Big Data in den Digital Humanities

Dr. Jochen Tiepmar

Abteilung Automatische Sprachverarbeitung, Universität Leipzig

Fragen

- Was bedeutet Digital Humanities & Big Data?
- Digital Humanities sind keine klassische Big Data Anwendung. Wo sind die Unterschiede?
- Welche Besonderheiten der Digital Humanities sind dabei zu berücksichtigen?
- Was sind praktische DH-Anwendungen für Big Data

Digital Humanities (DH)

- Supporting research in humanities with tools from computer science

  Datenbank-Spektrum, Springer Verlag, 2015.
  http://dx.doi.org/10.1007/s13222-014-0177-7

- Schnittstelle zwischen Informatik und Geisteswissenschaften
- Kombiniert Methoden von

  Archeologie, Kunst, Kultur & Sozialwissenschaften, Geschichte, Linguistik, Literatur, Philosophie, Musik, Politik,…

- Mit Tools aus

  Datenvisualisierung, Data Mining, Mapping , Information Retrieval, Statistik, Text Mining,…

- Wir betrachten hier nur textorientierte DH

DH Anfänge

- Index Thomisticus:
  - Konkordanz der Schriften des Thomas von Aquin
  - Texte mit 11 Millionen laufenden Wortformen
  - entspräche auf Lochkarten gespeichert einem Papiergewicht von mehr als 100 Tonnen
  - Erstes Projekt im Bereich Humanities Computing
  - Kooperation mit IBM, kontinuierliche Nutzung neuer Speichertechnologien (Magnetband, Festplatten, CD)
  - Erstellung von 56 gedruckten Bänden mit 70.000 Seiten

DH Anfänge

Roberto Busa Archive

Big Data

- Volume -> Umfang der Daten
- Velocity -> Verarbeitungs/Berechnungsgeschwindigkeit
- Veracity -> Verlässlichkeit der Informationen/Fehlerrate
- Variety -> Heterogenität von Daten/Tools/,…

- Je nach Anwendung:
  - Visualization, Value, Volatility, Vulnerability, Validity, Variability, (Quality)

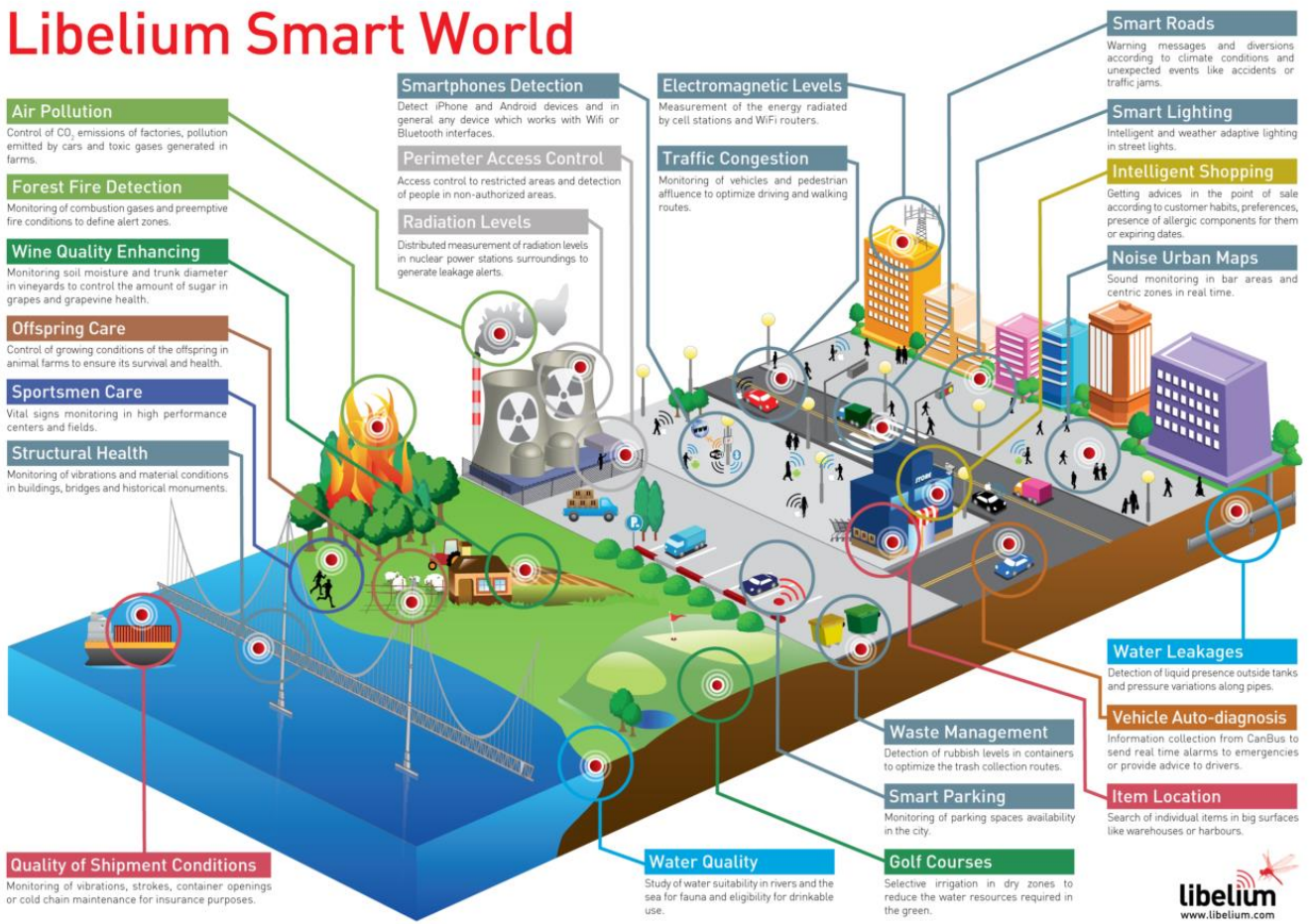
Big Data

- In klassischen Big Data-Anwendungsbereichen liegt der Fokus nicht gleichmäßig auf allen Vs
- Häufig werden Verbesserungen in Volume und Velocity als die einzigen Aufgaben der Informatik angesehen
- Kernanwendungen: Sensornetzwerke und Umweltdaten, Simulationen in Wissenschaft und Industrie

Big Data

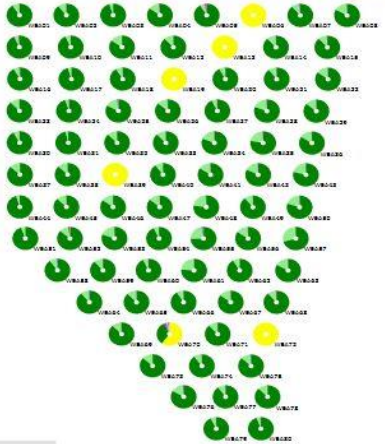
- Ein paar Anwendungsbeispiele

Big Data

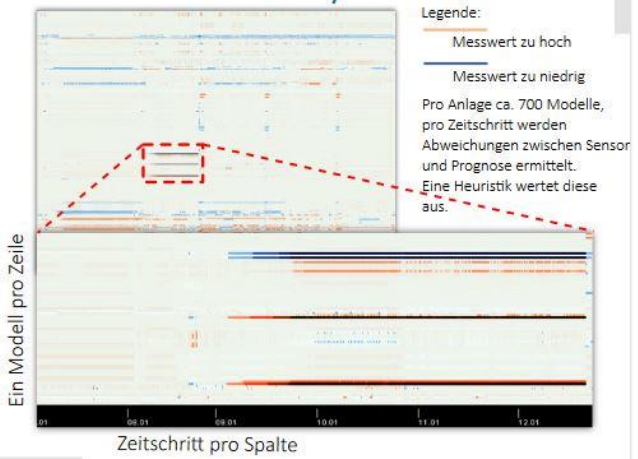

Libelium Smart World
Air Pollution
Control of CO_2 emissions of factories, pollution emitted by cars and toxic gases generated in farms.
Forest Fire Detection
Monitoring of combustion gases and preemptive fire conditions to define alert zones.
Wine Quality Enhancing
Monitoring soil moisture and trunk diameter in vineyards to control the amount of sugar in grapes and grapevine health.
Offspring Care
Control of growing conditions of the offspring in animal farms to ensure its survival and health.
Sportsmen Care
Vital signs monitoring in high performance centers and fields.
Structural Health
Monitoring of vibrations and material conditions in buildings, bridges and historical monuments.
Quality of Shipment Conditions
Monitoring of vibrations, strokes, container openings or cold chain maintenance for insurance purposes.
Smartphones Detection
Detect iPhone and Android devices and in general any device which works with Wifi or Bluetooth interfaces.
Perimeter Access Control
Access control to restricted areas and detection of people in non-authorized areas.
Radiation Levels
Distributed measurement of radiation levels in nuclear power stations surroundings to generate leakage alerts.
Electromagnetic Levels
Measurement of the energy radiated by cell stations and WiFi routers.
Traffic Congestion
Monitoring of vehicles and pedestrian affluence to optimize driving and walking routes.
Water Quality
Study of water suitability in rivers and the sea for fauna and eligibility for drinkable use.
Golf Courses
Selective irrigation in dry zones to reduce the water resources required in the green.
Smart Roads
Warning messages and diversions according to climate conditions and unexpected events like accidents or traffic jams.
Smart Lighting
Intelligent and weather adaptive lighting in street lights.
Intelligent Shopping
Getting advices in the point of sale according to customer habits, preferences, presence of allergic components for them or expiring dates.
Noise Urban Maps
Sound monitoring in bar areas and centric zones in real time.
Water Leakages
Detection of liquid presence outside tanks and pressure variations along pipes.
Vehicle Auto-diagnosis
Information collection from CanBus to send real time alarms to emergencies or provide advice to drivers.
Waste Management
Detection of rubbish levels in containers to optimize the trash collection routes.
Smart Parking
Monitoring of parking spaces availability in the city.
Item Location
Search of individual items in big surfaces like warehouses or harbours.
libelium
www.libelium.com

