ENGINEERING A XML-BASED CONTENT HUB FOR ENTERPRISE PUBLISHING

Elias Weingärtner

Christoph Ludwig

HAUFE GROUP – QUICK FACTS

- Software Company and Media Publishing House
- Head Office: Freiburg, Germany
- Business Domains: Law, Tax, Human Resources, Talent Management, Trainings
- 150 Software Developers

HAUFE. LEXWARE HAUFE. AKADEMIE Haufe umantis Talent Management smart steuer

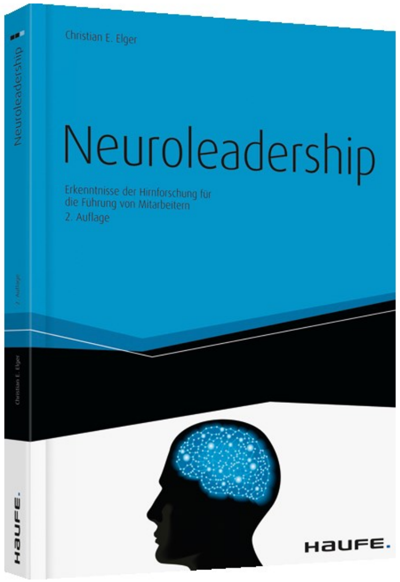
VCW Haufe-Lexware Real Estate AG HAUFE. Hammonia SCHÄFFER POESCHEL LSL NEW TIMES semigator.de Das Seminar- und Weiterbildungsportal

HAUFE: THE ROOTS

Krause/Zander
Arbeitssicherheit
Praxishandbuch für Arbeits- und Gesundheitsschutz
Haufe
Arbeitsschutz Office Professional
Haufe

Loose-leaf editions

Desktop content databases (1990s)

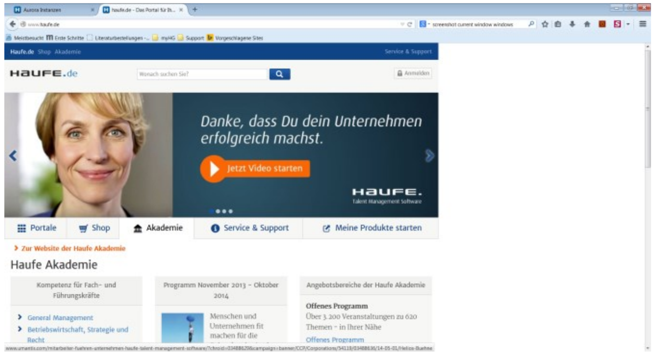

Christian E. Elger
Neuroleadership
Erkenntnisse der Hirnforschung für die Führung von Mitarbeitern
2. Auflage
HAUFE.

