

Lehrveranstaltung

Semantic Web Technologien

WS 2009/10

HTWG Konstanz

Vorstellung

- **Dipl.-Inf (FH) Steffen Schlönvoigt**
 - 2001–2003 Studium Mathematik
Uni-Konstanz
 - 2003–2006 Software-Entwicklung (C#/.Net-Umfeld)
 - 2004–2008 Studium Software-Engineering
HTWG Konstanz
 - 2006–2008 Software-Entwicklung (Enterprise CMS & Web 2.0)
Young Media Concepts, Kreuzlingen
 - 2008–2009 Diplomarbeit “Konzept und Implementierung von
Integrationsszenarien mit Semantic Web
Technologien” bei der TechniData AG, Markdorf
 - Seit 2009 Software-Entwicklung (ABAP, SAP Workflow)
TechniData AG, Markdorf

Organisation

- Wöchentliche Lehrveranstaltung
 - Wann:
 - Montag Nachmittag 15:45 - 17:15
 - Wo:
 - Raum F033
 - Was:
 - Vorträge kombiniert mit praktischen Einheiten
 - Prüfung:
 - Mündlich am Ende des Semesters
 - Sonstiges:
 - 3 ECTS
 - Material unter <http://www.schloenvoigt.de>
 - Kontakt per E-Mail: steffen@schloenvoigt.de

Literaturliste:

- Buch “Semantic Web: Grundlagen”, eXamen.Press
Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph, York Sure
- Seiten des W3C zum Semantic Web unter <http://www.w3.org/2001/sw/>
 - RDF-Primer: <http://www.w3.org/TR/rdf-primer/>
 - RDF-Schema: <http://www.w3.org/TR/rdf-schema/>
 - OWL-Overview: <http://www.w3.org/TR/owl-features/>
- Protégé OWL Tutorial:
 - <http://www.co-ode.org/resources/tutorials/protege-owl-tutorial.php>

Themen

- Was ist das Semantic Web?
- Sprachen des Semantic Web
- Ontologien
- Bibliotheken, Frameworks und andere Tools
- Logik und Inferenz
- Anwendungen des Semantic Web

Geschichte des Internets

- Anfang der 1960er Jahre:
 - Es entsteht die Idee eines dezentralen Netzwerks.
 - Computer sollen darüber kommunizieren können.
- 1968/1969 und 1970er: ARPA net
 - Basiert auf technischen Spezifikationen in RFC-Dokumenten
 - Großrechner der UCLA, SRI, UCSB und der University of Utah
 - Viele grundlegende Technologien des heutigen Netzes entstehen:
 - TCP/IP, E-Mail, FTP, Telnet ...
- 1980er
 - Gründung der IETF
 - Das Web besteht vor allem als Plattform für den Austausch von Infos unter Akademikern
 - Kommerzielle Interessen werden kaum verfolgt

Geschichte des Internets

- 1989: Tim Berners-Lee entwickelt am CERN das WWW
 - Basiert technisch auf
 - HTTP, URLs, HTML
 - Verteilung und Verlinkung formatierter Texte
 - Kommerzielle Nutzung des Internets beginnt
- Anfang der 90er: Erste Browser (Mosaic)
- Mitte der 90er: Durchbruch des WWW
 - Hype um Netscape Navigator
 - Microsoft integriert den Internet Explorer in sein OS
- Seitdem: unbändiges Wachstum
 - Momentan 1,668,870,408 Benutzer (Stand 2009)

- Unabhängiges Konsortium zur Entwicklung der technischen Standards des WWW
- 1994 gestartet, 1995 Beitritt erster Unternehmen
- Strukturiert in Aktivitäten
 - HTML, XML, CSS, XSLT, RDF, OWL ...
- Jede Aktivität besteht aus
 - Working Groups: Erstellen die “Standards”
 - Interest Groups: Beraten die Mitglieder der Wgs
- Standards werden als “Recommendation” veröffentlicht:
 - <http://www.w3.org/TR/>

- Das “Mitmach-Web”
- Jeder kann Informationen ins Netz stellen
 - Und viele tun das auch im großen Maße:
 - StudiVZ, Facebook, Flickr, twitter, del.icio.us, Blogs ...
- Unternehmen und Organisationen veröffentlichen ihre Daten frei verfügbar im Netz
- Inhalte werden von Benutzern in Zusammenarbeit erstellt
 - Wikipedia ...
- Benutzer “taggen” Inhalte

Warum Semantic Web ?

- Das Internet birgt eine immense Fülle an Daten
- Wichtige Informationen unter diesen Daten zu finden wird immer schwieriger
- Heutige Suchen basieren auf Textvergleich
- Daten sind hauptsächlich in HTML codiert und mit Hyperlinks verbunden
 - Computer sehen zwar...
 - Text + Formatierung
 - Grafiken
 - Multimedia-Inhalte
 - ...verstehen aber deren Bedeutung nicht
- Ohne Kontext sind die Ergebnisse oft zu ungenau

Warum Semantic Web ?

The screenshot shows a news website layout with several sections:

- Top Left:** Navigation links for "Bilder: Oktoberfest", "Bilder: Wien 2008", "Bilder: Wien-Meister 2007", and "Artikel senden".
- Top Right:** Advertisements for "WELT GRUPPE" and "Anzeigen-Sonderveröffentlichung".
- Middle Left:** Three small images with captions: "Vettel: Glaubt wieder an den WM-Titel", "Erotik: Kennen Sie alle Geheimnisse?", and "Wetten, dass...?: Top-Quote dank Michelle".
- Middle Right:** Three "Private Krankenkasse" advertisements for "TOP - Testsieger Private Krankenkasse".
- Bottom Left:** "NACHRICHTEN IM ÜBERBLICK" section with sub-sections for "POLITIK" (Islamisten rufen zu Terror in Deutschland auf), "WIRTSCHAFT" (Auftragsflaute: Autohändler leiden unter Ende des Abwrackbooms), "GESCHICHTE IN KARTEN" (Zweiter Weltkrieg), and "SATIRE" (Glasauge).
- Bottom Right:** "POLITIKER-HÖRQUIZ" section with a photo of a man and the text "Wer sagte was? Lustige Patzer & harte Attacken".

