

Workshop

Text Mining mit Open Semantic Search

Markus Mandalka, Berlin

Jörg Wettlaufer, Göttingen

Stefan Ziehe, Göttingen

Block I: 90 Minuten Workshop

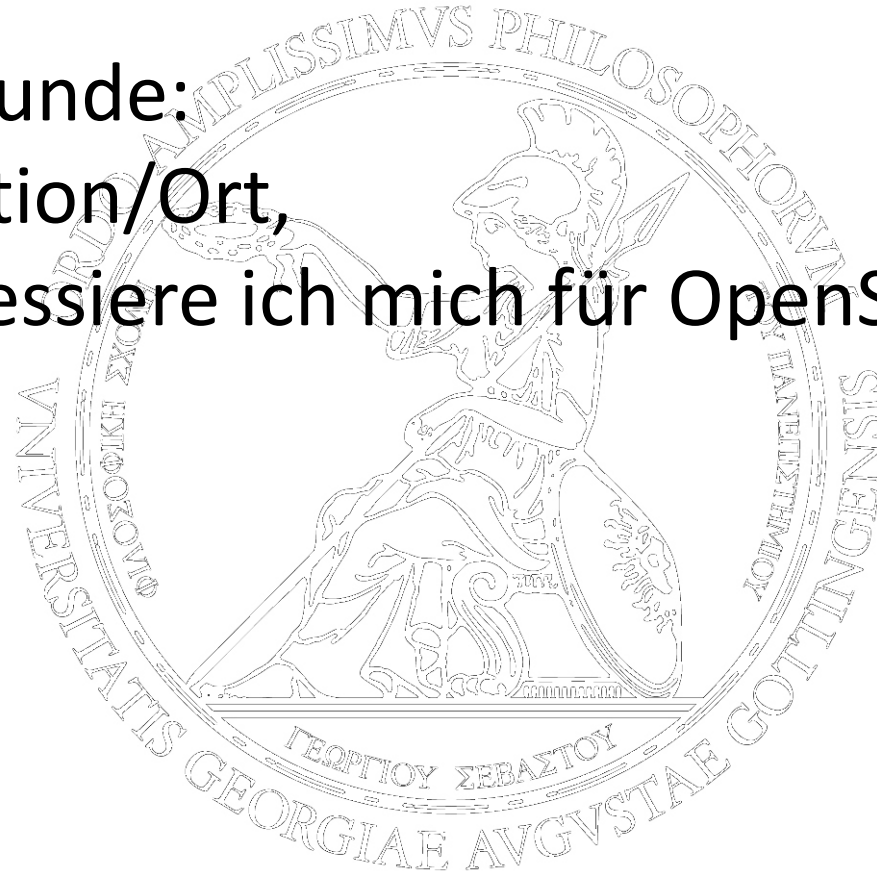
- 15 Minuten Begrüßung / Einführung
- 30 Minuten Hands On - Installation von Virtual Box / OSDS auf eigenen Geräten, sofern dies noch nicht vorher durchgeführt wurde. Behebung von Problemen.
- 30 Minuten Lecture zur Funktionalität von OSDS / Parallel: Indexierung des Beispielkorpus
- 15 Minuten Hands On / Ausprobieren des UI / Suchinterface anhand einfacher Abfragen mit boolschen Operatoren, Einsatz der Suchfacetten

