Zürcher Hochschule
für Angewandte Wissenschaften

Finanzen & Services
Hochschulbibliothek
Publikationsdienste

Keine Mauer nötig – Dublettenkontrolle bei der Datenmigration

DSpace Anwendertreffen, Bamberg, 11. April 2019

Friederike Gerland

Hochschulbibliothek ZHAW

12.000 Studierende – 3.000 Mitarbeitende ZHAW

8 Departemente / Fachrichtungen

3 Standorte – 3 Bibliotheken

49 Mitarbeitende (35 FTE)

1 Team Publikationsdienste – 2,6 FTE (reg. Betrieb), 1,3 FTE (Migration)

Nationale Open Access Strategie (Januar 2017)

[…]
Bis 2024 sollte Wissenschaftliches Publizieren in der Schweiz OA sein, alle mit öffentlichen Geldern finanzierten wissenschaftlichen Publikationen müssen im Internet frei zugänglich sein. Die OA-Landschaft wird aus verschiedenen OA-Modellen bestehen.
[…]

Keine Festlegung auf grünen oder goldenen Weg

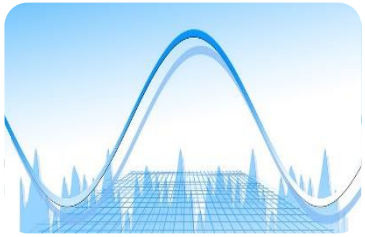
Projekt «Relaunch Forschungsdatenbanken»
Projektziele

ZHAW-Forschungsleistungen im Internet darstellen:

- Publikationen
- Forschungsprojekte

Anerkannte Qualitätsstandards einhalten

State-of-the-art Technik



Auswertungen und Reportings bereitstellen

Zürcher Hochschule
für Angewandte Wissenschaften

ZHAW digitalcollection und Migrationsprojekt

ZHAW digitalcollection *vor* Migration

- < 2.000 Einträge
- Abschlussarbeiten
- Working Papers ZHAW
- Open-Access-Publikationen (Zweitveröffentlichungen)

