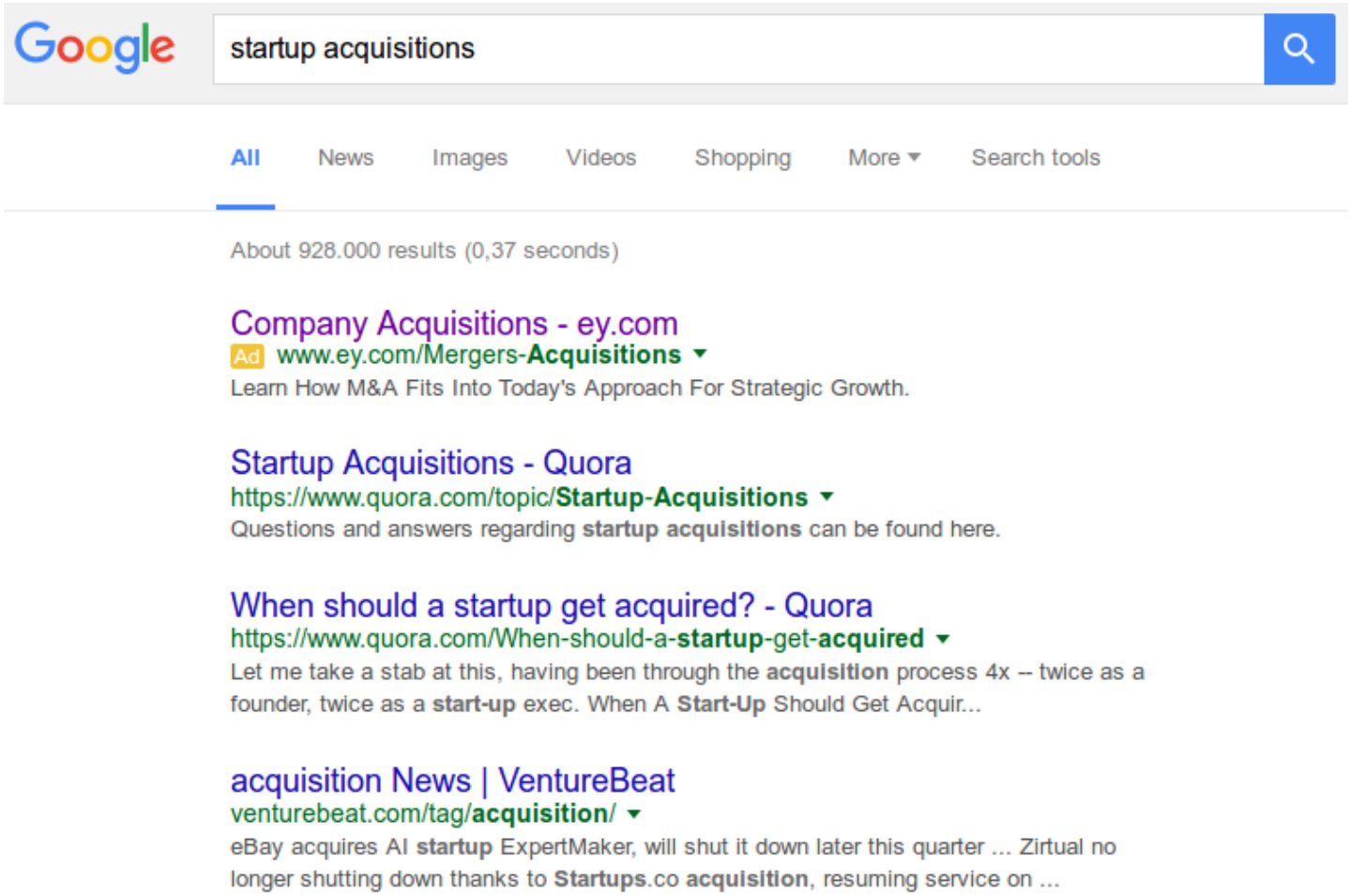
# Towards Monitoring of Novel Statements in the News

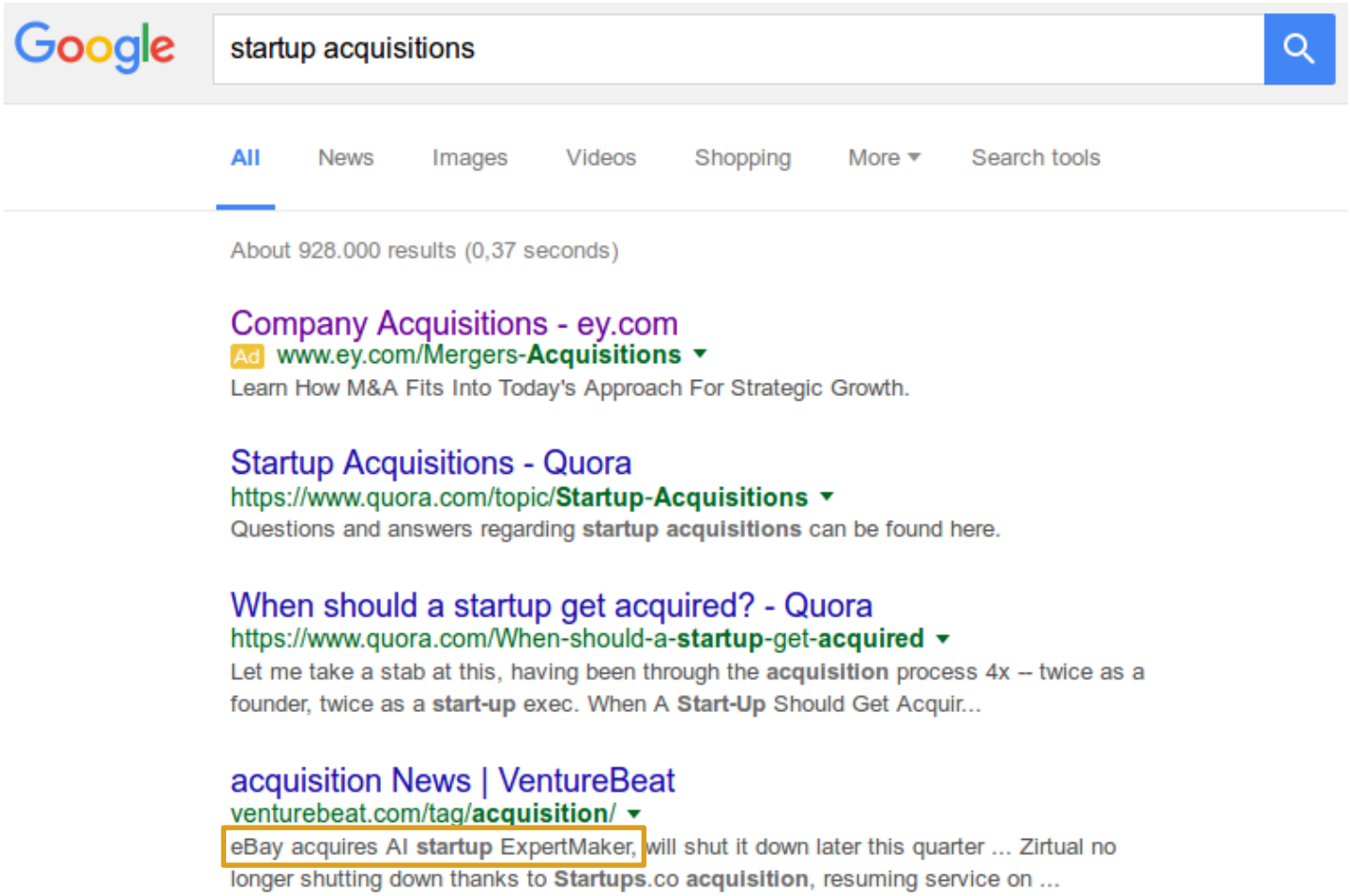
**Michael Färber,** Achim Rettinger, Andreas Harth
michael.faerber@kit.edu
**Institute AIFB, Karlsruhe Institute of Technology (KIT), Germany**



# Motivation: Search for Acquisitions of Startups



# Motivation: Search for Acquisitions of Startups



# Motivation: Search for Acquisitions of Startups

# Motivation

- Traditional IR systems only use *relevance* as criterion.
- Novelty detection systems need to consider both *relevance* and *novelty*.
- Existing novelty detection approaches are solely statistically (e.g., temporal TF-IDF).

