# Metadaten modellieren und Schnittstellen nutzen A



Die Lernveranstaltung zum Thema Metadaten modellieren und Schnittstellen nutzen teilt sich in drei Termine auf. Dieser Blogbeitrag behandelt den Teil A, welcher sich aus den Vorlesungen vom 23. und 30. April zusammensetzt. In diesen ersten Teil behandelten wir die Metadatenmodellierung am Beispiel von OpenRefine.

Der Beitrag zum Teil B befindet sich hier.

## OpenRefine

[OpenRefine](https://openrefine.org/) hat einen Diamanten als Logo, weil es Rohdaten so aufbereitet und bereinigt, dass sie am Ende geschliffen und wertvoll sind. Zu den Hauptfunktionen zählen:

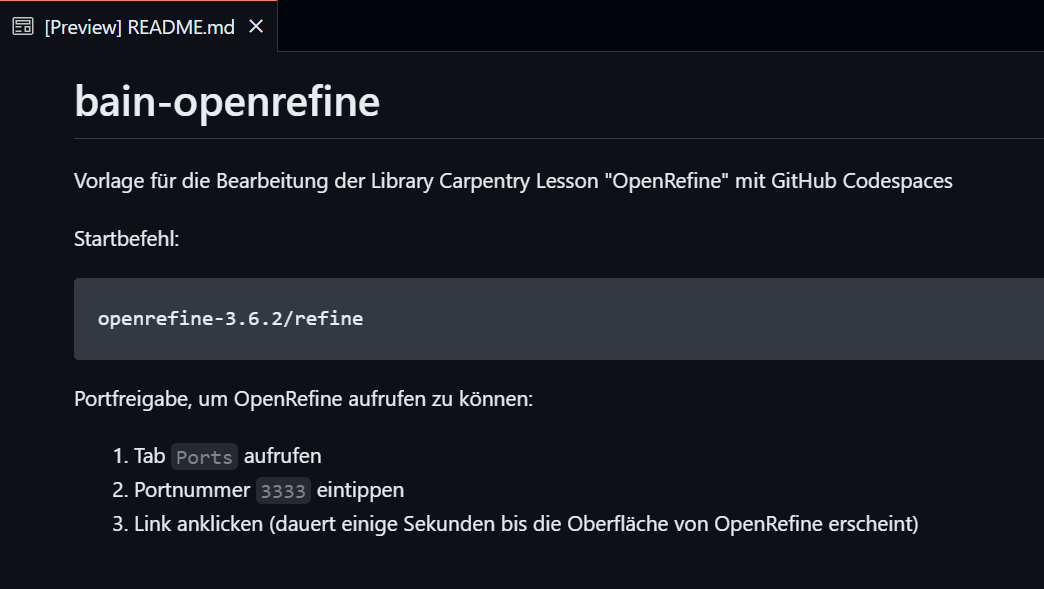
● [Faceting](https://openrefine.org/docs/manual/facets)  
● [Clustering](https://openrefine.org/docs/manual/cellediting#cluster-and-edit)  
● [Reconciliation](https://openrefine.org/docs/manual/reconciling)  
● [Infinite undo/redo](https://openrefine.org/docs/manual/running#history-undoredo)  
● [Privacy](https://openrefine.org/privacy)  
● [Wikibase](https://openrefine.org/docs/manual/wikibase/overview)

Die Software gehörte zu Google und hat dadurch eine ziemlich aktive Community. Unterstütz werden folgende Formate:

● tabellarische Daten (CSV, TSV, XLS, XLSX, inkl. TXT mit Trennzeichen oder festen Spaltenbreiten)  
● Einfaches “flaches” XML (z.B. MARCXML) oder JSON mit Übung einfach zu modellieren  
● Komplexes XML mit Hierarchien (z.B. EAD) möglich mit Zusatztools  
● [in Kombination mit MarcEdit](https://blog.reeset.net/archives/1873) für Analyse und Transformation von MARC21