Books

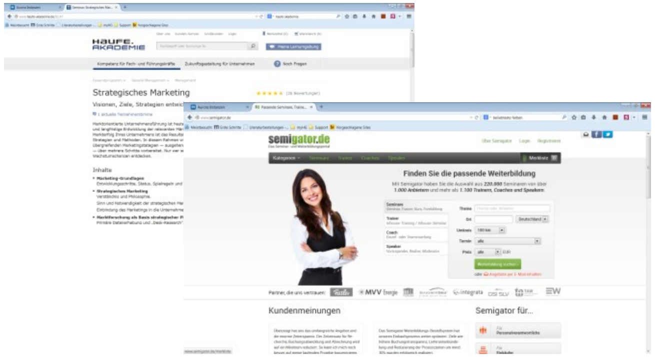
HAUFE TODAY

HAUFE.Gruppe

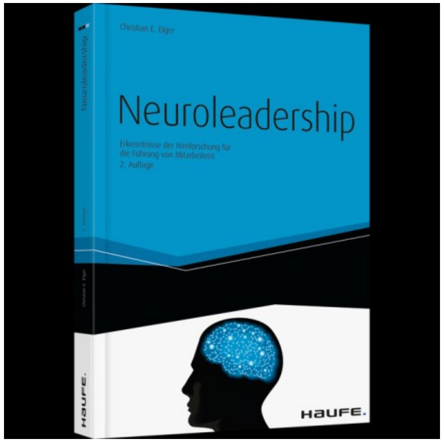
Online Content Databases


HAUFE.de
Danke, dass Du dein Unternehmen erfolgreich machst.
Jetzt Video starten
HAUFE. Talent Management Software
Portale
Shop
Akademie
Service & Support
Meine Produkte starten
Zur Website der Haufe Akademie
Haufe Akademie
Kompetenz für Fach- und Führungskräfte
Programm November 2013 – Oktober 2014
Angebotsbereiche der Haufe Akademie
General Management
Betriebswirtschaft, Strategie und Recht
Menschen und Unternehmen fit machen für die
Offenes Programm
Über 3.200 Veranstaltungen zu 620 Themen – in Ihrer Nähe