30 Minuten Pause

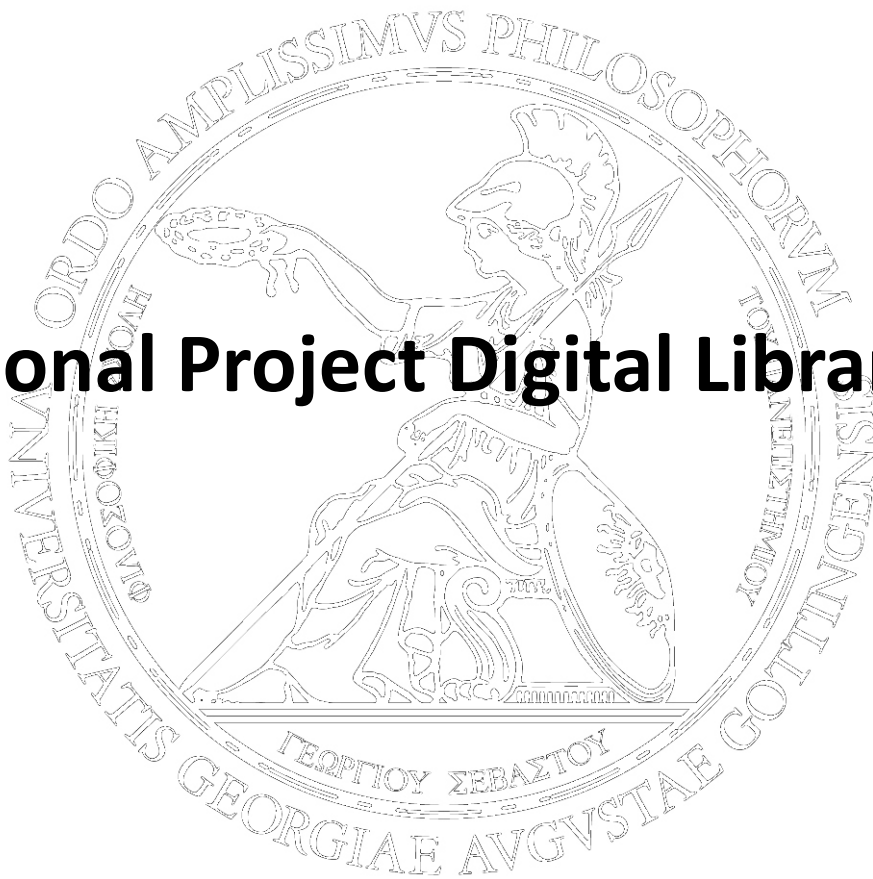
Block II: 90 Minuten Workshop

- 30 Minuten Vorstellung der erweiterten Funktionen von OSDS (OCR, Semantic Tagging, NER, Suche mit Listen, neo4j, Visualisierungen)
- 30 Minuten Hands On: Bearbeitung verschiedener Aufgabenstellungen zu diesen Funktionen (unscharfe Suche, Einbindung von Listen und Ontologien)
- 30 Minuten gemeinsame Diskussion der Erfahrungen mit OSDS, Ausblick des Entwicklers auf weitere Funktionalitäten und zukünftige Entwicklungen sowie Fragen.

Vorstellungsrunde:
Name, Affiliation/Ort,
Warum interessiere ich mich für OpenSemanticSearch?



