

# Unstructured information management

Research Cluster in SAP Research

Data Management and Analytics Program

Alexander Löser

THE BEST-RUN BUSINESSES RUN SAP™





- 2004 Doctorial student researcher with HP Research Labs Bristol
- PhD 2005 at DFG Graduate School Berlin-Brandenburg
- 2006 Postdoc IBM Almaden Research Center
- Since 02/2007 Senior Researcher & Project lead at SAP Research Dresden

Research interests: Enterprise search and Unstructured information management

- **Intelligent Find: SAP project with BMW Munich**

Aggregate unstructured product information for BMW's engineers

- **OKKAM: EU project on entity management**

Information extraction, aggregation and search of SAP documentation portals.  
E.g., SDN.SAP.COM, HELP.SAP.COM

- **High precision enterprise search:**

SAP project on ranking business objects with Netweaver Enterprise Search

- **Contributed to BMBF/DFG/EU proposals**

Aletheia / OKKAM / SOKNOS 2 / D-GRID 3

See [www.sap.com/research](http://www.sap.com/research)

# Unstructured information in the enterprise



The screenshot displays two overlapping windows. The background window is Microsoft Internet Explorer showing the SAP NetWeaver Enterprise Search interface. The foreground window is Microsoft Excel, titled "List of product and component versions.D.XLS", containing a table of product and component versions.

**Microsoft Excel - List of product and component versions.D.XLS**

	A	B
541	SAP XRPM 4.0	SAP xApp Resource and Portfolio Management , Release 4.0
542	SAP XSF 1.0	SAP xApp Service Flow 1.0 by Digital Fuel (Delivery Postponed - No date)
543	SAP XVIP 1.0	SAP xApp VIP 1.0 (Delivery Postponed. No New Date)
544	SAPM ANALYTICS 2.0	ANALYTICS 2.0
545	SAPM COLLABORATIVE DESIGN 2.0	SAPM COLLABORATIVE DESIGN 2.0
546	SAPM COLLABORATIVE PLANNING2.0	COLLABORATIVE PLANNING 2.0
547	SAPM DA/DB 3.0	SAPM DA/DB 3.0
548	SAPM E-PROCUREMENT 2.0	SAPM E-PROCUREMENT
549	SAPM E-PROCUREMENT 2.5	SAPM E-PROCUREMENT 2.5
550	SAPM MARKETSET DPE 2.0	SAPM MARKETSET DPE 2.0
551	SAPM MSET CORE INFRASTRUC. 2.0	SAPM Bulletin Board / Info Services / User Mgmt 2.0
552	SAPM ORDER MANAGEMENT 2.0	SAPM ORDER MANAGEMENT 2.0
553	SAPM ORDER MANAGEMENT 2.5	SAPM ORDER MANAGEMENT 2.5
554	SAPMARKETS INTERNET SALES 3.0	SAPMARKETS INTERNET SALES 3.0
555	SAPSCRIPT FORMS	SAPSCRIPT FORMS
556	SBO SOFTWARE DEV. KIT 2004	B1 SOFTWARE DEVELOPMENT KIT 2004
557	SBO SOFTWARE DEV. KIT 2005	SBO SOFTWARE DEV. KIT 2005
558	SBO SOFTWARE DEV. KIT 6.01	SBO Software Development Kit 6.01
559	SBO SOFTWARE DEV. KIT 6.01	SBO Software Development Kit 6.01
560	SBO SOFTWARE DEV. KIT 6.2	SBO SOFTWARE DEVELOPMENT KIT 6.2
561	SBO SOFTWARE DEV. KIT 6.2	SBO SOFTWARE DEV. KIT 6.2
562	SBO SOFTWARE DEV. KIT 6.5	SBO SOFTWARE DEVELOPMENT KIT 6.5
563	SBO SOFTWARE DEV. KIT 6.5	SBO SOFTWARE DEV. KIT 6.5
564	SBO-MYSAP INTEGRAT.TOOLKIT 1.0	SAP Business One Integration Toolkit for mySAP Business Suite - 1.0

**SAP NetWeaver Enterprise Search - Microsoft Internet Explorer**

Address: http://vmw2092.wdf.sap.com:53000/webdisp/pro/dispatcher/sap.com/...  
Display Standard Order 14289: Overview

**RE: SAP NetWeaver Innovation News - November 2007**

Von: Deissner, Klaus  
An: Loeser, Alexander  
Cc:  
Betreff: RE: SAP NetWeaver Innovation News - November 2007

2007 report. Check it out -- you'll feel like you were almost there. See more on the [Event Reports](#) page or check the wiki for the...

