

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

Two Variations on Ontology Alignment Evaluation: Methodological Issues.

**Laura Hollink, Mark van Assem, Antoine Isaac, Shenghui Wang,
Guus Schreiber.**

Vrije Universiteit Amsterdam, The Netherlands
ESWC 2008

5th June 2008

Ontology Alignments

Context

Ontology Alignments

Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria

WordNet

Case 3: SVCN

Results and
Conclusions

Evaluation methods

Alignments

- Growing number of different and heterogeneous ontologies on the Semantic Web.
- Need to interconnect these ontologies
- Alignment between ontologies: set of mappings between concepts.
- State-of-the-art of ontology alignment tools:
 - partial alignments
 - varying quality

Evaluation of Alignments

Context

Ontology Alignments

Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria

WordNet

Case 3: SVCN

Results and
Conclusions

Evaluation methods

Alignments

Three common methods:

- Judging individual mappings (precision).
- Comparison to a reference alignment (precision and recall).
- End-to-end evaluation (measure determined by end-application).

Two Variations on Ontology Alignment Evaluation

Context

Ontology Alignments

Evaluation of Alignments

Frequency-evaluation

Semantic-distance-evaluation

Applying the evaluation methods

Aria

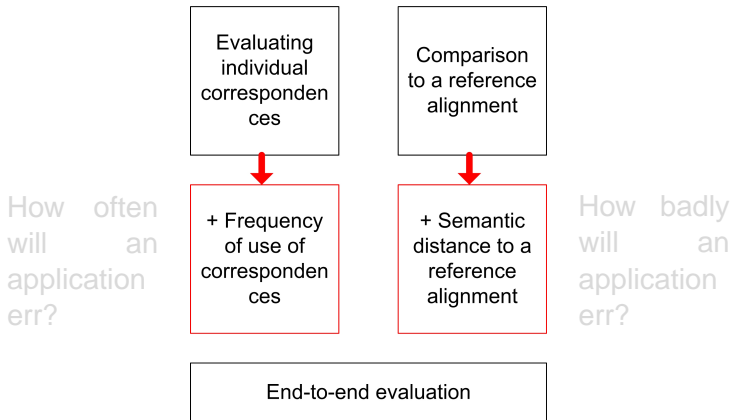
WordNet

Case 3: SVCN

Results and Conclusions

Evaluation methods

Alignments



Two Variations on Ontology Alignment Evaluation

Context

Ontology Alignments

Evaluation of Alignments

Frequency-evaluation

Semantic-distance-evaluation

Applying the evaluation methods

Aria

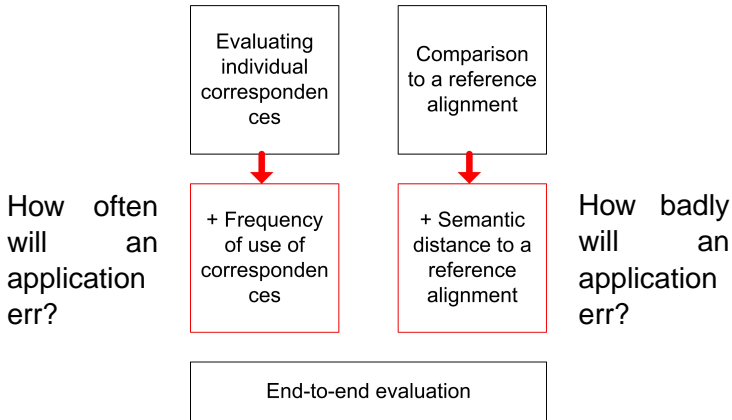
WordNet

Case 3: SVCN

Results and Conclusions

Evaluation methods

Alignments



Frequency evaluation: why

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

- Some mappings affect the quality of an end-application more than others.
- An end-to-end evaluation can take this into account but an evaluation of individual mappings cannot.
- Evaluating the most important mappings better approximates the outcome of an end-to-end evaluation.
- How to measure ‘Importance’?
- Proposal: use the estimated frequency of use of each mapping as a weighting factor in the computation of performance measures.

Frequency evaluation: why

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

- Some mappings affect the quality of an end-application more than others.
- An end-to-end evaluation can take this into account but an evaluation of individual mappings cannot.
- Evaluating the most important mappings better approximates the outcome of an end-to-end evaluation.
- How to measure ‘Importance’?
- Proposal: use the estimated frequency of use of each mapping as a weighting factor in the computation of performance measures.

