

Open Source Software zur Verarbeitung und Analyse von Metadaten

Prof. Magnus Pfeffer
`pfeffer@hdm-stuttgart.de`

- Hintergrund und Anforderungen
- Konkrete Softwarepakete
 - Einzelne Programme
 - Toolsets
 - Web-basierte Software
- Ausblick

Hintergrund und Anforderungen

- Metadatenmanagement „früher“
 - Ein Datenformat (MAB2)
 - Ein Regelwerk (RAK)
 - Ein Datenlieferant (Verbund)
 - Ein Datenempfänger (integriertes Bibliothekssystem oder lokaler OPAC)

- Datenmanagement
 - Viele Datenformate
 - MAB2, Marc21, Dublin Core, METS/MODS, ...
 - Viele Regelwerke
 - RAK, AACR, RDA, ...
 - Viele Datenquellen
 - Eigene Datenbanken, Verbund, Konsortium, Lieferanten, Anbieter, ...
 - Mehrere Datenempfänger
 - Integriertes Bibliothekssystem
 - Resource Discovery System

- Gewünschte Kompetenzen
 - Validierung und einfache Analyse von Datenlieferungen
 - Konsistente Feldbelegungen
 - Erkennen unvollständiger/korrupter Datensätze
 - Statistiken
 - Anpassung von Datenlieferungen
 - Filtern von Records aufgrund von Feldinhalten
 - Anpassen/Löschen/Ergänzen einzelner Felder
- Durchführen eines ETL-Prozesses
 - Extract: z.B. aus einem Repository
 - Transform: Anpassung und Formatwandlung
 - Load: z.B. in einem Index

- Anforderung an die Software
 - Keine Kosten für Anschaffung und Nutzung
 - Open Source
 - Klare Lizenzsituation
 - Eigene Anpassungen möglich
 - Einbringen in die Community (Forum, Bugtracker)
 - Nutzbar auch ohne Kenntnisse in Programmierung
 - Dokumentation mit Beispielen
 - Konfiguration über Dateien oder GUI
 - Umsetzung praxisrelevanter Szenarien möglich

- Dateien und Schnittstellen
 - Öffnen von MARC21 und MAB2 Dateien (nicht-XML)
 - Download von Daten über OAI-PMH und z39.50
- Analyse und Anpassung
 - Zählen der Records
 - Ausgeben der Titel, Verfasser, Jahr als Liste
 - Ersetzen eines Feldinhaltes
 - Zahl → Text aus einer Tabelle
 - Konvertierung
 - Dublin Core als CSV
 - JSON

Software



Malcolm Douglas McIlroy,
Head of Bell Labs in 1978

This is the Unix philosophy:

Write programs that do one thing and do it well.

Write programs to work together.

Write programs to handle text streams, because that is a universal interface.

■ Unix-“Philosophie“

- Ein Programm löst ein bestimmtes Problem
- Aufruf über die Kommandozeile
- Konfiguration über Aufrufparameter/Datei
- Ausgabe und Eingabe über Dateien und Pipes

- Bereitgestellt vom KOBV
- Java-basiert, Quellcode auf github
 - MABLE+: MAB2-Dateien (Bandformat)
 - Validierung und Fehleranalyse
 - Zählen von Sätzen
 - Indexierung
 - MARCEL: MARC21-Dateien (Bandformat)
 - Validierung
 - Feldstatistiken
 - MySQL-Import
- Keine Konfiguration, Einschränkungen beim Zeichensatz

- Bereitgestellt von der Deutschen Nationalbibliothek
- Java-basiert, Quellcode auf github
 - MabToMabxml
 - Konvertierung von MAB2-Datensätzen nach MABxml
 - MabxmlToMab
 - Konvertierung von MABxml-Dokumenten nach MAB2
 - XMabToUtf8
 - Konvertierung von MAB2-Standard-Zeichensatz ("x-Mab") nach UTF-8
- Keine Validierung, keine Analyse

- Bereitgestellt durch die UB Leipzig
- Go-basiert, Quellcode auf github
 - Anzeigen und Aufteilen von Dateien
 - Zählen von Records
 - Eliminieren von doppelten Einträgen
 - Konvertierung nach TSV und JSON
 - Laden in eine SQLite Datenbank
 - Arbeitet mit Marc21 und MarcXML Dateien
- Keine Konfiguration