**SAP NetWeaver Innovation Hits YouTube**

SAP NetWeaver Imagineering now has a [channel on YouTube](#) in its showcase (public) innovations happening on the SAP NetWeaver platform. See the [full Eventus demo on YouTube](#).

Got a video? [Submit it!](#)

**Poll of the Day - Do You YouTube?**

Do you consume, produce, direct, star in, or otherwise participate in the YouTube phenomenon? [Visit our poll](#) and make your voice heard.

**Thanks for contributing!**

Remember, these links go to a wiki. Free to edit or post comments to any page, or see below for more ways to contribute. Great ideas come from everywhere!

[SAP NetWeaver Product Management Imagineering](#)

# Mining unstructured information is costly!

## Selected enablers and potential research goals



Define a small set of reusable algebraic extraction operators to

- create simple information extraction flows outside of the application code
- discover relationships
- compose atomic entities to complex entities
- bind complex entities to existing relational data, e.g. from ERP system

Intuitive UI for algebraic information extraction

- lower barrier to “click” & compose information extraction flows
- immediately visualize extraction results

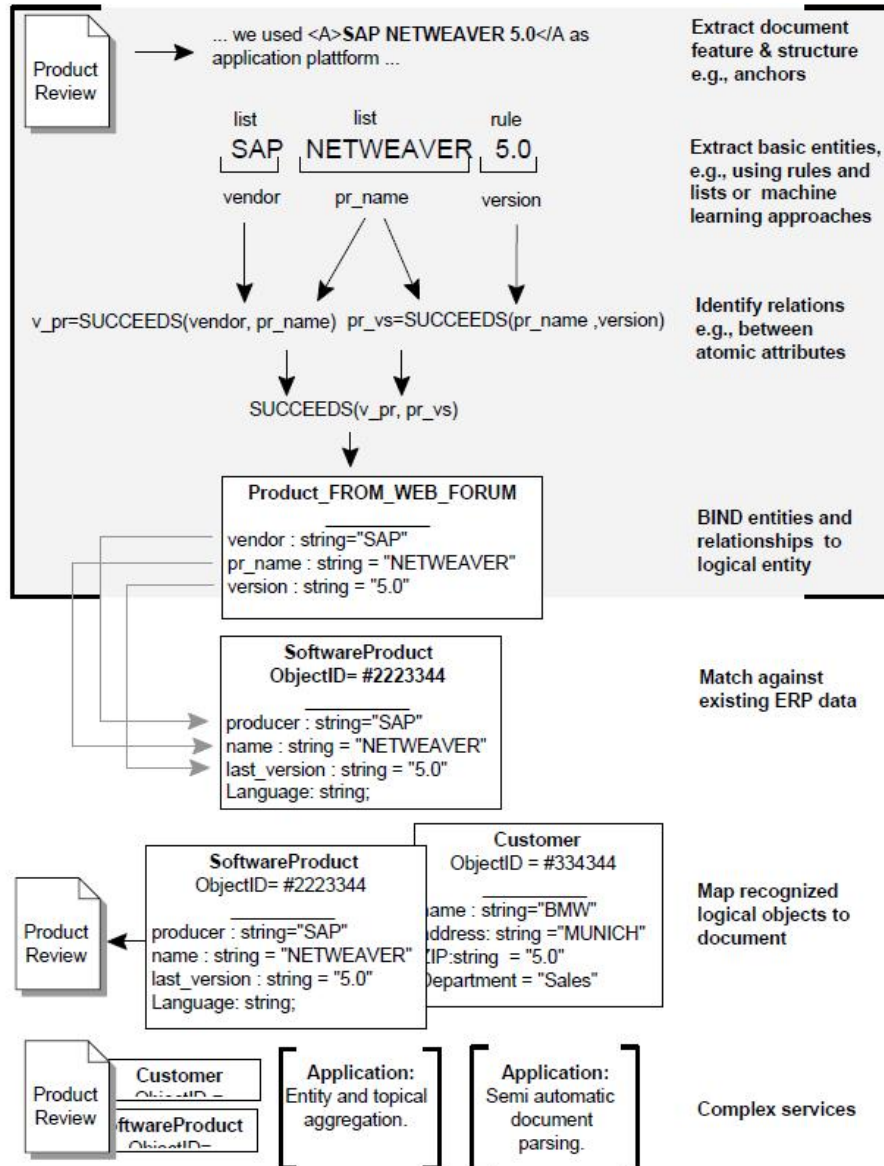
Recognizing entities based on existing relational data

