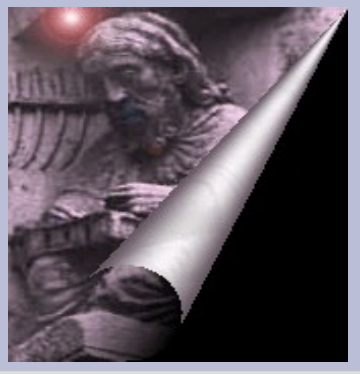
Integration von Donatus und Pollux in den MPDL-Prototypen: was fehlt (noch)

Inhalt: Lücken und Priorisierungen

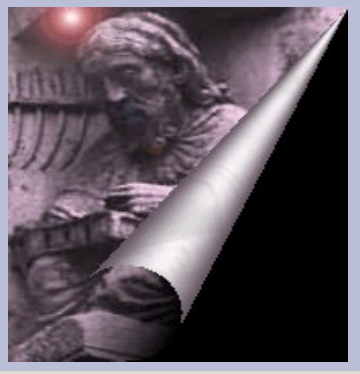
1. Donatus
2. Pollux

Donatus



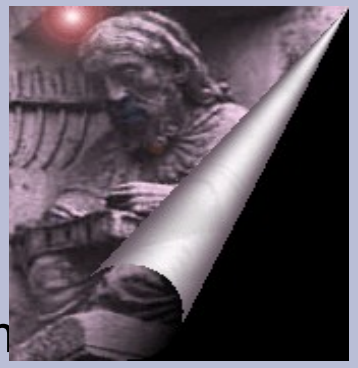
- Migration der Daten
  - 1. Schritt: Laden der primären Daten nach BerkeleyDB (Originaldaten, BackendDB's)
    - Perseus: aus den Sourcen von http://www.perseus.tufts.edu/hopper/ , dabei Betacode und Buckwalter Konvertierung nach Unicode (3 XML-Dateien, Stand Febr. 2010)
      - arabisch: 97.249 Formen
        - **1. Lücke**: Formen werden nicht korrekt erkannt: zu unterscheiden sind sog. Wurzeln und Formen von Wörtern (unterschiedl. Dateien in Perseus)
          - Lösung (zus. mit Mark Schiefsky): Mark sollte die Daten von Perseus besorgen. Danach Analyse wie sie importiert werden können. Wissen in Arabisch gefragt.
      - griechisch: 1.020.846 Formen
      - lateinisch: 710.620 Formen
    - Celex: aus den Daten der CD geholt von archimedes.fas.harvard.edu (6 CD-Text-Dateien, Stand: ?)
      - niederländisch: 381.275 Formen und 124.136 Lemma
      - englisch: 160.595 Formen und 52.447 Lemma
      - deutsch: 365.530 Formen und 51.728 Lemma

Donatus



- Migration der Daten (Forts.)
  - 1. Schritt (Forts)
    - Donatus-Speziell: von archimedes.fas.harvard.edu (2 Text-Dateien, Stand: ?)
      - französisch: 306.795 Formen
      - **2. Lücke:** italienisch: 49.265 (?) Formen (werden noch nicht importiert und verwendet, kompliziertere Struktur der Datei: ital.hash);
        - Lösung
          - BackendDB (BerkeleyDB-C) von archimedes.fas.harvard.edu holen und importieren oder
          - Datei ital.hash konvertieren und importieren
  - 2. Schritt: Laden der sekundären Daten nach BerkeleyDB (Donatus supplements, Fallback-DB's)
    - alle Sprachen wie bei primären Daten außer niederländisch von archimedes.fas.harvard.edu (7 CSV-Dateien, Stand: ?)

