



Das SCAPE Projekt: Langzeitarchivierung und Skalierbarkeit

Teil 1: Überblick über das SCAPE Projekt

Dr. Sven Schlarb

Österreichische Nationalbibliothek

SCAPE ½ Informationstag

05. Mai 2014, Österreichische Nationalbibliothek.

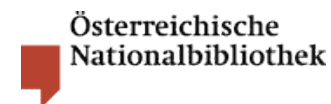
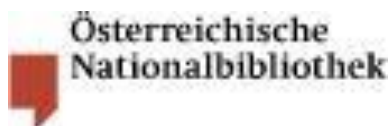
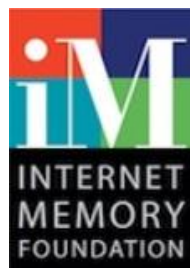
- “Big data” in Bibliotheken
- SCAPE-Projekt Überblick
- Themenbereiche

- „Big data“
- VVV: Volume (Datenmenge), Variety (Datenvielfalt), Velocity (Geschwindigkeit)
- Skalierbarkeit in mehrererlei Hinsicht
 - Anzahl
 - Größe
 - Komplexität
 - Vielfalt der Datenformate
- IT Umgebungen (Hardware, Software, Systeme), welche eine (kostengünstige) Verarbeitung großer Datenmengen ermöglichen

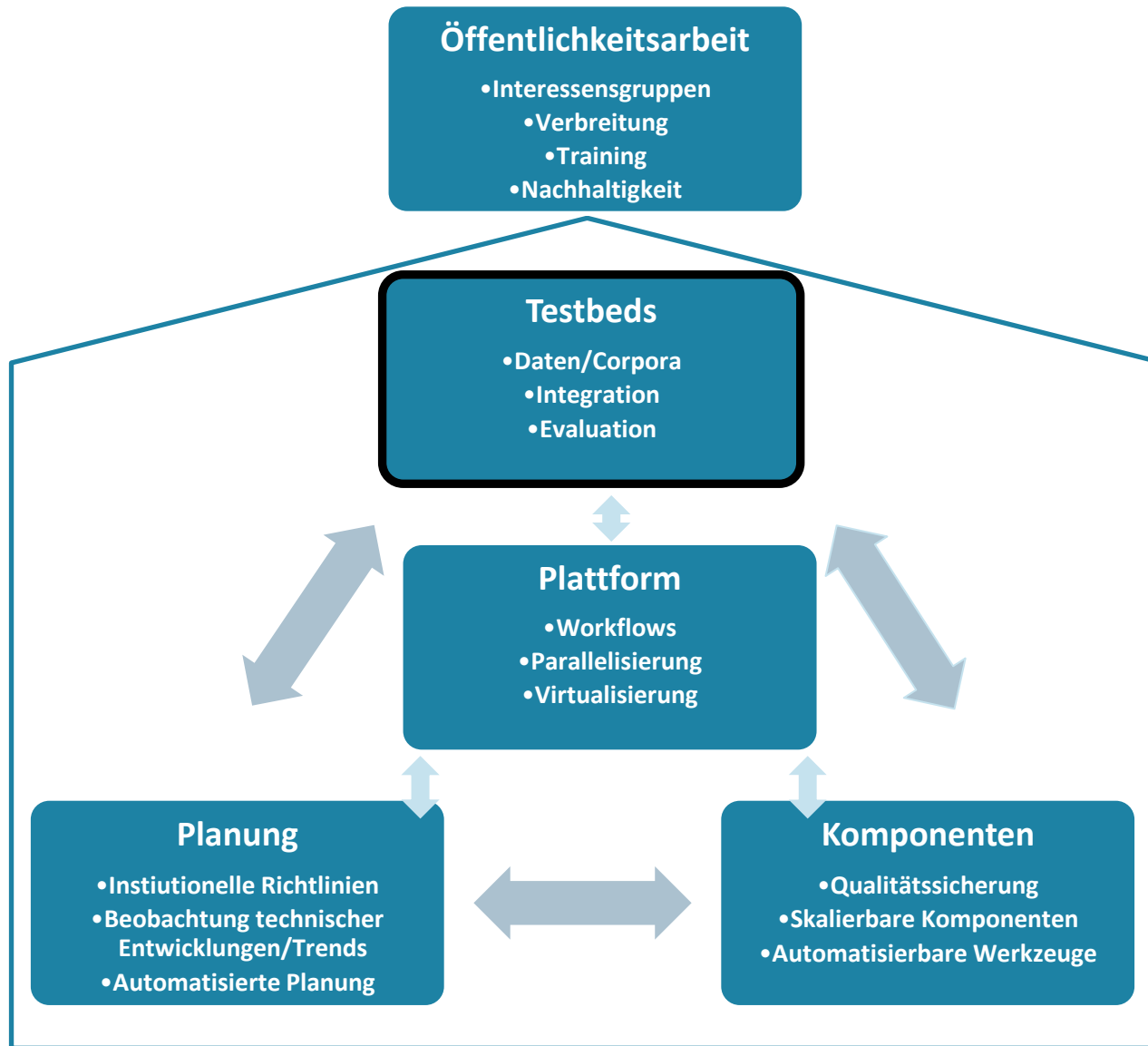
- Google-Books-Projekt: 30 Millionen digitale Bücher
 - <http://www.nybooks.com/articles/archives/2013/apr/25/national-digital-public-library-launched>
- Europeana: Metadaten von über 24 Millionen Objekten
 - Europeana annual report and accounts 2012, Europeana Foundation, April 2013
- Hathi Trust: 10 Millionen Bände (über 5,6 Millionen Werke) mit insgesamt über 3,7 Milliarden Buchseiten
 - http://www.hathitrust.org/statistics_info
- Internet Archive: 364 Milliarden Seiten, die ca. 10 Petabyte Speicher belegen.
 - <http://archive.org> und <http://archive.org/web/petabox.php>

