Nicht immer ist der SOLR schuld!

Qualitätskontrollen bei Metadaten und Optimierungsmöglichkeiten

Viktor Holzwert, Steffen Illig, Stefan Philipp, **Philipp Rumpf** | Universitätsbibliothek Bamberg

DSpace Praxistreffen

04.–05.04.2024 | Mainz

Die Otto-Friedrich-Universität Bamberg

- seit 1979 (wieder) Universität
- 4 Fakultäten
- ca. 12.000 Studierende
- in der Weltkulturerbestadt Bamberg

Praxistreffen

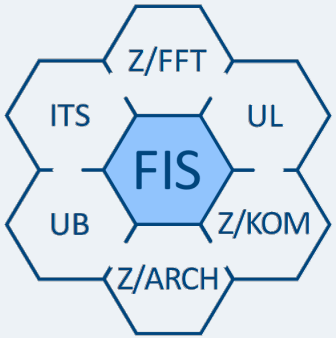
04.–05.04.2024
Mainz

DSpace in Bamberg – fis.uni-bamberg.de

2016	Entscheidung der Universität Bamberg für DSpace-CRIS
2018	Beitritt zum DSpace-Konsortium-Deutschland
2018	Go-Live mit DSpace-CRIS 5.10
2023	Umstieg auf DSpace-CRIS 2023.01.01 (DSpace 7.5)
2024	In Vorbereitung: DSpace-CRIS 2023.02.02 (LTS, DSpace 7.x) Fokus auf: Stabilisierung des Systems

Das FIS ist ein Kooperationsprojekt:

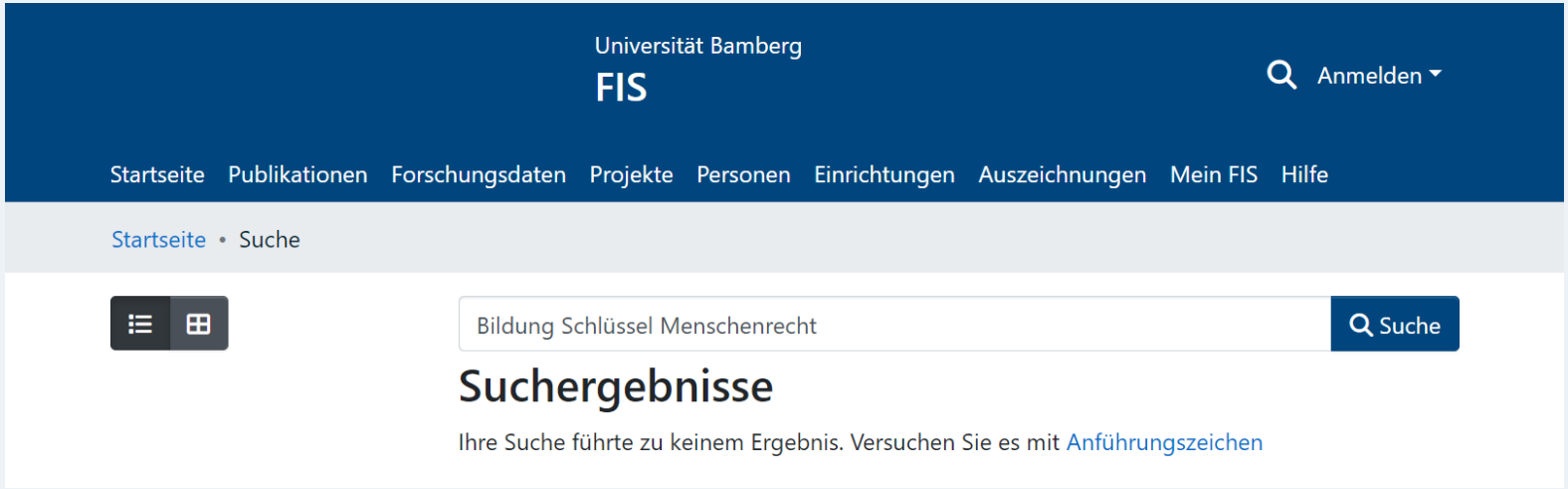
- Forschungsförderung (Z/FFT)
- IT-Service (ITS)
- Universitätsbibliothek
- u.a.



es beinhaltet:

- Bibliografie
- Publikationsserver
- ...

