



Web Data Integration Lab: 2 Jahre praxisnahe Forschung

Prof. Dr. Erhard Rahm

21.12.2011

WDI Lab

- 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

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

ForMaT

UNTERNEHMEN
REGION

Die BMBF-Innovationsinitiative
Neue Länder



WDI-Lab: Zielsetzungen

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

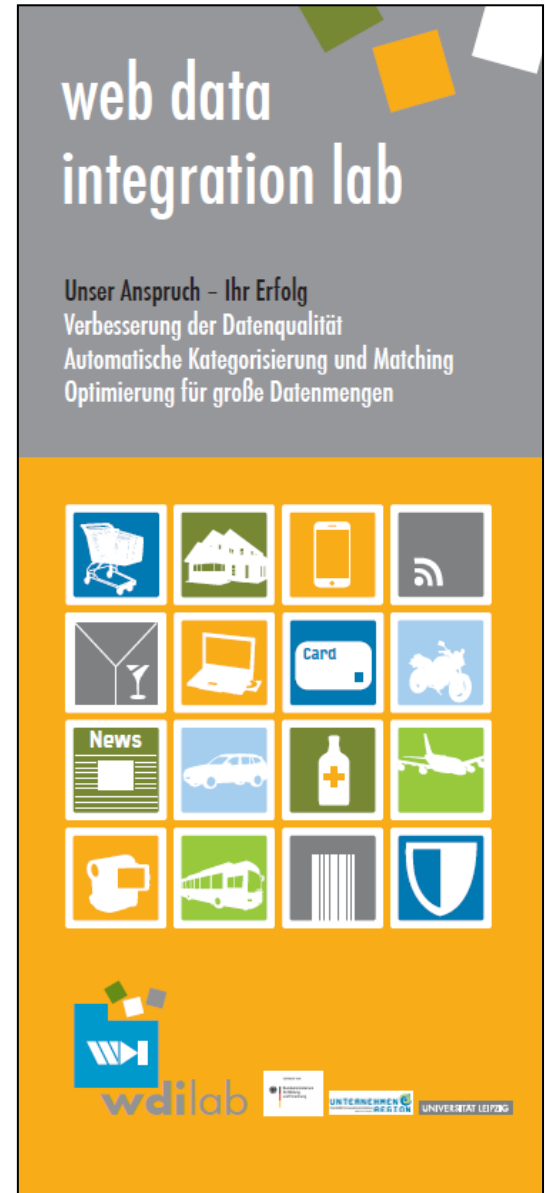
ForMaT

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






web data
integration lab

Unser Anspruch – Ihr Erfolg
Verbesserung der Datenqualität
Automatische Kategorisierung und Matching
Optimierung für große Datenmengen


wdilab

Teamstruktur

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

AG 1: Workflowbasierte
Datenintegration und
Informationsextraktion
Leitung: Christian Wartner