Wo sind die Informationen?

Was bedeuten die Informationen?

Was ist Unterhaltung?

Was ist nur Werbung?

Welche Information ist relevant?

Wie glaubhaft ist die Information?

Menschen hingegen haben "Kontext- /Weltwissen"

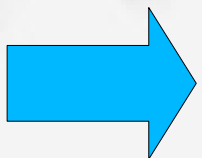
- Erschließen sich fehlende Informationen
- Können mentale Zusammenhänge knüpfen
- Verwenden mehrere Sinne gleichzeitig
- Finden und „verstehen“ die Informationen (meistens)

Warum Semantic Web ?

- Aufgaben erfordern oft, dass wir Informationen aus verschiedenen Quellen im Web verknüpfen
 - Wir verwenden mehrere Suchmaschinen
 - Wir suchen auf mehreren Portalen nach der idealen Reise
- Menschen können dies (relativ) problemlos!
 - Sogar wenn verschiedene Begriffe für die gleichen Dinge verwendet werden
 - Oder wenn der gleiche Begriff für verschiedene Dinge verwendet wird

Warum Semantic Web ?

- Computer stehen hierbei jedoch vor enormen Problemen:
 - Lückenhafte Informationen sind nutzlos
 - Verstehen nur schwer den Inhalt von Bildern / Videos
 - Erkennen nur schwer Ähnlichkeiten
 - Können verschiedene Informationen nur schwer kombinieren
- Ist ein Auto das gleiche wie ein PKW ?
- Das heutige WWW ist auf Menschen ausgerichtet!



Den Daten fehlt die SEMANTIK

Warum Semantic Web ?

Definiton Semantik:

Die **Semantik** (griech. σημαίνειν sēmainein = Bedeutungslehre) ist das Teilgebiet der Sprachwissenschaft (Linguistik), das sich mit

- Sinn und
- Bedeutung

von Sprache beziehungsweise sprachlichen Zeichen befasst.

Die Semantik kümmert sich um die Frage, wie Sinn und Bedeutung von komplexen Begriffen aus Sinn und Bedeutung von einfachen Begriffen abgeleitet werden können und stützt sich dabei in der Regel auf die **Syntax** (Zusammenstellung, Satzbau, steht für die normative Struktur von Daten).

Beispiel:

$$3 + \ddot{u} = \%$$

Syntax+Semantik falsch

$$3 + 4 = 12$$

Semantik falsch

$$3 + 4 = 7$$

Syntax und Semantik korrekt

Beispiele

Wo könnten uns Computer helfen, wenn ihnen die Daten und deren Semantik offenliegen würden?

- Reiseplanung
 - Flug / Reise suchen -> buchen
 - Urlaub beantragen / Kalendereintrag
 - Freunden / Familie Bescheid geben
 - Wer passt auf die Wohnung auf?
- Daten-Integration
 - Verschiedene Schemata kombinieren
 - Daten verschiedener DBs verschmelzen
- Social Networks
 - Daten eingeben?
 - Freunde mitnehmen?

Problemfelder

- Informationssuche: Wo finde ich ... ?
 - Keyword-Suche findet zu viele nicht relevante und nicht alle relevanten Informationen
 - => Semantisches Tagging
- Informationsextraktion: Was ist wichtig ... ?
 - Nur Menschen können **korrekt** Informationen extrahieren
 - Computern fehlt **“Welt-/Kontextwissen”**
 - => Semantische Annotation
 - Implizites Wissen muss durch logische Schlussfolgerung aus Kombination vorhandener Informationen ermittelt werden
 - => Einbindung formaler Logik