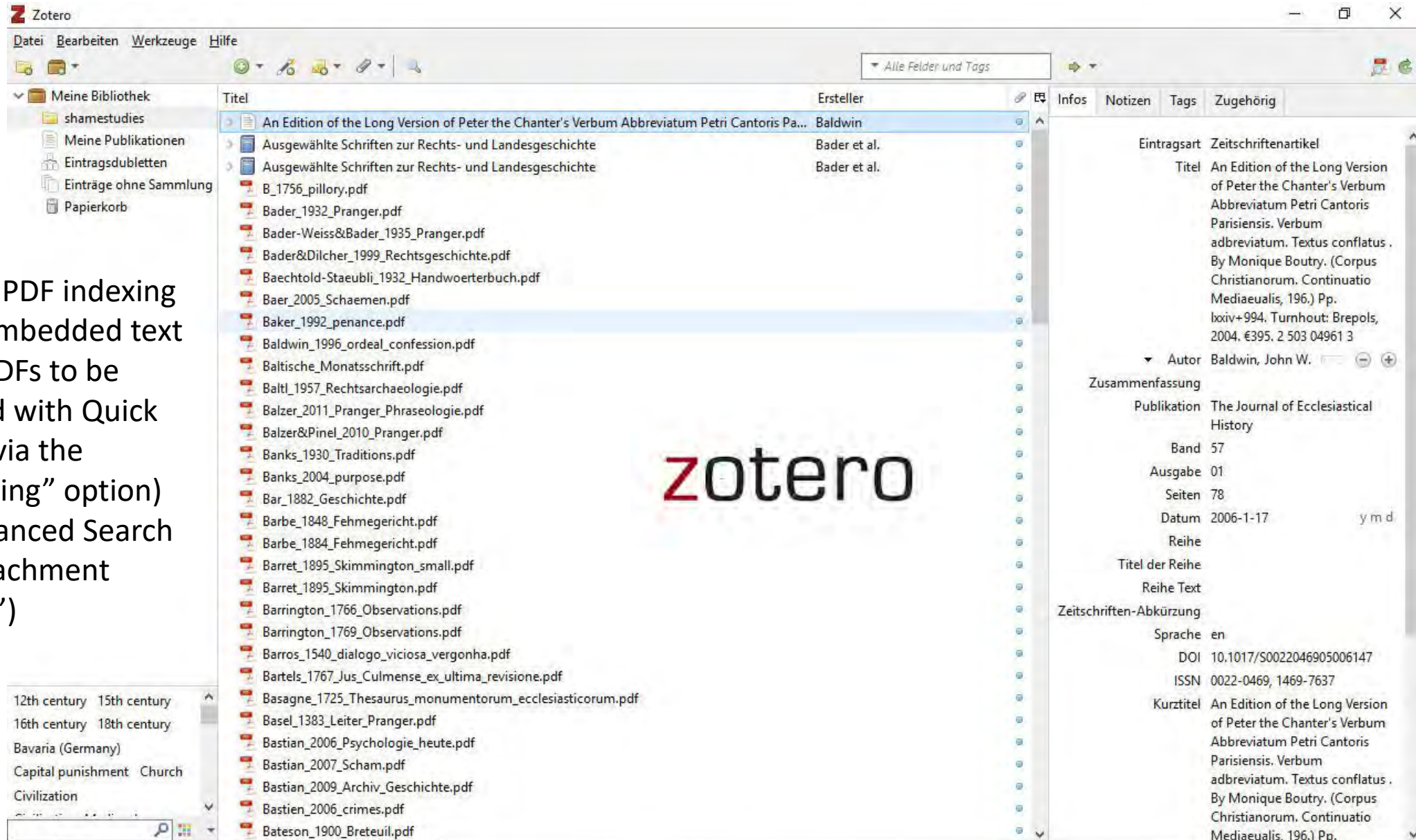
PHILOSOPHISCHE FAKULTÄT



“Personal Project Digital Libraries”

PHILOSOPHISCHE FAKULTÄT

Full-text PDF indexing
allows embedded text
within PDFs to be
searched with Quick
Search (via the
“Everything” option)
and Advanced Search
(via “Attachment
Content”)