Haufe.de Portal Site



Booking platforms for seminars & trainings

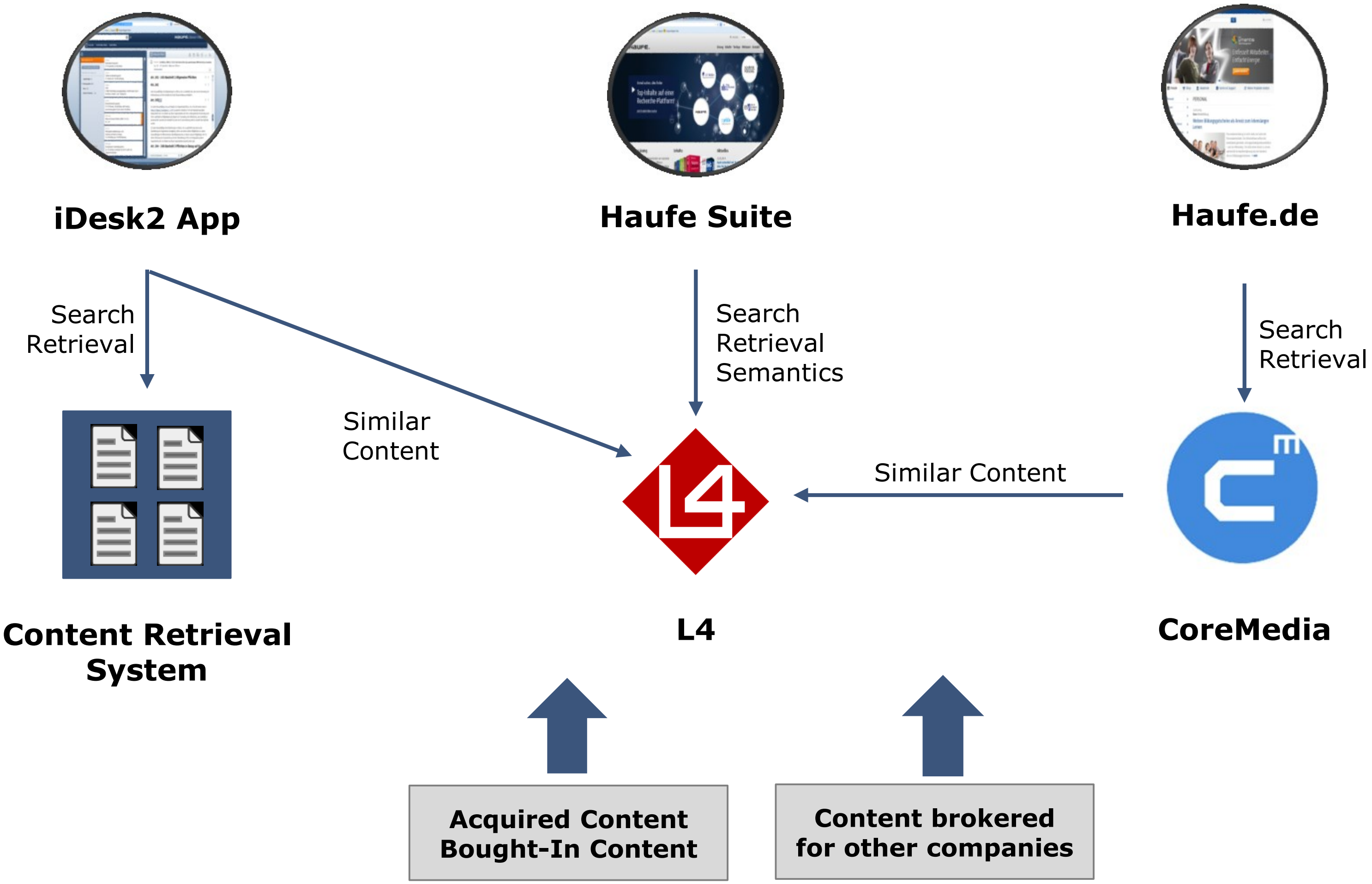

Christian E. Elger
Neuroleadership
Erkenntnisse der Hirnforschung für die Führung von Mitarbeitern
2. Auflage
HAUFE.

Books & Print Products

CONTENT @ HAUFE

- **50 million XML documents (Haufe Content)**
  - Own set of domain-specific DTDs
  - Proprietary Python-based publishing pipeline
    - Conversion to XML
    - Conversion to target formats (PDF, Database files)
- **Auxiliary content: PDFs, audio-visual content, forms, embedded applications**
- **News Posts**
- **Seminar descriptions**