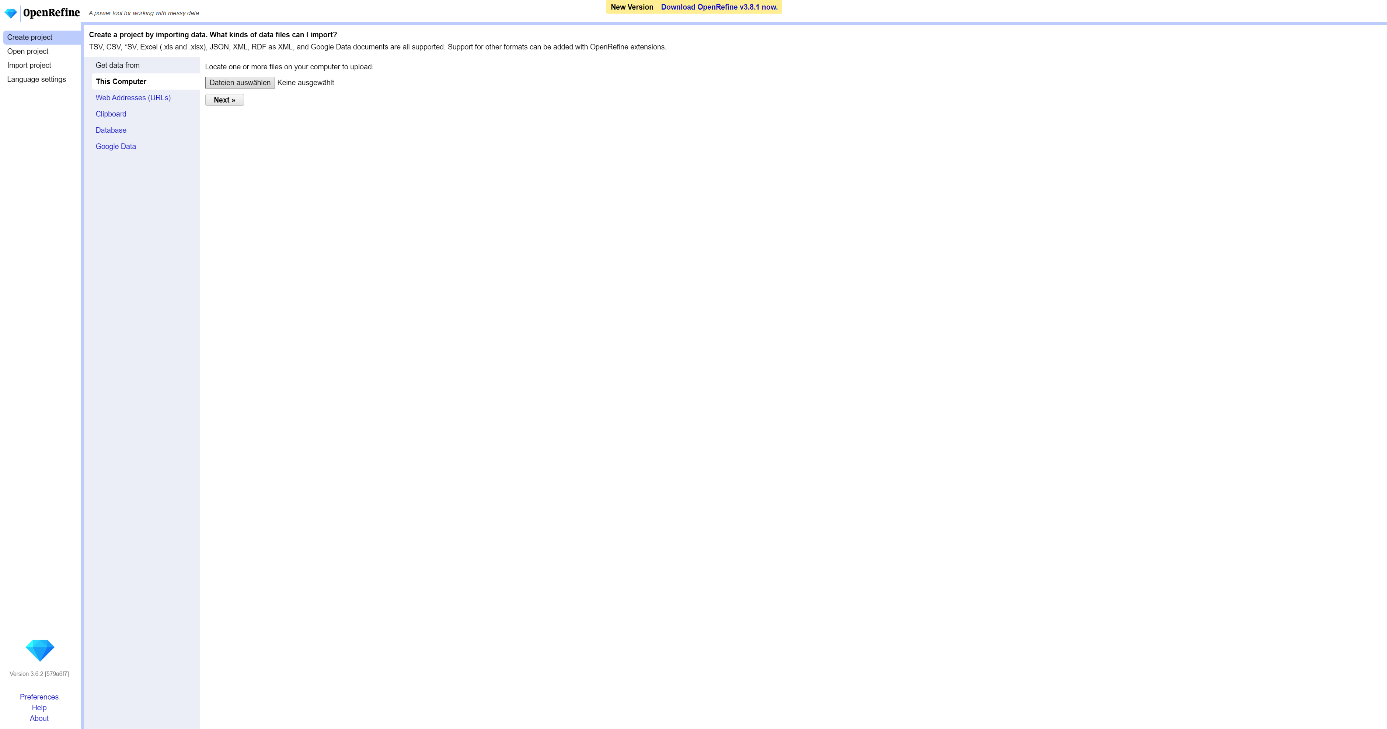
## Installation OpenRefine

Mit einem Codespace [(analog zu Blog 2)](https://anthonyxhd.github.io/lerntagebuch/post/3)_technische-grundlagen-02.html)konnten wir OpenRefine benutzen, was uns die lokale Installation gespart hat



## Basisfunktionen

Zu Beginn haben wir Daten anhand eines Links importiert und daraus ein neues Projekt erstellt.



Im Anschluss wurden Basisfunktionen vorgeführt und Fingerübungen durchgeführt, die zu folgenden Ergebnissen führten:

