Indexierung von Fedora Datensätzen in Apache Solr mit Apache Camel

(Ein Hands-on Tutorial)

Ralf.Claussnitzer@slub-dresden.de
Rainer.Gnan@ub.uni-muenchen.de
Jaime.Penagos@ub.uni-muenchen.de

Organisatorisches

- **Zeitrahmen**
  - Block 1: 14:30 - 16:00 Uhr
  - Pause : 16:00 - 16:30 Uhr
  - Block 2: 16:30 - 18:00 Uhr

- **Inhalt**
  - Generelle Präsentation der Anwendungen
  - Spezielles Setup der Anwendungen

- **Vorstellungsrunde**
  - Name
  - Tätigkeitsbereich
  - Motivation

- **Technische Vorbereitung**
  - Hilfe notwendig?

Agenda

- **Einführung in Fedora**
  - Fedora GUI
  - Fedora HTTP REST API
  - Fedora Events und Messaging
- **Einführung in Apache Solr**
  - Indexierung
  - Index-Schema und Solr-Konfiguration
  - Sucheinstieg
- **Einführung Apache Camel**
  - Apache Camel in der VM
  - Visualisieren und Steuern mit Hawtio
- **Integration von Fedora und Solr mit Apache Camel**
  - Konsumieren von Fedora Events
  - Ansteuern der Camel Routen für Ingest, Update, Delete
  - Extrahierung von Informationen
  - Beschicken von Solr
- **Fedora spezifischer Ausblick**
  - Linked Data Platform
  - Zukünftige Entwicklung
  - Alternative Implementationen

Einführung in Fedora

- Was ist Fedora?

*Quelle: “Fedora Commons Repository” http://docs.fcrepo.org/

Einführung in Fedora

- Warum Fedora?
  - Standard
    - LDP (Linked Data Platform)
    - RDF (Resource Description Framework)
  - Modular
    - Integration externer Komponenten / Systeme*

Einführung in Fedora

- Grundbegriffe
  - URI: Uniform Resource Identifier
  - RDF: Resource Description Framework
  - LDP: Linked Data Platform
    - LDPR: LDP Resources
    - LDPC: LDP Containers

- Das wichtigste in kürze
  - https://wiki.duraspace.org/display/FEDORA50/Fedora+5.0+Documentation
  - https://www.w3.org/Addressing/URL/uri-spec.html
  - https://www.w3.org/RDF/
  - https://www.w3.org/TR/ldp/
  - https://www.w3.org/TR/activitystreams-core/

Einführung in Fedora

- Web Resource
  - Alles ist ein Web Resource mit URI (Uniform Resource Identifier)
  - RDF Triples (Subjekt-Prädikat-Objekt)
  - Web Resource
    - Container
    - Binary

Einführung in Fedora

- Web Resource (Beispiel*)
  - Alles in Fedora ist eine Web Resource mit URI

    Berlin: <http://dbpedia.org/page/Berlin>

About: Berlin

An Entity of Type : Stadt, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Berlin Zum Anhören bitte klicken! [bɛɐ̯ˈliːn] ist die Bundeshauptstadt der Bundesrepublik Deutschland und zugleich eines ihrer Länder. Die Stadt Berlin ist mit gut 3,5 Millionen Einwohnern die bevölkerungsreichste und mit 892 Quadratkilometern die flächengrößte Gemeinde Deutschlands. Sie bildet das Zentrum der Metropolregion Berlin/Brandenburg (6 Millionen Einw.) und der Agglomeration Berlin (4,4 Millionen Einw.). Der Stadtstaat unterteilt sich in zwölf Bezirke. Neben den Flüssen Spree und Havel befinden sich im Stadtgebiet kleinere Fließgewässer sowie zahlreiche Seen und Wälder.

*Quelle: “DBpedia” https://wiki.dbpedia.org/OnlineAccess

Einführung in Fedora

- Web Resource (Beispiel*)
  - RDF Triples (Subjekt-Prädikat-Objekt)

Subjekt
<http://dbpedia.org/resource/Berlin>

Prädikat
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

Objekt
<http://dbpedia.org/ontology/City>

RDF:

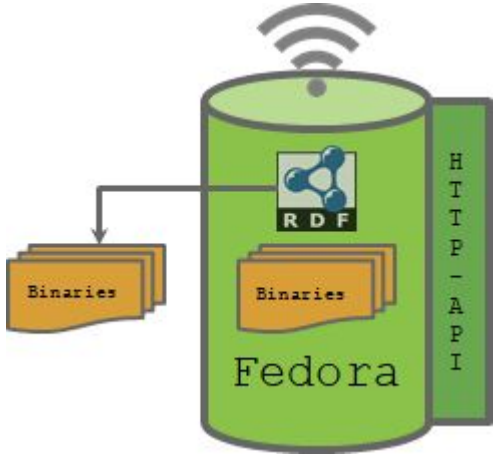
<http://dbpedia.org/resource/Berlin> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/City> .

*Quelle: “DBpedia” https://wiki.dbpedia.org/OnlineAccess

Einführung in Fedora

Web Resource (Beispiel*)

- Container
- Binary

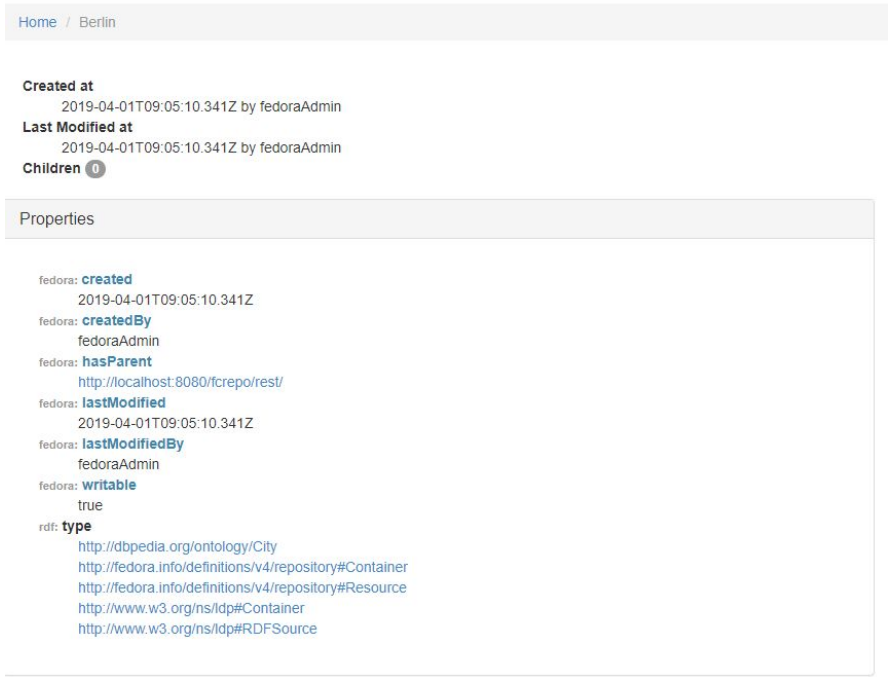


*Quelle: “Introduction to Fedora” https://wiki.duraspace.org/display/FF/2018-11+Fedora+and+Samvera+Camp+Berlin

