OCAS: Ontology-Based Corpus and Annotation Scheme

Alexander Grothkast, University of Kaiserslautern
Benjamin Adrian, DFKI
Kinga Schuhmacher, DFKI
Andreas Dengel, DFKI

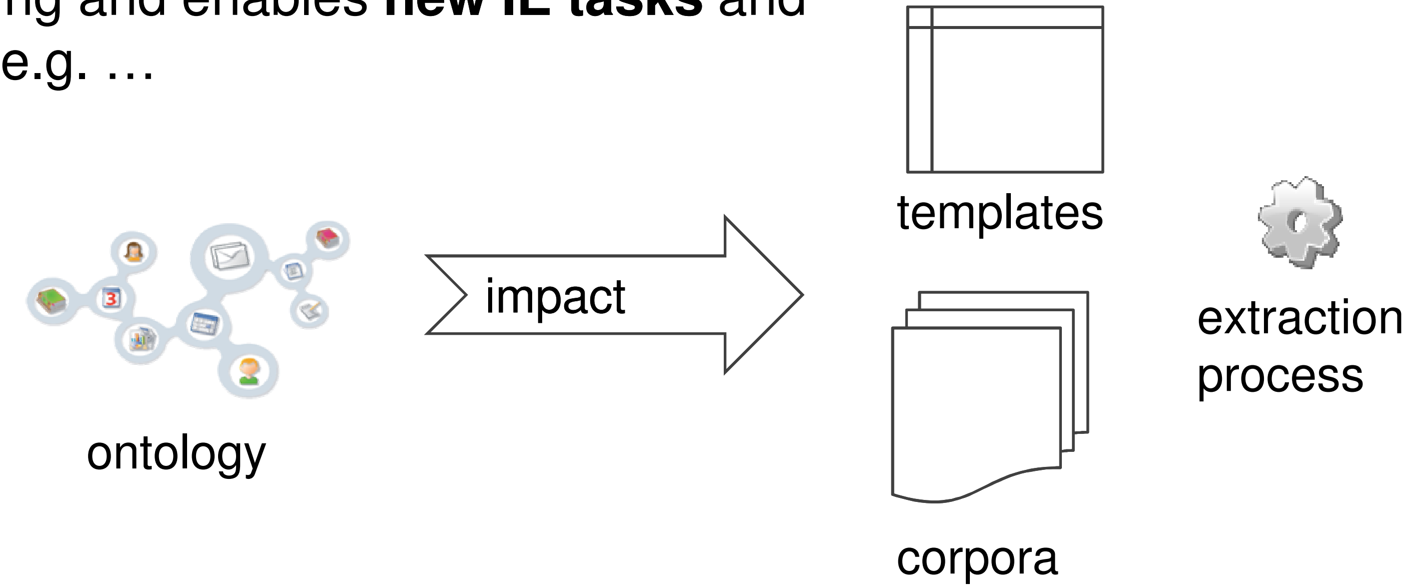
Outline

1. Evaluation of Ontology-based Information Extraction
2. The Process of an Ontology-based Corpus and Annotation Scheme
3. A Review about the OCAS 2008 corpus
4. Summary
5. Outlook

Evaluating OBIE systems

OBIES: In IE, **Ontologies** are used for enhancing results by the use of **symbolic domain knowledge**.

Compared to traditional IE, Ontology-based IE changes existing and enables **new IE tasks** and architectures, e.g. …

