Heiko Jansen / Pascal Christoph

Dynamische Kataloganreicherung auf Basis von Linked Open Data

101. Deutscher Bibliothekartag
Hamburg, 24.05.2012

Dieser Vortrag steht unter einer Creative Commons Namensnennung - Keine Bearbeitung 3.0 Deutschland Lizenz (CC BY-ND 3.0).
http://creativecommons.org/licenses/by-nd/3.0/de/

Davon ausgenommen sind die verwendeten, nicht von den Autoren erstellten Grafiken, Screenshots und Bilder, deren jeweilige Rechte und Lizenzbedingungen fortgelten.

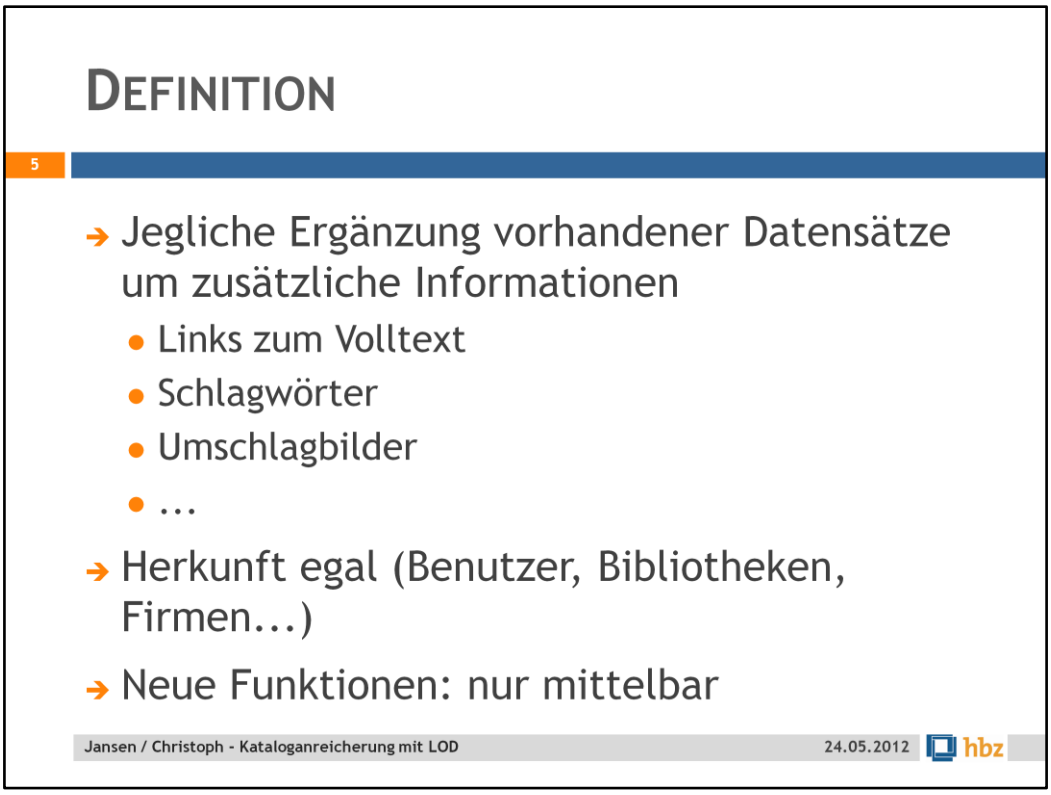
Insbesondere erfordert eine Verwendung der aus dem Portfolio von iStockphoto stammenden Grafiken auf den Folien 4, 5 und 6 den Erwerb einer entsprechenden Lizenz.

Übersicht

- Definition
- Methoden
- Projekt-Praxis
  - Technik
  - Matching
  - Verlinkung
- Mögliche Weiterentwicklung
- Fazit & Anwendungsbeispiel

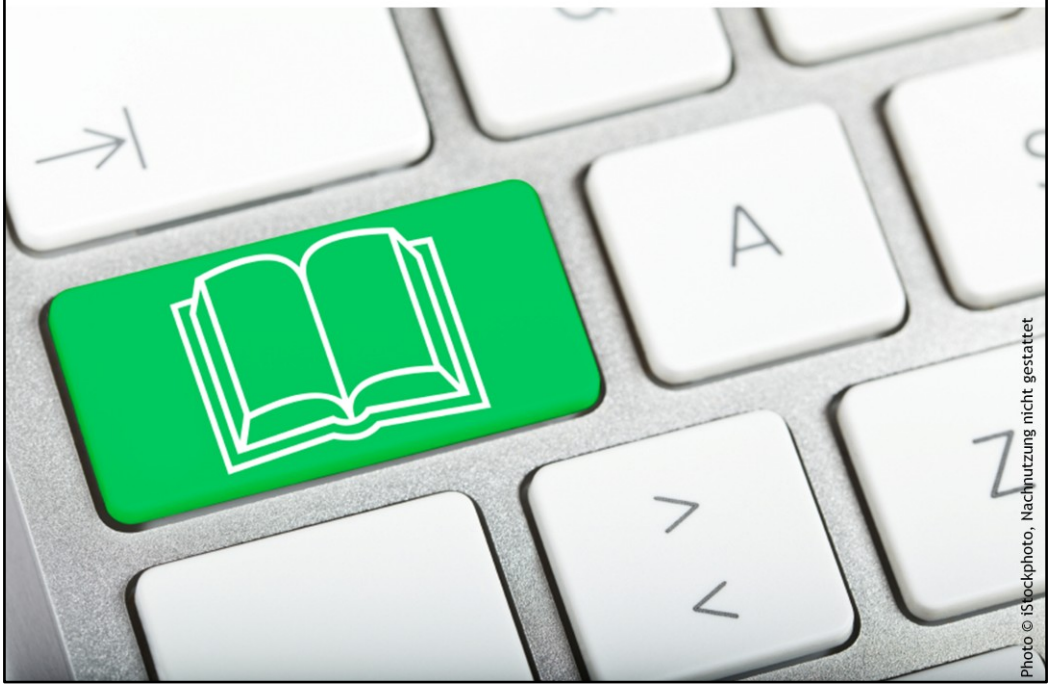
KATALOGANREICHERUNG?

Photo © iStockphoto, Nachnutzung nicht gestattet



Auch der eigene Katalog als Quelle: Zusammenführung von elektronischer und gedruckter Fassung; verschiedenen Auflagen usw.: Fließender Übergang zu FRBR

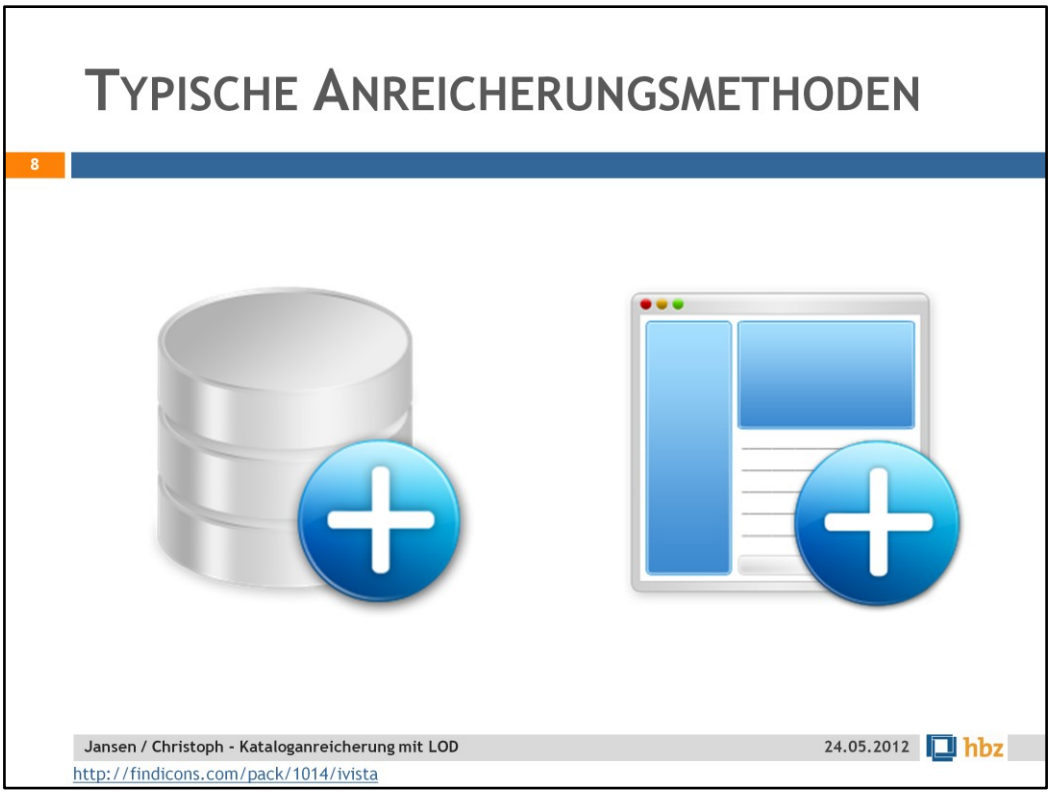
„Instant Gratification“



Begriff stammt aus: Coffmann, Steve: The response to „Building Earth´s Largest Library“. In: Searcher 7.7 (1999), S. 29-32. ISSN 1070-4795

Nutzer erwarten direkten Zugang zu den gefundenen Dokumenten, müssen bei physischen Medien aber in die Bibliothek kommen: wird als hohe Hürde wahrgenommen.
Je sicherer sie sein können, etwas zu bekommen, das ihren Interessen entspricht, desto eher werden sie bereit sein, die „Kosten“ zu tragen: ausführliche Info hilft.
Vgl. umfangreiche Info bei Amazon zur Erleichterung der Kaufentscheidung.
„Kosten“ bei Bibliotheken = Zeit und Weg

