Datenqualität in Bibliothekskatalogen

Michel Piguet – Josef Wandeler
Trialog AG

Congrès Bibliosuisse, 30.08.2018, Montreux

TRIALOG

TRIALOG

Der Anlass

- Zeitschriftenbestände mussten aus fünf Bibliotheken abgeglichen werden. Dazu mussten die Daten ausshalb der Verbundsysteme zusammengeführt werden.

- **Schritt A:** Aufgrund der MARC-Daten aus verschiedenen Katalogen wurden die Duplikate erkannt.

- **Schritt B:** Aus den Bestanddaten wurde ermittelt, welche Bibliothek den vollständigeren Bestand hatte.

Erfolg mit einem Aber

Schritt A (Titelabgleich):

- Duplikate liessen sich erst aufgrund von elaborierter Datenabgleichtechniken erkennen.

Resultat:

- Von 181’000 Titelaufnahmen wurden ca. 40’000 als Duplikate erkannt.

Aber:

- Eine sichere Zuordnung gelang zu etwa 90%.

➔ Es war also zusätzlich intellektuelle Arbeit nötig, um Zweifelsfälle zu bearbeiten.

TRIALOG

Bestandsabgleich

Für **Schritt B** (Bestandsabgleich) waren die Bestandsdaten erst aufgrund linguistischer Erkennungstechnik brauchbar.

- Die Angaben sind für Lesewesen geschrieben, nicht für Maschinen.
- Es gibt mehrere hundert Schreibweisen, wie Holdings notiert werden.
- Oft hiess es «unvollständig» - eine ungenügende Information.

Fazit:

- Erst mittels Datenanalyse war der Abgleich erfolgreich, aber nicht vollständig.

  Und, wen wundert's? Die Angaben entsprachen nicht immer dem physischen Bestand.

➔ Es war zusätzlich Arbeit im Magazin nötig, um die Angaben zu überprüfen.

TRIALOG

Es blieben ungute Gefühle ...

So viel Informationen stehen in bibliographischen Datensätzen!!

Dennoch ...

- ... konnten die Zeitschriften nur teilweise mit Sicherheit zugeordnet werden.
- ... konnten die Bestände nur teilweise aus den Daten ermittelt werden.

Warum liess sich die maschinelle Erkennung nicht einfach verbessern? **Woran liegt dies?**

TRIALOG

Eigentlich könnten viele Fragestellungen beantwortet werden …

Wie zum Beispiel

- Welche Titel sind in mehreren Ausgabeformen vorhanden (auf Papier und digital)?
- Wie entwickelte sich die Erscheinungsweise von Zeitschriften?

… wenn die Qualität stimmen würde…

TRIALOG

Warum lässt sich eine maschinelle Erkennung nicht einfach verbessern?

Kleiner Exkurs:

- Was ist Datenqualität? Über welche Aspekte sprechen wir?
- Was leistet die Katalogisierung? - Was nicht?
- Was für Daten braucht es? Wofür?

TRIALOG

Dimensionen der Datenqualität

Nach: Pipino, Lee, & Wang, 2002

Dimensions	Definitions
Accessibility	The extent to which data is available, or easily and quickly retrievable
Appropriate Amount of data	The extent to which the volume of data is appropriate for the task at hand
Believability	The extent to which data is regarded as true and credible
Completeness	the extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Concise Representation	The extent to which data is compactly represented
Consistent Representation	The extent to which data is presented in the same format
Ease of manipulation	The extent to which data is easy to manipulate and apply to different tasks
Free-of Error	The extent to which data is correct and reliable
Interpretability	The extent to which data is in appropriate languages, symbols, and units, and the definitions are clear
Objectivity	The extent to which data is unbiased, unprejudiced, and impartial
Relevancy	Hie extent to which data is applicable and helpful for the task at

Die hier relevante Aspekte

Für den Datenabgleich interessieren hier folgende Aspekte:

	Dimension	Erklärung
1	Vollständigkeit	Daten fehlen in keinem Datensatz
2	Konsistente Darstellung	Daten werden immer im gleichen Format dargestellt
3	Übersichtliche Darstellung	Daten werden kompakt dargestellt
4	Einfache Handhabung	Daten sind einfach zu manipulieren und auf verschiedene Aufgaben anzuwenden

TRIALOG

Befunde

In den MARC-Datensätzen befanden sich nur wenige Felder, die sich für einen Abgleich eignen.