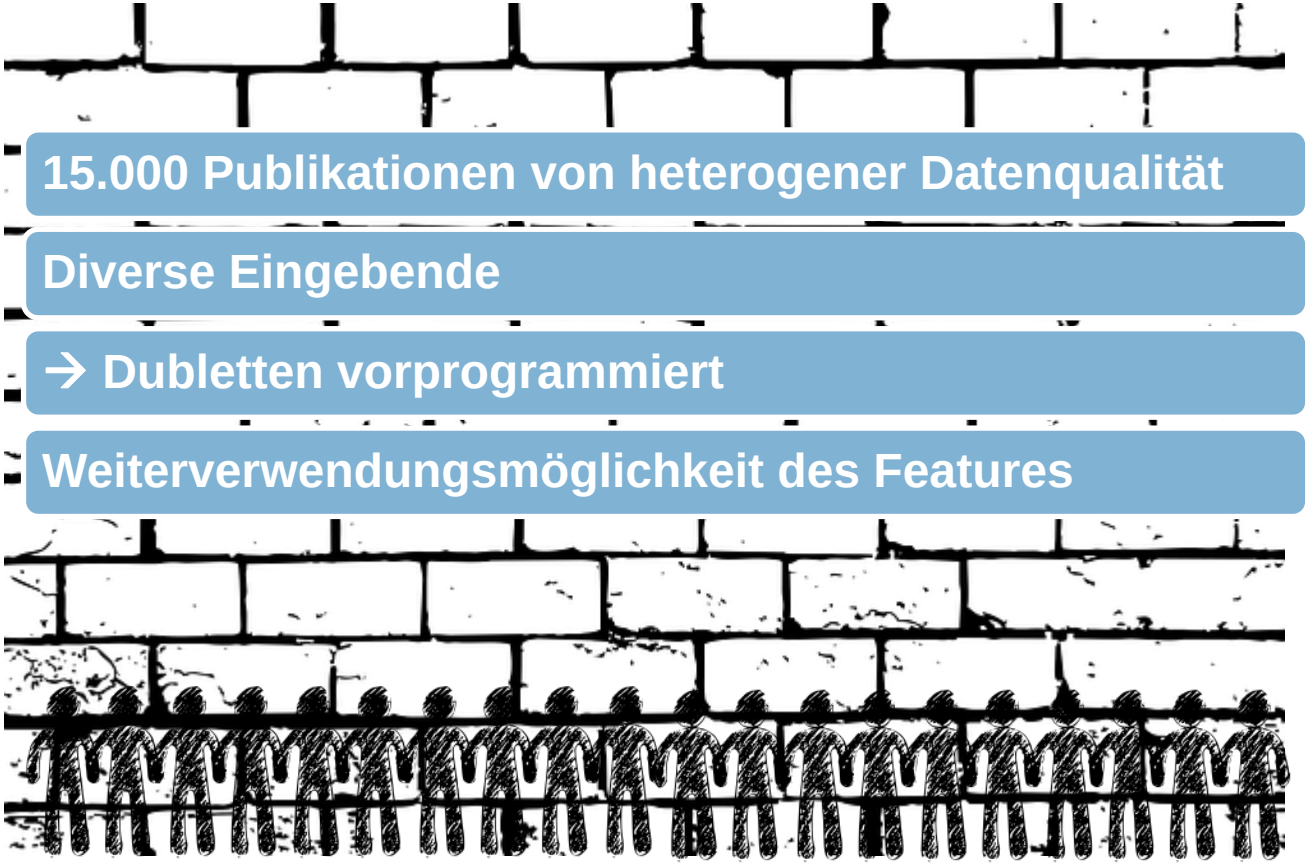
Migrationsprojekt

- 15.000 Einträge aus ehemaliger Publikationsdatenbank
- Qualität sehr heterogen
- Nicht migrierbare Titel vorhanden
- Dubletten vorhanden
- Zeit für Detailanalyse beschränkt
- Diverse Autorinnen und Autoren an Migration beteiligt

ZHAW digitalcollection *nach* Migration

- Weitere > 12.000 Einträge aus der Publikationsdatenbank
- Grundlage für **Hochschulbibliographie** und damit des **Reportings**

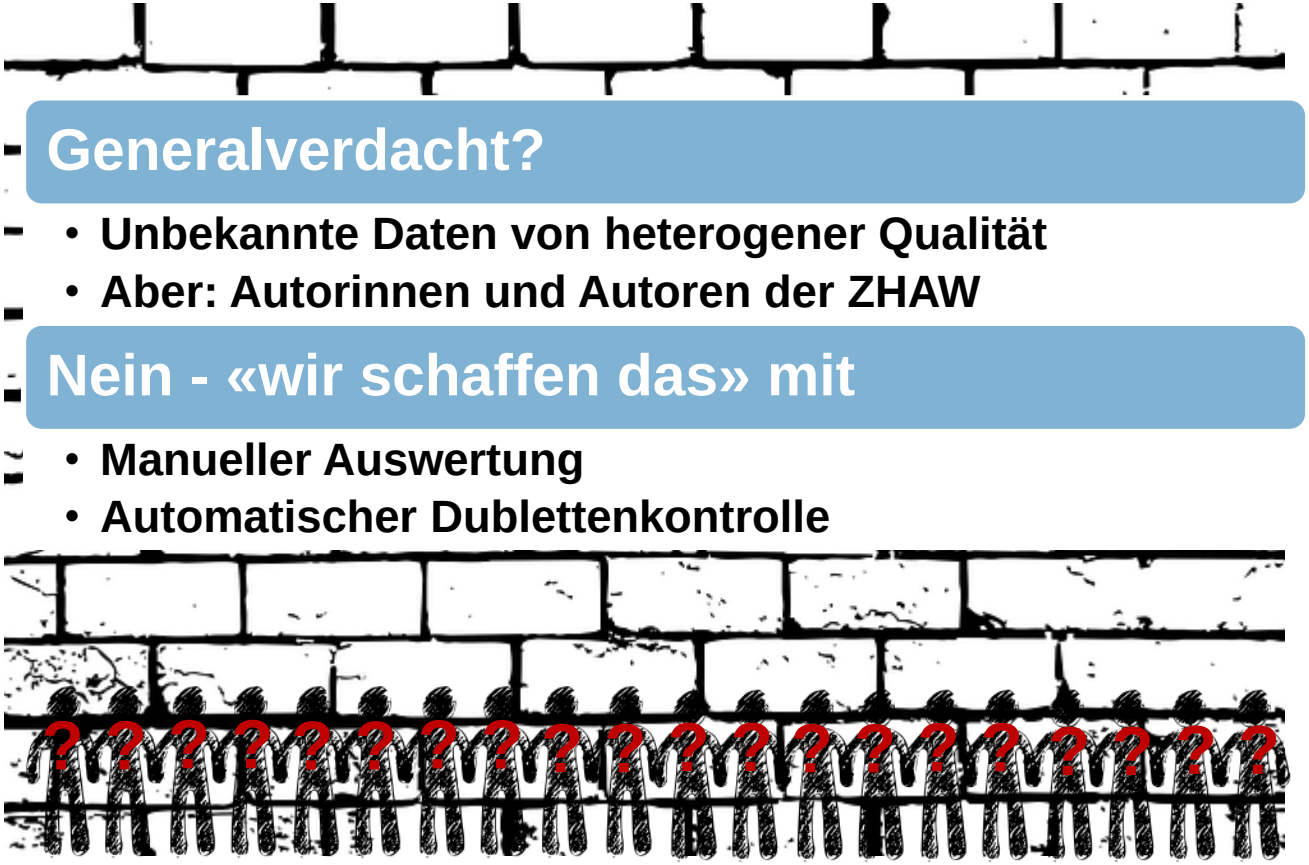
Migration: weshalb Dublettenkontrolle?



