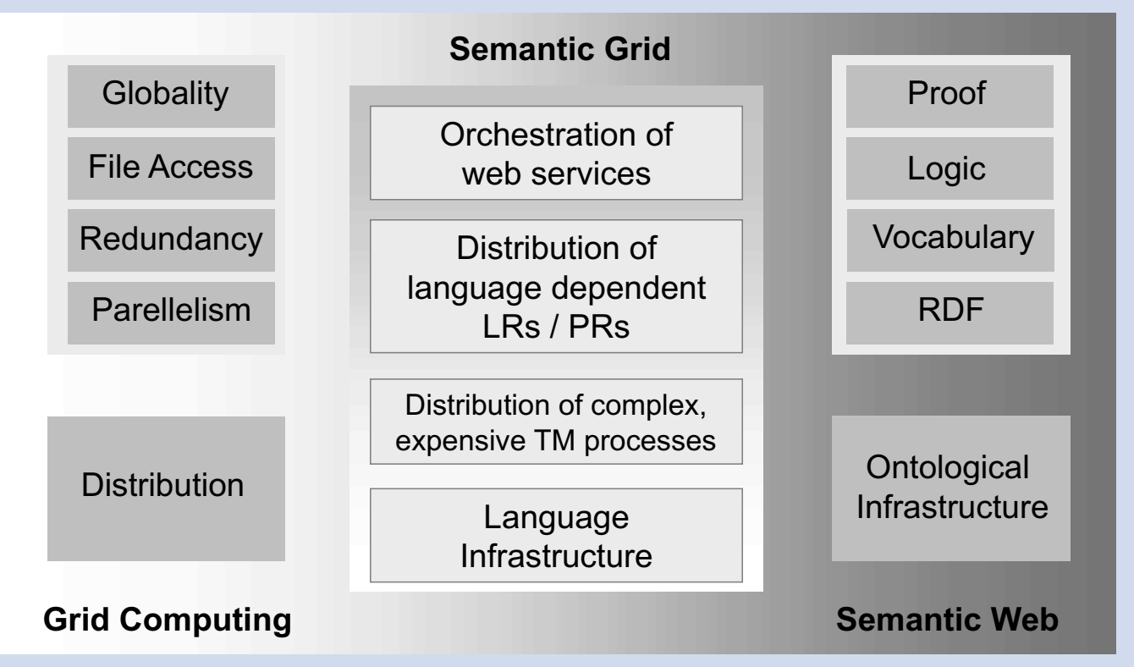
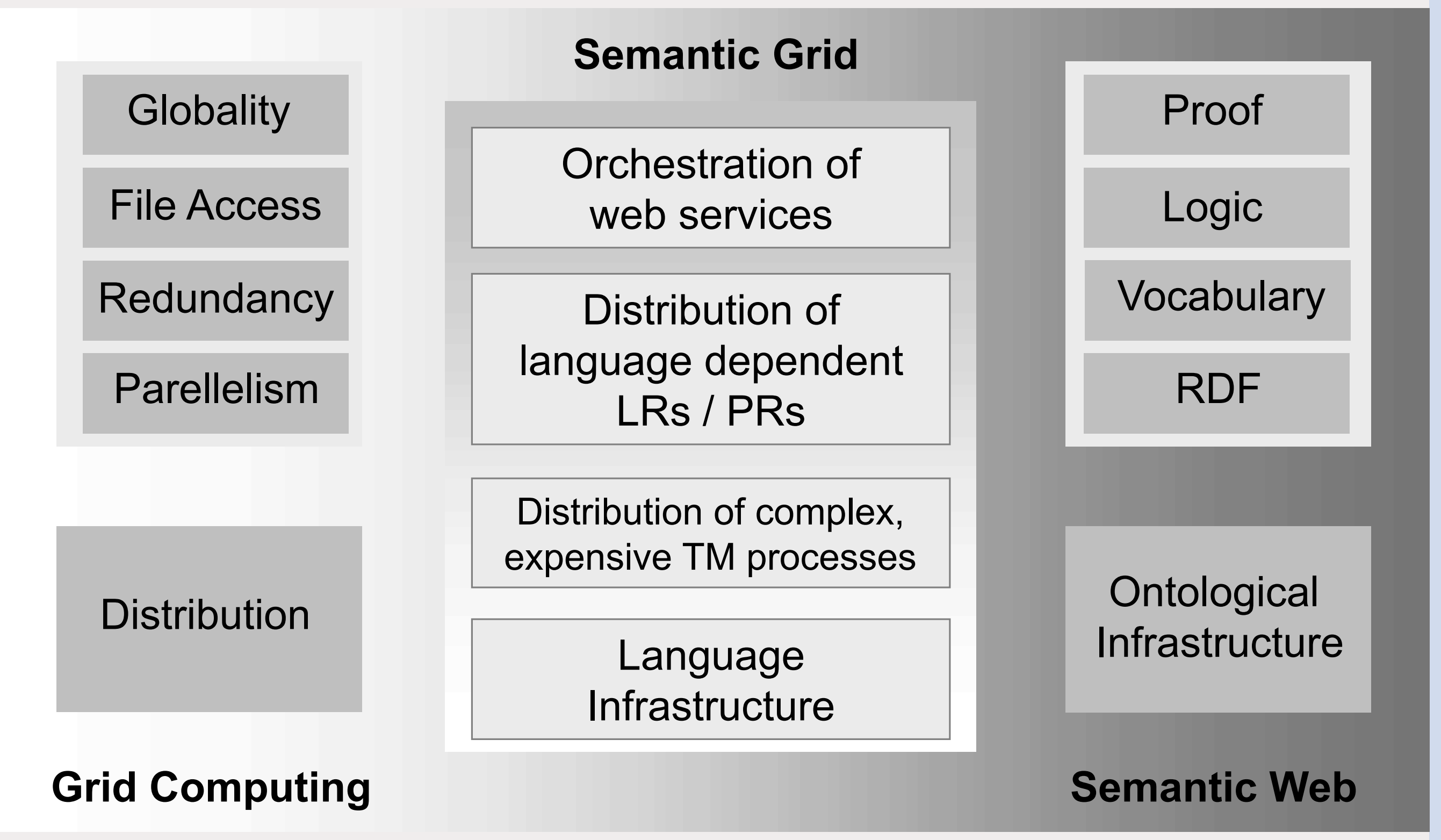
Embedded Distributed Text Mining and Semantic Web Technology