Frequency evaluation: why

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

- Some mappings affect the quality of an end-application more than others.
- An end-to-end evaluation can take this into account but an evaluation of individual mappings cannot.
- Evaluating the most important mappings better approximates the outcome of an end-to-end evaluation.
- How to measure ‘Importance’?
- Proposal: use the estimated frequency of use of each mapping as a weighting factor in the computation of performance measures.

How to estimate frequency of use of a mapping?

- Depends on the application!
 - In our case: a retrieval application
- Usage data of the application?
- Ask an expert?
- Naive case: we assume that in our retrieval system every concept has an equal probability of being chosen as query.

Frequency evaluation: scenario of use

- A user has a query in one ontology but the data is annotated with another ontology.
- Reformulation from one ontology to another.
- Obvious case: if the query concept is mapped to the annotation ontology.
- What if the query concept is not mapped to the annotation ontology?
- Take the closest concept in the query ontology that does have a mapping.

Frequency evaluation: scenario of use

- A user has a query in one ontology but the data is annotated with another ontology.
- Reformulation from one ontology to another.
- Obvious case: if the query concept is mapped to the annotation ontology.
- What if the query concept is not mapped to the annotation ontology?
- Take the closest concept in the query ontology that does have a mapping.

Frequency evaluation: use scenario

Context

Ontology Alignments
Evaluation of Alignments

Frequency-evaluation

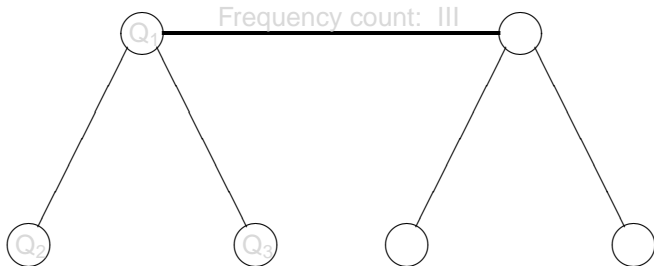
Semantic-distance-evaluation

Applying the evaluation methods

Aria
WordNet
Case 3: SVCN

Results and Conclusions

Evaluation methods
Alignments



Some mappings are used by many queries.

Frequency evaluation: use scenario

Context

Ontology Alignments
Evaluation of Alignments

Frequency-evaluation

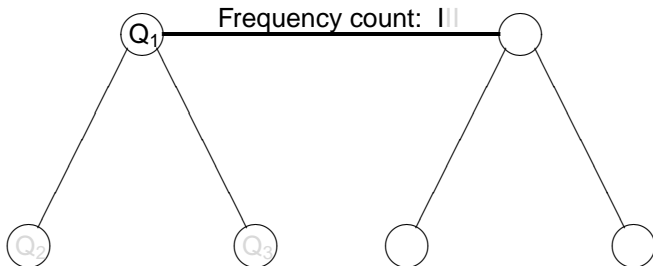
Semantic-distance-evaluation

Applying the evaluation methods

Aria
WordNet
Case 3: SVCN

Results and Conclusions

Evaluation methods
Alignments



Some mappings are used by many queries.

Frequency evaluation: use scenario

Context

Ontology Alignments
Evaluation of Alignments

Frequency-evaluation

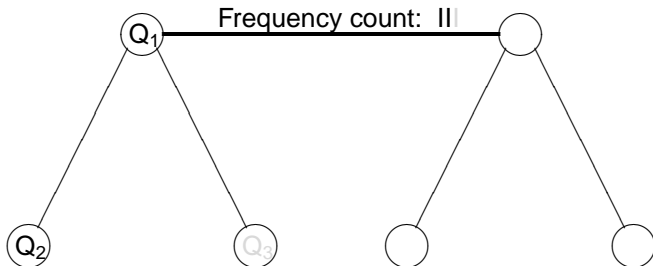
Semantic-distance-evaluation

Applying the evaluation methods

Aria
WordNet
Case 3: SVCN

Results and Conclusions

Evaluation methods
Alignments



Some mappings are used by many queries.

