Risikobetrachtung Crawler Modul

Kurze Erläuterung zur funktionsweise:

Das Crawler Modul lädt aktiv daten aus dem Internet, wenn diese noch nicht in der Datenbank vorhanden sind.

Risiken

1. Internet Ausfall bedeutet das der Crawler nicht arbeiten kann.
2. Quellen[1] von den der Crawler Daten bezieht sind offline
3. Quellen von den der Crawler Daten bezieht ändern ihre API
4. IP Block / IP Blacklist wegen zu aggressiver Nutzung[2]
5. Ausfall der Hardware, Webserver oder sonstiger Software
6. Crawler lässt kann unter bestimmten umständen die Festplatte fluten mit temporär Daten für die DataRefinery
7. Nicht genügend Speicherplatz verfügbar, um temporär Daten zu erstellen.
8. Inhalt aus dem Crawler stimmt nicht mit dem Seiteninhalt überein

[1] Wikipedia, Wikimedia, Wikidata

[2] Aggressiver Nutzung ist hier gleichzusetzen mit zu viele Anfragen in einer zu kurzen Zeit.

Maßnahmen, um das Risiko zu minimieren

Zu 1. Nicht abfangbares Risiko, da Provider dafür verantwortlich sind. Support Nummern ablegen, um einen schnellen Informationsaustausch zu gewährleisten.

Zu 2. Da daten „on demand“[3] nachgeladen werden bleibt immer ein gewisses Rest Risiko, jedoch wird dieses im falle von den Wikipedia Projekten[1] als sehr gering eingeschätzt. Alternativ kann eine lokale Kopie angefordert werden, jedoch ist das mit einem hohen Hardwareaufwand verbunden.

Zu 3. Risiko immer vorhanden jedoch unwahrscheinlich da die API dauerhaft weiterentwickelt wird und somit eine langfristige Kompatibilität gewährleistet

Zu 4. API Calls können Bots identifizieren und reichen diese durch. Solche Projekte sind erwünscht und verstoßen nicht gegen die Wikipedia Richtlinien. Abfragen die nicht über die API realisiert werden sollten optional den Anwendungsnamen sowie eine Emailadresse des Ansprechpartners falls klärungsbedarf besteht.

Zu 5. Risiko immer gegeben.

Zu 6. Daten sind nur Textdaten im JSON Format und somit kompakt im Speicherbedarf. Crawler triggert sich nicht Selbständig, sondern wird von einem anderen Modul getriggert. Daten sind nur temporär und werden gelöscht sobald diese aufbereitet wurden und für die Datenbank bereitgestellt wurden.

Zu 7. Aus Sicht des Crawlers wird das Risiko immer geringer desto mehr Datensätze in die Datenbank eingepflegt wurden, da immer nur neue Daten generiert werden, wenn ein Datensatz nicht vorhanden ist.

Zu 8. Risiko besteht ist aber nicht automatisch überprüfbar. So lange der Crawler Ergebnisse im richtigen Format liefert wird keine Fehlermeldung verursacht.

[3] Nach bedarf