Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

INFORMATIONSSYSTEM

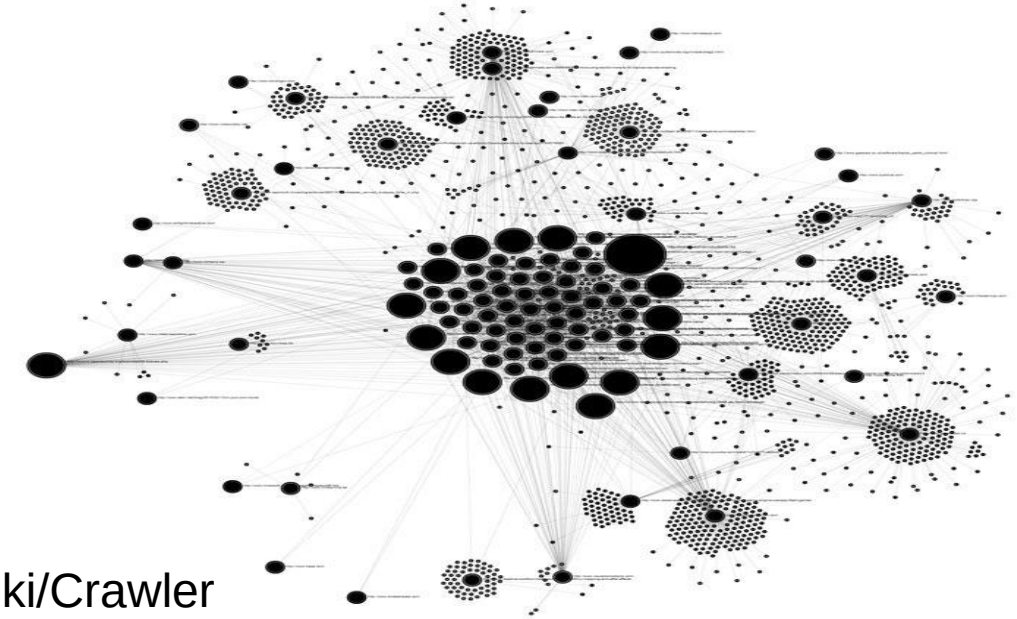
Erstellung eines Crawler zum Laden und Extraktion von Daten aus Bundestagsprotokollen

Gruppe I

Was ist ein Crawler?

Computerprogramm, das automatisiert Dokumente im Web durchsucht.

Hier werden wiederholende Aktionen programmiert, damit das Durchsuchen gänzlich automatisiert abläuft



https://de.ryte.com/wiki/Crawler

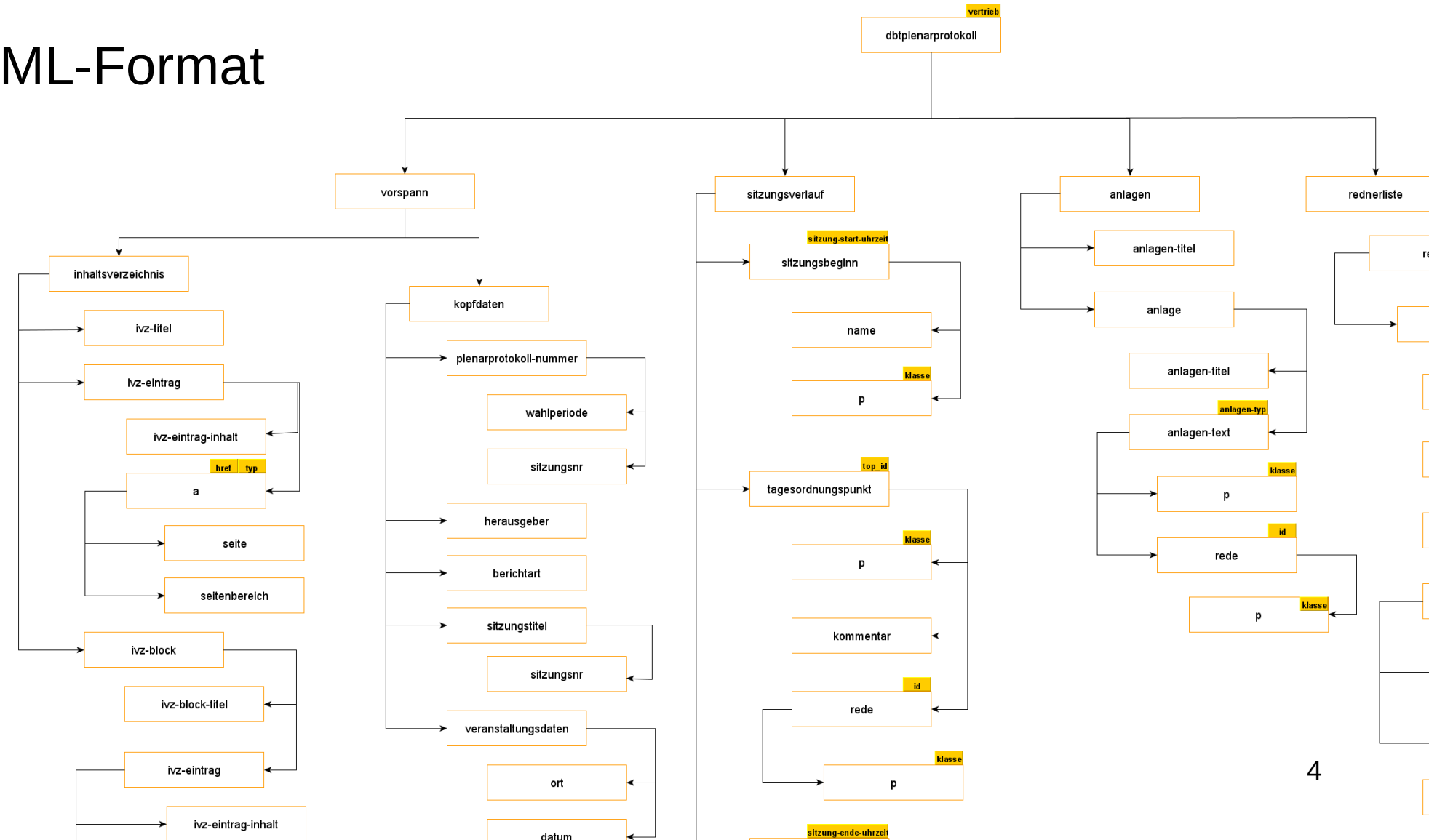
Was sind unsere Inputs?