- Felder sind ungenügend gefüllt
  ➔ 1 - «Vollständigkeit».
- Sie enthalten uneinheitliche Daten
  ➔ 2 - «Konsistente Daten»
- Oft sind Informationen in Bemerkungsfeldern abgelegt
  ➔ 3 - «übersichtliche Darstellung»
- Die Daten sind für Leser, nicht für Maschinen
  ➔ 4 - «Einfache Handhabung»

UND: Ein Datensatz kann eine, mehrere oder nur einen Teil eines Zeitschften-Objekts beschreiben (Bezugseinheit).

TRIALOG

1 - Vollständigkeit

Feld	Bezeichnung	Füllgrad
008_	Fixfeld (Erscheinungsjahre)	100%
022$a	ISSN	40%
245$a	Titelangabe	100%
245$b	… Zusatz	40%
245$c	… Verfasserangabe	40%
260$a	Publikation - Ort	100%
260$b	… Verlag	90%
310$a	Erscheinungsweise	40%
362$a	Erscheinungszeitraum	80%
710$a	Herausgebende Körperschaft	50%

TRIALOG

2 - Konsistente Daten

Beispiel:
Feld 310$a (= Erscheinungsweise)

Quizfrage: ?

In wie vielen Varianten wurde eine «vierteljährliche» Erscheinungsweise formuliert?

TRIALOG

Antwort

«vierteljährlich»

409 Varianten

TRIALOG

4 Mal pro Jahr
4 x/Jahr
4x jährl.
4xJahr
4xjährlich
4 Hefte im Jahr
Erscheint 4mal jährlich
Erschien 4mal jährlich
Published quarterly
Pubbl. 4x/anno
Publ. cuatrimestral
Publ. quarterly
Pubbl. Quadrimestrale
...

4 iss. per year
4 times a year
4x/an
Quarterly

Vierteljährlich mit jährlicher Kumulation
4x/Jahr + Spezialausg.
Vierteljährlich ; ab 1992 unregelmässig
Monatlich im Internet, vierteljährlich in gedruckter Form
4-6 mal jährlich ; anfangs: vierteljährlich
...

- unklare Form
- unterschiedliche Sprachen
- Überladen mit weiterer Info
- Tippfehler

TRIALOG

Katalogisieren versus Systematisieren

- Devise in der Katalogisierung ist:

  Die Informationen werden so erfasst, wie sie vorgefunden werden.

- Systematisiert wird (nur) dort, wo der Datensatz mit Normdaten verknüpft wird.

Weniger ist mehr.

1) Je **mehr Datenfelder** verfügbar sind,
   - ➔ desto mehr werden die Informationen darin verteilt
   - ➔ desto geringer ist der “Füllgrad” der Felder
   - ➔ desto weniger lässt sich die Information verwerten

2) Je **variantenreicher** eine Information geschrieben wird,
   - ➔ desto weniger kann die Information genutzt werden.
   - ➔ desto wertloser wird sie (für Automaten).

3) Je mehr in **Bemerkungsfeldern** abgelegt wird,
   - ➔ desto weniger ist klar, wo sich eine Information befindet.
   - ➔ desto mehr vermischen sich verschiedene Information.

TRIALOG

Was sind daraus für Schlüsse zu ziehen?

- Das MARC-Format hat zu viele Datenfelder.
- Normalisierung von Daten ist essentiell.
- Pflichtfelder mit genau definiertem Inhalt sind wesentlich.

Hilft RDA?

TRIALOG

Wird mit RDA alles besser?

- Teilweise ja:
  Verknüpfungen mit der GND sind ein Stück Normalisierung – das hilft.

Aber:

- Die bibliografische Beschreibung systematisiert auch in Zukunft nicht
- Das “Cataloguers Judgment” schafft tendenziell noch mehr Vielfalt
- RDA gilt nur für neue Katalogisate

