EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Design Metadata Harvester
FID Theologie 2018

Dr. Johannes Ruscheinski / Johannes Riedl

29. September 2017

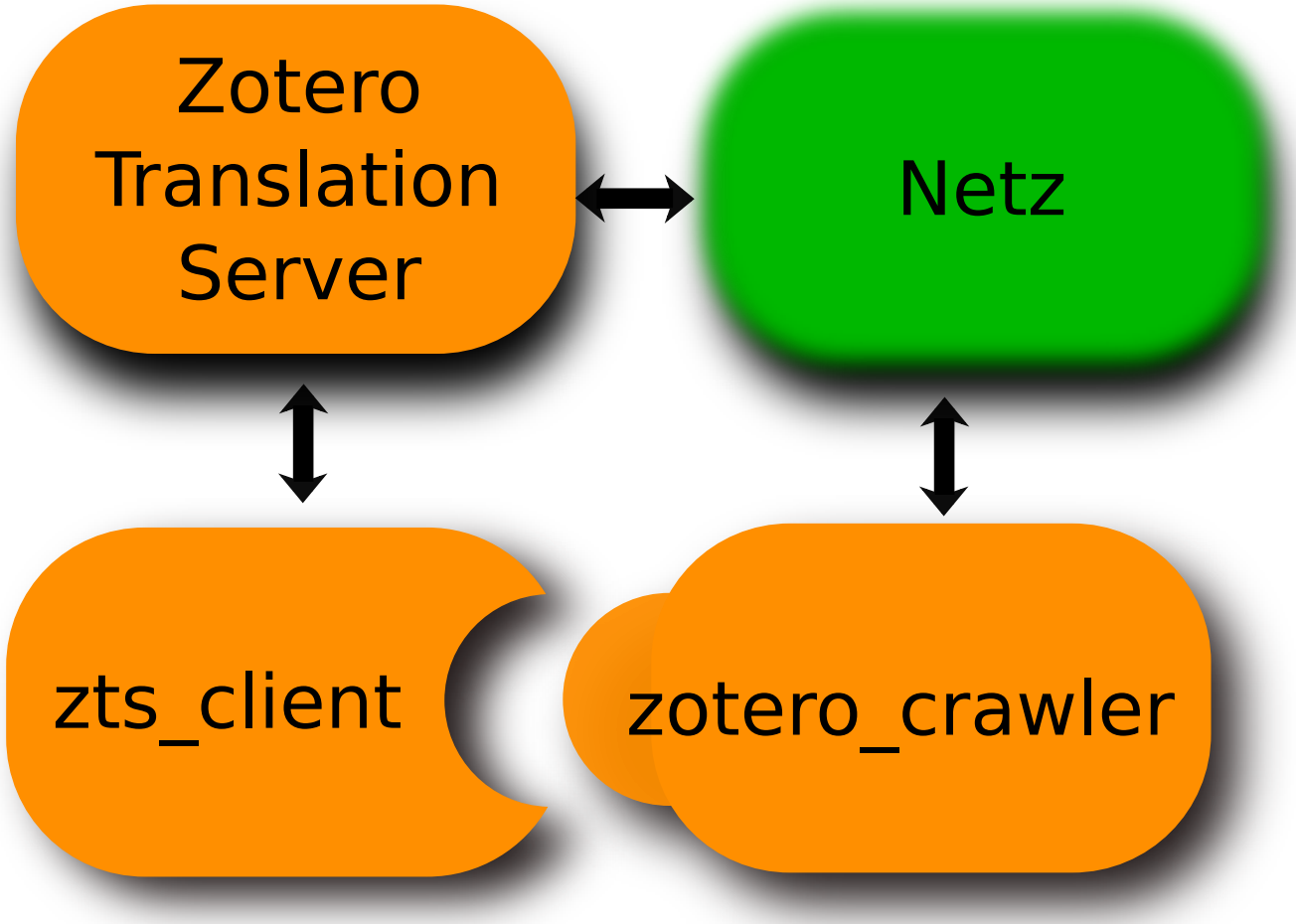
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Übersicht

Ein geplantes Tool zur Automatisierung des Harvesting und der Anreicherung von Metadaten aus diversen Webseiten mit Hilfe von Zotero-Plugins und des Exports in Form von MARC-XML oder MARC-21.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Übersicht (cont.)



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Zotero Translation Server

- ✓ Ursprünglich Browser-Plugin: "free, easy-to-use tool to help you collect, organize, cite, and share your research sources." (zotero.org)
- ✓ Translators: JS-Programme für die Extraktion von Metadaten aus Webseiten
- ✓ Translation Server: JSON-Protokoll

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Workflow

1. Eine Konfigurationsdatei für Einsprungspunkte für `zotero_crawler` muss angelegt werden
2. `zotero_crawler` identifiziert relevante Internetseiten und gibt diese auf `stdout` aus.
3. Anreicherungsdateien für `zts_client` müssen angelegt werden.
4. `zts_client` erzeugt MARC-XML und MARC-21.
5. Optional: Automatisierung mit `cron`.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

zotero_crawler

- ✓ Konfigurationsdatei:
  Jede Zeile spezifiziert eine Site und hat folgendes Format:
  Start-URL Crawl-Tiefe URL-Muster
- ✓ *Start-URL* gibt an wo das Harvesting anfängt
- ✓ *Crawl-Tiefe* gibt an wie viele Hops angefangen vom Start-URL gecrawlt werden sollen
- ✓ *URL-Muster* gibt an welche Seiten geharvestet werden sollen (Die Muster kann man Zotero-Plugins entnehmen)
- ✓ `zotero_crawler` crawlt dann alle Sites aus der Konfigurationsdatei und spuckt matchende auf `stdout` aus

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

zts_client

- ✓ Started `zotero_crawler` und arbeitet die dort ausgegebenen URLs ab.
- ✓ Für jedes URL wird ein Zotero Translation Server kontaktiert und die zurückgelieferten JSON-Daten übernommen.
- ✓ Metadaten können mit Mapping-Dateien angereichert werden, z.B. matchende ISSNs mit Schlagworten etc.
- ✓ `zts_client` verwendet eine Erkennung, um Duplikate von früher schon heruntergeladenen Metadatensätzen zu vermeiden.
- ✓ Als Ausgabe unterstüzt `zts_client` MARC-XML und MARC-21

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Fragen?