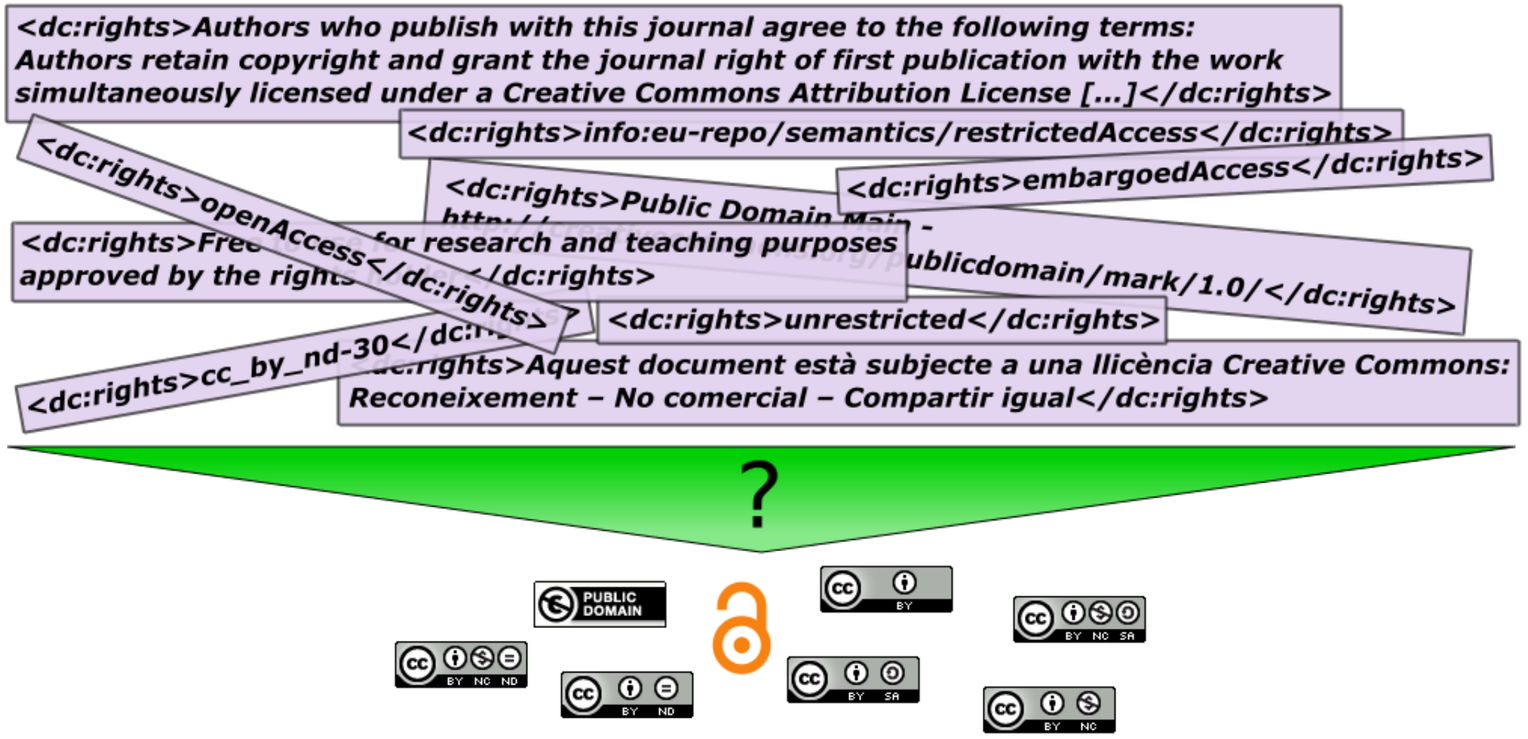
Normalisierung von Lizenzinformationen in OAI-Metadaten: Ein Beitrag zur Verbesserung der Open-Access-Statusanzeige in wissenschaftlichen Suchmaschinen



Christoph Broschinski, <broschinski@uni-bielefeld.de>

Universität Bielefeld
Information.plus! Universitätsbibliothek
BASE Bielefeld Academic Search Engine

Einführung

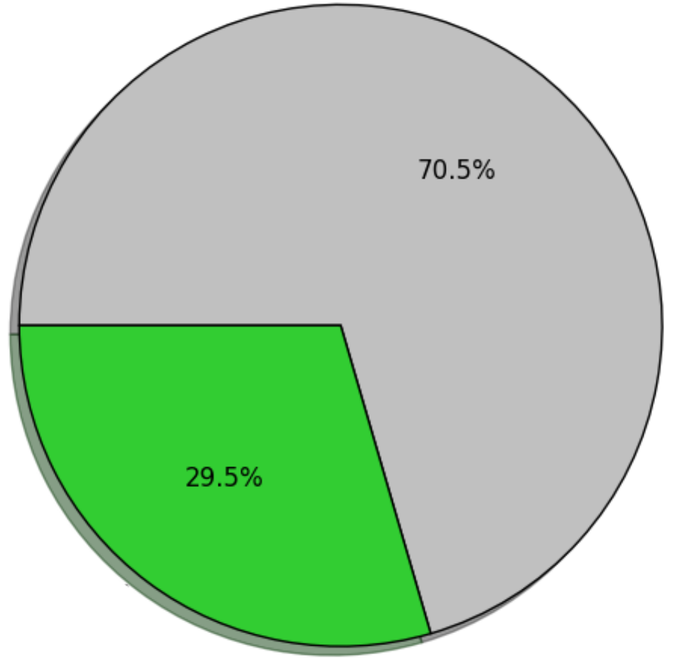
BASE: Bielefeld Academic Search Engine (https://base-search.net), seit 2004 entwickelt und betrieben von der Universitätsbibliothek Bielefeld