Zusammenstellung von „schmackhaften“ Infos aus vielen Quellen



Anreicherung der Datenbank zu beliebigen Zeiten

Anreicherung der Anzeige dynamisch im Moment des Aufrufs

Vor- und Nachteile

Dynamisches Mashup:

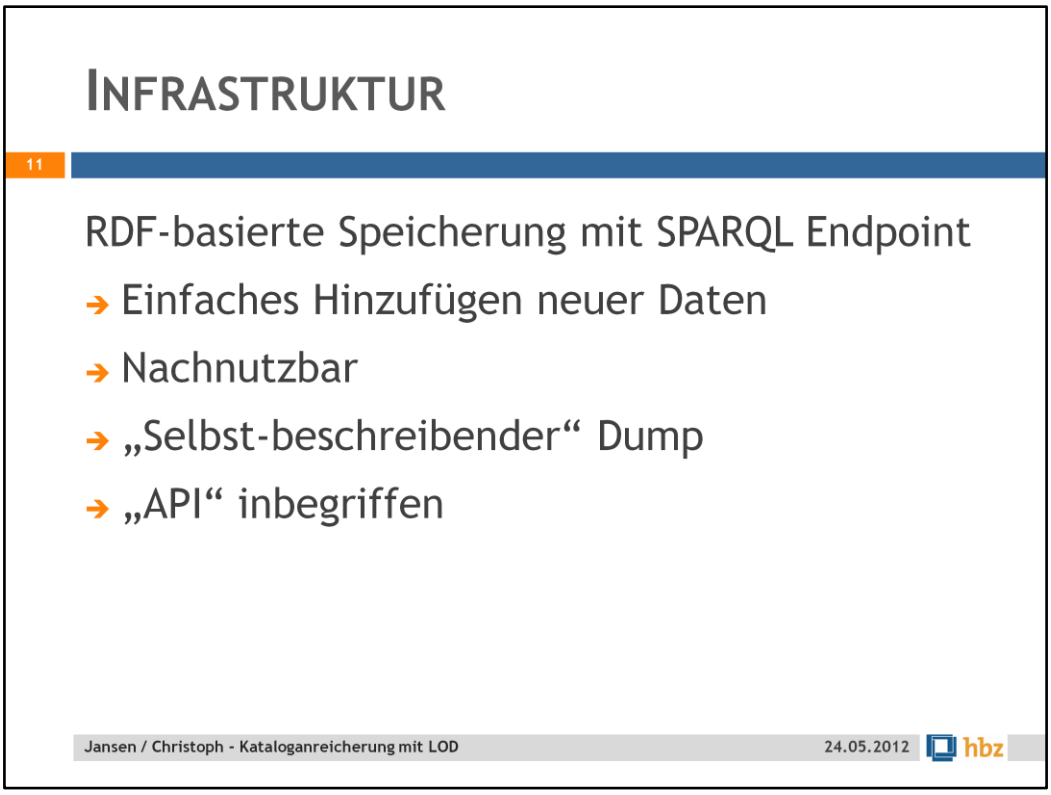
+ Aktuell
+ Relativ geringer Aufwand je Quelle
– Benötigt (performante) API
– Keine Recherche über die Daten

Lokale DB:

+ Aufwendigere Zusammenführungen möglich
+ Daten nutzbar für Recherche und weitere Funktionen
– Hoher kontinuierlicher Aufwand für jede Quelle

Grundidee des Projekts

- Aggregation von LOD aus verteilten Quellen
- Daten für beide Wege verfügbar machen
- Lokalen Aufwand minimieren
  - Homogenisierte Daten zum Import
  - Eine API für viele Bestände



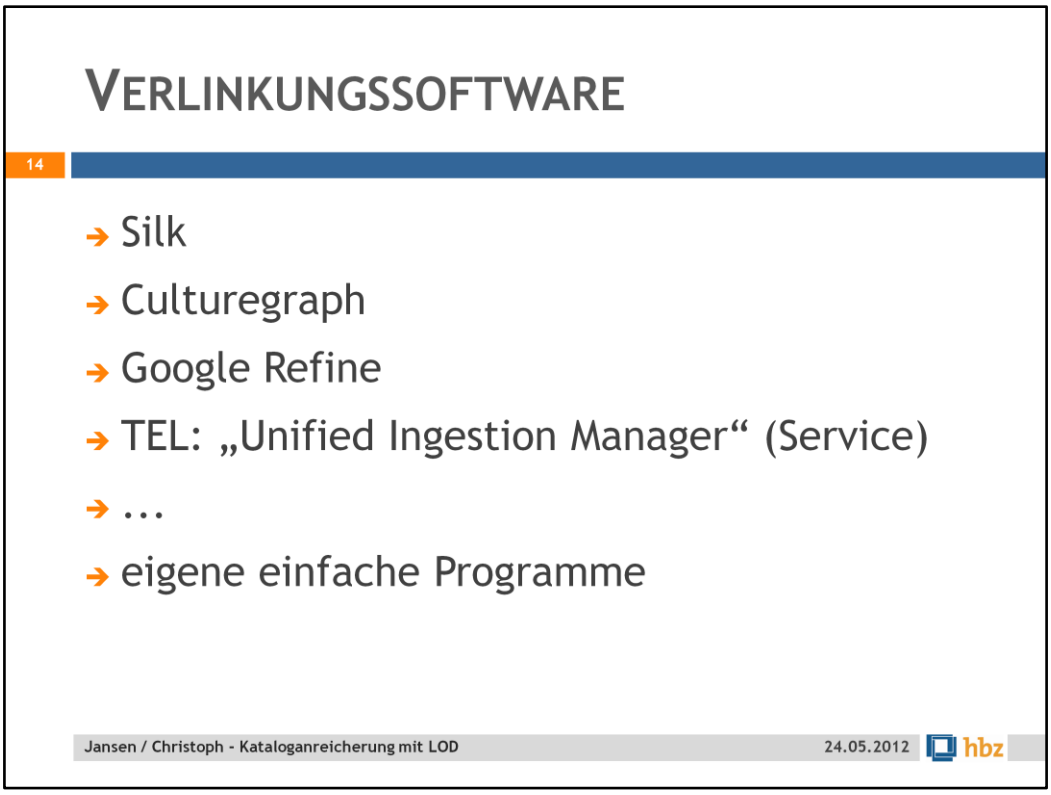
Nur LOD, denn

- sonst verbringt man viel Zeit mit Lizenzprüfungen
- muss man Daten viel aufwändiger Mappen

Praxis

AUSGANGSLAGE - LOBID.ORG

- RDF Store mit SPARQL Endpoint: 4store
- Open Data aus dem Verbundkatalog „hbz01“
- 10 M Titeldatensätze <=> 300 M Triple
- Bereits bestehende Verknüpfungen des Katalogs:
  - 58.000 zu b3kat
  - 103.000 zur Dewey DC
  - 192.000 zu DNB Titeldaten
  - 702.886 zur ZDB
  - 5.500.000 zur LOC Iso-639-2
  - 9.600.000 zur GND Normdatei
  - 18.000.000 zu lobid-organisation



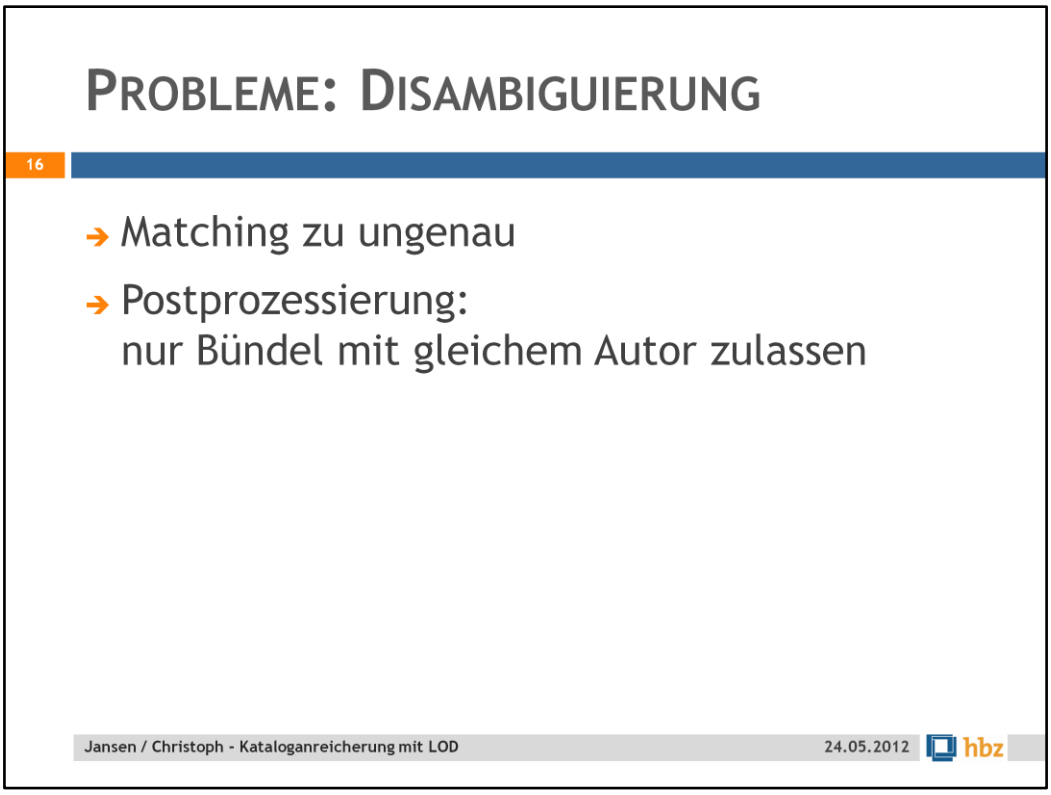
Diese Tools helfen die Verlinkungen zwischen verschiedenen Datenquellen herzustellen.

