

Slide 1

Vasari: Wissensextraktion mittels iterativ entwickelter Dokumentbeschreibungen

Norbert Gövert Norbert Fuhr

Universität Dortmund

<http://ls6-www.cs.uni-dortmund.de/vasari/>

- Das Vasari-Projekt
- VaLa: Abstrakte Beschreibung von Quelldokumenten
- Iterative Erstellung von VaLa-Beschreibungen
- Zusammenfassung, Ausblick

Slide 2

Das Vasari-Projekt

Vasari, Giorgio, * Arezzo 30. Juli 1511, † Florenz 27. Juni 1574, ital. Maler, Architekt und Kunstschriftsteller. Begründer der Kunstgeschichte. Seine Künstlerbiographien („Die Lebensbeschreibungen der berühmtesten italienischen Architekten, Maler und Bildhauer“, 1546–50, erw. Ausgabe 1568) gehören zu den wichtigsten Grundlagen der Kunstgeschichte.

Ziel: Erstellung eines Kunstportals für Fachleute und die interessierte Öffentlichkeit (*artregister.org*)

Quellen: OCR-Texte aus mehr als 20 einschlägigen Lexika

Aufgabe: Wissensextraktion aus unterschiedlich stark strukturierten Dokumenten

Das Vasari-Projekt II

<kuenstler>

<vorname>Giorgio</vorname><nachname>Vasari</nachname>

<geburt><datum>1511-07-30</datum><ort>Arezzo</ort></geburt>

<tod><datum>1574-06-27</datum><ort>Florenz</ort></tod>

<beschreibung>

ital. <beruf>Maler</beruf>, <beruf>Architekt</beruf> und

<beruf>Kunstschriftsteller</beruf>. Begründer der Kunstgeschichte. Seine

Künstlerbiographien (<werk><titel>Die Lebensbeschreibungen der berühmtesten

italienischen Architekten, Maler und Bildhauer</titel>,

<datum>1546–50</datum>, erw. Ausgabe <datum>1568</datum></werk>)

gehören zu den wichtigsten Grundlagen der Kunstgeschichte.

</beschreibung>

</kuenstler>

Slide 3

VaLa: Abstrakte Beschreibung von Quelldokumenten

[Vasari Language]

Ziel: automatische XML-Annotierung der Quelldokumente

XML Schema Definition von Datentypen und logischer Struktur

Kontextausdrücke Zuordnung Inhalt ↔ Struktur

- *Zeichenketten*: reguläre Ausdrücken
- *linguistische Konstrukte*: computerlinguistische Kategorien
- *Entitäten*: Wörterbücher
- *Rechts- / Linkskontext*

Slide 4

VaLa (II): Beispiel

Slide 5

```
<element name="geburt">
  <sequence>
    <element name="ort">
      <pre type="regexp"> geb|. </pre>
      <match type="entity"> place </match>
    </element>
    <element name="datum">
      <match type="entity"> date </match>
      <post type="element">tod </post>
    </element>
  </sequence>
</element>
```

Vasari, Giorgio, geb. Arezzo
30. Juli 1511, <tod>gest.
Florenz 27. Juni 1574</tod>, ital.-Maler,
Architekt und Kunstschriftsteller.



Vasari, Giorgio, <geburt>geb. <ort>Arezzo</ort>
<datum>30. Juli 1511 </datum></geburt>, <tod>gest.
Florenz 27. Juni 1574</tod>, ital.-Maler,
Architekt und Kunstschriftsteller.

VaLa (III): Beispiel

Slide 6

Vasari, Giorgio, geb. <ort>Arezzo</ort>
30. Juli 1511, gest. <ort>Florenz</ort>
27. Juni 1574, ital.-Maler, Architekt
und Kunstschriftsteller.

Places
:
:
Arezzo
Caprese
Florenz
Leiden
Mailand
Malaga
Mougins
:
:

Picasso, Pablo, geb. <ort>Malaga</ort>
25. Oktober 1881, gest. <ort>Mougins</ort>,
Frankreich 8. April 1973, ...

Iterative Erstellung von VaLa-Beschreibungen

Ziel: Beschreibung soll möglichst viele Dokumente abdecken

Slide 7

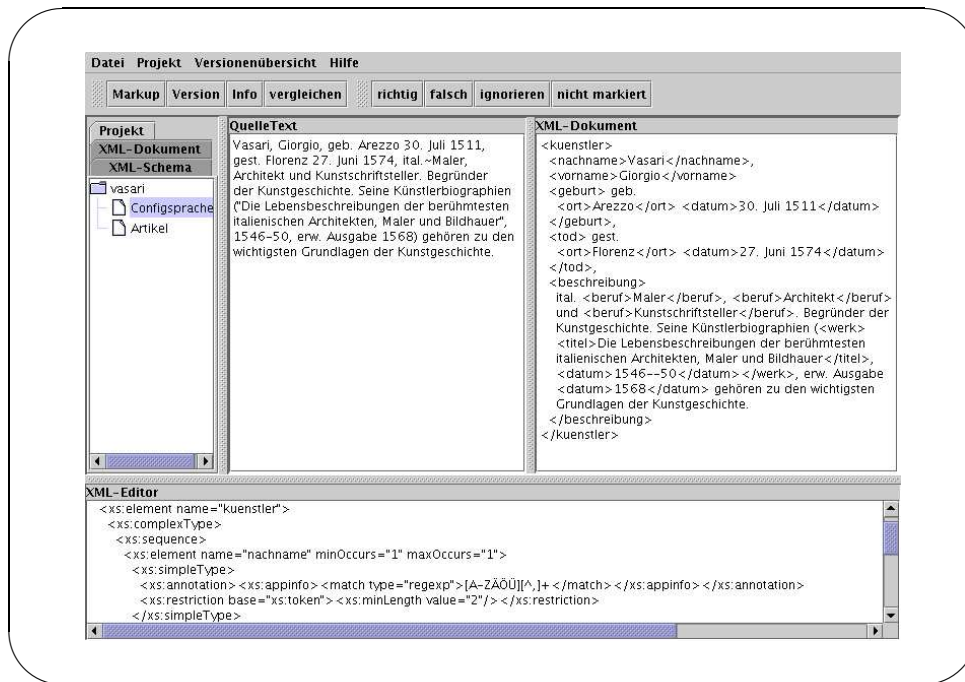
Interaktives Werkzeug zur Erstellung von VaLa-Beschreibungen:

- Entwicklung von Kontextausdrücken anhand von Beispielausdrücken
- unmittelbare Anwendung auf Beispieldokumente → Visualisierung
- iterative Verbesserung der Beschreibung
- Bewertung der Annotationen
- Versionierung der Beschreibungen und Bewertungen

Slide 8

Iterative Erstellung von VaLa-Beschreibungen

Slide 9



Zusammenfassung

Slide 10

- VaLa: Struktur-Beschreibungssprache für Wissensextraktion
→ Ausnutzung von Delimiterregeln, linguistische Kategorien
 - Werkzeug zur iterativen Erstellung von Dokument-Beschreibungen
- ⇒ Such- / Navigationsstrukturen für effektive Informationserschließung

Ausblick

Slide 11

Synthese: einheitliche Wissensrepräsentation

- Fusion von Dokumenten aus unterschiedlichen Quellen
- Zusammenhänge zwischen Dokumenten herstellen

→ Transformation nach *RDF*

Slide 12