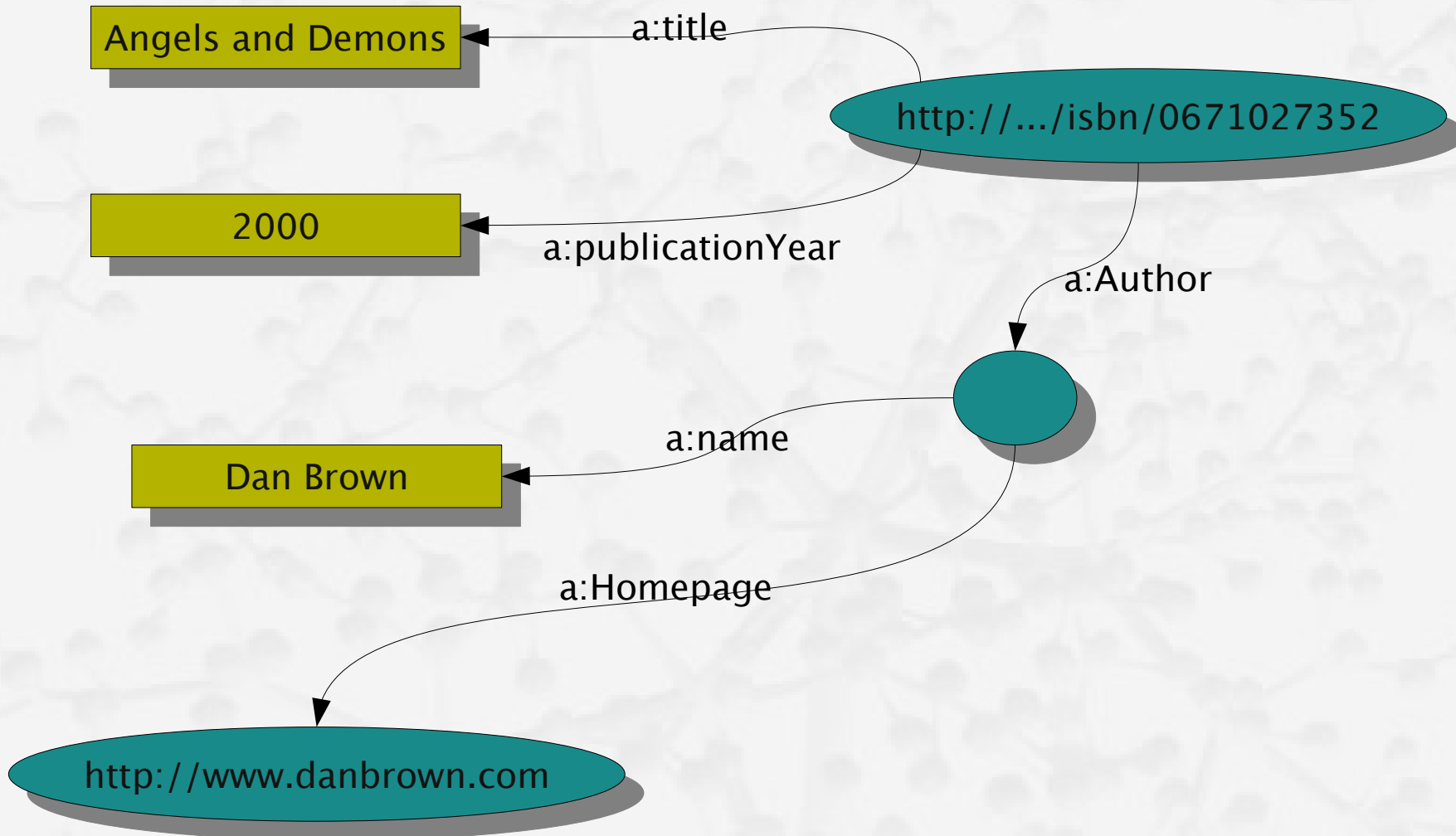
Problemfelder

- **Wartung: Sind meine Daten noch konsistent ?**
 - Konsistenz, Korrektheit, Aktualität von Daten
 - => Semantische Informationen **über** SEMANTIK
- **Personalisierung: Was interessiert mich ?**
 - Anpassung des dargestellten Informationsgehalts an persönliche Bedürfnisse und Anforderungen
 - Woher kommen die benötigten Daten?
 - => Repräsentation von SEMANTIK