Vorteil von Silk hadoop:
- Sehr einfache Bedienung/sehr gute Anleitung
- Eingrenzungen via SPARQL Queries
- mit grafischer Benutzeroberfläche „Workbench“
Nachteil von Silk:
- Geht immer nur über SPARQL-Endpoint => keine File-Dump Einspielung.
- Z.B.: Abfragen vom lobid-Endpoint von 9 M Ressourcen dauert 40 Stunden. (Ein einmal erzeugtes Binärfile kann aber durch einfaches kopieren mehrmals benutzt werden um z.B. einmal mit de-dbepdia zu verknüpfen und danach mit der internationalen dbpedia usw.

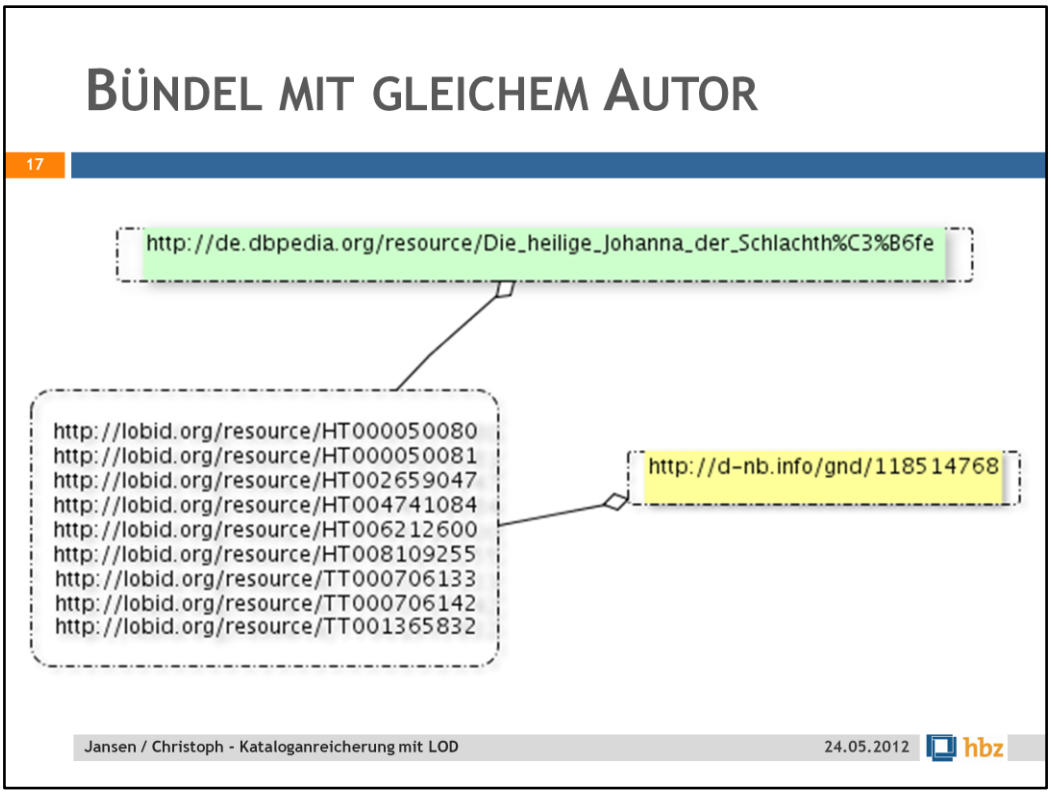
Vorteil Culturegraph hadoop:
- Geht über Filedumps => sehr schnell
- Datenhaltung egal (MARC, MAB, PICA, RDF ...)
Nachteil
- Anleitung noch nicht fertig

MATCHINGALGORITHMEN

- Abhängig von Daten
  - Spannende Daten liegen „woanders“
    => andere Regelwerke
- Beispiel DBpedia:
  - Autor, ISBN usw. fehlt meistens => bleibt Titel
  - Eingrenzungen:
    - deutsche DBpedia
    - category:Literarisches_Werk ,
      category:Lexikon,_Enzyklopädie

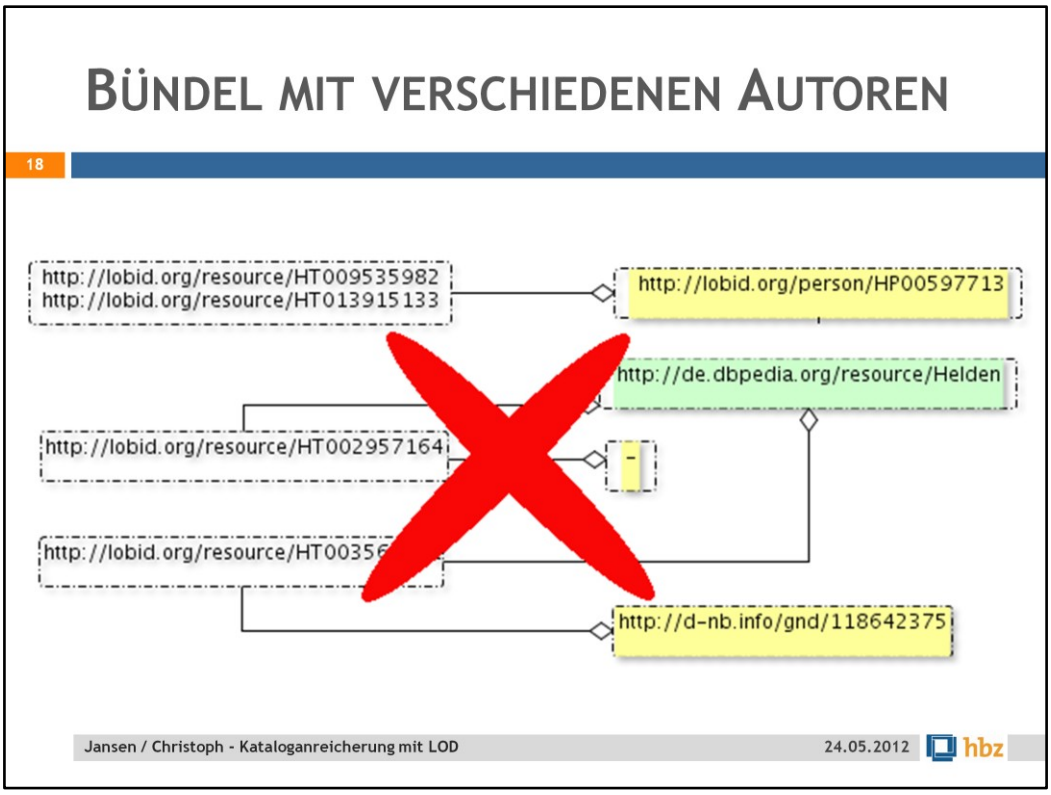


So gibt es z.B. mehrere unterschiedliche Bücher mit dem Titel „Helden“.



Alle Titeldaten im Katalog mit dem Titel „Die heilige Johanna der Schlachthöfe“ haben genau einen Autor.
Bündel in den Ressourcen drin sind die unterschiedliche Autoren haben werden **verworfen!**

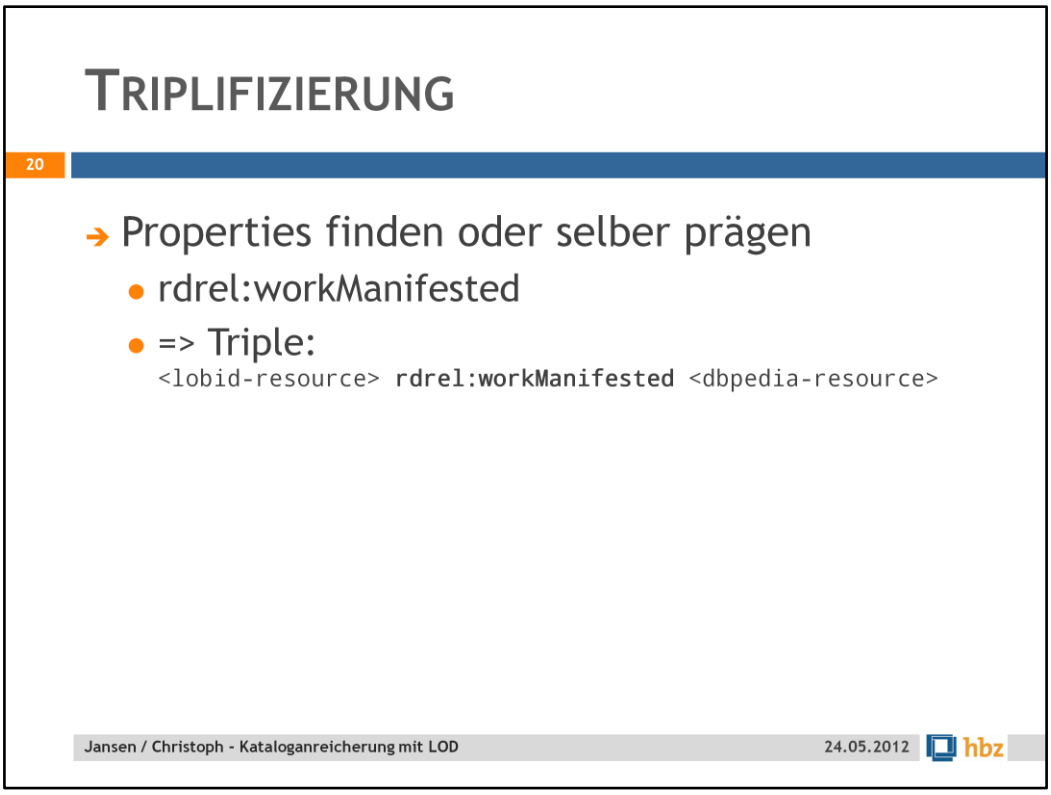
Titel: „Die heilige Johanna der Schlachthöfe“
Autor: http://d-nb.info/gnd/118514768
http://lobid.org/resource/HT000050080
http://lobid.org/resource/HT000050081
http://lobid.org/resource/HT002659047
http://lobid.org/resource/HT004741084
http://lobid.org/resource/HT006212600
http://lobid.org/resource/HT008109255
http://lobid.org/resource/TT000706133
http://lobid.org/resource/TT000706142
http://lobid.org/resource/TT001365832