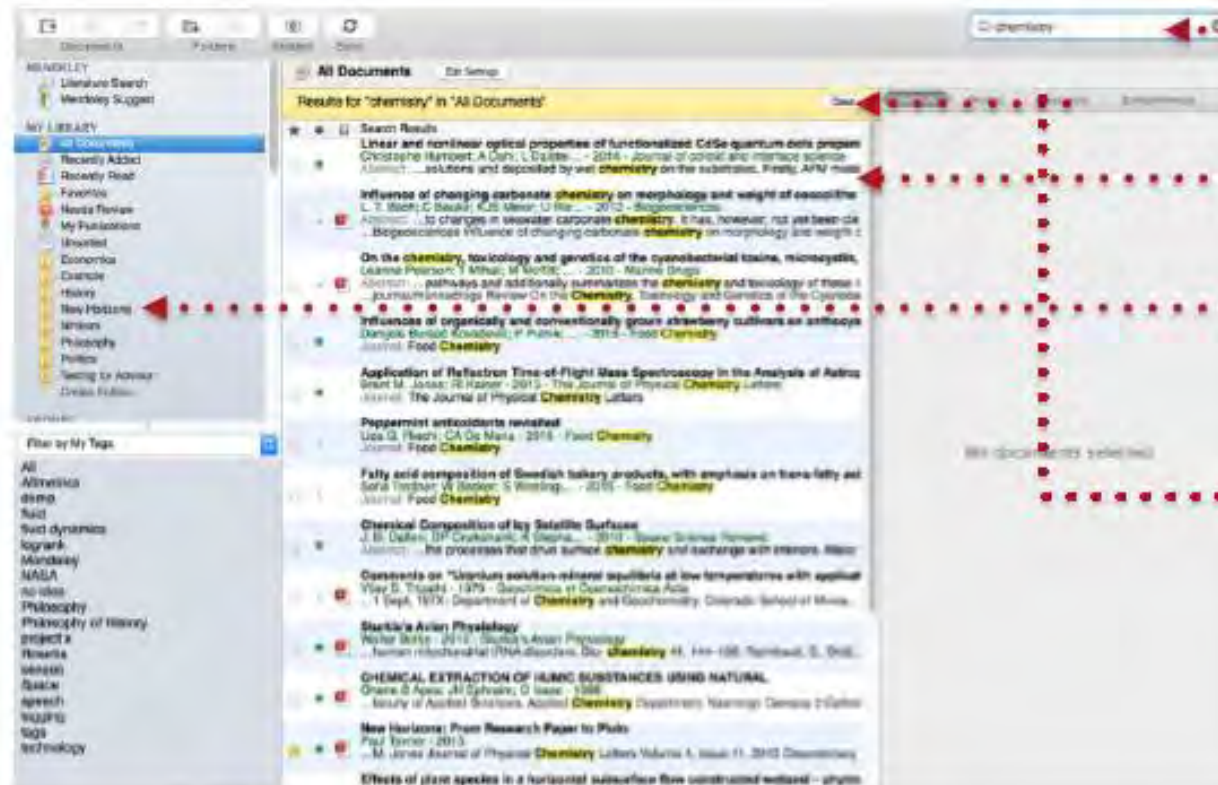


The screenshot displays the Zotero application window. On the left, a sidebar shows the library structure with 'Meine Bibliothek' and 'shamestudies'. The main pane lists various PDF files, including 'An Edition of the Long Version of Peter the Chanter's Verbum Abbreviatum Petri Cantoris Pa...' by Baldwin. The right pane provides detailed metadata for the selected entry, such as 'Eintragsart: Zeitschriftenartikel', 'Autor: Baldwin, John W.', and 'Publikation: The Journal of Ecclesiastical History'.

zotero



Search Your Documents



Enter your search term
in the search field

The main view will be
filtered accordingly

Click on a specific folder
to search within it

Use the clear button to
remove the search filter

Mendeley's search tool
will look at reference
metadata, but will also
search within the full text
of PDF papers.

Types of (research) data stored in PDF-format



- PDF with OCR
- Copyrights?
- Accuracy?
- PDF version?



- PDF from Scan
- Copyrights?
- no text layer
- No metadata
- PDF version?



- PDF from word processing or DTP
- Copyrights?
- With (some) metadata
- Standardized pattern for references

Open Semantic Search

Free Software for your own Search Engine, Explorer for Discovery of large document collections, Media Monitoring, Text Analytics, Document Analysis & Text Mining platform based on Apache Solr or Elasticsearch open-source enterprise-search and Open Standards for Linked Data, Semantic Web & Linked Open Data integration

[Download Open Semantic Search](#)

