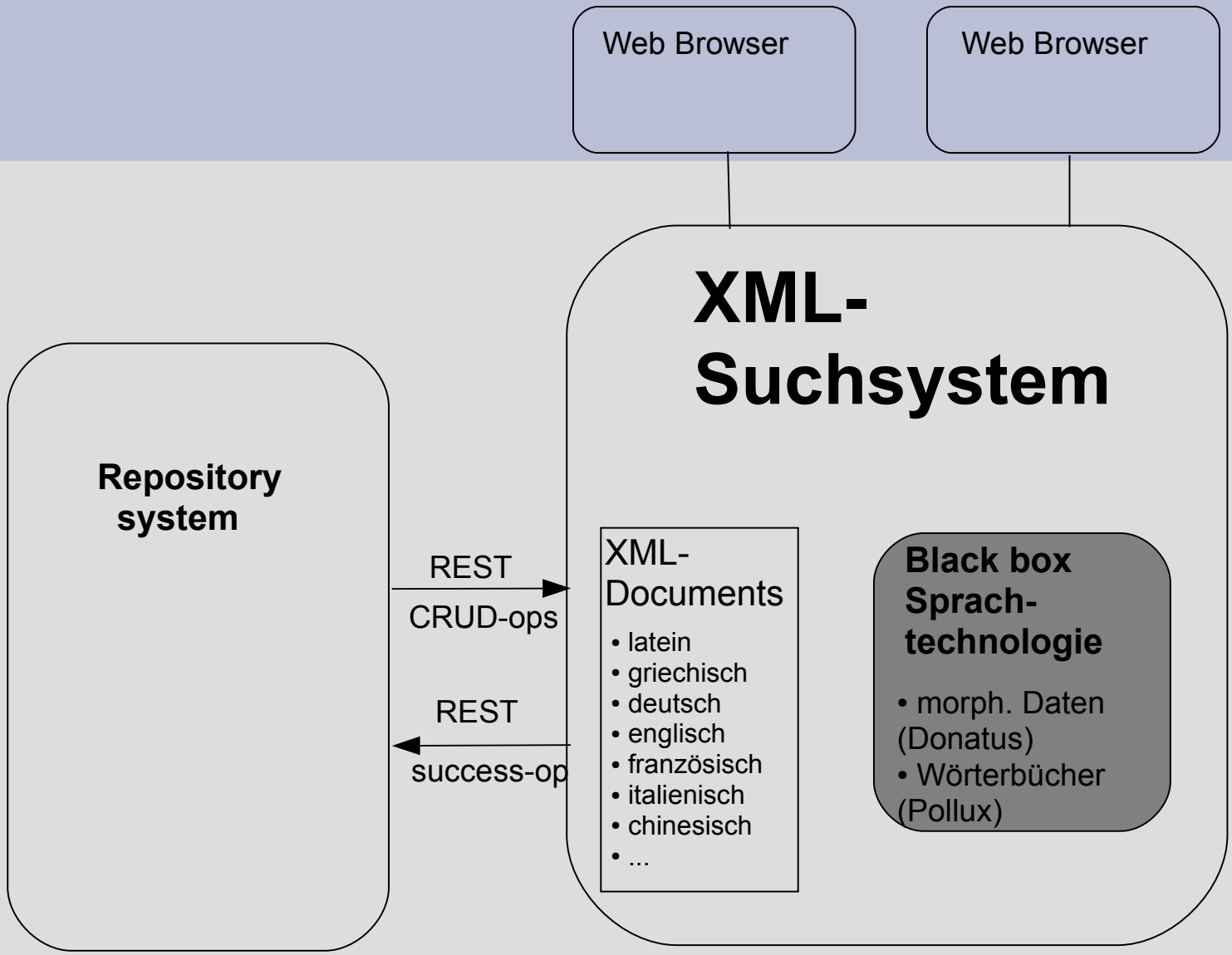
XML-Suchsystem/Sprachtechnologie

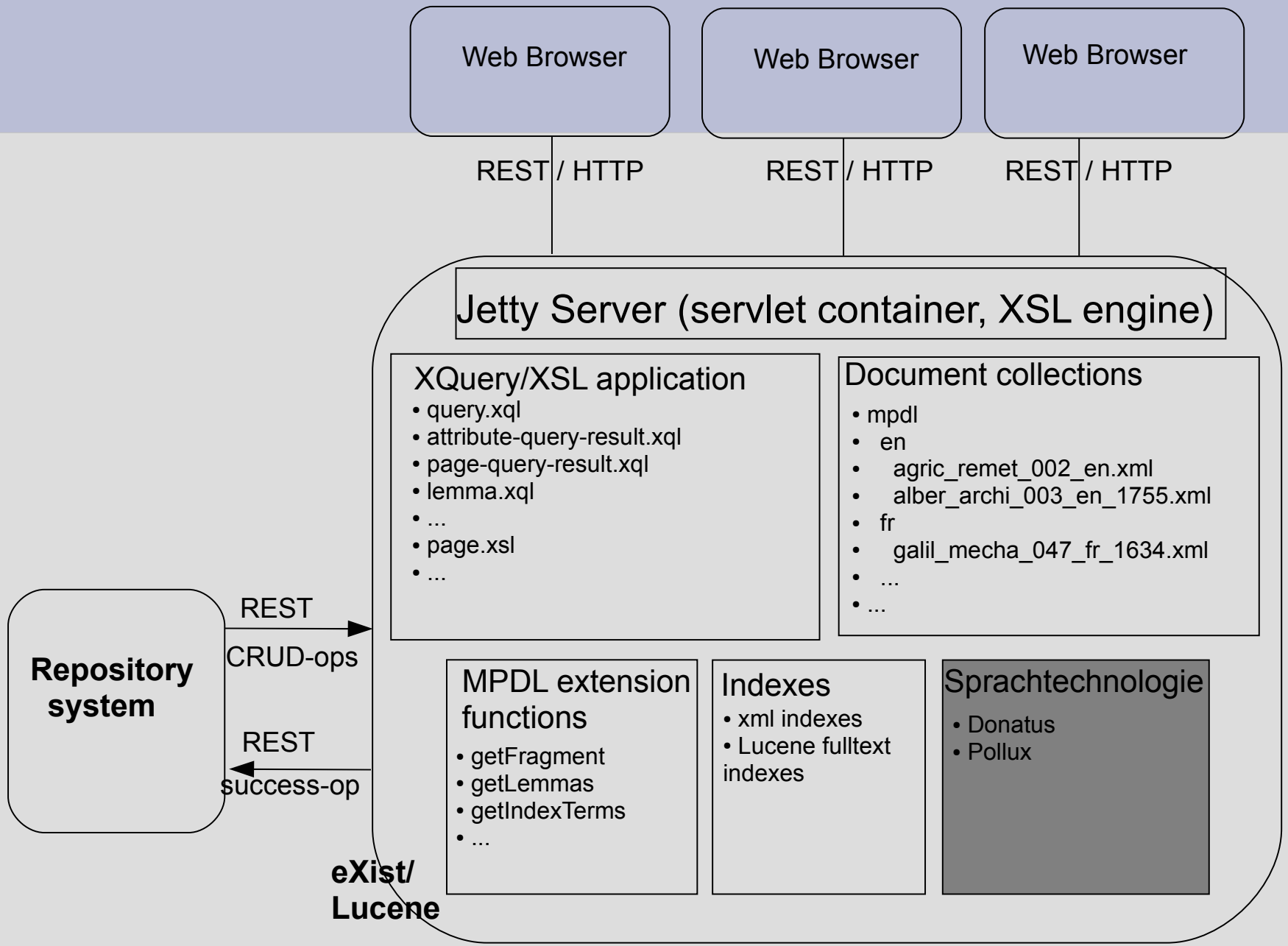
Inhalt

1. Architektur
2. Funktionalität
3. Sprachtechnologie
4. Vorführung

Architektur (1)



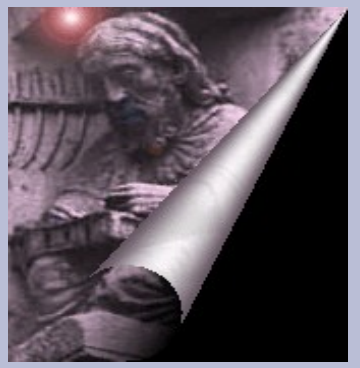
Architektur (2)



Funktionalität

- **Anzeige von XML Dokumenten**
  - Anzeige von einzelnen Dokumentseiten (XML fragment)
    - Anzeigemodi: Text, Text/Pollux, Image, XML
    - schnell: spezielle Erweiterung von eXist (getFragment)
  - Navigation im Dokument: Seite vor/zurück, gehe zu Seite
  - Integration von Digilib
  - Anreicherung von Buchseiten
- **Suchsystem**
  - Dokumentbasen: Archimedes, Echo, TEI
  - Suche in Dokumentbasen und in Dokumenten
    - boolesche Anfragen in Metadaten (Lucene-Anfragemächtigkeit)
    - (morphologische) Volltextanfragen (auch mit polyhierarchischen Lemmas)
    - strukturelle Anfragen: XPath und XQueries innerhalb von Dokumenten
    - Markierung in Suchergebnissen / Anzeige der morph. Varianten
  - Dokumentindex (angereichert mit Links zu Wörterbüchern wie Pollux oder Wikipedia und zu morph. Datenbasen)
  - Anzeige von morph. Daten (Donatus) und Wörterbüchern (Pollux)
- **CRUD Schnittstelle**
  - create, read, update, delete von Dokumenten
  - XML-RPC Schnittstelle (als Client Library verfügbar)
  - REST Schnittstelle zum Repository-System: Vorbereitungsphase