- instance data from SAP business objects
- relationships (aka foreign keys) to other business objects

**Overall goal: Reduce costs for information extraction projects, so more customers will be able to afford it!**

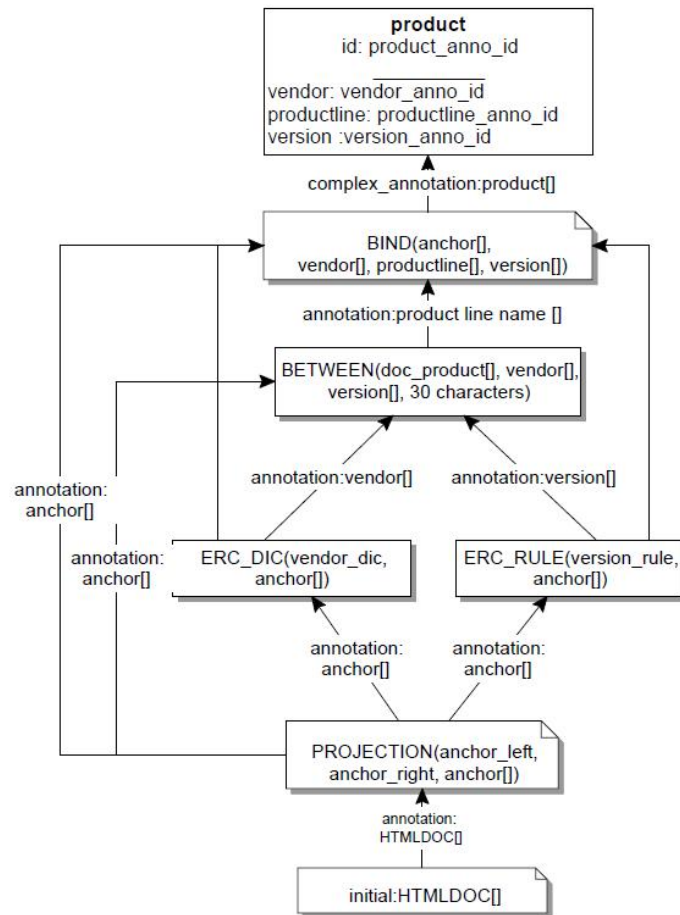
# Bind, aggregate and consolidate unstructured data

## Example application: Extract products from forum

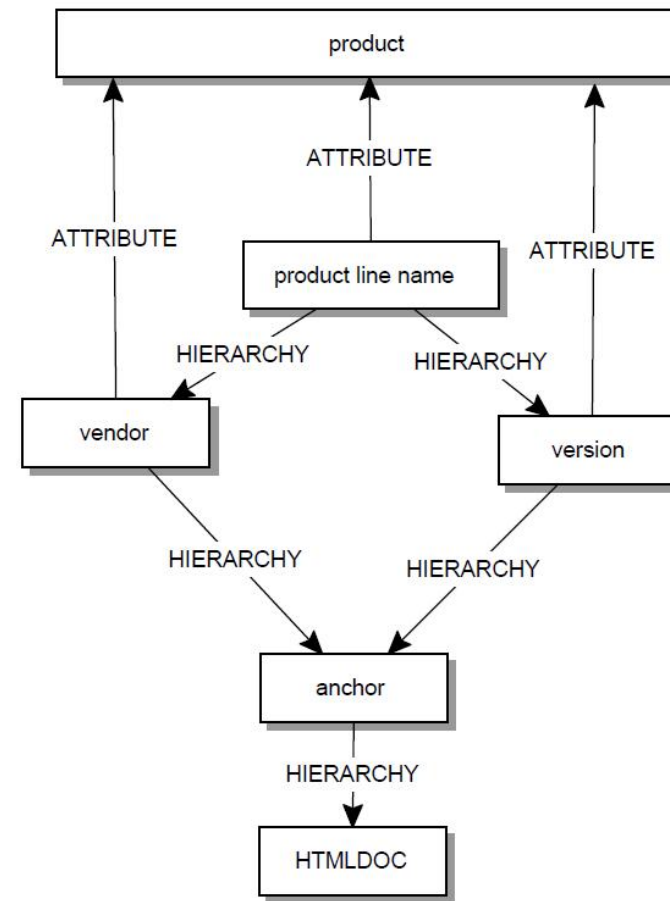


# Algebraic information extraction operators

## Example flow for recognizing a product



(a) Product line names from product list.