Donatus

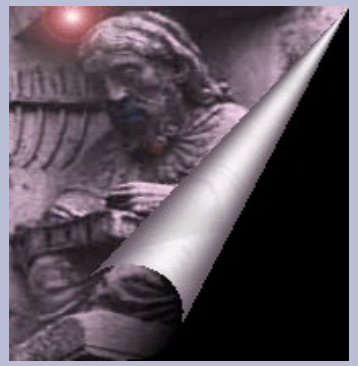


- **3. Lücke:** es fehlen in allen Sprachen Wortformen im Vergleich zum Archimedes-System: z.B. „proportionum“
  - Analyse
    - im Archimedes-System scheint ein „anderer“ primärer Datenbestand zu sein (sogenannte Backends). In ihm scheinen aber generell weniger Formen als in den Originaldaten enthalten zu sein. Z.B. sind in dem lateinischen Bestand (cache-la.db) nur ca. 154.000 Formen. Jedoch sind auch Formen enthalten, die nicht in den Perseus-Daten enthalten sind (z.B. „proportionum“).
  - Lösung: Konvertieren aller Backend-DB's und alle bisher noch nicht in der BerkeleyDB enthaltenen Wortformen hinzufügen (als Donatus-Supplements)
- **4. Lücke:** orthographische Normalisierung der Einträge
  - zum grossen Teil schon ok wie es jetzt ist, wollen wir das weiter verbessern ?
  - falls ja:
    - evtl. sind die Schreibweisen in den Pollux-Wörterbüchern anders, und dann müssten auch die Wörterbuch-Keys (Pollux) normalisiert werden
    - deutsch: Umlaute auflösen
    - französisch: Accents auflösen
    - ...

Donatus

- **5. Lücke:** grammatische Angaben sind im Originaldatenbestand verfügbar, werden aber noch nicht importiert und angezeigt (alle 8 Sprachen)
  - 1. Lösung: verschiedene Formate für Grammatiken vereinheitlichen in BerkeleyDB als Extra-Feld einfügen und in der Oberfläche anzeigen
    - möglich ?
    - falls ja: schwierig, aber dann keine Abhängigkeit mehr zum Datenbestand in den BackendDB's
  - 2. Lösung: sie sind evtl. vereinheitlicht in den BackendDB's enthalten
    - Problem: es sind in den BackendDB's aber wohl weniger Daten enthalten als in den Originaldaten
- **6. Lücke**: Browsing/Navigation in den morph. Beständen fehlt komplett, dies ist gerade auch für die MPDL-München wichtig
  - Lösung: realisieren
    - Anforderungsdefinition: alph. Suche, Links zwischen den Einträgen, gramm. Angaben, Text/Donatus-Anzeige einer Buchseite, GUI etc.
    - Realisierung: saubere Lösung implementieren, der als Software/Service auch für andere nutzbar ist
      - eXist: icke
      - GUI: Andrey