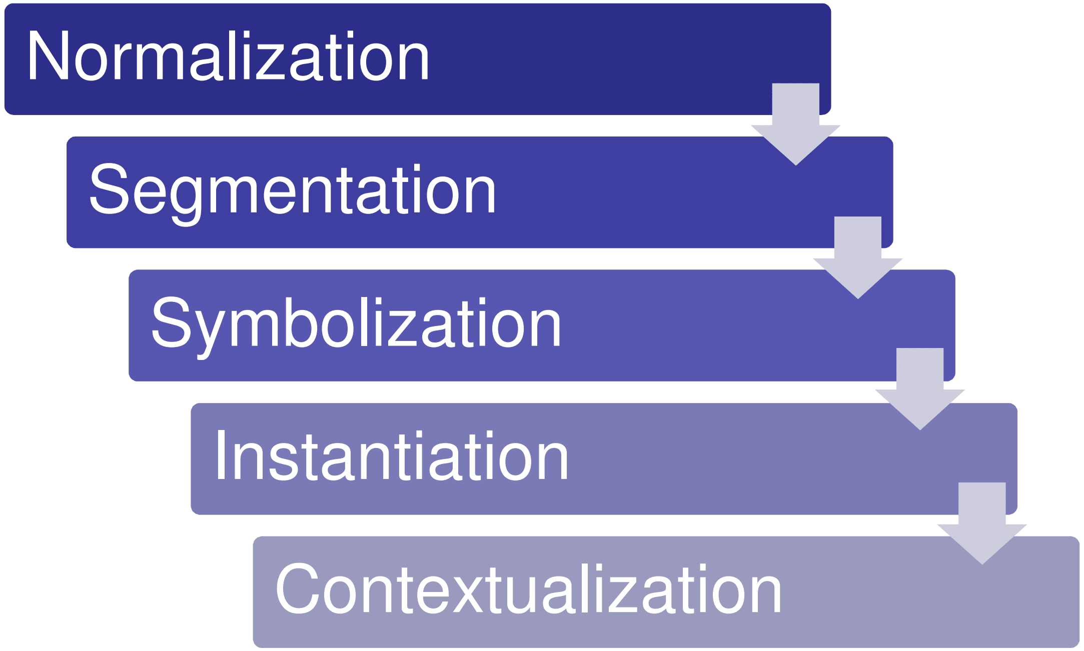


? Traditional IE evaluation methods do not suffice. (Maynard, AHM 2005)
How do we evaluate an ontology-based IE system?

Generic architecture of an OBIE system

A comparison of different OBIE systems needs a shared view on architectures.

The **OBIE layer cake** abstracts from a common OBIE process.



