An infrastructure for eHumanities

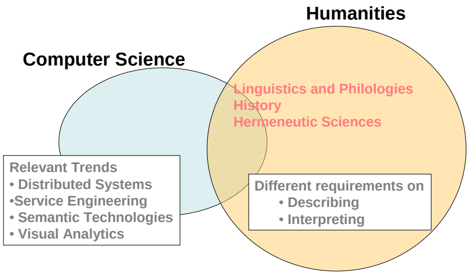
Workshop on Historical texts
Boston, 2010/01/13

Gerhard Heyer
(using slides by Marco Büchler, Volker Boehlke, and Charlotte Schubert)
Automatische Sprachverarbeitung
Computer Science Department
University of Leipzig

Agenda

- Computer Science and Humanities
- eAQUA
- eHumanities, Digital Humanities, and Humanities
- eHumanities and eScience Infrastructures

Computer Science and Humanities



Relevant Trends
- Distributed Systems
- Service Engineering
- Semantic Technologies
- Visual Analytics

Different requirements on
- Describing
- Interpreting

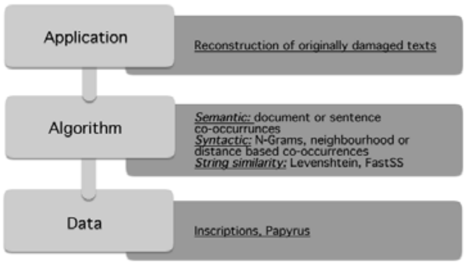
Computer Science and its applications

1940-1960	Scientific Computing
70ies	Data bases, digitizing business processes (Wirtschaftsinformatik)
80ies	Digitizing electro mechanical applications, SGML
90ies	Digitizing analogue media, linking distributed ressources: http, HTML, XML
since 2000	Internet based services, knowledge management

Computer Science and its applications

- Replacement of analogue by *digital* media and processes
- Increasing impact of digital media and proces-sing models on traditional work flows based on analogue media
- Digitizing media and work flows creates new methods and applications
- *eHumanities:* Digitizing media and work flows in the Humanities
- BMBF project eAQUA a good example

eAQUA Methodology


Application
Reconstruction of originally damaged texts
Algorithm
Semantic: document or sentence co-occurrunces
Syntactic: N-Grams, neighbourhood or distance based co-occurrences
String similarity: Levenshtein, FastSS
Data
Inscriptions, Papyrus

Subprojects

- **Atthidographers** – Classification of Atthidographers
- **Plato** - Aftermath of Plato's writings
- **Plautine metric** – Metrical analysis of Latin comedies
- **CAMENA** - Knowledge network of Latin CAMENA corpus
- **Inscriptions** - Extraction of significant templates for different kinds of inscriptions like release documents (slave trading)
- **Papyri** - Classification of papyri (e. g. slave trading contracts) and text completion of fragmentary texts
- **Mental maps** - Building of century based mental maps and highlighting differences

Course Text Mining for Classical Studies (TM4CS): http://www.eaqua.net/e_3.html

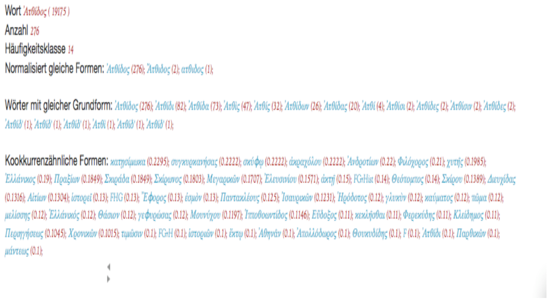
Main challenges

Challenges from Comuter Science point of view

- **Integrating textual ressources**
  Conversion of different formats and standards
- **Software engineering issues**
  Getting classicists to understand the intent and effects of text mining algorithms and clearly define their functional requirements
- **Semantic technologies that address the long tail**
  Modifying and amending statistical and pattern based methods to effectively deal with rare events

Example subproject Atthidographs

Searching for „Ἀτθίδος"