Einführung in Fedora

- Fedora GUI

  http://localhost:8080/fcrepo/rest/Berlin



Einführung in Fedora

- Fedora REST API

```
curl -X GET http://localhost:8080/fcrepo/rest/Berlin -u fedoraAdmin:secret3 -H "ld+triples"
```

```
@prefix premis:  <http://www.loc.gov/premis/rdf/v1#> .
@prefix test:  <info:fedora/test/> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ns001:  <http://dbpedia.org/ontology/> .
@prefix xsi:  <http://www.w3.org/2001/XMLSchema-instance> .
@prefix xmlns:  <http://www.w3.org/2000/xmlns/> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix fedora:  <http://fedora.info/definitions/v4/repository#> .
@prefix xml:  <http://www.w3.org/XML/1998/namespace> .
@prefix ebucore:  <http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#> .
@prefix ldp:  <http://www.w3.org/ns/ldp#> .
@prefix xs:  <http://www.w3.org/2001/XMLSchema> .
@prefix fedoraconfig:  <http://fedora.info/definitions/v4/config#> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix authz:  <http://fedora.info/definitions/v4/authorization#> .
@prefix dc:  <http://purl.org/dc/elements/1.1/> .

<http://localhost:8080/fcrepo/rest/Berlin>
        rdf:type                ns001:City ;
        rdf:type                fedora:Container ;
        rdf:type                fedora:Resource ;
        fedora:lastModifiedBy   "fedoraAdmin"^^<http://www.w3.org/2001/XMLSchema#string> ;
        fedora:createdBy        "fedoraAdmin"^^<http://www.w3.org/2001/XMLSchema#string> ;
        fedora:created          "2019-04-01T09:05:10.341Z"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
        fedora:lastModified     "2019-04-01T09:05:10.341Z"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
        rdf:type                ldp:RDFSource ;
        rdf:type                ldp:Container ;
        fedora:writable         "true"^^<http://www.w3.org/2001/XMLSchema#boolean> ;
        fedora:hasParent        <http://localhost:8080/fcrepo/rest/> .
```

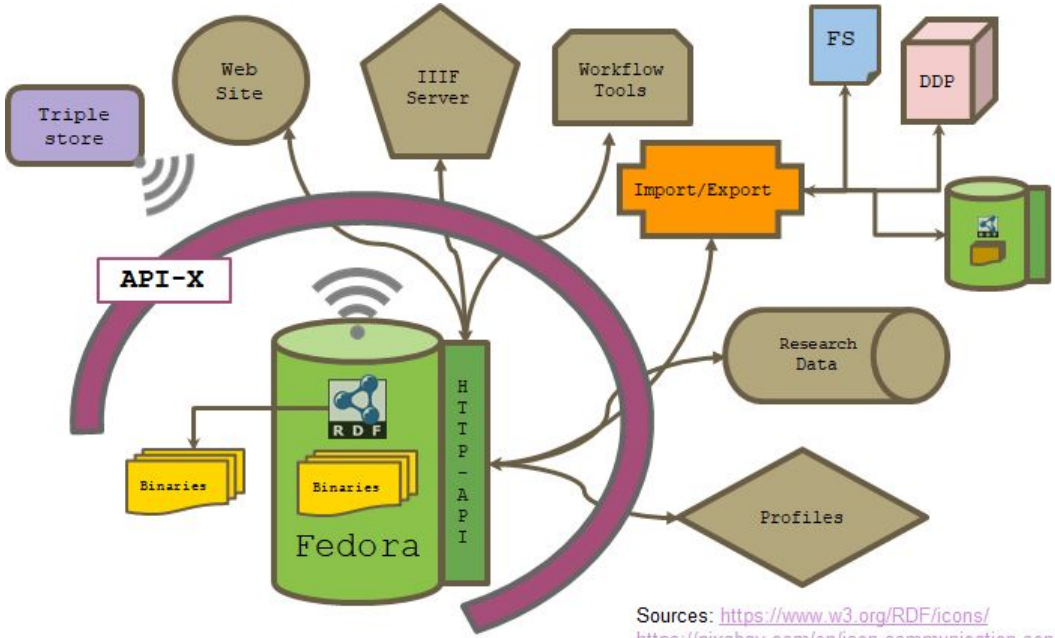
Einführung in Fedora

Core Services - Standards

- Create/Read/Update/Delete - Linked Data Platform
- Versionierung - Memento
- Zugriffsberechtigung - WebAC
- Fixity - Digest
- Messaging - Activity Streams 2.0

