Unstructured information management

Research Cluster in SAP Research

Data Management and Analytics Program

Alexander Löser

Alexander Löser

- 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

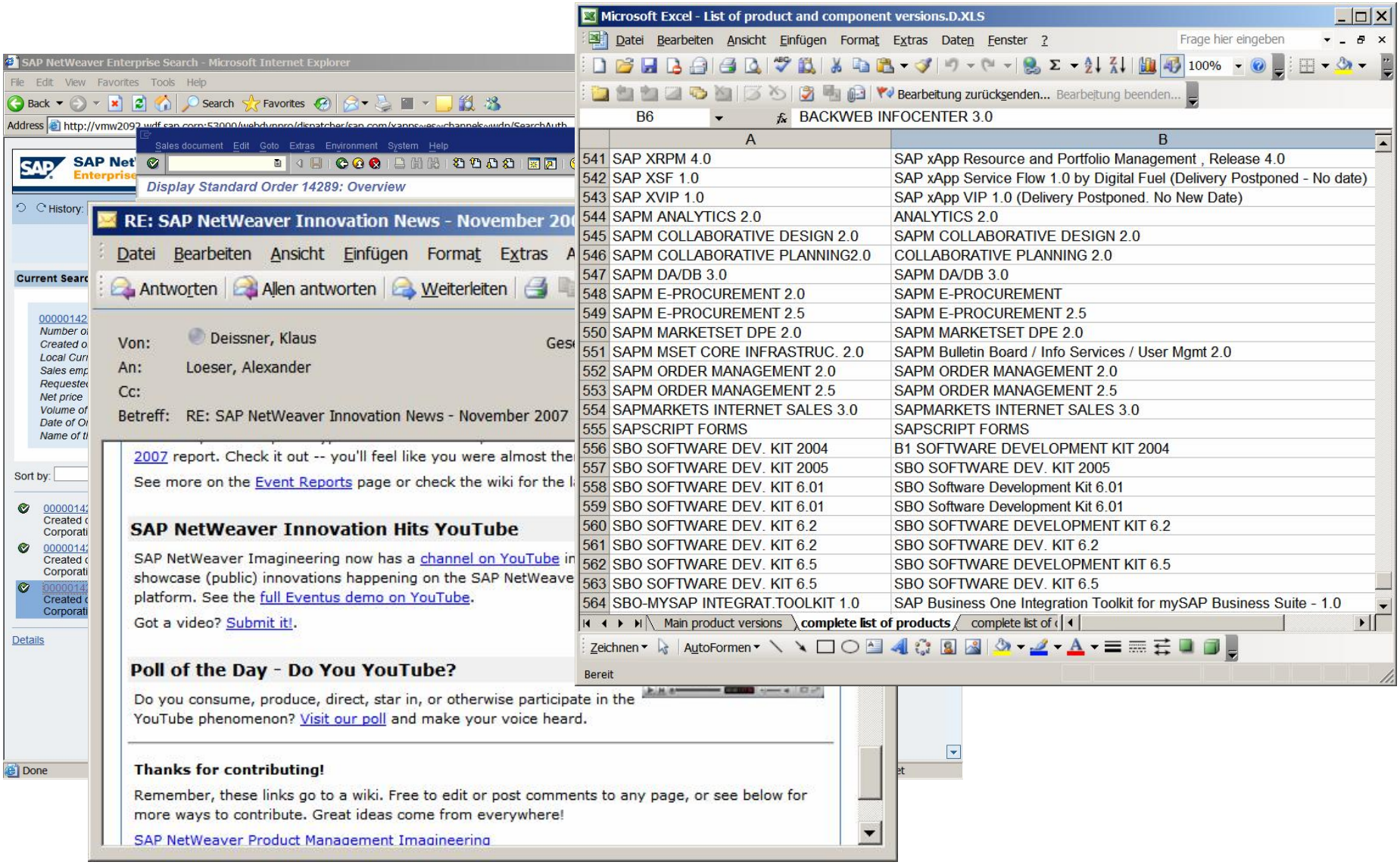
- **OKKAM: EU project on entity management**
Information extraction, aggregation and search of SAP documentation portals.
E.g., SDN.SAP.COM, HELP.SAP.COM

