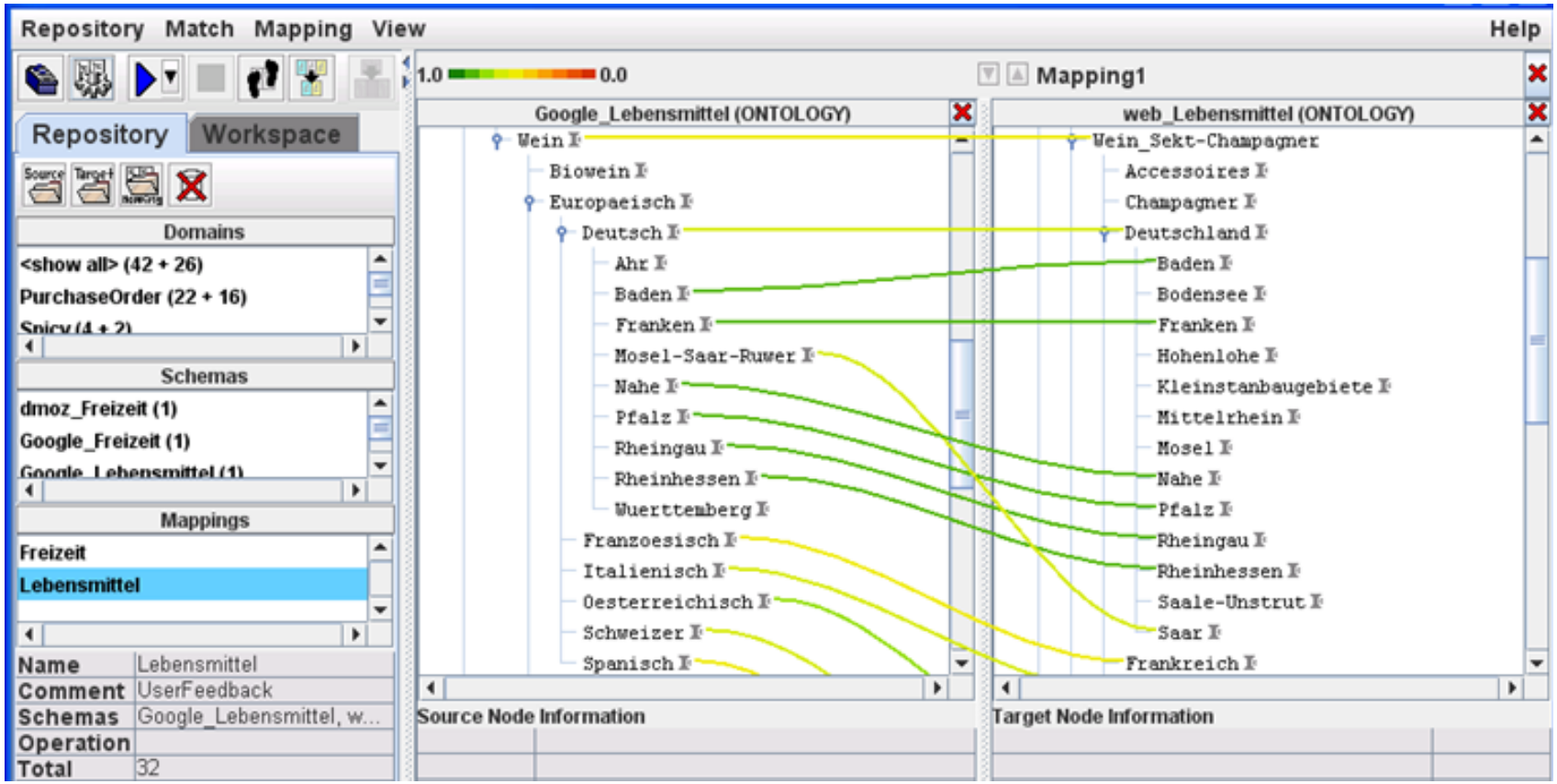

```
<ItemSearchResponse xmlns="http://webservices.amazon.co
<OperationRequest>...
<Items><TotalResults>126</TotalResults><TotalPages>13</
  <Item>
        <ASIN>3834804614</ASIN>
        <Author>Klaus Wüst</Author>
        <Edition>3., aktualisierte und erweiterte Aufl
        <ISBN>3834804614</ISBN>
        <FormattedPrice>EUR 29,95</FormattedPrice>
        <PublicationDate>2008-10-28</PublicationDate>
        <Publisher>Vieweg+Teubner Verlag</Publisher>
         ...
```