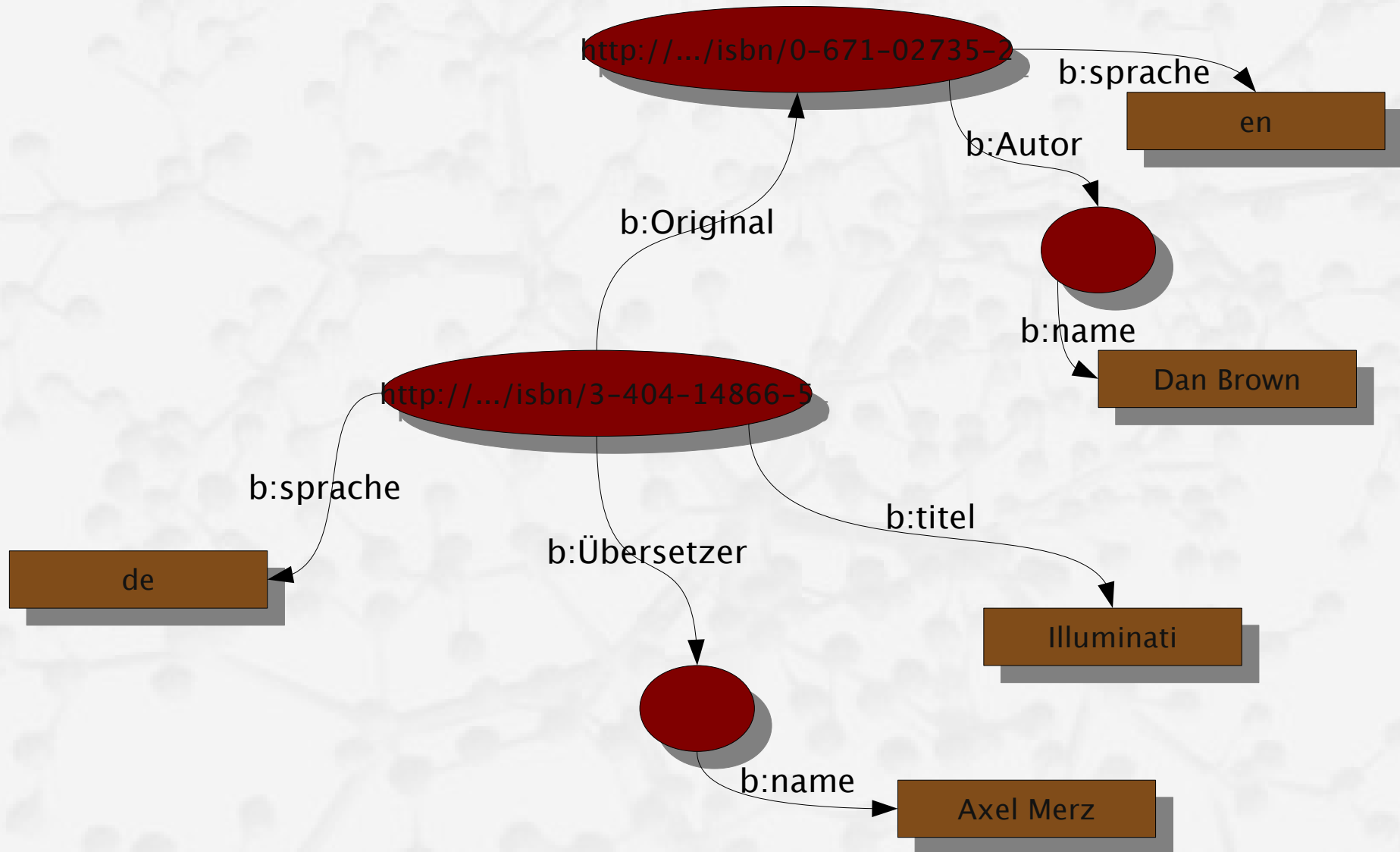
Zusammenfassung

- Das WWW ist (heute) zur Nutzung durch den Menschen bestimmt
- Lösungsansatz:
 - Beschreibung der Bedeutung der dargestellten Information mit Hilfe standardisierter Formalismen
 - Formalismus = maschinenlesbar
 - Standard = interoperabel
 - Metadaten = "Daten über Daten"
 - strukturierte Daten, mit deren Hilfe eine Informationsressource beschrieben und dadurch besser auffindbar gemacht wird