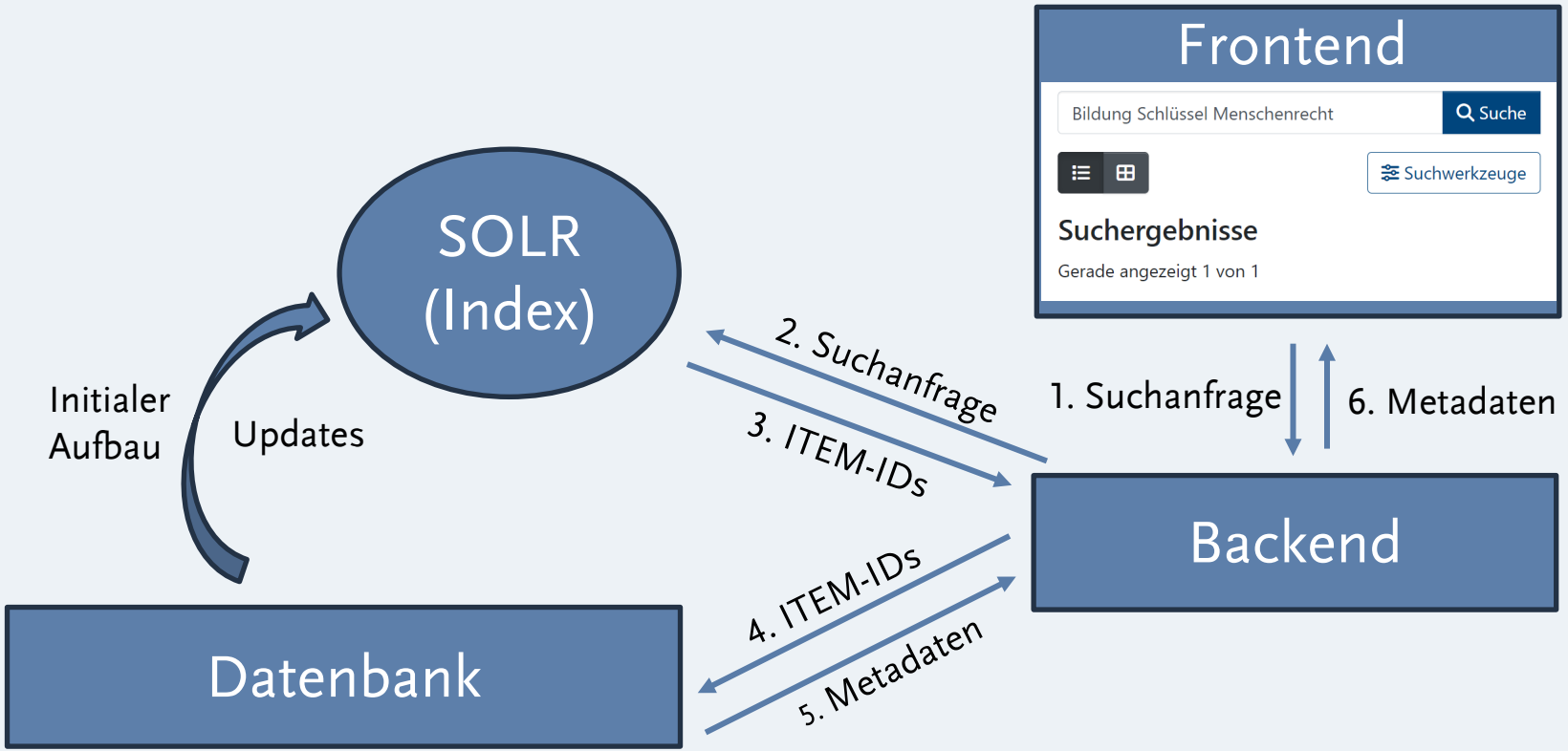
Problemstellung



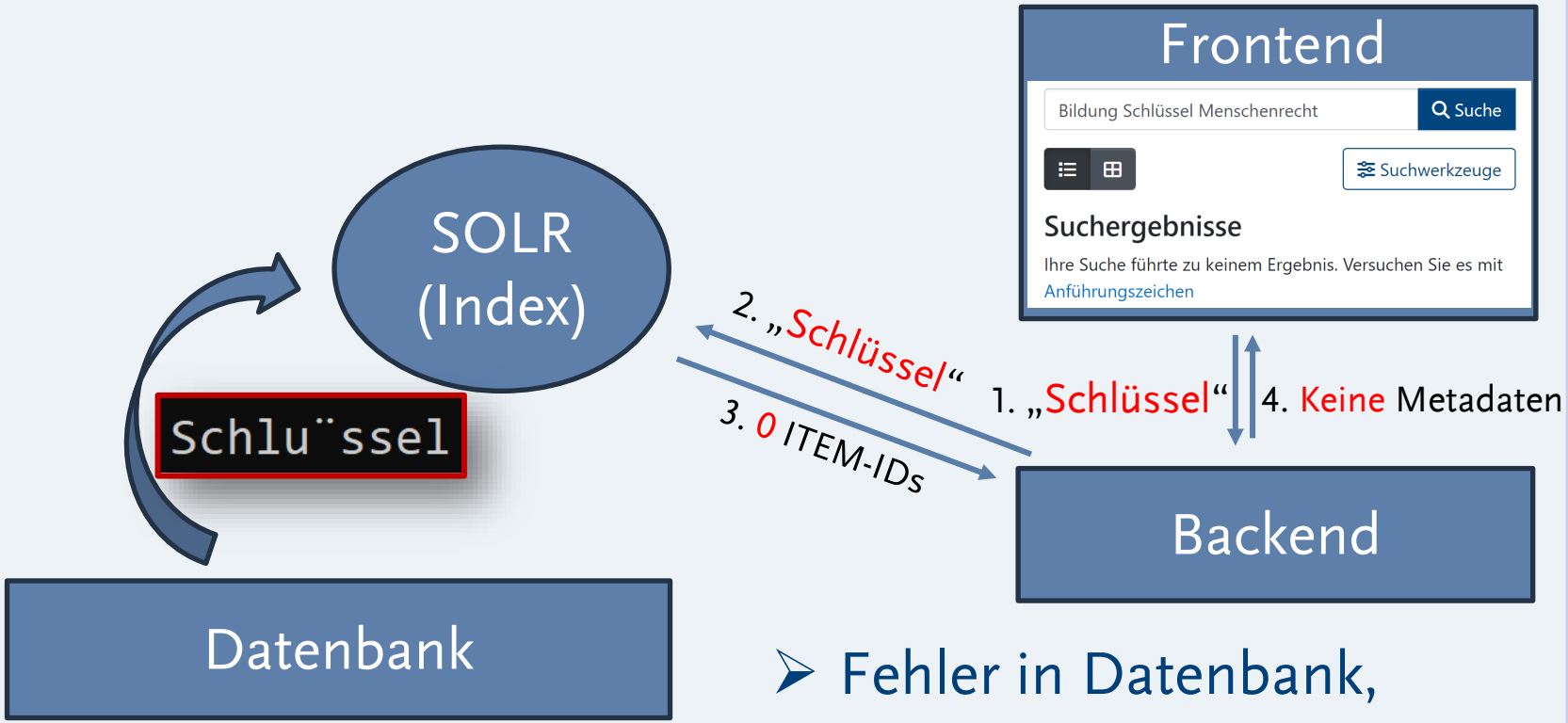
Suche im FIS führt nicht zum gewünschten Treffer, obwohl das ITEM existiert.