Mauer: https://pixabay.com/vectors/bricks-wall-brick-wall-background-2246406/

Menge: https://svgsilh.com/image/1458869.html

Migration: Generalverdacht?



Generalverdacht?

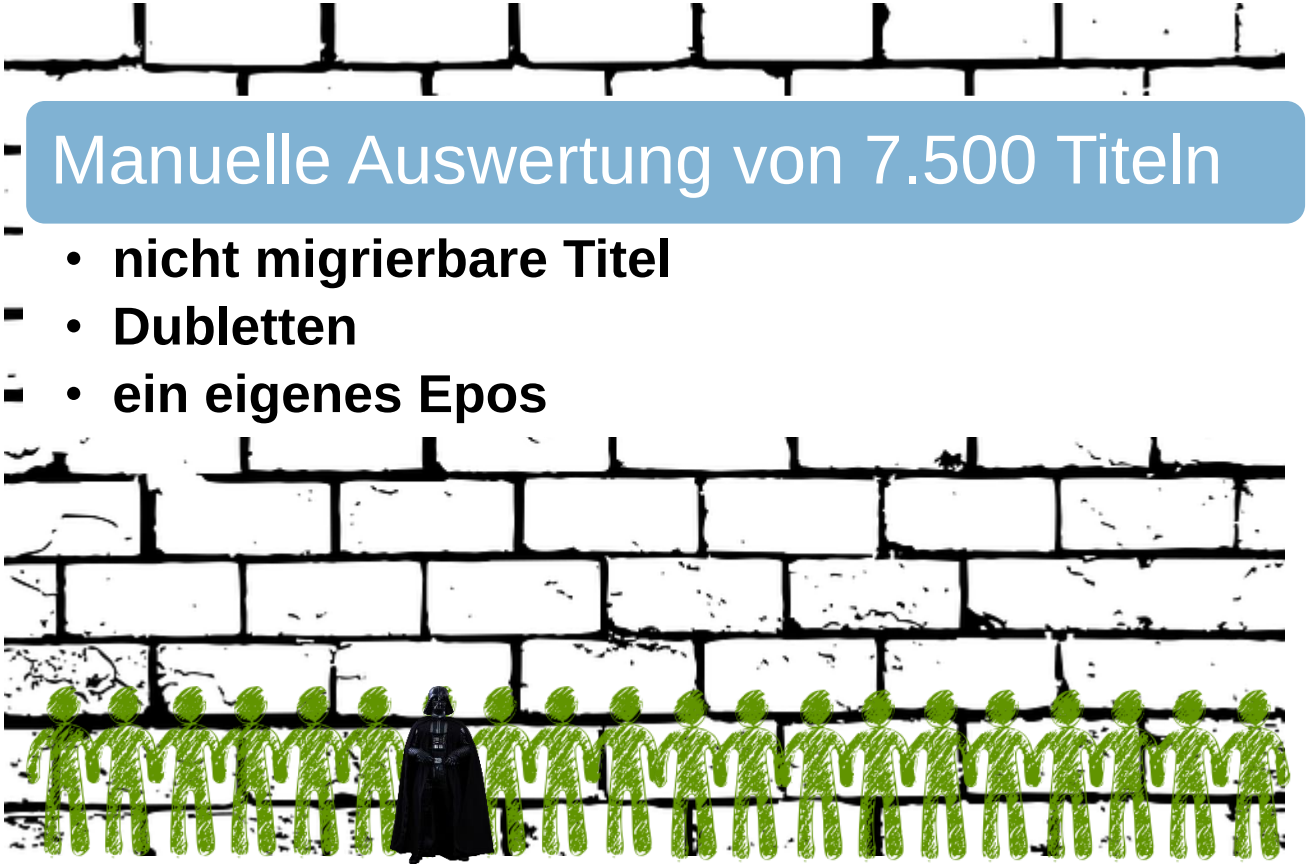
- Unbekannte Daten von heterogener Qualität
- Aber: Autorinnen und Autoren der ZHAW