Daniel Sonntag

German Research Center for Artificial Intelligence, Saarbrücken/Germany, daniel.sonntag@dfki.de


Semantic Grid
Globality
File Access
Redundancy
Parellelism
Distribution
Grid Computing
Orchestration of web services
Distribution of language dependent LRs / PRs
Distribution of complex, expensive TM processes
Language Infrastructure
Proof
Logic
Vocabulary
RDF
Ontological Infrastructure
Semantic Web

Common View

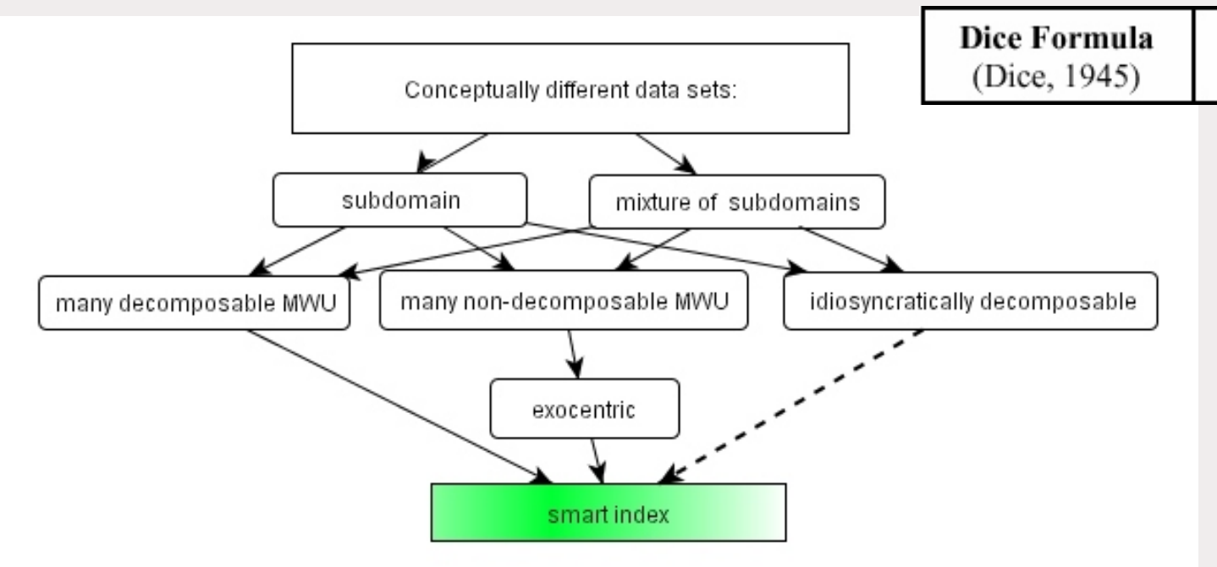

Semantic Grid
Globality
File Access
Redundancy
Parellelism
Distribution
Grid Computing
Orchestration of web services
Distribution of language dependent LRs / PRs
Distribution of complex, expensive TM processes
Language Infrastructure
Proof
Logic
Vocabulary
RDF
Ontological Infrastructure
Semantic Web

NLP for the Semantic Web

- Language Infrastructure (1): Ontology Conceptualisation and Mapping by multi word unit identification. Use frequency and information based methods. Collocation Finding: knock (at) the door, make up, prime minister, etc.

METHOD	FORMULA
Frequency (Guiliano, 1964)	f_{XY}