Normalization

Segmentation

Symbolization

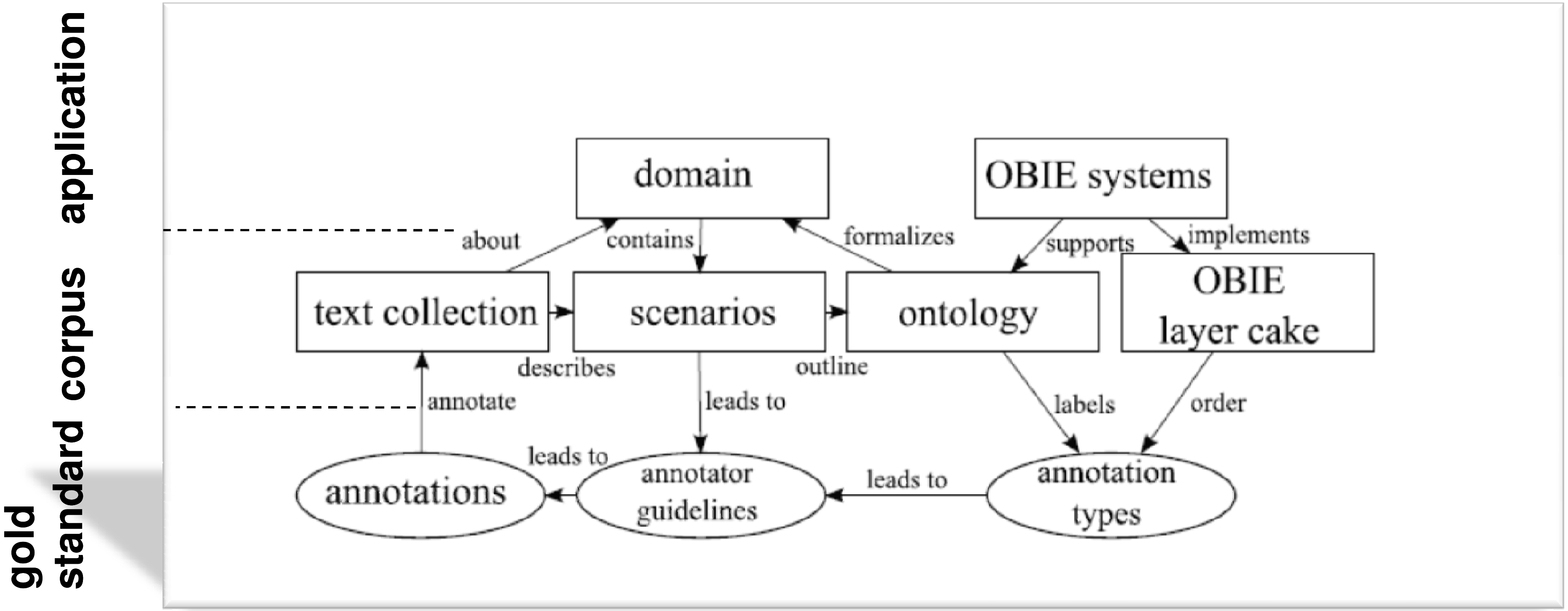
Instantiation

Contextualization

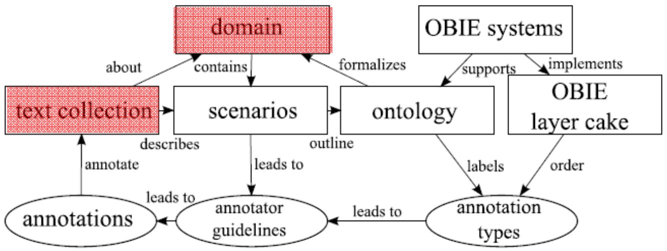
❗ For evaluating a certain OBIE systems with OCAS, its tasks have to be mapped to the OBIE layer cake.

Ontology-Based Corpus and Annotation Scheme

The process for creating high quality OBIE corpora:



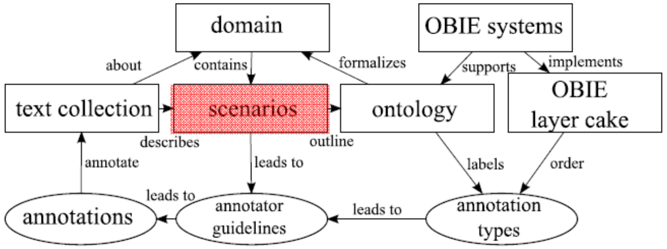
Domain and Text Collection



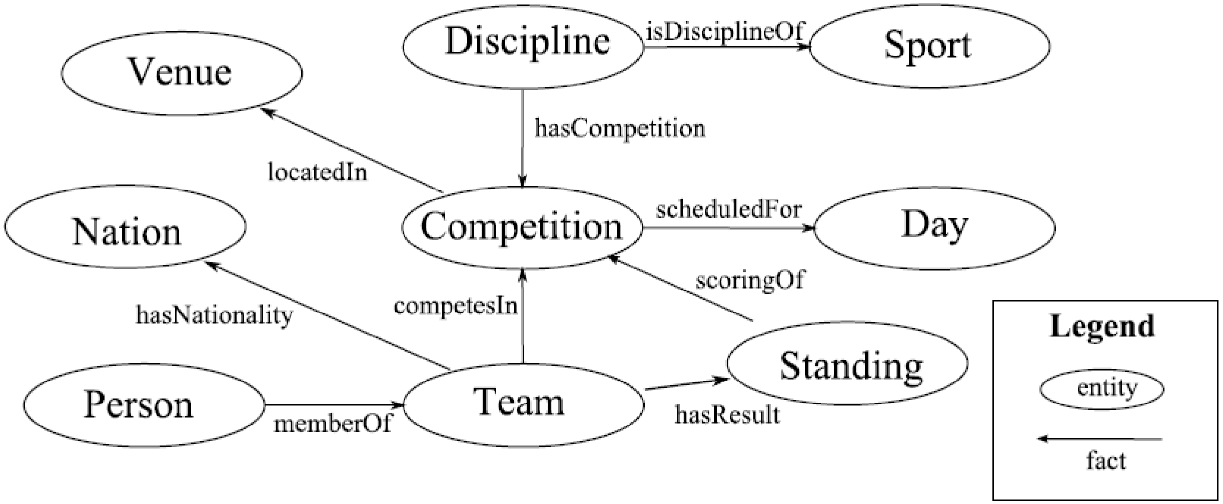
News articles (from ABC, BBC) about the Olympic summer games 2004 comprise several benefits …

- **Closeness**, limited to a few, but strictly defined concepts
- **Compactness**, concepts of domain model are highly coherent
- **Richness**, concepts can be covered completely by a text corpus

Scenario Maps



Formal representation frame of a text.

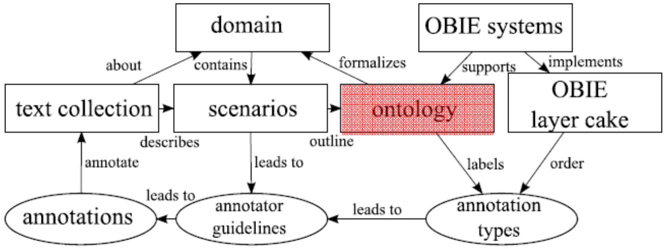


In OCAS, we used two scenario maps:

I. Single person

II. Team

Ontology



Ontology was evolved from aggregated scenario maps.