Vom Bundestag stehen uns sowohl Plenarprotokolle ab der 1. Wahlperiode als auch einige ergänzende Informationen unter https://www.bundestag.de/services/opendata zur Verfügung.

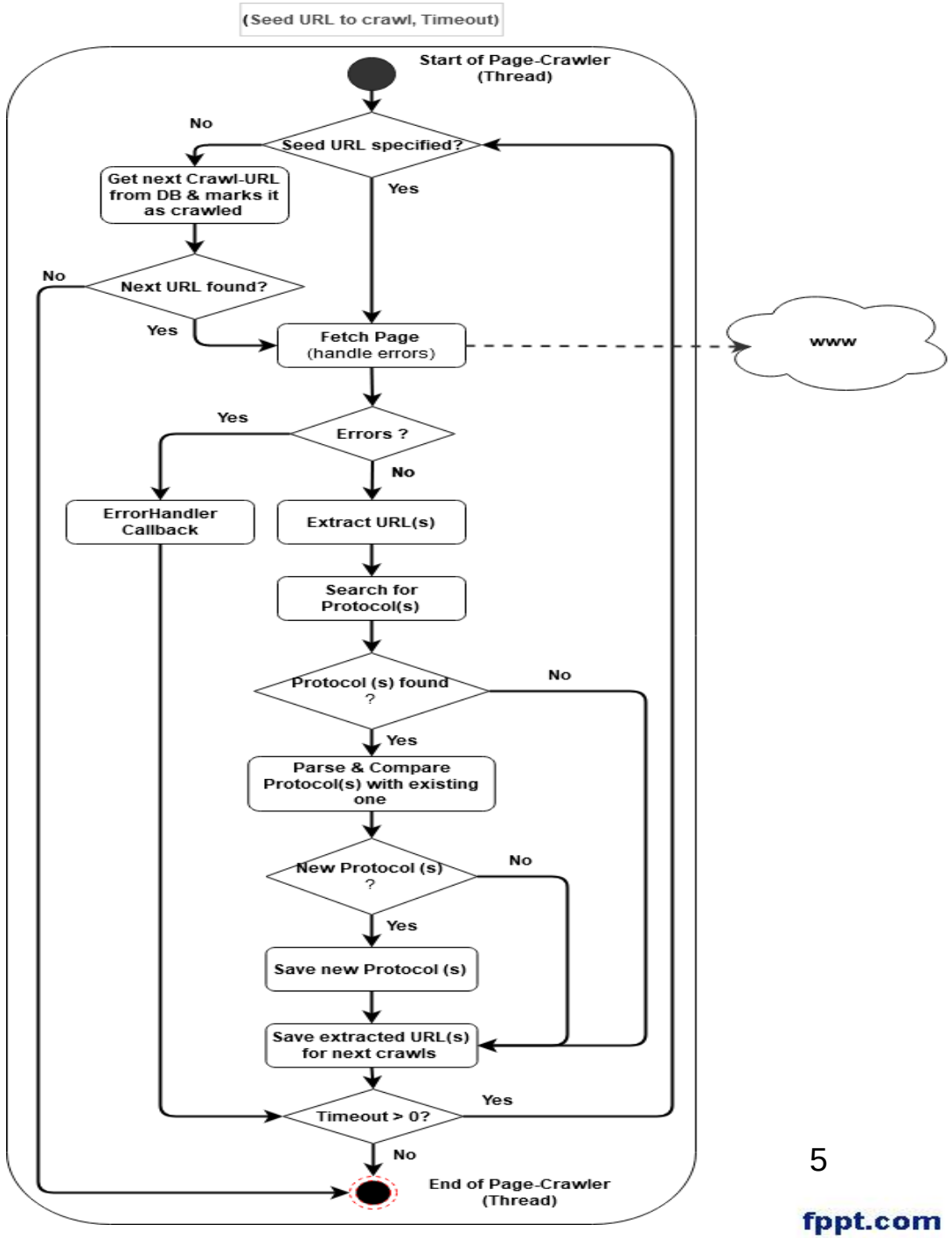
XML-Format

Was sind unsere Inputs?