# Related Work

| | | Extraction of and queries on statements (graph/tupel) | Grounding of extracted entities in a KB | Grounding of extracted relations in a KB | Novelty detection task implemented |
|---|---|---|---|---|---|
| IE with KB grounding | Presutti et al. 2012 | ✓ | ✓ | ✓ | |
| | Carvhalho et al. 2013 | ✓ | ✓ | | |
| | Augenstein et al. 2012 | ✓ | ✓ | ✓ | |
| | Fader et al. 2011 | ✓ | | | |
| Open IE | Del Corro et al. 2013 | ✓ | | | |
| | Mausam et al. 2012 | ✓ | | | |
| Novelty detection systems | Zhang et al. 2012 | | | | ✓ |
| | Gabrilovich et al. 2004 | | | | ✓ |
| | Karkali et al. 2013 | | | | ✓ |
| | Li et al. 2005, 2008 | | | | ✓ |
| Novelty detection tracks | Systems for TREC Nov. Track | | | | ✓ |
| | Systems for TREC KBA | ✓ | | | ✓ |

# Related Work

| | | Extraction of and queries on statements (graph/tupel) | Grounding of extracted entities in a KB | Grounding of extracted relations in a KB | Novelty detection task implemented |
|---|---|---|---|---|---|
| IE with KB grounding | Presutti et al. 2012 | ✓ | ✓ | ✓ | |
| | Carvhalho et al. 2013 | ✓ | ✓ | | |
| | Augenstein et al. 2012 | ✓ | ✓ | ✓ | |
| | Fader et al. 2011 | ✓ | | | |
| Open IE | Del Corro et al. 2013 | ✓ | | | |
| | Mausam et al. 2012 | ✓ | | | |
| Novelty detection systems | Zhang et al. 2012 | | | | ✓ |
| | Gabrilovich et al. 2004 | | | | ✓ |
| | Karkali et al. 2013 | | | | ✓ |
| | Li et al. 2005, 2008 | | | | ✓ |
| Novelty detection tracks | Systems for TREC Nov. Track | | | | ✓ |
| | Systems for TREC KBA | ✓ | | | ✓ |

# Related Work

| | | Extraction of and queries on statements (graph/tupel) | Grounding of extracted entities in a KB | Grounding of extracted relations in a KB | Novelty detection task implemented |
|---|---|---|---|---|---|
| IE with KB grounding | Presutti et al. 2012 | ✓ | ✓ | ✓ | |
| | Carvhalho et al. 2013 | ✓ | ✓ | | |
| | Augenstein et al. 2012 | ✓ | ✓ | ✓ | |
| | Fader et al. 2011 | ✓ | | | |
| Open IE | Del Corro et al. 2013 | ✓ | | | |
| | Mausam et al. 2012 | ✓ | | | |
| Novelty detection systems | Zhang et al. 2012 | | | | ✓ |
| | Gabrilovich et al. 2004 | | | | ✓ |
| | Karkali et al. 2013 | | | | ✓ |
| | Li et al. 2005, 2008 | | | | ✓ |
| Novelty detection tracks | Systems for TREC Nov. Track | | | | ✓ |
| | Systems for TREC KBA | ✓ | | | ✓ |