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

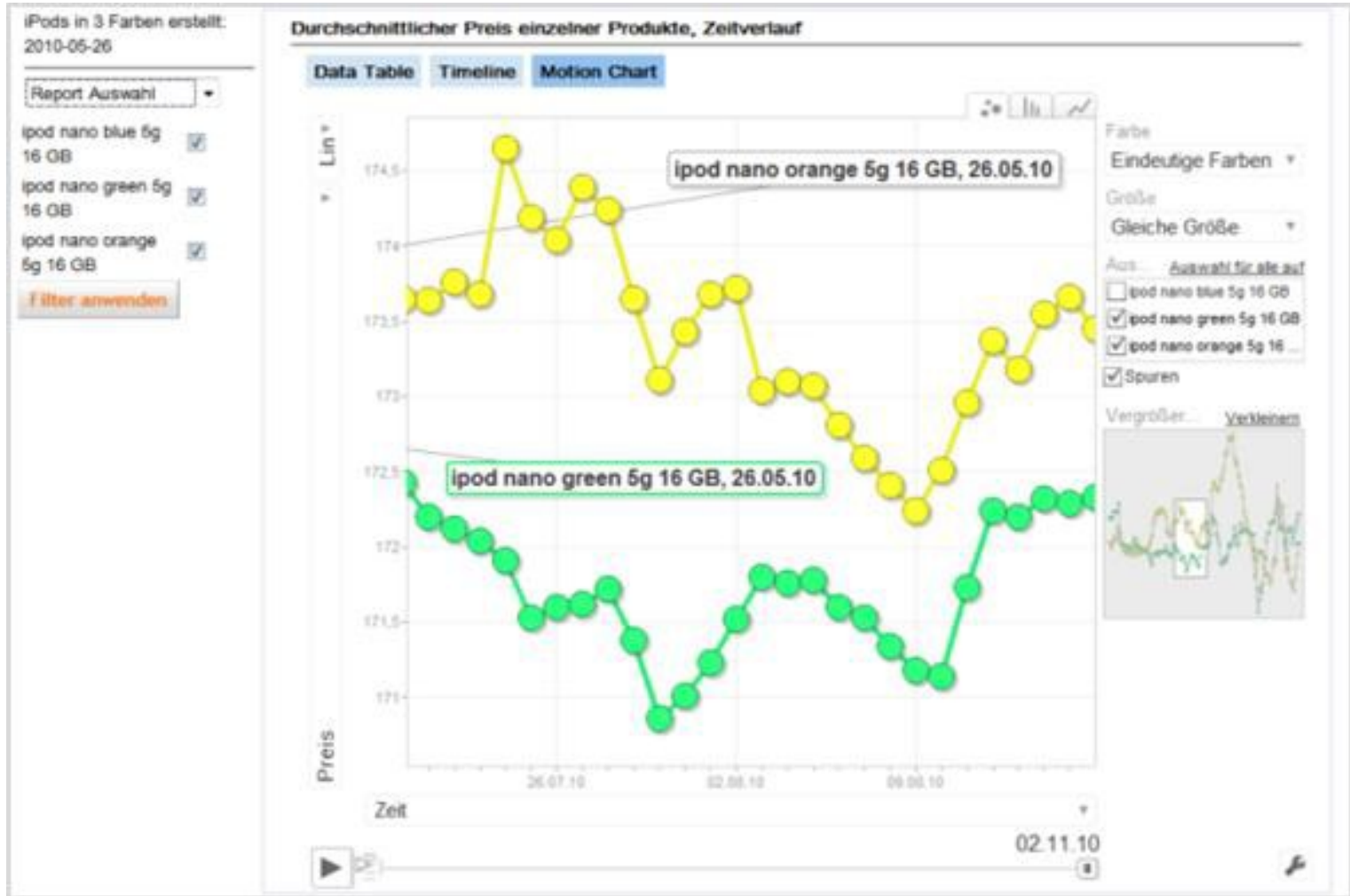


Unternehmenseigene Daten

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



Elektronik	>	Audio & Hi-Fi
Haus & Garten	>	Computer
Freizeit & Sport	>	Foto & Camcorder
Schmuck & Beauty	>	Handy & Organizer
Sammeln	>	Haushaltsgeräte
WOW! Angebote	>	PC- & Videospiele
Mode	>	Software
		TV, Video & Elektronik

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

Matching mit COMA 3.0

The screenshot displays the 'Repository Match Mapping View' in COMA 3.0. The interface is divided into several sections:

- Repository Match Mapping View:** The main window title.
- Repository:** A sidebar on the left containing:
 - Workspace:** A tab for the current workspace.
 - Domains:** A list of domains including 'PurchaseOrder (22 + 16)', 'Spicy (4 + 2)', and 'Lebensmittel' (highlighted).
 - Schemas:** A list of schemas including 'dmoz_Freizeit (1)', 'Google_Freizeit (1)', and 'Google_Lebensmittel (1)'.
 - Mappings:** A list of mappings including 'Freizeit' and 'Lebensmittel'.
 - Summary Table:**