Big Data

Vollautomatische Zustandsüberwachung von Windparks



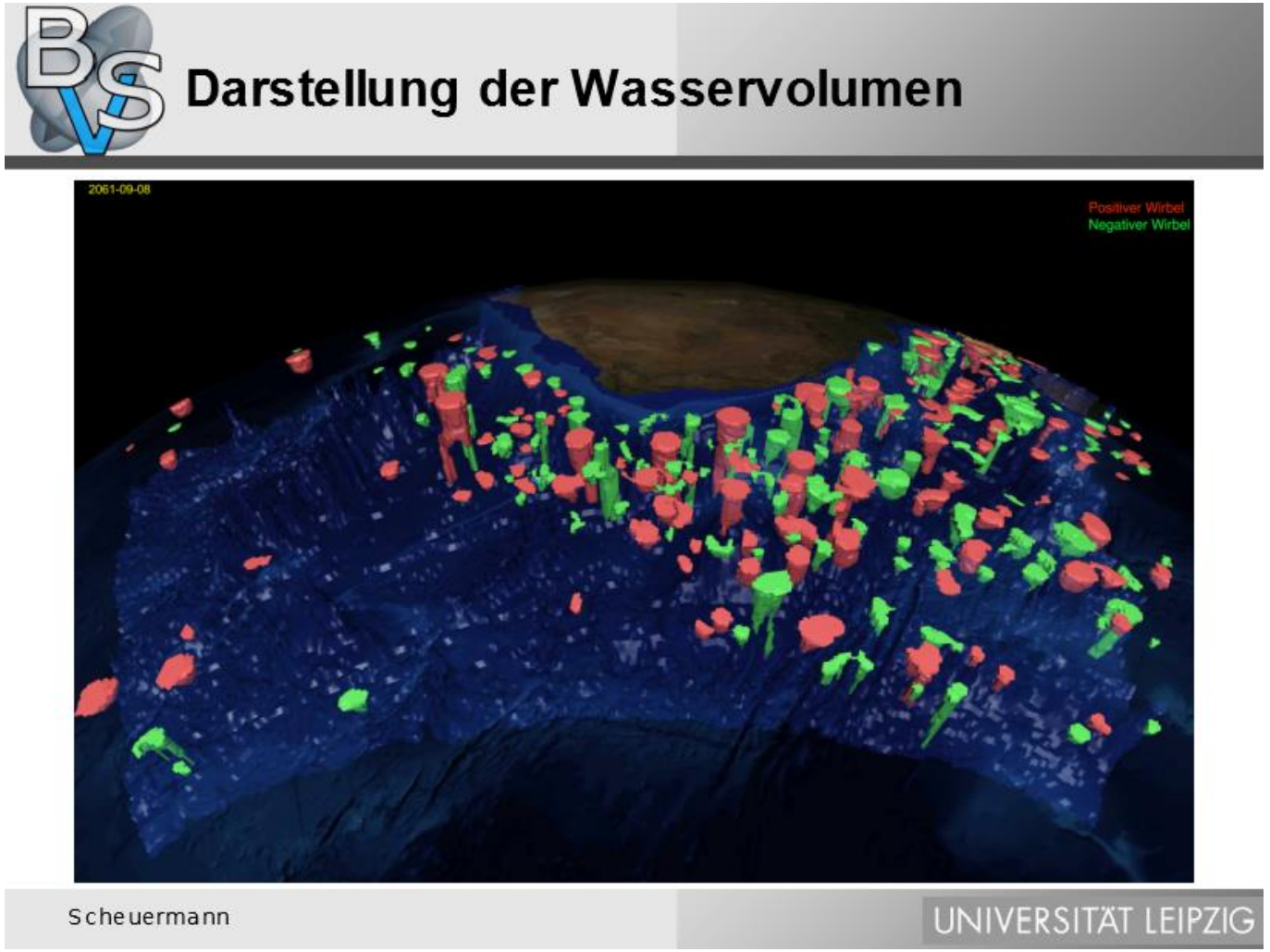
Datenstromanalyse



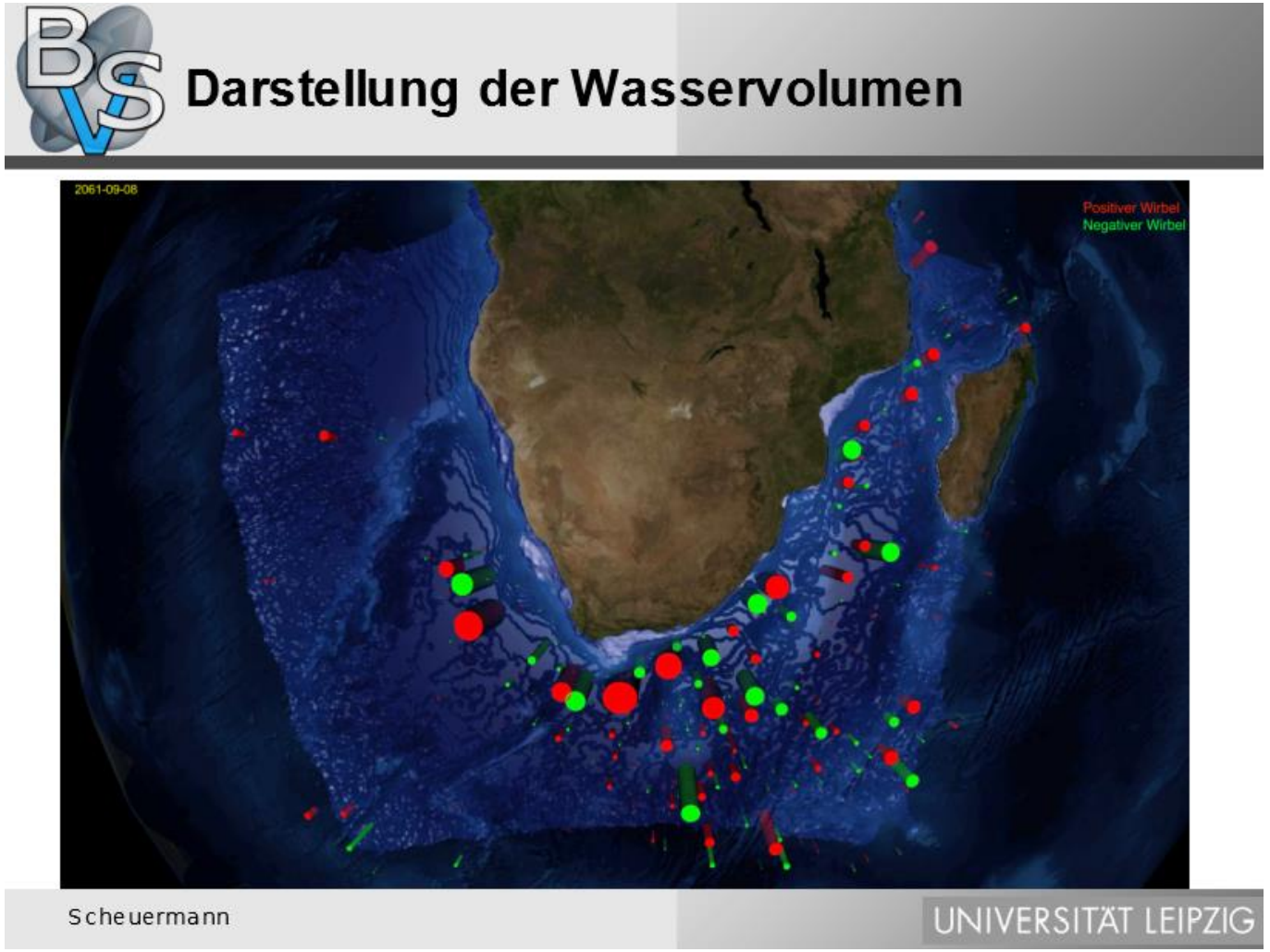
Ergebnisse und Mehrwerte

- Erkennung von Sensordefekten, Trend-änderungen, Druckverluste, Beschädigunge an Baugruppen, Leckagen, Ausrichtungs-fehler
- Ermöglicht prädiktive Wartung
- Frühzeitige Erkennung von schleichenden Ertragsverlusten: Fehlausrichtung einer WEA im Offshorebereich ca. 40k € Schaden pro Monat