Warum? 🧐

Funktionsweise des SOLR in DS 7 (vereinfacht)



Funktionsweise des SOLR in DS 7 (vereinfacht)



- Fehler in Datenbank, führt zu Fehler im SOLR

Verschiedene Darstellungen in Unicode
(trotz UTF-8 als Standard)

Beispieltitel: Bildung ist der entscheidende Schlüssel

1. Aktuell durch copy-paste „Schlüssel“
→ ü = [75 cc 88] u + Umlaut über vorherigem Buchstaben

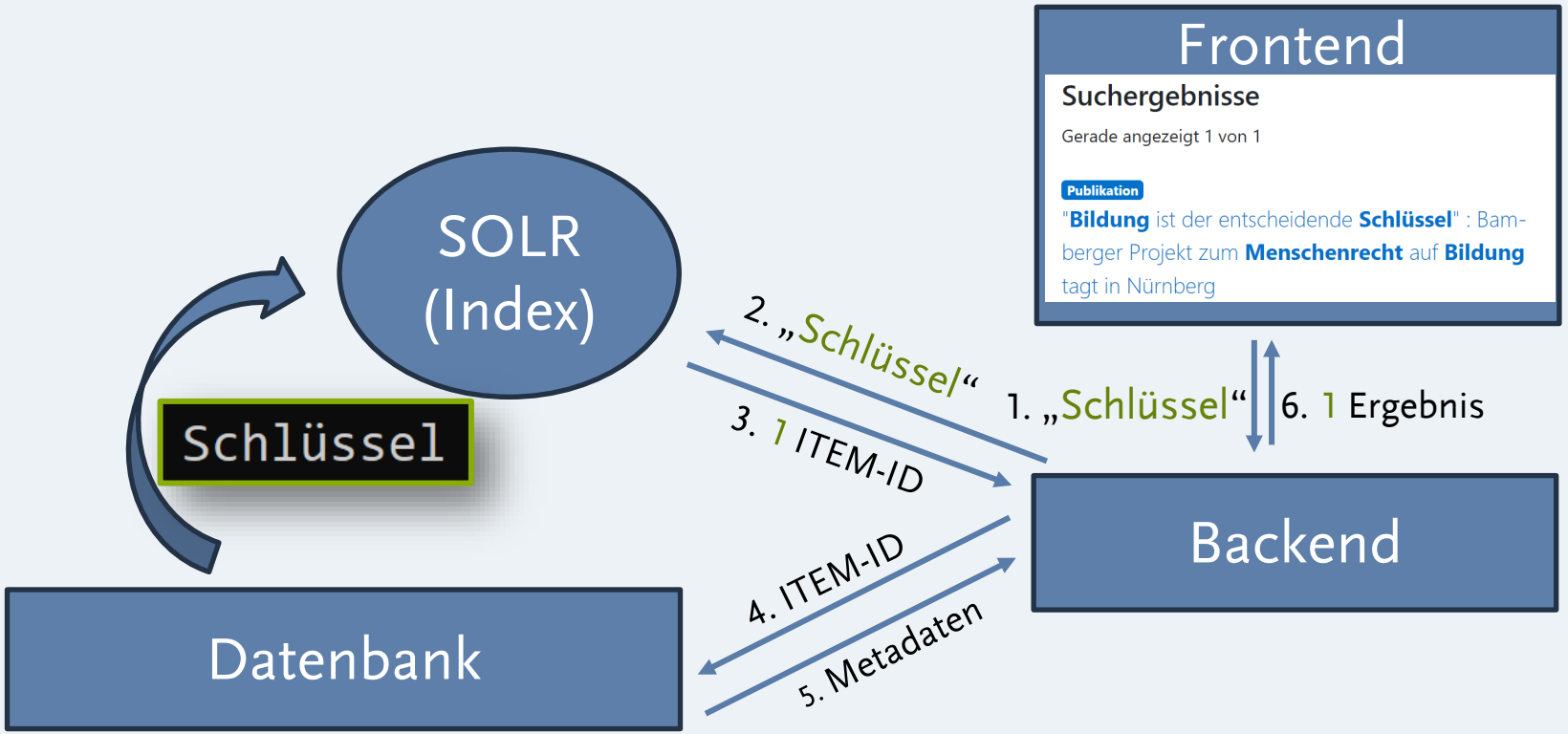
der	entscheidende	Schlüssel
[64 65 72]	[65 6e 74 73 63 68 65 69 64 65 6e 64 65]	[53 63 68 6c 75 cc 88 73 73 65 6c]

2. Eingabe über Tastatur (deutsch) „Schlüssel“ → ü = [c3 bc]

der	entscheidende	Schlüssel
[64 65 72]	[65 6e 74 73 63 68 65 69 64 65 6e 64 65]	[53 63 68 6c c3 bc 73 73 65 6c]

➔ Selbes Problem bei ö und ä

Funktionsweise des SOLR in DS 7 (vereinfacht)



Weiteres Fehlerbild: Leerzeichen

- Durch falsche Nutzereingabe: Ein oder mehrere Leerzeichen zu viel
- Durch copy-paste: Andere Leerzeichenarten

➢ Gut wäre ein einzelnes, einheitliches Leerzeichen.