Selectional Association (Resnik, 1996)	$\frac{P_{X\mid Y} * MI_{XY}}{\sum_Z Pr_{Z\mid Y} * MI_{ZY}}$
Log-likelihood (Dunning, 1993; Daille, 1996)	$-2\log\frac{[P_X P_Y P_{\bar{X}} P_{\bar{Y}}]^{f_Y}}{[P_{XY} P_{\overline{XY}}]^{f_{XY}} [P_{X\bar{Y}} P_{\bar{X}Y}]^{f_{\bar{X}Y}}}$
Student's t-Score (Church and Hanks, 1990)	$\frac{f_{XY} - \xi_{XY}}{\sqrt{f_{XY}(1-(f_{XY}/N))}}$

METHOD	FORMULA
Pointwise Mutual Information (MI) (Fano, 1961; Church and Hanks, 1990)	$\log_2 (P_{XY} / P_X P_Y)$
Symmetric Conditional Probability (Ferreira and Pereira, 1999)	$P_{XY}^2 / P_X P_Y$
Chi-squared (χ^2) (Church and Gale, 1991)	$\sum_{i\in\{X,\bar{X}\},\, j\in\{Y,\bar{Y}\}} \frac{(f_{ij}-\xi_{ij})^2}{\xi_{ij}}$
Dice Formula (Dice, 1945)	$2 f_{XY} / (f_X + f_Y)$

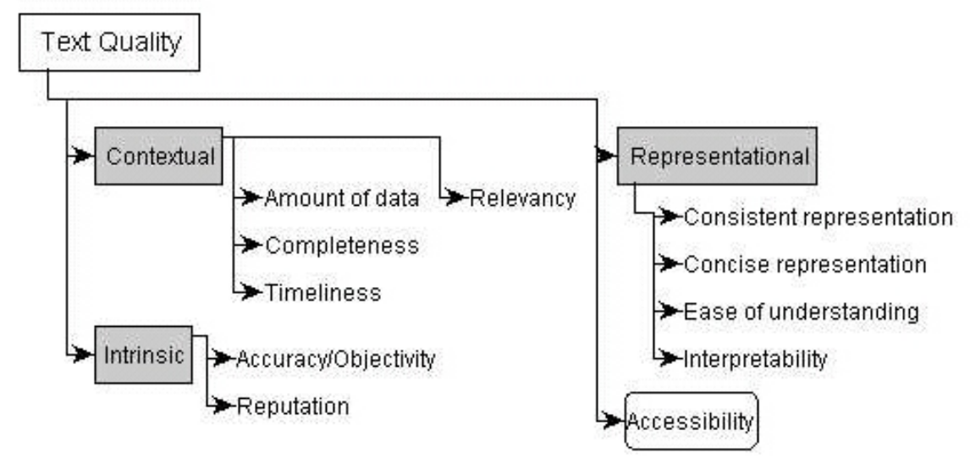

Conceptually different data sets:
subdomain
mixture of subdomains
many decomposable MWU
many non-decomposable MWU
idiosyncratically decomposable
exocentric
smart index

- Language Infrastructure (2): Ontology population, (semi-)automatic knowledge markup through Information Extraction, i.e. named entity class tagging and semantic relation tagging.