Einführung in Fedora

Integration externer Systeme / Komponenten

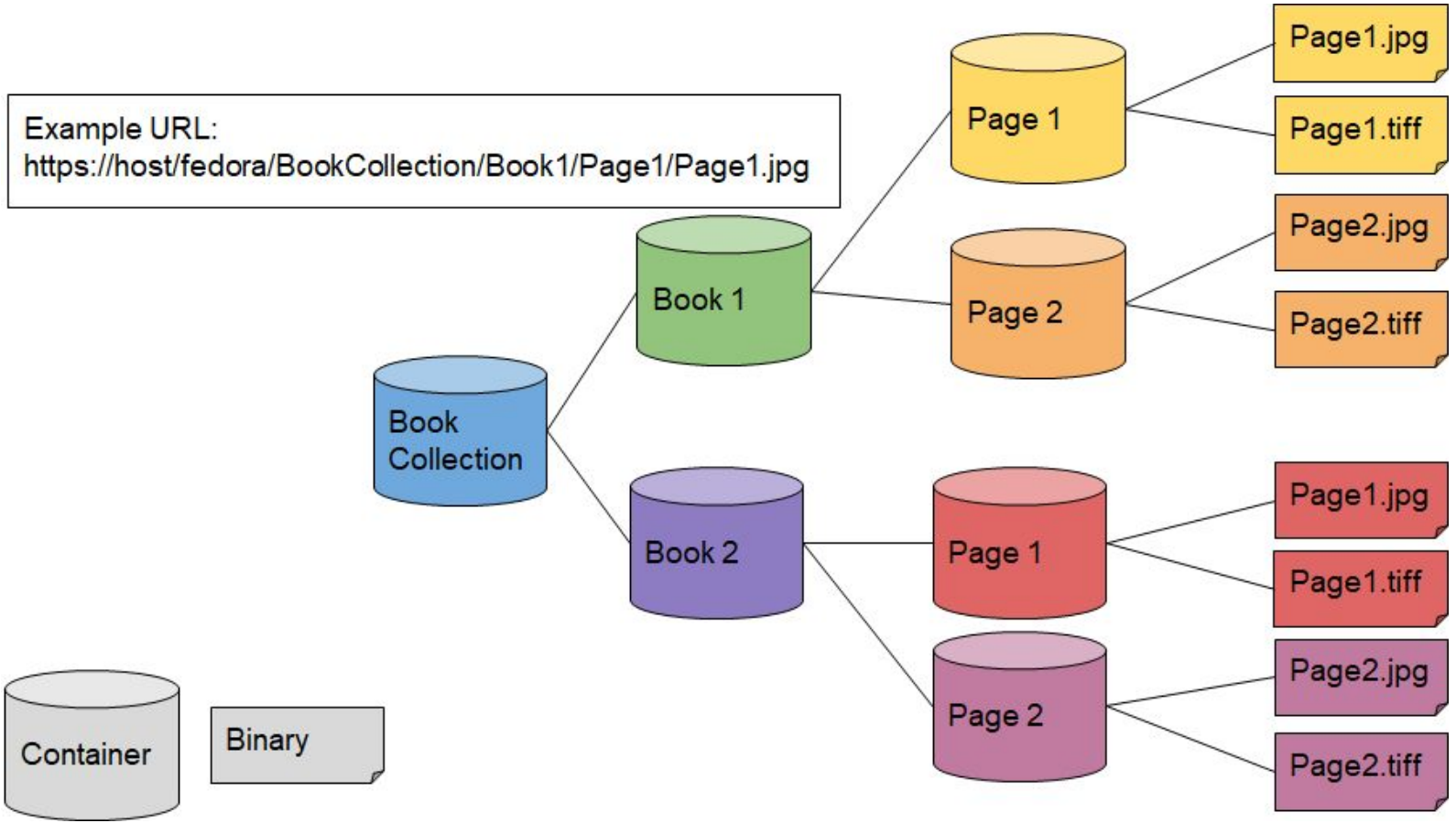


Sources: https://www.w3.org/RDF/icons/
https://pixabay.com/en/icon-communication-sender-antenna-157359/

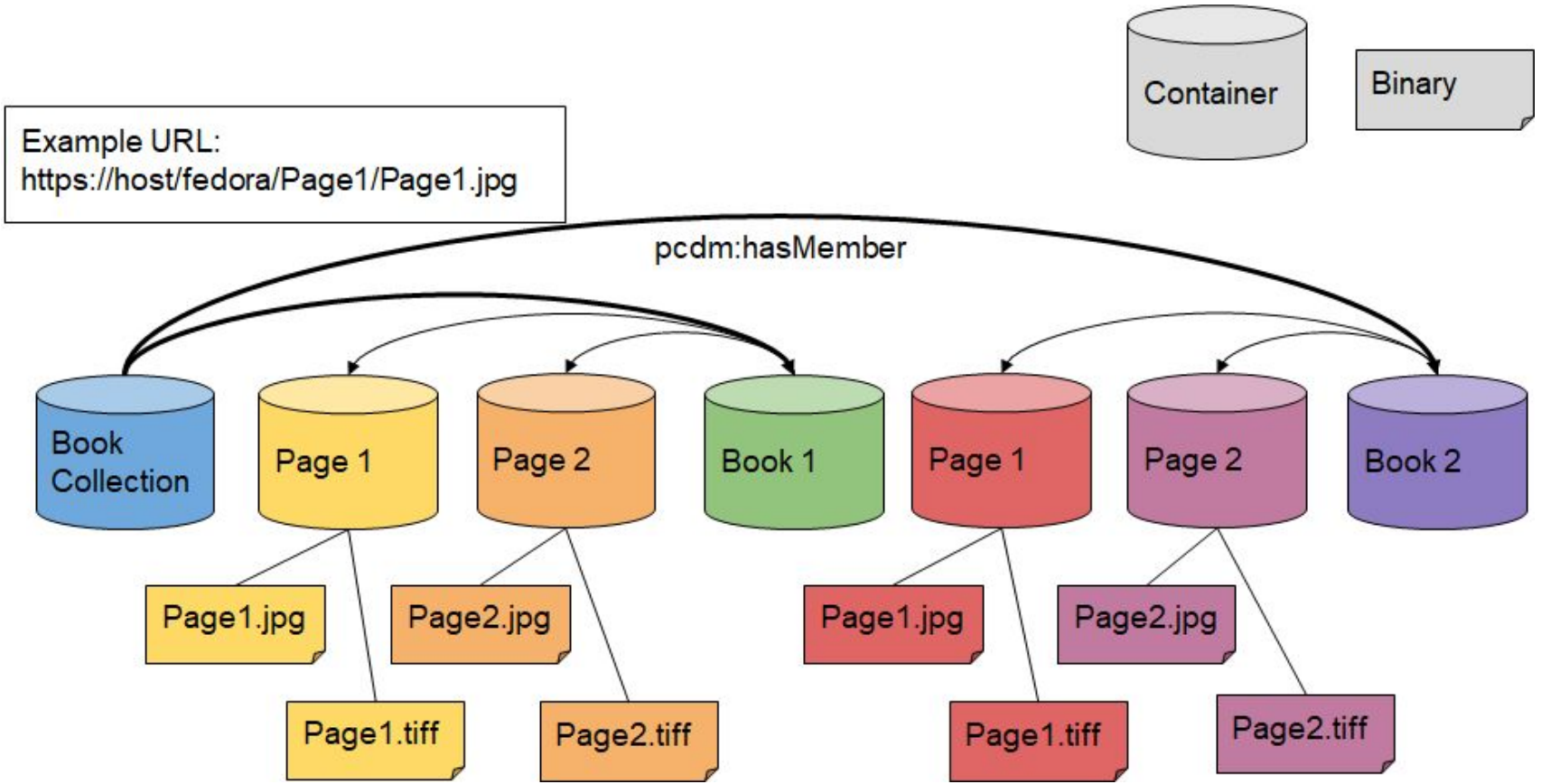
*Quelle: “Fedora and Samvera Overview” https://wiki.duraspace.org/display/FF/2018-11+Fedora+and+Samvera+Camp+Berlin

Hands-on Fedora

- Objekte anlegen
  - Container
  - Binary
- Abbildung eines Models



*Quelle: “Introduction to Fedora” https://wiki.duraspace.org/display/FF/2018-11+Fedora+and+Samvera+Camp+Berlin



*Quelle: “Introduction to Fedora” https://wiki.duraspace.org/display/FF/2018-11+Fedora+and+Samvera+Camp+Berlin

Hands-on Fedora

- Fedora Events

“A mechanism for notifying interested parties

that something has happened”

*Quelle: “Messaging and Versioning” https://wiki.duraspace.org/display/FF/2018-11+Fedora+and+Samvera+Camp+Berlin

