Web Information Extraction Systems for Web Semantization

Jan Dědek[1,2]

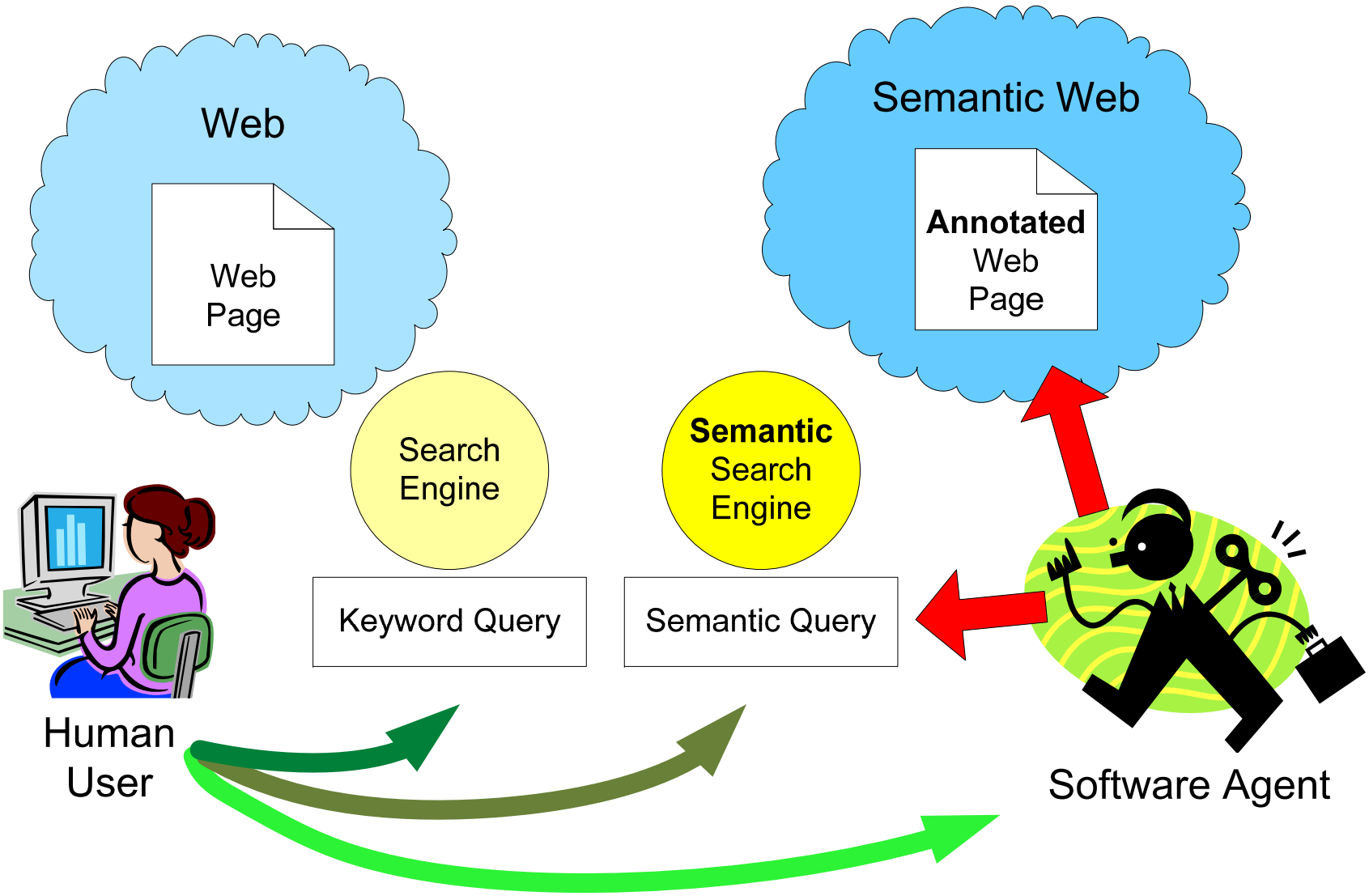
[1]Department of Software Engineering, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