- Suchmaschine für wissenschaftliche Dokumente und Publikationen
- Fokus auf Open-Access-Publikationen/frei zugängliche Inhalte
- Derzeit knapp 90 Mio Dokumente aus über 4000 Quellen indexiert (Beispiel)

Quellen

- Quellen für BASE sind beispielsweise Publikationsserver an Hochschulen und Instituten, Online-Journals, Aggregationsdienste
- Beispiel 1: OPUS FAU - Online-Publikationssystem der Friedrich-Alexander-Universität Erlangen-Nürnberg (5307 Datensätze)
- Beispiel 2: ETH E-Collection (Dokumentenserver der ETH Zürich) (30865 Datensätze)
- Regelmäßiges Aggregieren von Metadaten über das Protokoll OAI-PMH (Beispiel)

Open Access in BASE - bisheriger Stand



Problem: "Echter" Open Access-Anteil liegt deutlich höher (60-70%, bekannt aus regelmäßigen Stichproben)

- 30-40% Open Access-Dokumente sind nicht entsprechend markiert
- Keine Auszeichnung von "Nicht-OA"-Dokumenten, keine automatische Erkennung von weitergehenden Lizenzen (Creative Commons, gemeinfreie Inhalte)

Grund: Bisherige Auswertung von OA-Status ausschließlich auf Quellen-Ebene ("reine" OA-Quellen/OAI-Sets)

Neuer Ansatz: Individuelle Normalisierung

Seit Mitte 2015: Lizenzinformationen werden auf der Artikelebene für jedes Dokument einzeln klassifiziert

- Quellen in BASE werden über OAI-PMH geharvestet, das Format ist dabei üblicherweise Dublin Core (DC)
- Dublic Core enthält ein optionales Feld "dc:rights", das Informationen über Nachnutzungsrechte des Datensatzes enthalten kann.
- **Problem:** Feldinhalte sind völlig uneinheitlich (Art der Information, Notation, Sprache...)

Beispiel für dc:rights-Inhalte aus BASE:

- "info:eu-repo/semantics/openAccess"
- "Open Access"
- "The Public Domain Mark (PDM)"
- "local access"
- "Es gilt das UrhG"
- "Alle Rechte vorbehalten"
- "Volltextzugriff: nur innerhalb des Universitäts-Campus"
- "CC-BY-NC-ND 4.0"
- "Aquest document està subjecte a una llicència Creative Commons: Reconeixement – No comercial – Compartir igual"

Neuer Ansatz: Individuelle Normalisierung (2)

Ziel: Normalisierung (Abbildung auf ein festes Vokabular) der Inhalte des DC-Elements "dc:rights". Vorgehensweise:

- Entwicklung einer internen Platform, um Übersicht über dc:rights-Inhalte zu gewinnen
- Implementierung eines mehrstufigen Regelsatzes zur Klassifizierung (reguläre Ausdrücke)

Frage: Auf welche Zielkategorien soll normalisiert werden?

- 1) Gemeinfreie Inhalte/Public Domain
- 2) Creative Commons (mit jeweils exaktem Mapping auf eine der 6 möglichen CC-Lizenzen: CC-BY, CC-BY-SA, CC-BY-ND, CC-BY-NC, CC-BY-NC-SA, CC-BY-NC-ND)
- 3) Open Access
- 4) Kein Open Access
- 5) Unbekannt (kein dc:rights-Feld oder Inhalt nicht normalisierbar)

Grundannahme: Gemeinfreiheit (1) und Creative Commons (2) bedingen implizit Open Access. Daher wird bei diesen Normalisierungen auch der Open-Access-Status in den Metadaten gesetzt.