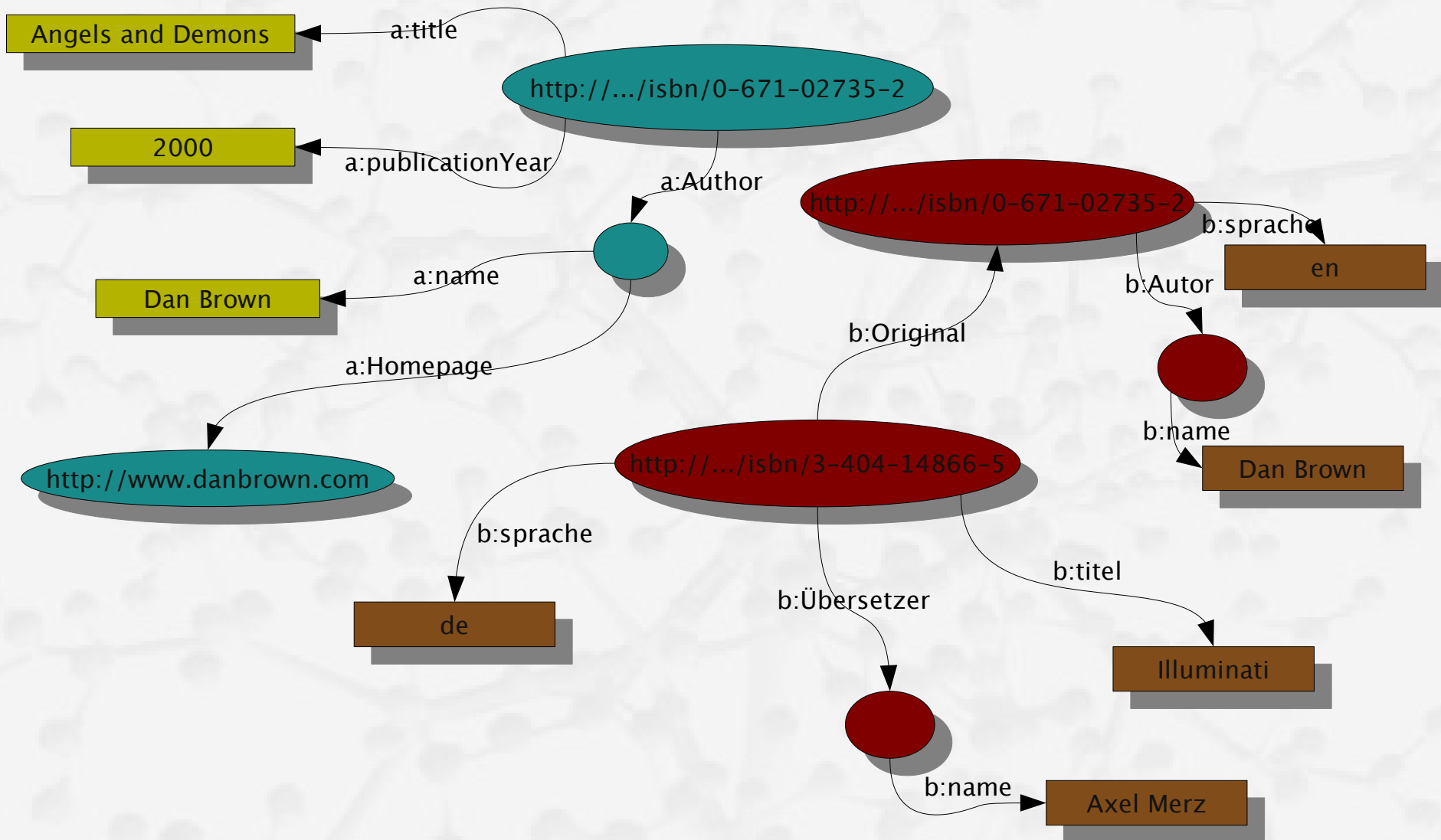
Beispiel: Datenintegration



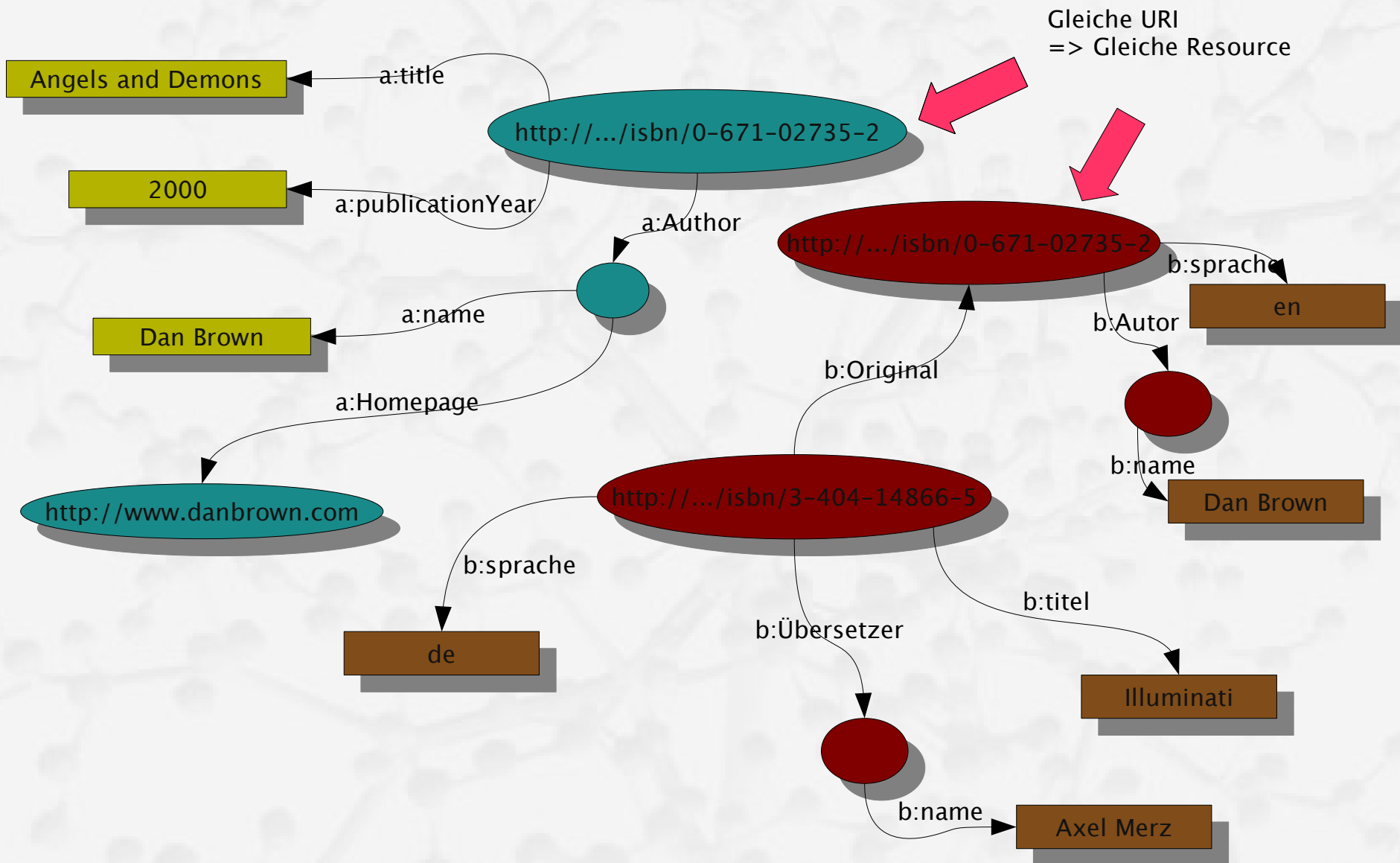
Beispiel: Datenintegration



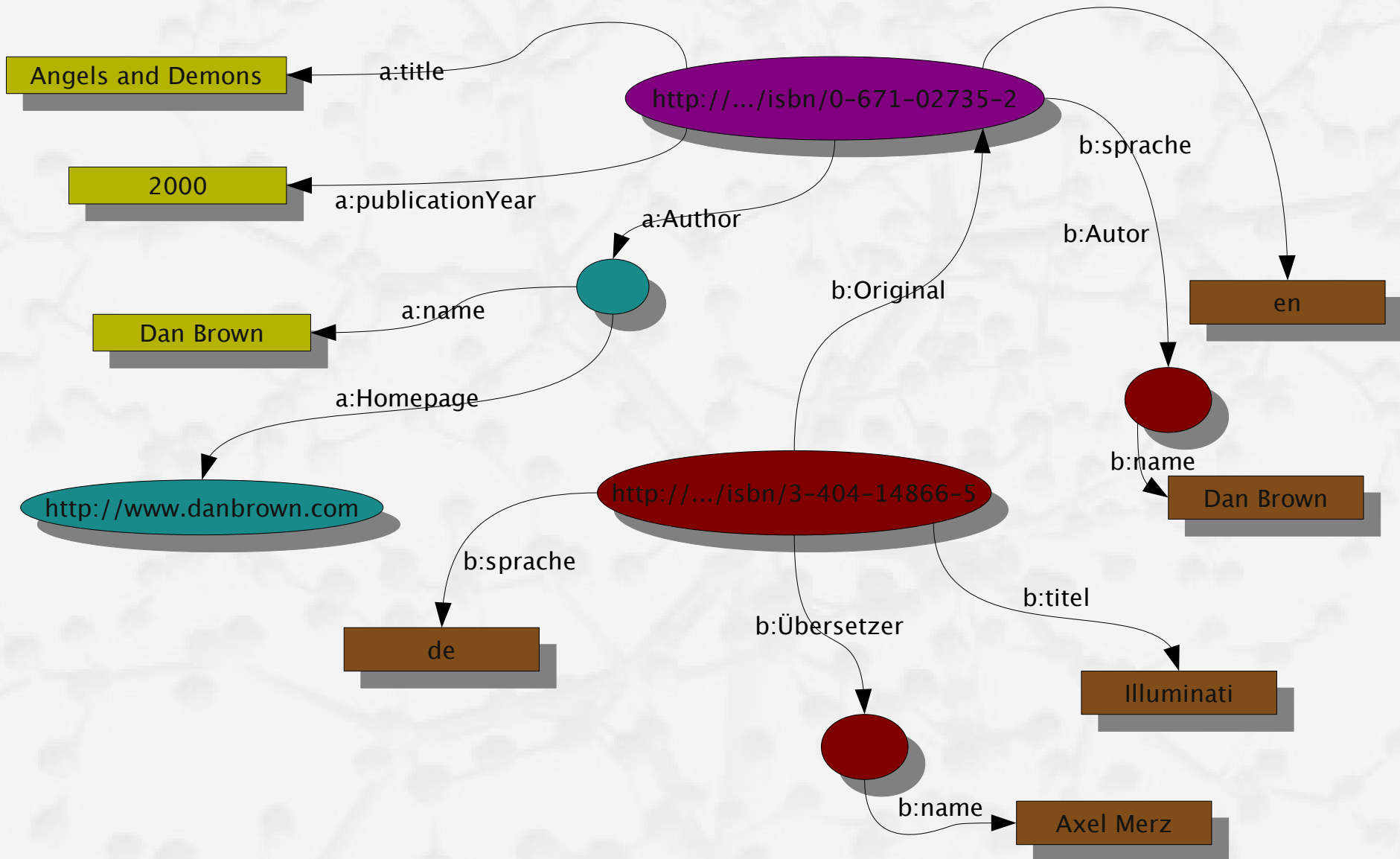