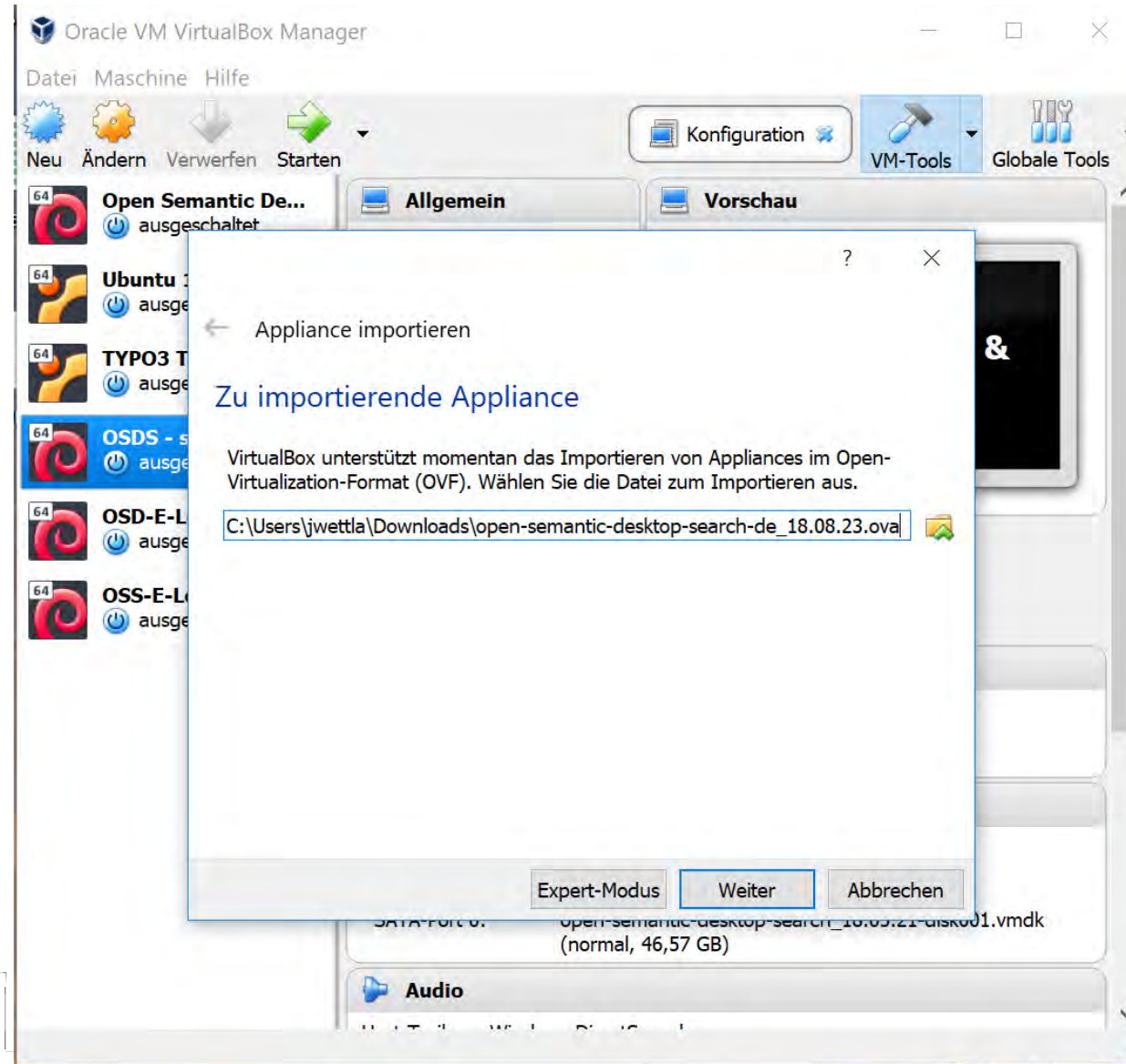
Integrated research tools for easier searching, monitoring, analytics, discovery & text mining of heterogenous & large document sets & news with free software on your own server

Search engine
(Fulltext search)

Thesaurus & Grammar
(Semantic search)

Interactive filters
(Faceted search)

- 
1. [Open Semantic Desktop Search](#)
 2. [Open Semantic Search Appliance](#)
 3. [Open Semantic Search Server \(Package\)](#)





← Appliance importieren

Appliance-Einstellungen

Dies sind die in der Appliance beschriebenen virtuellen Maschinen mit den entsprechenden Abbildungen für den Import in VirtualBox. Sie können Änderungen an vielen dieser Einstellungen mittels Doppelklick bzw. durch Auswahl der entsprechenden Checkbox ändern.

Virtuelles System 1		^
	Name	Open Semantic Desktop Search
	Produkt-URL	https://opensemantictsearch.org
	Gast-Betriebssystem	 Debian (64-bit)
	CPU	2
	RAM	5000 MB
		▼