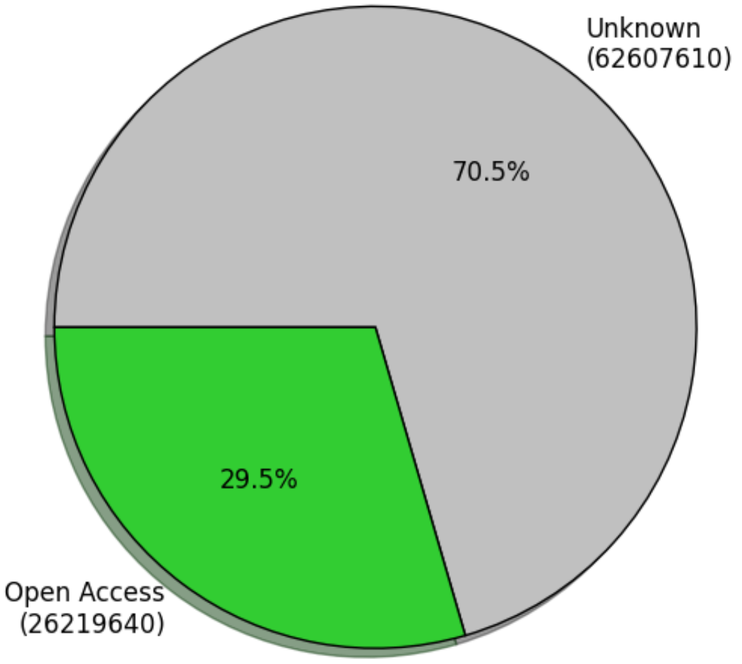
Individuelle Normalisierung: Beispiele

dc:rights	Normalisierung möglich?	Nachnutzung/Lizenz	Open Access
"info:eu-repo/semantics/openAccess"	Ja		Ja
"Open Access"	Ja		Ja
"The Public Domain Mark (PDM)"	Ja	PDM	Ja
"local access"	Ja		Nein
"Es gilt das UrhG"	Nein		Unbekannt
"Alle Rechte vorbehalten"	Nein		Unbekannt
"Volltextzugriff: nur innerhalb des Universitäts-Campus"	Ja		Nein
"CC-BY-NC-ND 4.0"	Ja	https://creativecommons.org/licenses/by-nc-nd/4.0	Ja
"Aquest document està subjecte a una llicència Creative Commons: Reconeixement – No comercial – Compartir igual"	Ja	https://creativecommons.org/licenses/by-nc-sa	Ja

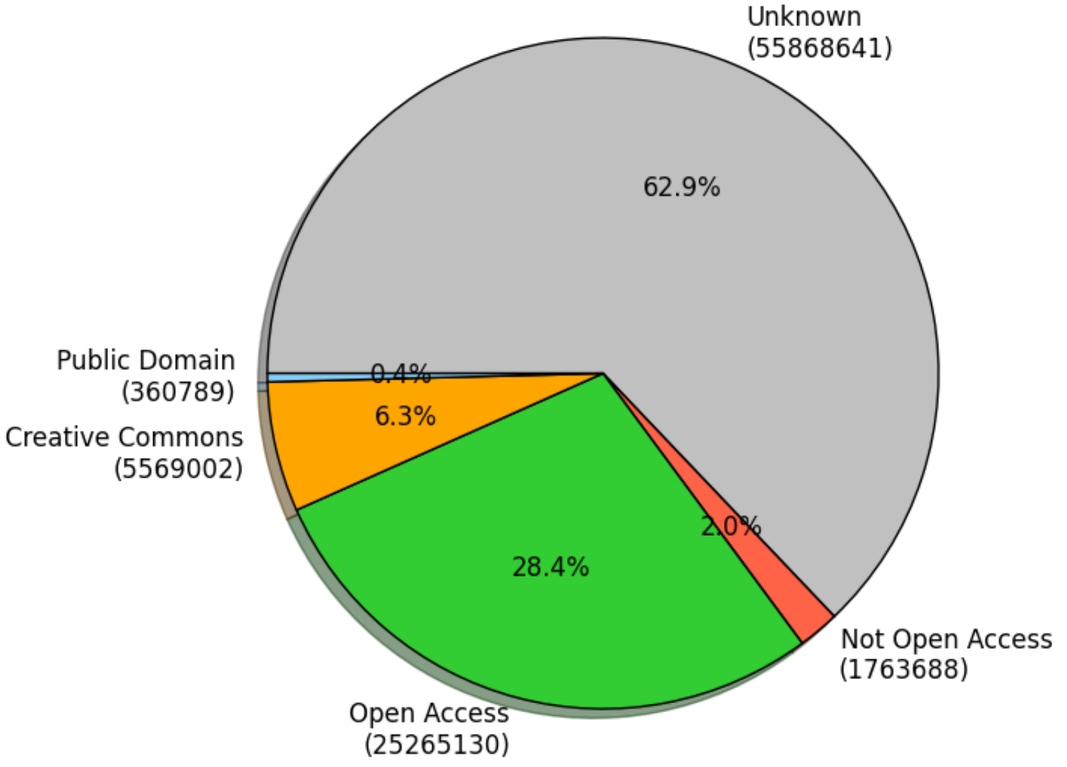
Ergebnisse: Gesamtbestand

Änderungen im Suchmaschinen-Index von BASE durch Rechtenormalisierung (Gesamtdaten, 88.827.250 Dokumente aus 4052 Quellen)