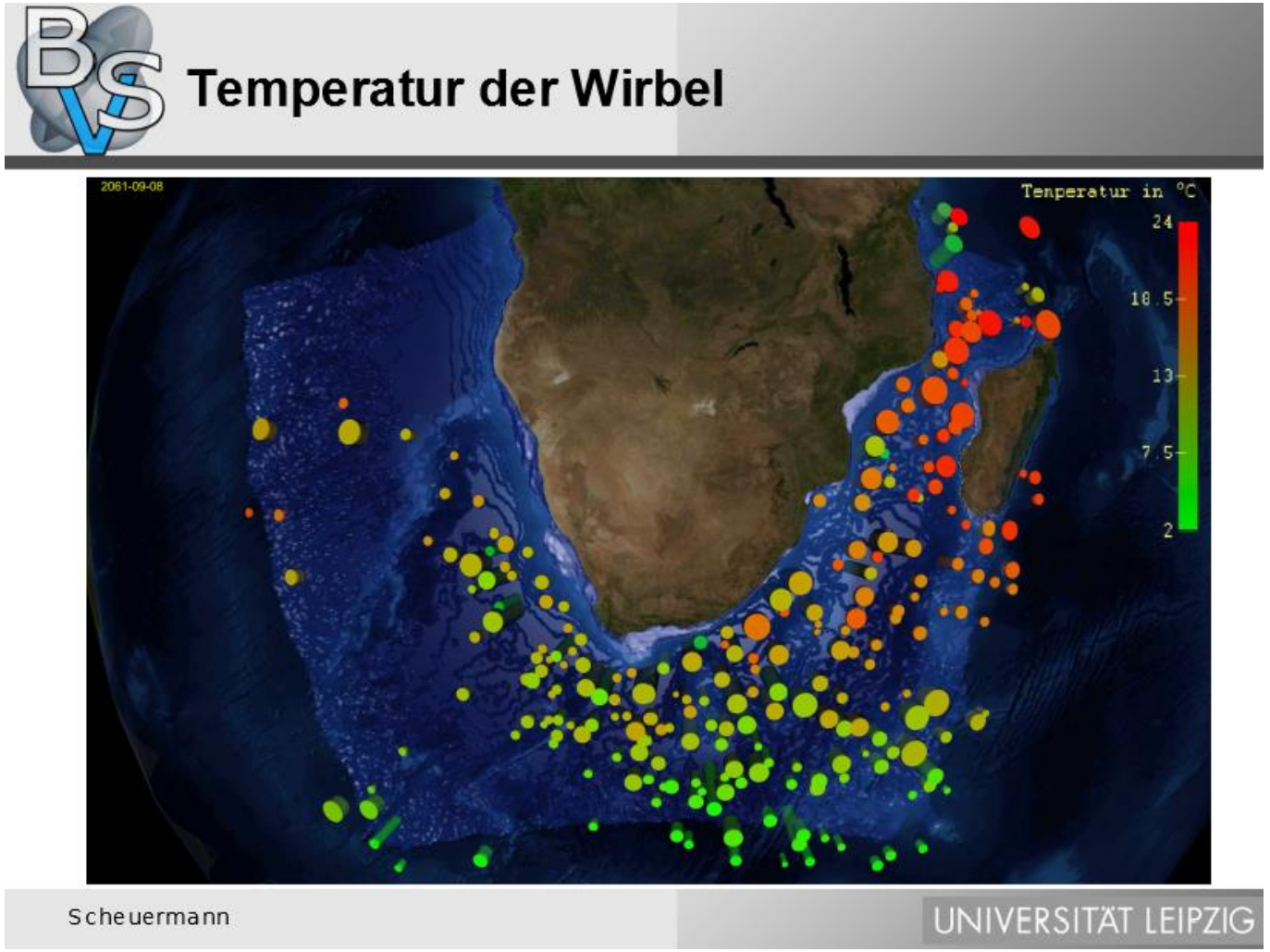
Big Data


Darstellung der Wasservolumen
2061-09-08
Positiver Wirbel
Negativer Wirbel
Scheuermann
UNIVERSITÄT LEIPZIG

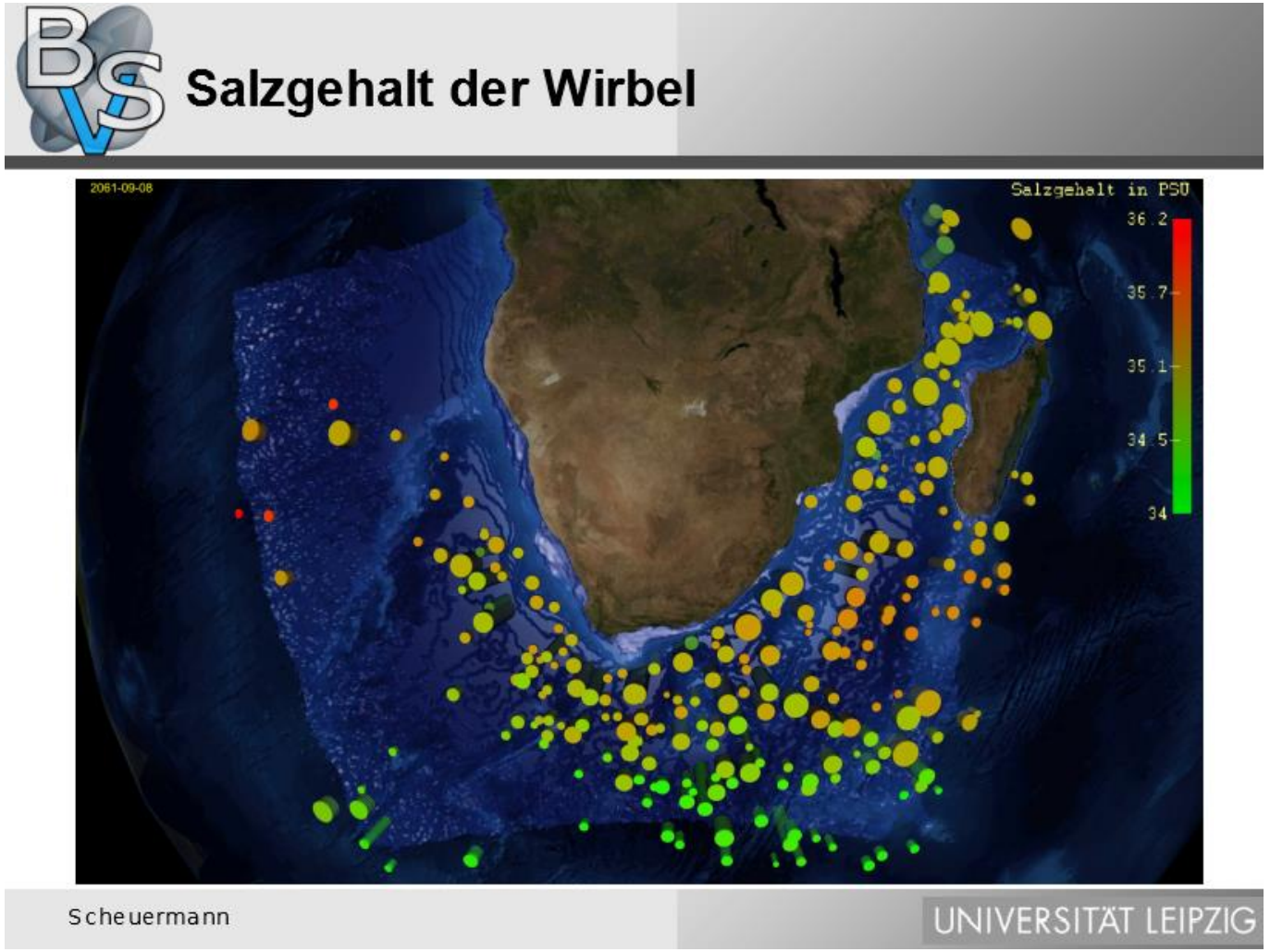
Big Data


Darstellung der Wasservolumen
2061-09-08
Positiver Wirbel
Negativer Wirbel
Scheuermann
UNIVERSITÄT LEIPZIG