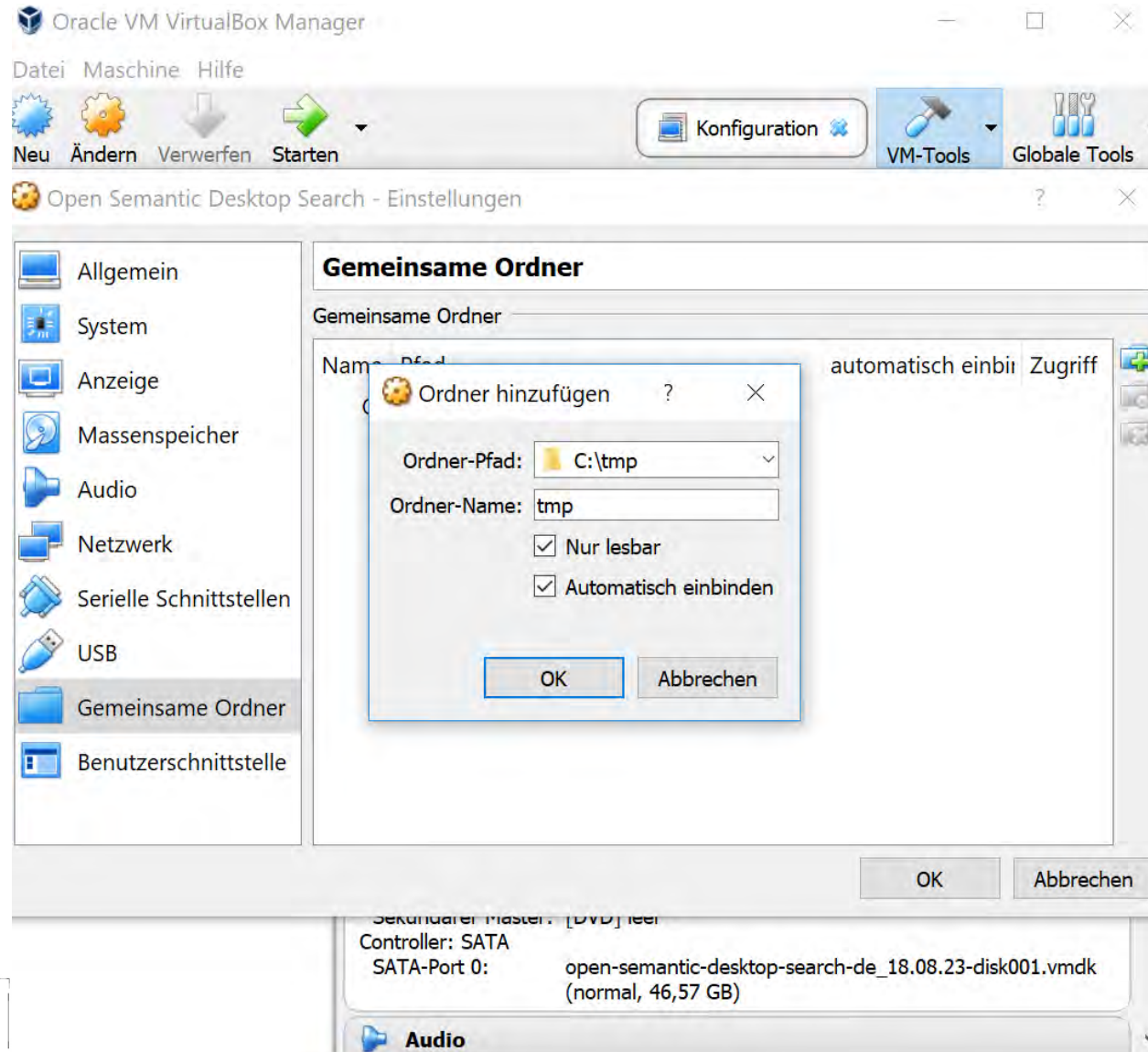
☐ Zuweisen neuer MAC-Adressen für alle Netzwerkkarten

Appliance ist nicht signiert

Standardeinstellungen

Importieren

Abbrechen



Gemeinsame Ordner

Gemeinsame Ordner

Name

Pfad

Automat



Ordner hinzufügen

?

×

Ordner-Pfad: C:\index

Ordner-Name: index

☐ Nur lesbar

☒ Automatisch einbinden

OK

Abbrechen

localhost/search-apps/datasources/

DatasourcesFiles & directories (filesystem)Websites & Intranet (HTTP)Newfeeds (RSS)Spreadsheets (CSV)Annotations (Hypothesis)

Queue

Show running and open imports and analysis tasks (ETL tasks)

Index datasource(s) periodic

If you want to (re)index file directories or web pages on a regular base, add it persistent using the manager apps on the topbar for crawling of files systems or web pages.