Alle Titeldaten im Katalog mit dem Titel „Helden“. Die Ressourcen haben verschiedene Autoren und werden deshalb **verworfen!**

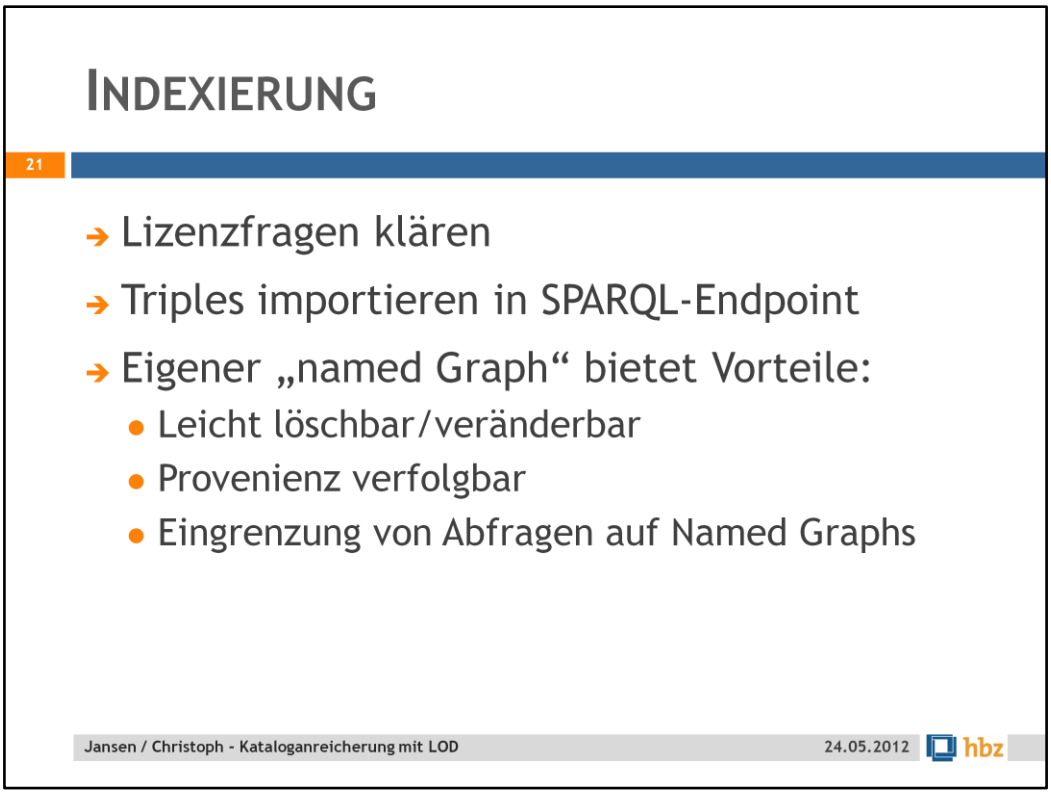
Titel: „Helden“
Autor: http://lobid.org/person/HP00597713
http://lobid.org/resource/HT009535982
http://lobid.org/resource/HT013915133
Autor: -
http://lobid.org/resource/HT002957164
Autor: http://d-nb.info/gnd/118642375
http://lobid.org/resource/HT003564841

„low hanging fruit“ = leicht erreichbare Ziele
Stärkere , komplexere Heuristiken können später verwendet werden



Triplifiziert = die Daten in das Linked Data Schema einpassen. Jedes Datum muss demnach als Triple beschrieben werden, dass heißt durch die Dreifaltigkeit von Subjekt, Prädikat und Objekt.

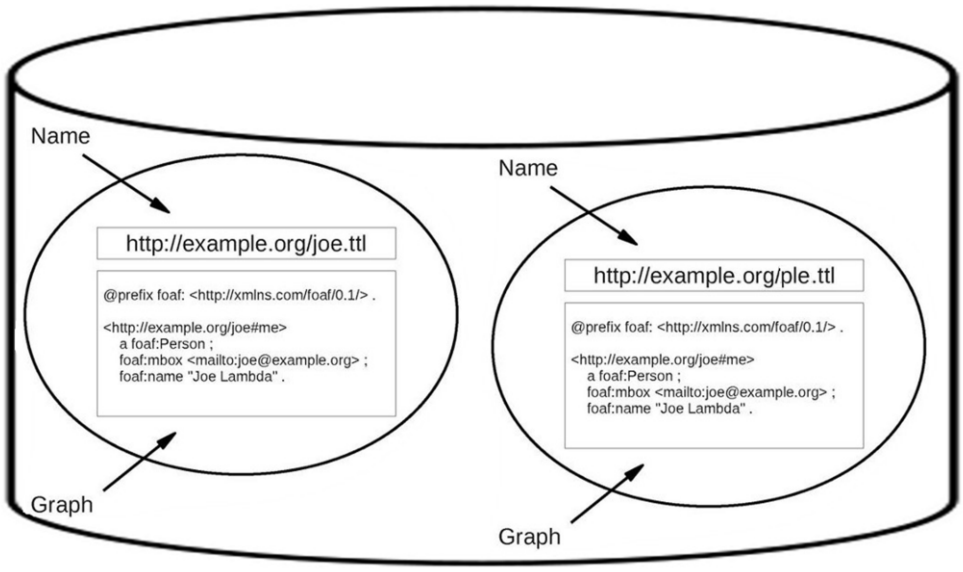
Properties: Wir haben uns dabei nur grob an FRBR orientiert und z.B. die Expressions im WEMI Modell überbrückt. rdrel:workManifested haben wir von der Open Library abgeschaut.

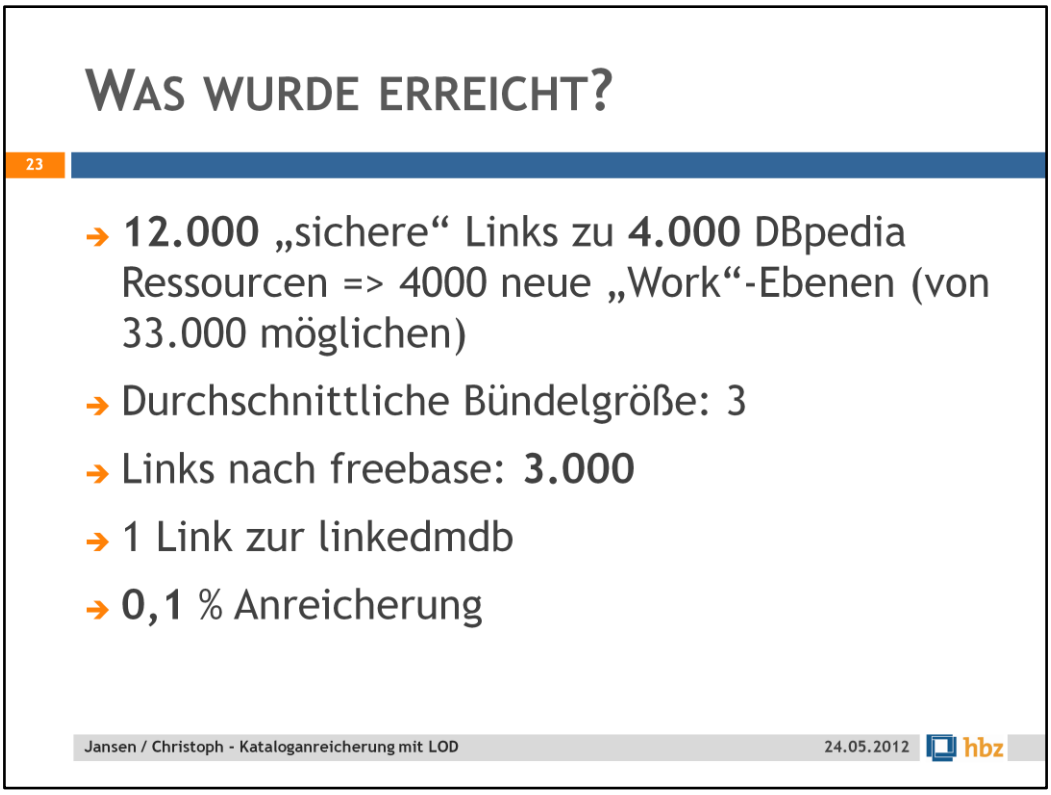


Eigener Graph = „eigener Datenraum“

Bedingt mindestens einen sogenannten Quad-Store (deshalb auch „4“-store). U.a. kann das auch Virtuoso.