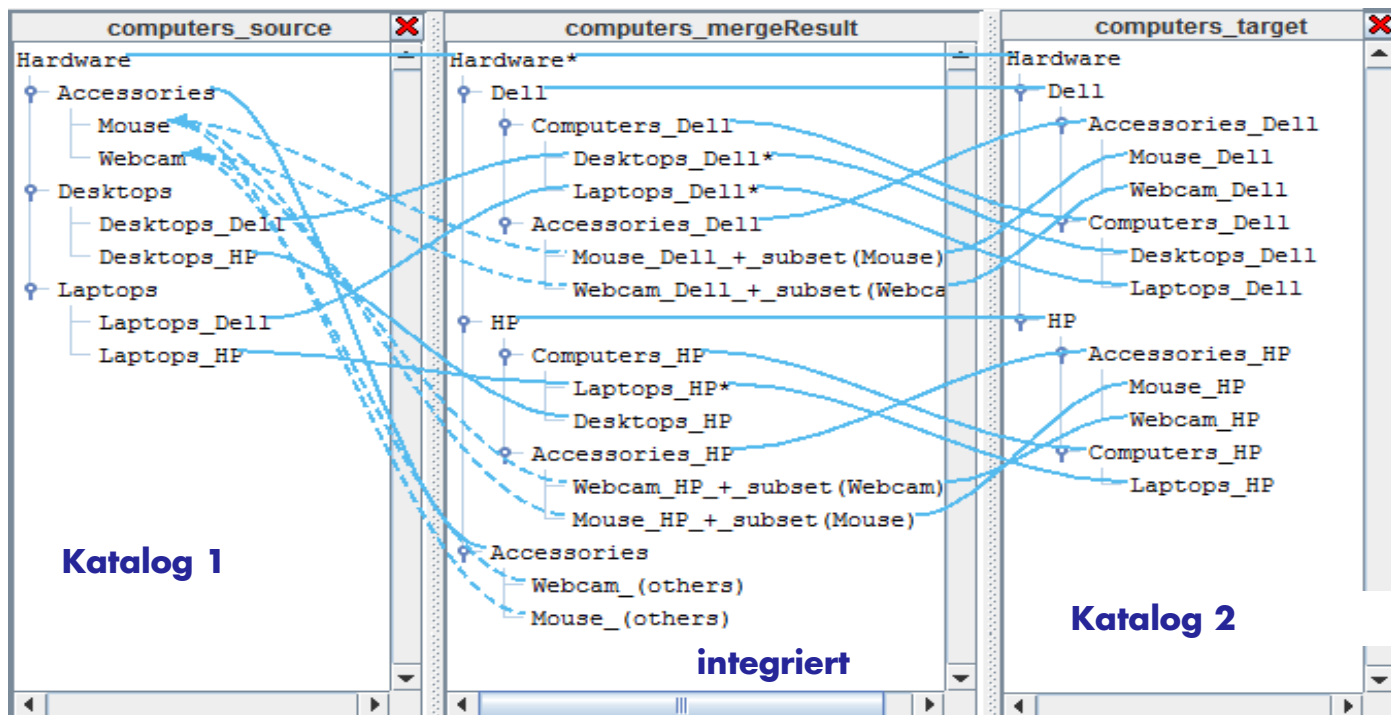
Name	Lebensmittel
Comment	UserFeedback
Schemas	Google_Lebensmittel, w...
Operation	
Total	32
- Mapping1:** The central area showing a mapping between two ontologies:
 - Source Ontology (Google_Lebensmittel):** A tree structure with nodes like 'Wein', 'Biowein', 'Europaeisch', 'Deutsch', 'Ahr', 'Baden', 'Franken', 'Mosel-Saar-Ruwer', 'Nahe', 'Pfalz', 'Rheingau', 'Rheinhausen', 'Wuerttemberg', 'Franzoesisch', 'Italienisch', 'Oesterreichisch', 'Schweizer', and 'Spanisch'.
 - Target Ontology (web_Lebensmittel):** A tree structure with nodes like 'Wein_Sekt-Champagner', 'Accessoires', 'Champagner', 'Deutschland', 'Baden', 'Bodensee', 'Franken', 'Hohenlohe', 'Kleinstanbaugebiete', 'Mittelrhein', 'Mosel', 'Nahe', 'Pfalz', 'Rheingau', 'Rheinhausen', 'Saale-Unstrut', 'Saar', and 'Frankreich'.
 - Connections:** Colored lines (yellow and green) connect corresponding nodes between the two ontologies. For example, 'Wein' maps to 'Wein_Sekt-Champagner', 'Deutsch' to 'Deutschland', and 'Ahr' to 'Baden'.
 - Color Scale:** A horizontal bar at the top of the mapping area shows a color gradient from 1.0 (yellow) to 0.0 (red), indicating the quality of the match.
- Source Node Information / Target Node Information:** Two empty tables at the bottom for detailed node data.