Nein - «wir schaffen das» mit

- Manueller Auswertung
- Automatischer Dublettenkontrolle

Migration: manuelle Auswertung vor Eingabe

nicht migrierbar



Darth Vader: http://pngimg.com/download/28330

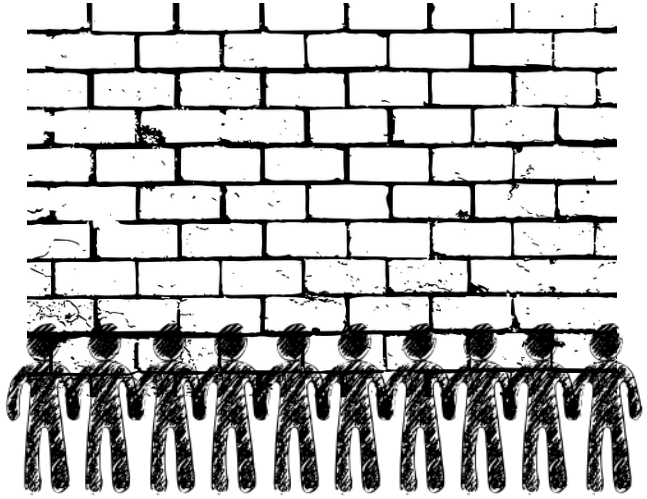
Migration: automatische Dublettenerkennung während Eingabe

Clone
Dublette

Clone: http://pngimg.com/download/28300

Was würde Yoda tun?

KEINE GENERALVERDACHT…



… SONDERN DUBLETTENKONTROLLE

Yoda: https://gamebanana.com/sprays/37154

Workflow Eingabe Publikation

Eingabe

Dublettenkontrolle

Qualitätskontrolle

Sacherschliessung

Planung Feature Dublettenkontrolle

Anforderungen

- Integration in bestehende Workflows (Datenübernahme/manuelle Eingabe)
- Frühe Dublettenwarnung (via Titelfeld)
- Einfache Dublettenerkennung (via Titelfeld)
- Übersichtliche Darstellung möglicher Dublette
- Bei erkannter Dublette: Möglichkeit für Abbruch, Speichern, Ignorieren
- Dublettenkontrolle für End-User UND Mitarbeiter (Berechtigungen!)
- Texte sollen an üblicher Stelle angepasst/übersetzt werden können
- Integration in Applikation/Nutzen der vh. Funktionalitäten in DSpace
- Nachnutzbarkeit für andere DSpace-Anwender

Kriterien Beauftragung

- Finanzierung aus Projektgeldern
- Vergabe an „akkreditierten“ DSpace Entwickler
- Lizenzierung (DSpace Source Code BSD License)
- Dokumentation
- Bereitstellung für Nachnutzung

Dublettenkontrolle

Bitte überprüfen Sie ob der Titel richtig ist oder passen Sie ihn gegebenenfalls an. Im nächsten Schritt wird geprüft ob bereits Einträge mit einem ähnlichen Titel vorhanden sind. Falls wir mögliche Dubletten finden, sehen Sie eine entsprechende Liste. Sie können dann entscheiden, ob Sie Ihre Veröffentlichung fortsetzen möchten oder nicht.

Titel * Interne PR-Arbeit als Instrument der Internen Kommunikation

Abbrechen/Speichern | Weiter >

Dublettenkontrolle

Wir haben einen Eintrag in der digitalcollection gefunden, der Ihrem Titel ähnelt (Interne PR-Arbeit als Instrument der Internen Kommunikation). Bitte entscheiden Sie, ob es sich bei dem gelisteten Eintrag bereits um Ihre Publikation handelt. Falls ja, brechen Sie die Einreichung ab. Falls nein, setzen Sie die Einreichung fort.

