Karlsruher Institut für Technologie

Automatische Metadatenanalyse im beweissicheren elektronischen Laborbuch

5. BeLab-Workshop 2011

Steinbuch Centre for Computing (SCC)

Agenda

- Klassifikation in BeLab
- Übersicht über implementierte Module
- Framework und Ergebnisspeicherung
- Umsetzung
- Fazit

Klassifikation in BeLab

- **Signatur**
  - B1 - keine
  - B2 - fortgeschritten
  - B3- qualifiziert
  - B4 - qualifiziert+ (signierter Zeitstempel)
  - B5 - akkreditiert (Signatur eines Zertifikatanbieters)
  - B6 - akkreditiert+
- **Dateiintegrität**
  - S1 - ungesicherte Datenerzeugung
  - S2 - gesicherte Datenerzeugung
- **Datenformat**
  - L1 - ungeeignet
  - L2 - geeignet (PDF, TIFF, DOC, DOCX, XLS, XLSX, ZIP)
  - L3 - empfohlen (ASCII, XML, PDF/A, PDF/E, PDF/UA, PDF/X, TIFF, SVG, ODF, TAR)

Modulübersicht

- Dateiformat
  - → Modul zur Analyse des Dateityps
- Signatur
  - → Module zur Verifikation unterschiedlicher Signaturen
- Dateiintegrität
  - bereits bestehendes Modul zur Verifikation von Kappa-Dateien
  - → Modul zur Dateifolgenanalyse