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

Duplikate

Heterogene Werte

Titel	Hersteller	Kategorie
Taschenradio ICF-304	Taschenradio ICF-304	Radio
Sony ICF-304 tragbares Radio (38459)	Sony	Radio
HP Photosmart C6180	HP	Kamera
HP Photosmart C6180	Hewlett Packard	Kamera
Kyocera FS 1300D	Kyocera	Drucker
Kyocera Toner TK-130	Kyocera Toner TK-130	Druckerzubehör
Kyocera Bildtrommel DK-130		Drucker

Fehlende Werte

Falsche Werte

- Unsere Ansätze:

- Datenbereinigung
- Automatische Kategorisierung
- Erkennung von Dubletten

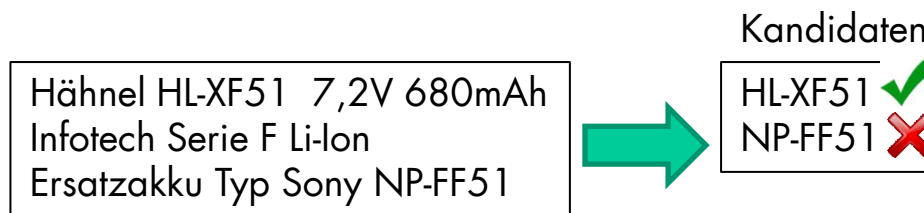
Datenbereinigung

- 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



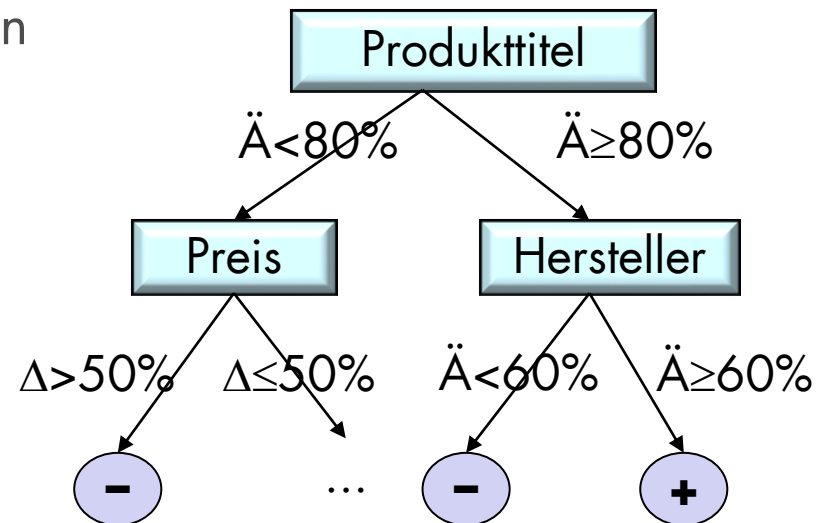
Title: 'Playstation 2 - Happy Pico Pad Retro mit 10 Spielen'
Description: 'SP-2370F/ P535Happy Pico Pad Retro, Lemon - Yellow, 10 Spiele/ Happy Pico Pad Retro, Lemon - Yellow, 10 Spiele/ Hersteller: Madrics MediaHersteller Originalnummer: 20567EAN-Code: 4041192205671 '