Ohne dc:rights-Normalisierung



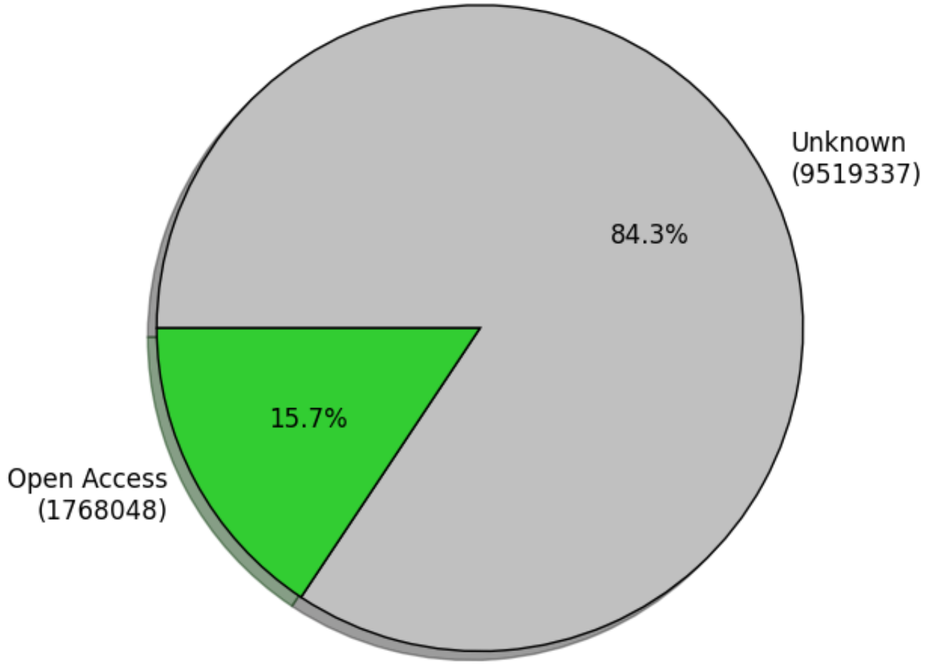
Mit dc:rights-Normalisierung



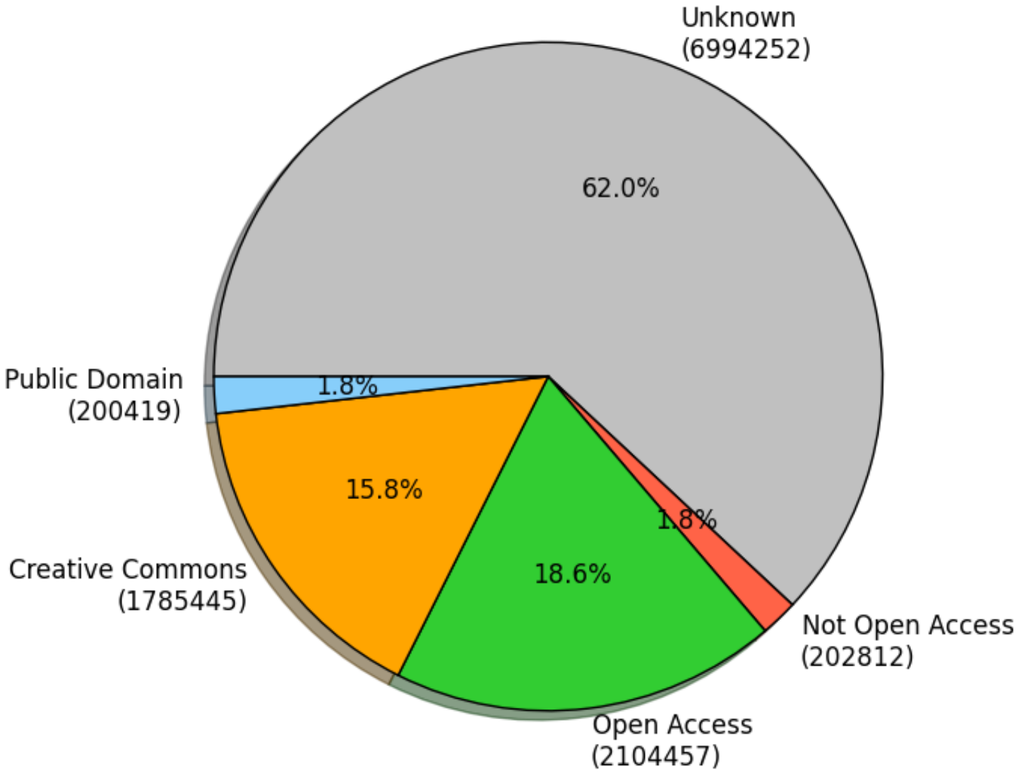
Ergebnisse: Deutschland

Änderungen im Suchmaschinen-Index von BASE durch Rechtenormalisierung (Land: Deutschland, 11.287.385 Dokumente aus 304 Quellen)