[2]Institute of Computer Science, Academy of Sciences of the Czech Republic

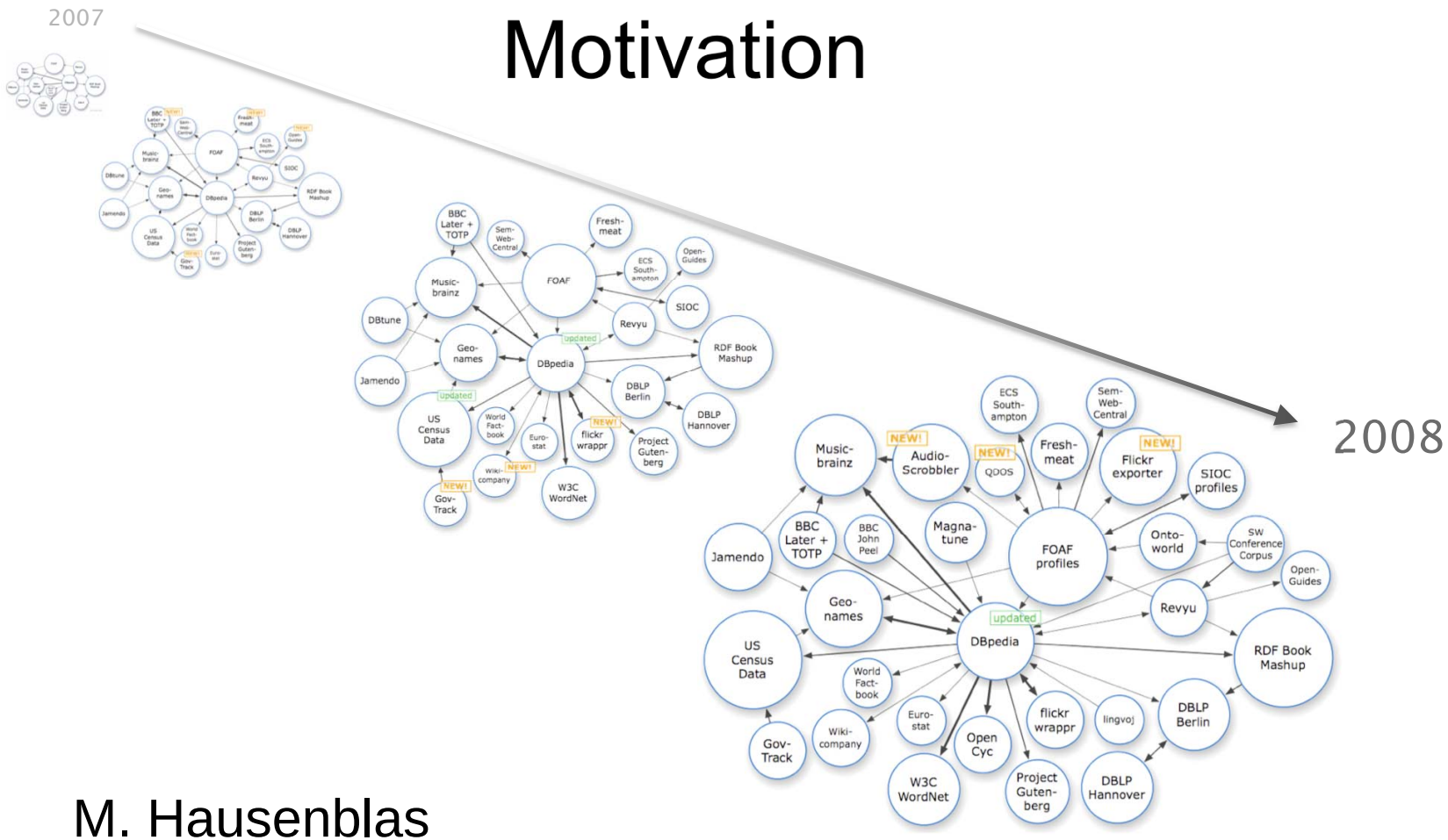
ITAT 2009
28 September, Kralova studna

Outline

The Semantic/Semantized Web in Use



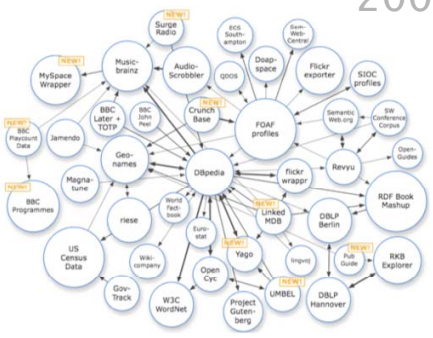