- MarcEdit
 - Editor mit GUI für Windows
 - Keine Lizenz, kein Quellcode („free“)
- MARC Record Translation Program
 - Kommandozeilentool für Windows und Linux
 - Keine Lizenz, kein Quellcode („enjoy“)
 - keine erkennbare Weiterentwicklung
- User Controlled Generic MARC Converter
 - British Library and the National Library of Finland
 - Eigenwillige nicht-standardisierte Lizenz
 - Persönliche Registrierung erforderlich

- MarcBreaker/MarcMaker
 - Library of Congress, Kommandozeilentools für DOS (!)
 - Keine Lizenz, kein Quellcode („free“)
 - Keine Weiterentwicklung

- Unzählige kommerzielle Tools

- Bündelung von einzelnen Programmen
 - Lokale Installation auf PC-Arbeitsplatz oder Server
 - Abgestimmter Funktionsumfang der Tools
 - Ähnliche Struktur und Konfiguration
- Unterstützung komplexer Aufgaben
 - Umfangreiche ETL-Workflows
 - Kombination von Datenquellen
 - Speichern von Daten in Datenbanken
- Unterstützung für gemeinsames Arbeiten
 - Austausch von Konfigurationen („Rezepte“)

- Entwickelt im Rahmen des Projekts Culturegraph
- Hauptentwickler: DNB und HBZ-NRW
 - Komponenten
 - Flux
 - Skriptsprache zum Aufbau von Verarbeitungs-Pipelines
 - Umwandlung, Speichern und Analysieren von Daten
 - Morph
 - Anwendungsspezifische Sprache zur Verarbeitung von Metadaten
 - Modellierung als „Pipeline“
 - Konfiguration in XML
 - Framework
 - Technische Umsetzung der einzelnen Komponenten in Java
 - Erweiterbar durch eigene Programme

- Besonderheiten
 - Skalierbar für große Datenmengen
 - Sehr komplexe Transformationen umsetzbar

- Eindrücke
 - Stark fokussiert auf die Transformation von MARC21 Dateien
 - Dokumentation sehr knapp
 - Hoher Grundaufwand: auch einfache Aufgaben sind vergleichsweise komplex in der Umsetzung

- Entwicklung der Universitäten Bielefeld, Lund und Ghent
- Sammlung von Werkzeugen zur Datenverarbeitung in Bibliotheken
 - Einlesen von Metadaten aus unterschiedlichen Quellen
 - Speichern von Metadaten
 - Suchen in Metadaten
 - Export und Umwandlung in unterschiedliche Formate
- Sprache „Fix“
 - Beschreibung von Transformationen und Bearbeitung von Metadaten
- Framework in Perl zur Entwicklung eigener Erweiterungen

- Besonderheiten
 - Speicherung der Daten in MongoDB möglich
 - Vorbereitung der Daten für ElasticSearch integriert
 - Spracherweiterungen in Perl über CPAN verfügbar

- Eindrücke
 - Sehr umfangreiche Funktionen
 - Viele unterstützte Datenformate
 - Viele unterstützte Schnittstellen
 - Dokumentation mit vielen praktischen Beispielen
 - Nahezu alle Szenarien direkt umsetzbar

- Entwicklung an der Universität Genf
- Sammlung von Algorithmen für den Vergleich von bibliografischen Datensätzen
 - Unterschiedliche Ähnlichkeitsfunktionen
 - Import von Daten aus Dateien oder über OAI-PMH
 - MarcXML
- Programme für konkrete Anwendungen
 - Dublettenerkennung
 - Vorschlagssysteme
- Erweiterungen in Python möglich

- Eindrücke
 - Ähnlichkeitsfunktionen sind Alleinstellungsmerkmal gegenüber den anderen Toolsets
 - Dokumentation ausführlich
 - Aktive Weiterentwicklung fraglich

- Keine Kommandozeile oder lokale GUI
- Zugang und Nutzung über Browser
- Funktionsumfang von einfach bis umfassend

- Zentrale Installation vorgesehen
- Angebot als Software-as-a-service
- (lokale Installation weiter möglich)

- Entwicklung von Google
- Web-Anwendung zur Arbeit mit tabellarischen Daten
 - Datenbereinigung
 - Facettierung und Clustering von Werten
 - Batch-Änderungen
 - Konvertierung
 - Listen und Tabellenformate, XML, JSON
 - Auflösen von Nesting
 - Anreicherung
 - Erweitern von Tabellen durch externe Daten
 - Reconciliation
 - Matching von Daten auf externe Vorgaben



A power tool for working with messy data.

- Create Project
- Open Project
- Import Project
- Language Settings

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Get data from

[This Computer](#)

Web Addresses (URLs)

[Clipboard](#)

[Google Data](#)

Enter one or more web addresses (URLs) pointing to data to download:

Add Another URL

Next »



A power tool for working with messy data.