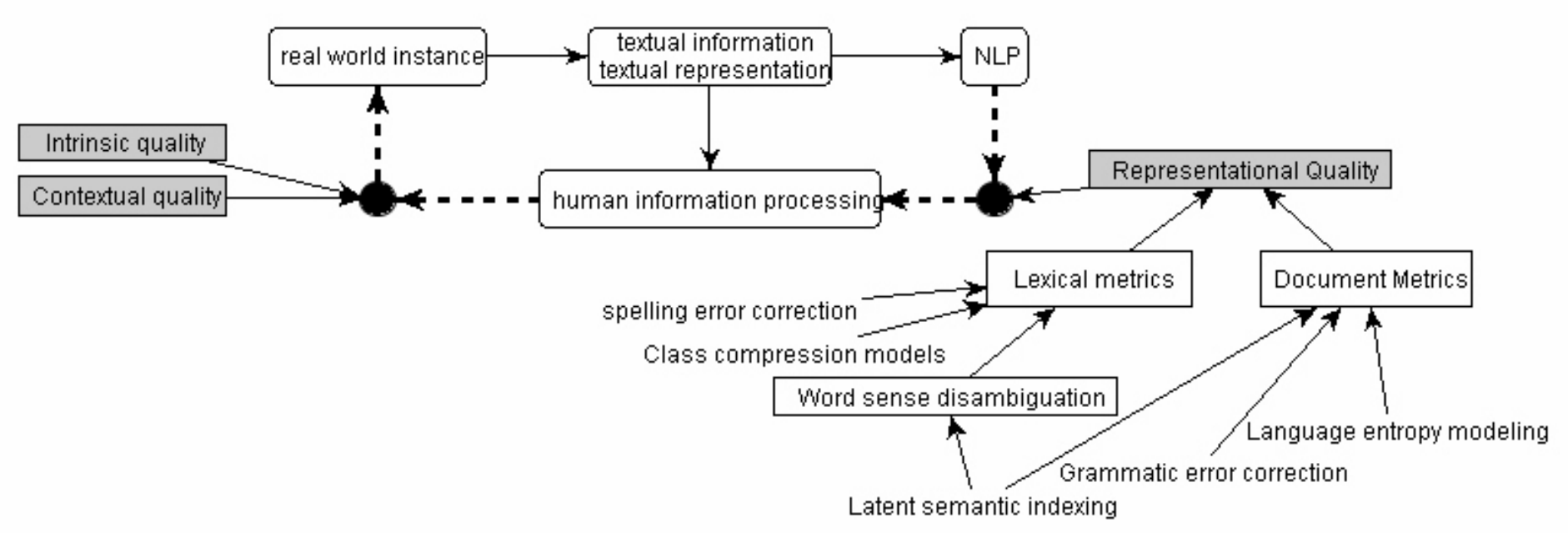
Ontology Learning provides the instrument for adapting to different application domains.

Representational Data Quality

- Representational data quality includes aspects of data formats in form of concise and consistent representation, and meaning in terms of interpretability and ease of understanding.