- SCAPE ist ein Nachfolge-Projekt des Planets-Projekts (FP6)
- Framework Program 7 (FP7) - Collaborative Project
- Laufzeit: 44 Monate
 - Februar 2011 – September 2014
- Budget: 12.0 Millionen Euro
 - Gefördert: 9.2 Millionen Euro

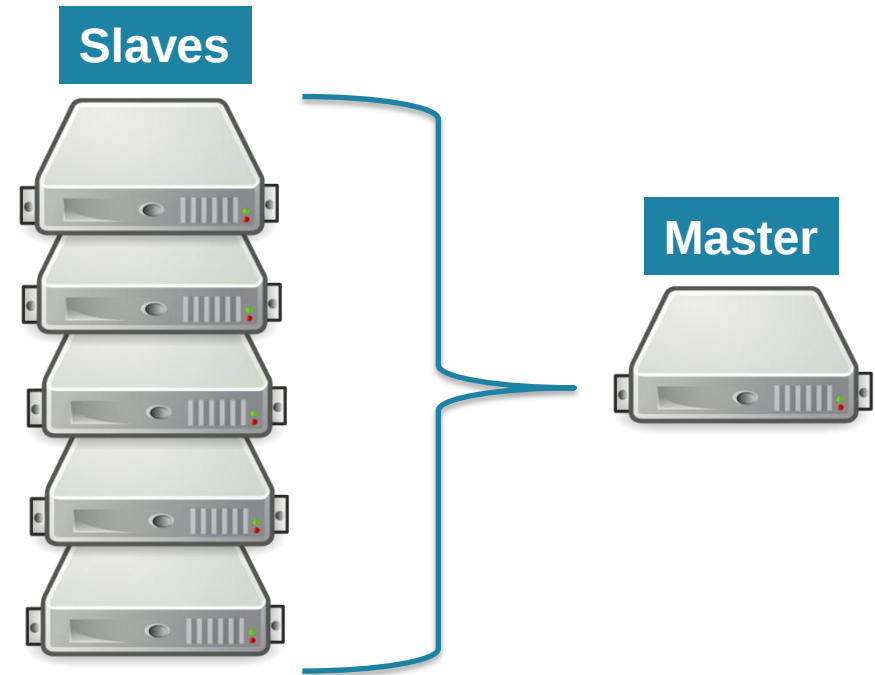
SCAPE Konsortium



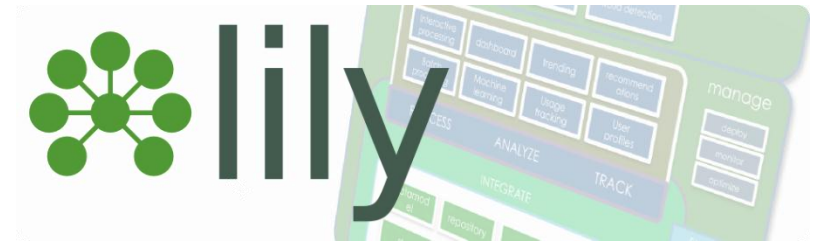
Übersicht über das SCAPE Projekt



- Basis
 - Computer-Cluster
 - Virtualisierung (XEN/Eucalyptus)
- Softwareverteilung
 - Debian Pakete
 - Tool Spezifikation
- Auftragsausführung
 - Apache [Hadoop](#) (HDFS/MapReduce) 
 - Apache [Oozie](#) (Workflows) 

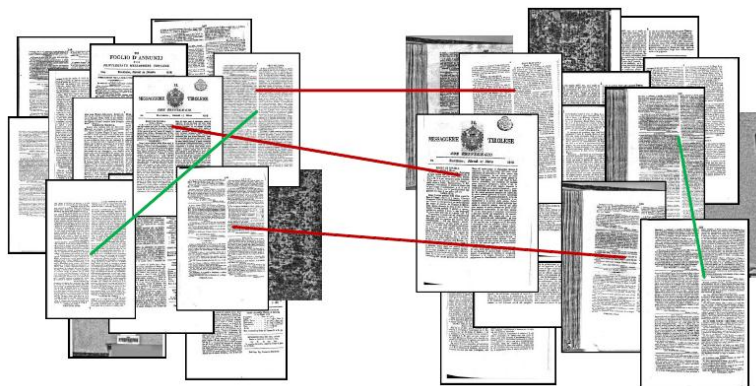


- Fedora 4.0.0
 - Nur REST, kein SOAP
 - Auf der Basis von RDF
 - JCR 2.0 Implementierung – ModeShape (Jboss)
 - Infinispan (Jboss) verteilter NoSQL Datenspeicher
- Lily 2.0
 - Basierend auf HBase/HDFS
 - Integrierte Datenverarbeitung und –speicherung
- MongoDB (NoSQL)
 - Dokumentenorientiert (JSON)