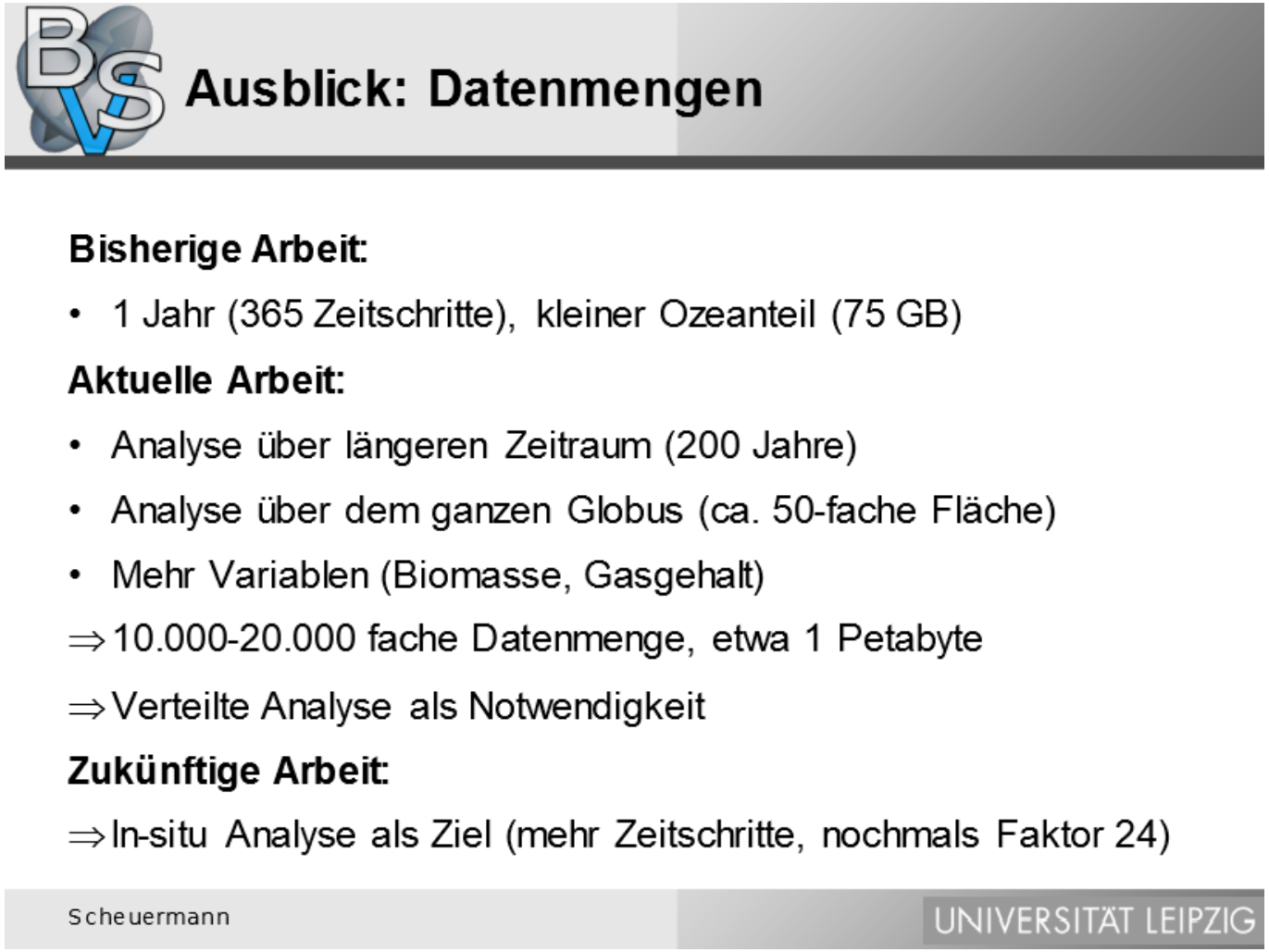
Big Data


Temperatur der Wirbel
2061-09-08
Temperatur in °C
24
18.5
13
7.5
2
Scheuermann
UNIVERSITÄT LEIPZIG

Big Data

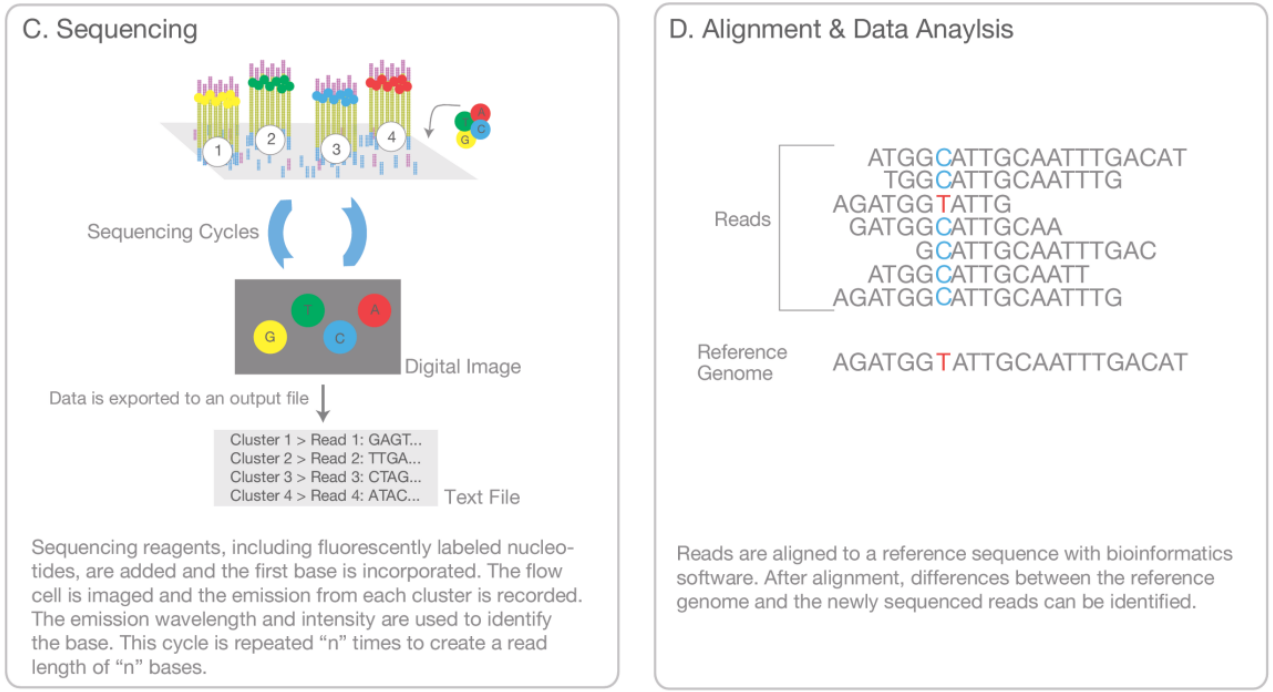

Salzgehalt der Wirbel
2061-09-08
Salzgehalt in PSU
36.2
35.7
35.1
34.5
34
Scheuermann
UNIVERSITÄT LEIPZIG

Big Data


Ausblick: Datenmengen
Bisherige Arbeit:
• 1 Jahr (365 Zeitschritte), kleiner Ozeanteil (75 GB)
Aktuelle Arbeit:
• Analyse über längeren Zeitraum (200 Jahre)
• Analyse über dem ganzen Globus (ca. 50-fache Fläche)
• Mehr Variablen (Biomasse, Gasgehalt)
⇒ 10.000-20.000 fache Datenmenge, etwa 1 Petabyte
⇒ Verteilte Analyse als Notwendigkeit
Zukünftige Arbeit:
⇒ In-situ Analyse als Ziel (mehr Zeitschritte, nochmals Faktor 24)
Scheuermann
UNIVERSITÄT LEIPZIG

Big Data

Next Generation Sequencing



Big Data

The Data Deluge

- Genome size (human) 3 Gb
- Transcripts:
  $\sim 20,000$ genes
  $10^6 ... 10^7$ RNA products (crude estimate)
- Sequencing run (Illumina HiSeq 2500) 6×10^{11} characters in 6×10^9 reads
- larger experiments: $\gg$ 1000 samples $\Longrightarrow$ $\sim 10^{15}$ characters, i.e., in the Petabyte range

Hintergrund

- Canonical Text Services Projekt in Leipzig
  - Dissertationsthema: Implementation and Evaluation of the Canonical Text Service Protocol as Part of a Research Infrastructure in the Digital Humanities
  - Webservice für persistente Textzitation über Projektgrenzen hinweg
  - Verschiedenste frei verfügbare Korpora
    - Deutsches Textarchiv, Textgrid, Voices of the Holocaust, Grimms Märchen, Perseus, Parallel Bible Corpus,….
  - http://cts.informatik.uni-leipzig.de/

Hintergrund

- Scalable Data Solutions (ScaDS) in Leipzig
  - Zusammenarbeit verschiedener Forschungsgruppen in Dresden & Leipzig
    - Visualisierung, Datenbanken, NLP, DH
  - Ziel: Aufbau eines nationalen Kompetenzzentrums für Big Data
  - CTS eine der entwickelten Lösungen

  - https://www.scads.de/

Motivation

- Eine der Hauptfragen in ScaDS:

- „Warum ist dein Projekt Big Data?“

Motivation

- Eine der Hauptfragen in ScaDS:

- „Warum ist dein Projekt Big Data?“

- Offensichtliche Antwort: „Weil ich mit Terabytes an Daten arbeite!“

Motivation

- Eine der Hauptfragen in ScaDS:

- „Warum ist dein Projekt Big Data?“

- Offensichtliche Antwort: „Weil ich mit Terabytes an Daten arbeite!“
- Mein Problem: Selbst große Textkorpora nur ein paar Gigabyte groß

Beispielanwendung CTS Index

Datensatz	Dokumente	Statsche Referenzen (CTS URNs)	HDD Größe Index/Daten (MB)	Sprache (n)
Textgrid	91133	7284050	2730/2621	Deu
Deutsches Text Archiv	8190	21911559	7841/12887	Deu
Perseus	1693	273126	490/65	Grc, Lat, Farsi, Eng
Parallel Bible Corpus	831	8818067	3330/4972	Ca 800
TED Talk Subttle Transripts	52987	15292408	2642/2507	105

Textkorpora idR thematisch/zeitlich/… zusammenhängend

Analysen idR nur über wenige Dokumentenparametrisierungen

- Trendanalyse benötigt bspw. nur Dokumente in 1 Sprache
- Eher dezentrale als zentrale Datenmenge

→ Abgesehen von (eher unüblichen) zentralen Archiven ist Volume für Textdaten (Primärdaten) kein Problem

The Big Vs

- Volume -> Umfang der Daten
- Velocity -> Verarbeitungs/Berechnungsgeschwindigkeit
- Veracity -> Verlässlichkeit der Informationen/Fehlerrate
- Variety -> Heterogenität von Daten/Tools/,…

- Je nach Anwendung:
  - Visualization, Value, Volatility, Vulnerability, Validity, Variability, (Quality)

4 Big V in DH

- Volume
  - Primärdaten vs Sekundärdaten
    - Text vs. Annotation/Metadata
  - Primärdaten generell unproblematisch bezogen auf Volume
  - Sekundärdaten können beliebig aufgebläht werden
    - Volume –Bezug projektabhängig

Daten im ASV-Wortschatz	Primärdaten (Bytes) (Sätze)	Sekundärdaten (Bytes) (alles andere)
deu_mixed_2011	37, 270, 576, 048	517, 020, 294, 364
deu_news_2011	3, 672, 898, 564	59, 421, 534, 187
deu_newscrawl_2011	3, 735, 178, 336	222, 879, 231, 073