Hands-on Fedora

- Fedora Events


```
expires = 0
org.fcrepo.jms.identifier = /Berlin/TestContainer01
org.fcrepo.jms.user = fedoraAdmin
org.fcrepo.jms.resourceType =
http://www.w3.org/ns/ldp#Container,http://fedora.info/definitions/v4/repository#Resource,http://fedora.info/definitions/v4/repository#Conta
iner,http://www.w3.org/ns/ldp#RDFSource
destination = /topic/fedora
ack = ID:fedora4-44117-1554107436039-6:1
org.fcrepo.jms.eventType =
http://fedora.info/definitions/v4/event#ResourceModification,http://fedora.info/definitions/v4/event#ResourceCreation
subscription = 1
priority = 4
org.fcrepo.jms.baseURL = http://localhost:8080/fcrepo/rest
org.fcrepo.jms.eventID = urn:uuid:9a7b6fb7-1aee-4472-8155-788707c5507b
org.fcrepo.jms.timestamp = 1554112288524
message-id = ID:fedora4-44117-1554107436039-4:1:1:1:3
persistent = true
org.fcrepo.jms.userAgent = Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/73.0.3683.86 Safari/
537.36
timestamp = 1554112288548
{"id":"http://localhost:8080/fcrepo/rest/Berlin/TestContainer01","type":["http://www.w3.org/ns/ldp#Container",
"http://fedora.info/definitions/v4/repository#Resource","http://fedora.info/definitions/v4/repository#Container",
"http://www.w3.org/ns/ldp#RDFSource","http://www.w3.org/ns/prov#Entity"],"isPartOf":"http://localhost:8080/fcrepo/rest","wasGeneratedBy":{
"type":["http://fedora.info/definitions/v4/event#ResourceModification","http://fedora.info/definitions/v4/event#ResourceCreation",
"http://www.w3.org/ns/prov#Activity"],"identifier":"urn:uuid:9a7b6fb7-1aee-4472-8155-788707c5507b","atTime":"2019-04-01T09:51:28.524Z"},
"wasAttributedTo":[{"type":"http://www.w3.org/ns/prov#Person","name":"fedoraAdmin"},{"type":"http://www.w3.org/ns/prov#SoftwareAgent",
"name":"Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/73.0.3683.86 Safari/537.36"}],"@context":{
"prov":"http://www.w3.org/ns/prov#","foaf":"http://xmlns.com/foaf/0.1/","dcterms":"http://purl.org/dc/terms/","xsd":
"http://www.w3.org/2001/XMLSchema#","type":"@type","id":"@id","name":{"@id":"foaf:name","@type":"xsd:string"},"identifier":{"@id":
"dcterms:identifier","@type":"@id"},"isPartOf":{"@id":"dcterms:isPartOf","@type":"@id"},"atTime":{"@id":"prov:atTime","@type":
"xsd:dateTime"},"wasAttributedTo":{"@id":"prov:wasAttributedTo","@type":"@id"},"wasGeneratedBy":{"@id":"prov:wasGeneratedBy","@type":"@id"
}}}
```

Hands-on Fedora

- Fedora Events


```
{
    "id": "http://localhost:8080/fcrepo/rest/Berlin/TestContainer01",
    "type": [
        "http://www.w3.org/ns/ldp#Container",
        "http://fedora.info/definitions/v4/repository#Resource",
        "http://fedora.info/definitions/v4/repository#Container",
        "http://www.w3.org/ns/ldp#RDFSource",
        "http://www.w3.org/ns/prov#Entity"
    ],
    "isPartOf": "http://localhost:8080/fcrepo/rest",
    "wasGeneratedBy": {
        "type": [
            "http://fedora.info/definitions/v4/event#ResourceModification",
            "http://fedora.info/definitions/v4/event#ResourceCreation",
            "http://www.w3.org/ns/prov#Activity"
        ],
        "identifier": "urn:uuid:9a7b6fb7-1aee-4472-8155-788707c5507b",
        "atTime": "2019-04-01T09:51:28.524Z"
    },
    "wasAttributedTo": [
        {
            "type": "http://www.w3.org/ns/prov#Person",
            "name": "fedoraAdmin"
        },
        {
            "type": "http://www.w3.org/ns/prov#SoftwareAgent",
            "name": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/73.0.3683.86 Safari/537.36"
        }
    ],
    "@context": {
        "prov": "http://www.w3.org/ns/prov#",
        "foaf": "http://xmlns.com/foaf/0.1/",
        "dcterms": "http://purl.org/dc/terms/",
        "xsd": "http://www.w3.org/2001/XMLSchema#",
        "type": "@type",
        "id": "@id",
        "name": {
            "@id": "foaf:name",
            "@type": "xsd:string"
        },
        "identifier": {
            "@id": "dcterms:identifier",
            "@type": "@id"
        },
        "isPartOf": {
            "@id": "dcterms:isPartOf",
            "@type": "@id"
        },
        "atTime": {
            "@id": "prov:atTime",
            "@type": "xsd:dateTime"
        },
        "wasAttributedTo": {
            "@id": "prov:wasAttributedTo",
            "@type": "@id"
        },
        "wasGeneratedBy": {
            "@id": "prov:wasGeneratedBy",
            "@type": "@id"
        }
    }
}
```

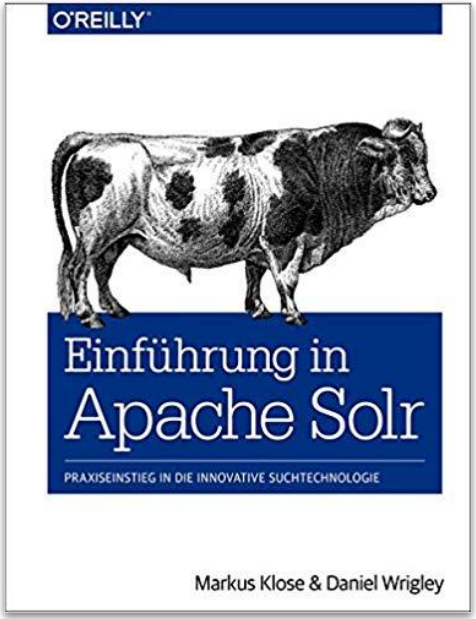
Apache Solr

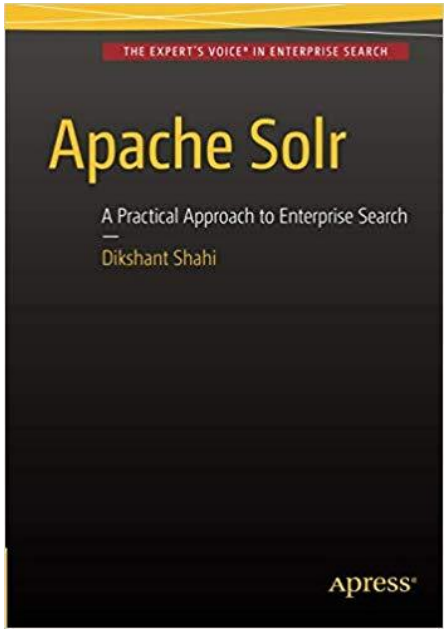
Apache Solr is a fast, scalable and easy to use open search engine built on top of the ever popular Apache Lucene Java library.