- **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

Unstructured information in the enterprise



	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

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

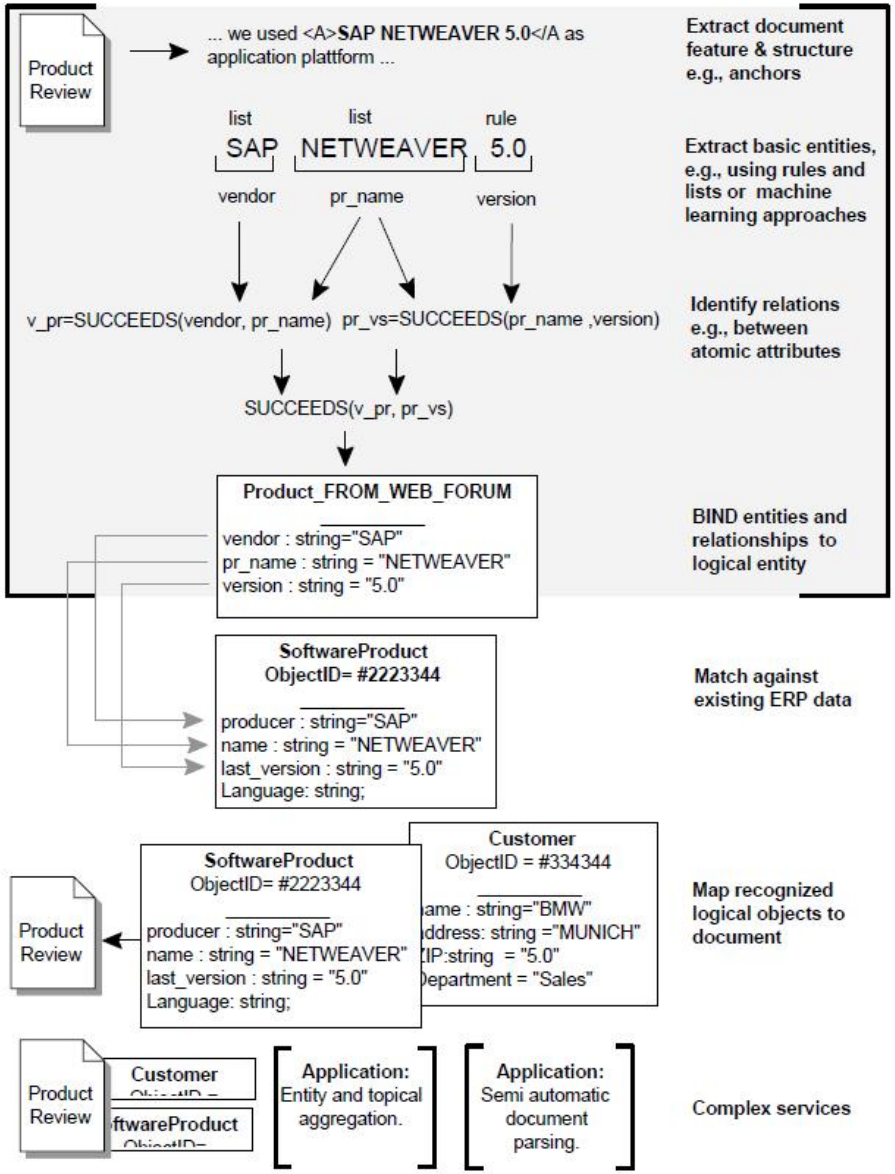
- lower barrier to “click” & compose information extraction flows
- immediately visualize extraction results

