elasticsearch

# Elasticsearch in Forschungsinfrastrukturen der Sozial- und Bibliothekswissenschaften

## Cloud-skalierbare Suche in Volltexten und strukturierten Daten

Oliver Schmitt

TMF Workshop, Text-Mining für die medizinische Forschung

28.01.2015, Berlin

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

# GWDG
## Vorstellung

### Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG)

- Gründung 1970 durch Max-Planck-Gesellschaft und Land Niedersachsen
- Heute: 125 Beschäftige inkl. 35 Wissenschaftler
- 30 drittmittelgeförderte Forschungsprojekten

### Kundenkreis

- **Max-Planck Gesellschaft**
  80 Max-Planck Institute in Deutschland – fünf in Göttingen – Beschäftige: 25.000
- **Universität Göttingen**
  200 Institute und 13 Fakultäten – 18.000 Beschäftigte und 25.000 Studenten
- **Externe Forschungseinrichtungen**

### Schwerpunkte

- **Rechenzentrumsbetrieb** (Scientific Computing, Storage, Netzw,erk, Cloud, Virtualisierung, Kollaborationsplattformen, Applikationshosting, …)
- **IT-Forschungseinrichtung**
  - Datenmanagement und Big Data (*Cloud4e*)
  - Cloud Computing, Virtualisierung, SDN, Resource Management
  - High Performance Computing, Scheduling

# Elasticsearch
## Outline

GWDG

### Elasticsearch (ES)

- Verteilte Such- und Datenanalyse-Plattform
- Designziele:
  - Hohe Performance
  - Skalierbarkeit und Flexibilität
  - Robustheit
  - Cluster-by-Design
- Anleihen aus NoSQL und Big Data Frameworks
- Architektur qualifiziert ES für Cloud-Betrieb
- Frei und OpenSource (optional: kommerziellen Support)
- Basis: Apache Lucene – Projektstart 2010
- Sehr große Community, gute Dokumentation
- Wohldefinierte APIs, Domain-Specific Language, Plugins

XING

GitHub

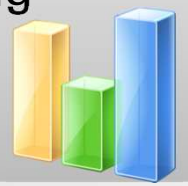
# Elasticsearch
## Highlights

### Erschließung und Analyse

- (Semi-) strukturierte Daten
- Volltexte
- Erschließung > 40 Dateiformaten
- Datenaggregation
- Skriptgesteuerte Auswertung

### Hohe Performance

- Hoher Datendurchsatz
- Rasche Indexupdates
- Rasche Ergebnisbearbeitung
- Monitoring einfach
- Stabiler Betrieb

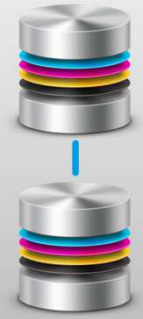


### Mandantenfähigkeit

- Bessere Auslastung phys. Server
- Trennung von Daten innerhalb des Clusters
- Datenhaltung-Policies
- Erweiterungen ggf. nötig.

### Verteilte Architektur

- Clusterbetrieb by-default
- Arbeitsteilung im Cluster
- Parallele Indizierung
- Verteilte Suchbearbeitung
- Hochverfügbarkeit



# Projekteinsatz
## Geistes- und Kulturwissenschaften

GWDG

### TextGrid

- Virtuelle Forschungsumgebung für Geisteswissenschaftler (Web / Rich-Client)
- ~ 800.000 TEI-annotierte Texte, Editionen und Scans
- Elasticsearch:
  - Volltext- und Metadatensuche
  - Metadaten-Harvesting (OAI-PMH)
- Herausforderungen im Bezug auf Textprocessing:
  - akkurate Suche unabdingbar für Geisteswissenschaftler
  - Mehrsprachigkeit
  - Konfiguration der Suchanalyzer und Tokenizer
  - Integration zahlreicher Textquellen
  - Klassische XML-Datenbankabfragen zu langsam
- ES-Clusterbetrieb: Loadbalancing + High Availibility
- Monitoring und Metriken per ES Dashboards

TextGrid

# Projekteinsatz
## Sozialwissenschaften

GWDG

### Soeb3 VFU

- Virtuelle Forschungsumgebung für Sozialwissenschaftler (Web: Liferay Portal)
- Syntaxen, Variablen, Studienbeschreibungen und Datensätze zur sozioökonomischen Entwicklung Deutschlands
- Elasticsearch:
  - Volltext- und Metadatensuche
  - Volltexterschließung in verschiedenen Dateiformaten
- Herausforderungen im Bezug auf Textprocessing:
  - Formatvielfalt (XML, Stata, SAS, R, PDF, MS Office, Text)
  - Rechte- und Zugriffsmodelle
  - Datenschutz im Bezug auf Datenaggregation
  - Unterstützung komplexer Metadaten
  - Abwägung: Anwenderfreundlichkeit VS. Schutzbedarfe
- Eigene Middleware-Plattform zur Datenintegration

# Danke für die Aufmerksamkeit

## Kontaktdaten


Oliver Schmitt
T +49 551 39 20512
E oliver.schmitt@gwdg.de

GWDG - Gesellschaft für
wissenschaftliche
Datenverarbeitung mbH Göttingen
Am Faßberg 11, 37077 Göttingen
http://www.gwdg.de