- Szyszka, Peter, 2006. **Interne PR-Arbeit als Instrument der internen Kommunikation** [Beitrag in Magazin oder Zeitung]. Eintrag öffnen

Eintrag abbrechen | Speichern, ich entscheide später | Weiter, kein Duplikat

Live Demo

Workflow Dublettenkontrolle

Zürcher Hochschule
für Angewandte Wissenschaften

ZHAW-Nutzer/-in erfasst Publikation

- Auswahl der Sammlung
- Eingabe des Titels
- Dublettenkontrolle bereits veröffentlichter Publikationen

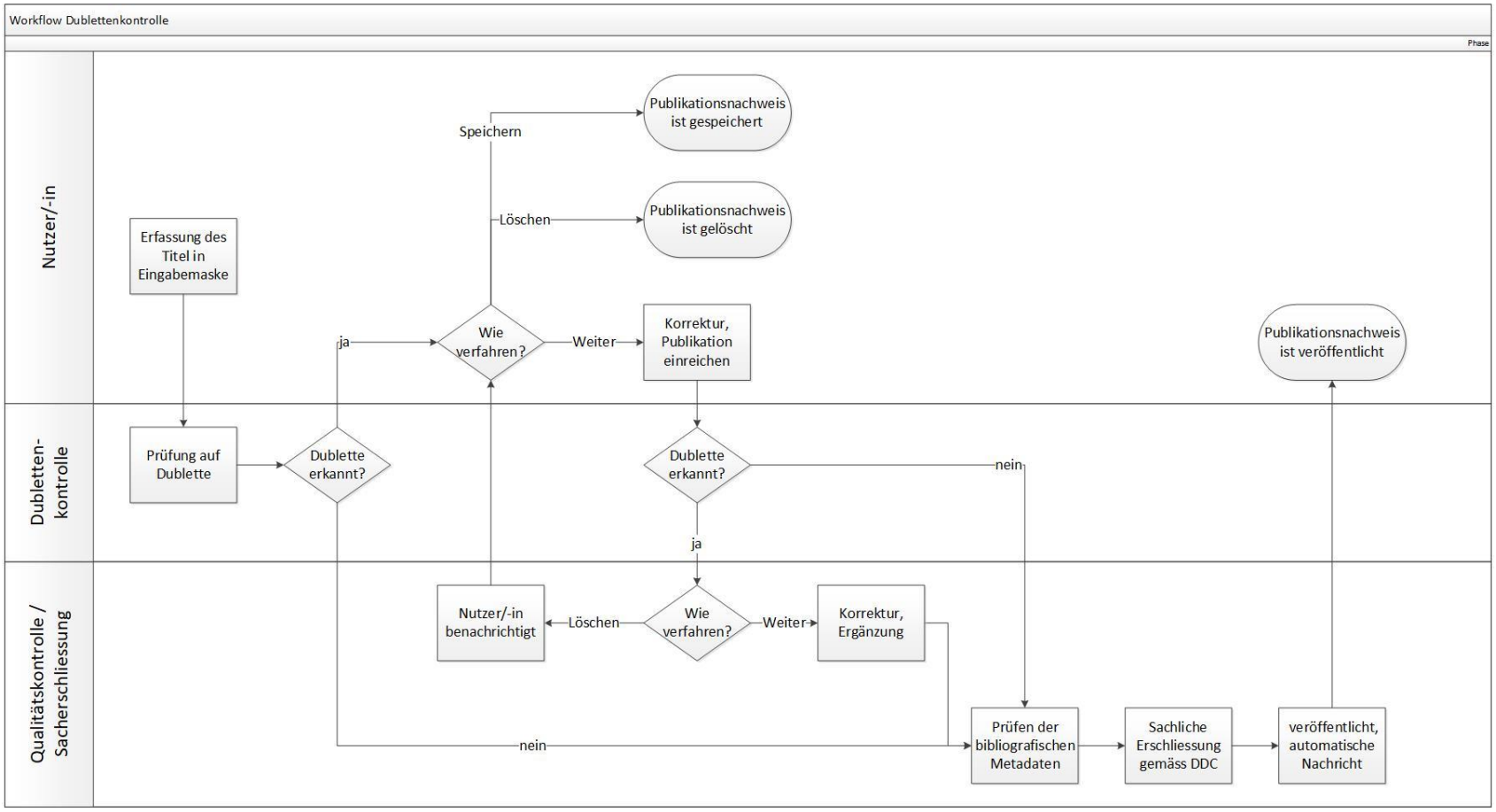
Schritt Qualitätskontrolle / Sacherschliessung

- Dublettenkontrolle bei Übernahme Aufgabe und Eingabe Titel
- Veröffentlichte Publikationen
- Publikationen im Geschäftsgang

Dublette?