(b) Annotation and aggregation flow

# A graphic & algebraic approach in IE

## Example of operators



### Regular expressions

- R1XR2 - PROJECTION(regex1, annotation, regex2): annotation[ ]  
PROJECTION(Regex('\w<title>\w '), HTML\_DOC , Regex('\w</title>\w ')) : **HTML\_TITLE**

Extract structure  
of document.

### Lists

- Dictionary - Dict(uri, annotation) : annotation[]  
DICT(SAP\_EMPLOYEE, annotation) : **SAP\_WORKERS**

Extract entities  
based on lists

### Context operators

- LC - LC(entity) : annotation[]  
LC (PERSON) : **LC\_PERSON**
- RC - RC(entity) : annotation[]  
RC (PERSON) : **RC\_PERSON**

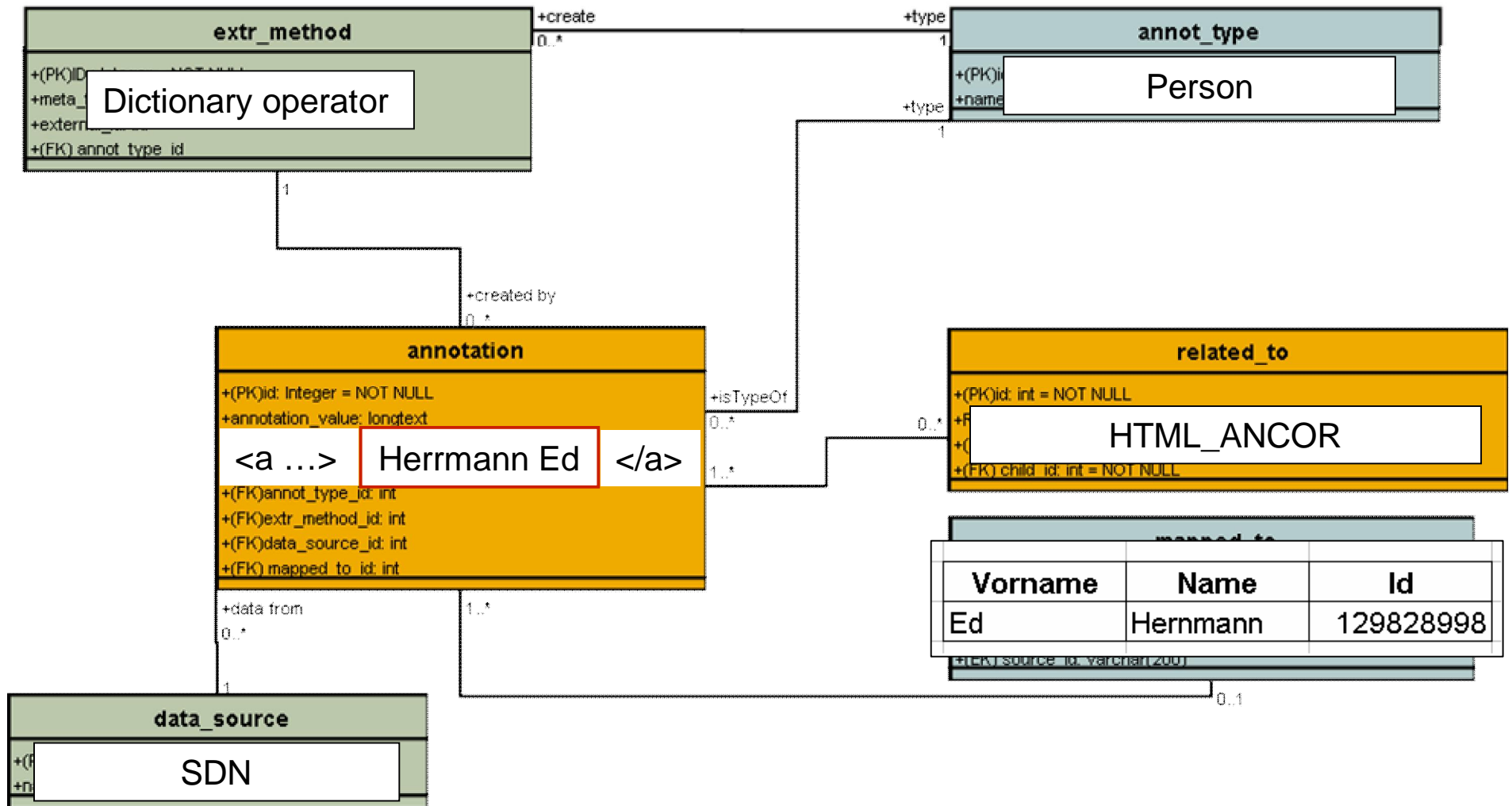
Context

### Relationships between annotations

- BETWEEN - BETWEEN(annotation1, annotation2) : annotation[]  
BETWEEN(PERSON, PROJECT) : **BE\_VENDOR\_VERSION**
- SUCCEED - SUCCEED(annotation1, annotation2, regular expression) : annotation[]  
SUCCEED (PERSON, PROJECT, Regex('\w+(works in|works with)\w+')):**BE\_PERSON\_PROJECT**

