Projektbeschreibung o++oWiki

Die Wikipedia ist ein großer Schatz. Sie beinhaltet nicht nur Text- und Faktendaten, sondern auch Bilder und Audiodaten. Videos sind ebenfalls für die Wikipedia geplant. Die deutsche Version enthält etwa 1.6 Millionen Einträge (2015). Zur Zeit kann man jedoch nur einzelne Einträge ansehen oder eine Volltextsuche vornehmen. Die deutsche Wikipedia ist jedoch bereits jetzt recht gut strukturiert, weshalb man weit mehr Informationen extrahieren könnte, wenn eine geeignete Anfragesprache zur Verfügung stehen würde. Dazu zählen beispielsweise folgende Probleme:
Gib alle Bilder der Wikipedia, deren Titel das Wort „Bauhaus“ enthält.
Gib die Nationalhymnen aller europäischen Staaten.
Gib zu jedem Ort des Bördekreises das Geschichtskapitel und bilde daraus ein neues Dokument.
Gib mir alle französischen Flüsse die länger als 200 km sind und ins Mittelmeer fließen.
Sortiere die größten deutschen Aktiengesellschaften nach Umsatz pro Mitarbeiter.
…
Auch wenn diese Anfragesprache zunächst nicht ganz so leicht zu nutzen sein wird wie die jetzige Wikipedia, gehen wir davon aus, dass aufgrund der neuen Möglichkeiten langfristig Hunderttausende Deutsche unsere Sprache o++o in einem Umfang erlernen werden, der vielfältige Anfragen ermöglicht. Da man o++o nicht nur auf die Wikipedia anwenden kann sondern auch auf CSV- und XML-Dateien und in Zukunft sollen Anfragen an beliebige Datenbanken ermöglicht werden, ist der Nutzer eher bereit eine Anfragesprache zu erlernen. Unserer Meinung nach ist o++o leichter zu erlernen als SQL, SparQL und erst recht XQuery. Die angestrebte Implementation o++oWiki kann als Basis für anderssprachige Wikipedias und als Ausgangspunkt für Firmenwikis gesehen werden.
Bei Erfordernis sind Benutzergruppen zu qualifizieren.

Metadaten der Wikipedia

Die folgende TTD (Tabment Type Definition=syntaktisch verbesserte DTD) stellt einen kleinen Teil der Metadaten (das sind Daten über Daten; z.B: Spaltennamen von Tabellen) der deutschen Wikipedia dar. Diese Metadaten muss der Benutzer kennen, oder er muss Zugriff auf diese Daten haben. Ansonsten ist er nicht in der Lage, bestimmte Anfragen zu formulieren oder zu verstehen.

Wiki.de: (KEY, INFOBOX, (STAAT | STADT | FLUSS | LEICHTATHLET | SOFTWARE | …),
EINLEITUNG, KAPITELl, SIEHEAUCHl, LITl, WEBLINKl,WIKILINKl,...m)
KAPITEL: TITEL, INHALT
EINLEITUNG, SIEHEAUCH, LIT, WEBLINK: INHALT
INHALT: (TEXT | FETT | KURSIV | WIKILINK | URL | BILD | UTITEL l)
BILD: BTITEL, JPG
KEY, TITEL, UTITEL: TEXT
STAAT: EINWOHNER, FLAECHE, OFFIZIELLESPRACHEl,...
STADT: STAAT?, BUNDESLAND?, PROVINZ?, LANDKREIS?, GEMEINDE?, HOEHE,
FLAECHE, EINWOHNER, ….
FLUSS: GEWAESSERKENNZEICHEN?, LAGEm, FLUSSSYSTEM?, LAENGE?,
EINZUGSGEBIETl,QUELLE,STADTl,LINKERNEBENFLUSSl,
RECHTERNEBENFLUSSl, …
LEICHTATHLET: GEBURTSDATUM, GEBURTSORT, STERBEDATUM?, …
OLYMPIAl, WELTMEISTERSCHAFTl,
EUROPAMEISTERSCHAFTl, HALLENWELTMEISTERSCHAFTl,...
OLYMPIA, WELTMEISTERSCHAFT, EUROPAMEISTERSCHAFT,
HALLENWELTMEISTERSCHAFT : TEXT
…

INFOBOX enthält alle Namen der Infoboxen der Wikipedia, wie Fluss Stadt, Staat, Firma,… . Die Kapitel der TTD wurde bewusst nicht-rekursiv gestaltet, um die Formulierung der Anfragen zu vereinfachen.