« Start Over Configure Parsing Options Project name: forecast Create Project »

	forecast - time - to	forecast - time - from	forecast - time - symbol - var	forecast - time - symbol - name	forecast - time - symbol - number	forecast - time - clouds - all	forecast - time - clouds - unit	for
1.	2016-03-16T09:00:00	2016-03-16T06:00:00	02d	few clouds	801	12	%	fev
2.	2016-03-16T12:00:00	2016-03-16T09:00:00	01d	clear sky	800	0	%	cle
3.	2016-03-16T15:00:00	2016-03-16T12:00:00	01d	clear sky	800	0	%	cle
4.	2016-03-16T18:00:00	2016-03-16T15:00:00	01n	clear sky	800	0	%	cle
5.	2016-03-16T21:00:00	2016-03-16T18:00:00	01n	clear sky	800	0	%	cle
6.	2016-03-17T00:00:00	2016-03-16T21:00:00	01n			0	%	cle
7.	2016-03-17T03:00:00	2016-03-17T00:00:00	01n			0	%	cle
8.	2016-03-	2016-03-17T03:00:00	01d			0	%	cle

Facet / Filter Undo / Redo 0

Refresh Reset All Remove All

✕ forecast - time - symbol - name change

5 choices Sort by: name count Cluster

- Load at most clear sky 14
- Preserve empty s few clouds 4
- Trim leading & tra light rain 17
- Parse cell text int numbers, dates, ... overcast clouds 1
- Store file source (file names, URLs in each row) scattered clouds 2

Facet by choice counts

Pick Record Elements Update Preview

0 row(s) of data

Parse data as

- Line-based text files
- CSV / TSV / separator-based files
- Fixed-width field text files
- PC-Axis text files
- JSON files
- RDF/N3 files
- XML files**
- Open Document Format spreadsheets (.ods)
- RDF/XML files

Version 2.8-beta.1 [TRUNK]

Help
About

38 rows

Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) rows « first

★	🗨	1.	2016-03-16T09:00:00Z	d	few clouds	801	12	%
★	🗨	2.	2016-03-16T12:00:00Z	d	clear sky	800	0	%
★	🗨	3.	2016-03-16T15:00:00Z			800	0	%
★	🗨	4.	2016-03-16T18:00:00Z					%
★	🗨	5.	2016-03-16T21:00:00Z					%
★	🗨	6.	2016-03-17T00:00:00Z					%
★	🗨	7.	2016-03-17T03:00:00Z					%
★	🗨	8.	2016-03-17T06:00:00Z	2016-03-17T03:00:00	01d	clear sky		%
★	🗨	9.	2016-03-17T09:00:00Z	2016-03-17T06:00:00	01d	clear sky		%
★	🗨	10.	2016-03-17T12:00:00Z	2016-03-17T09:00:00	01d	clear sky		%

- Facet ▶
- Text filter
- Edit cells ▶
 - Transform...
 - Common transforms ▶
 - Trim leading and trailing whitespace
 - Collapse consecutive whitespace
 - Unescape HTML entities
 - To titlecase
 - To uppercase
 - To lowercase
 - To number
 - To date
 - To text
 - Blank out cells
 - Edit column ▶
 - Transpose ▶
 - Fill down
 - Blank down
 - Split multi-valued cells...
 - Join multi-valued cells...
 - Cluster and edit...
- Edit column ▶
- Transpose ▶
- Sort...
- View ▶
- Reconcile ▶

- Eindrücke
 - Allgemeines Tool zur Datenanalyse und Datenverarbeitung
 - Sehr mächtig, intuitives Interface, überraschend schnell
 - Zahlreiche Tutorials und Anleitungen
 - Viele Erweiterungen , z.B.
 - Export als Linked Open Data
 - Nutzung bibliothekarischer Normdaten
 - Anwendung in Kombination mit anderem Toolset
 - Laden und Konvertierung: Toolset
 - Bereinigung und Anreicherung: Openrefine
 - Export: Openrefine/Toolset

- Datenintegrations- und -modellierungswerkzeug
 - Flexibles (elastisches), graphenbasiertes Datenmodell
 - Überführung von Daten aus heterogenen Datenquellen
- Middleware-Lösung
 - Bündelung aller Datenverarbeitungsprozesse
 - zwischen Datenmanagementsystemen und Webanwendungen (z.B. Discovery-System)
- Unterstützt u.a.
 - Analysen zur Verbesserung der Datenqualität
 - Deduplizierung und Zusammenführen von Titeldaten
 - FRBRisierung bibliografischer Daten