Ohne dc:rights-Normalisierung



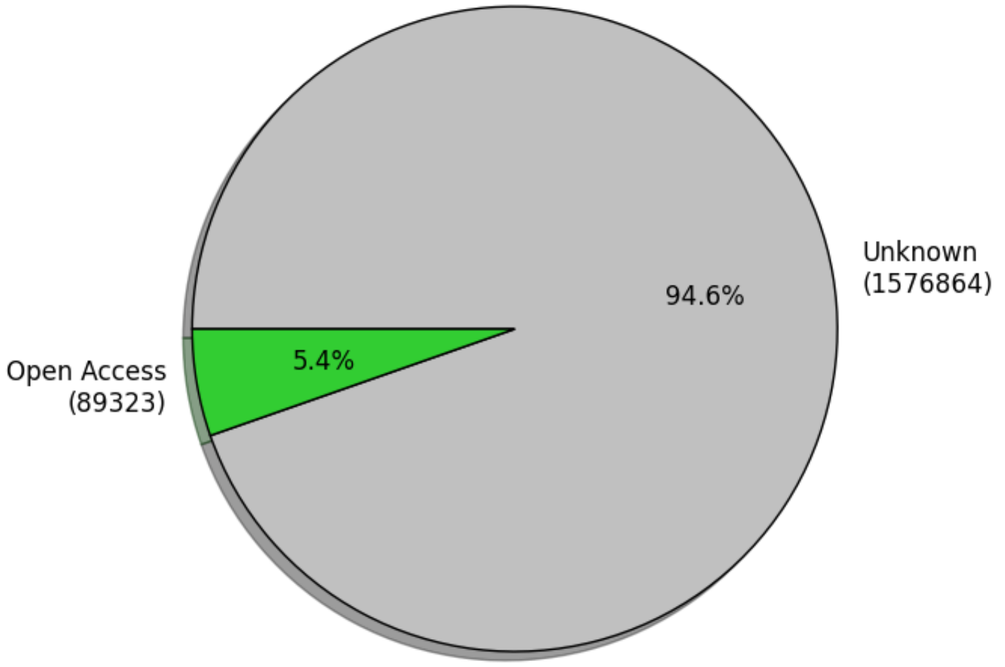
Mit dc:rights-Normalisierung



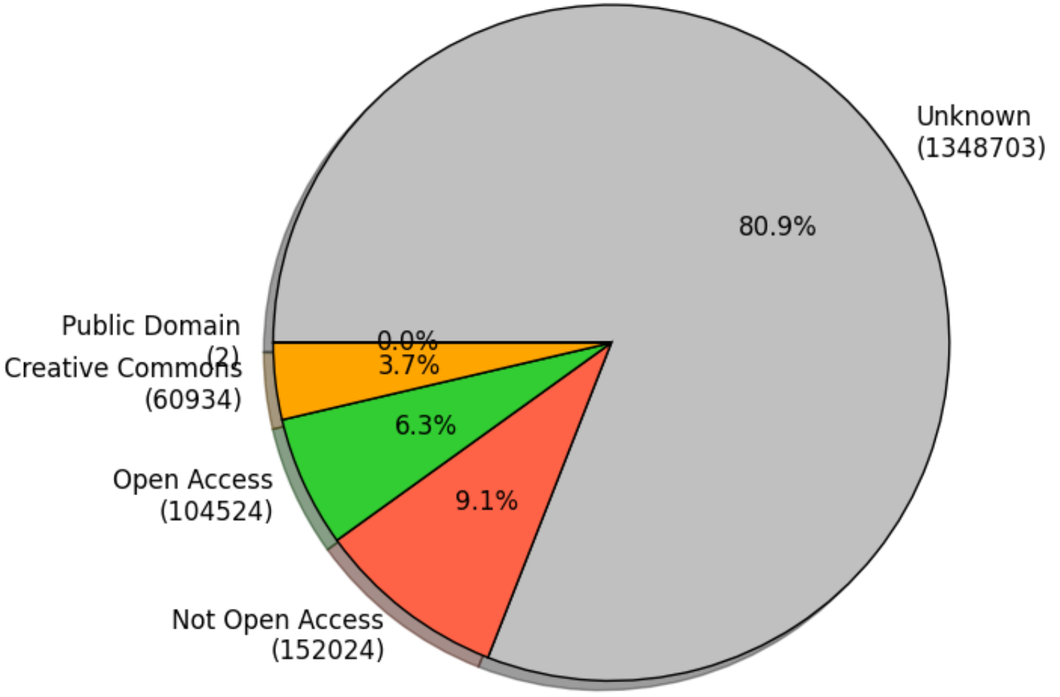
Ergebnisse: Schweiz

Änderungen im Suchmaschinen-Index von BASE durch Rechtenormalisierung (Land: Schweiz, 1.666.187 Dokumente aus 27 Quellen)