Beispiel: Datenintegration



Beispiel: Datenintegration



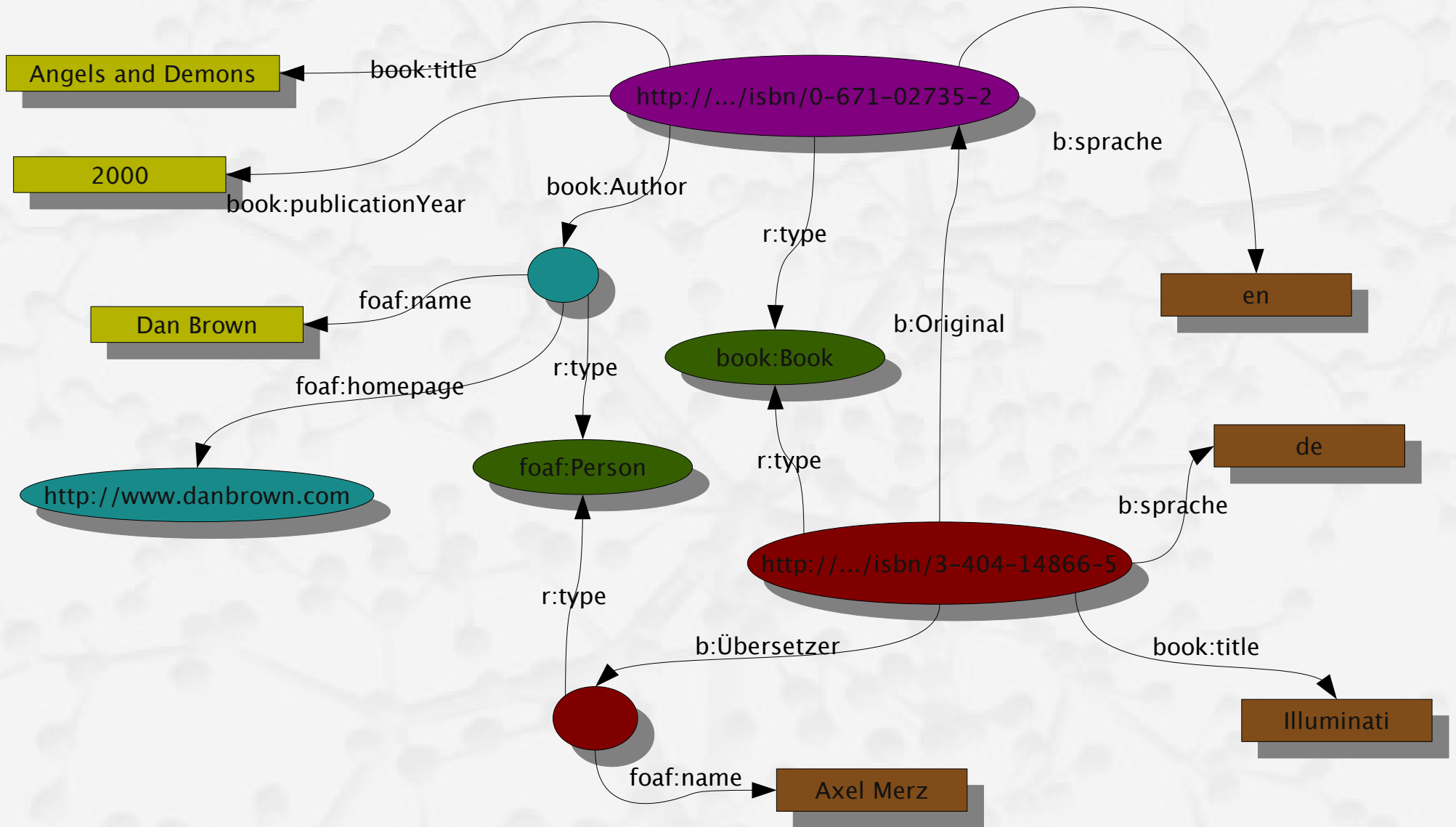
Beispiel: Datenintegration



Wow! -> Daten verknüpft!