Growing of LOD data set 2007–2008



M. Hausenblas

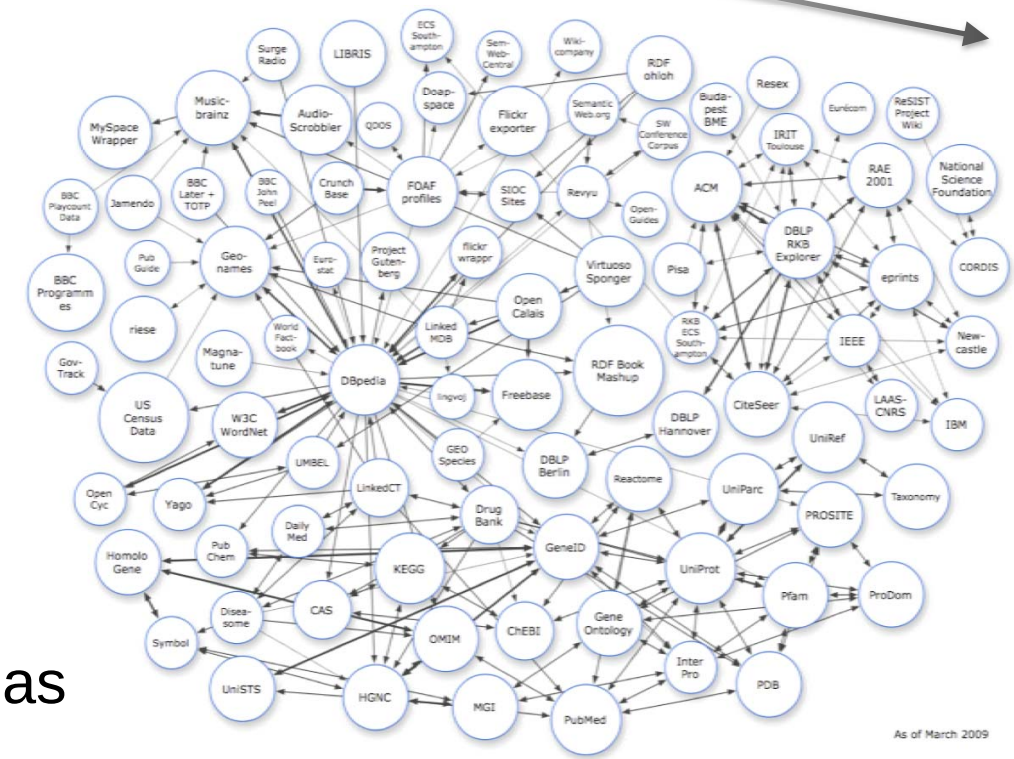
der Bundeswehr Universität München

Growing of LOD data set 2008–2009



Motivation

2009



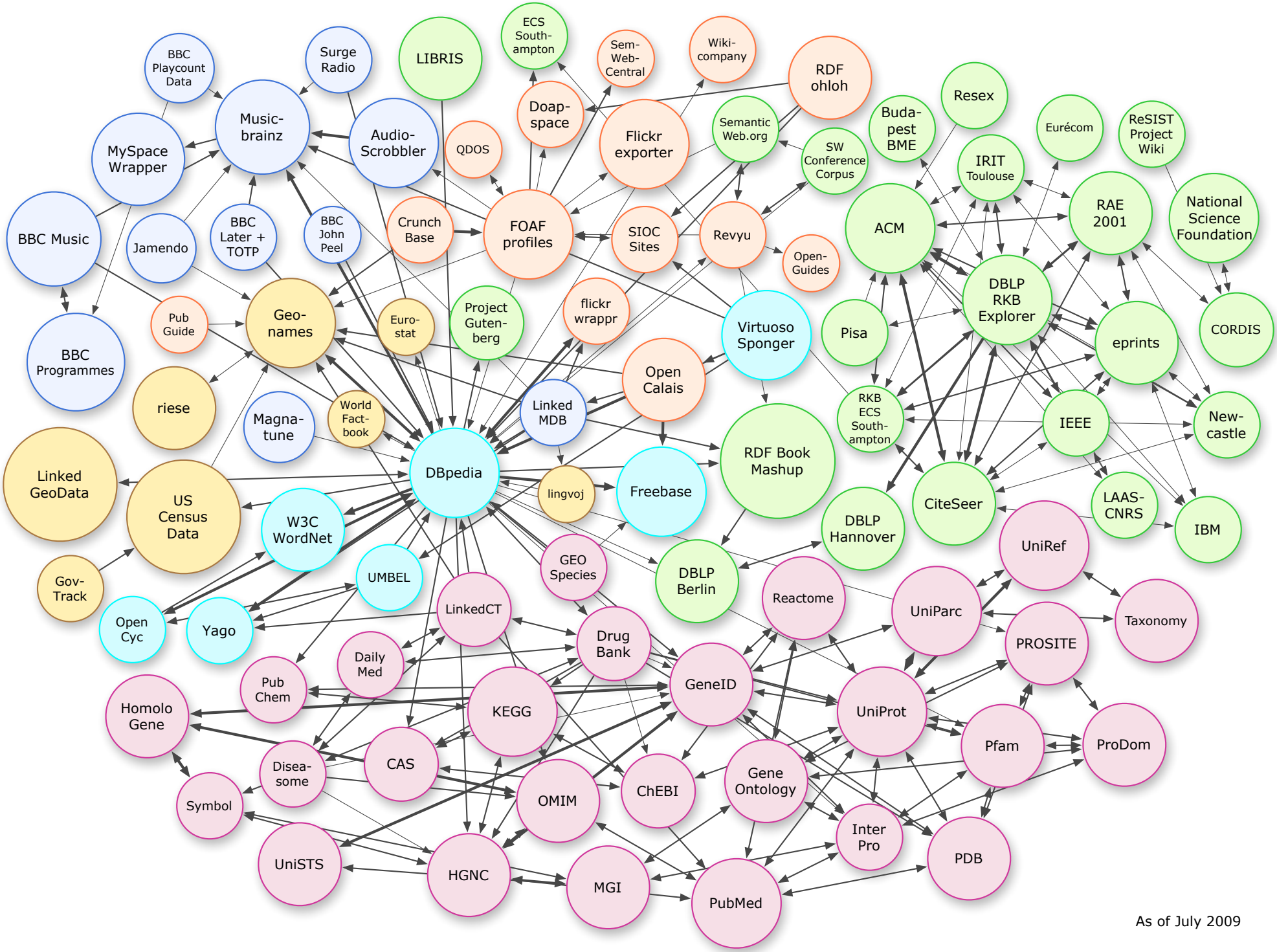
M. Hausenblas

der Bundeswehr Universität München