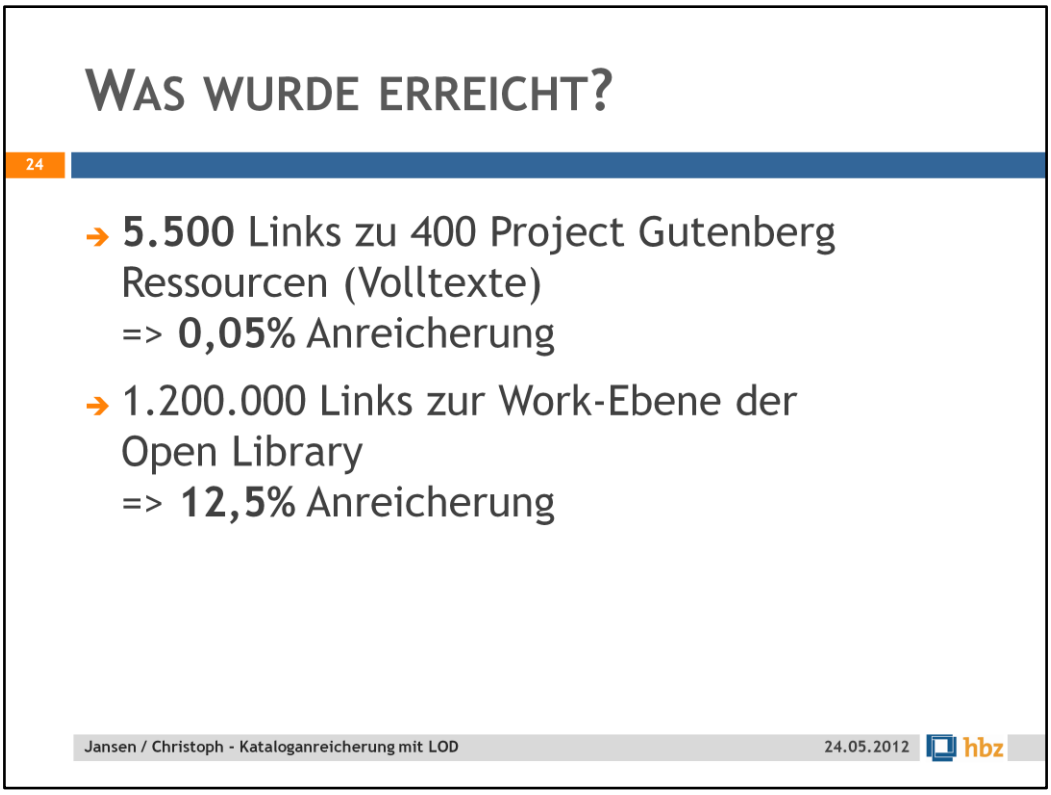
NAMED GRAPHS





„Sichere“ in Anführungszeichen, weil die Links eben nur maschinell hergestellt sind und sich auf Heuristiken beziehen die nicht unbedingt immer Richtiges ergeben müssen. Die Daten müssten also noch intellektuell validiert werden, z.B. durch supervisierten Crowdsourcing.

„Work“ in Anführungszeichen, da frbr:work nur gewisse Daten zulässt und die DBpedia Ressourcen natürlich sehr viel mehr Daten haben. Trotzdem lässt sich von „Work“-Ebene sprechen da alle Manifestationen sich die meisten Eigenschaften dieser DBpedia Ressource teilen (Schlagworte, Kategorien .. aber nicht z.B. Coverbilder).



Project Gutenberg Beispiel: <http://lobid.org/resource/HT001020262> lv:fulltext <http://gutenberg.org/ebooks/10365>

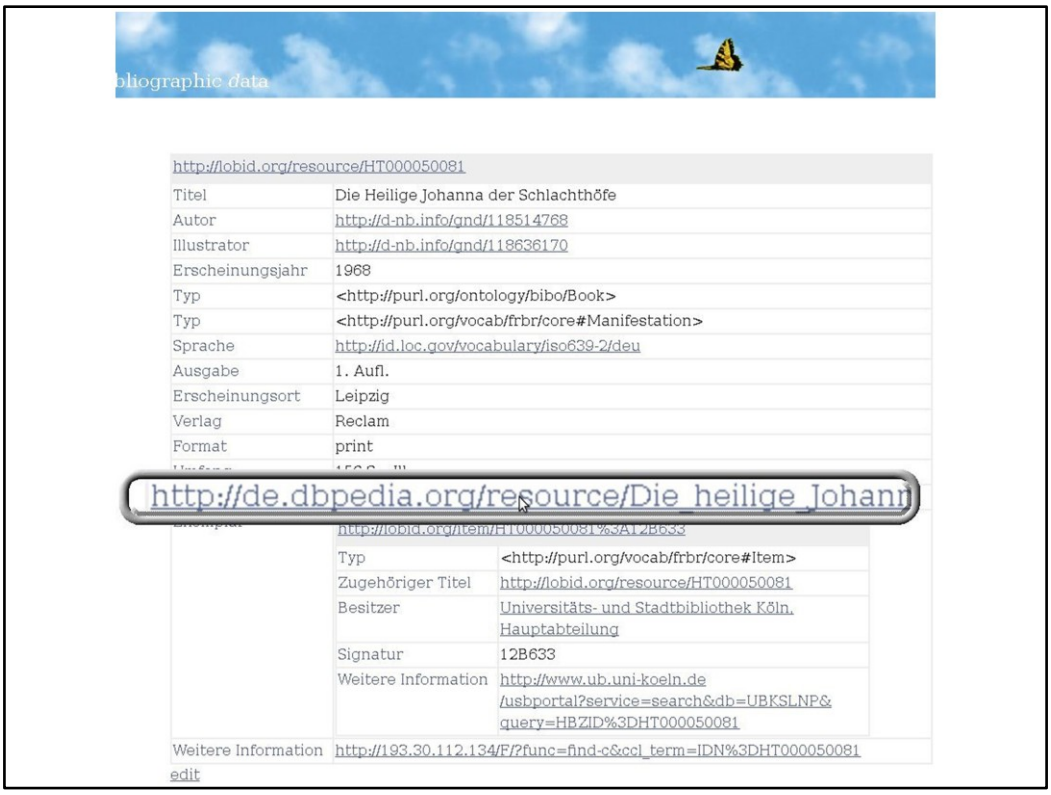
Open Library: die Verknüpfung zur Work-Ebene geschah lediglich auf Grundlage von ISBN 10.

Nochmal: es ging im ersten Schritt darum die „low hanging fruits“ zu ernten = leicht erreichbare Ziele erreichen.
Stärkere, komplexere Heuristiken/Algorithmen können später verwendet werden.

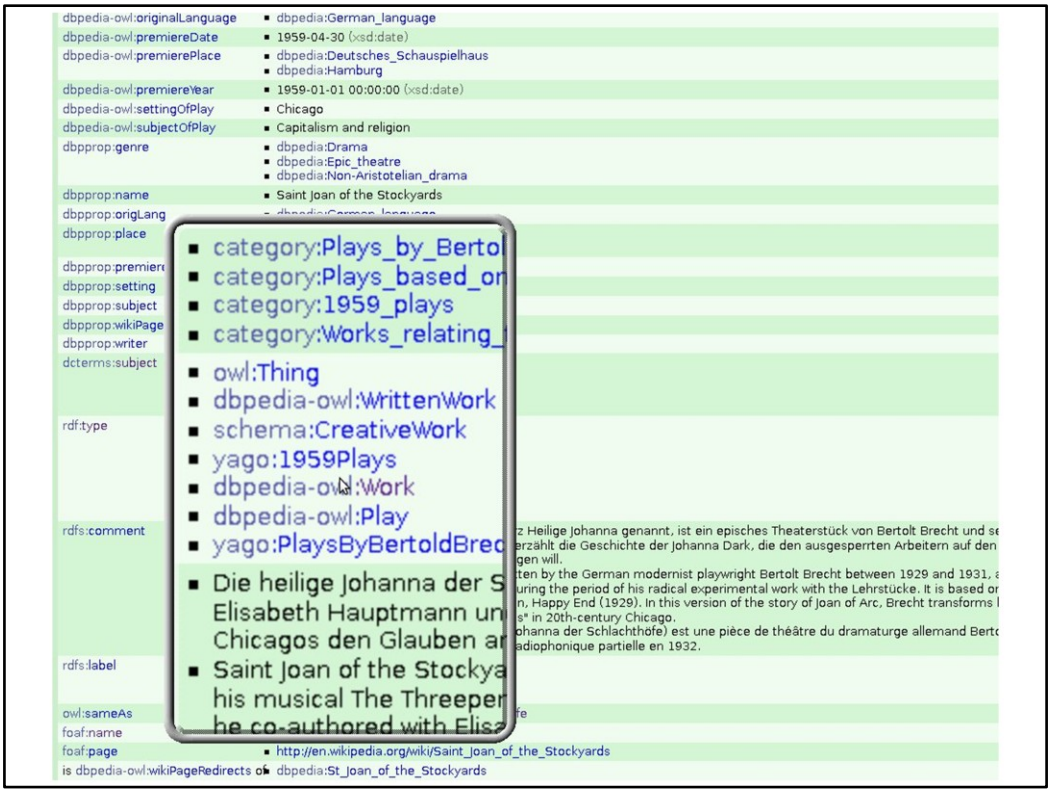
Beispiel

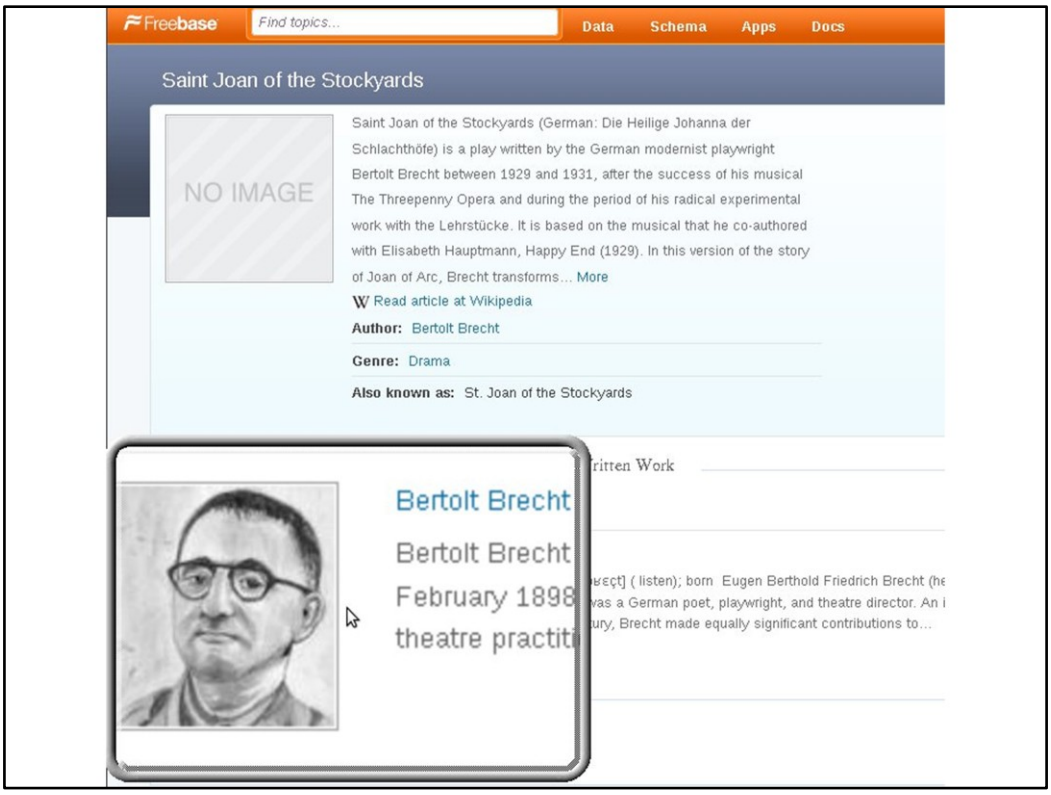
DBpedia:

„Die Heilige Johanna der Schlachthöfe“


bliographic data
http://lobid.org/resource/HT000050081
Titel
Die Heilige Johanna der Schlachthöfe
Autor
http://d-nb.info/gnd/118514768
Illustrator
http://d-nb.info/gnd/118636170
Erscheinungsjahr
1968
Typ
<http://purl.org/ontology/bibo/Book>
Typ
<http://purl.org/vocab/frbr/core#Manifestation>
Sprache
http://id.loc.gov/vocabulary/iso639-2/deu
Ausgabe
1. Aufl.
Erscheinungsort
Leipzig
Verlag
Reclam
Format
print
http://de.dbpedia.org/resource/Die_heilige_Johann
http://lobid.org/item/HT000050081%3A12B633
Typ
<http://purl.org/vocab/frbr/core#Item>
Zugehöriger Titel
http://lobid.org/resource/HT000050081
Besitzer
Universitäts- und Stadtbibliothek Köln, Hauptabteilung
Signatur
12B633
Weitere Information
http://www.ub.uni-koeln.de/usbportal?service=search&db=UBKSLNP&query=HBZID%3DHT000050081
Weitere Information
http://193.30.112.134/F/?func=find-c&ccl_term=IDN%3DHT000050081
edit

- Deutsche DBpedia-Ansicht: http://de.dbpedia.org/resource/Die_heilige_Johanna_der_Schlachth%C3%B6fe
- Kontrolliertes Vokabular => sehr gut geeignet um Facetten in Suchmaschinen zu bilden.


dbpedia-owl:originalLanguage
dbpedia:German_language
dbpedia-owl:premiereDate
1959-04-30 (xsd:date)
dbpedia-owl:premierePlace
dbpedia:Deutsches_Schauspielhaus
dbpedia:Hamburg
dbpedia-owl:premiereYear
1959-01-01 00:00:00 (xsd:date)
dbpedia-owl:settingOfPlay
Chicago
dbpedia-owl:subjectOfPlay
Capitalism and religion
dbpprop:genre
dbpedia:Drama
dbpedia:Epic_theatre
dbpedia:Non-Aristotelian_drama
dbpprop:name
Saint Joan of the Stockyards
dbpprop:origLang
dbpprop:place
dbpprop:premiere
dbpprop:setting
dbpprop:subject
dbpprop:wikiPage
dbpprop:writer
dcterms:subject
category:Plays_by_Bertol
category:Plays_based_or
category:1959_plays
category:Works_relating_
rdf:type
owl:Thing
dbpedia-owl:WrittenWork
schema:CreativeWork
yago:1959Plays
dbpedia-owl:Work
dbpedia-owl:Play
yago:PlaysByBertoldBrec
rdfs:comment
z Heilige Johanna genannt, ist ein episches Theaterstück von Bertolt Brecht und se
erzählt die Geschichte der Johanna Dark, die den ausgesperrten Arbeitern auf den
gen will.
tten by the German modernist playwright Bertolt Brecht between 1929 and 1931, a
uring the period of his radical experimental work with the Lehrstücke. It is based or
n, Happy End (1929). In this version of the story of Joan of Arc, Brecht transforms l
s" in 20th-century Chicago.
ohanna der Schlachthöfe) est une pièce de théâtre du dramaturge allemand Berto
adiophonique partielle en 1932.
Die heilige Johanna der S
Elisabeth Hauptmann un
Chicagos den Glauben ar
rdfs:label
Saint Joan of the Stockya
his musical The Threeper
he co-authored with Elisa
owl:sameAs
fe
foaf:name
foaf:page
http://en.wikipedia.org/wiki/Saint_Joan_of_the_Stockyards
is dbpedia-owl:wikiPageRedirects of
dbpedia:St_Joan_of_the_Stockyards

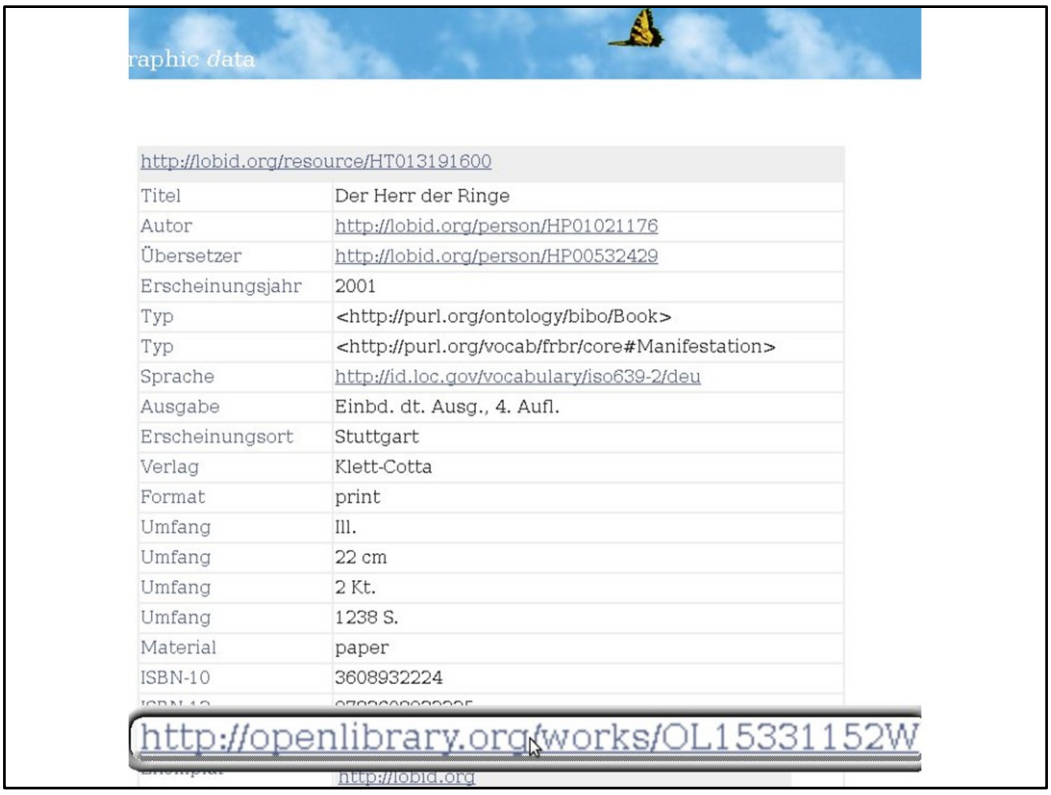


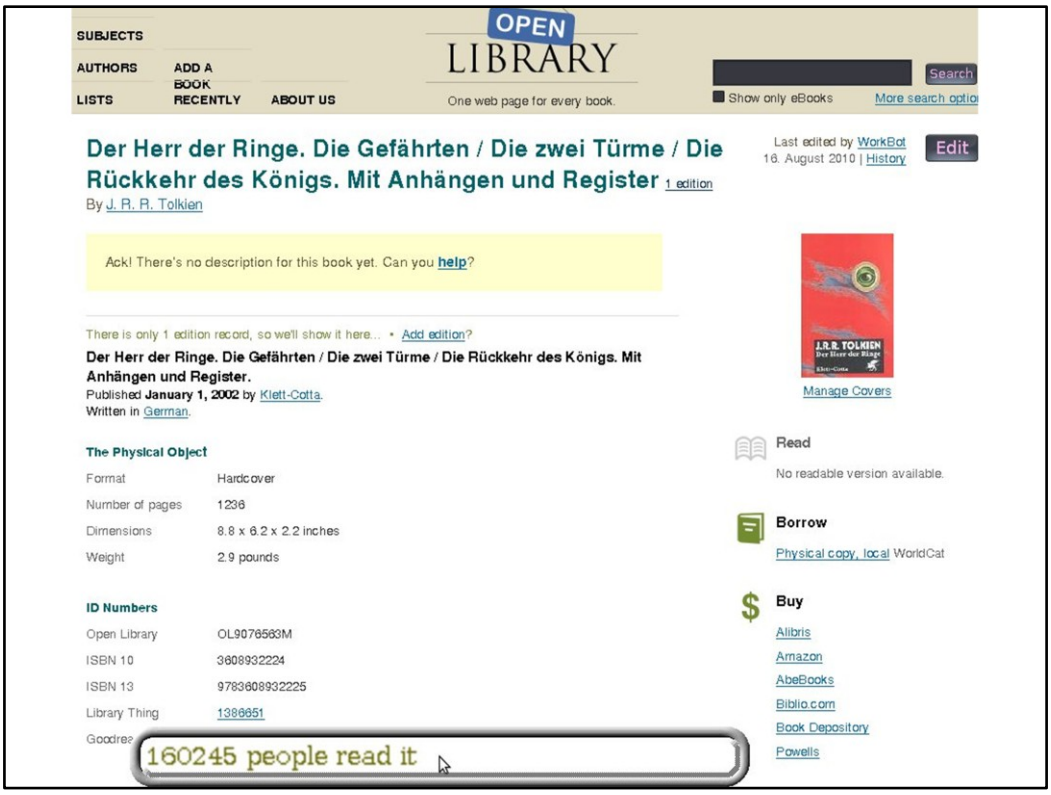
Freebase mit Links u.a. zu IMDB, Rotten Tomatoes => Rezensionen, Userbewertungen uvm., allerdings (noch) nicht als Linked Open Data, d.h. es lassen sich z.B. nicht automatische Suchanfragen stellen, und die dort liegenden Daten lassen sich nicht in die eigene Datenbank integrieren.

