IN02 Fachvortrag

iiRDS voll automatisch – mit Machine Learning zum intelligenten Austauschformat

Jan Oevermann, ICMS, Karlsruhe

Der neue tekom-Standard iiRDS bietet eine standardisierte und herstellerneutrale Möglichkeit, Inhalte mit klassifizierenden Metadaten auszuzeichnen. Erst durch diese semantischen Informationen werden intelligente Anwendungen im Umfeld von Industrie 4.0 und Service möglich. Doch in den meisten Fällen sind die Inhalte noch nicht aufbereitet und eine manuelle Nachklassifizierung zu aufwändig. Wie kann dieser Prozess automatisiert werden?

Im Beitrag wird ein Konzept gezeigt, wie mit Hilfe von Machine-Learning-Methoden der Standard iiRDS automatisiert mit intelligenten Metadaten angereichert werden kann. Dazu werden verschiedene Technologien miteinander kombiniert, um einen durchgängigen Prozess zu schaffen.

iiRDS – ein Überblick

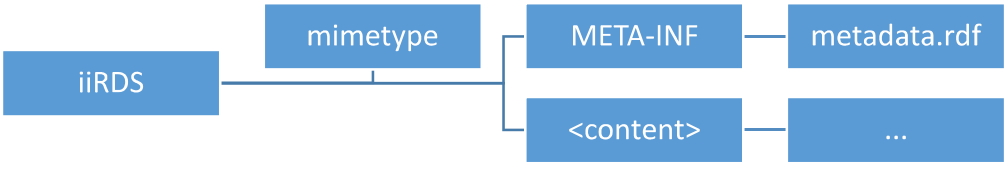
Bei iiRDS handelt es sich um einen neuen Standard, der unter der Schirmherrschaft der tekom von der Arbeitsgruppe „Information 4.0“ entwickelt wird. Die Spezifikation umfasst dabei zwei wesentliche Teile, die den Standard definieren (tekom e.V., 2017):

- Das Paketformat, das ein einheitliches Ausgabe-Dateiformat definiert und damit eine uneingeschränkte Austauschbarkeit zwischen Systemen ermöglicht.
- Die Metadatenmodellierung, die systemunabhängig Aussagen zu den Inhalten eines Pakets abbildet und z. B. durch ein Content-Delivery-Portal ausgewertet werden kann.

Der Standard selbst versteht sich als Austauschformat und nicht als Erfassungs- oder Publikationsformat und kann dadurch flexibel und unabhängig vom eingesetzten Informationsmodell eingesetzt werden. Obwohl iiRDS eine modularisierte Dokumentation propagiert, können damit auch dokumentbasierte Dateien ausgeliefert und mit Metadaten versehen werden. Weiterhin ist es möglich, nur Teile eines Dokuments auszuzeichnen.

Die Besonderheit der iiRDS-Metadatenmodellierung liegt in der Anwendung der sogenannten Linked-Data-Prinzipien (Berners-Lee, 2006), die auf der Formulierung logischer Aussagen beruhen. Diese Aussagen werden bei iiRDS in einer RDF-Datei mit Hilfe eines vorgefertigten Modells (dem RDF Schema) über den in einem Paket enthaltenen Content getroffen.

iiRDS wird in einem Paket ausgeliefert, das dem folgenden Aufbau entspricht:



Das Paket entspricht dem ZIP-Format. Alle Metadaten des Pakets werden in einer XML-Serialisierung als RDF-Datei gespeichert und im Ordner „META-INF“ im iiRDS-Paket abgelegt.

Metadaten durch Machine Learning

Eine automatisierte Klassifizierung von Texten kann mit Hilfe von Trainingsdaten, die händisch mit Metadaten ausgezeichnet wurden, realisiert werden. Nach einer erforderlichen Anpassung der Methoden des maschinellen Lernens (engl.: Machine Learning) auf die Charakteristiken der Technischen Dokumentation können Vorhersagen in ausreichender Qualität für einen produktiven Betrieb getroffen werden (Oevermann, 2016).