Wort Ἀτθίδος (19175)
Anzahl 276
Häufigkeitsklasse 14
Normalisiert gleiche Formen: Ἀτθίδος (276); Ἄτθιδος (2); ατθιδος (1);

Wörter mit gleicher Grundform: Ἀτθίδος (276); Ἀτθίδι (82); Ἀτθίδα (73); Ἀτθὶς (47); Ἀτθίς (32); Ἀτθίδων (26); Ἀτθίδας (20); Ἀτθί (4); Ἀτθίσι (2); Ἀτθίδες (2); Ἀτθίσιν (2); Ἀτθίδες (2); Ἀτθίδ' (1); Ἀτθίδ' (1); Ἀτθίδ' (1); Ἀτθὶ (1); Ἀτθίδ' (1); Ἀτθίδ' (1);

Kookkurrenzähnliche Formen: κατησίμωκα (0.2295); συγκυρκανήσας (0.2222); σκύφῳ (0.2222); ἀκραχόλου (0.2222); Ἀνδροτίων (0.22); Φιλόχορος (0.21); χυτῆς (0.1985); Ἑλλάνικος (0.19); Πραξίων (0.1849); Σκιράδα (0.1849); Σκίρωνος (0.1803); Μεγαρικῶν (0.1707); Ἐλευσινίου (0.1571); ἀκτῇ (0.15); FGrHist (0.14); Θεόπομπος (0.14); Σκίρου (0.1389); Διευχίδας (0.1316); Αἰτίων (0.1304); ἱστορεῖ (0.13); FHG (0.13); Ἔφορος (0.13); ἑσμὸν (0.13); Παντακλέους (0.125); Ἰσαυρικῶν (0.1231); Ἡρόδοτος (0.12); γλυκὺν (0.12); καύματος (0.12); πῶμα (0.12); μελίσσης (0.12); Ἑλλάνικός (0.12); Θάσιον (0.12); γεφυρώσας (0.12); Μουνύχου (0.1197); Ἱπποθοωντίδος (0.1146); Εὔδοξος (0.11); κεκλῆσθαι (0.11); Φερεκύδης (0.11); Κλείδημος (0.11); Περιηγήσεως (0.1045); Χρονικῶν (0.1015); τιμῶσιν (0.1); FGrH (0.1); ἱστοριῶν (0.1); ἕκτῳ (0.1); Ἀθηνᾶν (0.1); Ἀπολλόδωρος (0.1); Θουκυδίδης (0.1); F (0.1); Ἀτθίδι (0.1); Παρθικῶν (0.1); μάντεως (0.1);

Example subproject Atthidographs

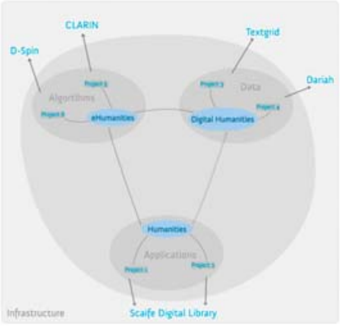
Looking at co-occurrences to find „interesting" neighbours

Signifikante rechte Kookkurrenzen für Ἀτθίδος

[illegible]

- νομάδας **is an interesting candidate**
- **Reference to quotation from the Atthis of Philochorus (3. / 2. century BC) in a normal word search hard to find**
- **Co-occurrence indicates that atthidographers may have invented a phase in Athenian history to show an evolutionary development from *nomadism* to settled life**

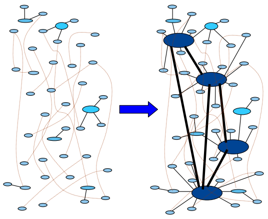
The three approaches in a more global context



There also is an infrastructure aspect …

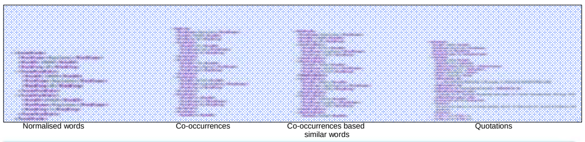
Infrastructures to share and reuse digital ressources