TRIALOG

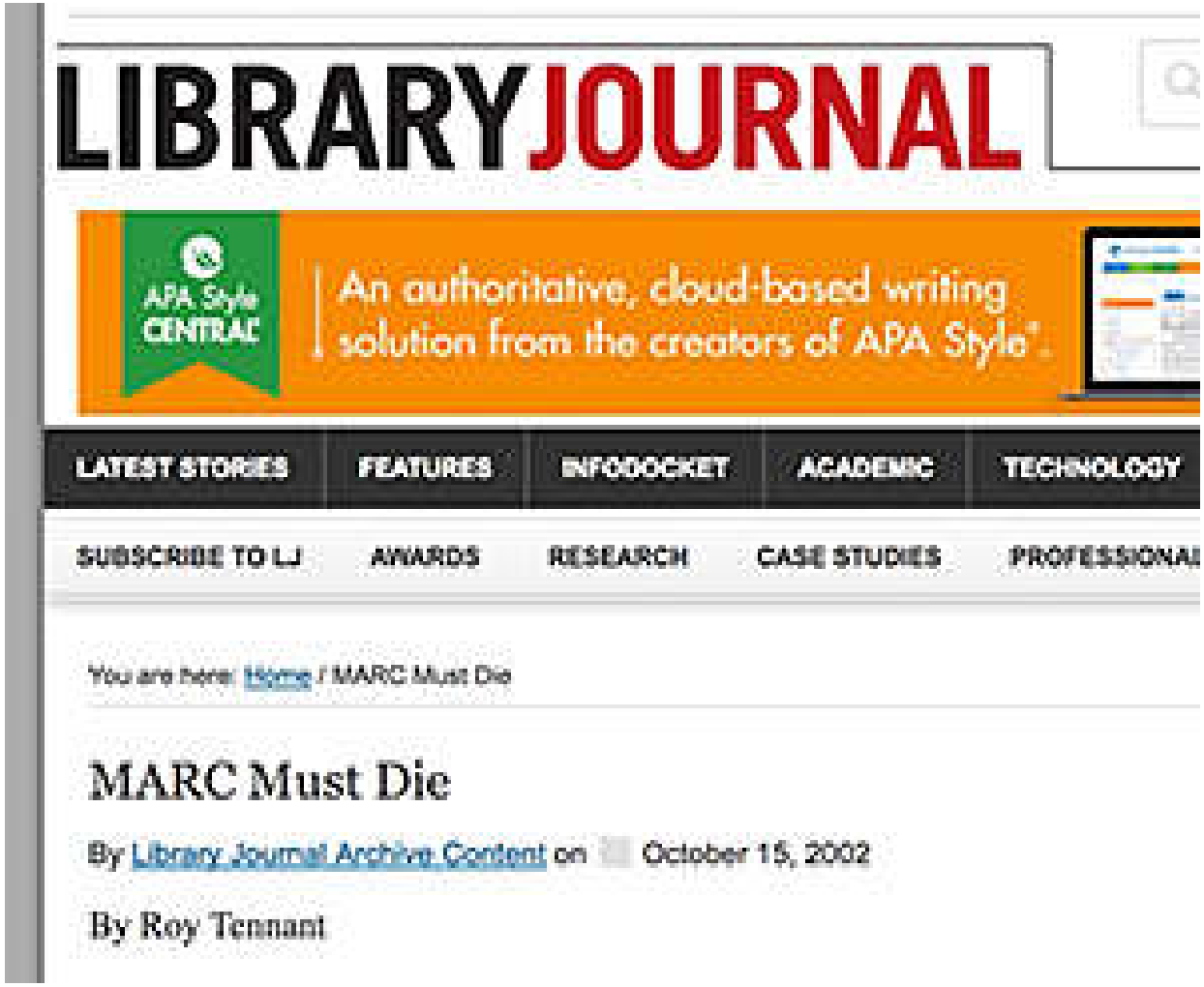
Was dann?

- Muss (noch) mehr Aufwand in die Katalogisierung gesteckt werden?
- Müssen die Daten besser bereinigt werden?

oder …

TRIALOG

Oder



BIBFRAME als Alternative?

Das von der Library of Congress propagierte Format ist eine Chance.

TRIALOG

Mögliche Wege in die Zukunft

- Computer sind «lernfähig» geworden.
- Maschinen unterstützen immer besser.

Wie können Bibliotheken diese Entwicklung nutzen?

