



www.data.europa.eu

Metadata Quality Assurance

03.05.2025

Wie data.europa.eu die Qualität von geharvesteten Metadaten misst

Das Metadata Quality Assessment (MQA) ist ein vom Konsortium des data.europa.eu Portals entwickeltes Werkzeug zur Untersuchung der Qualität der von data.europa.eu gesammelten Metadaten. Es soll Datenanbietern und Datenportalen helfen, ihre Metadatenqualität zu überprüfen und Verbesserungsvorschläge zu erhalten. Die Ergebnisse werden über die MQA präsentiert und stehen auch als Download zur Verfügung. Im Folgenden beschreiben wir die Funktionalität der MQA und die von ihr verwendete Methodik.

Wenn diese Seite immer noch nicht alle Ihre Fragen beantwortet, können Sie uns gerne über unser Feedback-Formular am Ende der Seite kontaktieren.

[Metadata Quality Assessment Methodologie](#)

Umfang der Untersuchung

Mit dem MQA wollen wir die folgende Frage beantworten:

Wie ist die Qualität der Metadaten für Daten des öffentlichen Sektors in der gesamteuropäischen Region und wo liegen die größten Hürden, um eine bessere Qualität zu erreichen?

Darauf aufbauend untersucht der MQA derzeit die folgenden konkreten Fragen:

- Übereinstimmung mit DCAT-AP und DCAT-AP-Derivaten
- Offenlegung von Informationen, zu denen DCAT-AP nicht verpflichtet ist
- Zugänglichkeit der in den Metadaten referenzierten Daten über die Access- und Download-URL
- Maschinenlesbarkeit der referenzierten Daten
- Verwendung von Lizenzen

Jede Frage führt wiederum zu individuellen Untersuchungen, die im Folgenden detailliert beschrieben werden.

Was wir nicht abdecken

Der MQA ist durch die Metadaten, die sie untersuchen kann, begrenzt. Die Untersuchung beschränkt sich ausschließlich auf die Metadaten, die data.europa.eu während des Harvesting-Prozesses sammelt. Wenn in den Quell-Metadaten Fehler auftreten, können diese das Gesamtergebnis verfälschen. Um dieses Fehlerpotential zu begrenzen, bietet der MQA einen

Validierungsdienst an, der von Datenanbietern genutzt werden kann, um ihre Metadaten vor der Integration in den Harvesting-Prozess auf gültige Formate und konformes DCAT-AP zu validieren.

[DCAT-AP SHACL Validierungsdienst Webseite](#)

[DCAT-AP SHACL Validierungsdienst API](#)

Der MQA Prozess

Mit jedem Harvesting werden auch die Metadaten vom MQA überprüft. Der MQA misst die Qualität der verschiedenen Indikatoren, jeder Indikator wird in den folgenden Tabellen erläutert. Die Ergebnisse der Überprüfungen werden als Data Quality Vocabulary ([DQV](#)) gespeichert. DQV ist eine Spezifikation des W3C, die zur Beschreibung der Qualität eines Datensatzes verwendet wird.

Da die Zugänglichkeit unbeständig sein kann, sind wiederholte Überprüfungen der accessURL und downloadURL notwendig. Aus diesem Grund überprüft die MQA regelmäßig die Zugänglichkeit aller Distributionen. Im Gegensatz zur Überprüfung der anderen Indikatoren hat diese eine höhere Laufzeit, da die Distributionen über HTTP geprüft werden und jede angeforderte URL eine längere Antwortzeit haben kann. Der MQA verwendet einen Mechanismus, der berücksichtigt, dass jede URL innerhalb weniger Wochen nach der letzten Überprüfung erneut auf ihre Zugänglichkeit hin überprüft wird.

Annahmen

Der MQA arbeitet mit den folgenden Annahmen.