- **What we need is a transition from accidental to organized cooperation**
- ***Centers of competence* might mediate technical and organisational issues**
- **A platform of *services* allows all participants to share digital ressources and establisch a culture of *best practices* (cf. CLARIN)**

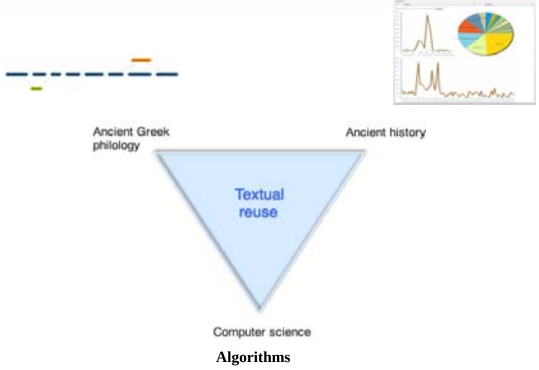


Reusibility of different aggregated results

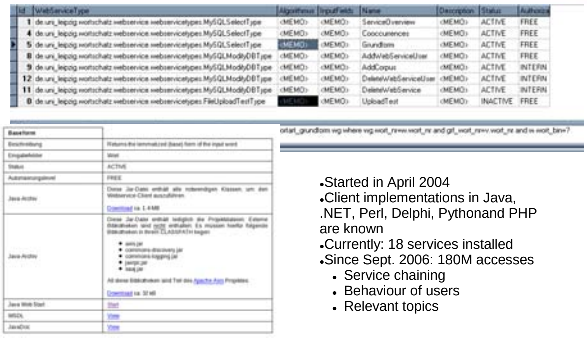
- ***Services*** **can be implemented as webservices for sharing data and algorithms (SOAP)**
- **Every participant needs to provide his data and algorithms in a standardized form (WSDL)**
- **Webservices allow to**
  - **reuse data from aggregated results**
  - **generate specific annotated data irrespective of the location of the resource**



Reusibility of data and results of algorithms for usage by different views



Webservice based infrastructure and service optimisation



Id	WebServiceType	Algorithmus	InputFields	Name	Description	Status	Authoriza
1	de.uni_leipzig.wortschatz.webservice.webservicetypes.MySQLSelectType	<MEMO>	<MEMO>	ServiceOverview	<MEMO>	ACTIVE	FREE
4	de.uni_leipzig.wortschatz.webservice.webservicetypes.MySQLSelectType	<MEMO>	<MEMO>	Cooccurrences	<MEMO>	ACTIVE	FREE
5	de.uni_leipzig.wortschatz.webservice.webservicetypes.MySQLSelectType	<MEMO>	<MEMO>	Grundform	<MEMO>	ACTIVE	FREE
8	de.uni_leipzig.wortschatz.webservice.webservicetypes.MySQLModifyDBType	<MEMO>	<MEMO>	AddWebServiceUser	<MEMO>	ACTIVE	FREE
9	de.uni_leipzig.wortschatz.webservice.webservicetypes.MySQLModifyDBType	<MEMO>	<MEMO>	AddCorpus	<MEMO>	ACTIVE	INTERN
12	de.uni_leipzig.wortschatz.webservice.webservicetypes.MySQLModifyDBType	<MEMO>	<MEMO>	DeleteWebServiceUser	<MEMO>	ACTIVE	INTERN
11	de.uni_leipzig.wortschatz.webservice.webservicetypes.MySQLModifyDBType	<MEMO>	<MEMO>	DeleteWebService	<MEMO>	ACTIVE	INTERN
0	de.uni_leipzig.wortschatz.webservice.webservicetypes.FileUploadTextType	<MEMO>	<MEMO>	UploadText	<MEMO>	INACTIVE	FREE