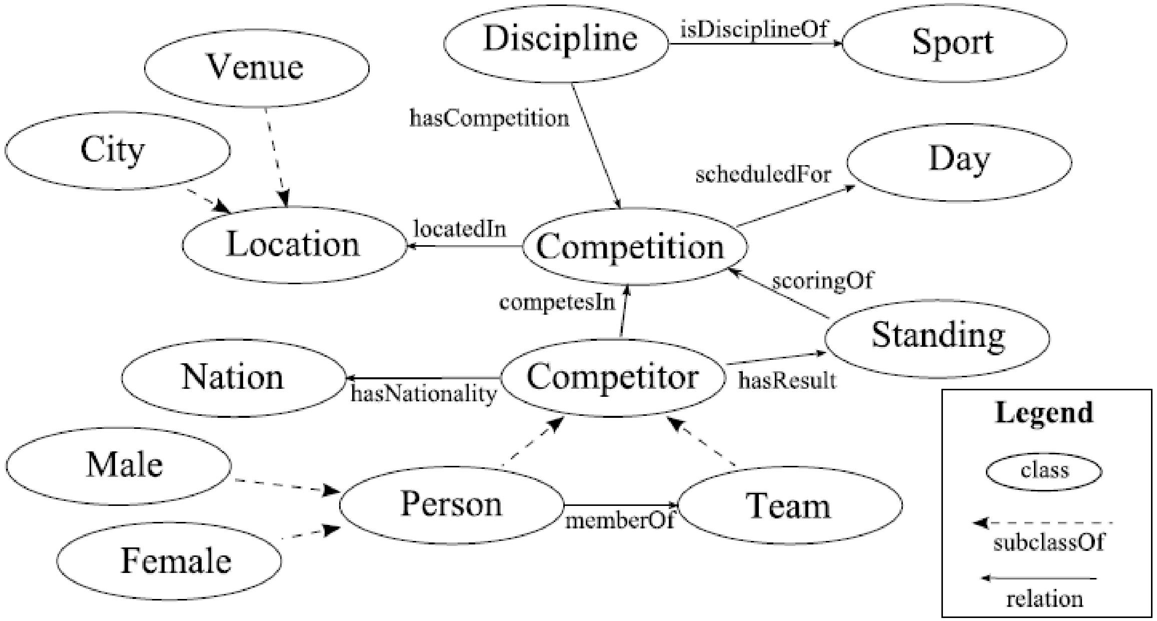
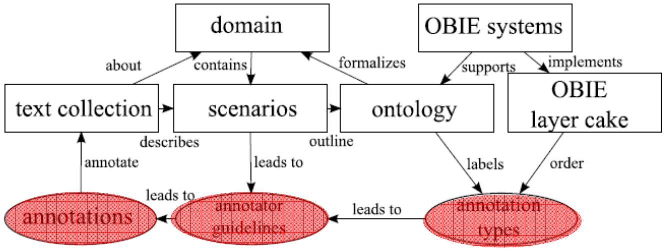
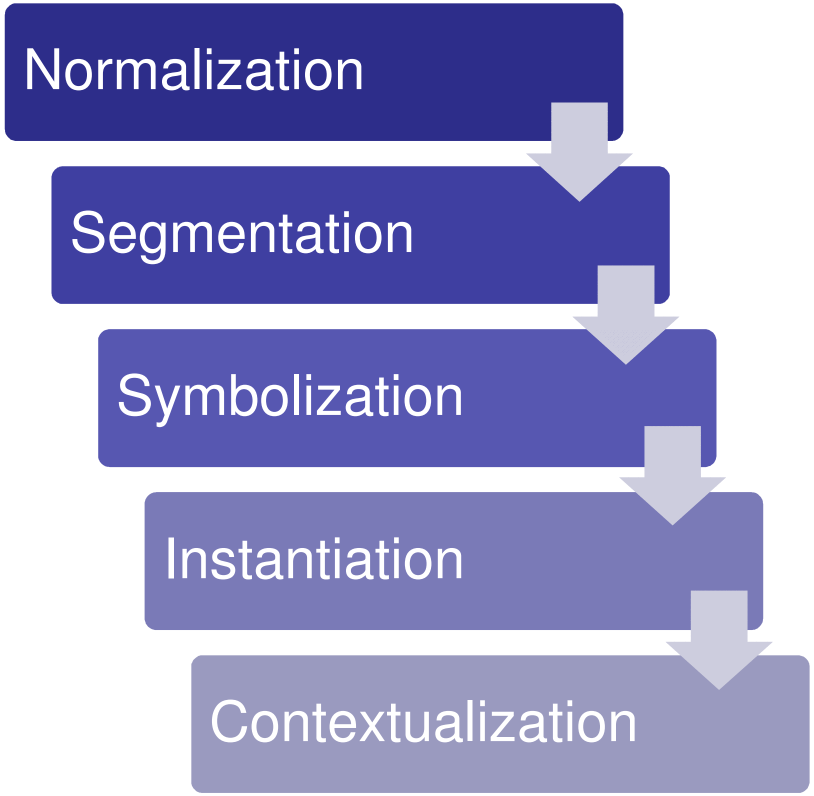