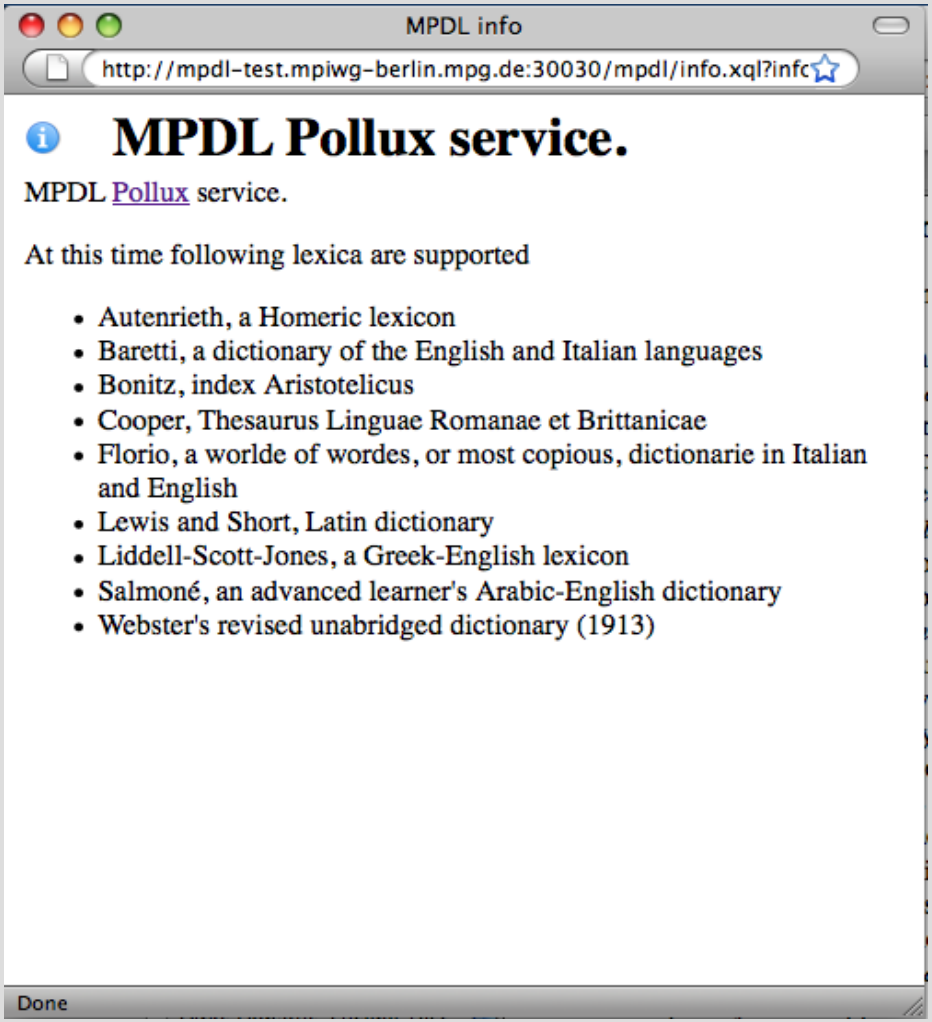
Donatus



- **Priorisierung**
  - noch relativ schnell zu beheben (und bringt auch gleich sichtbaren Erfolg)
    - 2. Lücke (italienische Formen)
    - 3. Lücke (BerkeleyDB-C: dumpen und neue Einträge finden und importieren)
  - langwierig sehr wichtig zur Akzeptanz bzw. für MPDL München etc.
    - 6. Lücke (Browsing/Navigation im Bestand)
  - langwierig aber eher speziell
    - 1. Lücke (arabische Formen): momentan nur 1 Archimedes-Dokument enthalten
    - 4. Lücke (orth. Normalisierung)
    - 5. Lücke (gramm. Angaben)

Pollux

- 9 Wörterbücher in den Sprachen: arabisch, englisch, griechisch, italienisch, lateinisch
- Migration
  - alle BerkeleyDB-C Wörterbücher nach BerkeleyDB-Java Datenbank konvertiert
  - HTML-Fehler in vielen Datensätzen bereinigt
  - Betacode und Buckwalter Kodierung nach Unicode durchgeführt



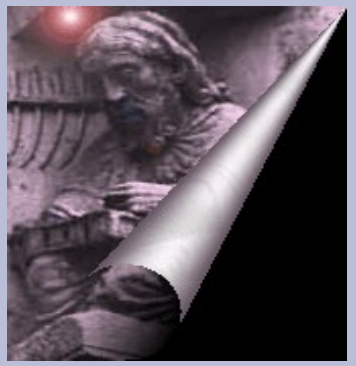
Pollux

- **1. Lücke**: es fehlen Wörterbücher, die in Pollux als Online-Schnittstelle zur Verfügung standen
  - deutsch
    - Wörterbuch der deutschen Gegenwartssprache
    - Deutsches Woerterbuch von Jacob und Wilhelm Grimm
  - französisch
    - Dictionnaire de l'Académie francaise, 6e éd
  - sumerisch
    - ePSD (Pennsylvania Sumerian Dictionary)
- Lösung
  - Prüfen, ob die Wörterbücher als Daten zur Verfügung stehen
  - Falls ja: importieren
  - Falls nein: evtl. eine Online-Schnittstelle vorsehen

Pollux

- **2. Lücke**: arabische Einträge werden über die morph. Keys nicht gefunden, da es Wurzeln (?) sind
  - Lösung: siehe 1. Lücke in Donatus (wird dort miterledigt)
- **3. Lücke**: Browsing/Navigation in den Wörterbüchern
  - Lösung: siehe 6. Lücke in Donatus

Pollux



- **Priorisierung**
  - langwierig aber sehr wichtig zur Akzeptanz
    - 3. Lücke (Browsing/Navigation im Bestand)
  - langwierig (evtl. später)
    - 1. Lücke (weitere Wörterbücher integrieren: deutsch, franz., sumerisch)