Unter den in iiRDS modellierten Metadaten kommen vor allem die Bereiche Informationsthema (`iirds:InformationSubject`), Topictyp (`iirds:TopicType`), Produktlebenszyklusphase (`iirds:ProductLifeCyclePhase`) und Komponente (`iirds:Component`) für eine automatische Vergabe in Frage. Diese Klassen entsprechen klassifizierenden und intrinsischen Metadaten, da sie eine vorgegebene Menge an möglichen Werten vorgeben und Eigenschaften des Inhalts selbst (und nicht etwa seiner Zugehörigkeit) abbilden. Damit lassen sich mit Mitteln der statistischen Sprachverarbeitung aus der Charakteristik der Texte die entsprechenden Werte für eine Klasse ableiten. Für die Erzeugung von Trainingsdaten kann ein Mapping von bestehenden Klassifikationen, die z. B. nach PI-Class® modelliert wurden (Drewer/Ziegler, 2011), nach iiRDS erstellt werden.

Der Klassifizierer vergibt nach erfolgtem Mapping die in iiRDS definierten Instanzen in Form von URIs. Beispiel: Für die PI-Klassifikation „Wartung“ der zweiten Ebene wird die iiRDS-Instanz „`iirds:Maintenance`“ (URI: `http://www.tekom.de/iirds#Maintenance`) der Klasse „`iirds:ProductLifeCyclePhase`“ zugeordnet. Die trainierten Informationen werden in einem Modell gespeichert, das dazu eingesetzt werden kann, bisher unbekannte (und nicht mit Metadaten versehene Daten) automatisiert zu klassifizieren.

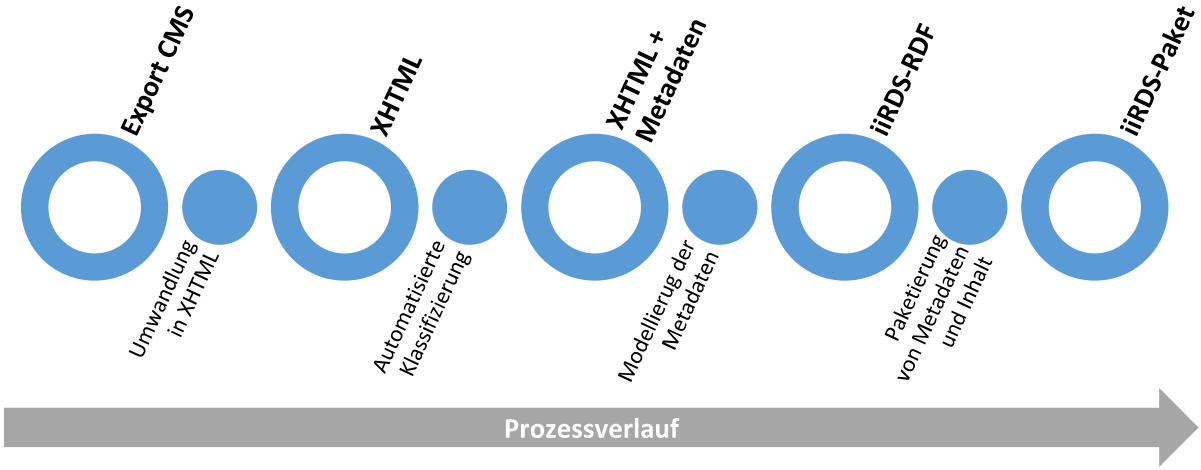
Für modularisierte Inhalte wird die Klasse direkt dem entsprechenden Topic zugeordnet. Soll ein dokumentbasierter Inhalt mit Metadaten angereichert werden, wird die semantische Struktur des Dokuments rekonstruiert und die erkannten Segmente einzeln mit Metadaten versehen. Dabei kommen die in iiRDS modellierten Selektoren zum Einsatz, die beispielsweise den Seitenbereich eines PDFs definieren, für den ein Metadatum gültig sein soll (`iirds:RangeSelector`).