Index file(s) or web page(s) once

Index an file or webpage only once.

URL or list of URLs:

Import

Aktivitäten

Firefox ESR

So 12:52

Flower - Mozilla Firefox

Suche: Brandenburg

Flower

localhost/search-apps/setup/

localhost:5555/tasks

Suchen

Flower

Dashboard

Tasks

Broker

Monitor

Logout

Docs

Code

Show 10 entries

Search:

Name	UUID	State	args	kwargs	Result	Received	Started	Runtime	Worker
etl.index_file	c8321d5a-9a25-4243-9936-f8cc0a1daca1	SUCCESS	[]	{'filename': '/home/user/Dokumente/user/VBox_GAs_6.0.4/AUTORUN.INF'}	'None'	2019-03-24 07:41:14.872	2019-03-24 07:41:14.876	0.012	celery@debian
etl.index_file	9f944d60-08d5-4c53-ae37-780d714160ec	SUCCESS	[]	{'filename': '/home/user/Dokumente/user/VBox_GAs_6.0.4/autorun.sh'}	'None'	2019-03-24 07:41:14.880	2019-03-24 07:41:14.881	0.009	celery@debian
etl.index_file	fb07f0a8-8cd9-4841-a9f7-7182a18dd0bc	SUCCESS	[]	{'filename': '/home/user/Dokumente/user/VBox_GAs_6.0.4/runasroot.sh'}	'None'	2019-03-24 07:41:14.891	2019-03-24 07:41:15.008	0.130	celery@debian
etl.index_file	8bdaad07-7dc9-4a9a-b915-7df726847aae	SUCCESS	[]	{'filename': '/home/user/Dokumente/user/VBox_GAs_6.0.4/VBoxLinuxAdditions.run'}	'None'	2019-03-24 07:41:15.012	2019-03-24 07:41:15.023	0.078	celery@debian
etl.index_file	7223dbfc-414c-485b-9381-e302a5b7360d	SUCCESS	[]	{'filename': '/home/user/Dokumente/user/VBox_GAs_6.0.4/TRANS.TBL'}	'None'	2019-03-24 07:41:15.009	2019-03-24 07:41:15.025	0.077	celery@debian
etl.index_file	65b7c772-8c41-4327-9115-688d04426aca	SUCCESS	[]	{'filename': '/home/user/Dokumente/user/VBox_GAs_6.0.4/VBoxWindowsAdditions.exe'}	'None'	2019-03-24 07:41:15.014	2019-03-24 07:41:15.112	0.112	celery@debian
etl.index_file	d8063d51-c4d6-412c-94f0-d0adacddd1cd	SUCCESS	[]	{'filename': '/home/user/Dokumente/user/VBox_GAs_6.0.4/VBoxWindowsAdditions-amd64.exe'}	'None'	2019-03-24 07:41:15.014	2019-03-24 07:41:15.275	0.054	celery@debian
etl.index_file	66daae36-9c2d-4e7f-b954-c25580a60461	SUCCESS	[]	{'filename': '/home/user/Dokumente/user/VBox_GAs_6.0.4/VBoxDarwinAdditionsUninstall.tool'}	'None'	2019-03-24 07:41:15.011	2019-03-24 07:41:15.276	0.055	celery@debian
etl.index_file	4bc7b126-e1e0-4c34-8f03-7fca3db55647	SUCCESS	[]	{'filename': '/home/user/Dokumente/user/VBox_GAs_6.0.4/cert/VBoxCertUtil.exe'}	'None'	2019-03-24 07:41:15.016	2019-03-24 07:41:15.393	0.112	celery@debian
etl.index_file	beff6c97-56ca-456f-8f9c-c105e5eeb53b	SUCCESS	[]	{'filename': '/home/user/Dokumente/user/VBox_GAs_6.0.4/cert/TRANS.TBL'}	'None'	2019-03-24 07:41:15.016	2019-03-24 07:41:15.398	0.053	celery@debian