Verwendung von nicht obligatorischen Feldern

Wir glauben, dass das Ausfüllen der DCAT-AP-Pflichtfelder allein nicht ausreicht, um qualitativ hochwertige Metadaten zu liefern. Aus diesem Grund prüft die Auswertung auch Felder, die nach DCAT-AP nicht als Pflichtfelder angegeben sind. Die genauen Felder, die geprüft werden, sind unten aufgeführt.

Identischer Inhalt für mehrere Distributions

Wenn ein Datensatz mehr als eine Distribution enthält, sind alle Verteilungen inhaltlich identisch, sie unterscheiden sich nur in der Darstellung der Daten. So kann ein Datensatz beispielsweise zwei Distributions haben, wobei die eine die Daten als PDF und die andere die identischen Daten als maschinenlesbare RDF/XML anbietet.

Dimensionen

Dieser Abschnitt beschreibt alle Dimensionen, die der MQA untersucht, um die Qualität zu bestimmen. Die Dimensionen werden aus den [FAIR-Prinzipien](#) abgeleitet.

Auffindbarkeit

Die folgende Tabelle beschreibt die Metriken, die Menschen und Maschinen beim Auffinden von Datensätzen helfen. In diesem Bereich können maximal 100 Punkte erzielt werden.

Indikator	Beschreibung	Metrik	Gewichtung
Schlüsselwörter	Schlüsselwörter unterstützen direkt die Suche und erhöhen damit die Auffindbarkeit des Datensatzes.	Der MQA prüft, ob Schlüsselwörter definiert sind. Die Anzahl der Schlüsselwörter hat keinen Einfluss auf die Punktzahl. Dataset dcat:keyword	30
Kategorien	Kategorien helfen den Benutzern, Datensätze thematisch zu erforschen.	Es wird geprüft, ob dem Datenbestand eine oder mehrere Kategorien zugeordnet sind. Die Anzahl der zugeordneten Kategorien hat keinen Einfluss auf die Punktzahl. Dataset dcat:theme	30
Ortsbezogene Suche	Die Nutzung räumlicher Informationen würde es den Nutzern ermöglichen, den Datensatz mit einer Suche mit räumlichen Bezug zu finden.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Dataset dcat:spatial	20
Zeitbasierte Suche	Die Verwendung von zeitlichen Informationen würde den Benutzern eine Suche mit Zeitbezug ermöglichen.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht.	20

Zugänglichkeit

Die folgende Tabelle beschreibt, welche Metriken verwendet werden, um festzustellen, ob der Zugriff auf die von den Distributions referenzierten Daten gewährleistet ist. In diesem Bereich können maximal 100 Punkte erzielt werden.

Indikator	Beschreibung	Metrik	Gewichtung
Erreichbarkeit der AccessURL	Die AccessURL ist nicht unbedingt ein direkter Link zu den Daten, sondern kann auch auf eine URL verweisen, die den Zugriff auf den Datensatz ermöglicht oder unter der weitere Informationen über den Datensatz verfügbar sind.	Die angegebene URL wird über eine HTTP HEAD-Anfrage auf Zugänglichkeit geprüft. Liegt der geantwortete Statuscode im Bereich 200 oder 300, wird die Zugänglichkeit der Ressource positiv bewertet. Distribution dcat:accessURL	50
DownloadURL	Die DownloadURL ist ein direkter Link zu den referenzierten Daten.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Distribution dcat:downloadURL	20
Erreichbarkeit der DownloadURL	Wenn eine DownloadURL existiert, wird sie auf ihre Erreichbarkeit geprüft.	Die angegebene URL wird über eine HTTP HEAD-Anfrage auf Zugänglichkeit geprüft. Liegt der geantwortete Statuscode im Bereich 200 oder 300, wird die Zugänglichkeit der Ressource positiv bewertet. Distribution dcat:downloadURL	30

Interoperabilität

In der folgenden Tabelle werden die Metriken beschrieben, die zur Bestimmung der Interoperabilität einer Distribution verwendet werden. Gemäß der Annahme "identischer Inhalt mit mehreren Distributionen" wird nur die Distribution mit der höchsten Punktzahl zur Berechnung der Punkte herangezogen. In diesem Bereich können maximal 110 Punkte erzielt werden.