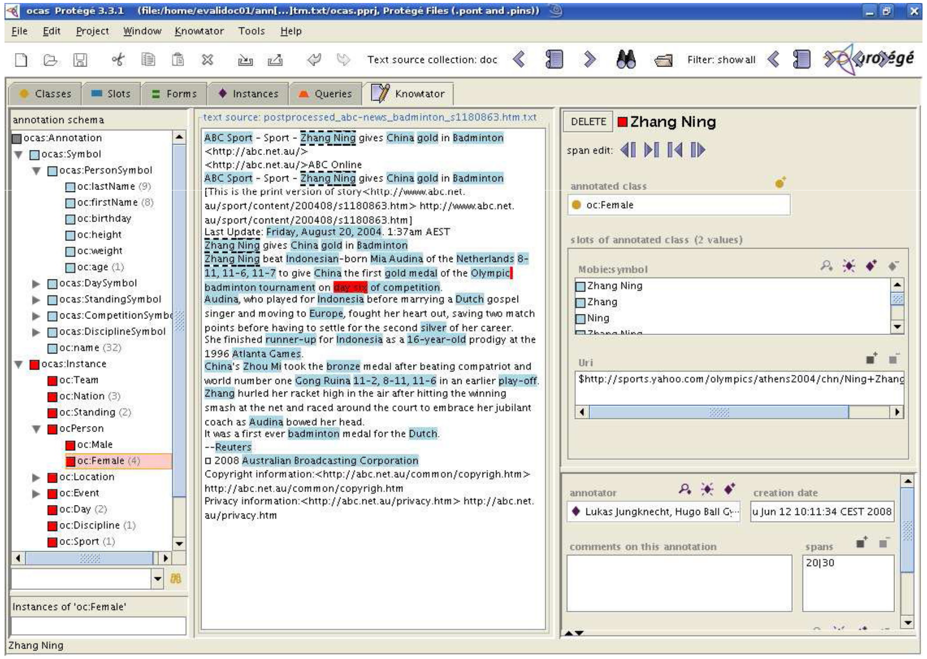
Fig. 2. The ontology for the OCAS2008 test corpus.

The instance base was populated with data from Wikipedia and Yahoo! sports.

Annotation Types







text source: postprocessed_abc-news_badminton_s1180863.htm.txt

ABC Sport - Sport - Zhang Ning gives China gold in Badminton
<http://abc.net.au/>
<http://abc.net.au/>ABC Online
ABC Sport - Sport - Zhang Ning gives China gold in Badminton
[This is the print version of story<http://www.abc.net.au/sport/content/200408/s1180863.htm> http://www.abc.net.au/sport/content/200408/s1180863.htm]
Last Update: Friday, August 20, 2004. 1:37am AEST
Zhang Ning gives China gold in Badminton
Zhang Ning beat Indonesian-born Mia Audina of the Netherlands 8-11, 11-6, 11-7 to give China the first gold medal of the Olympic badminton tournament on day six of competition.
Audina, who played for Indonesia before marrying a Dutch gospel singer and moving to Europe, fought her heart out, saving two match points before having to settle for the second silver of her career.
She finished runner-up for Indonesia as a 16-year-old prodigy at the 1996 Atlanta Games.
China's Zhou Mi took the bronze medal after beating compatriot and world number one Gong Ruina 11-2, 8-11, 11-6 in an earlier play-off.
Zhang hurled her racket high in the air after hitting the winning smash at the net and raced around the court to embrace her jubilant coach as Audina bowed her head.
It was a first ever badminton medal for the Dutch.
--Reuters
□ 2008 Australian Broadcasting Corporation
Copyright information:<http://abc.net.au/common/copyrigh.htm> http://abc.net.au/common/copyrigh.htm
Privacy information:<http://abc.net.au/privacy.htm> http://abc.net.au/privacy.htm