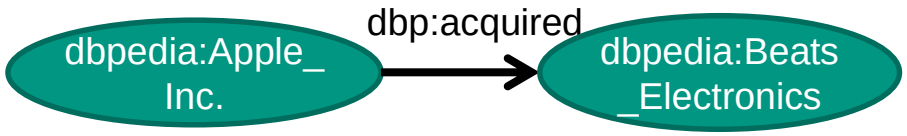
# Our Scenario

# Our Scenario

*Semantic Novel Fact Extraction System*

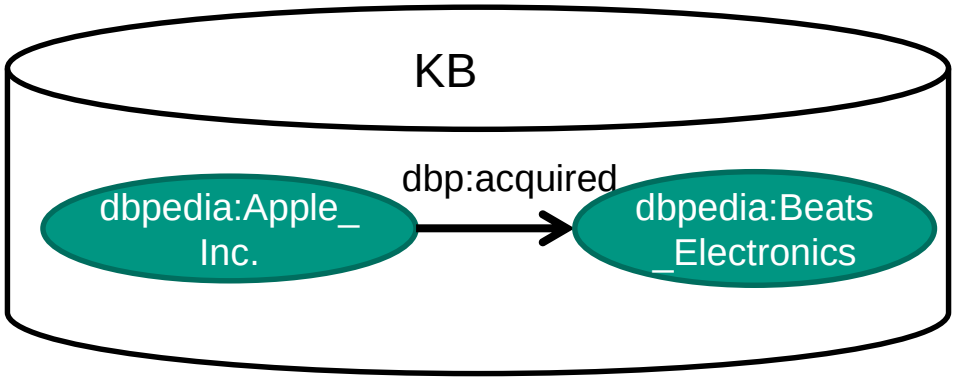
# Our Scenario

*Semantic Novel Fact Extraction System*

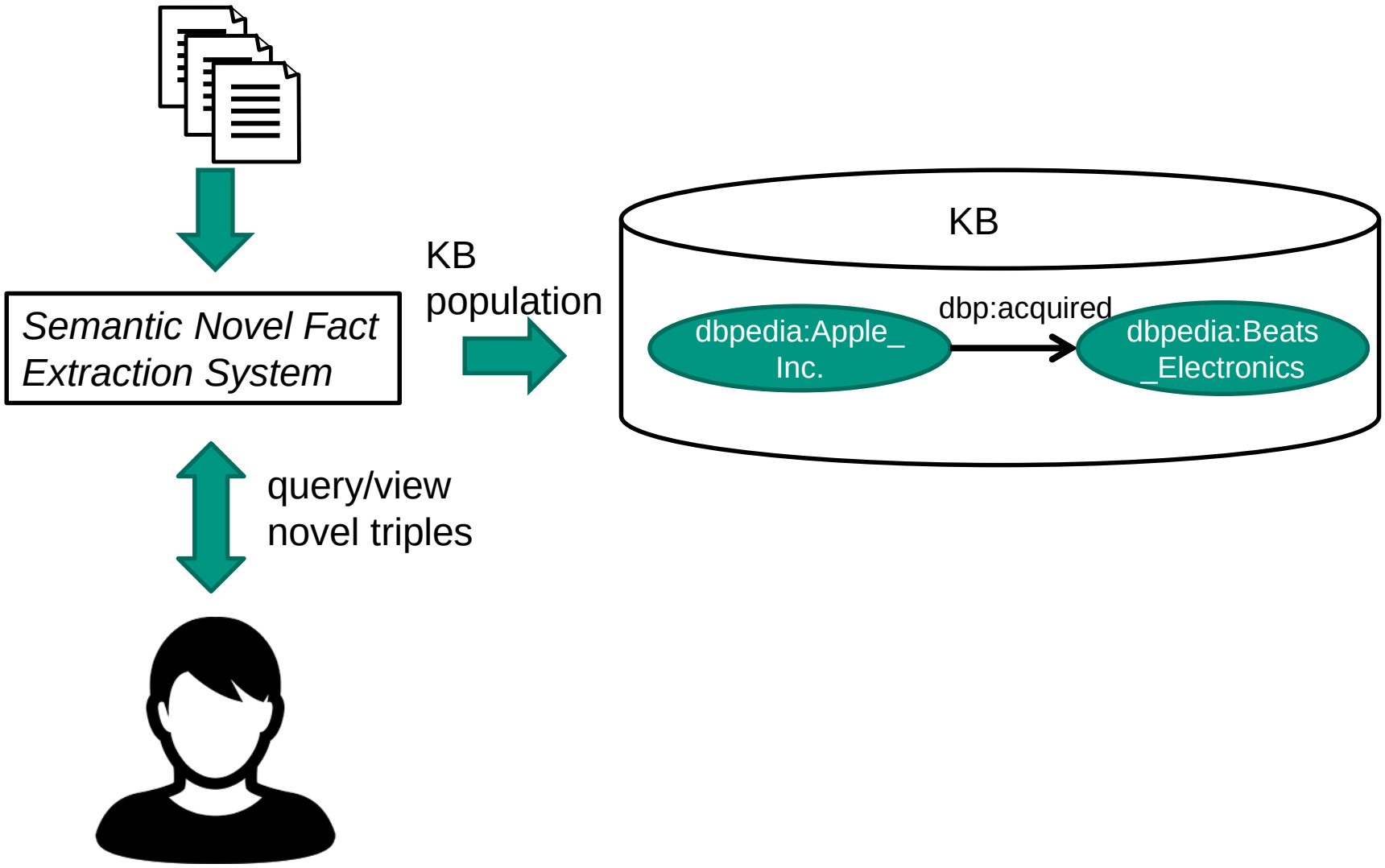


# Our Scenario

*Semantic Novel Fact Extraction System*

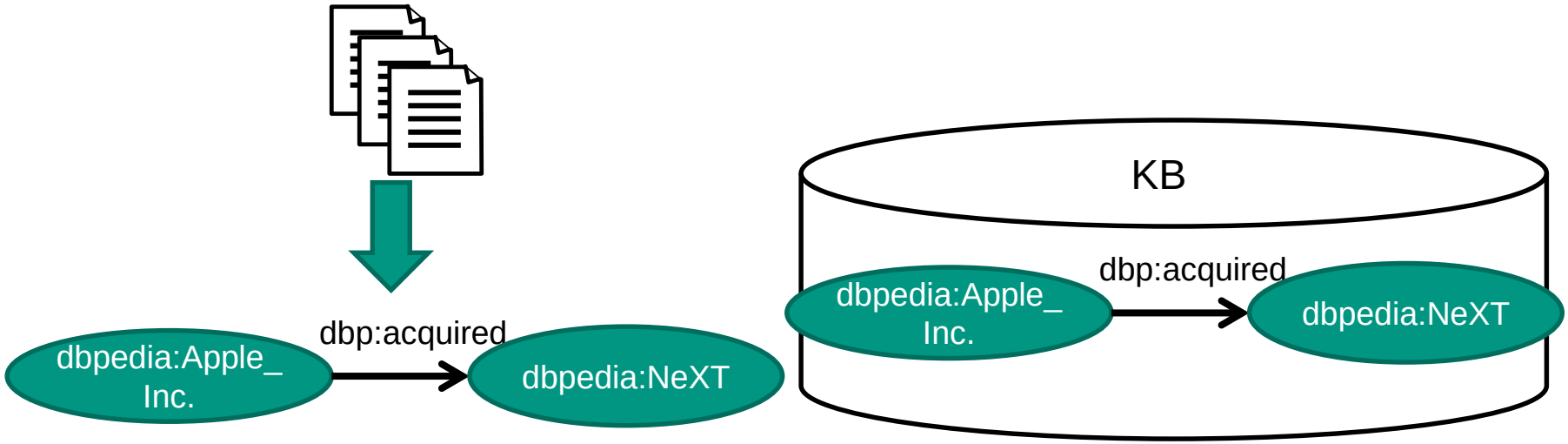


# Our Scenario



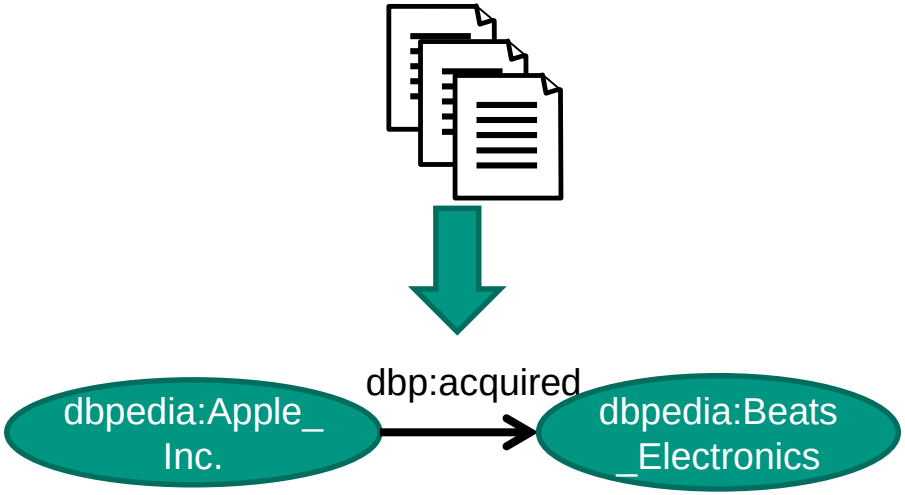
# Formalization of Novel Facts

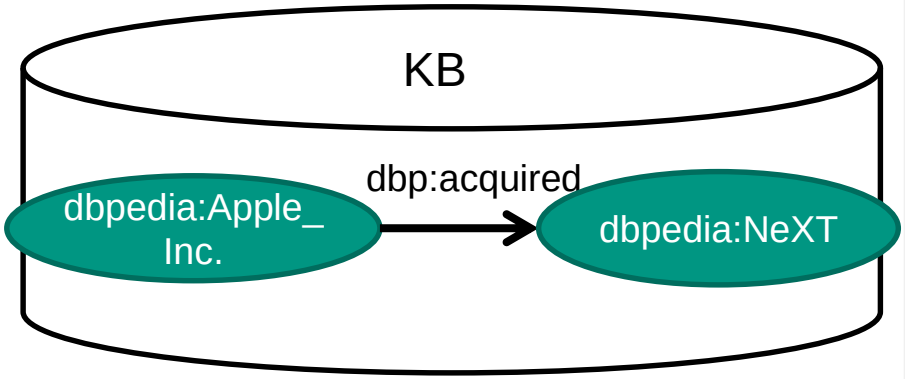
- Novel statements are relevant and novel, if they are not yet in the KB, but if they partially overlap with entries in the KB:
  - All 3 parts of triple in KB => Triple is not novel.