4 Big V in DH

- Velocity
  - Verarbeitungsgeschwindigkeit
  - Streamingbasierte Workflows
    - (bspw. real time Twitter Analyse)
    - Frameworks wie Apache Flink, Spark, …
  - Zeitintensive Berechnungen (bspw. Zitatanalyse, Topic Modeling)
    - Parallelisierung
    - Computer Cluster

Velocity - Parallelisierung

- Auf mehreren Ebenen durchführbar
- Dokumente, Kapitel, Sätze,…
- Über Daten oder Workflows oder Quellcode
- Erlaubt Beschleunigung über Computer-Cluster oder Threads
- Parallelisierung erzeugt immer Overhead (bspw. Verwaltungsaufwand)
- Beschleunigungsfaktor sollte etwa dem Parallelisierungsfaktor entsprechen

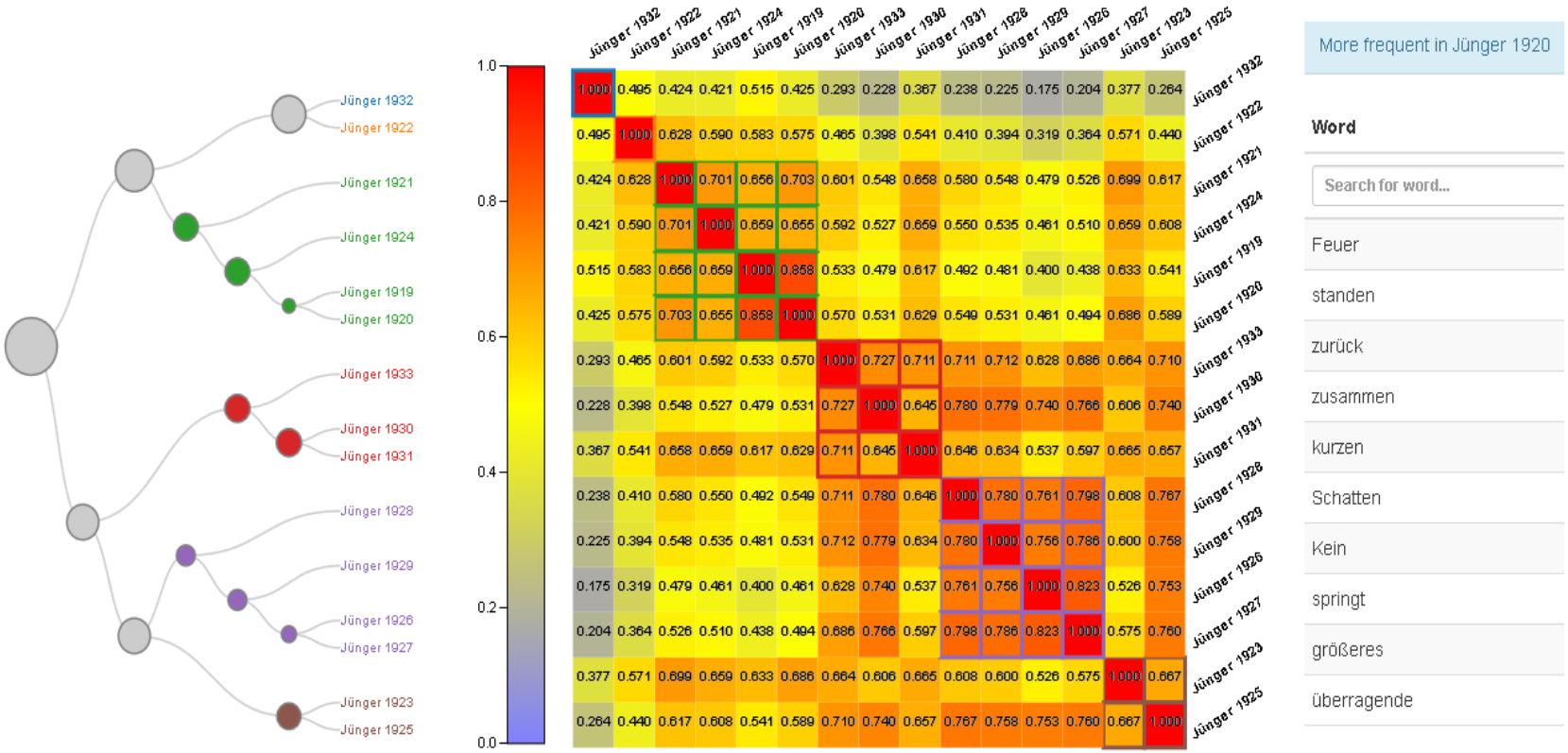
Velocity - Paralleles Wortzählen (Flink)

```java
public class WordCountExample {
    public static void main(String[] args) throws Exception {
        final ExecutionEnvironment env = ExecutionEnvironment.getExecutionEnvironment();

            DataSet<String> text = env.fromElements(
            "Who's there?",
            "I think I hear them. Stand, ho! Who's there?");

        DataSet<Tuple2<String, Integer>> wordCounts = text
            .flatMap(new LineSplitter())
            .groupBy(0)
            .sum(1);

        wordCounts.print();

        env.execute("Word Count Example");
    }

    public static class LineSplitter implements FlatMapFunction<String, Tuple2<String, Integer>> {
        @Override
        public void flatMap(String line, Collector<Tuple2<String, Integer>> out) {
            for (String word : line.split(" ")) {
                out.collect(new Tuple2<String, Integer>(word, 1));
            }
        }
    }
}
```

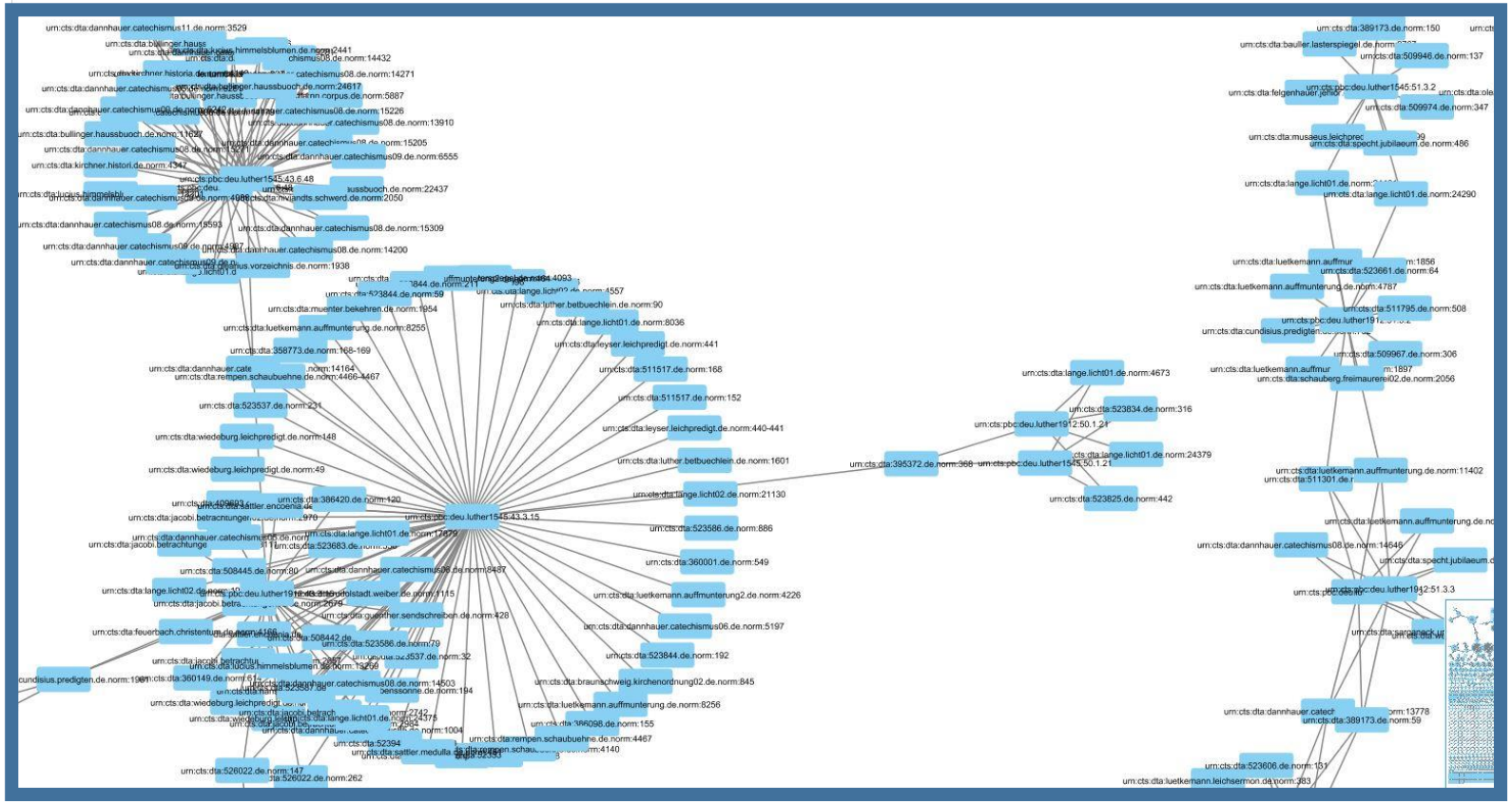
Hier werden die 1en pro Wort gruppiert und zusammenaddiert