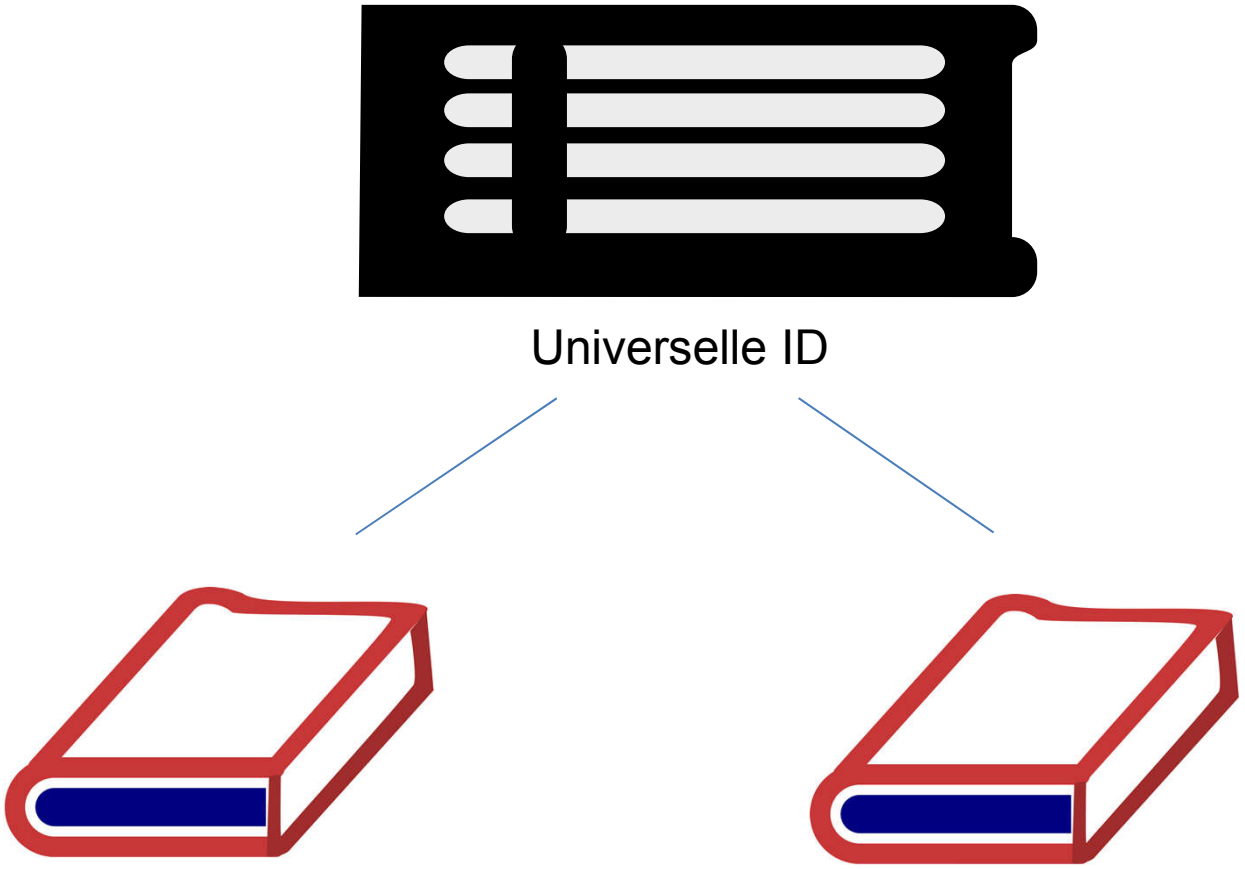
1. Eindeutige Objekt-Identifikation
2. Metadaten extrahieren statt abschreiben
3. Automatisierte Verlinkung

Und:

Auf Suchtechniken vertrauen

TRIALOG

Identifikation



TRIALOG

Identifikation

- Die (bibliografischen) Identitätsmethoden sind nur beschränkt geeignet, um weltweit über eine universelle ID zu verfügen (ISBN, EAN, LCCN, Worldcat OpenURL und all die zahlreichen Permalinks von Servern).

Moderne ID-Verfahren

- Bild- bzw. musterorientierte ID’s
- Beispiele, die Sie kennen:
  - Etikettenerkennung (Vivino, ...)
  - Gesichtserkennung (Zutrittsberechtigung, Fotoclustering z.B. in Google Picasa)
  - Plagiatserkennung
  - selbst: Musikerkennung (Shazam, SoundHound, ...)

Diese Verfahren werden im Bibliotheksbereich nicht angewendet.

WARUM NICHT?

Zukunft: Identifikation

- Grundidee: Das Abbild eines eindeutigen Teils eines Werkes (Titelseite + Impressum, ...) kann es identifizieren.
- ➔ bildet eine universelle ID

Voraussetzungen:

- Die «Titelseiten» wurden digitalisiert und sind abrufbar.
- Optional: Ein offener Algorithmus erzeugt daraus eine eindeutige Kennung (ID).

TRIALOG

Zukunft: Extraktion von Metadaten

Katalogdaten werden automatisiert übernommen, gegebenenfalls korrigiert.