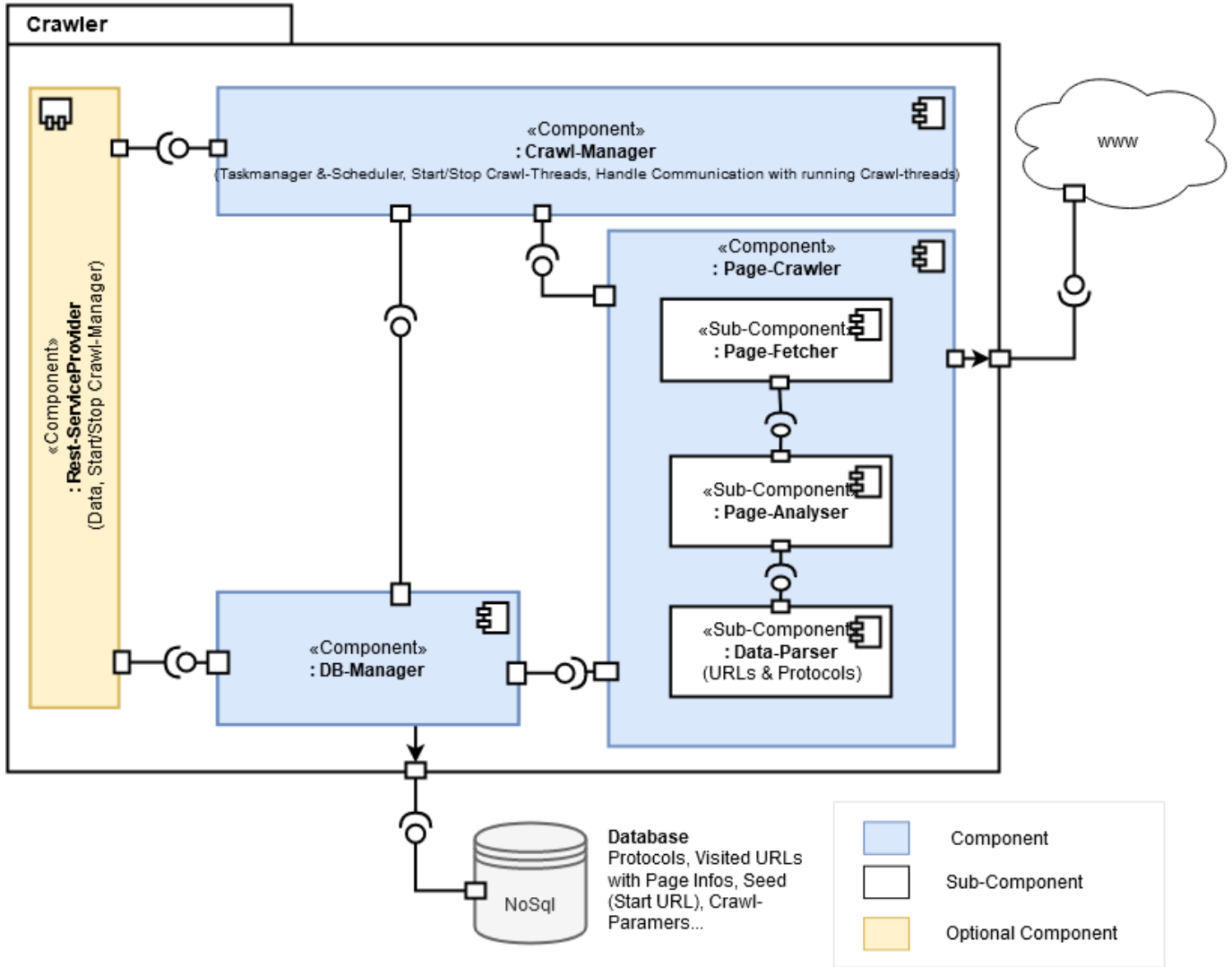
XML-Format



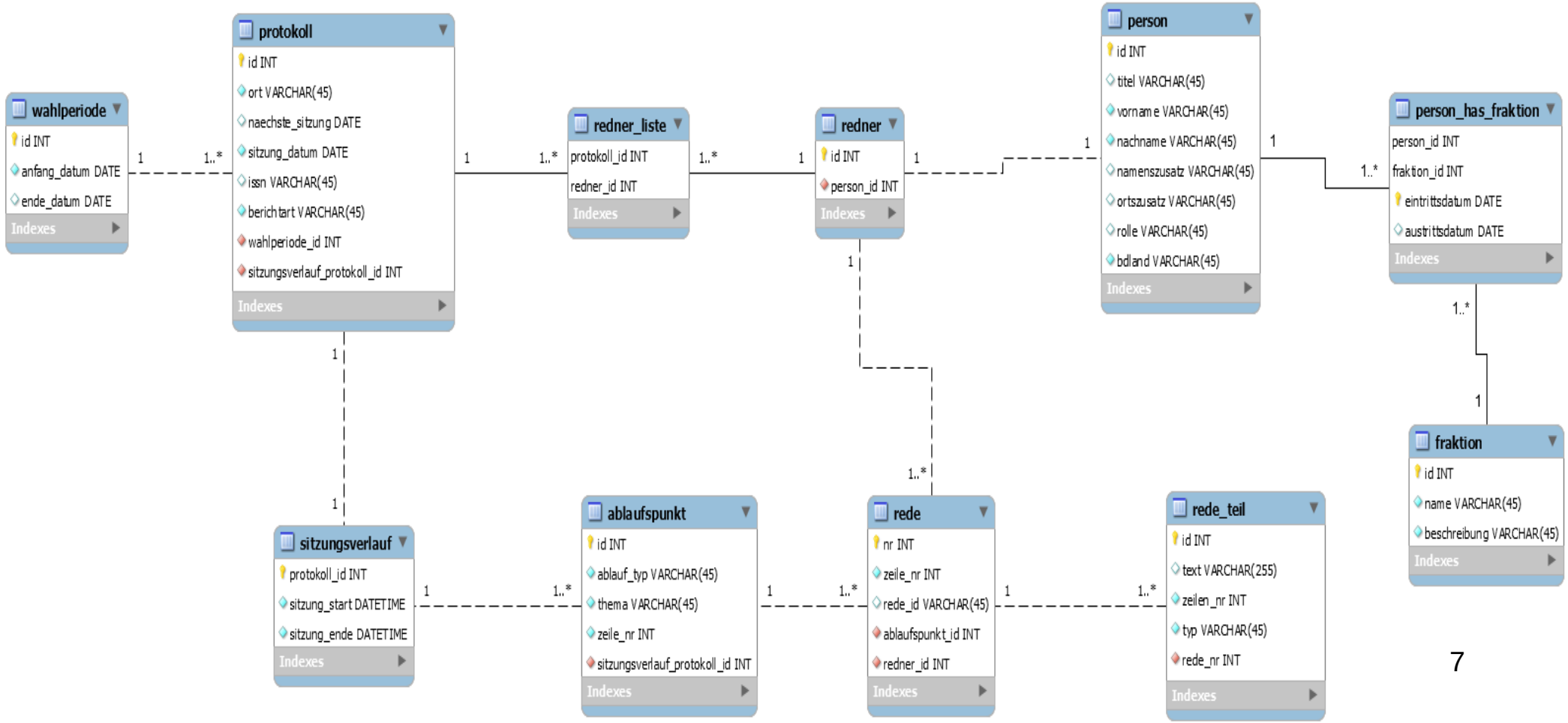
Wie wird unsere Crawler funktionieren?



Welsche Komponenten werden benötigt ?



Datenbank Model



Implementierung

1. **Service-Management** (Crawl-Manager & Rest-API)
2. **Logik** zum Page-Crawler
3. **Persistenz** (MongoDB & DB-Manager)

Welsche Technologien werden verwendet?

- Programmiersprache: Java
- Datenbank: MongoDB
- Sonstige noch zu erwähnen

Planung

Ziel bis zum **13. Nov**:

DB (MongoDB), DB-Manager : 50%
Page-Crawler (Erster Prototyp : 50%)
Task-Manager (Crawl-Manager & Rest-API : 50%)

Ziel bis zum **27. Nov**:

Vorläufiger Release (90%) zum Test an anderen Gruppen übergeben

Ziel bis zum **04. Dez**: finale Version und Anfang der Supervision

Danke für die Aufmerksamkeit !

Fragen ?