Showing 1 to 10 of 42 entries

Previous

1

2

3

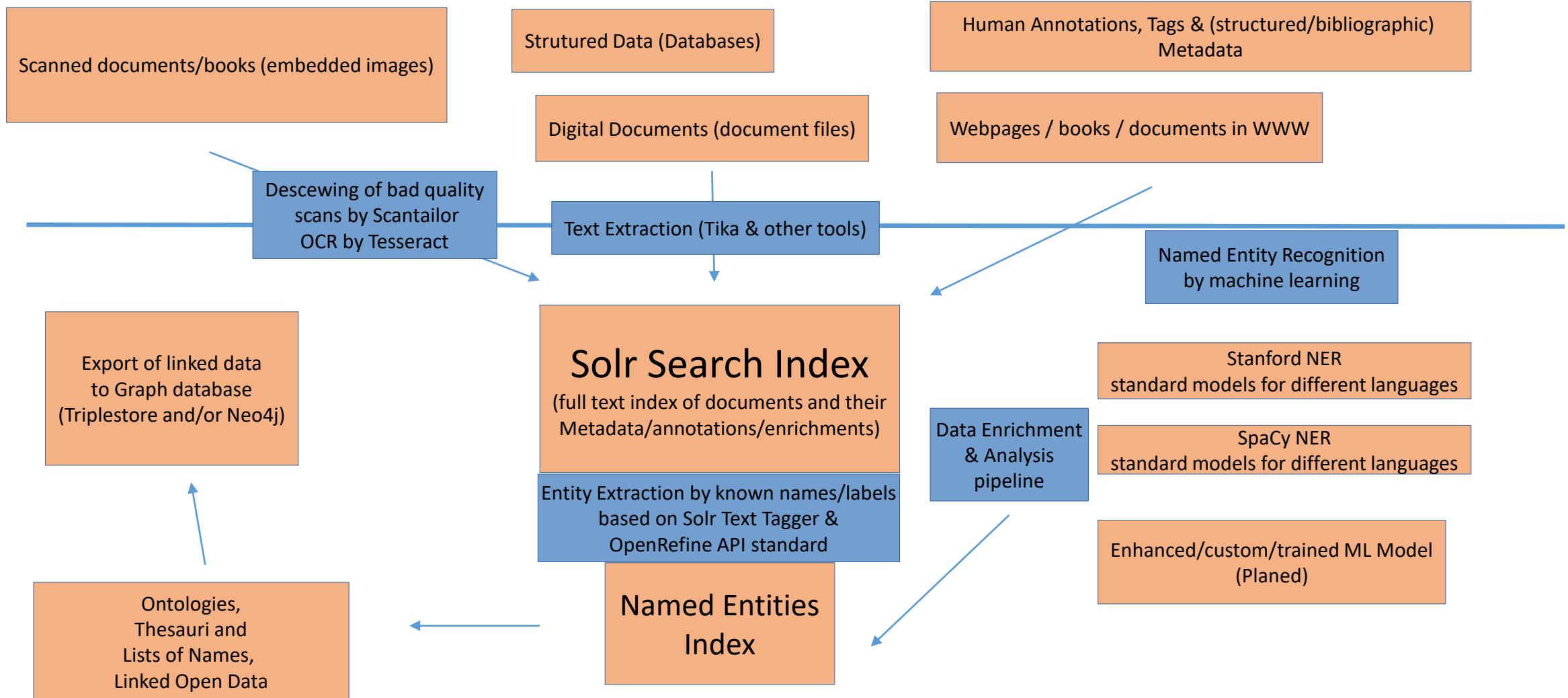
4

5

Next



Architektur und Funktionsübersicht von Open Semantic Desktop Search



Userinterface for full text search and faceted search by Solr

Recommender UI by fuzzy search in corpus

UI for text annotation / for Training ML model

UI for import of datasources

SKOS Thesaurus Editor (f.e. SKOSMOS)
for editing concept / entities

Integration of Wikibase as UI for editing ontologies

(Planned) Zotero DB as datasource
for full text document import

(Planned) Integration of Zotero as
Bibliographic metadata source for indexed documents