  - Kataloge, z.B. Produktkatalog



- Mappings notwendig zum Datenaustausch sowie zur Integration
- Hohe Komplexität durch semantische Heterogenität (Synonyme, Abkürzungen, Granularität) und regelmäßige Änderungen/Evolution

Matching mit COMA 3.0

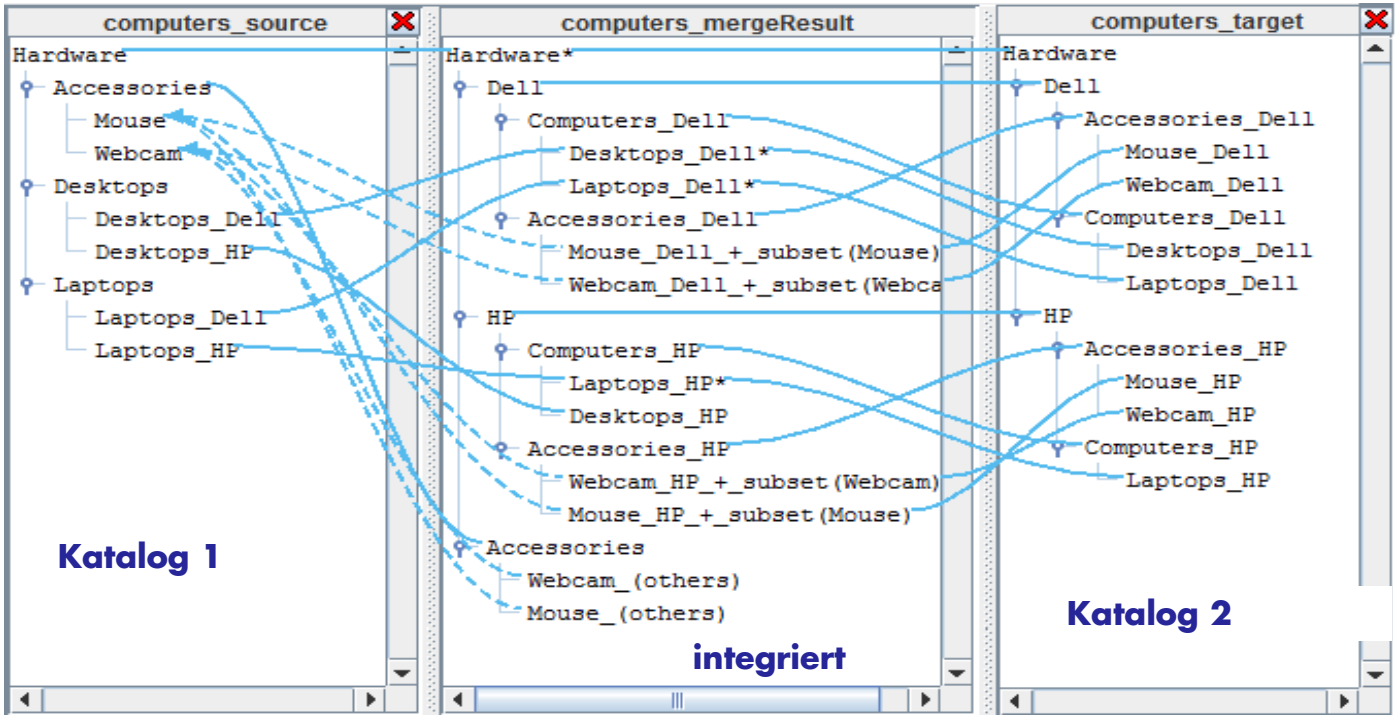


Unsere Lösung

- COMA 3.0 basiert auf international anerkannten Lösungen COMA und COMA++ zum (semi-)automatischem Schema- und Ontologie-**Matching**
  - COMA++ wird bereits an hunderten Forschungsinstituten genutzt
  - Matching von Schemas (XML, relational, CSV) und Ontologien (OWL, RDF)
  - Erfolgreiche Evaluierung für zahlreiche Einsatzfälle
  - Kombinierte Nutzung ausgefeilter Match-Algorithmen
  - Strategien zum Matchen von großen Schemas und Ontologien
  - Wiederverwendung bereits erstellter Mappings
  - Nutzbarkeit als Web Service
- **Integration** von Ontologien, z.B. Produktkatalogen
- GUI zur Konfigurierung und Nachbearbeitung