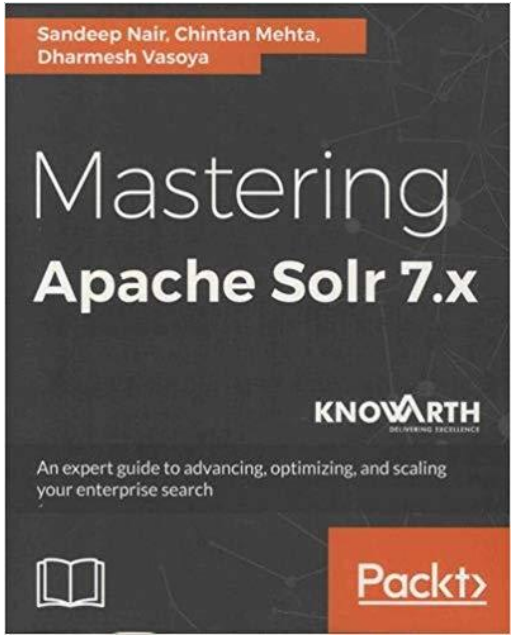
Solr and Lucene have been developing at a dizzying pace, rising to meet the challenges of modern big data applications with hundreds of billions of documents and tens of thousands of queries per second.

-- Einführung in Apache Solr, 1. Auflage 2014

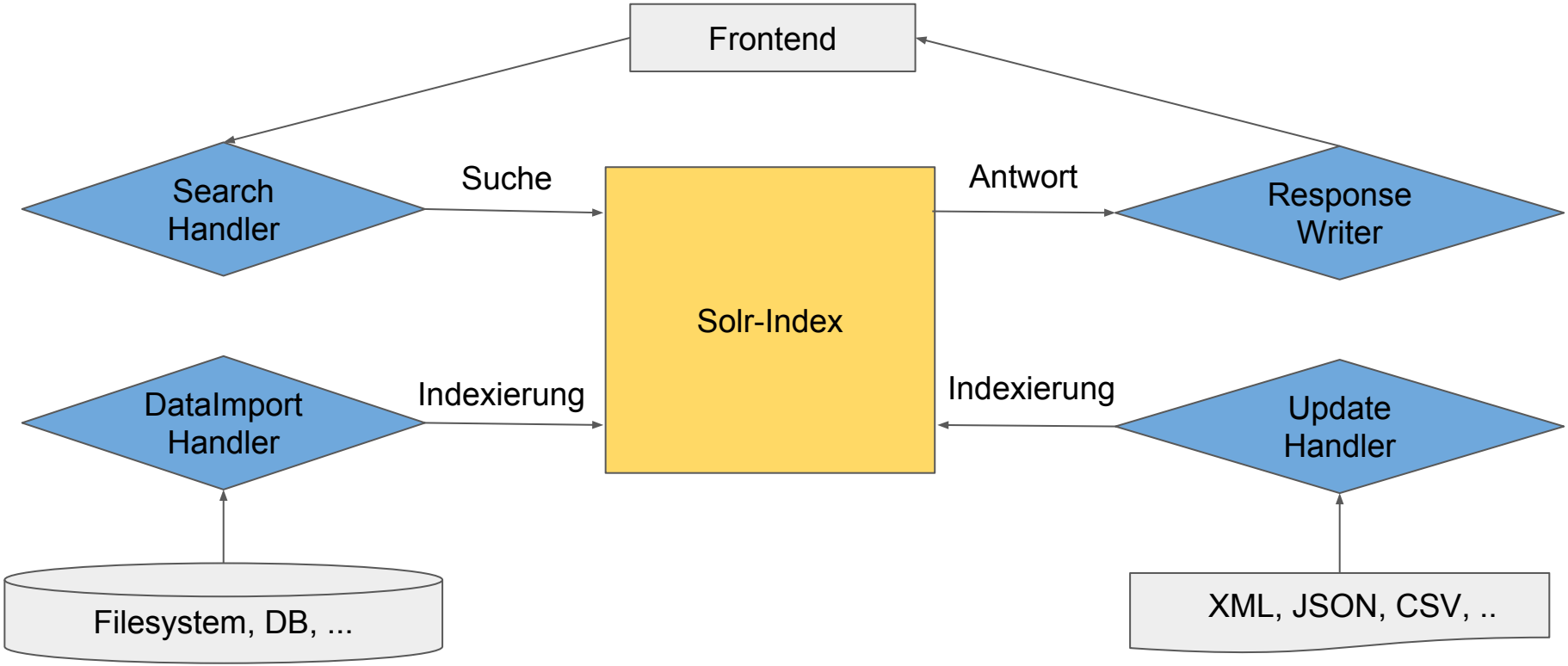
Apache Solr


O'REILLY
Einführung in
Apache Solr
PRAXISEINSTIEG IN DIE INNOVATIVE SUCHTECHNOLOGIE
Markus Klose & Daniel Wrigley


THE EXPERT'S VOICE® IN ENTERPRISE SEARCH
Apache Solr
A Practical Approach to Enterprise Search
Dikshant Shahi
Apress®


Sandeep Nair, Chintan Mehta,
Dharmesh Vasoya
Mastering
Apache Solr 7.x
KNOWARTH
An expert guide to advancing, optimizing, and scaling
your enterprise search
Packt›

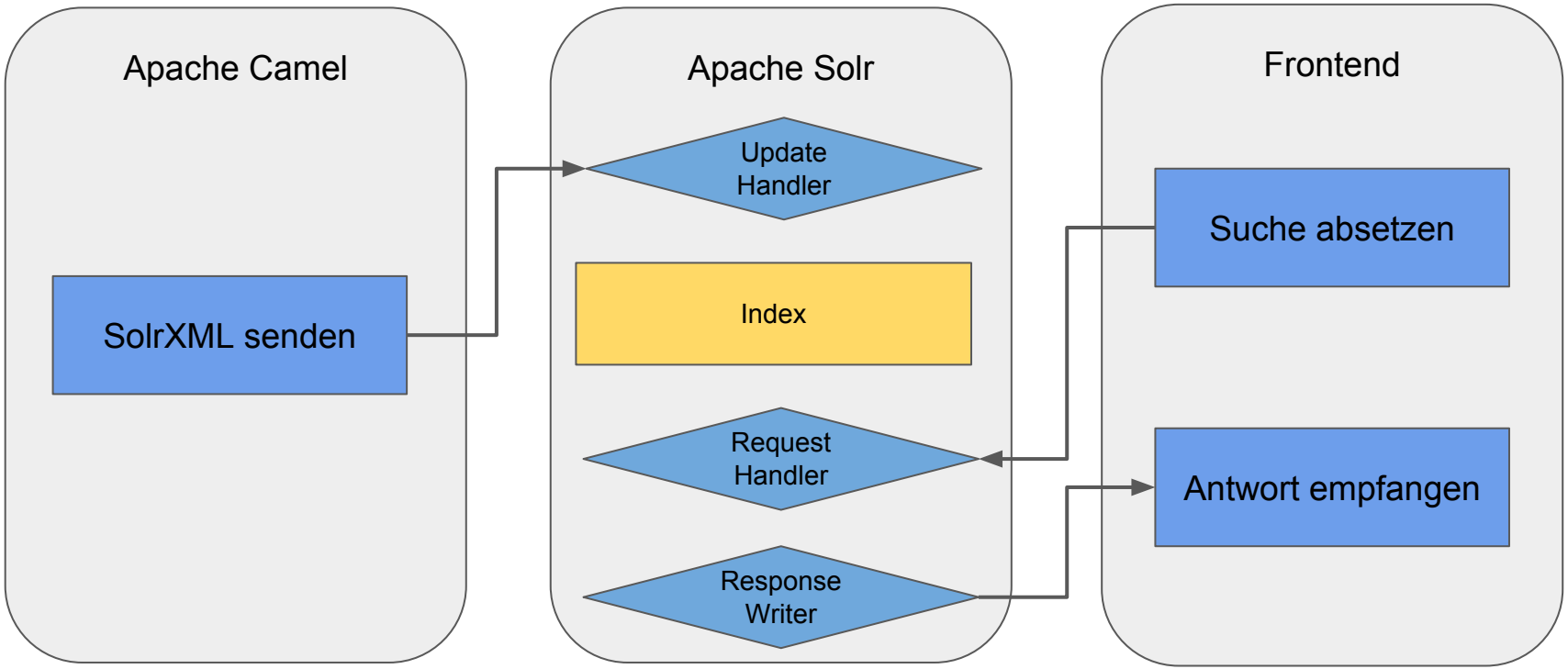
Apache Solr - innerhalb einer Applikation