- Keine Dublette: weiter im Workflow
- Mögliche Dublette wird per Link in neuem Tab geöffnet
- Prüfung durch Submitter/Bearbeiter/in
- Entscheidung: Abbruch, Speichern, Ignorieren

Workflow Dublettenkontrolle (Visio)



Zürcher Hochschule
für Angewandte Wissenschaften

Technische Implementation Dublettenkontrolle

Funktion

- Neues Eingabefeld für Titel geschaffen
- Step „Dublettenkontrolle“ im Workflow konfiguriert
- Überprüfung des Titels anhand des Levensthein Distance Algorithmus (Übereinstimmung ≤ 8 Zeichen inkl. Leerzeichen)

Voraussetzungen

- PostgreSQL Extension „fuzzystrmatch“ ist implementiert
- DSpace 6.2/6.3, JSPUI (nicht XMLUI)

Konfigurierbarkeit

- Levenshtein-Distanz (default = 8)
- Maximale Anzahl der gezeigten möglichen Dubletten (default = 10)
- Welches Feld überprüft wird (default=dc.title)

Zusammenfassung Dublettenkontrolle

Zürcher Hochschule
für Angewandte Wissenschaften

Vorteile

- Reduktion der Bereinigungsarbeit im Vorfeld bei Migration (schlechter) Daten
- Lediglich Eingabe des Titels nötig
- Übersichtliche Darstellung möglicher Dubletten (Titel, Autoren, Jahr, Publikationstyp)
- Für interne MA sind auch unveröffentlichte Publikationen sichtbar

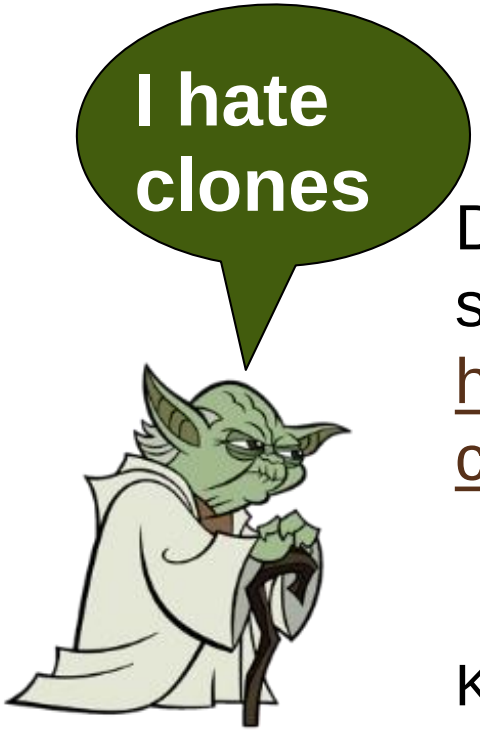
Nachteile

- Kurzer Titel (< 8 Zeichen) findet viele falsche Dubletten
- Jahr und Publikationstyp werden bei Prüfung nicht berücksichtigt

Fazit

- Sehr hilfreich (insb. bei vielen Co-Autorenschaften und bei Migration)
- Erkennt Dubletten zuverlässig und schnell
- Übersichtliche Darstellung möglicher Dublette erleichtert schnelle Prüfung

Vielen Dank für Ihre Aufmerksamkeit & Danke an das Team Publikationsdienste



Das kostenlose Feature Dublettenkontrolle steht frei zur Verfügung:
https://github.com/the-library-code/deduplication

Kontakt:
Friederike Gerland, Iris Hausmann
digitalcollection@zhaw.ch

Powerpoint-Foliengestaltung: Sabine ☼
Bestes Migrationsteam ever: Adrian, Clemens, Debbie, Iris, Kerstin, Sabine ☼