Recognizing entities based on existing relational data

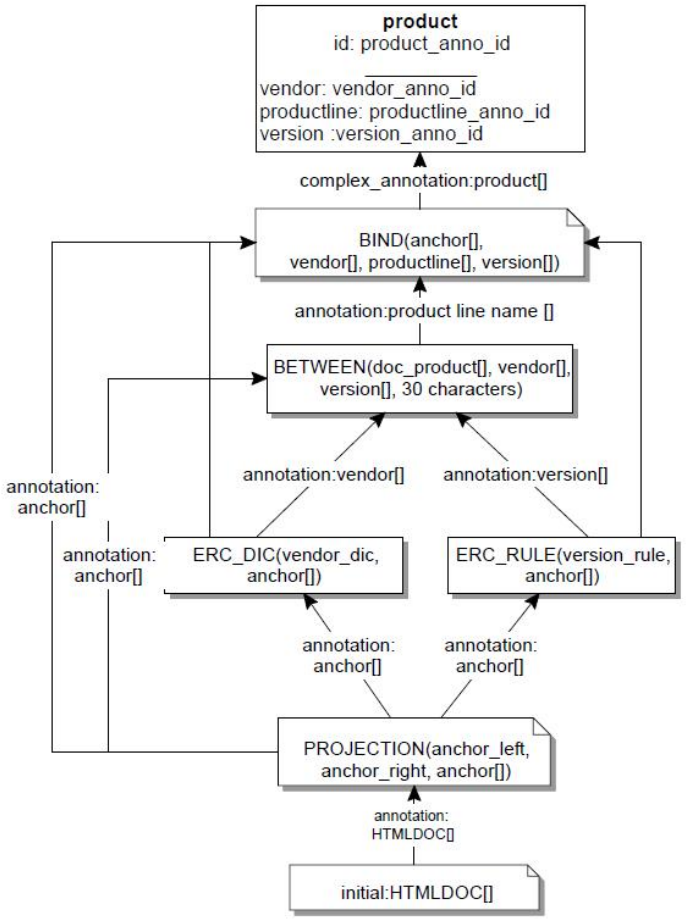
- instance data from SAP business objects
- relationships (aka foreign keys) to other business objects

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

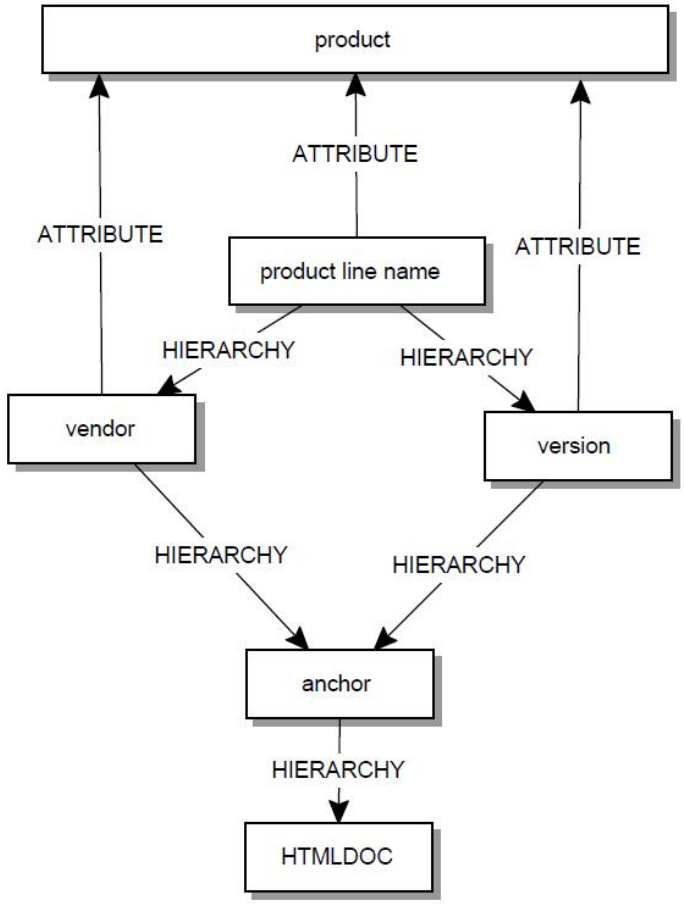
Bind, aggregate and consolidate unstructured data
Example application: Extract products from forum