Apache Solr - Grundbegriffe

- **Core**
  - Ein Core ist ein Index mit einer passenden Konfiguration
- **Index**
  - Medium, in dem die Daten für die Suche gespeichert sind.
- **Schema**
  - Beschreibung, wie der Index aufgebaut ist (z.B. Felder)
- **RequestHandler**
  - Solr-Komponenten, die Requests entgegennehmen und verarbeiten
- **UpdateHandler**
  - RequestHandler, der die Indexierung von Daten steuert
- **ResponseWriter**
  - Erstellt Response auf Request und bestimmt dessen Format
- **SolrXML**
  - Definierter Standard für Solr, um Dokumente in den Index aufzunehmen, zu aktualisieren oder zu löschen

Apache Solr - Camel/Frontend Kontext



Apache Solr - Advanced Topics

- SolrCloud
- ZooKeeper-Ensemble

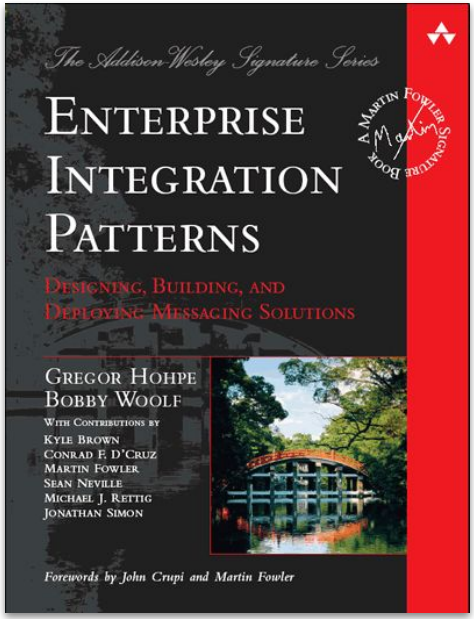
Apache Camel

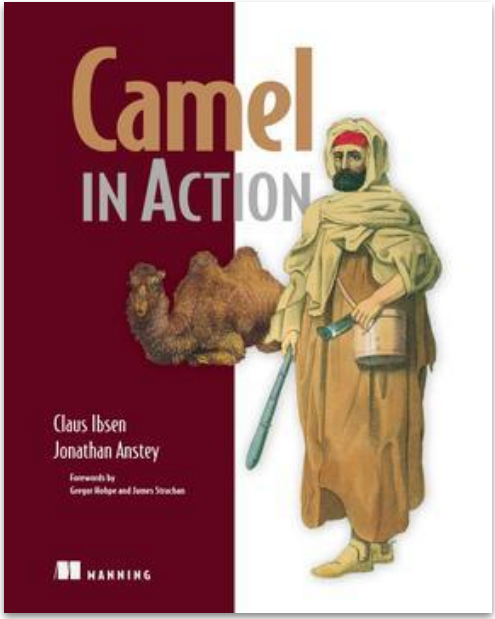
IT systems may not have been designed to be accessible from other systems, and if they were designed for interoperability, they may not speak the protocol you need.

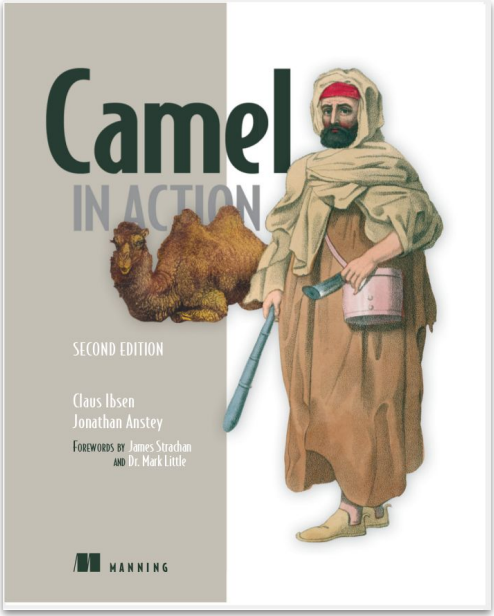
Apache Camel is an open source integration framework that aims to make integrating systems easier.