PROBLEM:
SATURATED CONTENT MANAGEMENT



iDesk2 App

Haufe Suite

Haufe.de

Search
Retrieval

Similar
Content

Search
Retrieval
Semantics

Search
Retrieval

Similar Content

Content Retrieval System

L4

CoreMedia

Acquired Content
Bought-In Content

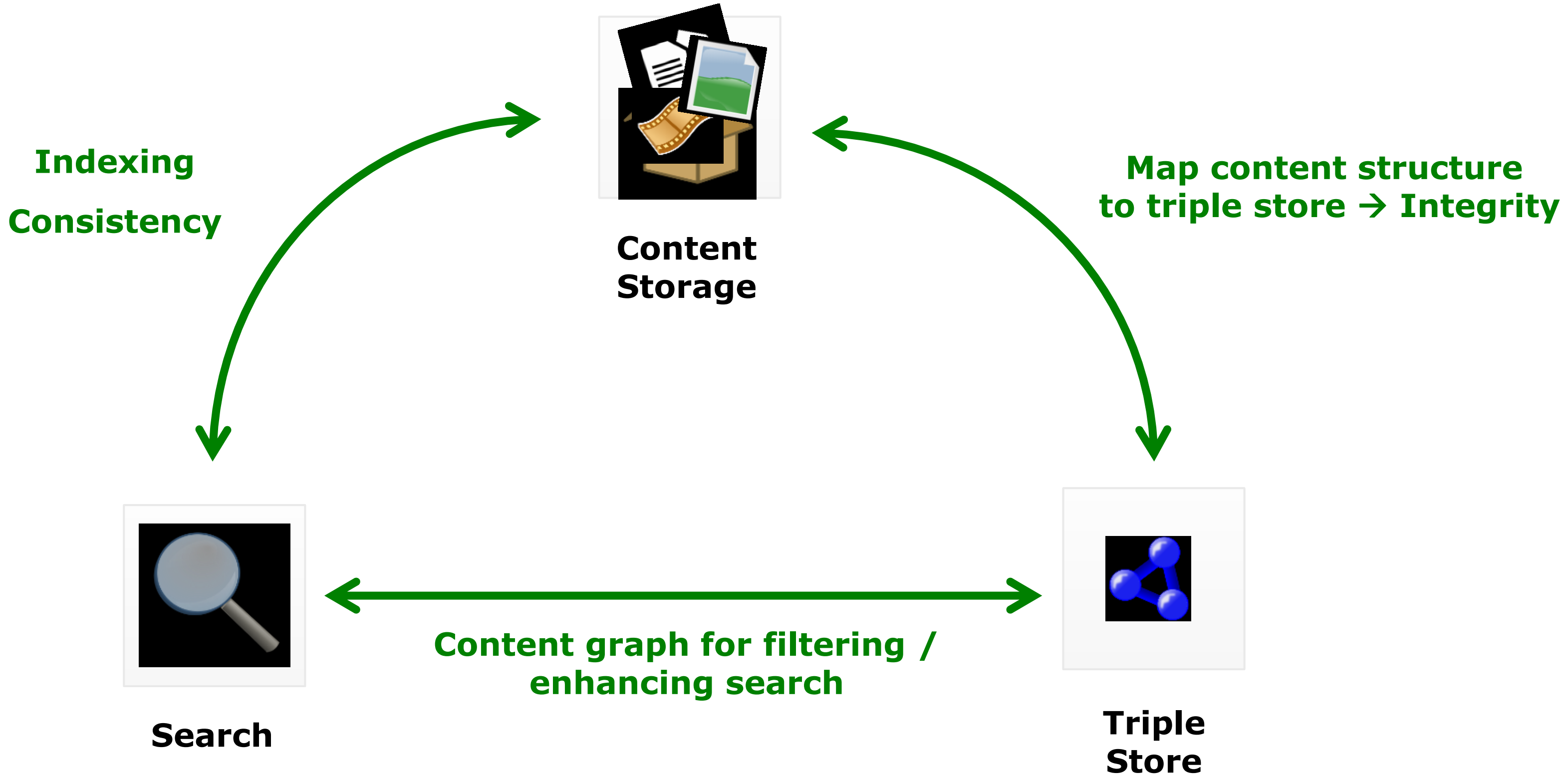
Content brokered
for other companies

PROBLEM:
SATURATED CONTENT MANAGEMENT

1. **Complicated Content Reuse / Cross-Referencing**
2. **Difficult Authorization**
3. **Massive Content Duplication**
4. **High System heterogeneity → Increased management efforts**