Hier wird jedes Wort mit einer 1 gesammelt

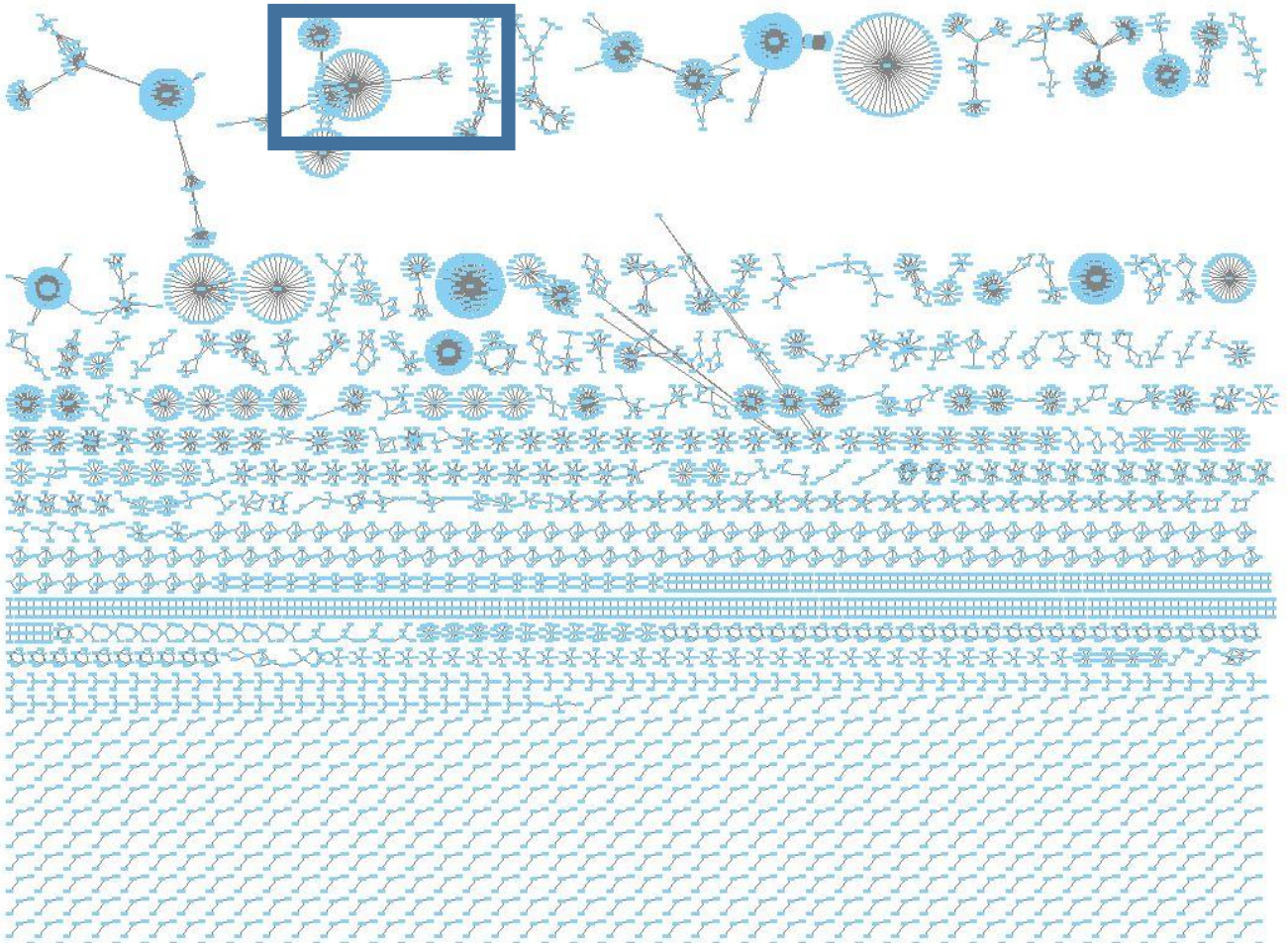
Velocity - Vergleichende Korpusanalyse



Velocity - Text Reuse/Zitationsanalyse 1



Velocity - Text Reuse/Zitationsanalyse 1

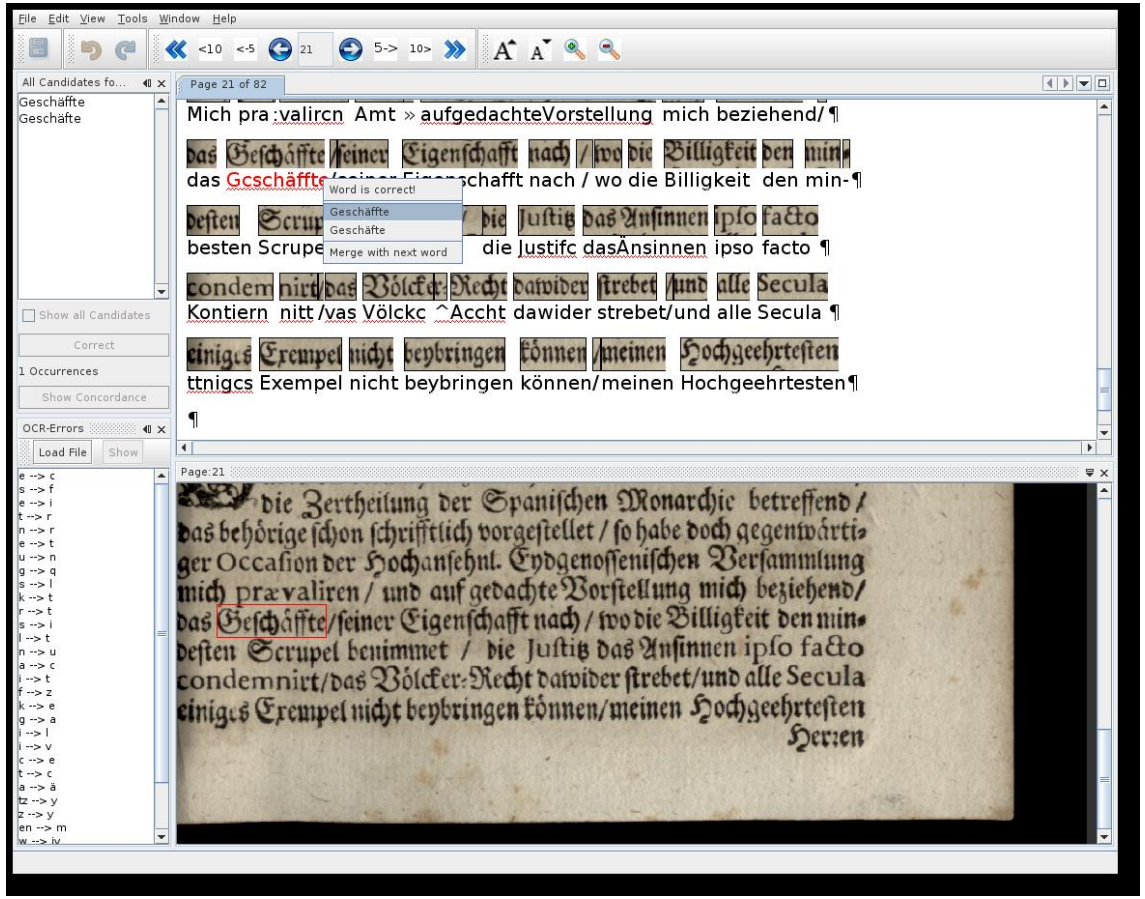


4 Big V in DH

- Veracity
  - Verlässlichkeit der Daten/Fehlerrate
  - OCR Korrektur automatisch/manuell
  - Schrifterkennung in Bildern
    - Machine Learning, Neural Networks
  - Normalisierung von Text
    - Punktuation, Rechtschreibfehler vs Korpusspezifische Abweichungen (zeitl. Kontext)