BEISPIEL

Open Library

„Der Herr der Ringe“

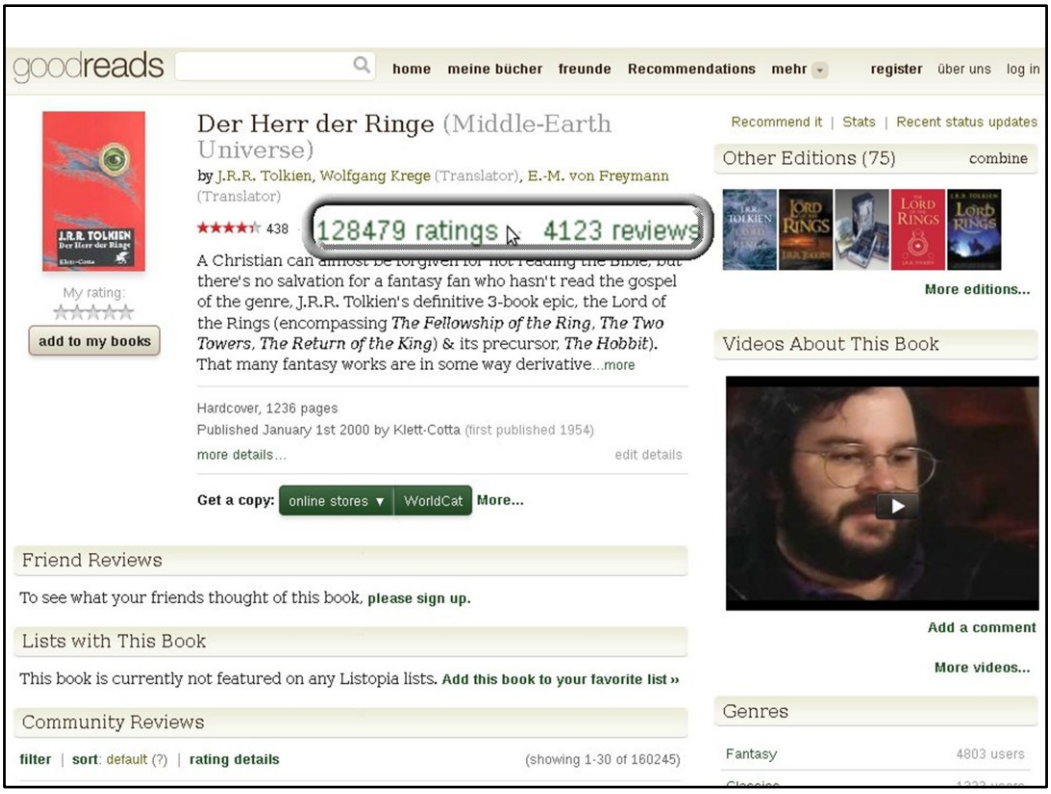

raphic data
http://lobid.org/resource/HT013191600
Titel
Der Herr der Ringe
Autor
http://lobid.org/person/HP01021176
Übersetzer
http://lobid.org/person/HP00532429
Erscheinungsjahr
2001
Typ
<http://purl.org/ontology/bibo/Book>
Typ
<http://purl.org/vocab/frbr/core#Manifestation>
Sprache
http://id.loc.gov/vocabulary/iso639-2/deu
Ausgabe
Einbd. dt. Ausg., 4. Aufl.
Erscheinungsort
Stuttgart
Verlag
Klett-Cotta
Format
print
Umfang
Ill.
Umfang
22 cm
Umfang
2 Kt.
Umfang
1238 S.
Material
paper
ISBN-10
3608932224
http://openlibrary.org/works/OL15331152W
http://lobid.org



Cover und weitere Links in http://openlibrary.org/works/OL15331152W/Der_Herr_der_Ringe._Die_Gef%C3%A4hrten_Die_zwei_T%C3%BCrme_Die_R%C3%BCckkehr_des_K%C3%B6nigs._Mit_Anh%C3%A4ngen_und_Register

Die Daten liegen in RDF vor und gehorchen ebenfalls einer OpenData Lizenz. Sie können also einfach den eigenen Daten zugefügt werden.

Sieht so aus als ob es hier ein Problem mit der Zuordnung von Manifestationen zur Werkebene gibt, denn u.a. fehlen anderssprachige Ausgaben.



Viele ratings und user reviews, ein video uvm:
http://www.goodreads.com/book/show/671212.Der_Herr_der_Ringe

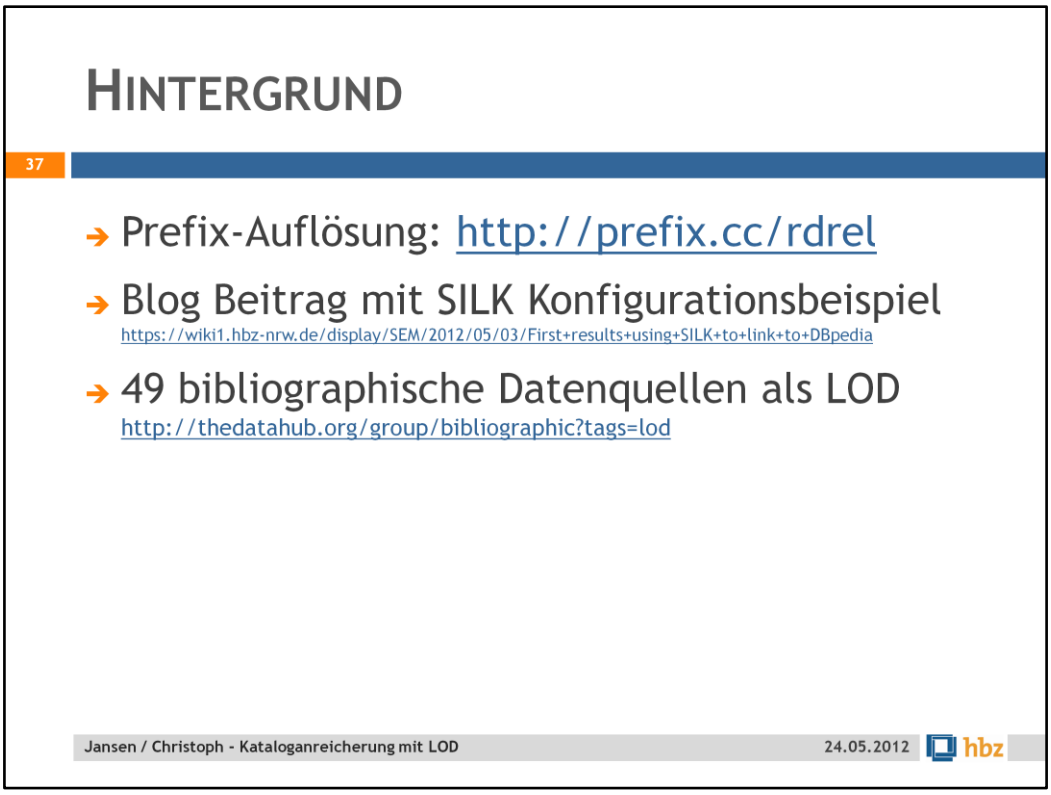
Allerdings liegen die Daten nicht als RDF vor und sind wohl auch nicht Open Data.

Integration in Kataloge

- Was *darf* übernommen werden, was nicht?
- Was *sollte* übernommen werden, was nicht?
- Menschen-lesbare Anzeige der Links/URIs
- (Manche) Daten *sollten* lokal indexiert werden (zur Recherche)
- ...

Warum Linked Open Data ?

- Richtig ist: „Warum *nicht* LOD?“
- Gute Tools wie SILK basieren auf SPAQRL
- Linked Data impliziert **kontrolliertes Vokabular, Identifier**, weitere **Links** … <= ideal!
- Linked Open Data: Daten lassen sich **nachnutzen**/zwischen**speichern/verändern** … <= ideal!