# Formalization of Novel Facts (2)

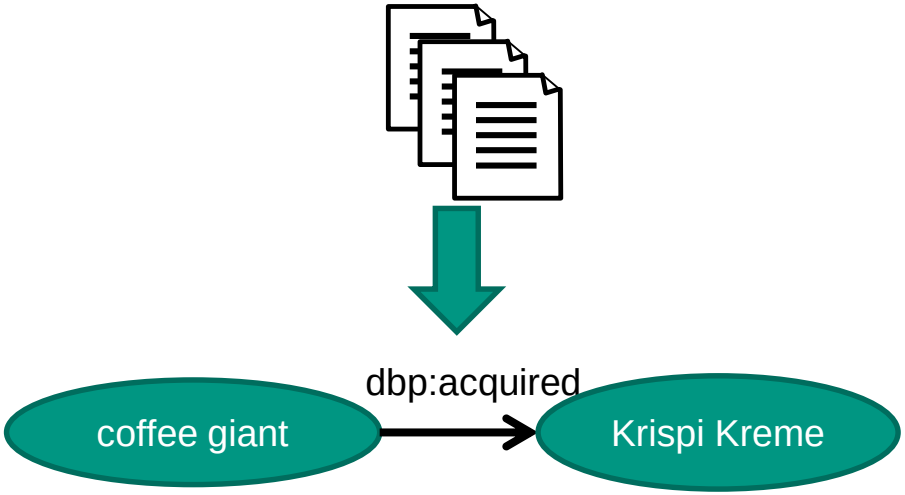
- 2 parts of triple in the KB => Triple is novel and relevant.

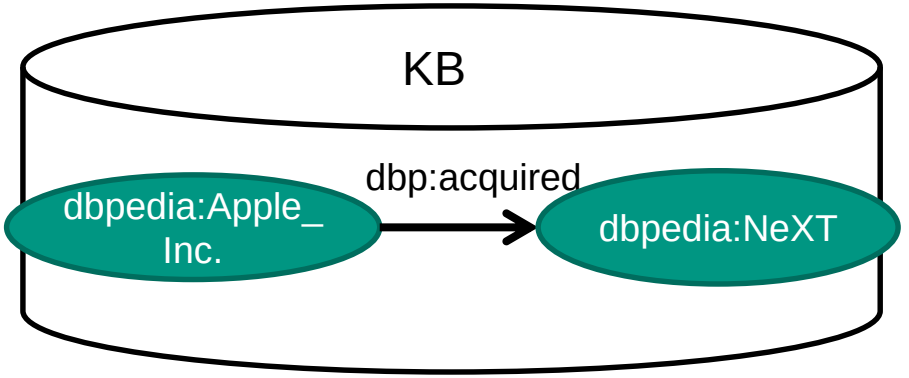


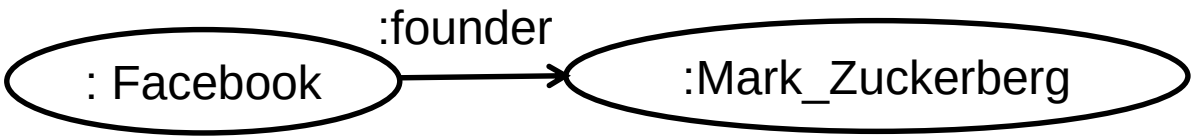
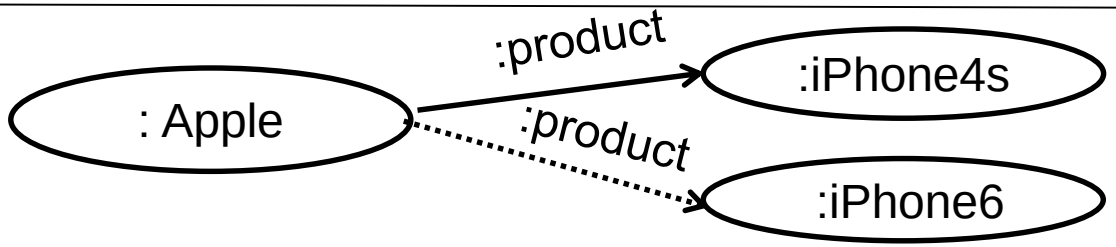
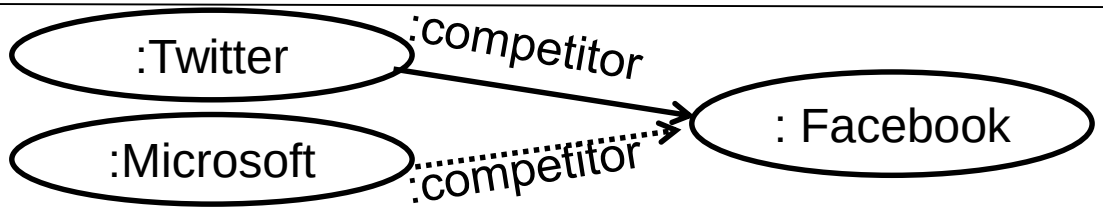
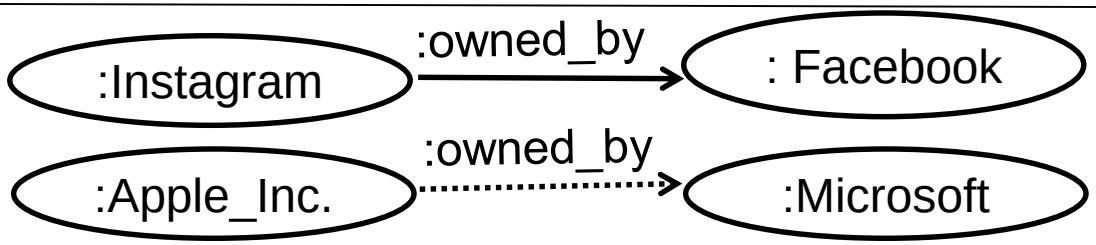
# Formalization of Novel Facts (3)

- 1 or no parts of triple in KB => triple is irrelevant.



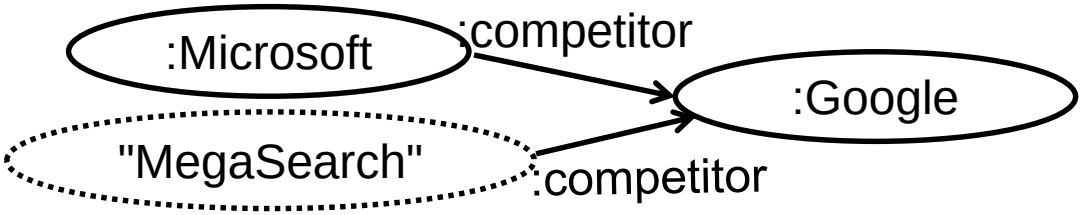



# Classification of Novel Triples

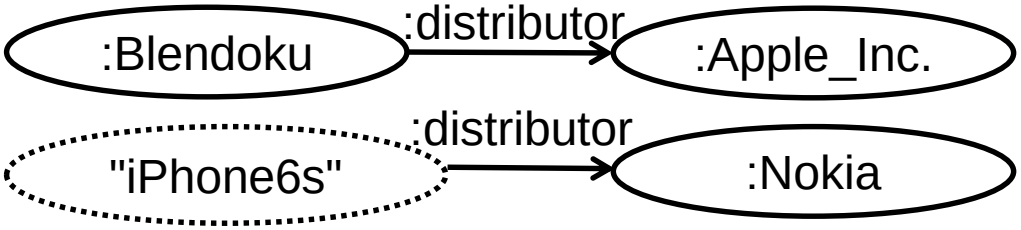
| | | |
|---|---|---|
| A | $(s,p,o) \in KB$ |   |
| $B_{1,i}$ | $(s,p,o) \notin K\ B \wedge \exists x \in U:(s,p,x) \in KB$ |   |
| $B_{1,ii}$ | $(s,p,o) \notin K\ B \wedge \exists x \in U:(x,p,o) \in KB$ |   |
| $B_2$ | $(s,p,o) \notin K\ B \wedge p \neq \emptyset \wedge \neg \exists x \in U:((s,p,x) \in KB \vee (x,p,o) \in KB)$ |   |
| C | $(s,p,o) \notin KB \wedge s \neq \emptyset \wedge o \neq \emptyset \wedge p \neq \emptyset$ |   |