- Aber wir können noch mehr, indem wir weitere Informationen hinzufügen:
 - a:Author sameAs b:Autor
 - a:title sameAs b:titel
 - Kategorisierung
 - Personen
 - Bücher
 - Usw
 - Verwendung von bereits vorhandenem Vokabular

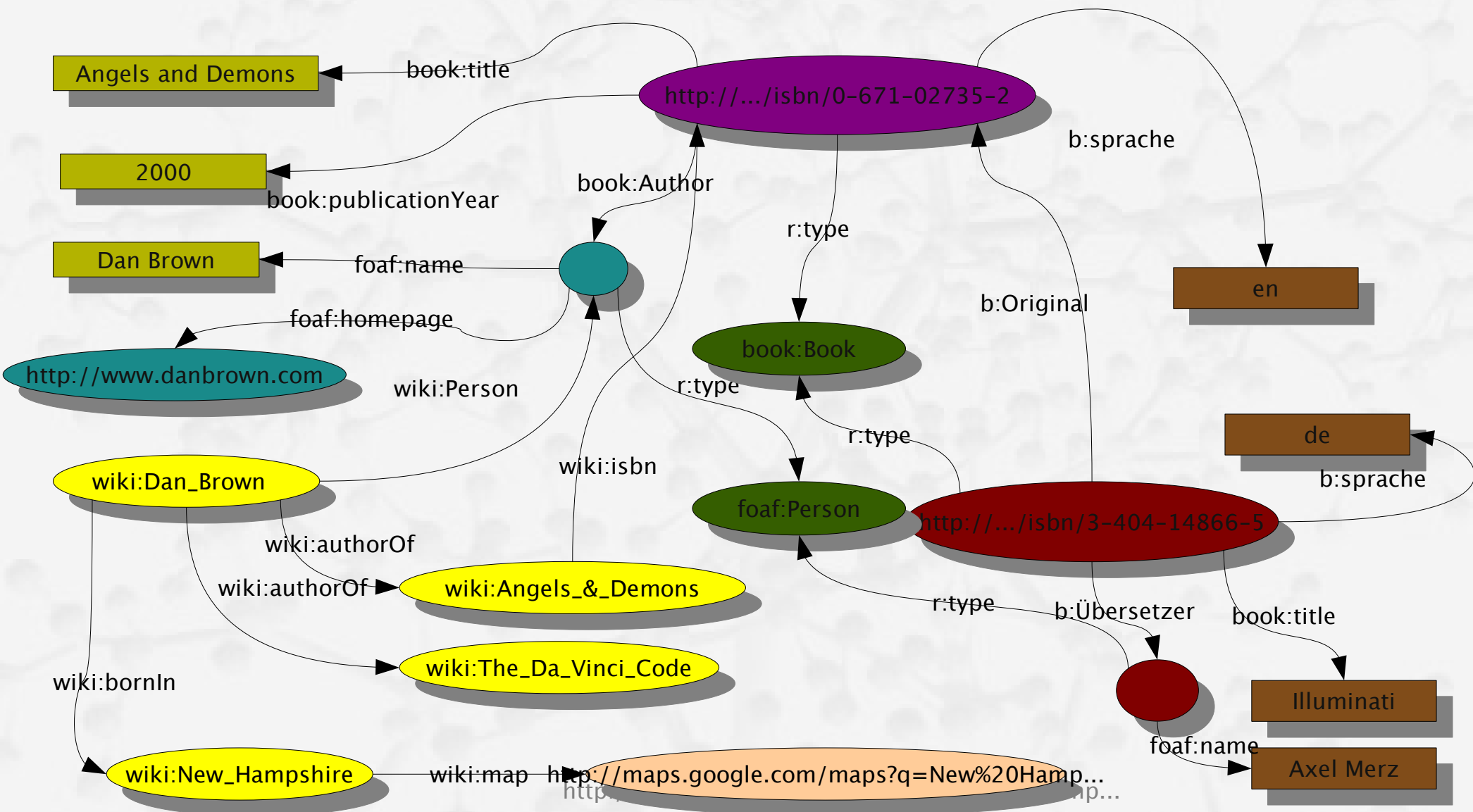
Beispiel: Datenintegration



Das ist noch nicht alles

- Durch Typisierung können wir evtl. weitere Informationen hinzufügen:
 - Zum Beispiel durch geeignete Tools extrahierte Informationen aus der Wikipedia
 - Oder Rezensionen von Amazon
 - ...

Beispiel: Datenintegration



Was ist das Semantic Web ?

- Das Semantic Web ...
 - ... ist ein **Web der Daten**, vergleichbar mit einer riesigen **globalen Datenbank**
 - ... besteht aus einer Ansammlung **semantischer Aussagen**
 - ... fördert die **intelligente Verwendung und Verarbeitung von Wissen**, das im Web bereitgestellt wird
 - ... ist eine **Erweiterung** des heutigen Internets, in der vorhandenen Informationen eine **wohldefinierte Bedeutung** gegeben wird, um die **Zusammenarbeit zwischen Menschen und Maschinen** zu verbessern.
- Programme müssen **heterogene Daten** austauschen und gemeinsam verarbeiten können, auch wenn diese Programme völlig unabhängig voneinander entwickelt worden sind

Semantic Web Roadmap

- **Stufe 1:** Textdokumente und Datenbankeinträge
- **Stufe 2:** XML-Dokumente mit einheitlichem Vokabular
- **Stufe 3:** RDF-Taxonomien und Dokumente mit unterschiedlichem Vokabular
- **Stufe 4:** OWL Ontologien und automatisches Schlussfolgern

Semantic Web Roadmap

- Stufe 1: Textdokumente und Datenbankeinträge
 - Daten sind anwendungsbezogen (proprietär)
 - Semantik steckt in Anwendungen – nicht in Daten
 - Beispiele:
 - Warenkatalog von Amazon
 - Index von Google
 - Datenbestand bei Ebay
 - ...

- Stufe 2: XML-Dokumente mit einheitlichem Vokabular
 - Innerhalb bestimmter vorgegebener Domäne erreichen Daten Unabhängigkeit von der Anwendung und können so zwischen Anwendungen ausgetauscht werden
 - Beispiel:
 - Dublin-Core Bibliografien zur unabhängigen Beschreibung von Dokumenten
 - Google API Webservice
 - Twitter API
 - ...