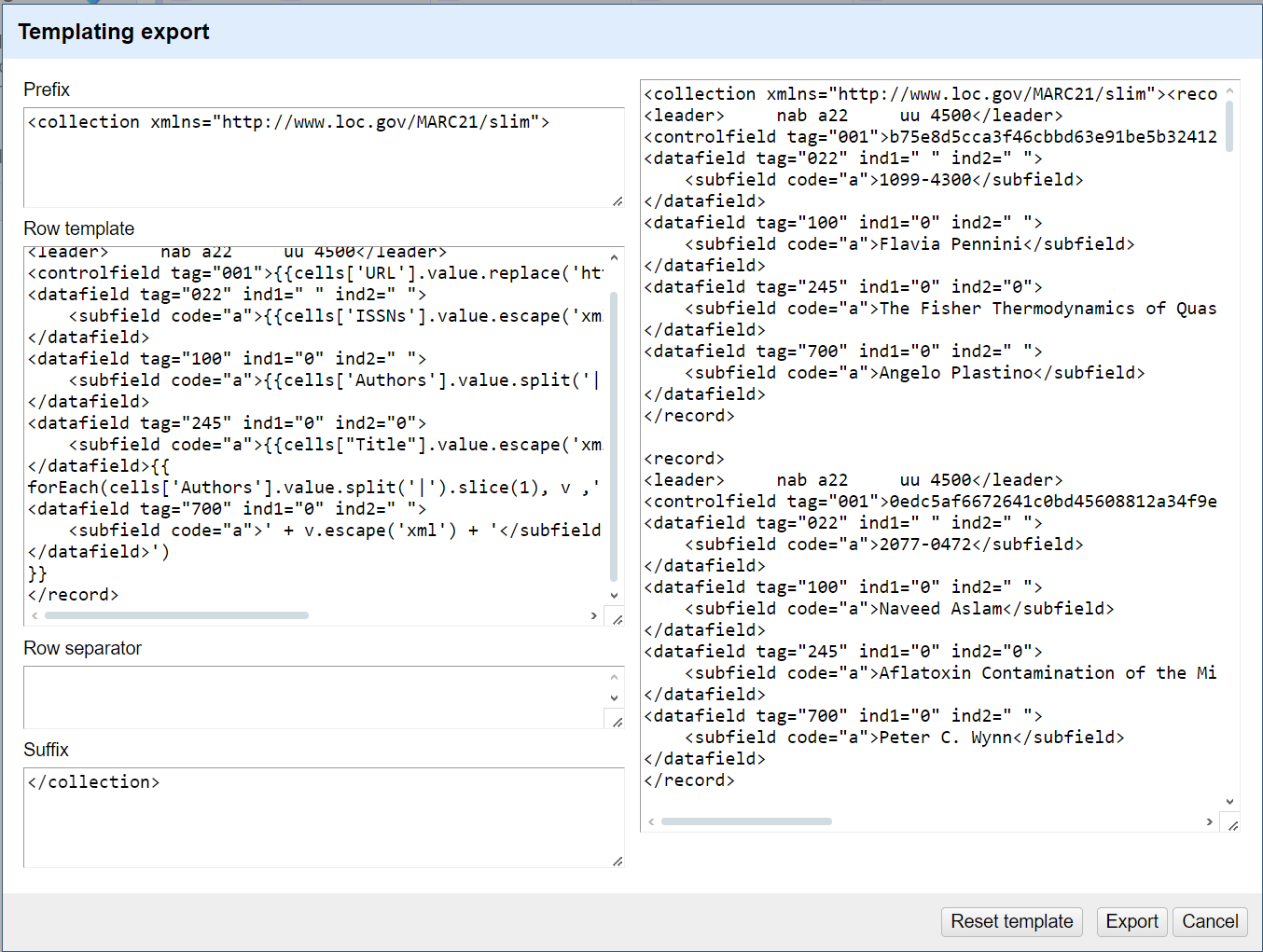
1) Was ist die am häufigsten vergebene Lizenz? CC BY (954 Zeilen)  
Wieviele Artikel haben keine Lizenz? 6 Zeilen

2) Warum erscheint MDPI AG zweimal und? Treffer hat 2x Leerschläge  
Wie lässt sich das Korrigieren 1x Leerschlag löschen

Zu Beginn der Lektion am 30. April wurde die [Reconciliation](https://pad.gwdg.de/GDKAQMKwT3u-VP6hZ6NtOA#Vorführung-Reconciliation) als Ergänzung der Besisfunktionen vorgezeigt, die es ermöglicht, Daten mit externen datenquellen wie [Wikidata](https://www.wikidata.org/wiki/Wikidata:Main_Page) abzugleichen, um zusätzliche Informationen abzurufen oder eigene Daten an diese Quellen zu spenden.

## Übung CSV nach MARCXML mit OpenRefine

[«The Templating Export window allows you to set your own separators, prefix, and suffix to create a complete dataset in the language of your choice.»](https://openrefine.org/docs/manual/exporting#templating-exporter) Mit die vorgegebene Vorlage als Ausgangsbasis haben wir die Aufgabe lösen können:



Man kann festlegen, dass nur der erste Wert der Zelle "Authors" in das Tag 100 (Lead Author) und die übrigen Autoren in das Tag 700 übertragen werden. Der Wert der Zelle "Title" wird in das Tag 245 eingefügt. Mit der Angabe von ".escape('xml')" werden Zeichen, die in XML nicht erlaubt sind, korrekt behandelt. Die Prüfung der ausgegebenen Datei zeigt, dass das XML valide ist.

[Fragen und Ergebnisse](https://pad.gwdg.de/GDKAQMKwT3u-VP6hZ6NtOA#Fragen--Ergebnisse)

## Learnings

Die Arbeit mit OpenRefine war eine bereichernde Erfahrung für mich. Sie vermittelte mir nicht nur technische Fähigkeiten zur Datenbereinigung und -analyse, sondern vertiefte auch mein Verständnis im Umgang mit großen, unstrukturierten Datenmengen. OpenRefine ist ein mächtiges Werkzeug, das sicherlich in vielen Bereichen der Archiv- und Informationswissenschaften äußerst nützlich sein wird. Die zusätzliche [Übungen 1 + 2](https://pad.gwdg.de/GDKAQMKwT3u-VP6hZ6NtOA#Aufgabe-2-Template-ergänzen) sowie die [die Lehrmaterialien von Library Carpentry zu OpenRefine](https://librarycarpentry.org/lc-open-refine/) wären zwar spannend gewesen, aber ich löste sie aufgrund von Zeitgründen nicht.

Im Grossen und Ganzen waren die Lerninhalte für mich ganz neu. Durch die viele Übungen war dieser Block sehr Praxisorientiert, was mir besonders gut gefallen hat :)