# Classification of Novel Triples (2)

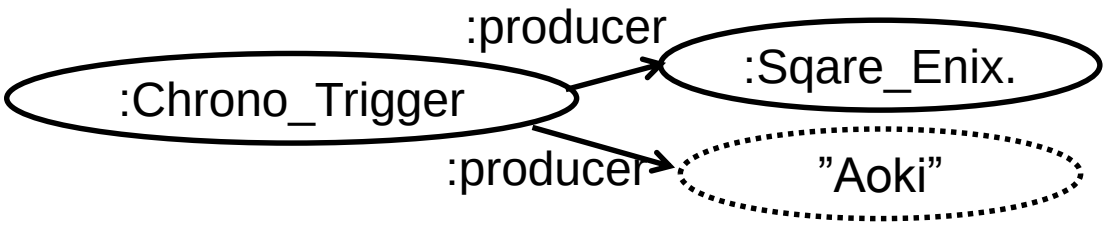
**D1** $(s,p,o) \notin K\ B \wedge \exists x \in U:(x,p,o) \in K\ B \wedge s \neq \emptyset$



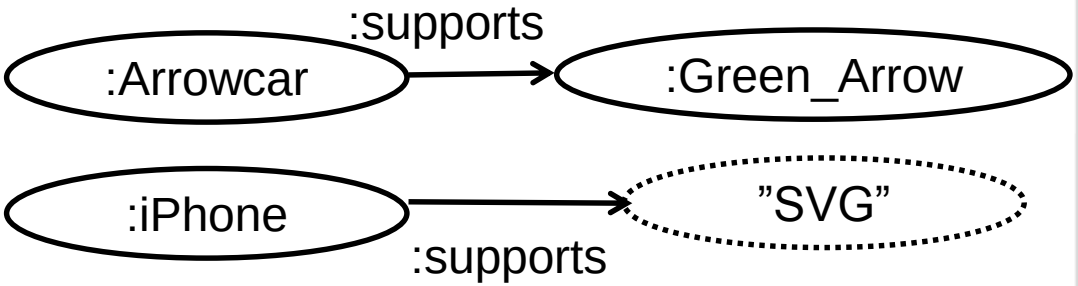

**D2** $(s,p,o) \notin KB \wedge p \neq \emptyset \wedge \neg \exists x \in U:(x,p,o) \in KB \wedge s = \emptyset$



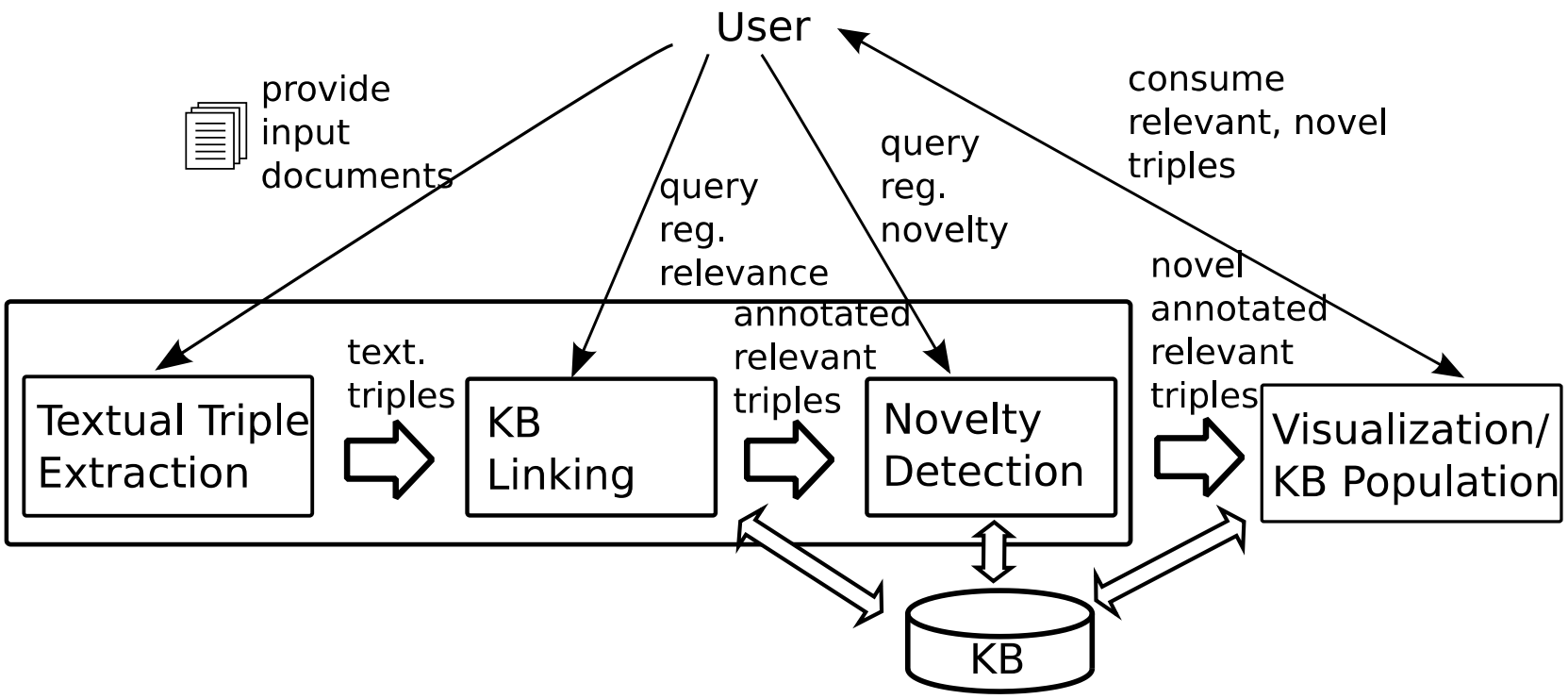

**E1** $(s,p,o) \notin KB \wedge \exists x \in U: (s,p,x) \in KB \wedge o = \emptyset$




**E2** $(s,p,o) \notin KB \wedge p \neq \emptyset \wedge \neg \exists x \in U:(s,p,x) \in KB \wedge o = \emptyset$



# Novel Triple Extraction System

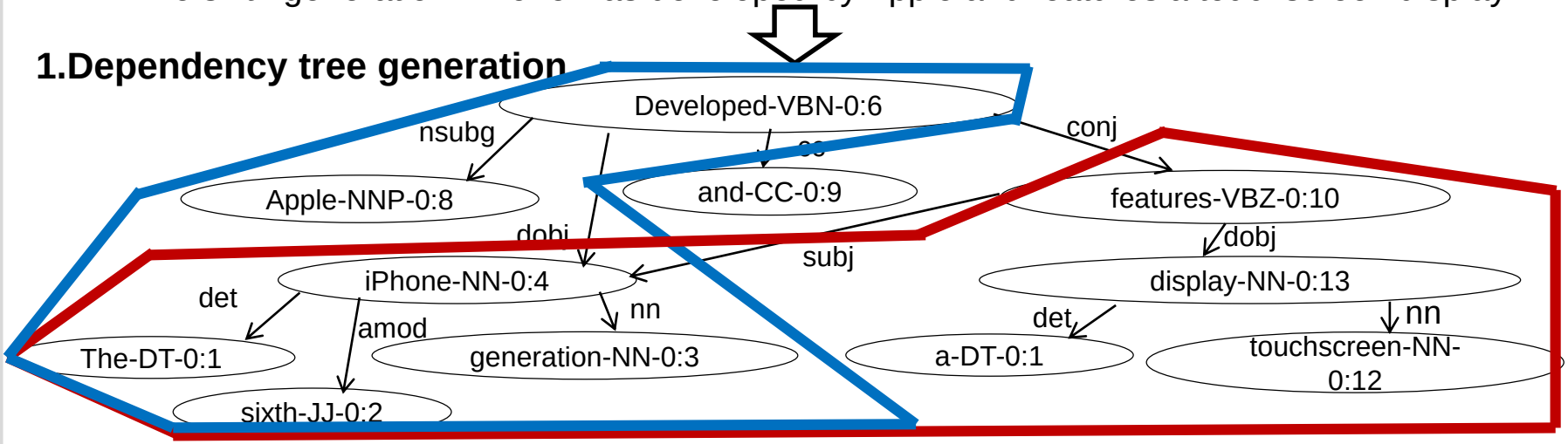


# Steps of the Novel Triple Extraction System (1)

- Step 1: Textual Triple Extraction
  - Based on *ClausIE**

"The sixth generation iPhone was developed by Apple and features a touchscreen display."