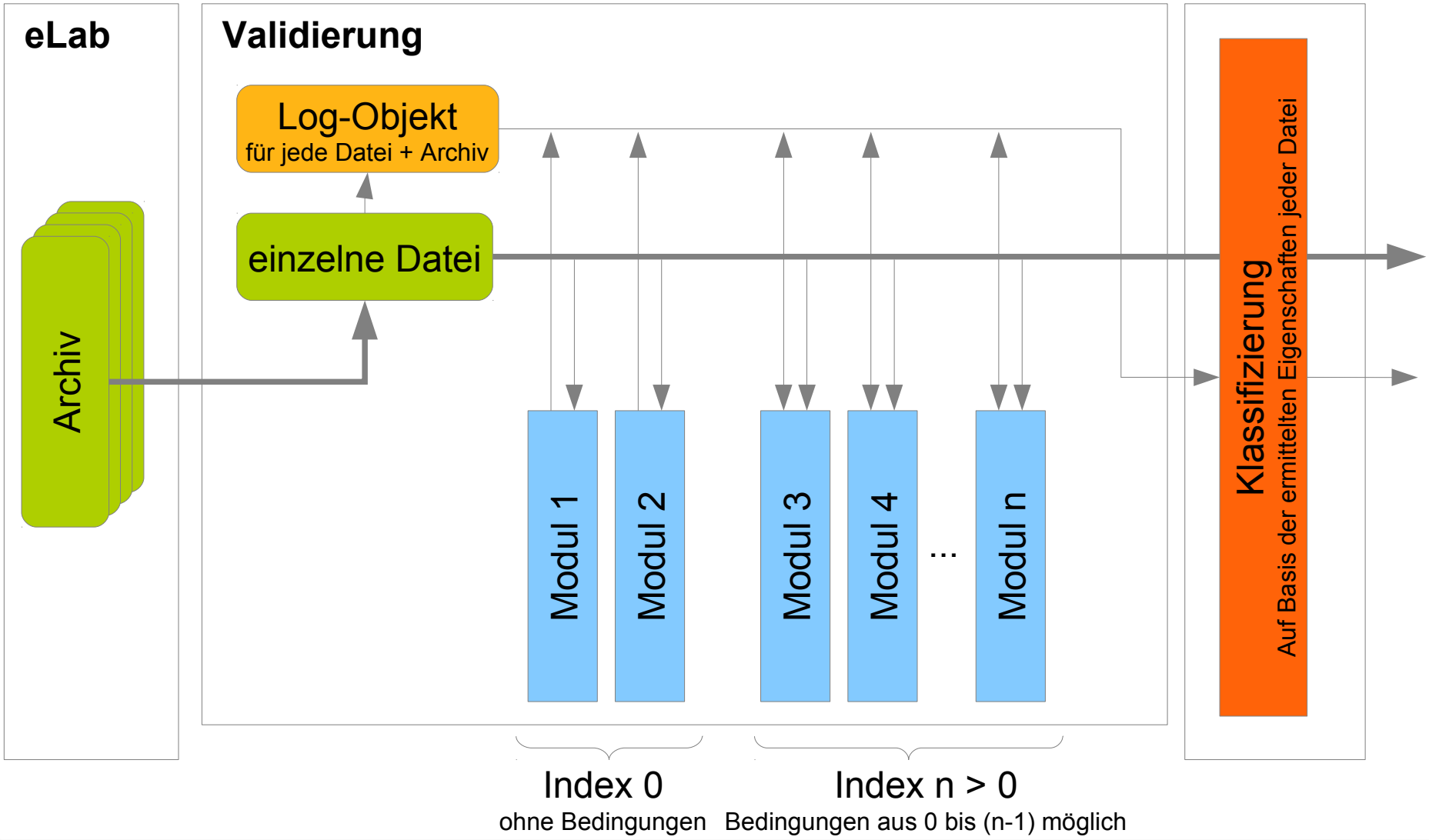
Framework

- Anforderungen:
  - Analyse auf Dateiebene
  - geeignete Auswahl entsprechender Module je Datei
  - Prüfungen bauen auf Ergebnissen anderer Prüfungen auf
    - Lese- und Schreibzugriff auf den Ergebnisspeicher
    - geeignete Reihenfolge der Prüfung notwendig

- Ergebnisspeicherung (Log-Objekt)
  - auf Dateiebene
  - enthält mehrere Einträge:
    - `(Class type, String key, Object content)`
  - Lese- und Schreibzugriff über Identifikation `key` möglich

- Modulverwaltung über XML-Datei
  - Beinhaltet jeweils Name der Klasse und Aufrufbedingungen

Framework


eLab
Validierung
Log-Objekt
für jede Datei + Archiv
einzelne Datei
Archiv
Modul 1
Modul 2
Modul 3
Modul 4
...
Modul n
Klassifizierung
Auf Basis der ermittelten Eigenschaften jeder Datei
Index 0
ohne Bedingungen
Index n > 0
Bedingungen aus 0 bis (n-1) möglich

Implementierung der Formatanalyse

- Benchmark verschiedener Frameworks
  - jMimeMagic, MimeUtil, MimetypesFileTypeMap
  - Erkennungsraten zwischen 10% und 56.6%
  - Fehlerraten zwischen 3.3% und 33.3%

- → Kombination zweier Bibliotheken
  - 1. Prüfung durch Bibliothek mit der geringsten Fehlerrate
  - 2. bei Standardrückgabewert Prüfung durch eine weitere Bibliothek
  - → Steigerung der Erkennungsrate auf 73,3% bei 3,3% Fehlerrate

- Problem: Erkennung von ODT und DOCX als ZIP, PKCS#7 als Text
  - → Einführung einer Ersetzungstabelle auf Basis der Dateiendung
  - z.B. erkanntes Format „text/plain“ und Endung „.sha1“
    - neues Format „application/pkcs7-mime“