D:SWARM >> IMPORT >> DATA >> EXPORT HELP

RECORD JSON PROJECT

REVERT SAVE PROJECT SELECT RECORDS DEFINE SKIP FILTER PREVIEW EXECUTE

SOURCE

record_config
type
DOI
ISBN
PMID
abstract
apparent_dup
changed
created
deskman
edition
editorial_status
id
issued
key_publication
language
license
note
number_of_pages

TARGET

Internal Data Model OAI-PMH + DC Elements
creator
subject
description
publisher
contributor
date
type
format
identifier
type
lang
value
source
language
type
lang
value
relation

isbn value

language value

■ Besonderheiten

- Entworfen als Software-as-a-service Lösung
- Extrem flexible Architektur und Datenbank
- Funktionen zum gemeinsamen Arbeiten im Kern integriert
- Produktiver Einsatz an der SLUB Dresden
- Streaming Variante für große Datenmengen

■ Eindrücke

- Frei zugänglicher Prototyp im alpha-Stadium
- Schwerpunkt auf Mapping und Transformation
- Gute Dokumentation im Wiki

Ausblick

- Idee
 - Sammeln von Open-Data Metadatenpaketen
 - Dokumentierte Ablage in lokalem Speicherdienst
 - Bereitstellung in mehreren Datenformaten

- Erhoffter Nutzen
 - Zentrale Anlaufstelle für Datennutzer
 - Vermeidung von Doppelarbeit
 - Einfache Nachnutzung, auch in der Lehre

■ Idee

- Bereitstellung von Werkzeugen und Programme zur Metdatenverarbeitung ohne aufwändige Installation
- Vorbereitete Server-Einrichtung für virtuelle Server
- Wenn möglich: Installation mit web-basiertem Zugang

■ Erhoffter Nutzen

- Niederschwelliger Zugang zu den Werkzeugen
- Einfache Evaluation der Möglichkeiten
- Nutzung in der Lehre

- Viele Projekte
 - Unterschiedliche Ansätze und Schwerpunkte
 - Sehr unterschiedliche Entwicklungsaktivität
 - Software teilweise sehr schwer zu finden
 - Viele Miniprojekte ohne großen Nutzwert
 - Dennoch: Viele Tools nur „intern“ und (noch?) nicht veröffentlicht

- Nutzung
 - Dokumentation nicht für Einsteiger geeignet
 - Teilweise sehr spezielle Systemvoraussetzungen
 - Anwendung teilweise frustrierend
 - Realistische Workflows nur mit Kombinationen von unterschiedlichen Programmen umsetzbar

- Konkrete Anwendung in der Lehre WS 2016
- Modul „Metadatenmanagement“
 - Datenquellen
 - Datenformate
 - Schnittstellen
 - Typische Workflows
 - Software
 - Librecat als allgemeines Toolset
 - Datenaggregation
 - Indexierung und Filterung
 - ETL
 - Openrefine zur Analyse und Datenbereinigung



- DINI KIM Workshop 2016
- UB Mannheim, 04. und 05. April 2016
 - Vorträge
 - Ganztägige Workshops
 - Catmandu
 - Openrefine
- Link: <https://dini.de/veranstaltungen/workshops/kim2016/>



Danke für Ihre Aufmerksamkeit!

Folien online unter
<http://www.slideshare.net/MagnusPfeffer/>

Dieses Werk bzw. Inhalt steht unter einer
[Creative Commons Namensnennung - Weitergabe unter gleichen Bedingungen 3.0 Unported Lizenz.](https://creativecommons.org/licenses/by-sa/3.0/)



- MABLE+: <https://www.kobv.de/entwicklung/software/mable/>
- MARCEL: <https://www.kobv.de/entwicklung/software/marcel/>
- DNB Tools: <https://sourceforge.net/projects/dnb-conv-tools/>
- Marctools: <https://github.com/ubleipzig/marctools>
- Metafacture: <https://github.com/culturegraph/metafacture-core>
- Catmandu: <http://librecat.org/Catmandu/>
- MarcXimiL: <http://marcximil.sourceforge.net/>
- Openrefine: <http://openrefine.org/>

- OCLC: MARC specialized tools. Website. <https://www.loc.gov/marc/marctools.html>
- Code4Lib Wiki: Working with MARC. Website. http://wiki.code4lib.org/Working_with_MARC
- Margret Heller: A Librarian's Guide to OpenRefine. ACRL Tech Connect Blog. Website. <http://acrl.ala.org/techconnect/post/a-librarians-guide-to-openrefine>