Welche weiteren Wiki-anfragen ermöglicht werden, soll an konkreten Anfragen illustriert werden. Da alle Anfragen an die gleiche Datenquelle gerichtet werden, kann man auf die Angabe der Quelle hier „wiki.de“ verzichten, was bei den folgenden Beispielanfragen eingespart wird:

Wiki 1.otto: Gib die Namen der Einträge, die das Wort *Hadmersleben* enthalten.

```
avec Hadmersleben
gib  KEYm
```

Wiki 2.otto: Gesucht ist der *Hadmersleben* Eintrag der Wikipedia.

```
avec KEY = Hadmersleben
```

Wiki 3.otto: Gesucht sind alle Städte, die einen Link auf *Hadmersleben* enthalten.

```
avec INFOBOX = Stadt
avec WIKILINK = Hadmersleben
gib  KEYm
```

Wiki 4.otto: Fasse die Geschichtskapitel der Orte Hadmersleben, Alikendorf und Oschersleben zu einem neuen Dokument zusammen..

```
avec KEY in “Hadmersleben Alikendorf Oschersleben“
avec TITEL = Geschichte
gib  KEY,KAPITEL m
```

Wiki 5.otto: Gesucht sind die Geschichtsinformationen aller Orte des *Bördekreises*.

```
avec INFOBOX=Stadt
avec LANDKREIS = Börde
avec TITEL = Geschichte
gib  KEY,KAPITEL m
```

Wiki 6.otto: Gesucht sind zu jedem Ort des Bördekreises die Einwohnerzahl, wobei die Orte nach den Einwohnern zu sortieren sind.

```
avec INFOBOX = Stadt
avec LANDKREIS = Börde
gib  EINWOHNER,KEY m
```

Wiki 7.otto: Gesucht ist die Anzahl der Städte Deutschlands, die in der Wikipedia vorkommen.

```
avec INFOBOX = Stadt
avec STAAT = Deutschland
++1
```

Wiki 8.otto: In wie vielen WIKI-Einträgen kommt das Wort Bauhaus vor.

```
avec Bauhaus
++1
```

Wiki 9.otto: Wie viele Städte gibt es in der deutschen Wikipedia?

```
avec INFOBOX=Stadt
++1
```

Wiki 10.otto: Was ist eine Stadt?

```
avec KEY = Stadt
```

Wiki 11.otto: Gesucht sind alle Zusammenfassungen von großen spanischen Flüssen.

```
avec INFOBOX=Stadt
avec LAGE = Spanien & LAENGE >200
gib KEY,EINLEITUNG m
```

Wiki 12.otto: Gesucht sind alle Städte, durch die mehrere große Flüsse fließen.

```
avec INFOBOX=Fluss
avec LAENGE > 80
gib  STADT,KEYl m
avec KEYl ++1 > 1
```

Wiki 13.otto: Gesucht sind alle Teilnehmer von Olympiaden mit mehr als 3 Medaillen.

```
avec INFOBOX=Person
avec OLYMPIAl ++1 > 3
gib  KEY,OLYMPIAl m
```

Wiki 14.otto: Gesucht sind zu jedem Staat alle fetten und kursiven Wörter.

```
avec INFOBOX=Staat
gib  KEY,FETTm,KURSIVm m
```

Wiki 15.otto: Gesucht sind alle Einträge mit einer Literaturstelle, die *Martin* und *Luther* enthält.

```
avec KEY! “Martin Luther“ in LITl
```

Wiki 16.otto: Gib alle Bilder der Wikipedia, deren Titel das Wort „Bauhaus“ enthält (ohne Duplikate).

```
avec Bauhaus in BTITEL
gib  BILDm
```

...