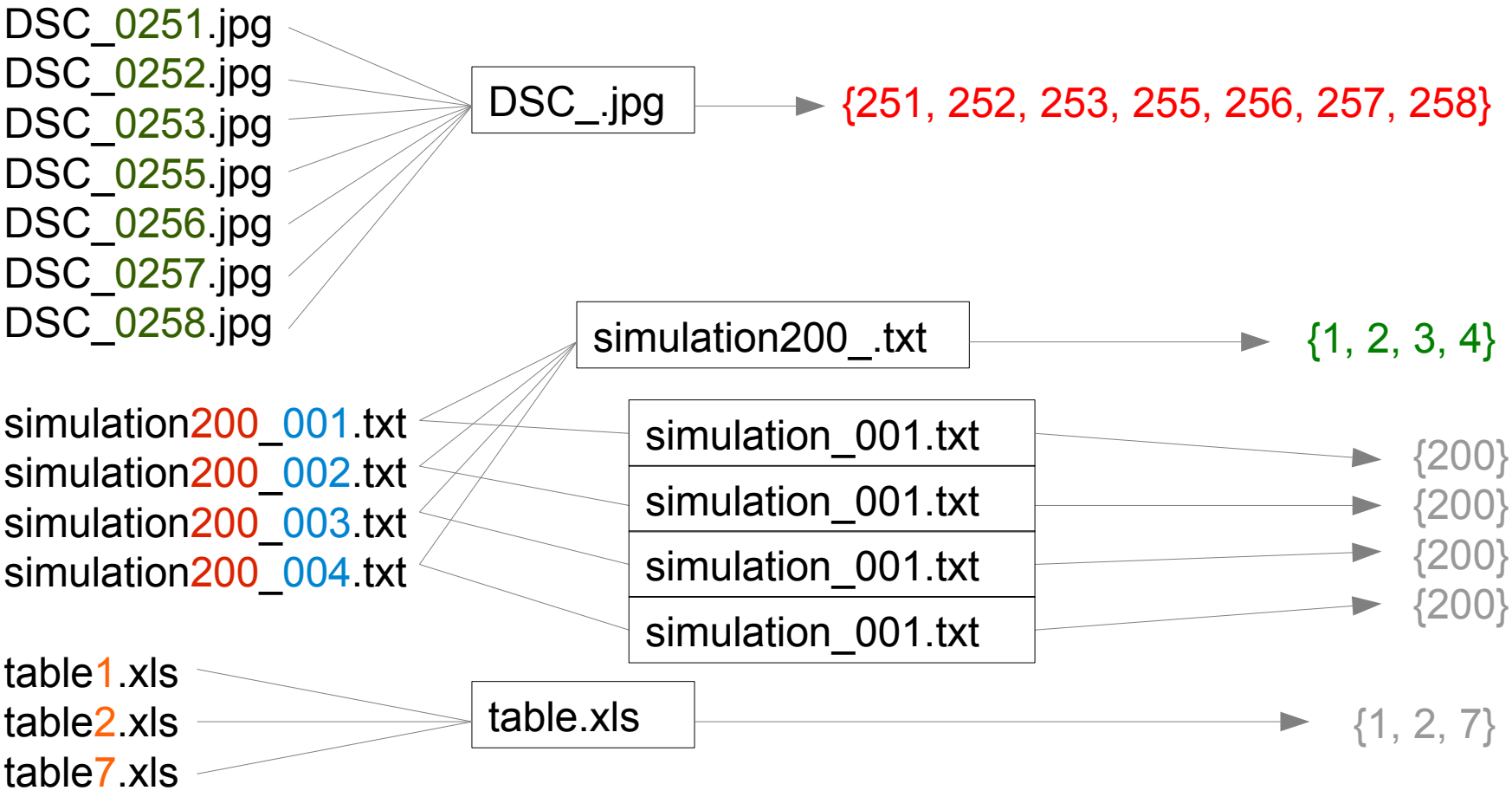
Implementierung der Signaturanalyse

- Problem: verschiedene Arten der Signatur möglich
  - XMLDSig in OpenDocument und Office Open XML
  - PKCS#7 in PDF
  - externe Signaturen auf PKCS#7-Basis

- BeLab-eigenes Werkzeug zur Verifikation von XMLDSig-Signaturen
  - Problem: Entpacken des ZIP-Archives in ein temporärer Verzeichnis nötig
  - Auswertung von Signaturen in Office Open XML nicht möglich
    - Änderung der Spezifikation zur Kanonialisierung durch Microsoft

- BouncyCastle Crypto API zur Verifikation von PDF-internen und externen Signaturen

Implementierung der Folgenerkennung



Fazit

- Aussage in den Klassifikations-Bereichen „Signatur“ und „Dateiformat“ möglich
  - Bis auf Office Open XML zuverlässige Erkennung
- Robuste Möglichkeit zur Bestimmung fehlender Dateifolgegliedern
  - Damit Erweiterung des Bereichs „Dateiintegrität“
- Leicht zu erweiterndes Framework, dadurch einfache Implementation weiterer Analysemodule möglich

Karlsruher Institut für Technologie

Fragen?