-- *Camel in Action, Second Edition*

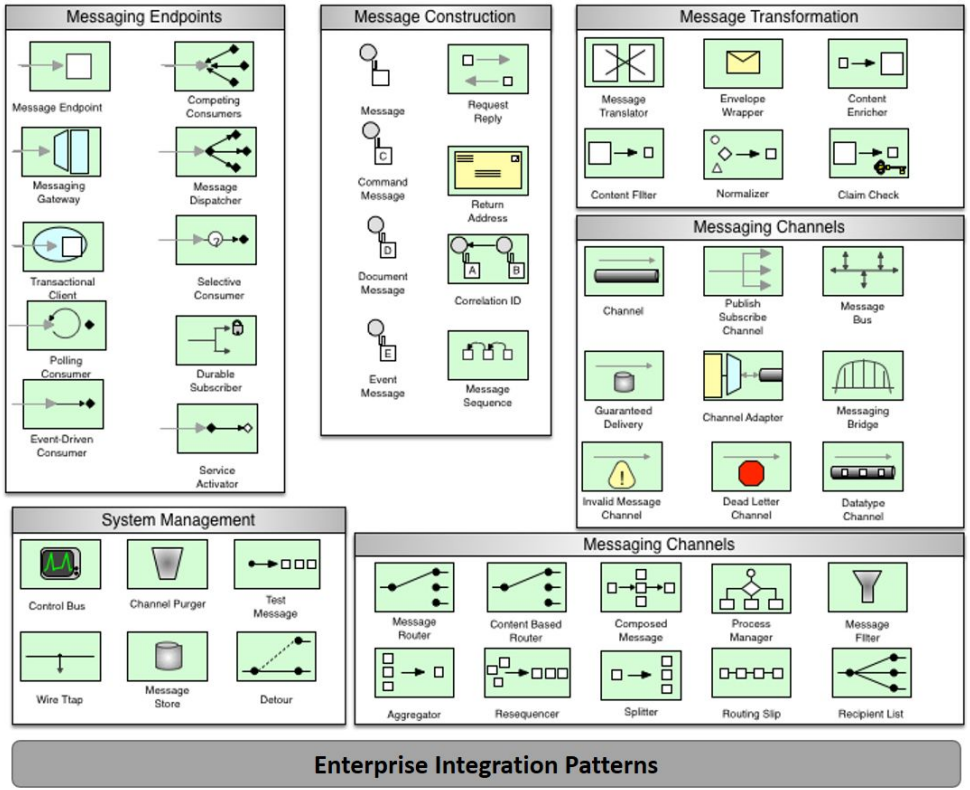
Apache Camel





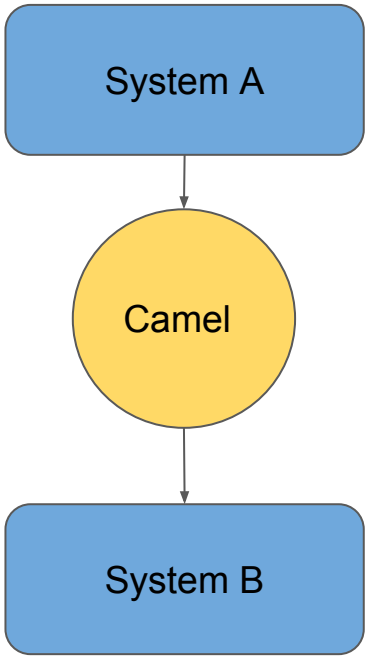


Apache Camel - Integration Patterns



https://www.oreilly.com/library/view/distributed-computing-in/9781787126992/assets/cdcd6a8b-218e-4c39-894b-4ae0f451d1c6.png

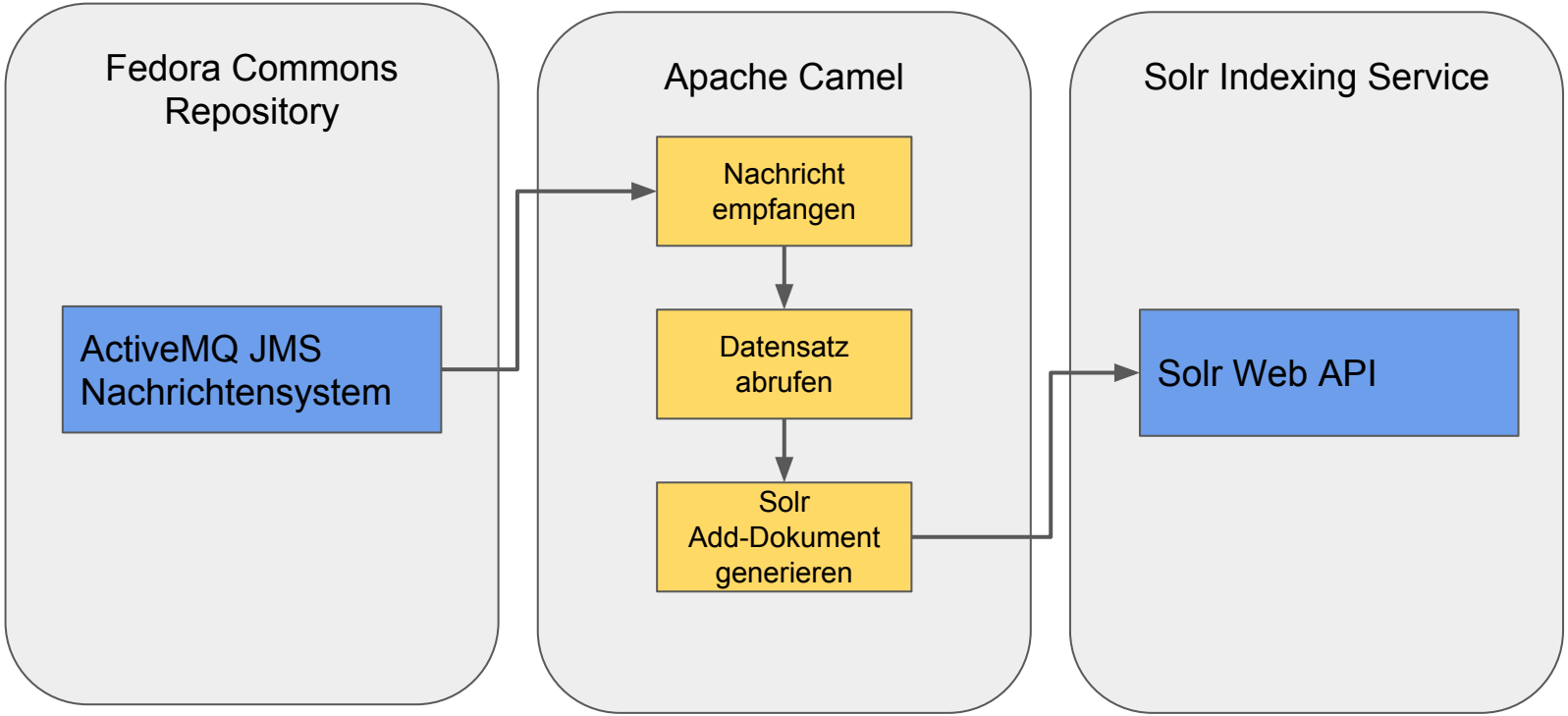
http://camel.apache.org/enterprise-integration-patterns.html



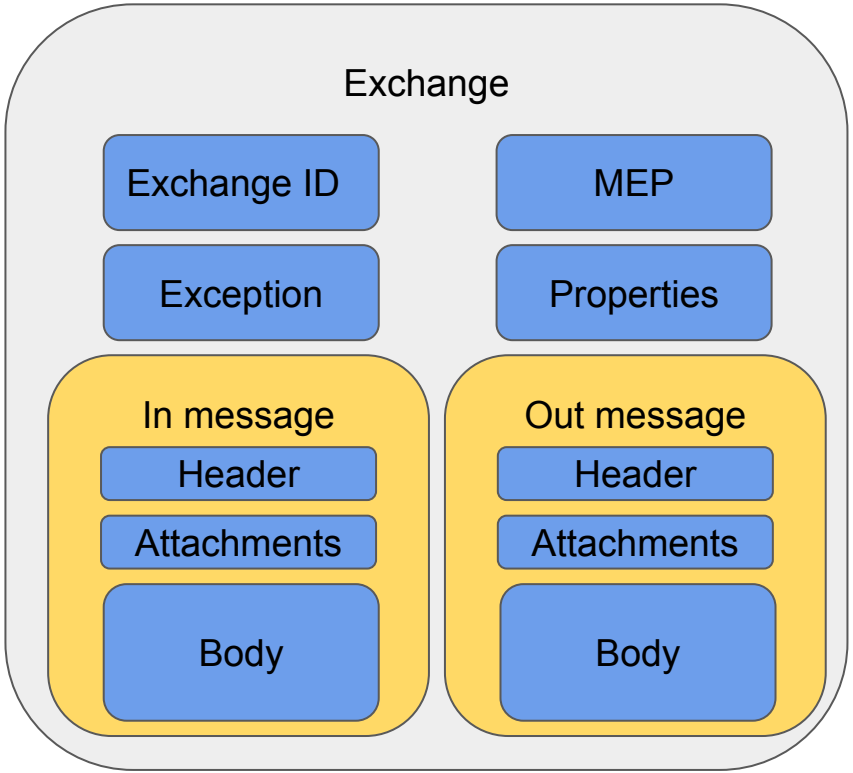
Apache Camel - Grundbegriffe

- **Endpoint**
  - Ein- bzw. Ausgang von/zu externen Systemen
- **Route**
  - Nachrichtenweg von Exchanges von Endpoint zu Endpoint
- **Context**
  - Zusammenfassung von Routen, meist auf separate Anwendungen verteilt
- **Processor**
  - Java Komponente zur Datenverarbeitung
- **Component**
  - Bieten Endpoints und Prozessoren für bestimmte Systeme und Aufgaben
- **Exchange**
  - Header
  - Properties
  - Body
  - (In/Out-Pattern)