(Folie wurde im Vortrag nicht gezeigt)

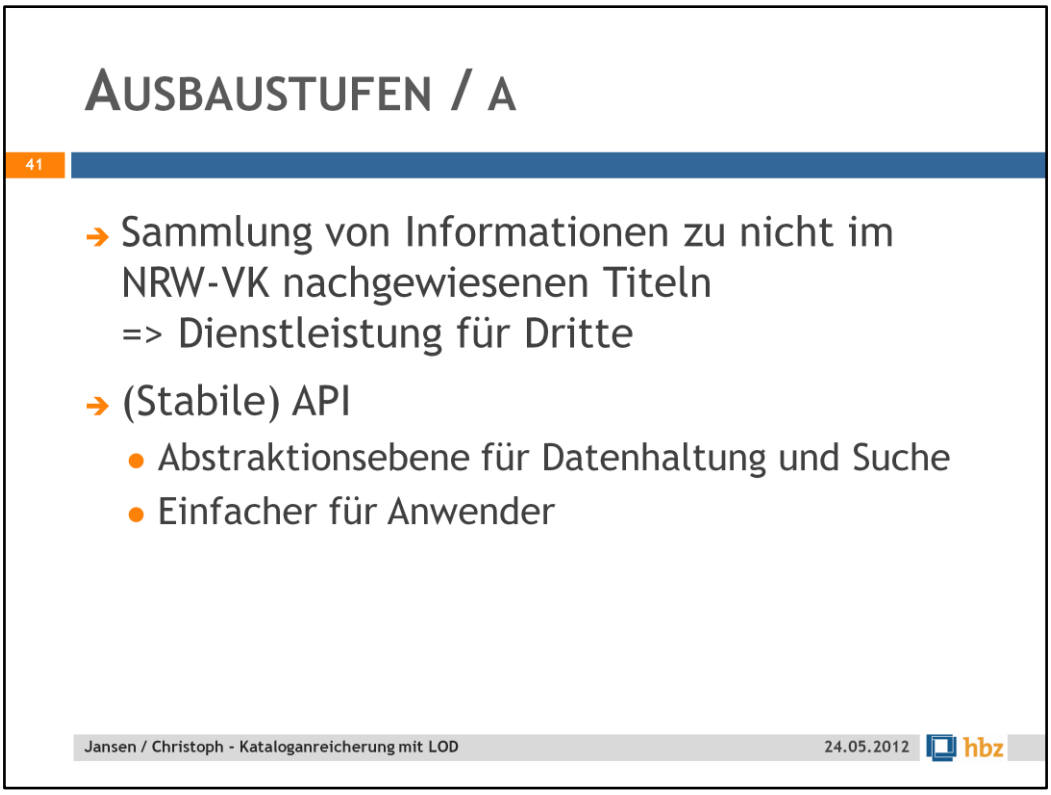
Weiterentwicklung

Jewish Historical Society of the Upper Midwest; http://www.flickr.com/photos/jhsum-commons/4419490136/

Werkstattbericht: „Work in Progress“, kein abgeschlossenes Projekt

POTENTIALE

- Mehr Datenquellen
- Vokabular-Mapping
- Mehr Verknüpfungen
  (FRBR-Werke oder coref oder ...)



Query-Parameter in einer Art „Rest-ful URL“ und JSON sind leichter zu verstehen bzw. bekannter als SPARQL und RDF

Ausbaustufen / b

- ~~Anreicherungs**datenbank**~~ => Anreicherungs**dienst**
- Integration nicht indexierbarer Bestände per dynamischer Abfrage
- Vgl. http://www4.wiwiss.fu-berlin.de/bizer/bookmashup/

AUSBAUSTUFEN / C

- Sammelstelle für kollaborativen Content, z. B. Tagging?
  - Upload von Identifier/Tag Informationen aus verteilten Systemen
  - „Kritische Masse“ gemeinsam erreichbar

Fazit

The Library of Congress, http://www.flickr.com/photos/library_of_congress/4037490394/

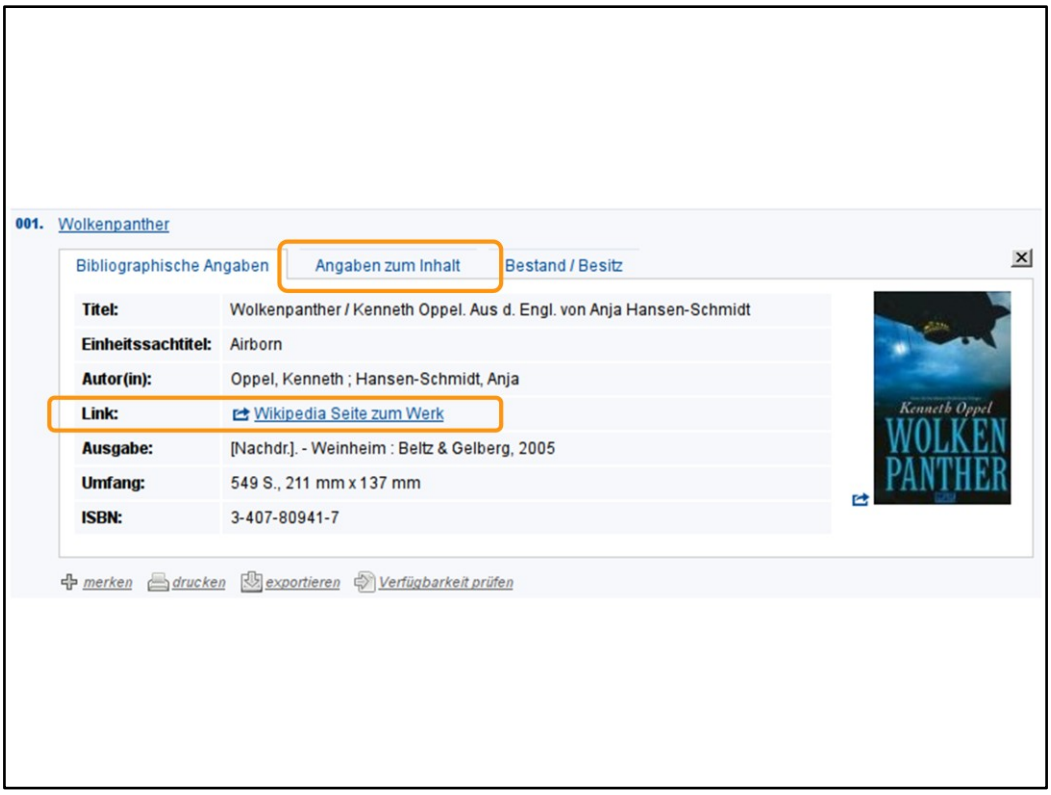
Stehen am Anfang
„Spielerisches“ Ausprobieren der Möglichkeiten
Andere können mitspielen: Was da ist, kann bereits genutzt werden

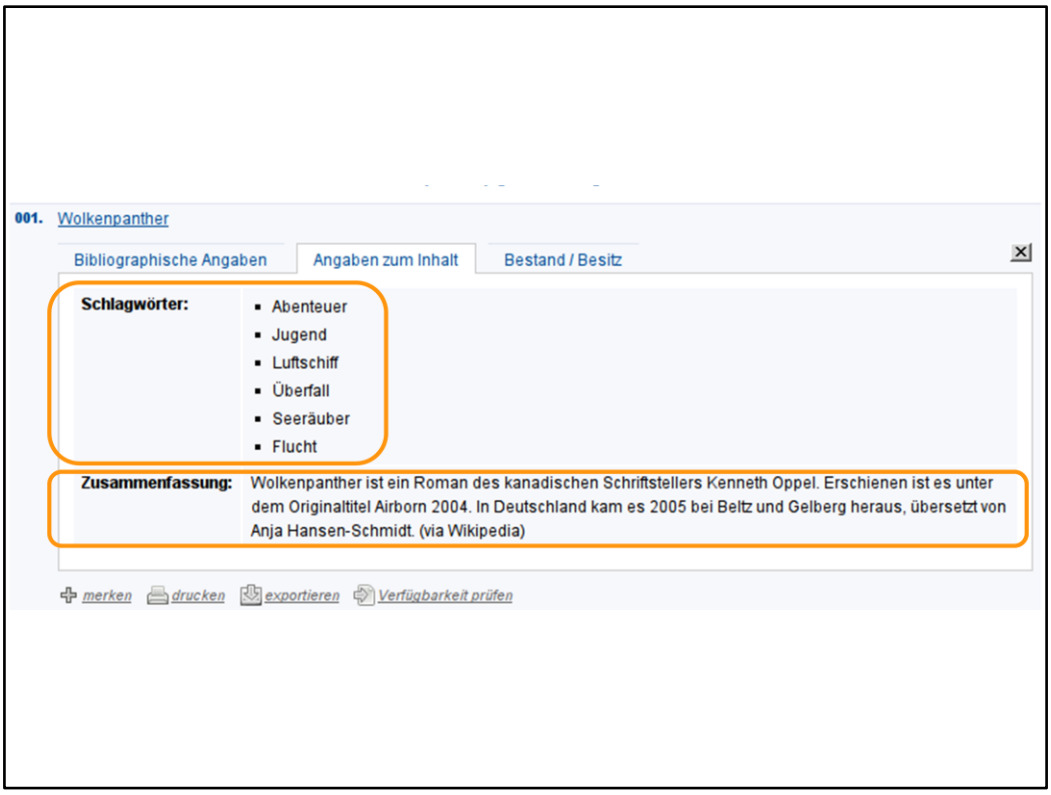
001. Wolkenpanther

Bibliographische Angaben | Bestand / Besitz

Titel:	Wolkenpanther / Kenneth Oppel. Aus d. Engl. von Anja Hansen-Schmidt
Einheitssachtitel:	Airborn
Autor(in):	Oppel, Kenneth ; Hansen-Schmidt, Anja
Ausgabe:	[Nachdr.]. - Weinheim : Beltz & Gelberg, 2005
Umfang:	549 S., 211 mm x 137 mm
ISBN:	3-407-80941-7

merken drucken exportieren Verfügbarkeit prüfen


001. Wolkenpanther
Bibliographische Angaben
Angaben zum Inhalt
Bestand / Besitz
Titel: Wolkenpanther / Kenneth Oppel. Aus d. Engl. von Anja Hansen-Schmidt
Einheitssachtitel: Airborn
Autor(in): Oppel, Kenneth ; Hansen-Schmidt, Anja
Link: Wikipedia Seite zum Werk
Ausgabe: [Nachdr.]. - Weinheim : Beltz & Gelberg, 2005
Umfang: 549 S., 211 mm x 137 mm
ISBN: 3-407-80941-7
Kenneth Oppel
WOLKEN PANTHER
merken
drucken
exportieren
Verfügbarkeit prüfen



Schlagwörter stammen aus dem entsprechenden Datensatz im Verbundkatalog des hbz (Datenbank HBZ01)

i has a question
ICANHASCHEEZBURGER.COM
http://app.cheezburger.com/TemplateView.aspx?ctid=1131803

Vielen Dank!

Rückfragen jederzeit:

jansen@hbz-nrw.de
+49 221 400 75 - 150

christoph@hbz-nrw.de
+49 221 400 75 - 139