Ohne dc:rights-Normalisierung



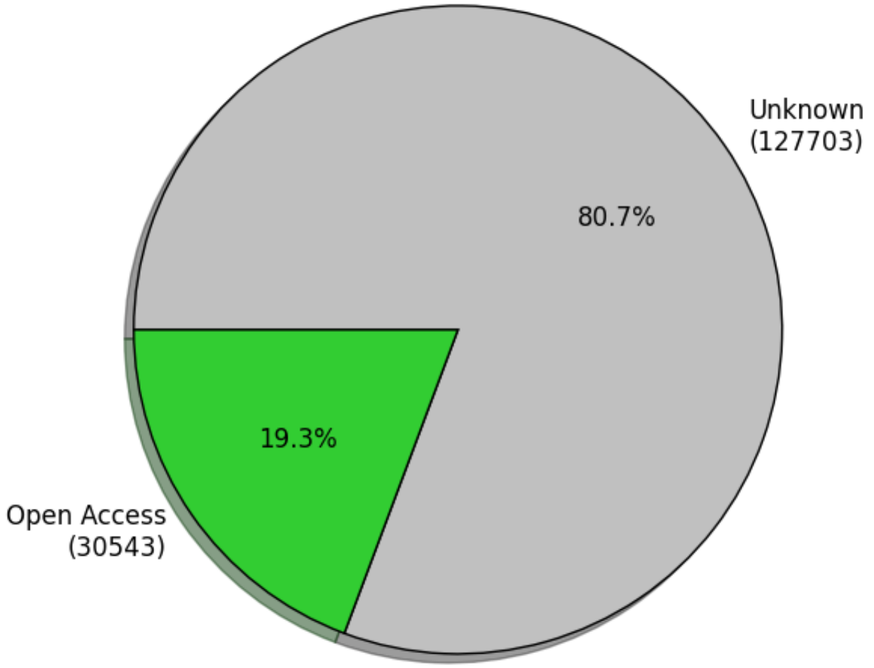
Mit dc:rights-Normalisierung



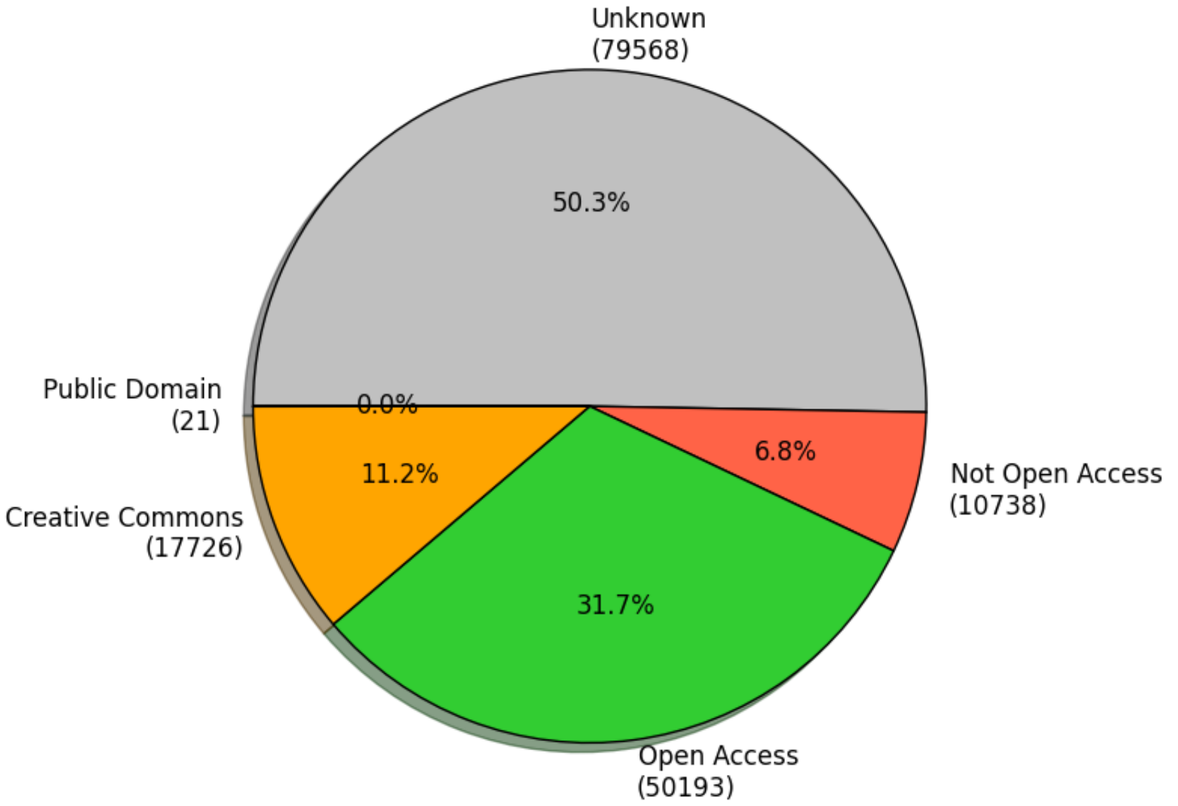
Ergebnisse: Österreich

Änderungen im Suchmaschinen-Index von BASE durch Rechtenormalisierung (Land: Österreich, 158.246 Dokumente aus 21 Quellen)