**1.Dependency tree generation**




**2.Clause detection and clause type determination**

Clause (i) of type subject-verb-object, passive voice

Clause (ii) of type subject-verb-object, active voice

**3.Textual triple generation**

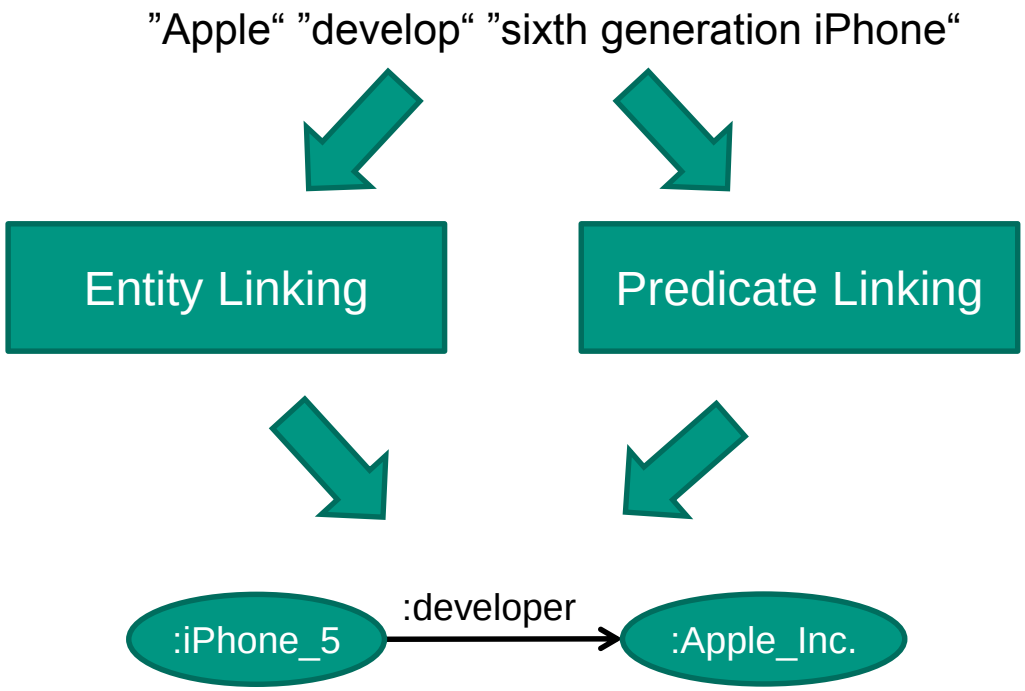
“Apple” “develop” “sixth generation iPhone”

“sixth generation iPhone” “feature” “touchscreen display”

* Luciano Del Corro and Rainer Gemulla. *ClausIE: Clause-Based Open Information Extraction*. WWW’13.

# Steps of the Novel Triple Extraction System (2)

- Step 2: Linking to KB

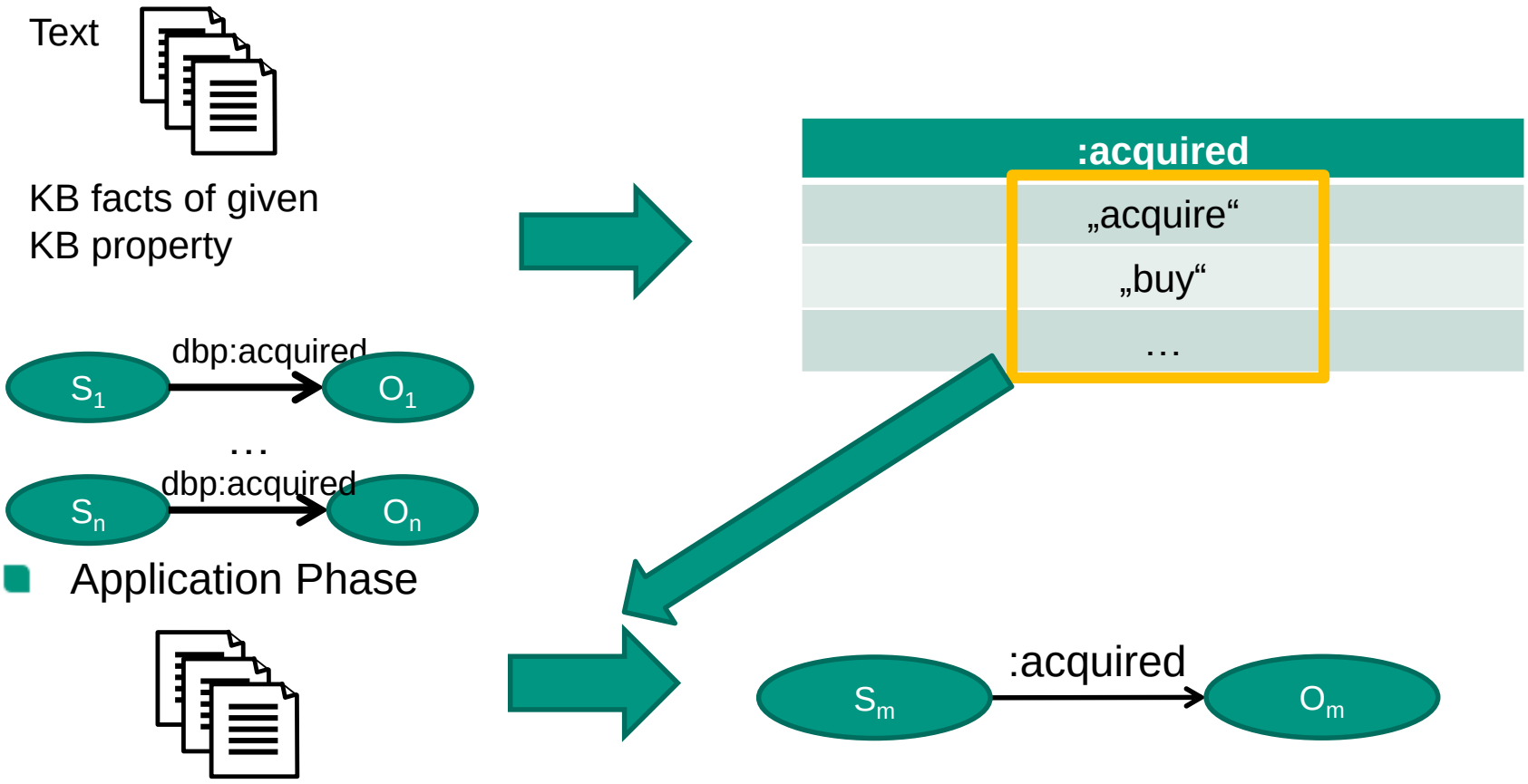



  - Entity Linking based on *x-LiSA**

* Lei Zhang and Achim Rettinger. *X-LiSA: cross-lingual semantic annotation*. VLDB 2014.