\u0020		space
\u00A0		no-break space
\u2000		en quad
\u2001		em quad
\u2002		en space
\u2003		em space
\u2004		three-per-em space
\u2005		four-per-em space
\u2006		six-per-em space

\u2007		figure space
\u2008		punctuation space
\u2009		thin space
\u200A		hair space
\u202F		narrow no-break space
\u205F		medium mathematical space
\u3000		ideographic space
\u0009		character tabulation
\u1680		ogham space mark

\u0020		space

Besonders gemein

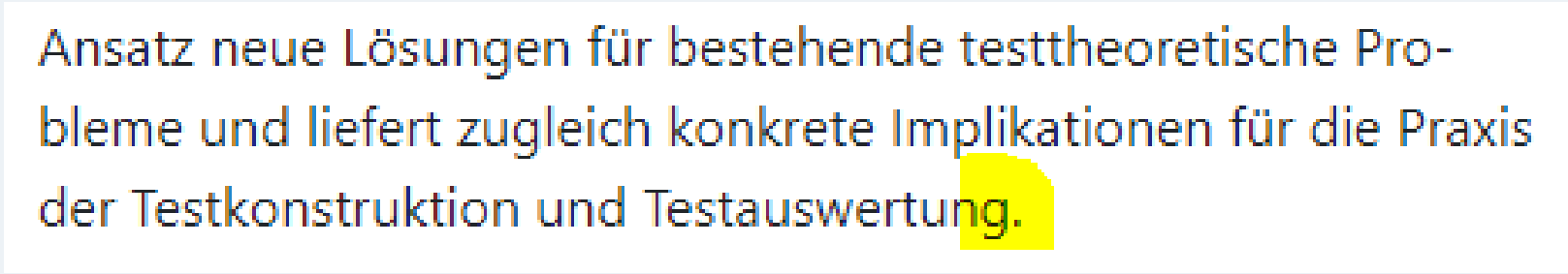
Manche Leerzeichenarten wie das geschützte Leerzeichen (non-breaking space) sind auch in der Konsole nicht als falsch zu erkennen:

Towards the Evaluation of Action Reversibility in STRIPS Using Domain Generators

```
-Towards the Evaluation of Action Reversibility in STRIPS Using Domain Generators
+Towards the Evaluation of Action Reversibility in STRIPS Using Domain Generators
```

```
Towards the Evaluation of Action Reversibility
in STRIPS Using Domain Generators
```

Weiteres Fehlerbild: Breitenlose Leerzeichen
(zero width space)



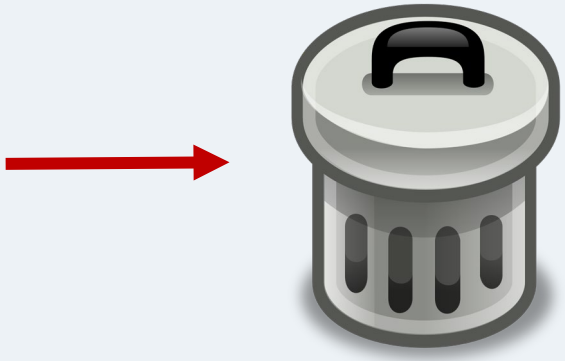

```
Implikationen für die Praxis der Testkonstruktion und Testauswertung.
Implikationen für die Praxis der Testkonstruktion und Testauswertung.
```

```
Implikationen für die Praxis der Testkonstruktion und
Testauswertung.&ZeroWidthSpace;
```

Weiteres Fehlerbild: Breitenlose Leerzeichen (zero width space)