Ohne dc:rights-Normalisierung



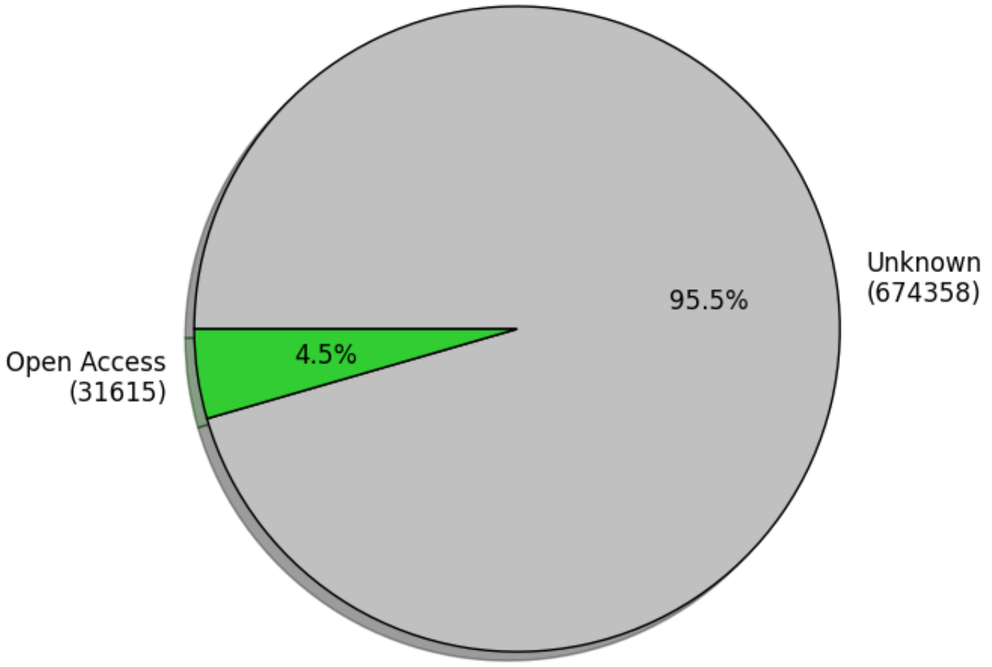
Mit dc:rights-Normalisierung



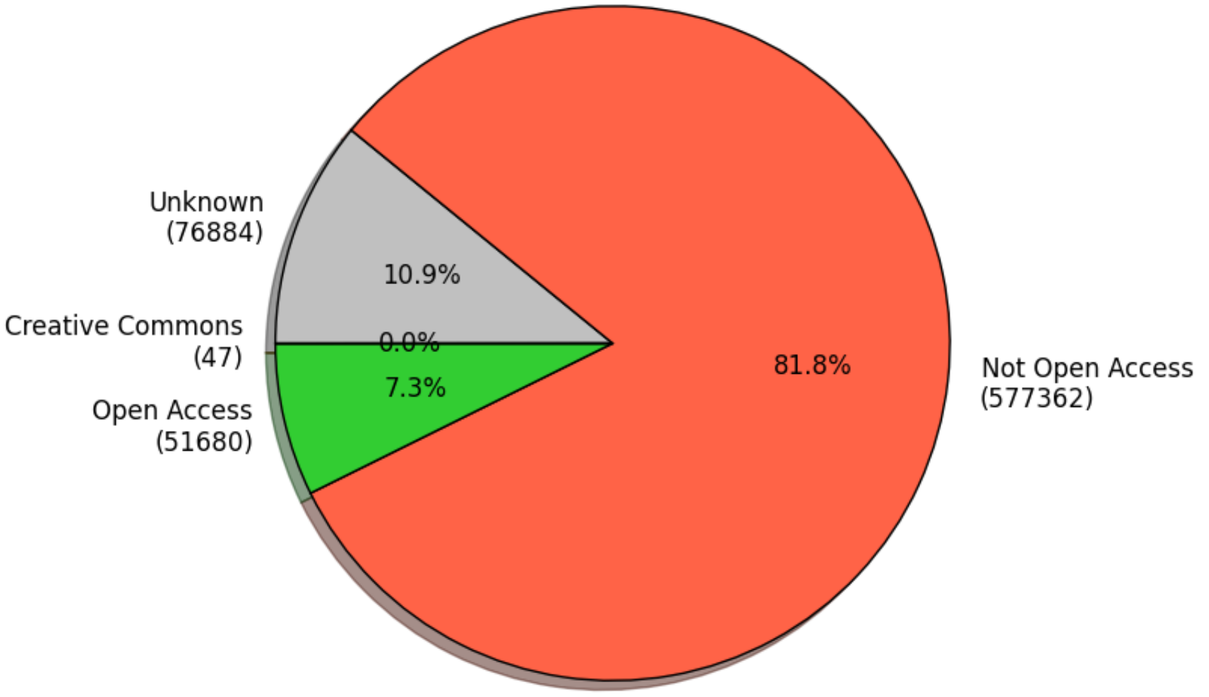
Ergebnisse: Dänemark

Änderungen im Suchmaschinen-Index von BASE durch Rechtenormalisierung (Land: Dänemark, 705.973 Dokumente aus 15 Quellen)

Ohne dc:rights-Normalisierung



Mit dc:rights-Normalisierung



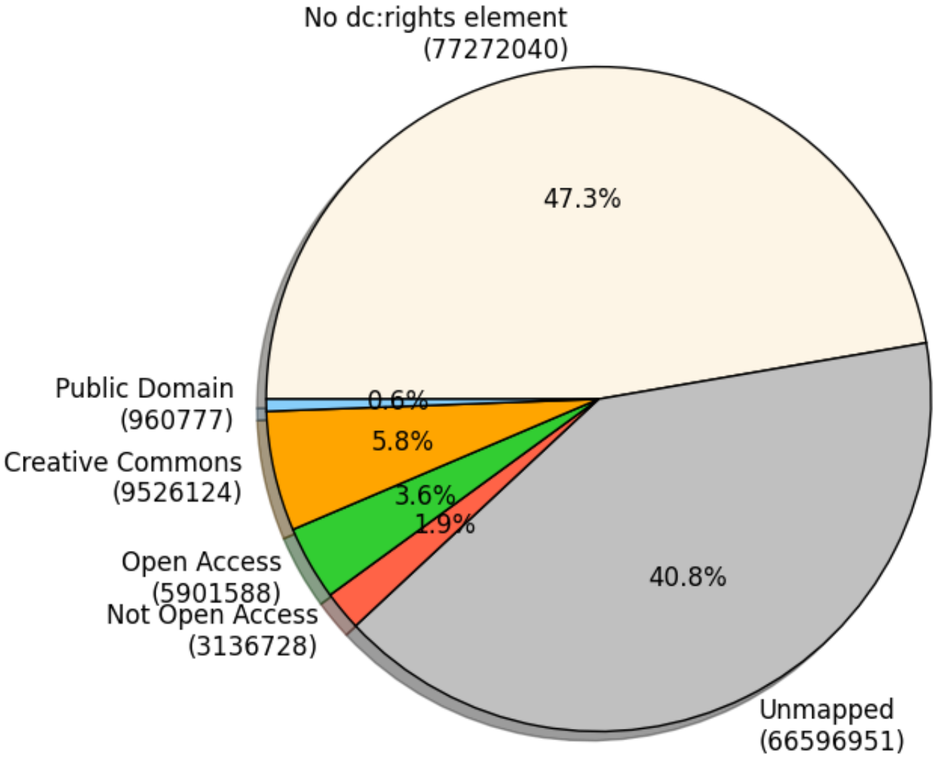
Zusammenfassung und Empfehlungen

Normalisierung von dc:rights-Feldinhalten in BASE seit Mitte 2015, dadurch:

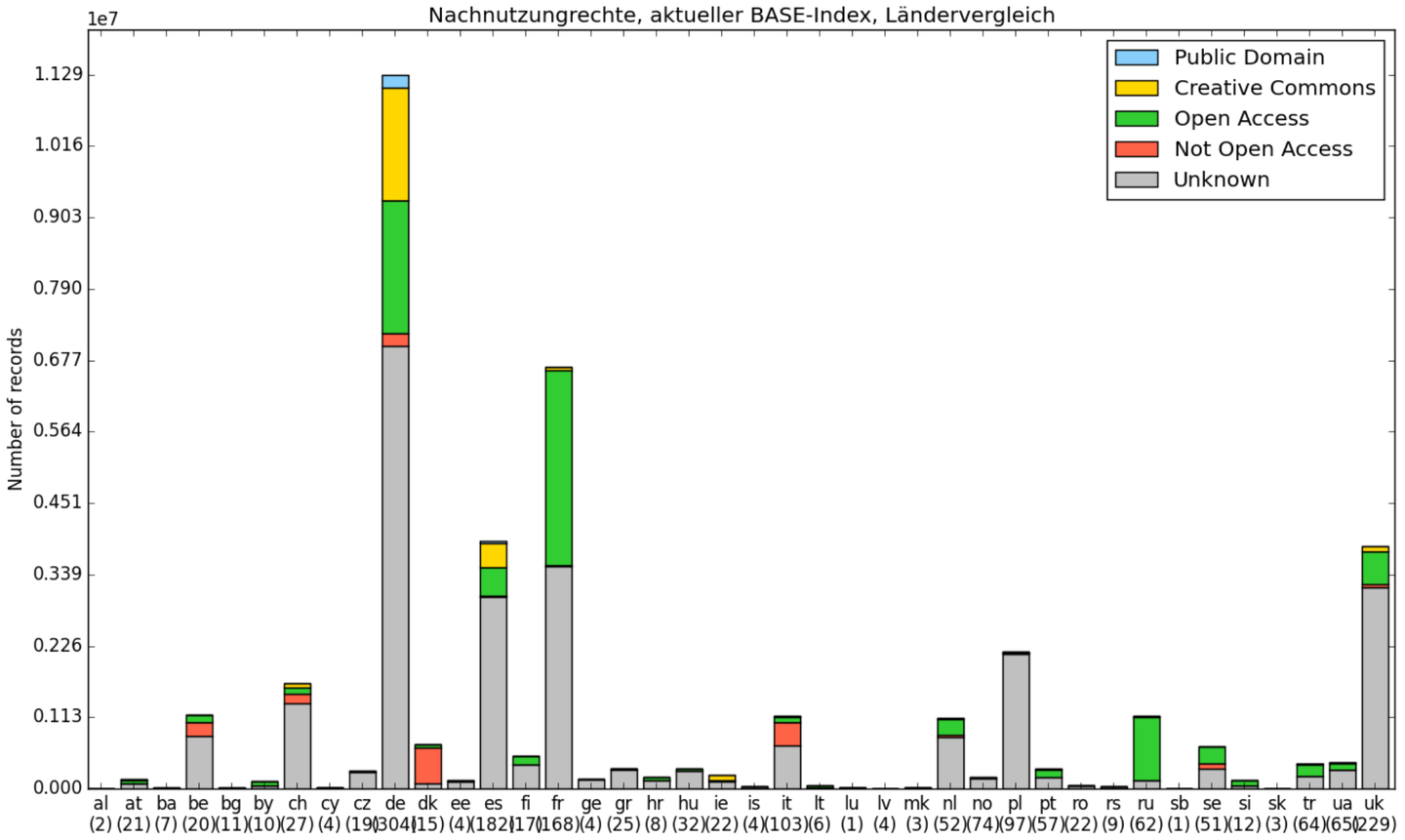
- Automatische Identifikation von Open Access-Status direkt auf Dokumentebene
- Automatische Identifikation von weitergehenden Lizenzen/Nachnutzungsrechten
- **Bestehendes Problem:** Fehlende/nicht-maschinenlesbare Feldinhalte!

Empfehlungen an Repository-Verwalter:

1. Das Feld dc:rights sollte unbedingt ausgefüllt werden.
2. Die Inhalte sollten mindestens standardisiert und bevorzugt maschinenlesbar sein (info-repo-Vokabular, Lizenz-URIs).
   Hinweis: dc:rights darf auch mehrfach vorkommen, um beispielsweise einen natürlichsprachlichen und einen maschinenlesbaren Rechtevermerk anzubringen.
3. Auch Hinweise wie Closed Access oder Restricted Access sind hilfreich und sollten vermerkt werden - "Negative" Information ist immer noch sehr viel besser als gar keine Information.



Letztes Diagramm...



Danke für ihre Aufmerksamkeit!