Frequency evaluation: use scenario

Context

Ontology Alignments
Evaluation of Alignments

Frequency-evaluation

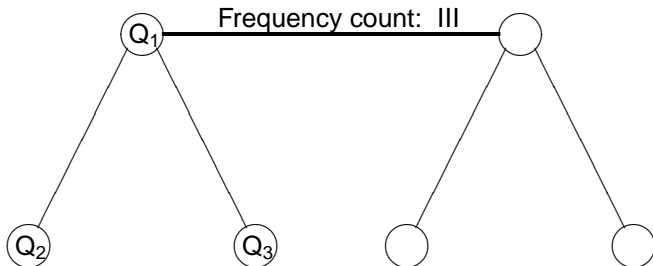
Semantic-distance-evaluation

Applying the evaluation methods

Aria
WordNet
Case 3: SVCN

Results and Conclusions

Evaluation methods
Alignments



Some mappings are used by many queries.

Frequency evaluation: how

Context

Ontology Alignments
Evaluation of Alignments

Frequency-evaluation

Semantic-distance-evaluation

Applying the evaluation methods

Aria
WordNet
Case 3: SVCN

Results and Conclusions

Evaluation methods
Alignments

- 1 Divide all mappings into two strata: infrequently and frequently used mappings (simplification).
- 2 Determine quality of each stratum.
- 3 Weigh the strata according to their expected frequency of use.

$$\hat{P} = \sum_{h=1}^L \frac{\sum_{a \in H} \text{freq}(a)}{\sum_{a \in A} \text{freq}(a)} \hat{P}_h \quad (1)$$

$\text{freq}(a)$ is the frequency of use of mapping a , H is the total set of mappings in stratum h , and A is the total set of mappings in the alignment.

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

Frequency evaluation: benefits

- If there is a difference in quality between strata → frequency-weighted precision will be more realistic.
- In a semi-automatic matching process, it makes sense to check and correct frequent mappings first.

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluationSemantic-
distance-
evaluationApplying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

Semantic-distance-evaluation: why

- Comparing an alignment A to a reference alignment R gives precision as well as recall scores.
- Incorrect mappings negatively affect the performance of an application.
- However, this effect varies depending on how incorrect the mapping is.
- Linking two completely unrelated concepts is more harmful than linking two closely related concepts.
- We use an semantic distance measure to capture this difference.
- Semantic distance to represent the distance between a mapping in A and a mapping in a reference alignment R .

Semantic-distance-evaluation: why

Context

Ontology Alignments
Evaluation of Alignments

Frequency- evaluation

Semantic- distance- evaluation

Applying the evaluation methods

Aria
WordNet
Case 3: SVCN

Results and Conclusions

Evaluation methods
Alignments

- Comparing an alignment A to a reference alignment R gives precision as well as recall scores.
- Incorrect mappings negatively affect the performance of an application.
- However, this effect varies depending on how incorrect the mapping is.
- Linking two completely unrelated concepts is more harmful than linking two closely related concepts.
- We use an semantic distance measure to capture this difference.
- Semantic distance to represent the distance between a mapping in A and a mapping in a reference alignment R .

Semantic-distance-evaluation: why

Context

Ontology Alignments
Evaluation of Alignments

Frequency- evaluation

Semantic- distance- evaluation

Applying the evaluation methods

Aria
WordNet
Case 3: SVCN

Results and Conclusions

Evaluation methods
Alignments

- Comparing an alignment A to a reference alignment R gives precision as well as recall scores.
- Incorrect mappings negatively affect the performance of an application.
- However, this effect varies depending on how incorrect the mapping is.
- Linking two completely unrelated concepts is more harmful than linking two closely related concepts.
- We use an semantic distance measure to capture this difference.
- Semantic distance to represent the distance between a mapping in A and a mapping in a reference alignment R .

Semantic-distance evaluation: how

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluationSemantic-
distance-
evaluationApplying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

Existing semantic distance measure of Leacock and Chodorow:

$$sim_{LC} = -\log \frac{len_{(c_1, c_2)}}{2D}$$

where $len_{(c_1, c_2)}$ is the shortest path between concepts c_1 and c_2 , D is the maximum depth of the hierarchy.

To calculate precision and recall, we normalize the semantic distance to a 0–1 scale.