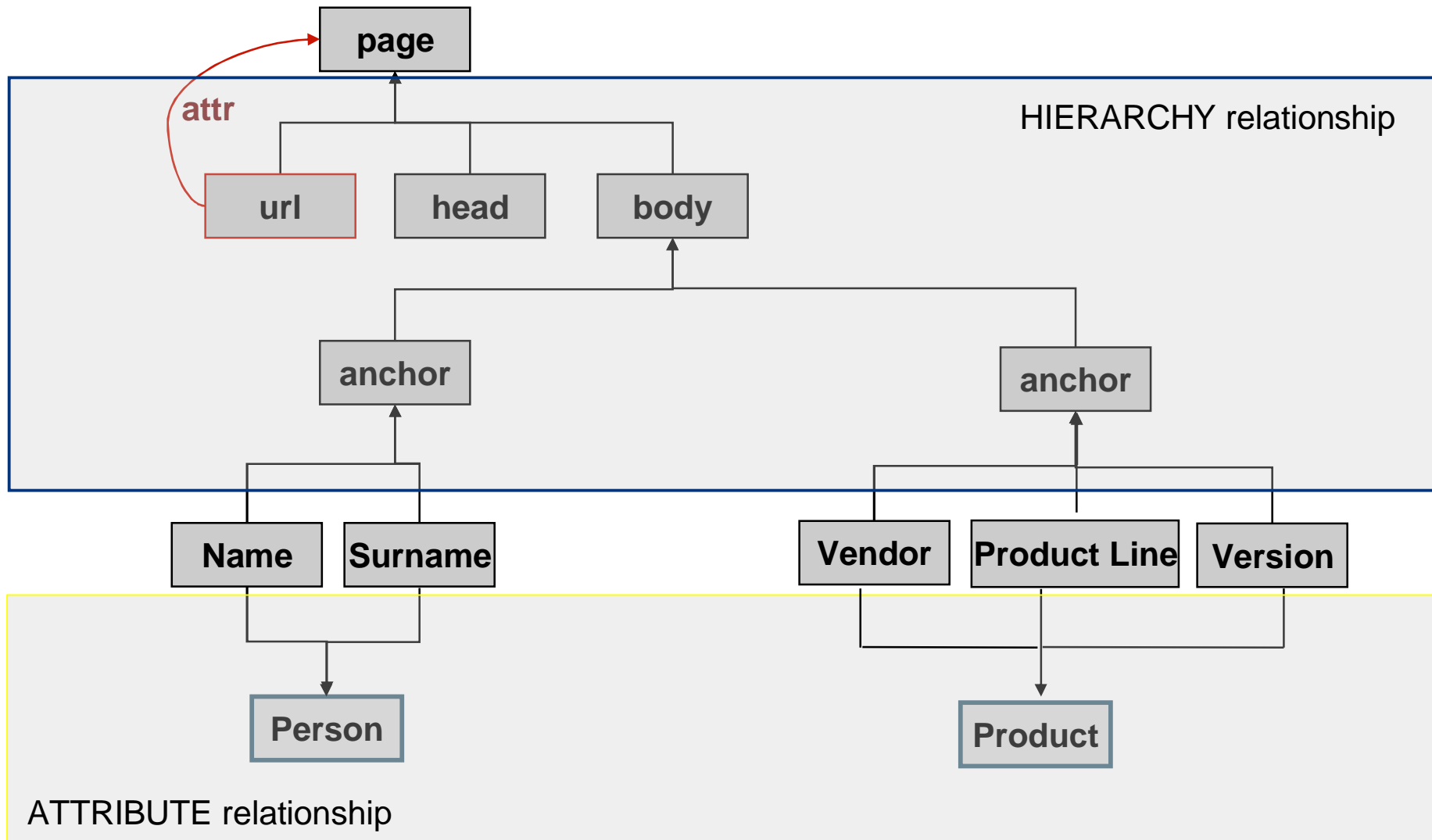
Find relations

# Basic storage scheme for RDBMS





# STORAGE: Hierarchical and attribute relationships between annotations



# Intuitive UI for algebraic information extraction



Unstructured Information Management

### Webpage (html)

**Developer's Journey Through the SDN Subscriptions Program**

**Ed Herrmann**, a developer for **Colgate-Palmolive**, recently joined the **SDN Subscriptions Program** for access to the software, tools, and services needed to develop for **SAP NetWeaver**. His latest blog describes his early experiences with the Subscriptions Program, with detailed pictures of his subscription kit and other useful information for potential subscribers. 30 Oct 2007

**What's New**

Topic of the Month: [Research](#)

Blog: [Pulse Check on the SDN/BPX Communities](#)

url:

Zoom

### Extracted Annotations

- Person
- Vendor

Operators : X R Y

**Input**

**Person**

Entity  Type

**Company**

Entity  Type

**Output**

Value

Type

**Regex**

### Webpage (txt)

```
<p> <a href="/irj/sdn/weblogs?blog=/pub/u/251">Ed Herrmann</a>, a developer for Colgate-Palmolive, recently joined the <a href="/irj/sdn/dev?promo=featuredblog">SDN Subscriptions Program</a> for access to the software, tools, and services needed to develop for SAP NetWeaver. His latest blog describes his early experiences with the Subscriptions Program, with detailed pictures of his subscription kit and other useful information for potential subscribers. <span class="article_date">30 Oct 2007</span> </p>
```