- 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



Online Product Manager

File Categorise Match Filter Remove Filter

category tree	number	Name	Description	Brand	Price	category
[-] Foto-, Videokamera & Co.	2	Schmusedecke, Sterntaler, »Hund«	Süße Babyzimmer-Ausstattung von STERN-TALE Sterntaler		34.99	Babyspielzeug
[-] Computer & Software	9	Wippe, Disney, »Bungee Deluxe«	Feststellbare Schaukelwippe. Weiche Auflage. I Disney		52.9	sonstige Baby- & Kinderartikel
[-] Schnäppchen & Gebrauch	0	Wickeltasche, Hauck, »Lady«	Einfaches und bequemes Wickeln für Mutter u. I Hauck		39.9	sonstige Baby- & Kinderartikel
[-] Audio, Video & TV	2	Beruhigungssauger Happy Days Silikon Gr. 1 2€	ACHTUNG! Dieser Artikel ist in verschiedenen F. NUK		4.49	sonstige Baby- & Kinderartikel
[-] Buch, Hörbuch & Kalende	6	Toiletten- Sitz weiß-rot	Die ergonomische weiche Form des BABYBJÖF BabyBjörn		29.99	sonstige Baby- & Kinderartikel
[-] Beauty, Wellness & Droge	0	Baby- Lammfell ca. 70-80 cm lang	Kuschelweiches Lammfell - temperatur- und feuc Kaiser		24.99	sonstige Baby- & Kinderartikel
[-] Downloads zum Verkauf &	0	Beruhigungssauger Trendline Latex Gr. 1 4er Pa	ACHTUNG! Dieser Artikel ist in verschiedenen F. NUK		9.99	sonstige Baby- & Kinderartikel
[-] Telekommunikation	14	Philips SCD 480/00	Hersteller: Philips HAUSTECHNIK	Philips	48.65	Funktechnik & Zubehör
[-] Game & Konsole	0	Babyfon, Philips, »SCD 480/00«	Mehr Sicherheit für Sie und Ihren Liebling. Der : Philips		69.99	Funktechnik & Zubehör
[-] Hobby & Spiel	1	Babyfon, Philips, »SCD 480/00«	Mehr Sicherheit für Sie und Ihren Liebling. Der : Philips		69.99	Funktechnik & Zubehör
[-] Haushalt & Wohnen	13	Schnullerbox Bär	In der süßen Schnullerbox aus hochwertigem Ku Euret		2.79	sonstige Baby- & Kinderartikel
[-] Sport & Freizeit	0	Topf marine/beige	Jetzt wird Sauberwerden ganz einfach. Die idea Lockweiler		4.99	sonstige Baby- & Kinderartikel
[-] Mode & Accessoires	0	Trittschemel marine/beige	Jetzt erreichen auch die Kleinen schnell und sic! Lockweiler		6.99	sonstige Baby- & Kinderartikel
[-] weitere Produkte	13	Toiletten- Sitz beige/aquamarin	Jetzt wird Sauberwerden ganz einfach. Dieser V Lockweiler		4.99	sonstige Baby- & Kinderartikel
		Trittschemel beige/aquamarin	Jetzt erreichen auch die Kleinen schnell und sic! Lockweiler		6.99	sonstige Baby- & Kinderartikel
		Wickeltasche Baby TravelBag schwarz	Auf Wochenend-Reisen beim Übernachten bei c AVENT		79.99	sonstige Baby- & Kinderartikel
		Beruhigungssauger Flex Comfort Flat Silikon Gr.	Ganz wie bei Mama! Der Flex Comfort-Sauger rr Nuby		2.99	sonstige Baby- & Kinderartikel
		Mein erstes Bildlexikon, Tiere im Wald Antje Klei	Die Tierwelt des Waldes ist für Kinder spannend Bertelsmann		9.95	Freizeit-Buch
		Telekommunikation - Headset »BT530« von Jabr	Headset »BT530«, Anschluss: Bluetooth, Trageve. JABRA		53.54	Headset & Freisprecheinrichtung

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)



SAP RESEARCH



AGENTURSCHADE