Der SOLR erkennt zwei Worte, wenn das Zeichen in einem Wort vorkommt:
Auflage → Auf+lage

\u200B	zero width space
\u200C	zero width non-joiner
\u200D	zero width joiner
\u2060	word joiner
\uFEFF	zero width non-breaking space



Weiteres Fehlerbild: Unterschiedliche Zeilenumbrüche

`\u000B`	line tabulation
`\u000C`	form feed
`\u0085`	next line
`\u2028`	line separator
`\u2029`	paragraph separator

Ausreißer

`\u000A`
line feed

Linux
macOS

`\u000D\u000A`
carriage return + line feed

Windows

Weiteres Fehlerbild: Unterschiedliche Zeilenumbrüche

Beispiele:

for focused disciplinary and interdisciplinary research projects.
Our approach presents a novel concept for data federation in

```
interdisciplinary research projects.\nOur approach pr
interdisciplinary research projects.\r\nOur approach
```

are responsible for reducing obsessive-compulsive symptoma-
tology. After an initial exploration of the literature, I hypothesi-

```
symptomatology. LSAfter an initial
symptomatology.\r\nAfter an initial
```

Wir analysieren unsere Datenbank ...

... mit einem selbstentwickelten Skript für **DSpace 7**

➢ **DSpace Pull Request #9427**

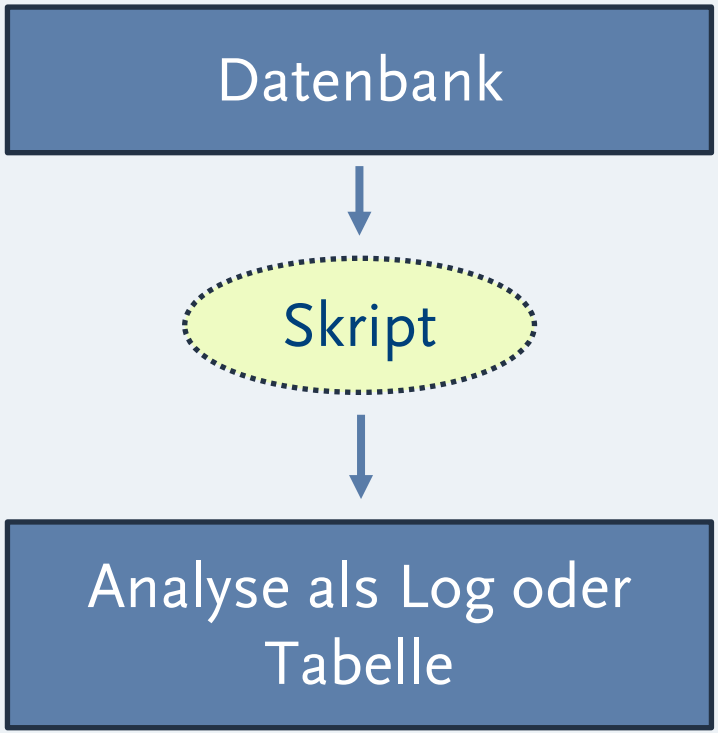
Das Skript verwendet die Java Klasse „Enhancer“ (nicht die Enhancer Funktionalität von DS-CRIS) und kontrolliert in unserer Datenbank folgende Felder:

- dc.title
- dc.title.alternative
- dc.description.abstract
- (weitere Felder sind möglich)