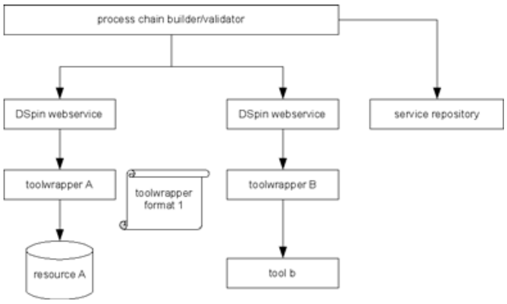
```
ortart_grundform wg where wg.wort_nr=w.wort_nr and grl_wort_nr=v.wort_nr and w.wort_bin=?
```

Baseform	
Beschreibung	Returns the lemmatized (base) form of the input word
Eingabefelder	Wort
Status	ACTIVE
Autorisierungslevel	FREE
Java-Archiv	Diese Jar-Datei enthält alle notwendigen Klassen, um den Webservice-Client auszuführen. Download ca. 1.4 MB
Java-Archiv	Diese Jar-Datei enthält lediglich die Projektklassen. Externe Bibliotheken sind nicht enthalten. Es müssen hierfür folgende Bibliotheken in ihrem CLASSPATH liegen: • axis.jar • commons-discovery.jar • commons-logging.jar • jaxrpc.jar • saaj.jar All diese Bibliotheken sind Teil des Apache Axis Projektes. Download ca. 32 kB
Java Web Start	Start
WSDL	View
JavaDoc	View

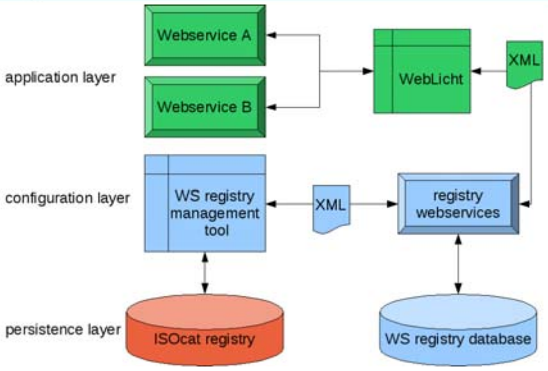
- Started in April 2004
- Client implementations in Java, .NET, Perl, Delphi, Pythonand PHP are known
- Currently: 18 services installed
- Since Sept. 2006: 180M accesses
  - Service chaining
  - Behaviour of users
  - Relevant topics

Source: *Marco Büchler and Gerhard Heyer: Leipzig Linguistic Services - A 4 Years Summary of Providing Linguistic Web Services. In: Gerhard Heyer (Editor):Text Mining Services – Building and applying text mining based service infrastructures in research and industry., Leipzig, Germany, 2009*

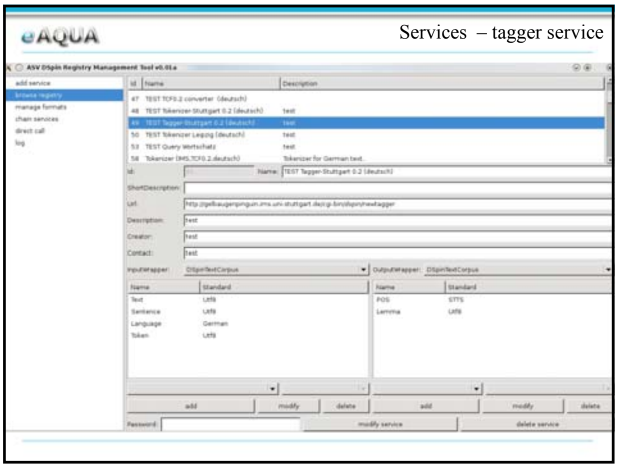
DSpin architecture



DSpin architecture III



Services – tagger service



Id	Name	Description
47	TEST TCF0.2 converter (deutsch)	
48	TEST Tokenizer Stuttgart 0.2 (deutsch)	test
49	TEST Tagger-Stuttgart 0.2 (deutsch)	test
50	TEST Tokenizer Leipzig (deutsch)	test
53	TEST Query Wortschatz	test
58	Tokenizer (IMS.TCF0.2.deutsch)	Tokenizer for German text.