### IE Process

# Map structured data (BO's, tables) to unstructured data (text, html, email etc)



A390	SAP NETWEAVER 04			
372	SAP MPR 3.0	SAP Mobile Procurement 3.0		
373	SAP MSA&MSE FOR HH WITH CRM4.0	SAP MSA&MSE FOR HH WITH CRM 4.0		
374	SAP MSA&MSE FOR HH WITH CRM4.0	SAP MSA&MSE for Handheld with CRM 4.0		
375	SAP MSON 1.0	SAP Mobile Sales Online 1.0 with CRM		
376	SAP MSON 1.0	MSO 1.0		
377	SAP MSR 1.0 FOR HH	MSR 1.0 FOR HH		
378	SAP MSR 1.0 FOR HH	SAP Mobile Sales 1.0 for Handheld (with R/3)		
379	SAP MSR 2.0 FOR HH	SAP Mobile Sales 2.0 for Handheld (with R/3)		
380	SAP MSR 2.0 FOR HH	MSR 2.0 FOR HH		
381	SAP MTS & MTR 1.0 FOR HH	MTS & MTR 1.0 FOR HH		
382	SAP MTS & MTR 1.0 FOR HH	SAP MTS & MTR 1.0 for Handheld		
383	SAP MTS & MTR 1.0 FOR LAPTOP	MTS 1.0		
384	SAP MTS & MTR 1.0 FOR LAPTOP	SAP MTS & MTR 1.0 for Laptop		
385	SAP MTS & MTR 1.5 FOR LAPTOP	MTS 1.5 FOR LAPTOP		
386	SAP MTS & MTR 1.5 FOR LAPTOP	SAP MTS & MTR 1.5 for Laptop		
387	SAP MTS & MTR 1.6 FOR LAPTOP	SAP MTS & MTR 1.6 for Laptop		
388	SAP MTS & MTR 1.6 FOR LAPTOP	MTS & MTR 1.6 FOR LAPTOP		
389	SAP MULTI SITE WFD 1.0	SAP MULTI SITE WORKFORCE DEPLOYMENT 1.0		
390	SAP NETWEAVER 04	SAP NETWEAVER 04		
391	SAP NETWEAVER 2005	SAP NET		
392	SAP NETWEAVER 2006	SAP NET		
393	SAP OGS 4.72	SAP OGS		
394	SAP OLM 4.6B	SAP OLM		
395	SAP OLM 4.6C	SAP OLM		
396	SAP ONE SERVER 1.0	SAP ONE		
397	SAP ONE SERVER 2.0	SAP ONE		
398	SAP PATIENT MANAGEMENT 4.03A	SAP Patie		
399	SAP PATIENT MANAGEMENT 4.03B	SAP Patie		
400	SAP PATIENT MANAGEMENT 4.03C	SAP Patie		
401	SAP PATIENT MANAGEMENT 4.03D	SAP Patie		
402	SAP PATIENT MANAGEMENT 4.61A	SAP Patie		