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**

Context

- RC - RC(entity) : annotation[]
  RC(PERSON) : **RC_PERSON**

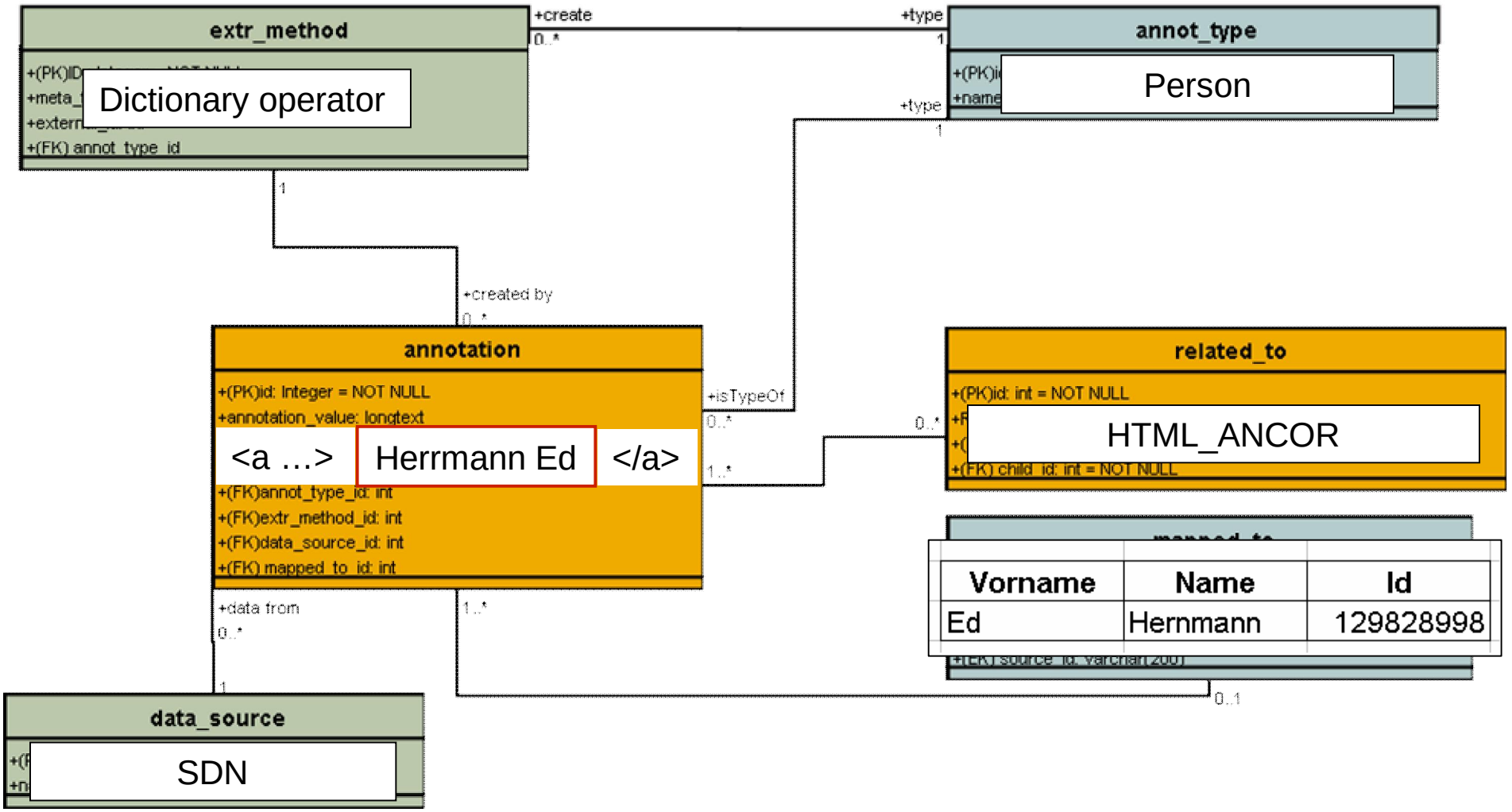
Relationships between annotations

- BETWEEN - BETWEEN(annotation1, annotation2) : annotation[]
  BETWEEN(PERSON, PROJECT) : **BE_VENDOR_VERSION**