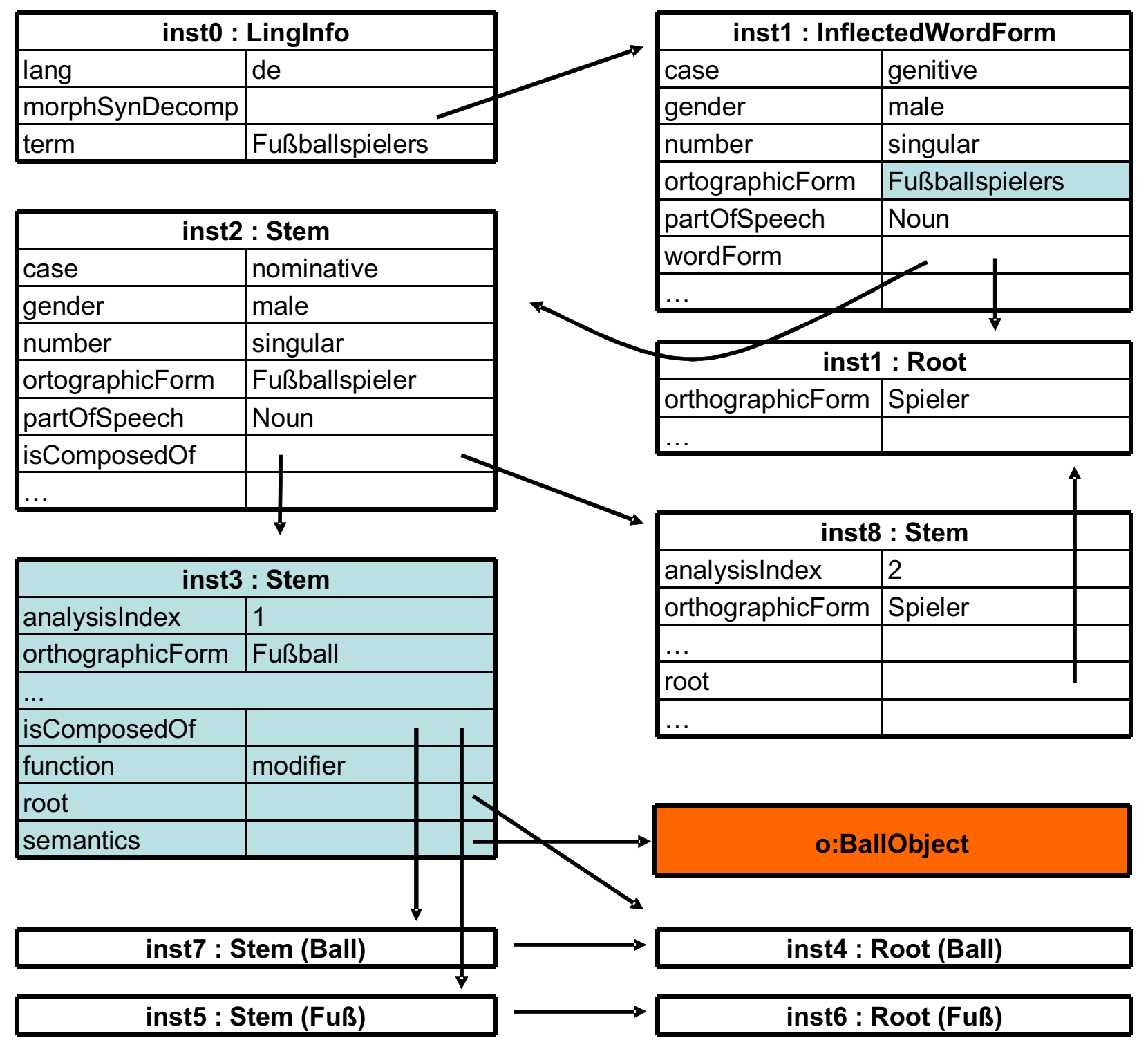


- Text Quality Audit Guideline. Better text representation means to make text more suitable for automatic NLP.


real world instance
textual information textual representation
NLP
Intrinsic quality
Contextual quality
human information processing
Representational Quality
Lexical metrics
Document Metrics
spelling error correction
Class compression models
Word sense disambiguation
Language entropy modeling
Grammatic error correction
Latent semantic indexing

Semantic Web for NLP

- Language infrastructure (2): Semantic Web data structures organise texts in different languages for better representational data quality.
- LingInfo morphosyntactic decomposition instance example of german compound: Fussballspielers (of the football player) according to a domain football player ontology (SWintO).