LOD data set statistics as of July 2009

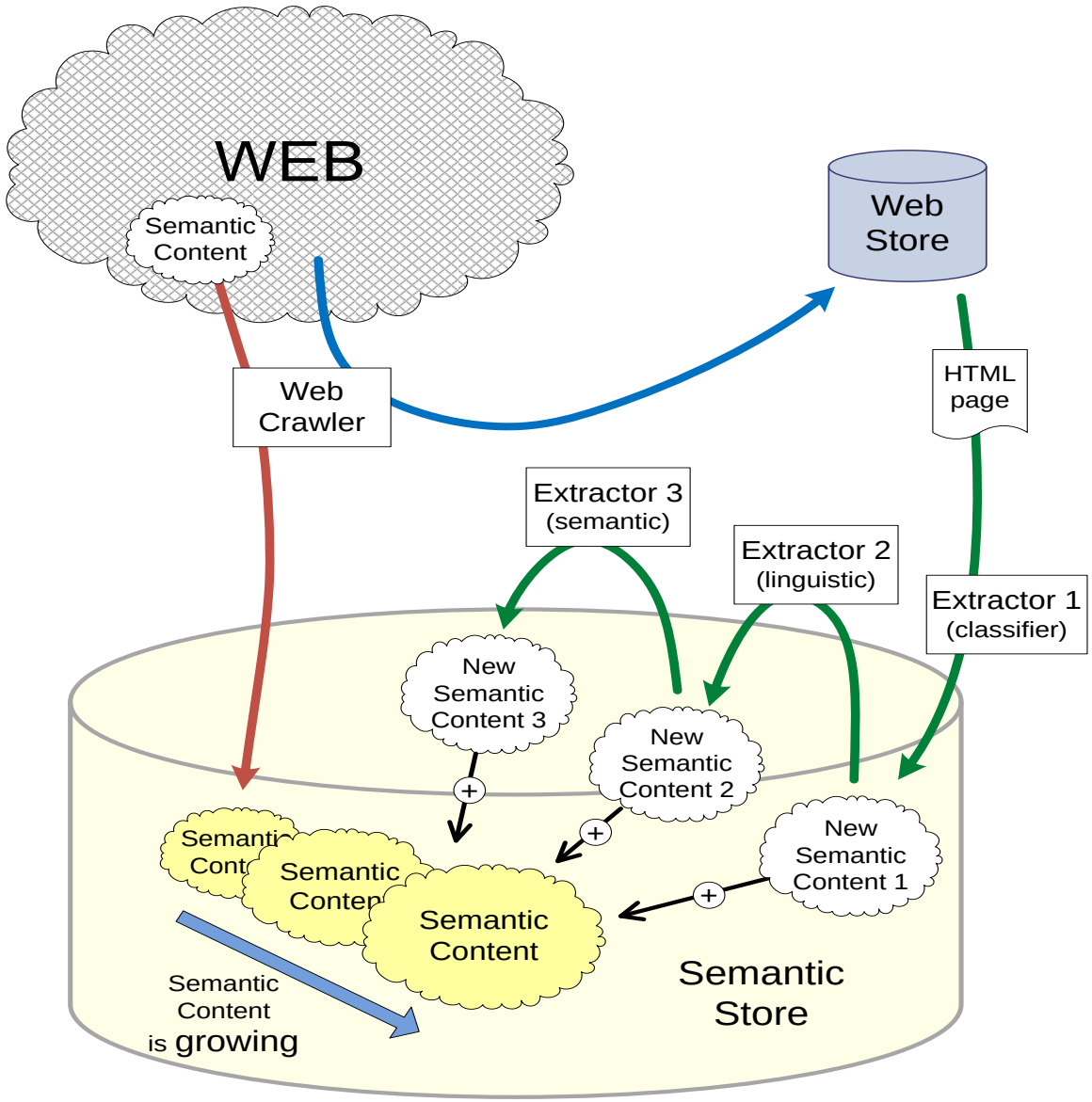
Domain	*No of Triples*	*% of Cloud*	*No of Links*	*% of Links*
Media	698.000.000	10,4%	1.238.000	0,8%
Publications	212.000.000	3,2%	4.922.000	3,3%
Life Sciences	2.429.000.000	36,1%	133.199.000	89,4%
Geographic Data	3.097.000.000	46,0%	4.038.000	2,7%
User Generate Content	76.000.000	1,1%	1.559.000	1,0%
Cross-Domain	214.000.000	3,2%	3.992.000	2,7%
Total	***6.726.000.000***		***148.948.000***	

Christian Bizer: The Web of Linked Data (26/07/2009)

