



Success Story | GVL

Mit Semantic Data Cleansing erfolgreich

Die Arbeit mit unstrukturierten Daten kann viel Zeit und Geld kosten – das weiß die Gesellschaft zur Verwertung von Leistungsschutzrechten (GVL) aus eigener Erfahrung. Regelmäßig steht sie vor der Herausforderung, Produktinformationen in Nutzungsmeldungen aus verschiedenen Quellen und Formaten mit ihrer zentralen Künstler- und Titeldatenbank abzugleichen. Lange war diese Aufgabe nur mit hohem Personalaufwand und manueller Prüfung zu bewältigen – bis die GVL die Semantik-Experten der Berliner Condat AG engagierte, die das Problem analysierten und eine Lösung zur automatisierten Datenbereinigung entwickelten.

Immer wenn der Musiktrack eines Künstlers genutzt wird, hat er Anspruch auf eine angemessene Vergütung. Das gilt auch für diejenigen, die als Hersteller die wirtschaftliche Grundlage für eine Produktion schaffen. Verwertungsgesellschaften wie die GVL übernehmen für Ihre Berechtigten die Lizenzierung und ziehen die anfallenden Gebühren ein. Diese werden dann jährlich auf der Grundlage der Nutzungsmeldungen ausgeschüttet. Das Problem: Nur wenn eine Meldung eindeutig einem Berechtigten zugeordnet werden kann, ist eine Weiterleitung der Vergütung an ihn möglich. Aufgrund der uneinheitlichen Struktur und Inhalt der Meldungen war dafür bislang ein zeit- und kostenintensives manuelles Sichten nötig.

Fact Box - Die GVL im Überblick.

Seit 1959 vertritt die Gesellschaft zur Verwertung von Leistungsschutzrechten mbH (GVL) die Interessen von ausübenden Künstlern und Tonträgerherstellern. Die treuhänderisch eingenommenen Gelder u.a. von Radio- und Fernsehsendern sowie für die öffentliche Wiedergabe leitet die Gesellschaft jährlich und direkt als Vergütung an die Berechtigten weiter. Mehr als 136.000 Künstler und über 49.000 Labels weltweit machen die GVL zu einer der größten Verwertungsgesellschaften für Leistungsschutzrechte in Europa.

Die Lösung: Semantic Data Cleansing.

Zunächst führten Berater der Condat AG Mitte 2016 erste Interviews mit den Fachabteilungen und untersuchten Struktur und Inhalt der Produktinformationen. Die gemeinsam entwickelte Idee: Gelänge es, die genannten Künstler und Titel automatisiert zweifelsfrei zu identifizieren, wäre die Auszahlung der Vergütungen ohne aufwändige manuelle Nacharbeit möglich.

Anhand eines ersten Sets von Testdatensätzen und dem fachlichen Know how der GVL-Mitarbeiter begannen die Experten von Condat mit der Arbeit. Um der großen Bandbreite an Notationen und Uneindeutigkeiten in der Datenbank Herr zu werden, setzten sie im Projekt auf eine semantische Datenanalyse auf Basis der Condat Smart Media Engine. Der Vorteil: Durch das Einbeziehen geeigneter Informationsquellen aus dem Internet ist die Engine in der Lage, selbst stark fragmentierte und heterogene Datensätze abzugleichen und den korrekten Namen des Künstlers in vielen Fällen zweifelsfrei zu ermitteln.

Die GVL über das Projekt



„Zugegeben: Am Anfang waren wir skeptisch was den Projekterfolg angeht. Immerhin arbeiten wir seit Jahren mit hohem Personalaufwand an der Verbesserung unserer Datenqualität. Was wir dann mit Condat und ihrer Smart Media Engine in der kurzen Projektlaufzeit erreicht haben, hat uns positiv überrascht.“

*Dr. Tilo Gerlach und Guido Evers
Geschäftsführer der GVL*

Das Ergebnis: Endlich Licht im Datenwald.

Bereits im Pilotprojekt gelang es, den Vorteil der semantischen Analyse für die Datenbereinigung eindrucksvoll zu demonstrieren. Obwohl als externe Datenquelle lediglich eine offene Musik-Datenbank eingebunden wurde, konnte die Smart Media Engine deutlich mehr als der Hälfte der Meldungen automatisch den korrekten Berechtigten zuordnen.

In dem nachfolgenden Beratungs- und Bereinigungsprojekt wurden elf Millionen Datensätze mit der Condat Smart Media Engine analysiert. Auch hier war das Ergebnis außerordentlich positiv: Acht Millionen Datensätze konnten in der ersten Projektphase korrigiert werden. In einem Folgeprojekt wird es nun die Aufgabe der Condat-Berater sein, eine Lösung für die verbliebenen, zum großen Teil qualitativ sehr schwierigen Datensätze zu erarbeiten.