Indikator	Beschreibung	Metrik	Gewichtung
Format	Dieses Feld gibt das Dateiformat der Distribution an.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Distribution dct:format	50
Media Type	Dieses Feld gibt den Media Type der Distribution an.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Distribution dcat:mediaType	10
Format / Media type aus Vokabular	Prüft, ob Format und Media Type zu einem kontrollierten Vokabular gehören.	Das Vokabular für das Format kann in dem data.europa.eu <u>GitLab repository</u> Repository gefunden werden. Der Medientyp wird gegen die <u>IANA-Liste</u> geprüft. Distribution dct:format dcat:mediaType	10
Non-proprietär	Prüft, ob das Format der Distribution nicht-proprietär sind.	Die Distribution wird als nicht-proprietär betrachtet, wenn das angegebene Format und Media Type in der entsprechenden data.europa.eu <u>GitLab repository</u> Vokabular enthalten ist. Distribution dct:format	20
Maschinenlesbarkeit	Prüft, ob das Format der Distribution maschinell lesbar ist.	Die Distribution wird als maschinenlesbar betrachtet, wenn das angegebene Format und Media Type in der entsprechenden data.europa.eu <u>GitLab repository</u> Vokabular enthalten ist. Distribution dct:format	20

Indikator	Beschreibung	Metrik	Gewichtung
DCAT-AP Konformität	DCAT-AP-Konformität wird über alle in einem Katalog verfügbaren Quellen und Datensätze berechnet. Diese Überprüfung wird nur durchgeführt, wenn die Metadaten ursprünglich als DCAT-AP oder als gültiges Derivat bereitgestellt wurden. DCAT-AP ist eine Spezifikation zur Beschreibung verlinkter öffentlicher Daten in Europa. Das data.europa.eu Portal kann auch Metadaten erfassen, die nicht vollständig mit DCAT-AP konform sind. Um die Konformität mit DCAT-AP zu erhöhen, überprüft die MQA alle Metadaten auf ihre DCAT-AP-Konformität.	Die Metadaten werden gegen einen Satz von <u>SHACL Regeln</u> validiert. Die Metadaten sind nicht konform, wenn die SHACL-Validierung mindestens ein Problem meldet. Der MQA verwendet den <u>data.europa.eu DCAT-AP SHACL</u> Validierungsdienst. SHACL ist eine Empfehlung des W3C und wird zur Validierung von RDF-Graphen gegen einen Satz von Regeln verwendet.	30

Wiederverwendbarkeit

Die folgende Tabelle beschreibt, welche Metriken zur Überprüfung der Wiederverwendbarkeit der Daten verwendet werden. In diesem Bereich können maximal 75 Punkte erzielt werden.

Indikator	Beschreibung	Metrik	Gewichtung
Lizenzangaben	Eine Lizenz ist eine wertvolle Information für die Wiederverwendung von Daten.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Distribution dct:license	20
Lizenzvokabular	Wir möchten die Angabe von falschen Lizenzinformationen einschränken. Zum Beispiel begegnen wir vielen CC-Lizenzen, denen eine Versionierung fehlt.	Dieser Abschnitt beschreibt alle Dimensionen, die der MQA untersucht, um die Qualität zu bestimmen. Die Dimensionen werden aus den <u>FAIR-Prinzipien</u> abgeleitet. Der MQA empfiehlt und würdigt die Verwendung von kontrollierten Vokabularen. Das data.europa.eu Portal veröffentlicht seine kontrollierten Vokabulare in <u>GitLab</u> . Die Vokabulare sind abgeleitet von den <u>EU Vocabularies</u> . Distribution dct:license	10
Zugangsbeschränkungsangaben	Dieses Feld gibt an, ob der Zugang zu den Daten öffentlich oder eingeschränkt ist.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Dataset dct:accessRights dcat:mediaType	10
Zugangsbeschränkungsangaben aus Vokabular	Die Verwendung eines kontrollierten Vokabulars erhöht die Wiederverwendbarkeit.	Es wird geprüft, ob das <u>kontrollierte Vokabular für Zugriffsrechte</u> verwendet wird. Dataset dct:accessRights	5
Kontaktinformation	Die Kontaktstelle enthält Informationen, an wen Sie sich bei Fragen zu den Daten wenden können.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Dataset dct:contactPoint	20
Herausgeber	Der Herausgeber ist eine Person oder Organisation, die die Daten veröffentlicht hat.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Dataset dct:publisher	10