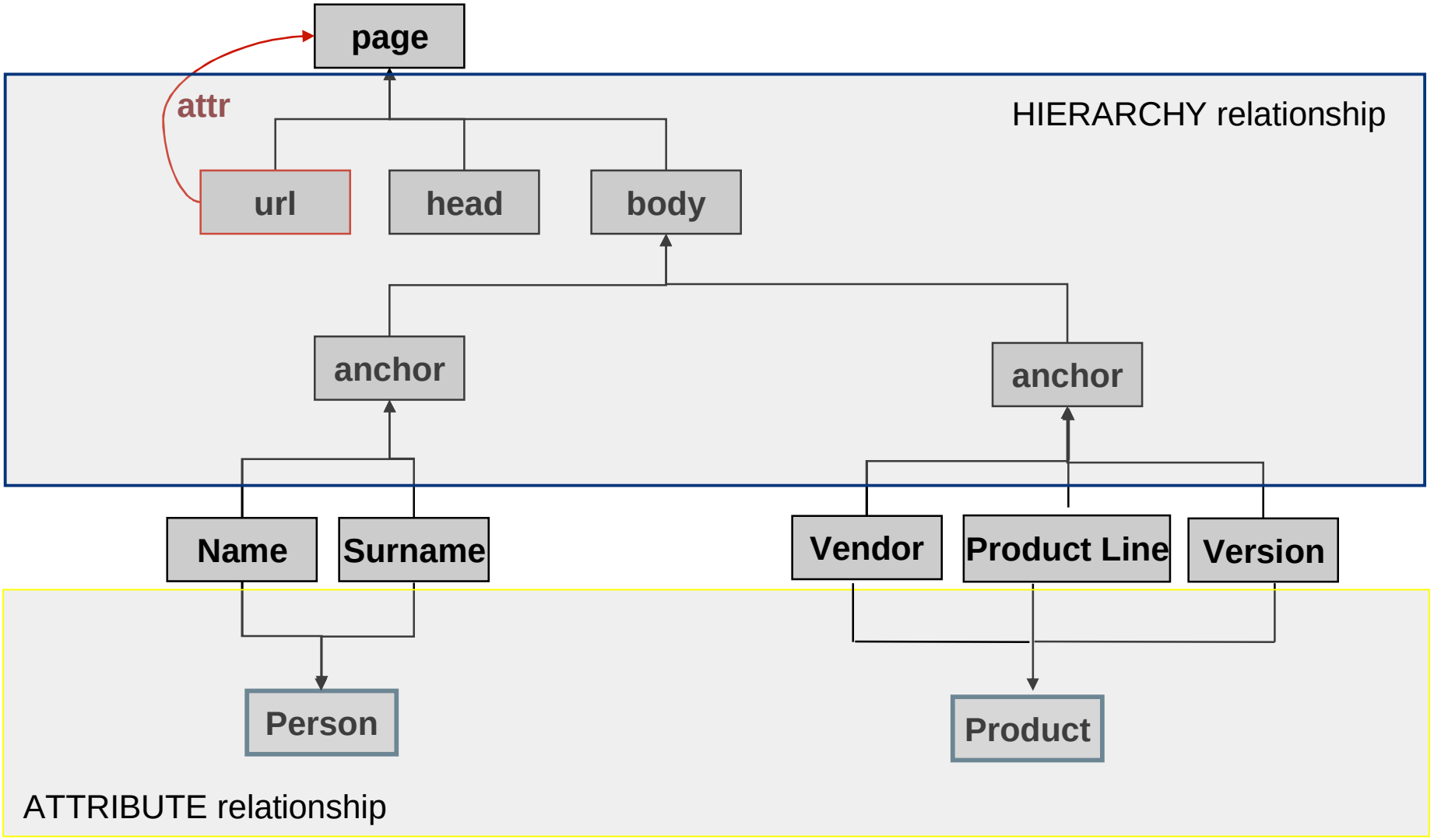
Find relations

- SUCCEED - SUCCEED(annotation1, annotation2, regular expression) : annotation[]
  SUCCEED (PERSON, PROJECT, Regex('\w+(works in|works with)\w+')):**BE_PERSON_PROJECT**