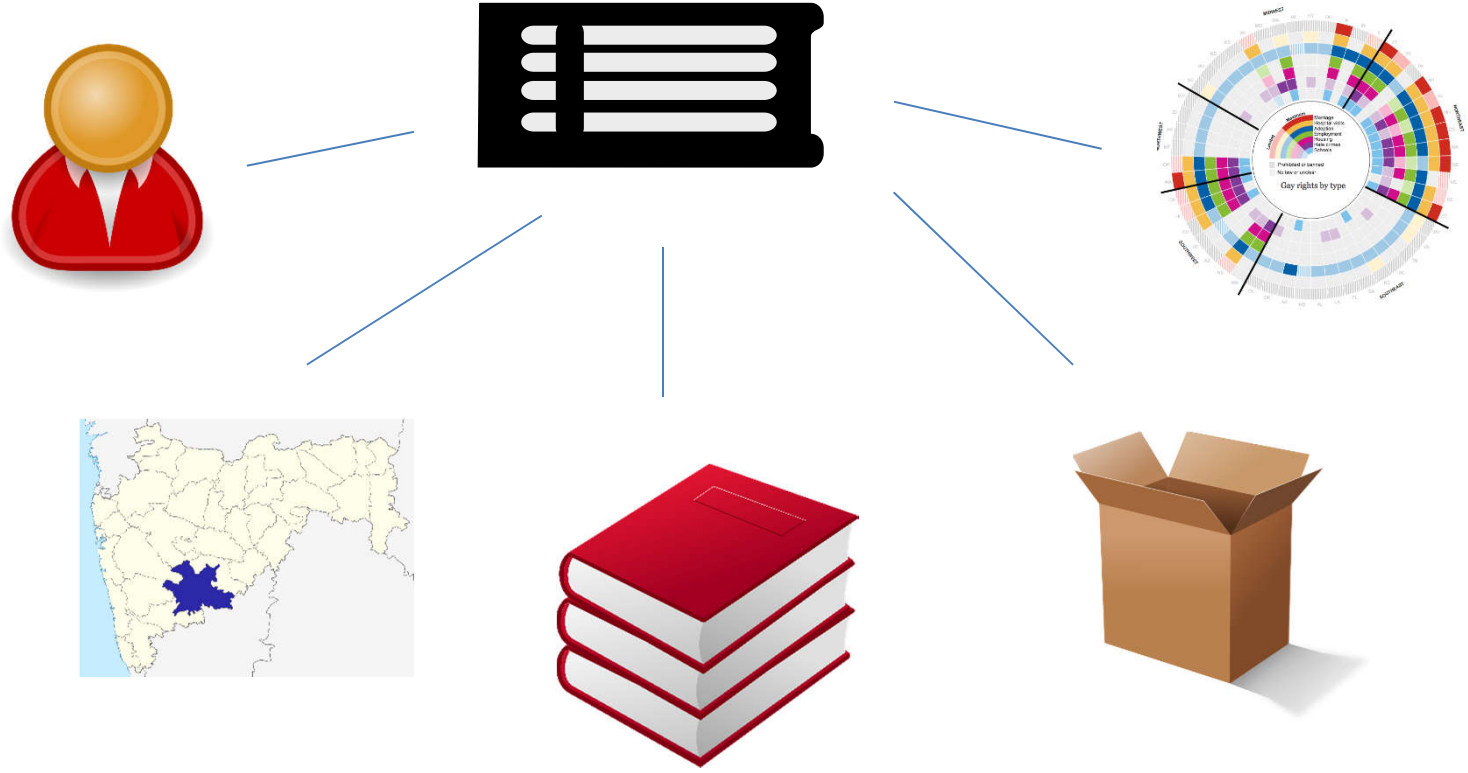
Bsp: moderne Dokumenten-Management-Systeme

- Automatische Verarbeitung von Belegen
  - Elektronische Belege werden analysiert
  - Das System lernt, wo welche Metadaten zu finden sind z.B. Belegdatum, Firmenname (Kunde, Lieferant), etc.
  - Metadaten werden ausgelesen und in die Datenbank abgelegt
  - Endkontrolle und Ergänzung statt manuelle Erfassung

TRIALOG

Zukunft: Verlinkung

Ein Objekt wird mit verlinkten Normdaten beschrieben – und zugleich angereichert.



TRIALOG

Der Weg zu mehr Qualität (Fazit)

1. Universelle Identifikation ist Grundlage.
2. Standardisierung von Metadaten schafft einheitlichere Daten.
3. Verlinkung zu universalen Diensten validiert.

➔ bibliothekarische Formalerfassung + plus ID-Generierung

Was bleibt vom bisherigen MARC-Datensatz?

- Beschreibende Daten (ohne Normdaten)
- Lokalinformationen
- Bemerkungen

TRIALOG

Danke für Ihre Aufmerksamkeit und das Mitdenken!

Ihre Fragen?

Ihre Erfahrungen?

Ihre Einwände?

Die Präsentation finden sie auf:
www.trialog.ch → Vorträge und Artikel

TRIALOG