Veracity - Schrifterkennung 1

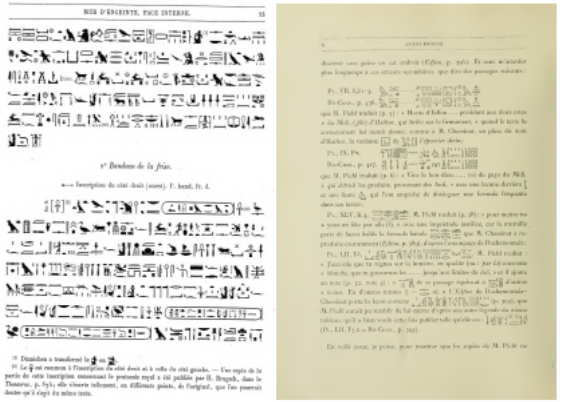
Interaktive OCR-Plattform
PoCoTo

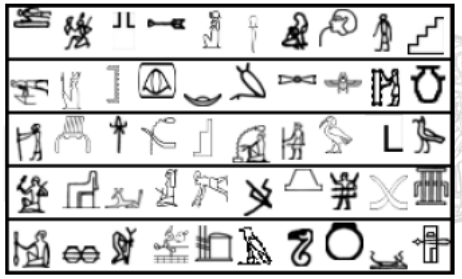


Veracity - Schrifterkennung 2

Erkennung ägyptischer Hieroglyphen

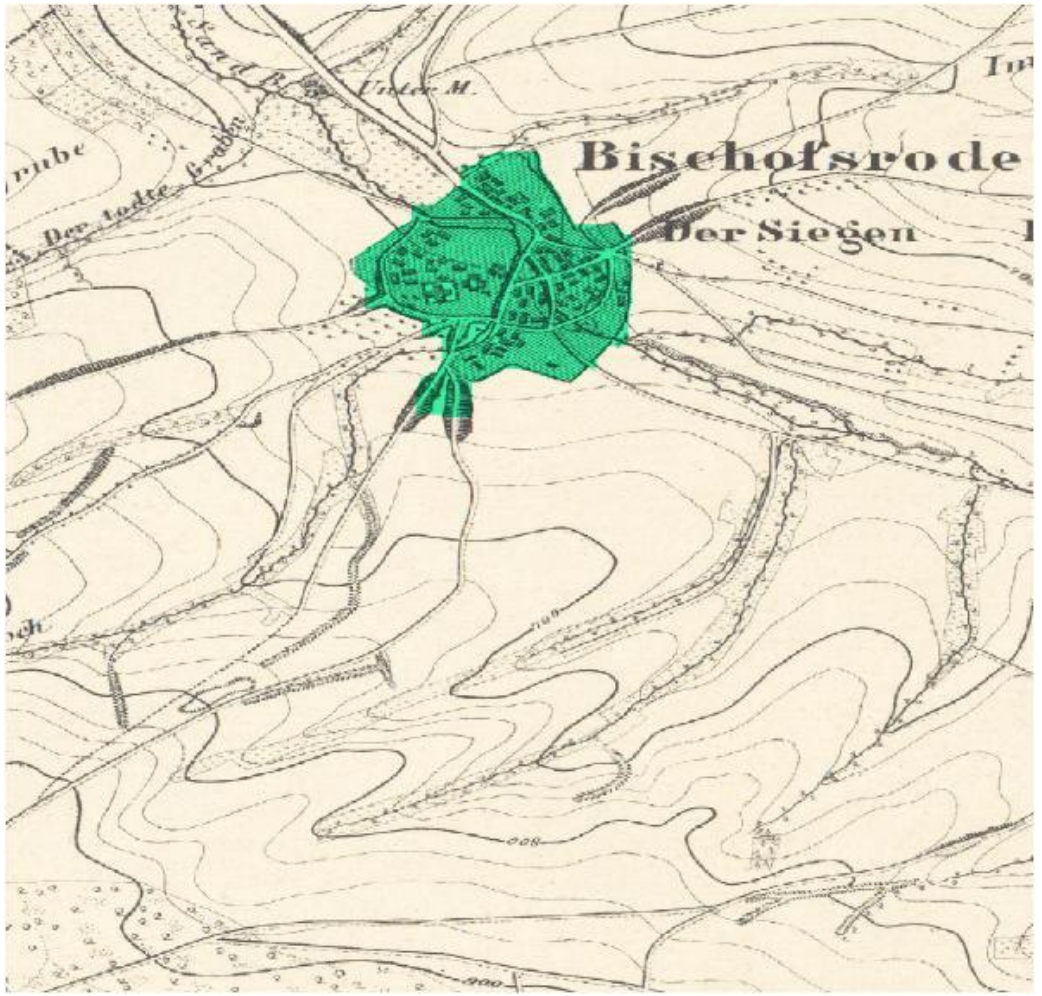
- Ziel: Erkennung der Hieroglyphen in digitalisierten Büchern der frühen Ägyptologe
- 6832 unterschiedliche Hieroglyphen
- Test der Objekterkennung an Sequenzen mit jeweils 10 zufällig ausgewählten Hieroglyphen



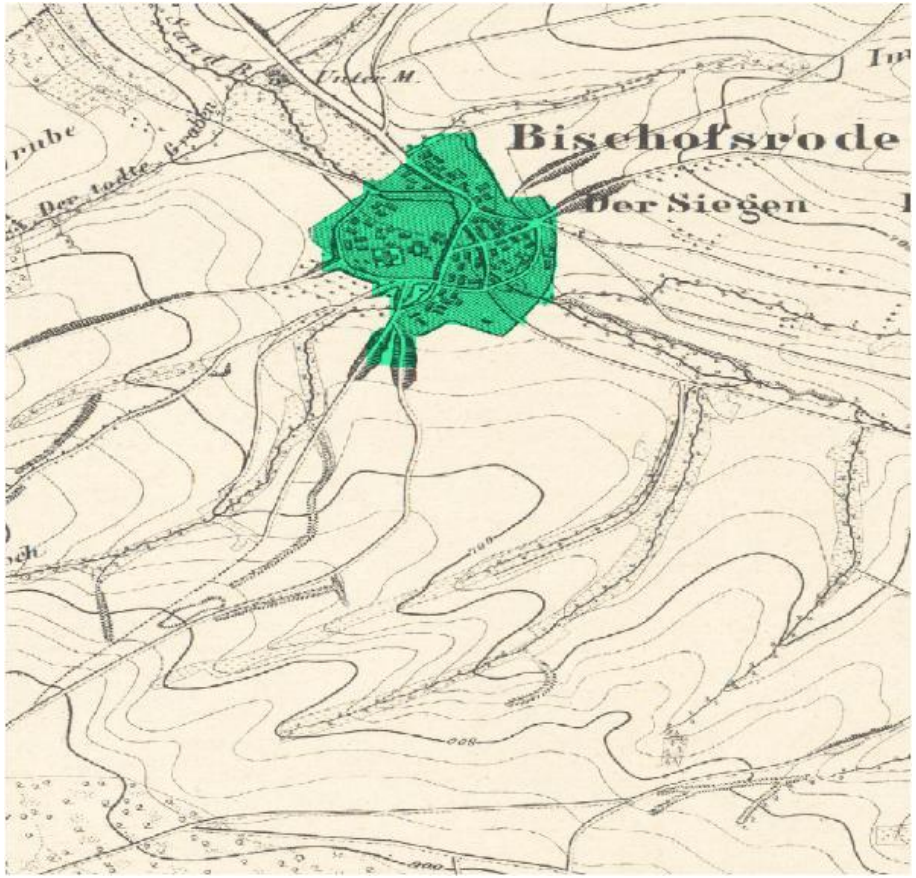


Veracity - Ortsnamenerkennung

- Ebenfalls OCR:
- Kartenbeschriftung erkennen



Veracity - Mustererkennung



4 Big V in DH

- Variety
  - Heterogenität & Interoperabilität
  - Dateiformate
    - TEI/XML, Docbook, …
  - Datenzugang
    - Online vs Offlne
    - Zip-Pakete, Git, Projektspezifische Lösungen
  - Tools
    - Input/Import-Formate

Variety - Forschungsinfrastrukturen

- Umgebungen (in DH) für
  - Evaluation
  - Auswahl
  - Wiederverwendung
  - Kombination
- von Tools, Daten und Metadaten

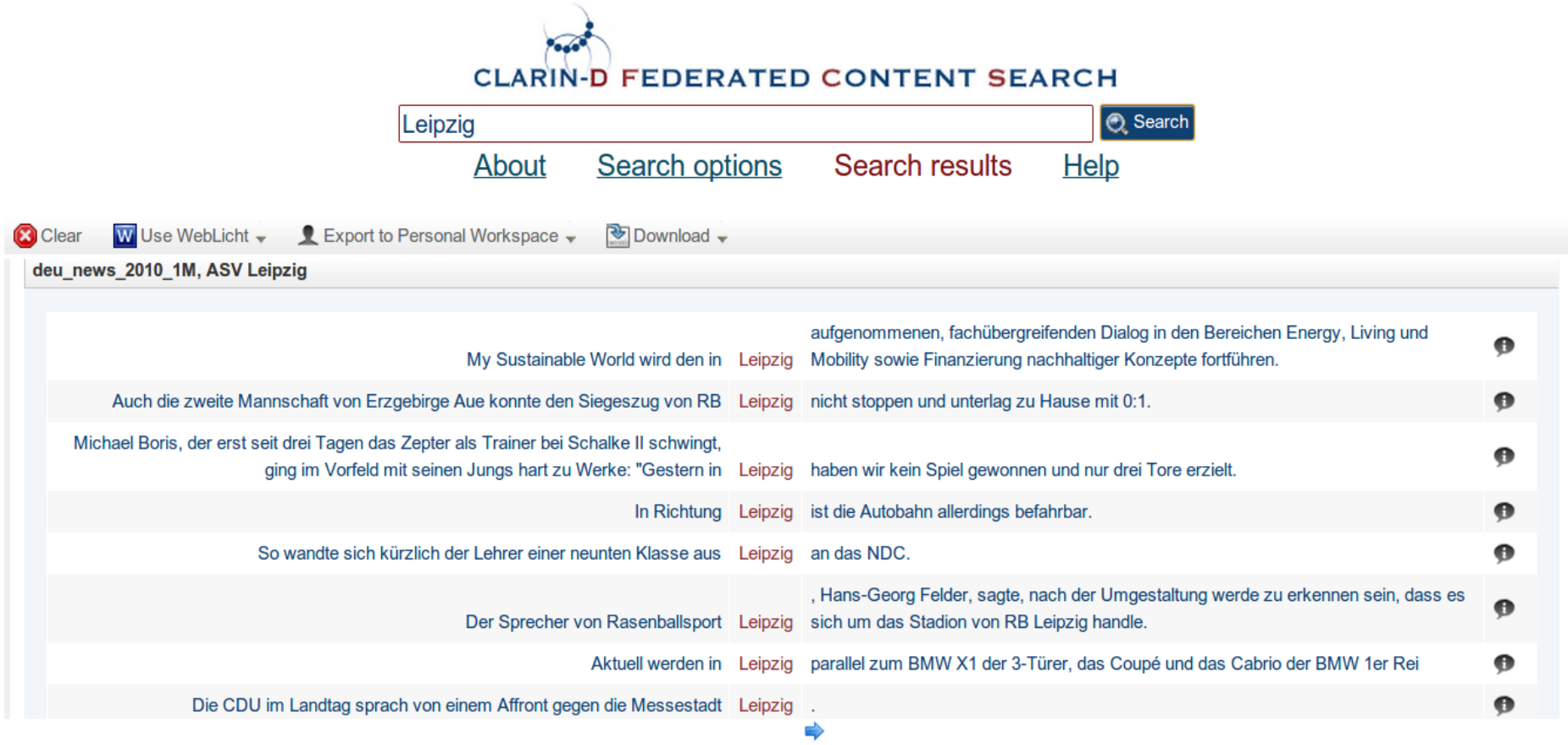
Variety - CLARIN

- Zugang zu verschiedensten Daten & Metadaten via webservices
- Inhalte über gemeinsamen Metadatenstandart vereinheitlicht (CMDI)
- Federated Search
- Verkettete workflows via WebLicht
  - webapplications, keine locale Installation nötig
- Persistente Zitierbarkeit durch Persistent IDentifiers (PID)
- Aggregierter Inhalt aus verschiedenen Quellen als Ergebnis