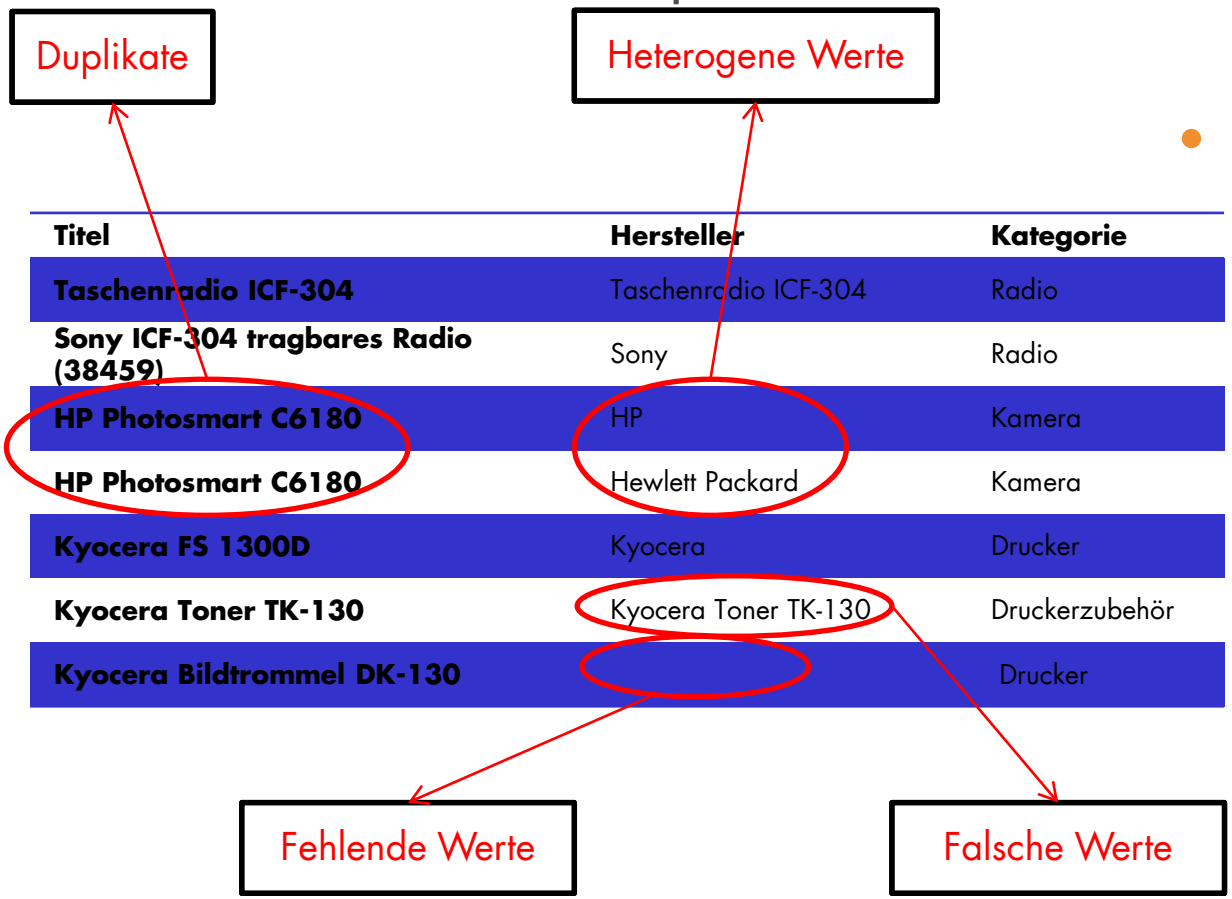
Katalogintegration

- (Semi-)Automatische Erstellung eines integrierten Katalogs
  z.B. Portale benötigen Produktkatalog, der alle Produkte von den Händlern umfassen soll und somit idealerweise deren Produktkategorien beinhaltet
- Algorithmus **ATOM** (Automatic Target-Driven Ontology Merging)