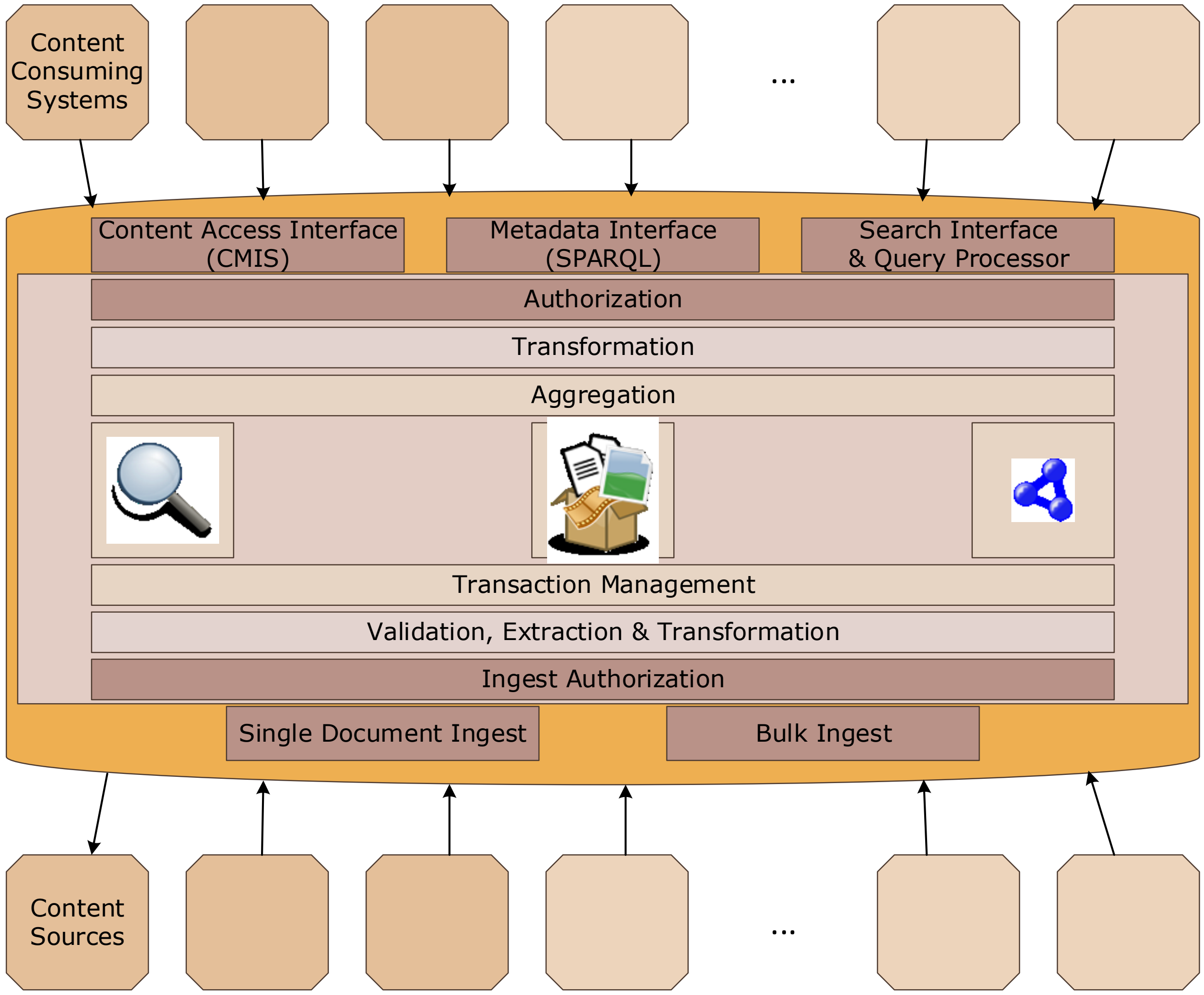
Vision: Unified Content Hub

FUNCTIONAL BUILDING BLOCKS

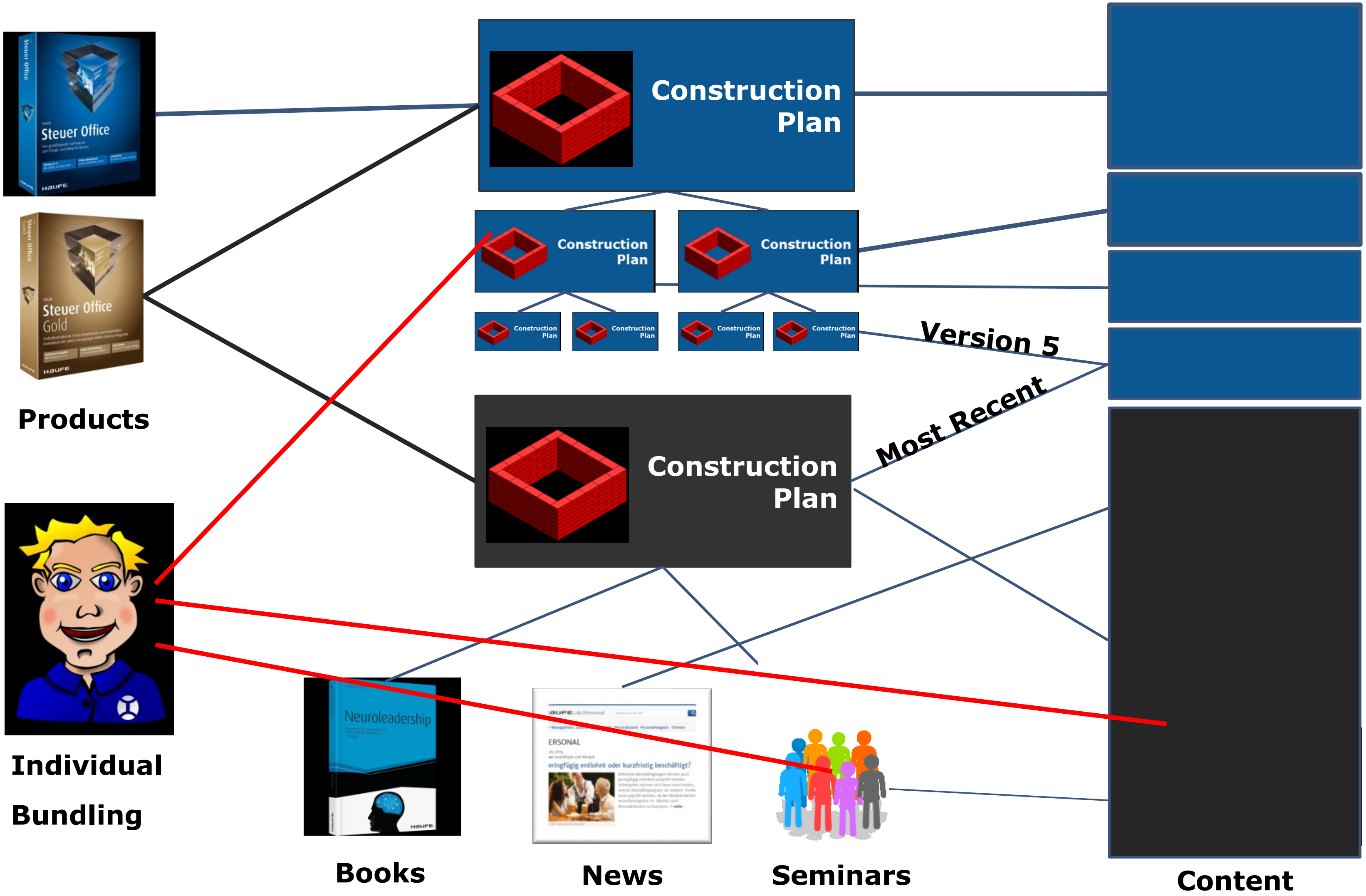

Content Storage
Indexing
Consistency
Map content structure to triple store → Integrity
Search
Triple Store
Content graph for filtering / enhancing search

CONTENT HUB ARCHITECTURE

HAUFE.Gruppe


Content Consuming Systems
...
Content Access Interface (CMIS)
Metadata Interface (SPARQL)
Search Interface & Query Processor
Authorization
Transformation
Aggregation
Transaction Management
Validation, Extraction & Transformation
Ingest Authorization
Single Document Ingest
Bulk Ingest
Content Sources
...

WHY TRIPLES?



Construction Plan

Construction Plan

Construction Plan

Construction Plan

Construction Plan

Construction Plan

Construction Plan

Construction Plan

Version 5

Most Recent

Products

Individual Bundling

Books

News

Seminars

Content

WHY TRIPLES?

Enables fast answers to complex questions

- Display all seminars that discuss „Neuroleadership?"
- Enable cross references from free content (news posts) to relevant paid products

RDF and triples for modeling relationships

SPARQL 1.1 for graph traversal

EXISTING EXPLICIT RELATIONS

```
<link.norm bezeichner="paragraph" kuerzel="EStG" zahl="32">
 § 32 des Einkommensteuergesetzes
</link.norm>

<link.text zielid="HI39751.gen1">
Über dieses Dokument
</link.text>

<kuerzel basis="Einkommensteuer-Richtlinien 1999">
 Einkommensteuer-Richtlinien 1999