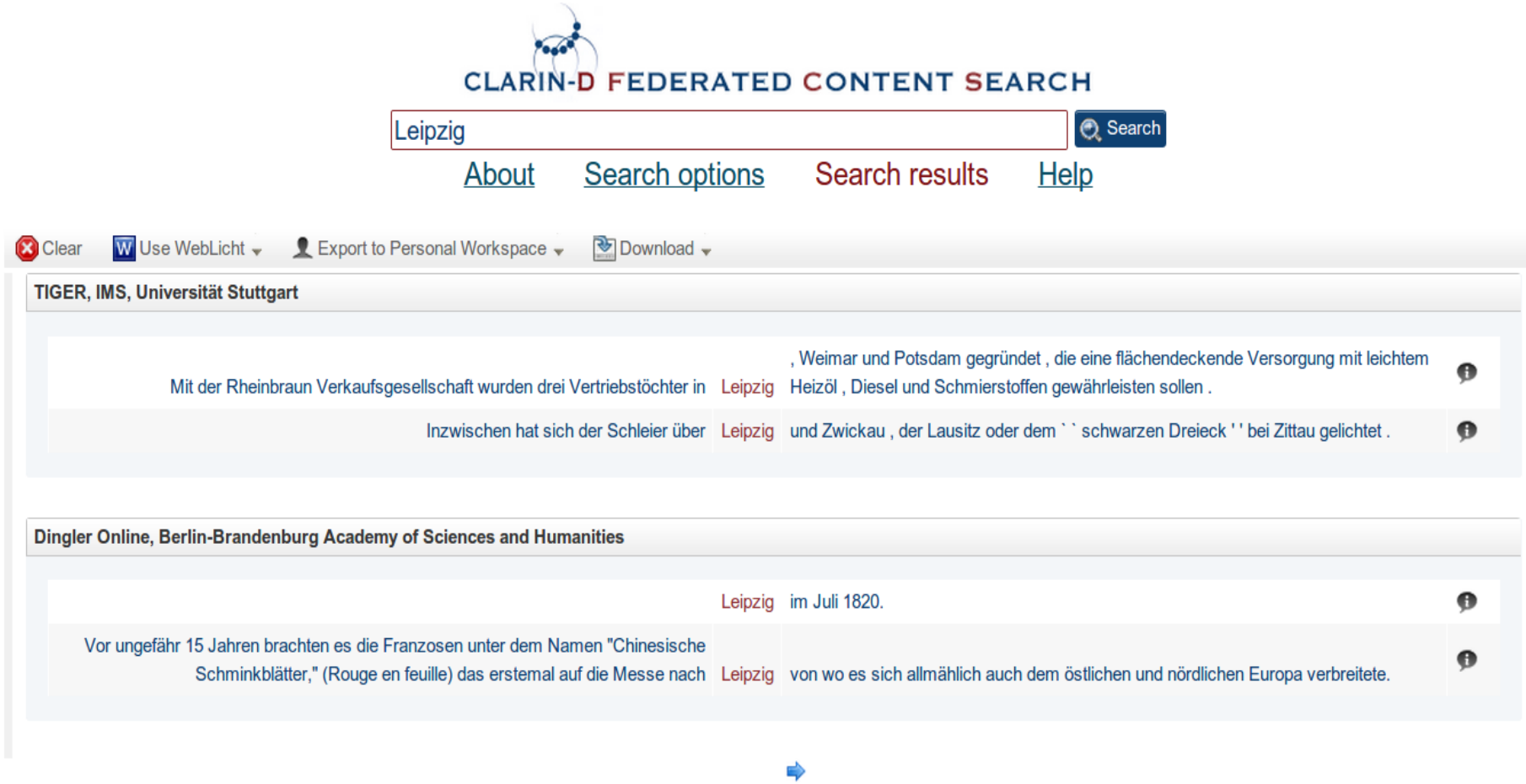
Variety - VLO


Virtual Language Observatory
VLO
Explore the world of language resources and technology from different perspectives
CLARIN
WALS
DoBeS
VLO > Faceted search > Search: "Leipzig" > Selections: Written Corpus ✖ German ✖
Permalink | Report | Help
SEARCH
Leipzig
Search
SEARCH RESULTS
6 results
Showing 1 to 6
Wortschatz
Expand
Collected from newspaper texts, webcrawling, etc.: words (+frequency), cooccurrences (+graph), left/right neighbours, example sentences
Resources: | 1 other |
deu_news_2008_100K
Expand
100.000 sentences of a German newspaper corpus based on material from 2008
Resources: | 1 text document | 1 other |
NARROW DOWN
Use the categories below to limit the search results to those matching the selected value(s).
+ LANGUAGE
German ✖
+ COLLECTION
- RESOURCE TYPE
Written Corpus ✖
+ COUNTRY
+ GENRE
+ FORMAT

Variety - Federated Content Search (FCS)

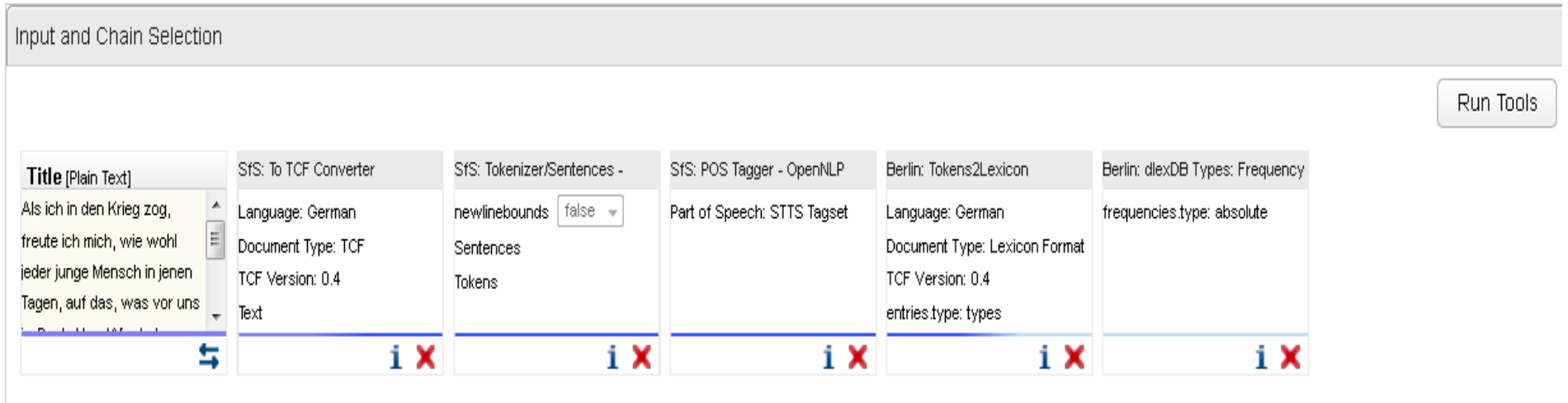


Variety - Federated Content Search (FCS)



Variety - WebLicht

- Webservice zur Verkettung und Ausführung von Workflows
  - Workflowübergreifende Metabeschreibung
    - Kompatibilität zwischen Ausgabe und Eingabe verschiedener Elemente
    - CMDI



Variety - Interoperabilität

- Austauschbarkeit und Wiederverwendbarkeit von Daten und Workflows
- Erlaubt „blindes“ Erstellen von Tools & Daten
- Erfordert übergeordnete, anwendungsunabhängige Regeln (Metaebene) für Interface & Access
- Reduziert Variety
- Für textbasierte DH: Canonical Text Services (nächste Woche)

Zusammenfassung

- Datenvolumen ist nur 1 Aspekt von Big Data
  - Komplexe Probleme können auch mit kleineren Datensätzen aufwendig werden
- Big Data Herausforderungen können sein:
  - Texterkennung, Schrifterkennung, Fehlererkennung, Normalisierung
  - Zeitanforderungen, Durchssatz, in sich komplexe Berechnungen
  - Umgang mit Annotation, Metadata, Meta-metadata,…
  - Interoperabilität
- Vor allem in DH ist die Bestimmung von “Big Data-igkeit” einer bestimmten Fragestellung nicht trivial und universell übertragbar