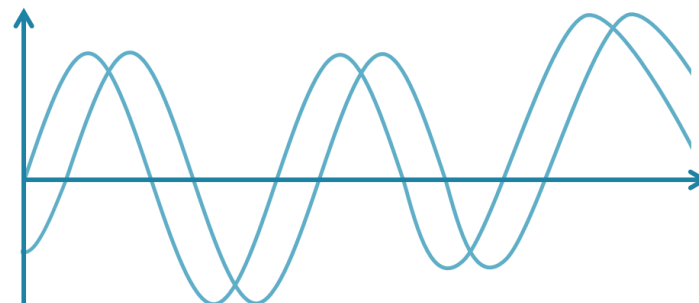


In SCAPE entwickelte Software-Werkzeuge

Matchbox – Duplikat-Bilderkennung



XCorrSound – Audio-Vergleich

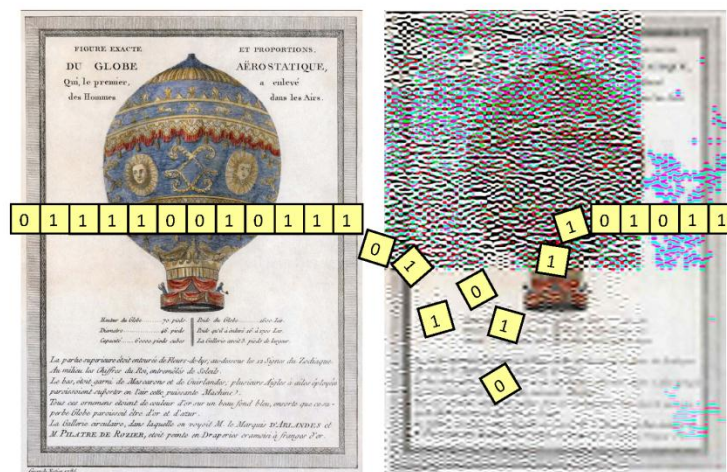


Pagelyzer – Web-Inhalte vergleichen



■ no change between versions
■ changes between versions

Jpylyzer – JPEG 2000 Validierung



- **Tool Wrapper**

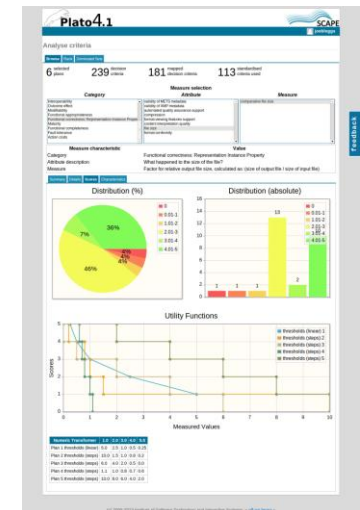
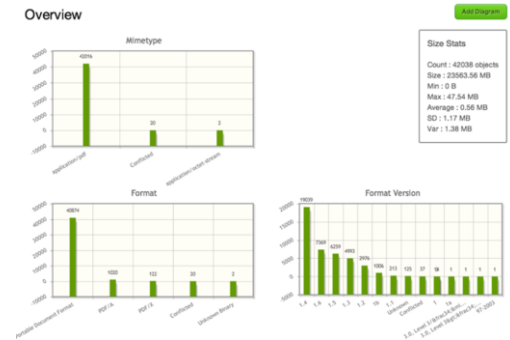
- Integration bestehender Software-Werkzeuge in die SCAPE Plattform
 - **Standard-Namensgebung** von Werkzeugen für Dateianalyse, Migration und Qualitätssicherung
 - **Standardisierte Ausführung** (Kommandozeile)
 - **Einfache Einrichtung** auf dem Cluster dank der Bereitstellung von Debian-Paketen
 - **Unterstützt Datenströme**, die für Hadoop-Verarbeitungsaufträge nützlich sind.
- Generiert Langzeitarchivierungskomponenten
 - **Taverna Workflow-Komponenten** inklusive Metadaten, die das Auffinden erleichtern
 - **Automatischer Veröffentlichung** von Komponenten auf **myExperiment**
 - **Standardisierte Eingabe und Ausgabe**, die das Kombinieren der Workflow-Komponenten vereinfacht

- **Langzeitarchivierungswerkzeugkasten**

- Werkzeugsammlung für die Langzeitarchivierung
- Über 80 Operationen für Migration, Dateianalyse, und Qualitätssicherung
- Einfaches deployment unter Linux (via *apt-get*)

- *apt-get install digital-preservation-tools*

- **SCOUT: Automatisierte Überwachung der Langzeitarchivierung**
 - Einbindung externer Informationsquellen
 - Fall-basierte Benachrichtigungsfunktionen
- **c3po: Analyse von Archivinhalten**
 - Scale-out MongoDB (Anzahl der Knoten beliebig erweiterbar)
 - Auswertung von FITS-Datei-Analyseergebnissen
 - Detaillierte Statistiken
 - Möglichkeiten der Stichprobenbildung
- **PLATO 4.1: Planung in der Langzeitarchivierung**
 - www.ifs.tuwien.ac.at/dp/plato
 - Weiterentwicklung der Version aus dem Planets-Projekt
 - Neue Funktionalität
 - Gemeinsame Planung der Langzeitarchivierung in Gruppen
 - Integration maschinenlesbarer Richtlinien



SCAPE Architektur