Durch dieses Vorgehen lassen sich heterogene Datenquellen mit einem einheitlichen Metadatenformat auszeichnen, wodurch sich integrierte Informationsplattformen realisieren lassen, die semantisch strukturierte Abfragen über alle Informationsarten hinweg ermöglichen (Bader/Oevermann, 2017). Diese Standardisierung ist eines der Hauptziele von iiRDS und bedient ein erkennbares Bedürfnis nach unternehmensweiten Informationsportalen (Ziegler, 2017).

Automatisierung

Für die Automatisierung des Prozesses werden mehrere Phasen miteinander verknüpft:
- Umwandeln des Erfassungsformats in neutrales XHTML (bei modularen Inhalten)
- Einbinden einer Verarbeitung durch eine Klassifizierungs-API (Schnittstelle)
- Generieren der korrekten iiRDS-Metadatendatei (metadata.rdf)
- Erzeugen des iiRDS-Paketformats

Dieser Prozess kann aufbauend auf den Exportformaten marktüblicher Content-Management-Systeme z. B. mit den auf vergangenen Jahrestagungen vorgestellten Lösungen „infoflow“ (Oevermann, 2015) und „fastclass“ (Oevermann, 2016) realisiert werden. Eine Einbindung alternativer System oder eine Prozesssteuerung durch Batch-Dateien ist natürlich auch möglich.



Sollen PDF-Dokumente zu intelligentem iiRDS überführt werden, kommt statt des ersten Schritts die oben beschriebene segmentbasierte Auszeichnung zum Einsatz. Das PDF bleibt dabei als Dokument bestehen, es werden jedoch Metadaten zu bestimmten (Seiten-)Bereichen erzeugt, die z. B. bei einer Suche ausgewertet werden können.

Fazit

Mit geeigneten Mitteln ist es bereits heute möglich, „voll automatisch“ und mit Hilfe von Machine Learning das semantische Austauschformat iiRDS zu erzeugen. Nachdem mit entsprechenden Trainingsdaten ein Modell trainiert wurde, kann innerhalb dieses Prozesses der Content automatisiert mit Metadaten angereichert und ein standardisiertes Inhaltspaket erzeugt werden. Einem intelligenten Austausch steht dann nichts mehr im Wege!

Literatur

- Bader, Sebastian / Oevermann, Jan (2017): „Semantic Annotation of Heterogeneous Data Sources: Towards an Integrated Information Framework for Service Technicians“. In: Proceedings of the 13th International Conference on Semantic Systems. SEMANTiCS 2017. ACM, NewYork.
- Berners-Lee, Tim (2006): „Linked Data“. http://www.w3.org/DesignIssues/LinkedData.html
- Drewer, Petra / Ziegler, Wolfgang (2011): „Technische Dokumentation“. Vogel, Würzburg.
- Oevermann, Jan (2017): „Intelligente Klassifizierung von technischen Inhalten – Automatisierung und Anwendungspotenziale“. In: Tagungsband zur tekom Jahrestagung 2016. tekom, Stuttgart.
- Oevermann, Jan (2015): „infoflow – Publikationsstrecken nach dem Baukastenprinzip“. In: Tagungsband zur tekom Jahrestagung 2015. tekom, Stuttgart.
- tekom e.V. (2017): „iiRDS Specification – intelligent information Request and Delivery Standard“. First Public Working Draft. April 2017. tekom, Stuttgart. http://iirds.tekom.de
- Ziegler, Wolfgang (2017): „Verteilen leicht gemacht“. In: technische Kommunikation, Ausgabe 03/2017, 39. Jahrgang, S. 30–34.

Kontakt:
jan.oevermann@icms.de