inst0 : LingInfo	
lang	de
morphSynDecomp	
term	Fußballspielers

inst1 : InflectedWordForm	
case	genitive
gender	male
number	singular
ortographicForm	Fußballspielers
partOfSpeech	Noun
wordForm	
…	

inst2 : Stem	
case	nominative
gender	male
number	singular
ortographicForm	Fußballspieler
partOfSpeech	Noun
isComposedOf	
…	

inst1 : Root	
orthographicForm	Spieler
…	

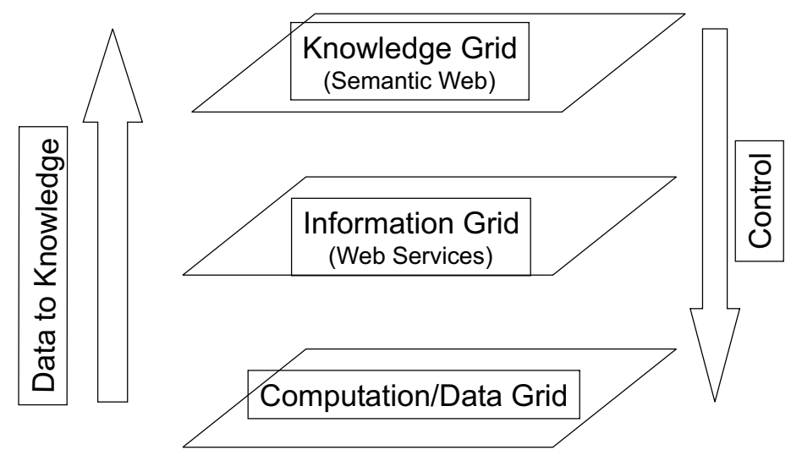
inst3 : Stem	
analysisIndex	1
orthographicForm	Fußball
...	
isComposedOf	
function	modifier
root	
semantics	

inst8 : Stem	
analysisIndex	2
orthographicForm	Spieler
…	
root	
…	

- Close Semantic Gap by ontological text annotations.

Semantic Grid Computing

- Semantic Grid is defined to emphasise the use of Semantic Web technology in the Grid. The Grid is the seamless access to distributed computing and information resources (semanticgrid.org).
- Distribution of Language and Processing Resources over geographically dispersed systems. LRs refer to data resources such as lexica, corpora, thesauri, and ontologies. PRs refer to programmatic or algorithmic sources such as distributed classifiers, POS taggers, NE recognisers, or grammatical parsers.
- Computational Grid vs. Data Grid and three layer Grid abstraction.


Knowledge Grid (Semantic Web)
Information Grid (Web Services)
Computation/Data Grid
Data to Knowledge
Control