Apache Camel - Fedora/Solr Kontext

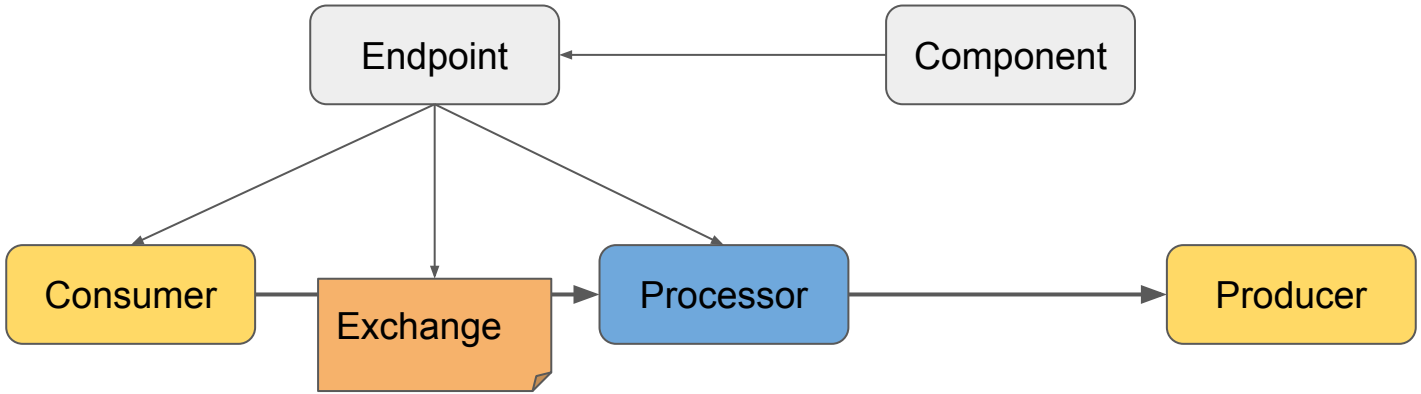


Apache Camel - Exchange



- Exchange ID
  - eindeutige ID
- MEP
  - Message Exchange Pattern
  - (InOut / InOnly)
- Exception
  - Routing Error
- Properties
  - Schlüssel-Wert Paare, gültig im Exchange Lebenszyklus
- In Message
- Out Message

Apache Camel - Routing



Apache Camel - Demo

Code

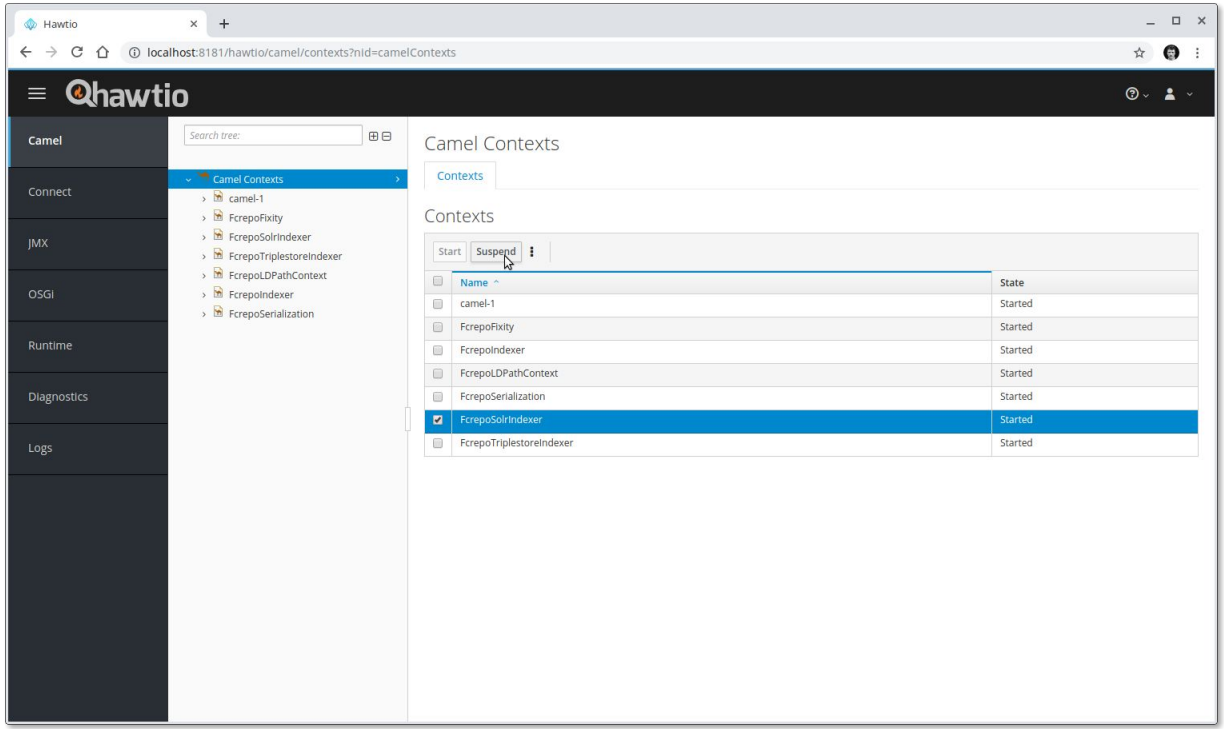
- https://github.com/fcrepo4-exts/fcrepo4-vagrant
  - git checkout -b kimws19 a93c313
- https://github.com/UB-LMU/kimws19-fedoracamelsolr
  - git reset --hard step1
  - git reset --hard step2
  - git reset --hard step3
  - ...

Webanwendungen

- http://localhost:8080/fcrepo
  - fedoraAdmin:secret3
- http://localhost:8080/solr
- http://localhost:8181/hawtio
  - karaf:karaf

Apache Camel - Demo - Hawtio

Vorkonfigurierte SOLR Routen der VM abschalten



Apache Camel - Advanced Topics

- Parallelisierbarkeit
- Splitter und Aggregationen
- Error und Exception Handling
  - Retry Verfahren
  - Dead Letter Queues
- Route Policies
- ...

Vielen Dank für Ihre Aufmerksamkeit!