</kuerzel>
```

IMPLEMENTATION OPTIONS

Use Document Repository

Think Fedora or JCR

Scale-Out?

XML-agnostic

→ No XPath, XQuery…

Integrate XML-DB

With scalable search & triple store

Certainly doable

Core Problem: Data Integrity

Use XML-based, integrated solution

Don`t reinvent the wheel

Focus on developing add-on functionality

License fees

Dependency on vendor

TIMELINE: PAST, PRESENT, FUTURE

September 2013: **Business department wants TWO new systems:**

- Global Content Search
- Unified Content Hub

Fall 2013 **Three Software architects create two architectural drafts**

Outcome: Search without docs? Store without search?

Data Integrity? How to deal with graph structure?

Winter 2013/2014 **Consolidation of Drafts → Unified Content Hub**

Spring 2014 **Proof of Concept with major XML NoSQL vendor**

- Identification of additionally required external services
- Further elaboration of triple use

Summer 2014- **Start of Implementation**

SUMMARY & CONCLUSION

- **Consolidation of saturated storage and search services**

  → Avoid content duplication

  → No duplicated indexing

  → Reduce infrastructure and management costs

- **Indexing XML Structure is vital**

  → Faceted search & complex search using XPath / XQuery

- **Triples for relationship management**

  → Will allow querying structure in real-time

  → Triples for modeling

  → SPARQL1.1 for querying and graph traversal

- **Currently working towards first implementation**