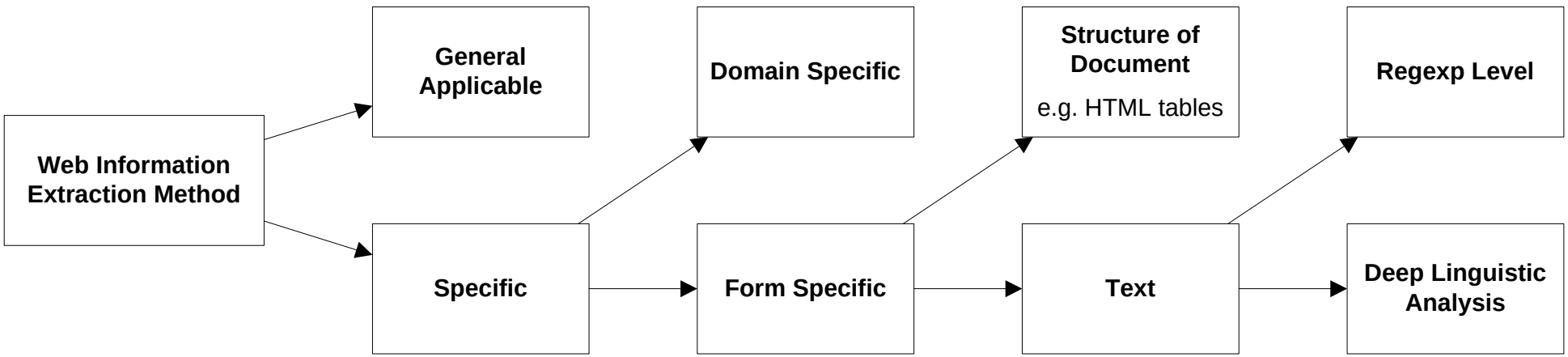


As of July 2009

Web Semantization

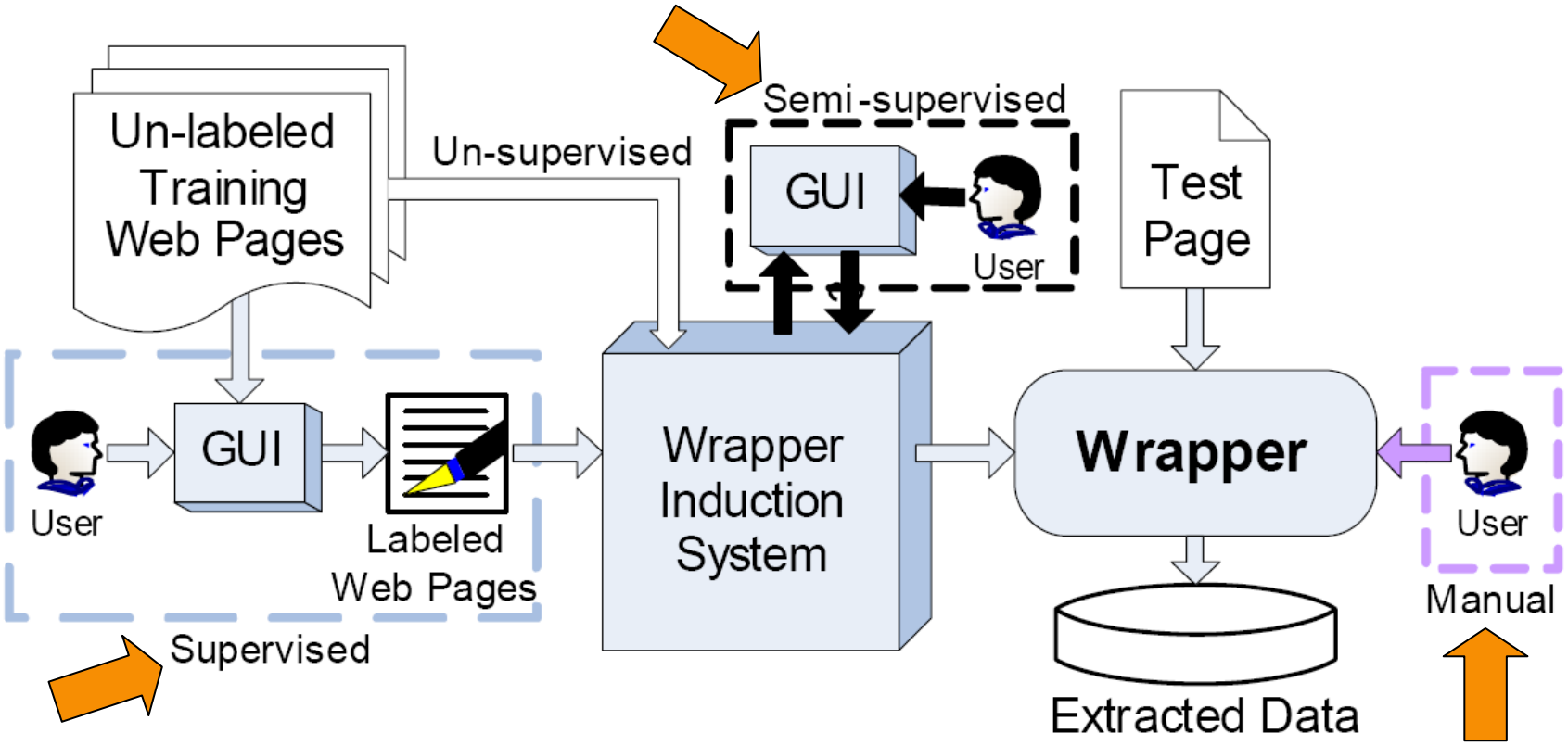


Division of extraction methods



- General Applicable
- Domain Specific
- Form Specific

A general view of WI systems



Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled F. Shaalan, "**A Survey of Web Information Extraction Systems**," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 10, pp. 1411-1428, October, 2006.

Classical tasks of text preprocessing and linguistic analysis

Text Extraction – e.g from HTML, PDF or DOC,

Tokenization – detection of words, spaces, punctuations, etc.,

Segmentation – sentence and paragraph detection,

POS Tagging – part of speech assignment, often including lemmatization and morphological analysis,

Syntactic Analysis (often called linguistic *parsing*) – assignment of the grammatical structure to given sentence with respect to given linguistic formalism (e.g. formal grammar),

Coreference Resolution (or *anaphora resolution*) – resolving what a pronoun, or a noun phrase refers to. These references often cross boundaries of a single sentence.