Datenqualität und Objekt-Matching

- Schlechte Datenqualität



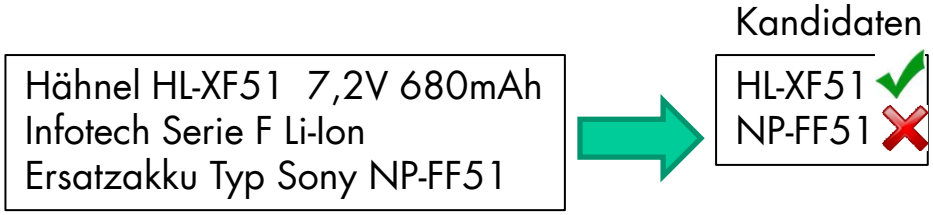
- Unsere Ansätze:
  - Datenbereinigung
  - Automatische Kategorisierung
  - Erkennung von Dubletten

Datenbereinigung

- Cleaning kritischer Attribute, z.B. von Hersteller- oder Markenangaben
  - Automatisierte Erstellung von Lookup-Tabellen mit Synonym-Verwaltung, zB für
    allied telesis, *allied telesyn*, *allied*, *allied telesyn international gmbh*
- Feature-Extraktion und -Verbesserung
  - Produkt-Ids, zB EAN, GTIN

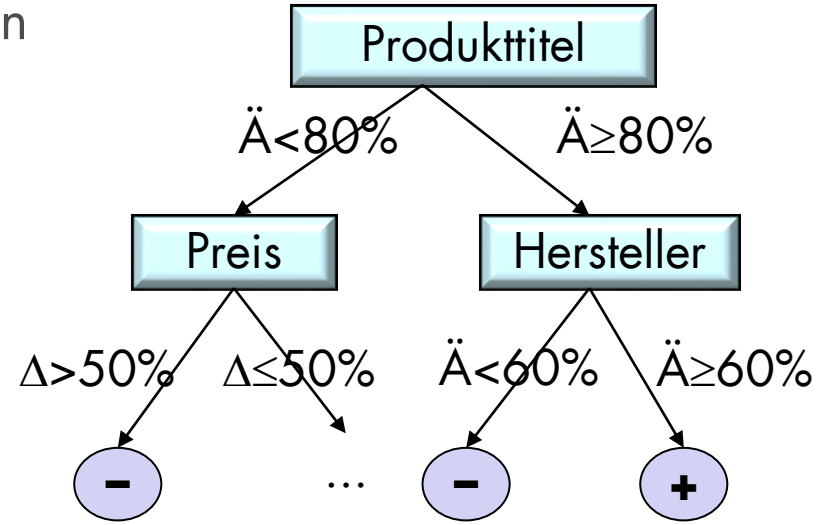
Title: 'Playstation 2 - Happy Pico Pad Retro mit 10 Spielen'
Description: 'SP-2370F/ P535Happy Pico Pad Retro, Lemon - Yellow, 10 Spiele/ Happy Pico Pad Retro, Lemon - Yellow, 10 Spiele/ Hersteller: Madrics MediaHersteller Originalnummer: 20567EAN-Code: 4041192205671 '