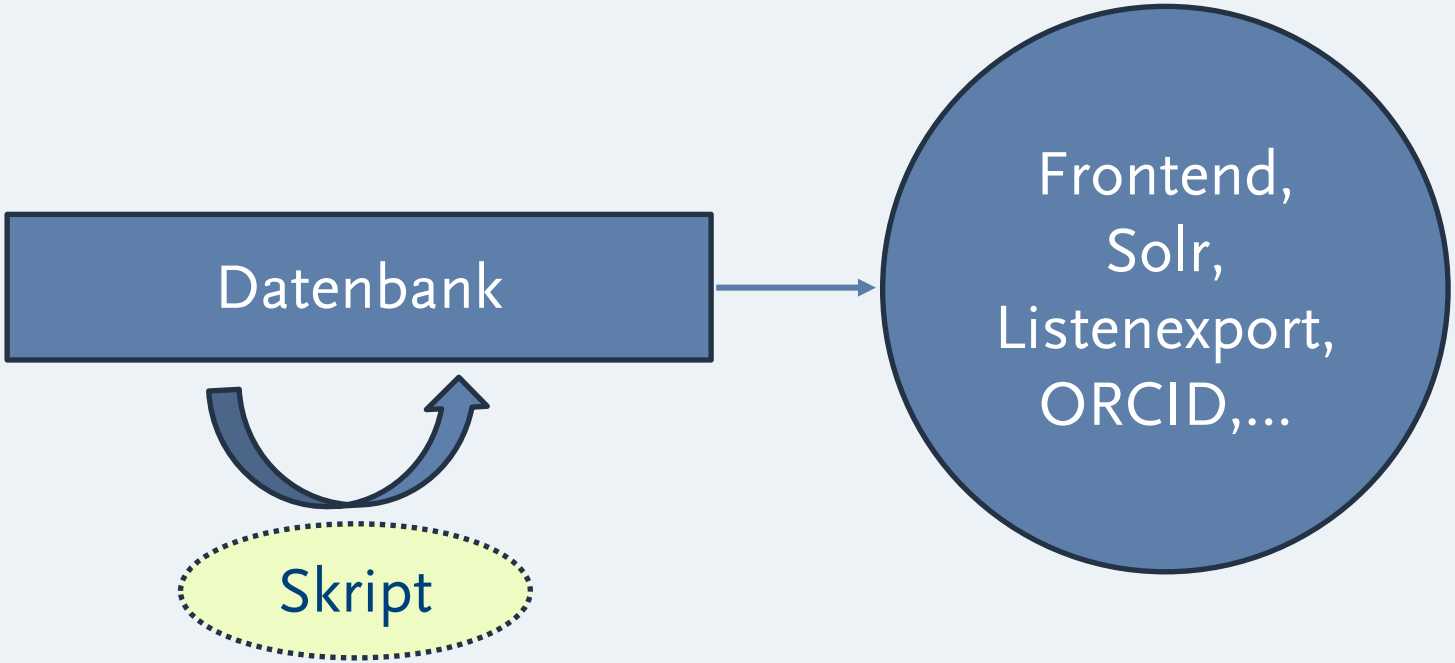
Das Skript kann:

- Eine Ergebnisliste anzeigen => Für manuelle Prüfung
- Fehler automatisch bereinigen => wenn gewünscht

Das Skript analysiert ...



… und korrigiert auf Wunsch (optional)



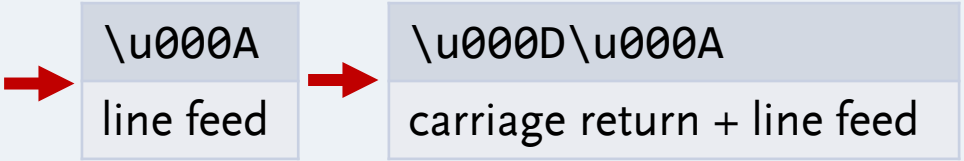
Zum Beispiel: Unterschiedliche Zeilenumbrüche

- Zeilenumbrüche werden vom Skript in 2 Schritten korrigiert:

1. Alle Zeichen, die Zeilenumbrüche markieren, werden durch Zeilenvorschub (englisch: line feed) ersetzt.
2. Zeilenvorschübe werden zu Zeilenumbrüchen (Windows Standard) kombiniert

\u000B	line tabulation
\u000C	form feed
\u0085	next line
\u2028	line separator
\u2029	paragraph separator

Ausreißer



Linux
macOS

Windows

Lessons learned: Fehler in unserer Datenbank

Datensätze mit:	dc.title	dc.title.alternative	dc.description.abstract
Gesamtzahl der Datensätze	59.836	986	5.545
Betroffen	1.126	9	738
Zeilenvorschübe			52
Spezielle Zeilenumbrüche			2
Leerzeichen vor dem Zeilenumbruch			417
Multiple Leerzeichen	1.075	9	250
Spezielle Leerzeichen	43		118
Umlaute	11		18

- Das Bamberger FIS ist Bibliografie und Publikationsserver.
- Bei Einträgen mit Dateien (mit Abstract) häufen sich die Fehler. Größeres Problem bei Publikationsservern!