Semantic Web Roadmap

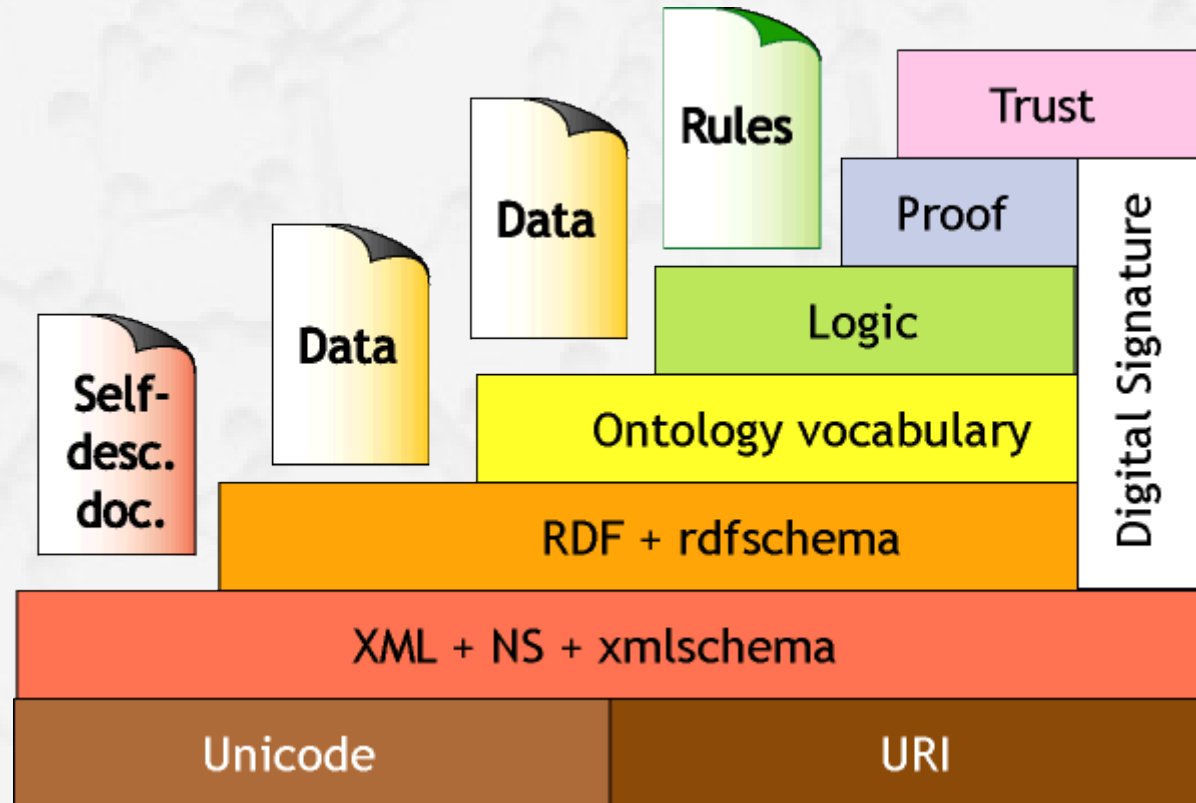
- Stufe 3: RDF-Taxonomien und Dokumente mit unterschiedlichem Vokabular
 - Daten stammen aus unterschiedlichen Domänen und werden durch hierarchisch strukturierte Taxonomien/Partonomien klassifiziert.
 - Einfache Beziehungen zwischen den einzelnen Kategorien dienen der Kombination von Daten.
 - Beispiel:
 - „Semantic Web Technologien“ ist eine „Vorlesung“ ist eine „Lehrveranstaltung“ des „Software Engineering“
 - „Hans Mustermann“ studiert „Software Engineering“ ist ein „Studiengang“ an der „Fakultät Informatik“ gehört zur „HTWG Konstanz“

- Stufe 4: OWL-Ontologien und automatisches Schlussfolgern
 - Neue Daten können aus bestehenden Daten gewonnen werden durch die Anwendung von Schlussfolgerungsregeln.
 - Daten, Eigenschaften von Daten und Beziehungen zwischen Daten lassen sich semantisch beschreiben.
 - Beispiele
 - automatisches Übersetzen eines Dokuments von einer Domäne in eine andere
 - Formulieren komplexer Abhängigkeiten und Bedingungen
 - Anwendung „versteht“ Daten

• Kann man semantischen Daten trauen?

- Authentizität
 - Ist der mutmaßliche Urheber (Autor) auch der tatsächliche?
- Integrität
 - Wurden die Daten manipuliert?
- Public Key Kryptografie und digitale Signaturen
 - Semantische Daten müssen signiert werden, damit Inferenzsysteme deren Zuverlässigkeit mit in Betracht ziehen können
- Semantisches Dokument = Sammlung von Aussagen + zugehörige Signaturen

Architektur des Semantic Web



• Was ist das Semantic Web nicht?

- „The Semantic Web is not AI“
 - RDF und OWL sind
 - (im Vergleich zur Künstlichen Intelligenz)relativ simple Beschreibungssprachen
 - einfache Art, untereinander austauschbare Daten zu formulieren und zu beschreiben
 - eine Möglichkeit, einfache Aussagen zu treffen
 - bieten primitive Möglichkeit der Inferenz innerhalb eines beschränkten Rahmens
 - ...and that's it!
- „One goal in SW is to keep things relatively simple and not necessarily seek absolute completeness“

- Nicht Teil des Semantic Webs sind (jetzt und in naher Zukunft)
 - assoziatives Schlussfolgern
 - räumliches Denken
 - Erkennung von Text, Bildern, Gesten, etc...
 - Komplexe Entscheidungsprozeduren
 - Schlussfolgern aus unvollständigen oder kontextabhängigen Informationen
 - „Strong AI“
- „Just as Prolog is not AI but merely a useful tool for it, SW might be just a good tool for AI“