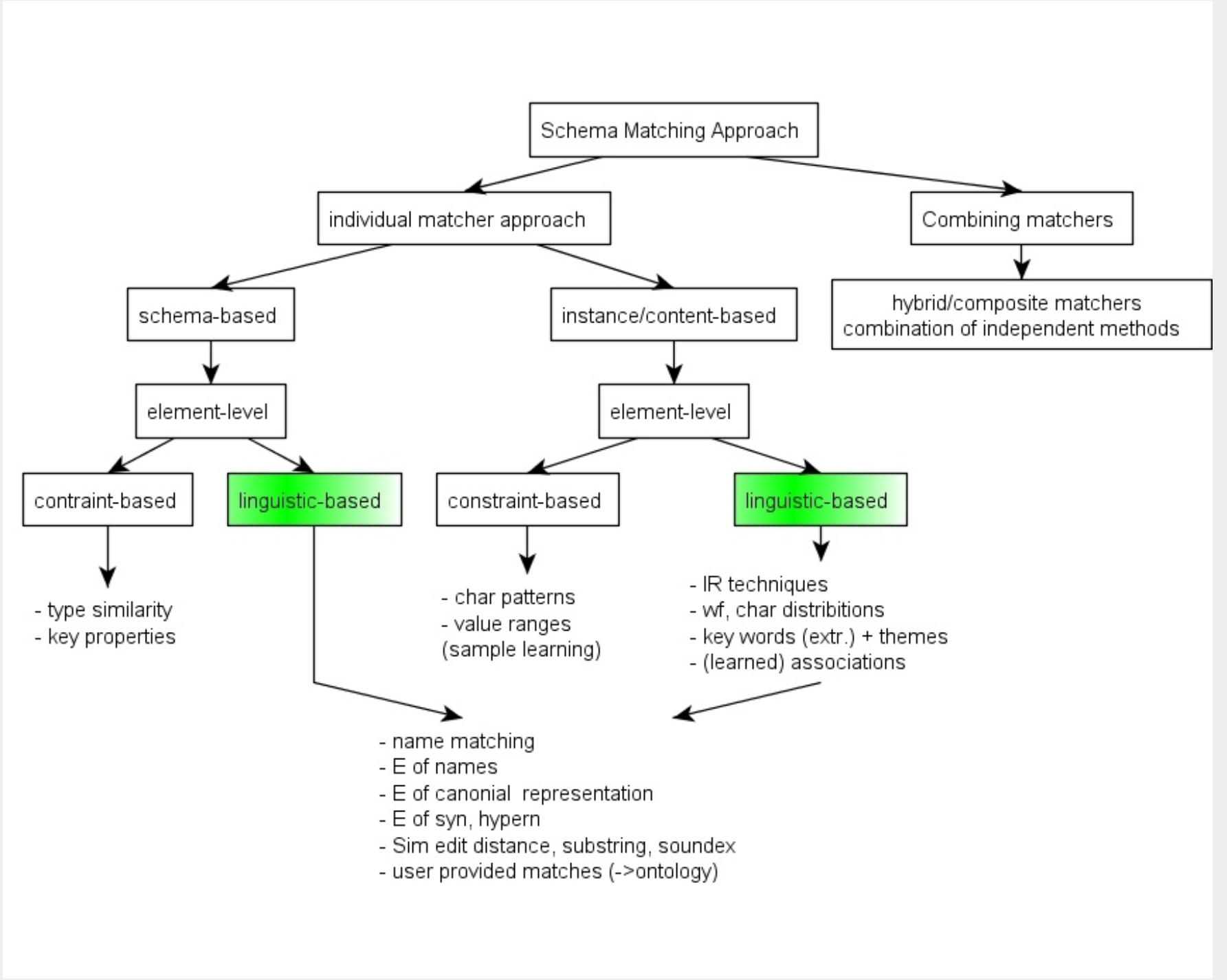
- Resource Reasoning: discovery, selection, composition

Embedded Text Mining Workflow

- (1) Build local (semantic web) services for e.g. text classification by building special purpose classifiers for different natural languages realised through distributed access to a multitude of such classifiers.
- (2) Model PRs and LRs communities (e.g. equivalence classes) by ontologies.
- (3) Annotate texts and spin your Semantic Web.
- (4) Enable e.g. distributed document access and retrieval to confidential data.
- (5) Decide on different NLP pipelines.
- (6) Deploy embedded knowledge assisted Text Processing services.
- (7) Automatise previous stages by Text Mining, i.e. automatic document markup, automatic ontology population, ontology concept and relation learning, and automatic ontology mapping, i.e. automatic schema mapping.

Database Technology

- Schema Matching Problems: unknown synonyms/hyponyms, foreign-language data material, cryptic attributes
- Automatic (Linguistic) Schema Matching Approaches



- Use DB operators and ontologies for data scalability and data representation, respectively.
- Expand queries by ontological content operators. These operators are useful in keyword searching, collocation searches, and pattern matching searches.

The table below shows some of the more complex operators available in the search contexts of the vendor implementations.

Query expansion operator	DB2	Oracle	SQL Server	Informix
Fuzzy term matches to include words that are spelled similarly to the query term.	•	•	-	•
Taxonomy search to include more specific or more general terms.	•	•[24]	-	-
Proximity search to test whether two words are close to each other, i.e. near positions.	•	•	•	•
Related term matches to expand the query by related terms defined in a thesaurus.	•	•	•	•
Term replacement to replace a term in a query with a preferred term defined in a thesaurus. Could also be used for synonym searches.	•	•	•	•

Standards

- RDF(S), OWL, SOAP, WSDL, DAMSL, Topic Maps
- MPEG-7 for describing multimedia content data.
- SmartWeb SWEMMA: W3C EMMA Extension for representing ontological queries, results, and status objects.
- Open Grid Services Architecture (OGSA), Service-Oriented Architecture (SOA), ebXML (ISO/TS15000)
- Predictive Model Markup Language (PMML)

NATO Advanced Study Institute on Mining Massive Data Sets for Security