Name: TEST Tagger-Stuttgart 0.2 (deutsch)
ShortDescription:
Url: http://gellbaugenpinguin.ims.uni-stuttgart.de/cgi-bin/dspin/newtagger
Description: test
Creator: test
Contact: test
InputWrapper: DSpinTextCorpus
OutputWrapper: DSpinTextCorpus

Name	Standard
Text	Utf8
Sentence	Utf8
Language	German
Token	Utf8

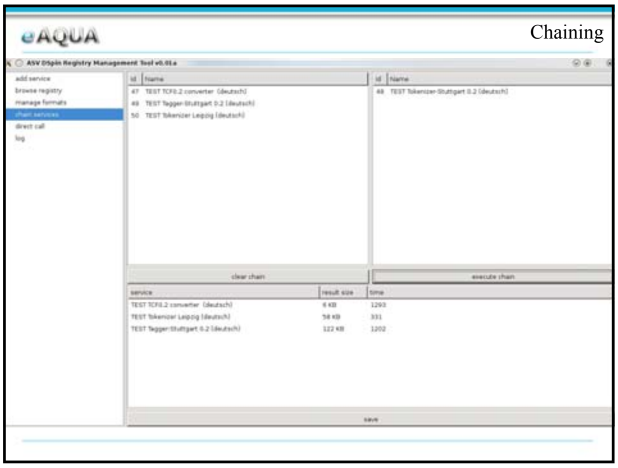
Name	Standard
POS	STTS
Lemma	Utf8

Password:

Matching

- Chaining consideren an instance of functional composition: c(b(a()))
- Basically the same problem to be solved like on the type checking level of a compile run: Check if all input needs of a certain function are satisfied: correct number and type of parameters (format is given by programming language).
- Our "parameters" (NLP-specific):
  - document is encoded using a certain format (TEI, …)
  - a certain concept/kind of information inside of a valid document (tokens, tags, …) => parameter
  - this concept/information is encoded using a certain datatype (utf8, tagset A, ...)

Chaining

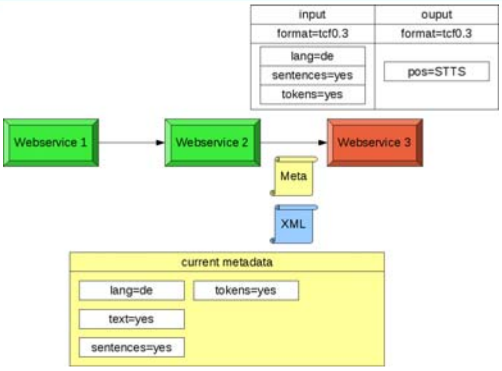

ASV DSpin Registry Management Tool v0.01a

Id	Name
47	TEST TCF0.2 converter (deutsch)
49	TEST Tagger-Stuttgart 0.2 (deutsch)
50	TEST Tokenizer Leipzig (deutsch)

Id	Name
48	TEST Tokenizer-Stuttgart 0.2 (deutsch)

service	result size	time
TEST TCF0.2 converter (deutsch)	6 KB	1293
TEST Tokenizer Leipzig (deutsch)	58 KB	331
TEST Tagger-Stuttgart 0.2 (deutsch)	122 KB	1202

Chaining in detail



input	ouput
format=tcf0.3	format=tcf0.3

lang=de
sentences=yes
tokens=yes
	pos=STTS

current metadata

lang=de	tokens=yes
text=yes	
sentences=yes	

Milestones

- **Already done:**
  - **Implementation of a basic registry for webservices (url, descriptions etc + metadata on format & input/ouput)**
  - **Several services of the DSpin prototype available**
  - **Implementation of a first version of the matchmaking algorithm**
  - **First successfull test in two different workflow/chaining tools (Tübingen, Leipzig)**
- **Future developments:**
  - **Implementation of a generic chain builder: automatically suggest a chain from startpoint (resource) to endpoint (tool, certain information, ...)**
  - **Open up the registry for metadata harvesting**
  - **Integrate other communities (like digital classics) and services**