# Steps of the Novel Triple Extraction System (3)

- Predicate Linking based on own *distant supervision* method
  - Training Phase




  - Application Phase

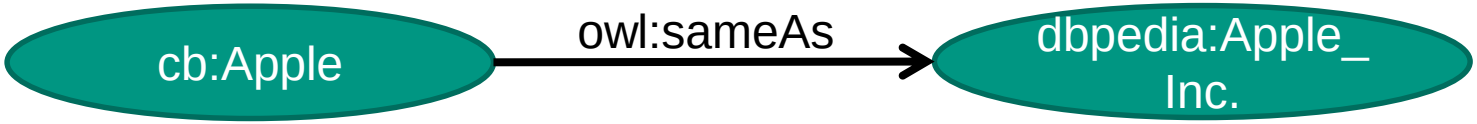
# Steps of the Novel Triple Extraction System (4)

- Step 3: Novelty Detection
  - Classification of novel triples
  - Filtering based on user query (w.r.t. relevance and novelty)

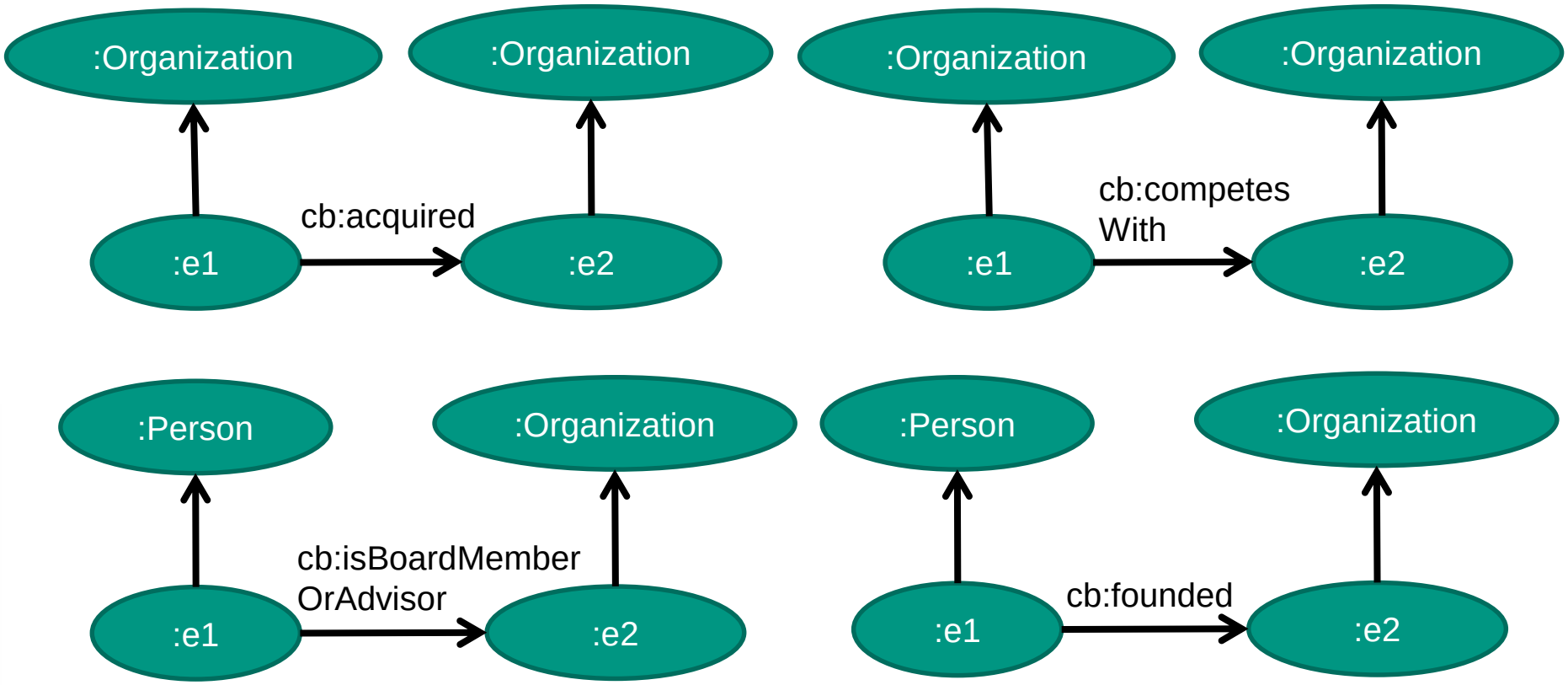
# EVALUATION

# Evaluation Data – KB entities

- CrunchBase in RDF (as of beginning of Oct 2015)
  - 16,706 entities of type *organization*
  - 26,468 entities of type *person*
- `owl:sameAs` links to DBpedia
  - Example:

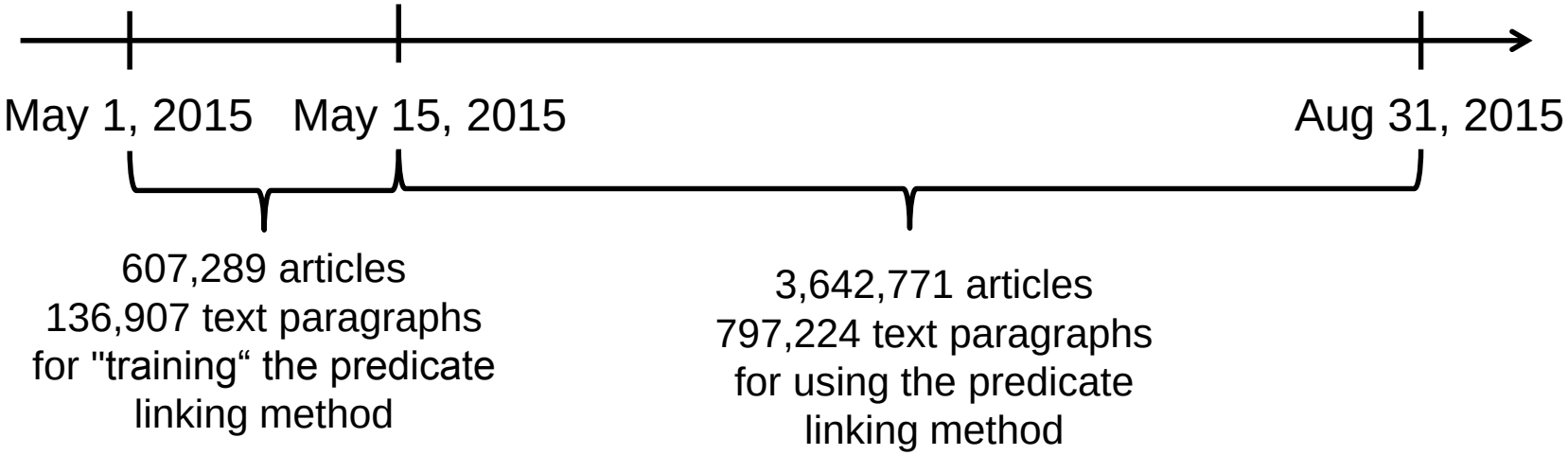


# Evaluation Data – KB properties



# Evaluation Data - Documents

- English news articles from the IJS newsfeed*




* http://newsfeed.ijs.si/

# Tasks to Evaluate

1. **Fact Forecast**: Detect facts before they are officially announced.
2. **Improved KB Population**: Extract new facts in a semantically-structured format, with links to the news articles where mentioned, and faster than if they were manually added to the KB.
3. **Impact Quantification**: Track in-KB facts over time.