Classical domain dependent IE tasks

Named Entity Recognition: This task recognizes and classifies named entities such as persons, locations, date or time expression, or measuring units. More complex patterns may also be recognized as structured entities such as addresses.

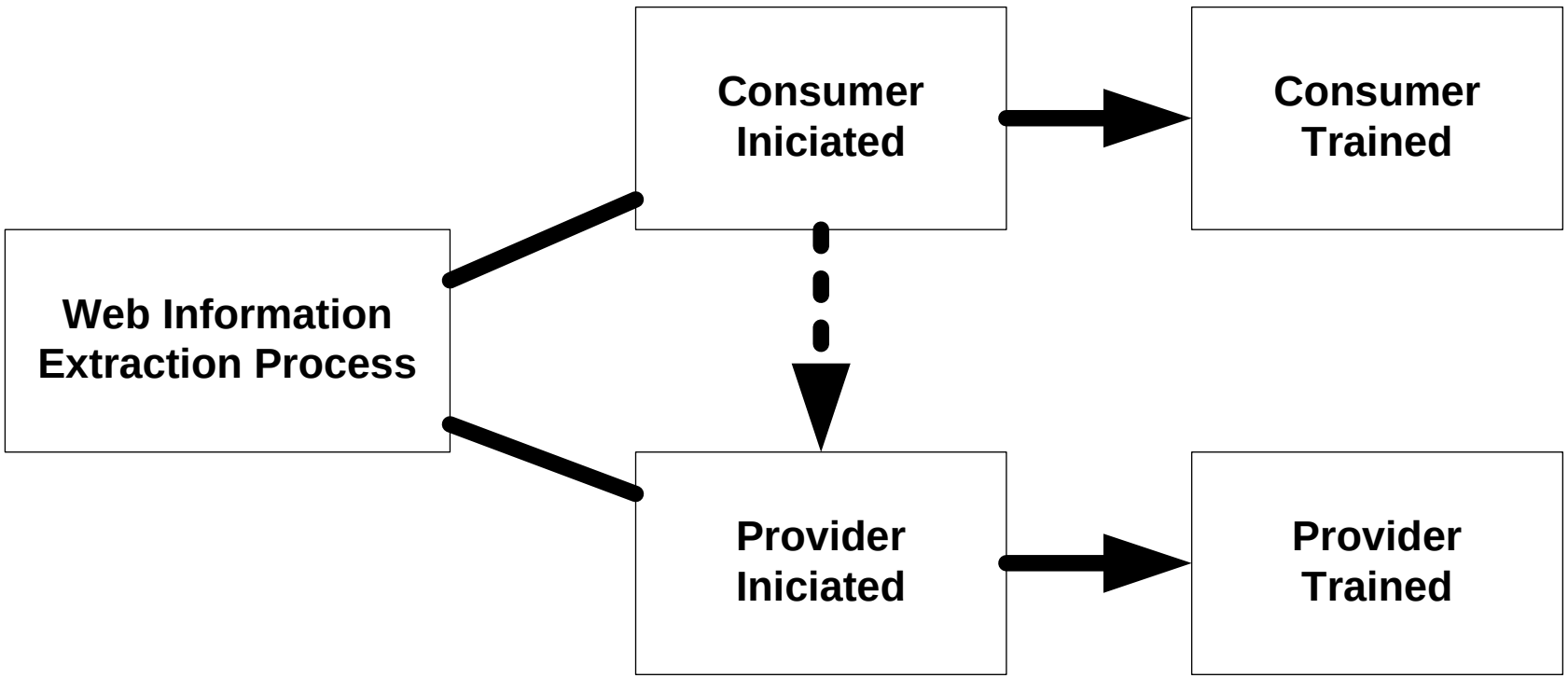
Template Element Construction: Populates templates describing entities with extracted roles (or attributes) about one single entity. This task is often performed stepwise sentence by sentence, which results in a huge set of partially filled templates.

Template Relation Construction: As each template describes information about one single entity, this tasks identifies semantic relations between entities.

Extraction Based on Structural Similarity

Linguistic Information Extraction

User initiative and effort



Conclusion

- Future development of WIE tools and work on their adaptability to new domains.
- Integration of WIE tools to the web semantization system.
- Development of the methodology and software to support the extension of the semantization system to a new domain for a non-expert user.