Sprachtechnologie (1) Donatus



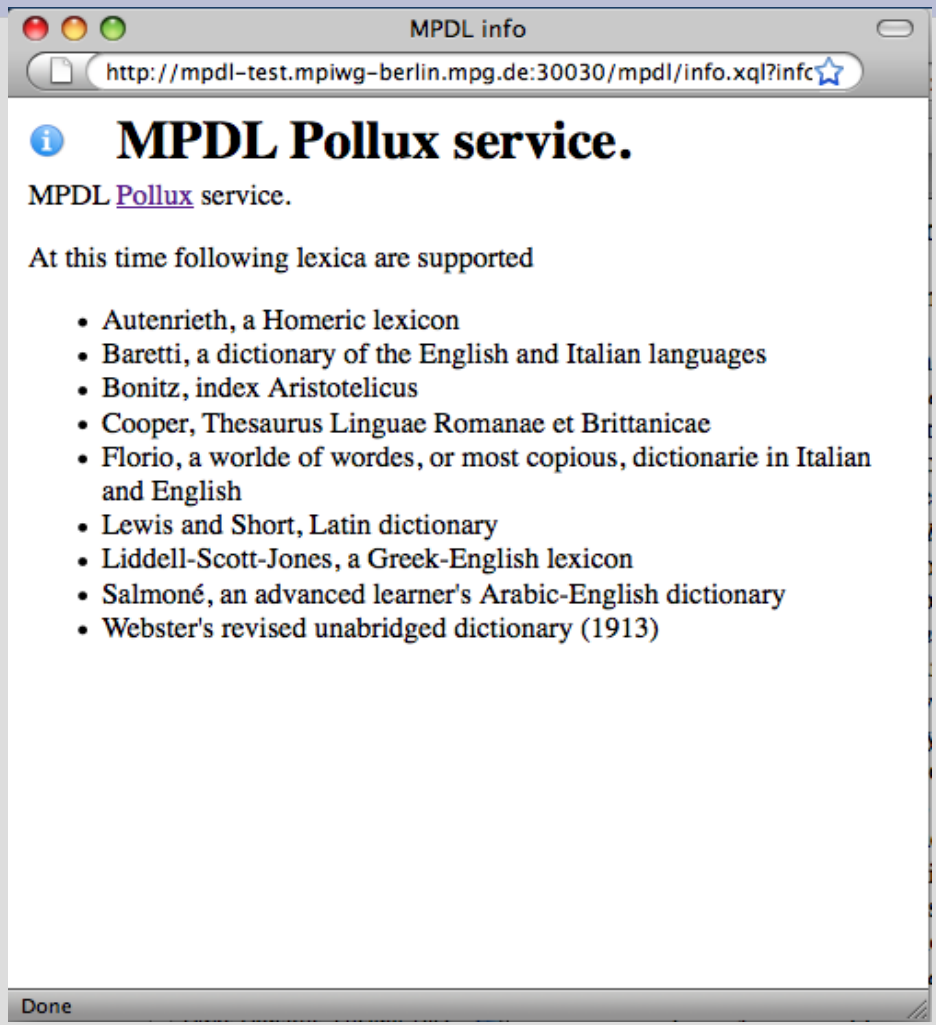
Was ist die Donatus Sprachtechnologie (siehe: archimedes.fas.harvard.edu)

- morphologische Datenbasis in den Sprachen arabisch, deutsch, englisch, französisch, griechisch, italienisch, lateinisch, niederländisch
- insgesamt ca. 3 Mio Wordformen mit ihren Lemmas und grammatik. Angaben
- Einsatz bei der morph. Indexierung von Dokumenten und der Suche mit morph. Einträgen (reine Java-Implementierung)
- spez. griechische und arabische Konvertierung nach Unicode
- Anzeige der morph. Erweiterung bei der Suche
- Anzeige von morph. Information im Dokumentindex
- Web-Schnittstelle:
  - z.B.: http://mpdl-proto.mpiwg-berlin.mpg.de/mpdl/lt/lemma.xql?language=la&form=accedo

Sprachtechnologie (2)
Pollux

Was ist die Pollux Sprachtechnologie (siehe: archimedes.fas.harvard.edu)

- neun digitale Wörterbücher/Thesauri in den Sprachen arabisch, englisch, griechisch, italienisch, lateinisch
- insgesamt ca. 460.000 Einträge



Sprachtechnologie (3)
Pollux

Was ist die Pollux Sprachtechnologie

- Datenbasis mit neun Wörterbüchern (BerkeleyDB)
- spez. griechische und arabische Konvertierung nach Unicode
- Anzeige einer Dokumentseite: Polluxmodus mit Links zu den Pollux-Einträgen (gewonnen über die morph. Datenbasis)
- Dokumentindex: Link zum Pollux-Eintrag
- Web-Schnittstelle
  - z.B.: http://mpdl-proto.mpiwg-berlin.mpg.de/mpdl/lt/lex.xql?language=la&form=accedo

Sprachtechnologie (4)

- Sprachtechnologie (Pollux, Donatus) ist modular (reines Java) implementiert und kann flexibel eingesetzt werden (in Clients oder in Servern)

- Gesamte entwickelte Software ist frei (open source)

Vorführung

- http://mpdl-proto.mpiwg-berlin.mpg.de/mpdl/query.xql
- http://mpdl-test.mpiwg-berlin.mpg.de:30030/mpdl/query.xql