Kontext

Die folgende Tabelle zeigt einige Leichtgewicht-Eigenschaften, die dem Benutzer mehr Kontext bieten. In diesem Bereich können maximal 20 Punkte erzielt werden.

Indikator	Beschreibung	Metrik	Gewichtung
Rechte	In einigen Fällen kann eine bestimmte Lizenz nicht auf einen Datensatz angewendet werden. Das Feld "Rechte" kann verwendet werden, um einen Verweis auf eine Ressource anzugeben, der einen Benutzer über die Rechte informiert, die er bei der Nutzung des Datasets hat.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Distribution dct:rights	5
Dateigröße	Gibt die Größe der Datei in Bytes an.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Distribution dct:byteSize	5
Ausstellungsdatum	Das Datum, an dem der Datensatz oder die Distribution freigegeben wurde.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Dataset and Distribution dct:issued	5
Änderungsdatum	Das Datum, an dem der Datensatz oder die Distribution zuletzt geändert wurde.	Es wird geprüft, ob diese Eigenschaft gesetzt ist oder nicht. Dataset and Distribution dct:modified	5

Bewertung

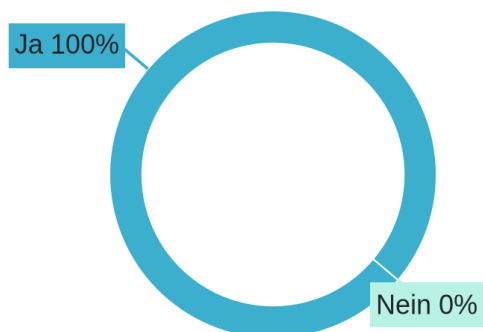
Die endgültige Bewertung geschieht über vier Bewertungsgruppen. Die Zuordnung der Punkte zur Bewertungskategorie ist in der folgenden Tabelle dargestellt. Die Darstellung der Bewertung im MQA wird ausschließlich über die Bewertungskategorien ausgedrückt. Dies ermöglicht es den Anbietern, die höchste Bewertung trotz eines leichten Punktabzug zu erreichen.

Dimension	Maximale Punktzahl
Auffindbarkeit	100
Zugänglichkeit	100
Interoperabilität	110
Wiederverwendbarkeit	75
Kontext	20
Summe	405

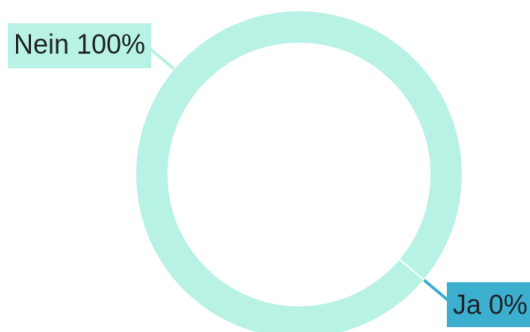
Bewertung	Punktebereich
Excellent	351 - 405
Gut	221 - 350
Ausreichend	121 - 220
Mangelhaft	0 - 120