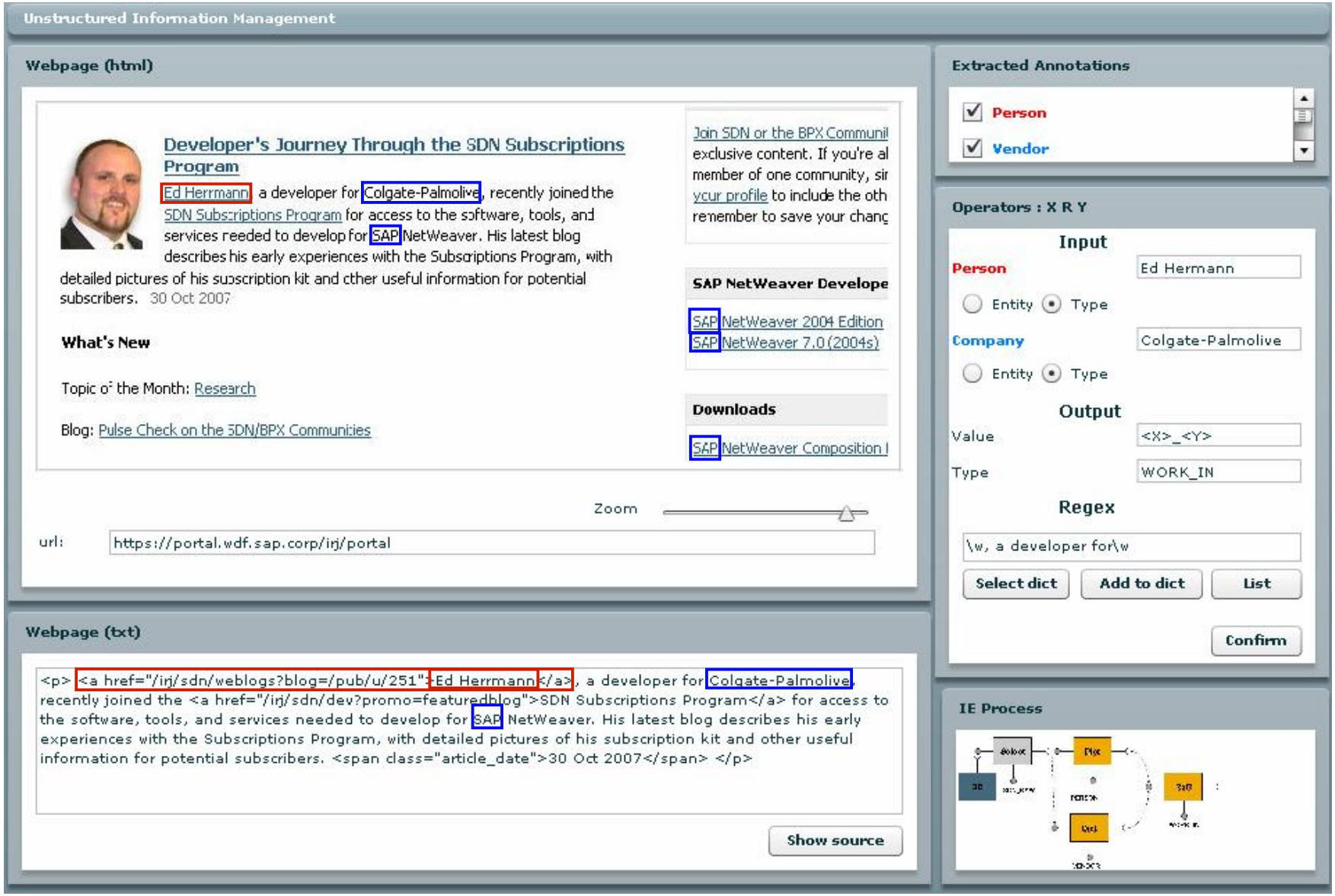
Basic storage scheme for RDBMS



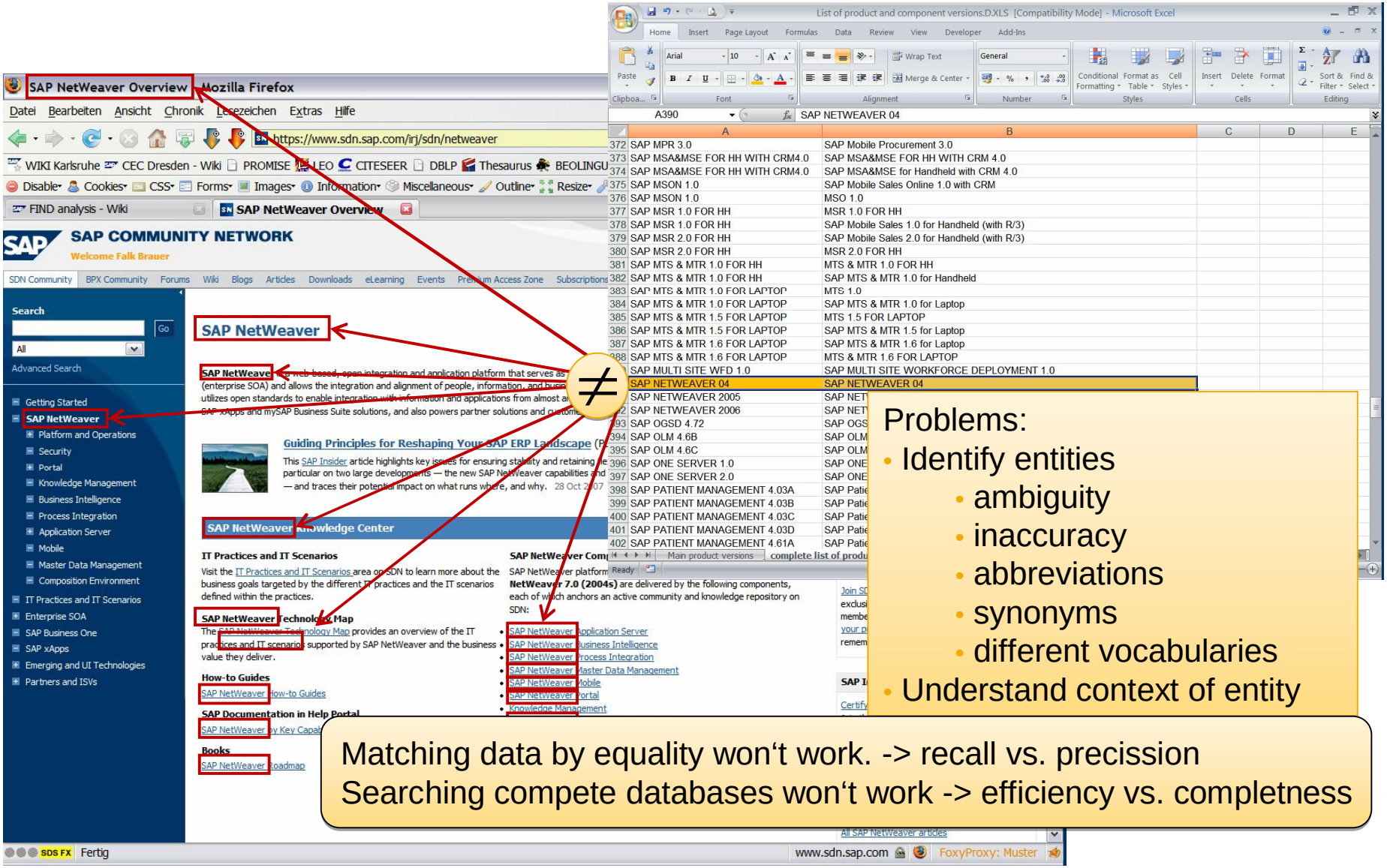
STORAGE: Hierarchical and attribute relationships between annotations



Intuitive UI for algebraic information extraction



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



Selected literature

[ZLRV07] ***Navigating the Intranet with high precision.*** **Huaiyu Zhu Alexander Löser, Sriram Raghavan, Shivakumar Vaithyanathan. 16th. 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: 16th. 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

Copyright 2005 SAP AG. All Rights Reserved

Copyright 2005 SAP AG. Alle Rechte vorbehalten