🟥 Zhang Ning

annotated class: oc:Female

slots of annotated class (2 values)

Mobie:symbol: Zhang Ning, Zhang, Ning, Zhang Ning

Uri: $http://sports.yahoo.com/olympics/athens2004/chn/Ning+Zhang

annotator: Lukas Jungknecht, Hugo Ball G…

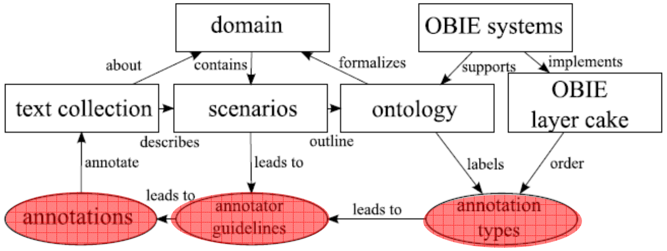
creation date: u Jun 12 10:11:34 CEST 2008

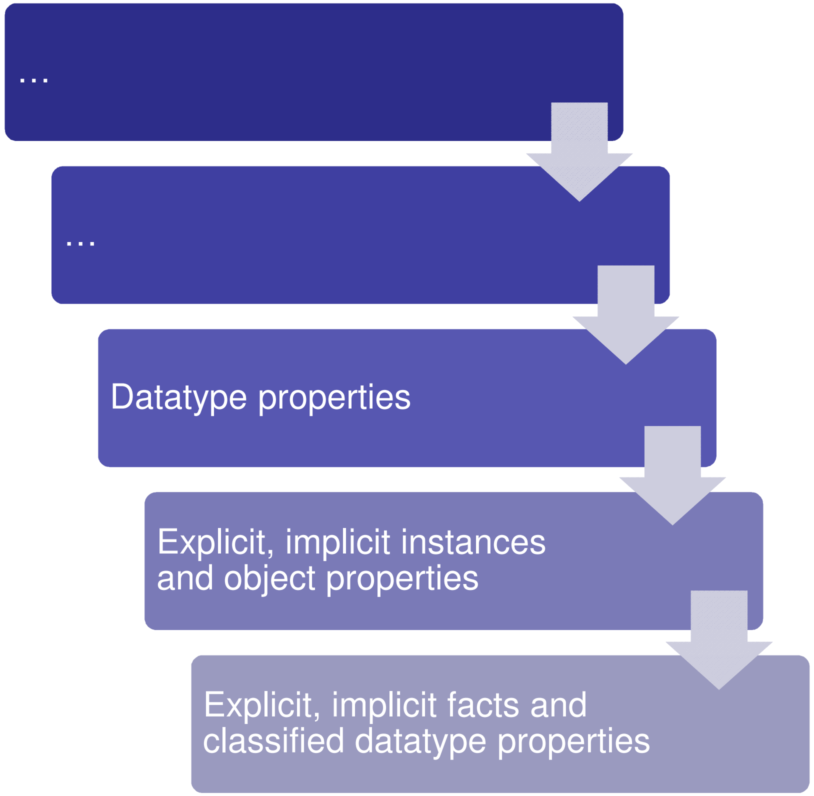
comments on this annotation:

spans: 20|30

Knowtator, Ogren, P.V., Proc. of North American Chapter of ACL 2006

Annotation Types



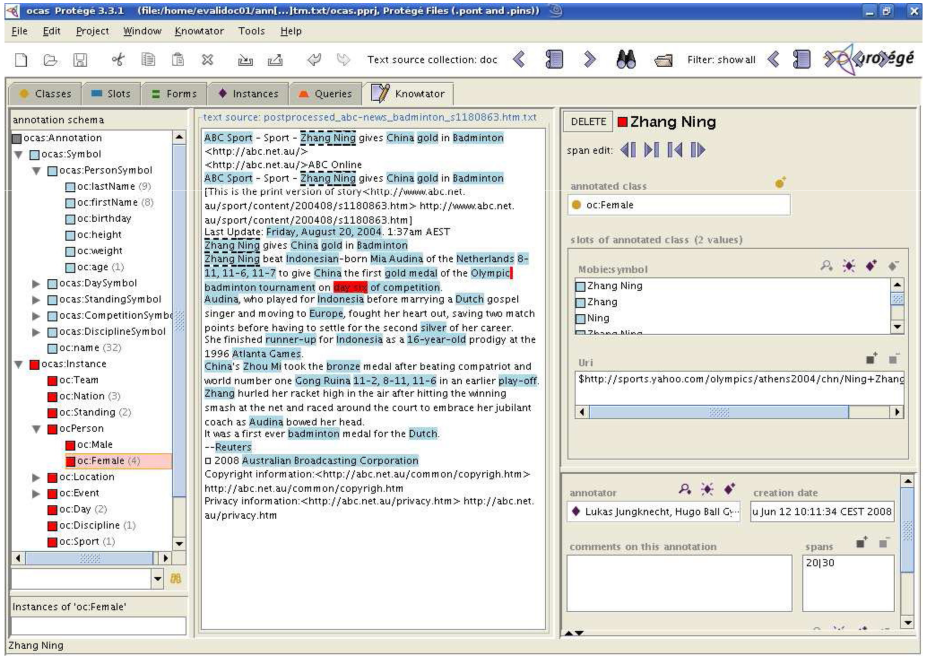

…

…

Datatype properties

Explicit, implicit instances and object properties

Explicit, implicit facts and classified datatype properties



text source: postprocessed_abc-news_badminton_s1180863.htm.txt

ABC Sport - Sport - Zhang Ning gives China gold in Badminton
<http://abc.net.au/>
<http://abc.net.au/>ABC Online
ABC Sport - Sport - Zhang Ning gives China gold in Badminton
[This is the print version of story<http://www.abc.net.au/sport/content/200408/s1180863.htm> http://www.abc.net.au/sport/content/200408/s1180863.htm]
Last Update: Friday, August 20, 2004. 1:37am AEST
Zhang Ning gives China gold in Badminton
Zhang Ning beat Indonesian-born Mia Audina of the Netherlands 8-11, 11-6, 11-7 to give China the first gold medal of the Olympic badminton tournament on day six of competition.
Audina, who played for Indonesia before marrying a Dutch gospel singer and moving to Europe, fought her heart out, saving two match points before having to settle for the second silver of her career.
She finished runner-up for Indonesia as a 16-year-old prodigy at the 1996 Atlanta Games.
China's Zhou Mi took the bronze medal after beating compatriot and world number one Gong Ruina 11-2, 8-11, 11-6 in an earlier play-off.
Zhang hurled her racket high in the air after hitting the winning smash at the net and raced around the court to embrace her jubilant coach as Audina bowed her head.
It was a first ever badminton medal for the Dutch.
--Reuters
□ 2008 Australian Broadcasting Corporation
Copyright information:<http://abc.net.au/common/copyrigh.htm> http://abc.net.au/common/copyrigh.htm
Privacy information:<http://abc.net.au/privacy.htm> http://abc.net.au/privacy.htm

Zhang Ning

annotated class: oc:Female

Mobie:symbol: Zhang Ning, Zhang, Ning, Zhang Ning

Uri: $http://sports.yahoo.com/olympics/athens2004/chn/Ning+Zhang

annotator: Lukas Jungknecht, Hugo Ball G···

creation date: u Jun 12 10:11:34 CEST 2008

spans: 20|30

Zhang Ning

Knowtator, Ogren, P.V., Proc. of North American Chapter of ACL 2006

Reviewing OCAS

Facts

- Selected **121 rich articles** from 5,000 candidates. The selected articles contain a total of 31,102 words.
- **6 persons** annotated during **8 days**.
- Annotation process took a total of **176 person hours**.
- Results in **5.66 hours** for annotating **1,000 words**.

Planned Quality Assurance

Completeness - documents have to be annotated completely with respect to scenario maps.

Consistency between annotations, e. g. each annotated instance must also be marked as symbol.

Correctness of annotations, i.e. every annotated instance in the corpus respects domain ontology’s instance set and uses the same URIs.