Kontext

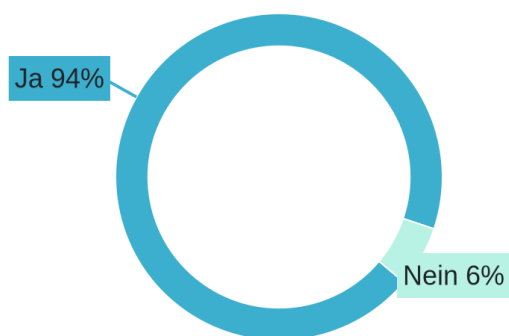
Rechte



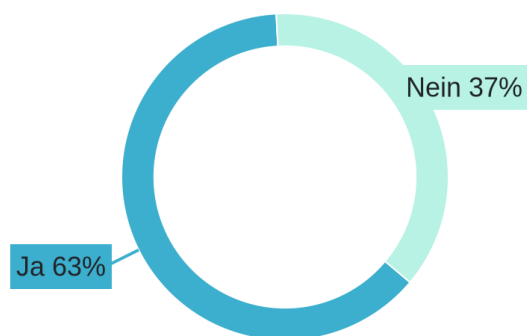
Dateigröße



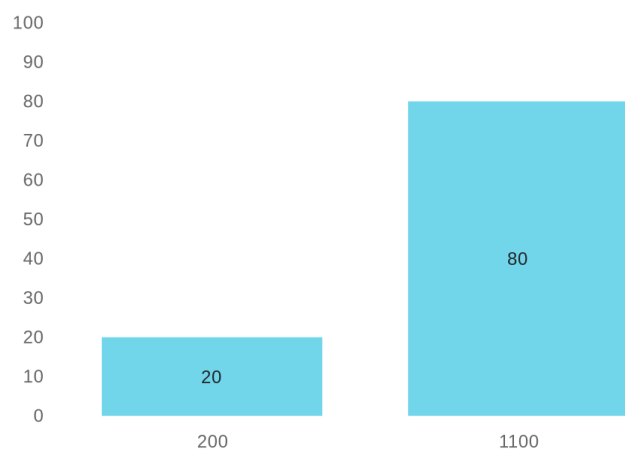
Ausstellungsdatum



Änderungsdatum

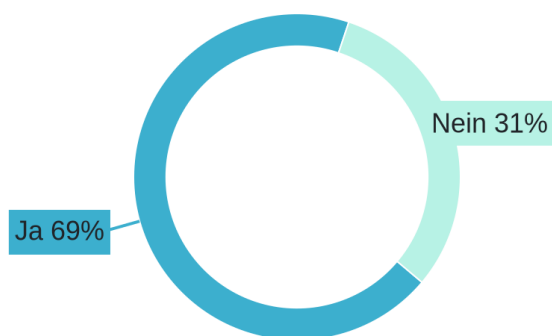
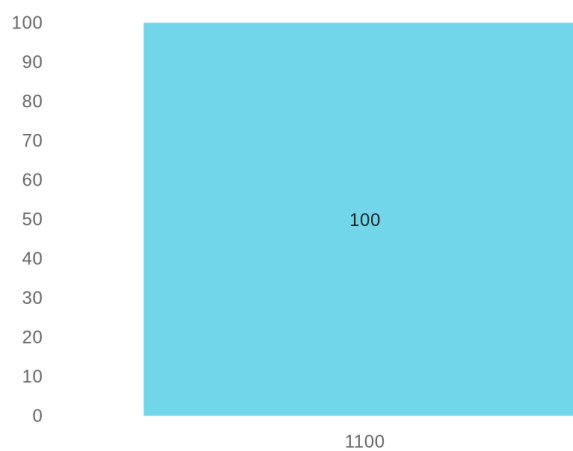