  - Produkt-Codes

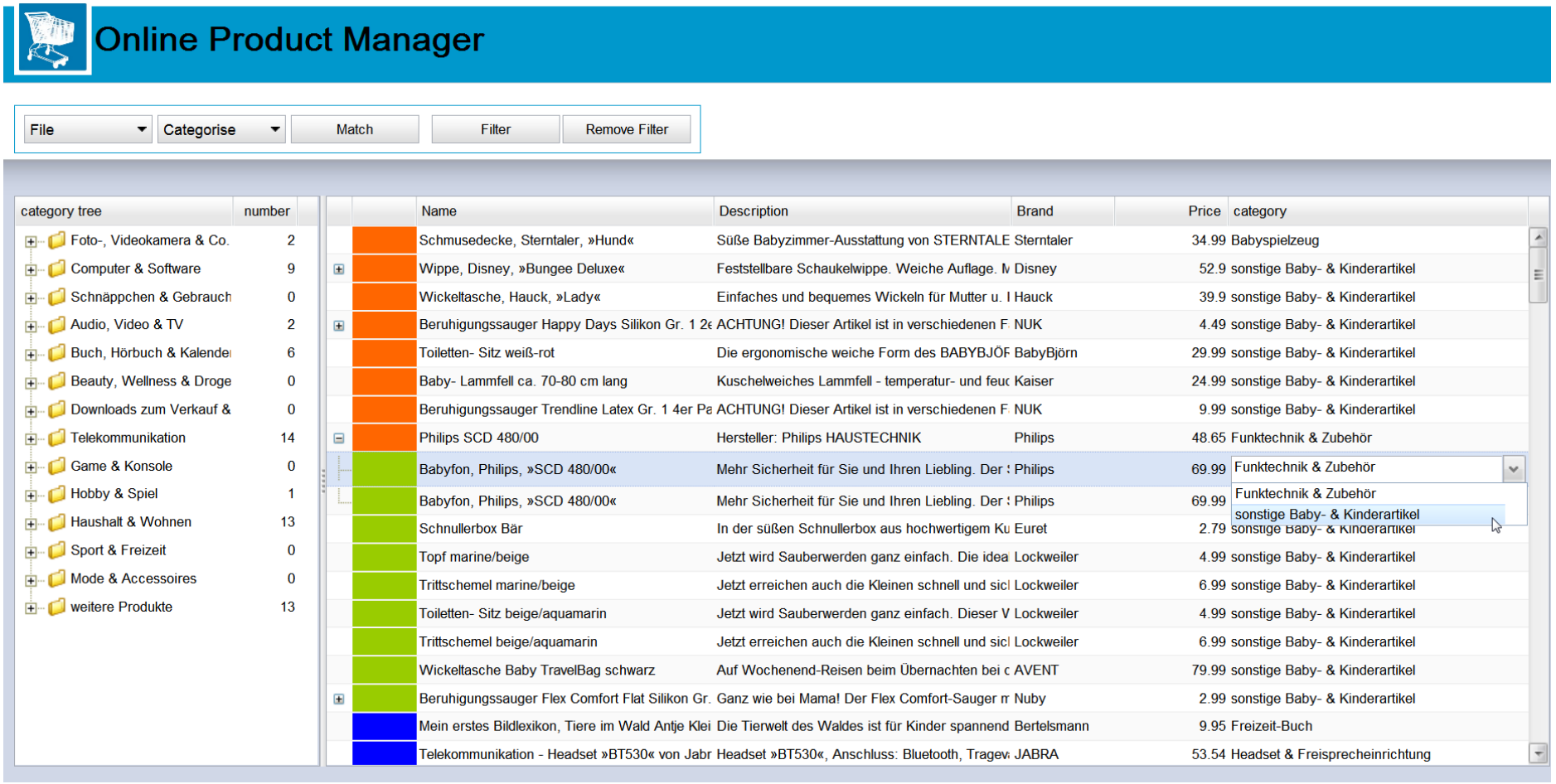


Erkennung von Duplikaten

- Identifikation von Objekten, die dasselbe Realweltobjekt repräsentieren
- Komplexe Strategien zur Kombination mehrerer Match-Techniken
- Nutzung von maschinellem Lernen, u.a. SVM, Entscheidungsbäume
- Performance-Optimierungen
- Cloud-Nutzung



Prototyp: Online Product Manager



Vorteile unserer Technologien

- (Semi-)automatisches Matching von Daten und Metadaten
- Strategien für große Datenmengen
- Branchenübergreifend einsetzbar
- Matchen von Produktangeboten
  - zu Produkten in einem Produktkatalog
  - mit anderen Angeboten
- Aufbau /Erweiterung von Produktkatalogen
- Nutzung von Trainingsdaten zur vereinfachten Konfigurierung
- Nutzung von Cloud-Infrastrukturen
- Umfassende Evaluierung und Vergleich mit Standard-Lösungen
- Hohe Effektivität und Effizienz

Ausblick

- Spinoff **Webdata Solutions GmbH** (ab Jan. 2012)

- WDI-Lab wird an der Universität Leipzig fortgeführt
  - weitere Forschung und Entwicklung
  - Industriekooperationen

Kooperationspartner (Auswahl)