Summary

OCAS comprises:

a **generic OBIE architecture** called *OBIE layer cake for comparing OBIE systems* by similar subtasks,

a document **corpus of 121 news articles** (ABC, BBC) with 31,000 words about a closed domain (Olympic Summer Games 2004),

a compact **domain ontology** about the Olympic Summer Games 2004 including more than 40,000 instances from Wikipedia and Yahoo!,

two **annotation scenarios** that extend traditional template-based evaluations,

an **annotation set** that contains typed annotations according to the ontology and the *OBIE layer cake,*

annotations that concern **symbols**, **instances**, explicitly occurring **facts** and implicit induced facts

human created annotations according to predefined specifications.

Outlook

Verify quality of OCAS 2008

Evaluate iDocument with OCAS 2008

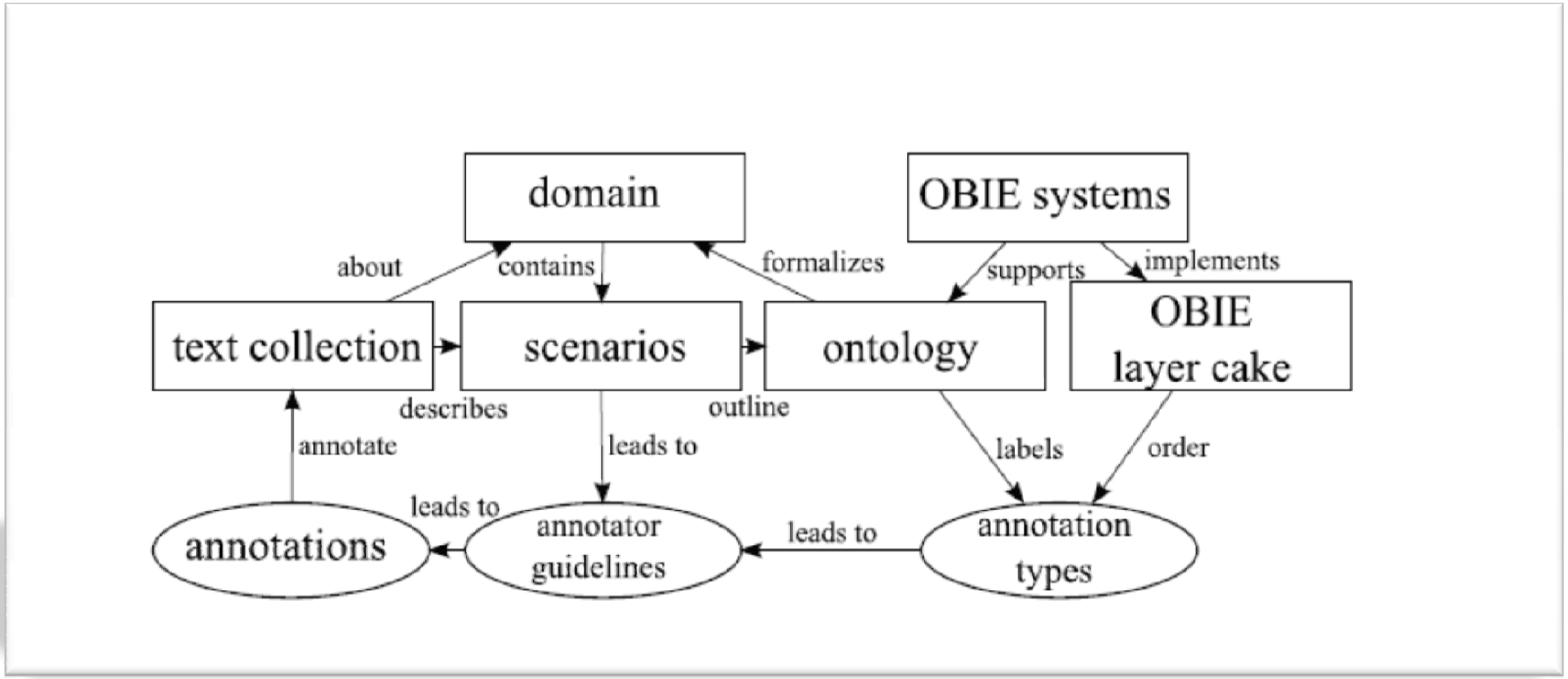
Evaluate GATE’s OBIE facilities with OCAS 2008

Publish OCAS
(RDF dump of ontology, annotated text corpus, test results)

Other ideas?

Thanks for attention

Any questions, tips, comments or remarks?



More information on OCAS will appear soon at http://idocument.opendfki.de