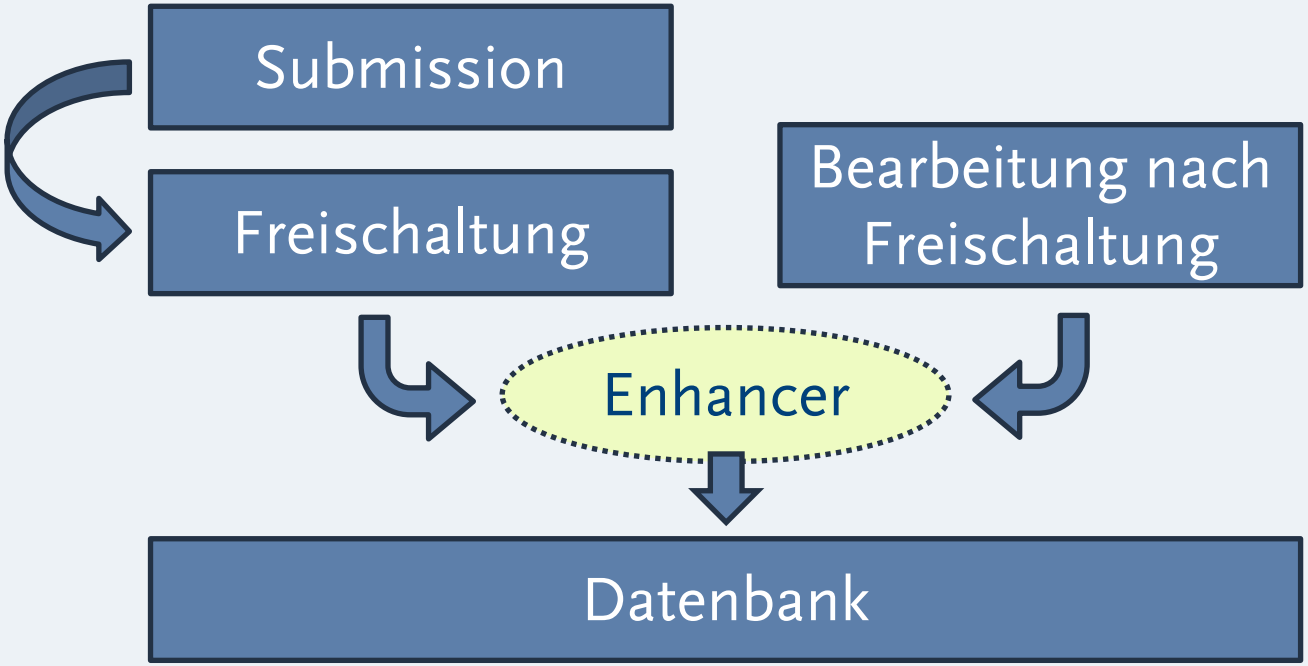
Ist mit dem Skript das Problem gelöst?

- Die Fehler in unserer Datenbank sind bereinigt. Wir zeigen die Metadaten korrekt an und geben diese korrekt weiter.
- Es werden aber weiterhin neue Daten aus unterschiedlichen (fehlerbehafteten) Quellen per copy-paste eingegeben.

➢ Das Problem kehrt wieder. 

Lösung: Wir nutzen die DSpace-CRIS-Funktionalität des „Enhancers“!

Der Enhancer korrigiert (anders als das Skript) fortlaufend ab Freischaltung



DSpace-CRIS Pull Request #444

Ist ein Enhancer für DSpace 7 denkbar?

Anders als DSpace-CRIS kann DSpace keine Enhancer ausführen.

Es braucht dazu einen neuen Listener.

Dieser Listener würde unseren Enhancer nachnutzen.

Freiwillige gesucht:

- ➔ Issue eröffnen
- ➔ PR schreiben
- ➔ Feedback der Community erbitten/einarbeiten.

Vielen Dank für die Aufmerksamkeit!
Sind noch Fragen offen?