Query: "Retrieve all novel facts with the KB properties cb:acquired, cb:competesWith, cb:founded, and cb:isBoardMemberOrAdvisor."

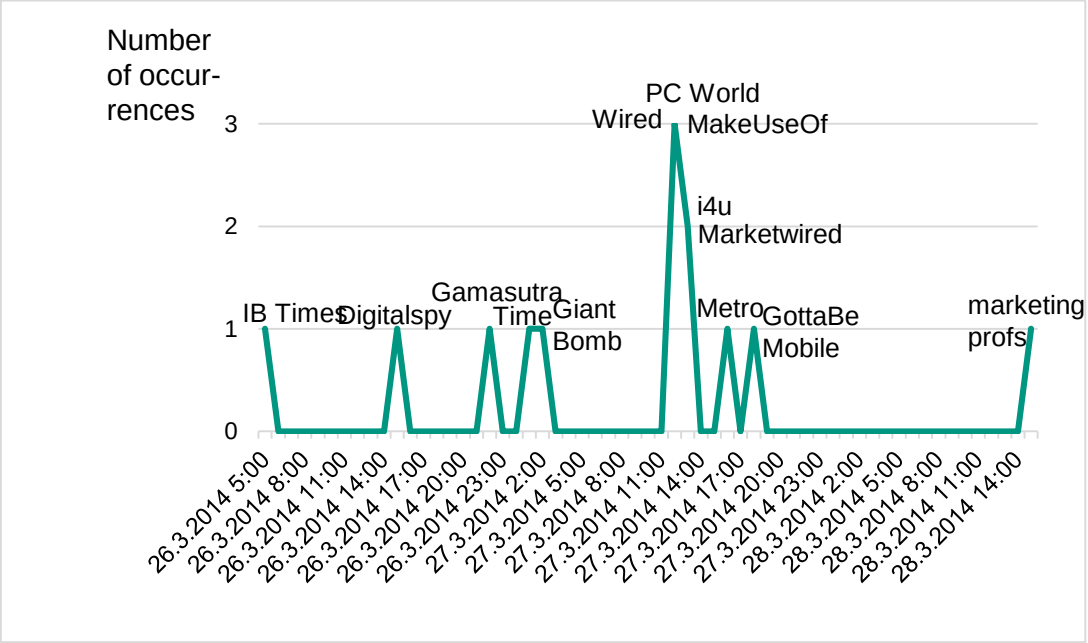
Query: "Extract all novel triples (considering all novelty classes) with the relation cb:acquired."

# Evaluation Results – Part 1

- System extracted 32 different *:acquired* facts (89 in total) which were also in CrunchBase.
- *Lower bound for recall*: *14.3%* for :*acquired* facts.
- Regarding *1. Fact Forecast*:
  - Out of the 32 facts, 2 acquisitions were announced in the time range and detected by the system before inserted into CrunchBase manually.
- Regarding *2. Improved KB Population*:
  - Out of the 32 facts, 4 facts were announced and inserted into CrunchBase in time frame, i.e. those facts are novel considering the CrunchBase KB as of May 15, 2015.

# Evaluation Results – Part 1

- Regarding *3. Impact Quantification*:
  - Most repeated facts:
    :Facebook :acquired :Oculus_VR (18 times),




:Verizon_Communications :acquired :AOL (7 times),
:Apple_Inc. :acquired :Beats_Electronics (7 times)

# Evaluation Results – Part 2

| KB property | A | $B_{1,i}$ | $B_{1,ii}$ | $B_2$ | $D_1$ | $D_2$ | $E_1$ | $E_2$ |
|---|---|---|---|---|---|---|---|---|
| :acquired | 89/89 | 33/36 | 4/4 | 13/13 | 24/86 | 14/71 | 63/636 | 48/398 |
| :competesWith | 7/8 | 35/35 | 11/13 | 19/20 | 72 | 57 | 171 | 240 |
| :founded | 33/33 | 0/0 | 1/2 | 3/3 | 63 | 73 | 145 | 311 |
| :isBoardMemberOrAdvisor | 1/2 | 7/9 | 18/18 | 7/8 | 58 | 70 | 70 | 450 |

- Precision for $B_x$ novel triples 95.9%,
- Precision for other ($D_x$, $E_x$) class triples: 12.5%
- However, wrong triples mostly almost correct
- Issues:
  a) Ling. proposition results in correct and additional wrong triples
  b) Coreference resolution (15%)
  c) Fact not representable as triple (31 %)

# Conclusion

- Task: Targeted search for novel, grounded facts in unstructured text
- Approach: Measure novelty w.r.t. a KB and semantic novelty classes. Our prototypical system can facilitate
  i. fact forecast
  ii. improved KB population
  iii. impact quantification.
- Future work: Improve recall via
  i. improving the Textual Triple Extraction step (e.g., extract also nominal phrases)
  ii. implementing co-reference resolution.

Thank you for your attention.

Any questions?

**Michael Färber**
michael.faerber@kit.edu
**Institute AIFB, Karlsruhe Institute of Technology (KIT), Germany**

| | |
|---|---|
| Presutti et al. 2012 | Presutti, V., Draicchio, F., Gangemi, A.: Knowledge Extraction Based on Discourse Representation Theory and Linguistic Frames. In: Knowledge Engineering and Knowledge Management. Springer Berlin Heidelberg (2012) 114–129 |
| Carvhalho et al. 2013 | Carvalho, D.S., Freitas, A., Silva, J.: Graphia: Extracting Contextual Relation Graphs from Text. In: The Semantic Web: ESWC 2013 Satellite Events. Springer Berlin Heidelberg (2013) 236–241 |
| Augenstein et al. 2012 | Augenstein, I., Pado, S., Rudolph, S.: LODifier: Generating Linked Data from Unstructured Text. In Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V., eds.: The Semantic Web: Research and Applications. Volume 7295 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 210–224 |
| Fader et al. 2011 | Fader, A., Soderland, S., Etzioni, O.: Identifying Relations for Open Information Extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP ’11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 1535–1545 |
| Del Corro et al. 2013 | Del Corro, L., Gemulla, R.: ClausIE: Clause-Based Open Information Extraction. In: Proceedings of the 22nd international conference on World Wide Web. WWW ’13, Republic and Canton of Geneva, Switzerland, ACM (2013) 355–366 |

| | |
|---|---|
| Mausam et al. 2012 | Mausam, Schmitz, M., Bart, R., Soderland, S., Etzioni, O.: Open Language Learning for Information Extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning. EMNLP-CoNLL '12, Stroudsburg, PA, USA, ACL (2012) 523–534 |
| Zhang et al. 2012 | Zhang, Y., Callan, J., Minka, T.: Novelty and redundancy detection in adaptive filtering. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '02, New York, NY, USA, ACM (2002) 81–88 |
| Gabrilovich et al. 2004 | Gabrilovich, E., Dumais, S., Horvitz, E.: Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In: Proceedings of the 13th international conference on World Wide Web. WWW '04, New York, NY, USA, ACM (2004) 482–490 |
| Karkali et al. 2013 | Karkali, M., Rousseau, F., Ntoulas, A., Vazirgiannis, M.: Efficient Online Novelty Detection in News Streams. In Lin, X., et al., eds.: Web Information Systems Engineering – WISE 2013. Springer Berlin Heidelberg (2013) 57–71 |
| Li et al. 2008 | Li, X., Croft, W.B.: An information-pattern-based approach to novelty detection. Information Processing & Management 44(3) (2008) 1159–1188 |

# ”Rumour facts”

- http://dbpedia.org/resource/Aetna cb:acquired http://dbpedia.org/resource/Humana
- http://dbpedia.org/resource/Intel cb:acquired http://dbpedia.org/resource/Altera