Häufigste AccessURL Statuscodes

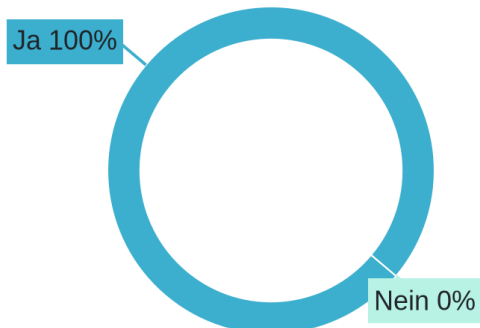


Download URL

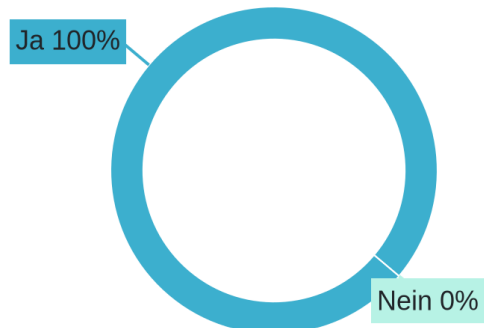
Häufigste DownloadURL Statuscodes



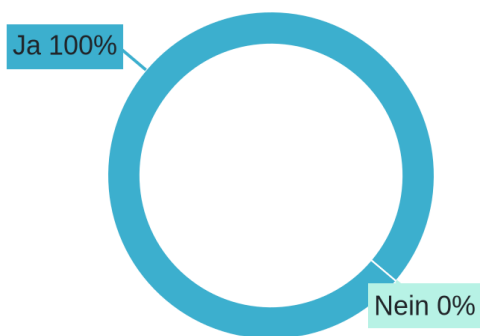
Lizenzangaben



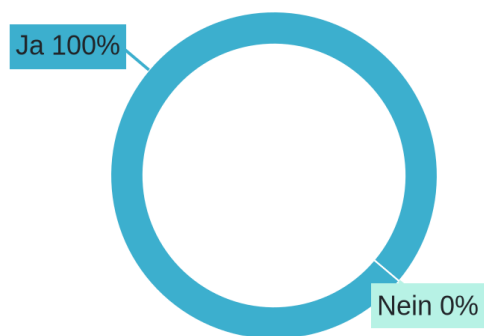
Lizenzvokabular



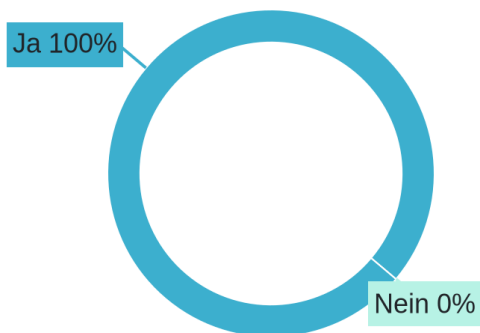
Zugangsbeschränkungsangaben



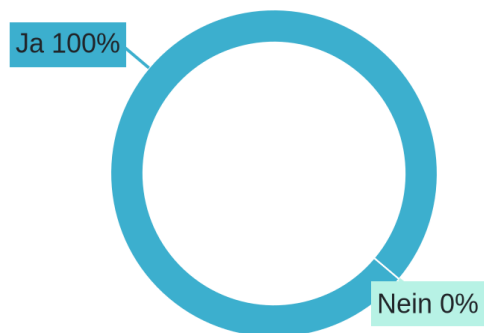
Zugangsbeschränkungsangaben aus Vokabular