**Problems:**

- Identify entities
  - ambiguity
  - inaccuracy
  - abbreviations
  - synonyms
  - different vocabularies
- Understand context of entity

Matching data by equality won't work. -> recall vs. precision  
 Searching complete databases won't work -> efficiency vs. completeness

**[ZLRV07] *Navigating the Intranet with high precision.* Huaiyu Zhu Alexander Löser, Sriram Raghavan, Shivakumar Vaithyanathan. 16<sup>th</sup>. WWW Conference 2007**

**[KLM06] *Challenges and Trends in Information Management.* Karin Kailing, Alexander Löser, and Volker Markl. Journal Datenbank-Spektrum, 19/2006**

**[WEIK07] *DB & IR: Both Sides Now.* Gerhard Weikum, SIGMOD/PODS Keynote'07**

**[RSCL07] *DBLife: A Community Information Management Platform for the Database Research Community,* CIDR 07 (Demo)**

**[SKW07] *YAGO: Yet Another Great Ontology.* Fabian Suchanek, Gjergji Kasneci, Gerhard Weikum: 16<sup>th</sup>. WWW Conference 2007**

**[DRV06] *Managing Information Extraction.* Anhai Doan, Raghu Ramakrishnan, Shiv Vaithyanathan. SIGMOD 2006 Tutorial**

**[TREX] SAP Search Engine TREX in Wikipedia**  
**[http://en.wikipedia.org/wiki/TREX\\_search\\_engine](http://en.wikipedia.org/wiki/TREX_search_engine)**



- No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG. The information contained herein may be changed without prior notice.
  - Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors.
  - Microsoft, Windows, Outlook, and PowerPoint are registered trademarks of Microsoft Corporation.
  - IBM, DB2, DB2 Universal Database, OS/2, Parallel Sysplex, MVS/ESA, AIX, S/390, AS/400, OS/390, OS/400, iSeries, pSeries, xSeries, zSeries, z/OS, AFP, Intelligent Miner, WebSphere, Netfinity, Tivoli, and Informix are trademarks or registered trademarks of IBM Corporation in the United States and/or other countries.
  - Oracle is a registered trademark of Oracle Corporation.
  - UNIX, X/Open, OSF/1, and Motif are registered trademarks of the Open Group.
  - Citrix, ICA, Program Neighborhood, MetaFrame, WinFrame, VideoFrame, and MultiWin are trademarks or registered trademarks of Citrix Systems, Inc.
  - HTML, XML, XHTML and W3C are trademarks or registered trademarks of W3C®, World Wide Web Consortium, Massachusetts Institute of Technology.
  - Java is a registered trademark of Sun Microsystems, Inc.
  - JavaScript is a registered trademark of Sun Microsystems, Inc., used under license for technology invented and implemented by Netscape.
  - MaxDB is a trademark of MySQL AB, Sweden.
  - SAP, R/3, mySAP, mySAP.com, xApps, xApp, SAP NetWeaver and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries all over the world. All other product and service names mentioned are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.
- 
- The information in this document is proprietary to SAP. No part of this document may be reproduced, copied, or transmitted in any form or for any purpose without the express prior written permission of SAP AG.
  - This document is a preliminary version and not subject to your license agreement or any other agreement with SAP. This document contains only intended strategies, developments, and functionalities of the SAP® product and is not intended to be binding upon SAP to any particular course of business, product strategy, and/or development. Please note that this document is subject to change and may be changed by SAP at any time without notice.
  - SAP assumes no responsibility for errors or omissions in this document. SAP does not warrant the accuracy or completeness of the information, text, graphics, links, or other items contained within this material. This document is provided without a warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose, or non-infringement.
  - SAP shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials. This limitation shall not apply in cases of intent or gross negligence.
  - The statutory liability for personal injury and defective products is not affected. SAP has no control over the information that you may access through the use of hot links contained in these materials and does not endorse your use of third-party Web pages nor provide any warranty whatsoever relating to third-party Web pages