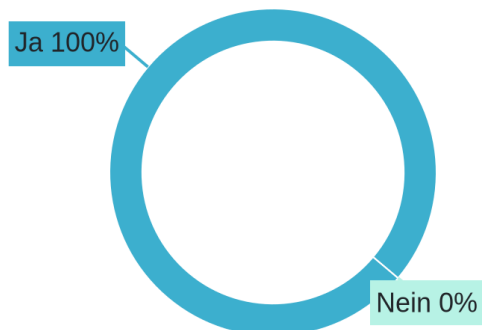
Kontaktinformation



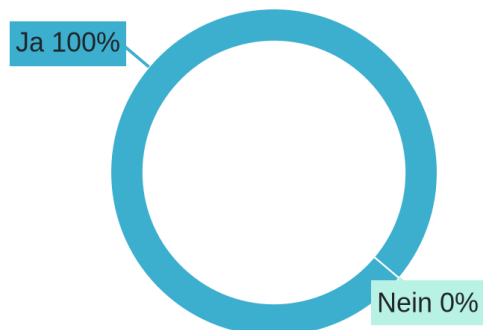
Herausgeber



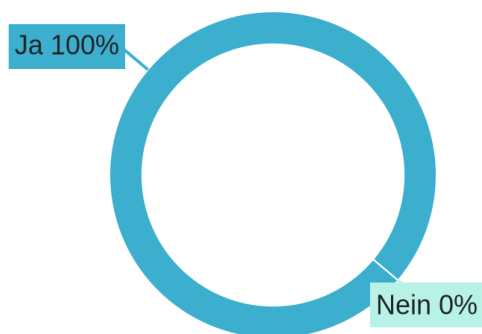
Format



Media Type



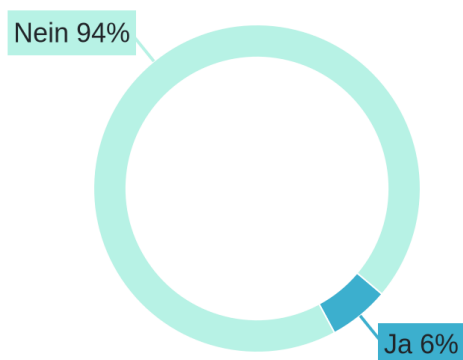
Format / Media type aus Vokabular



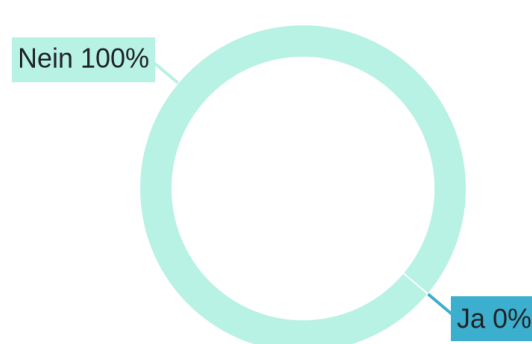
Nicht-proprietär



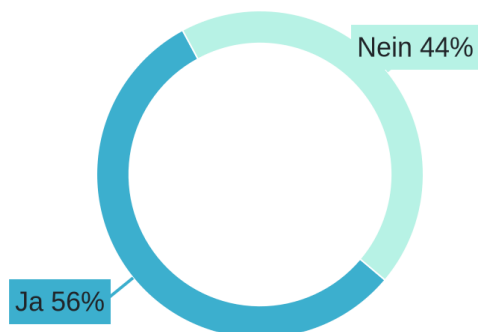
Maschinenlesbarkeit



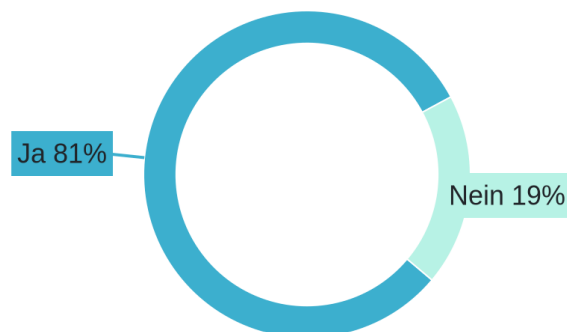
DCAT-AP Konformität



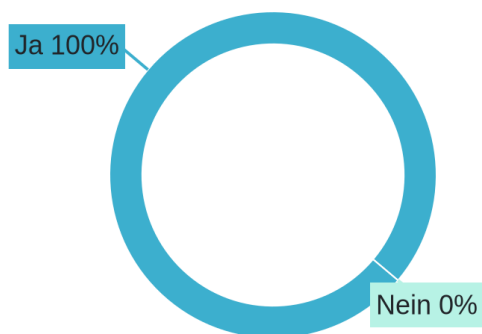
Schlüsselwörter



Kategorien



Ortsbezogene Suche



Zeitbasierte Suche