- Weitergabe und Vervielfältigung dieser Publikation oder von Teilen daraus sind, zu welchem Zweck und in welcher Form auch immer, ohne die ausdrückliche schriftliche Genehmigung durch SAP AG nicht gestattet. In dieser Publikation enthaltene Informationen können ohne vorherige Ankündigung geändert werden.
  - Die von SAP AG oder deren Vertriebsfirmen angebotenen Softwareprodukte können Softwarekomponenten auch anderer Softwarehersteller enthalten.
  - Microsoft, Windows, Outlook, und PowerPoint sind eingetragene Marken der Microsoft Corporation.
  - IBM, DB2, DB2 Universal Database, OS/2, Parallel Sysplex, MVS/ESA, AIX, S/390, AS/400, OS/390, OS/400, iSeries, pSeries, xSeries, zSeries, z/OS, AFP, Intelligent Miner, WebSphere, Netfinity, Tivoli, und Informix sind Marken oder eingetragene Marken der IBM Corporation in den USA und/oder anderen Ländern.
  - Oracle ist eine eingetragene Marke der Oracle Corporation.
  - UNIX, X/Open, OSF/1, und Motif sind eingetragene Marken der Open Group.
  - Citrix, ICA, Program Neighborhood, MetaFrame, WinFrame, VideoFrame, und MultiWin sind Marken oder eingetragene Marken von Citrix Systems, Inc.
  - HTML, XML, XHTML und W3C sind Marken oder eingetragene Marken des W3C®, World Wide Web Consortium, Massachusetts Institute of Technology.
  - Java ist eine eingetragene Marke von Sun Microsystems, Inc.
  - JavaScript ist eine eingetragene Marke der Sun Microsystems, Inc., verwendet unter der Lizenz der von Netscape entwickelten und implementierten Technologie.
  - MaxDB ist eine Marke von MySQL AB, Schweden.
  - SAP, R/3, mySAP, mySAP.com, xApps, xApp, SAP NetWeaver und weitere im Text erwähnte SAP-Produkte und -Dienstleistungen sowie die entsprechenden Logos sind Marken oder eingetragene Marken der SAP AG in Deutschland und anderen Ländern weltweit. Alle anderen Namen von Produkten und Dienstleistungen sind Marken der jeweiligen Firmen. Die Angaben im Text sind unverbindlich und dienen lediglich zu Informationszwecken. Produkte können länderspezifische Unterschiede aufweisen.
- 
- The information in this document is proprietary to SAP. No part of this document may be reproduced, copied, or transmitted in any form or for any purpose without the express prior written permission of SAP AG.
  - This document is a preliminary version and not subject to your license agreement or any other agreement with SAP. This document contains only intended strategies, developments, and functionalities of the SAP® product and is not intended to be binding upon SAP to any particular course of business, product strategy, and/or development. Please note that this document is subject to change and may be changed by SAP at any time without notice.
  - SAP assumes no responsibility for errors or omissions in this document. SAP does not warrant the accuracy or completeness of the information, text, graphics, links, or other items contained within this material. This document is provided without a warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose, or non-infringement.
  - SAP shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials. This limitation shall not apply in cases of intent or gross negligence.
  - The statutory liability for personal injury and defective products is not affected. SAP has no control over the information that you may access through the use of hot links contained in these materials and does not endorse your use of third-party Web pages nor provide any warranty whatsoever relating to third-party Web pages