Applying the evaluation methods

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

Goal 1: Insight in methodological issues.

- How do methods work in practice?
- Compare alternative methods to common methods.
- Are conclusions different?

Goal 2: Insight into the effect of the characteristics of the ontologies on the quality of the alignment, and on the best evaluation method to choose.

- Compare three very different ontologies.
- Is one alignment better than another and why?
- Which evaluation method shows these differences?

Applying the evaluation methods

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

Goal 1: Insight in methodological issues.

- How do methods work in practice?
- Compare alternative methods to common methods.
- Are conclusions different?

Goal 2: Insight into the effect of the characteristics of the ontologies on the quality of the alignment, and on the best evaluation method to choose.

- Compare three very different ontologies.
- Is one alignment better than another and why?
- Which evaluation method shows these differences?

Three alignments

ARIA-AAT, WordNet-AAT, SVCN-AAT, all made with Falcon-AO

AAT:

- Used by museums around the world for indexing works of art.
- Size: 16,436 concepts in the 'Object facet'.
- The broader/narrower hierarchy of this facet is ontologically clean.
- maximum hierarchy depth is 17.

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria

WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

Case 1: ARIA

- Developed by the Rijksmuseum for their website
- It contains 491 concepts which are all art-related object types.
- Its hierarchy is at most 3 concepts deep and is arranged in a polyhierarchy
- ARIA is the smallest and most weakly structured of the three source vocabularies.

AAT-ARIA frequency of use of mappings

Context

Ontology Alignments
Evaluation of Alignments

Frequency-evaluation

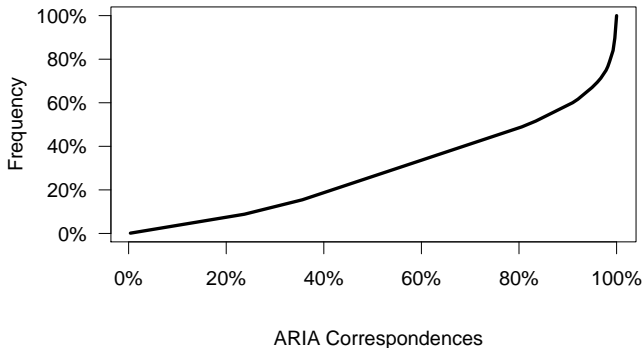
Semantic-distance-evaluation

Applying the evaluation methods

Aria
WordNet
Case 3: SVCN

Results and Conclusions

Evaluation methods
Alignments



The total number of mappings is 278.

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluationSemantic-
distance-
evaluationApplying the
evaluation
methods

Aria

WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

AAT-ARIA results

Evaluation Type	Precision	Recall
Judging of individual mappings	0.74 ± 0.03	
Variation: Weighted by frequency of use	0.70 ± 0.03	
Comparison to a reference alignm.(RA)	0.66 ± 0.09	0.63 ± 0.09
Variation: semantic distance to an RA	0.80 ± 0.08	0.76 ± 0.08

Taking into account semantic distance does change the results.

Evaluation Type	Precision
After random correction	0.74 ± 0.03
After correction of frequent stratum	0.85 ± 0.02

Correction of the most frequent stratum gives a higher precision than random correction.

AAT-ARIA results

Evaluation Type	Precision	Recall
Judging of individual mappings	0.74 ± 0.03	
Variation: Weighted by frequency of use	0.70 ± 0.03	
Comparison to a reference alignm.(RA)	0.66 ± 0.09	0.63 ± 0.09
Variation: semantic distance to an RA	0.80 ± 0.08	0.76 ± 0.08

Taking into account semantic distance does change the results.

Evaluation Type	Precision
After random correction	0.74 ± 0.03
After correction of frequent stratum	0.85 ± 0.02

Correction of the most frequent stratum gives a higher precision than random correction.

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria

WordNet

Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

Case 2: Wordnet

- WordNet is a freely available thesaurus of the English language developed at Princeton.
- Size: 31,547 concepts under 'Object'.
- The main hierarchy is formed by the polyhierarchic hyponym relation which contains more ontological errors than AAT's hierarchy.
- The topical overlap with AAT is reasonable.

AAT-WordNet frequency of use of mappings

Context

- Ontology Alignments
- Evaluation of Alignments

Frequency-evaluation

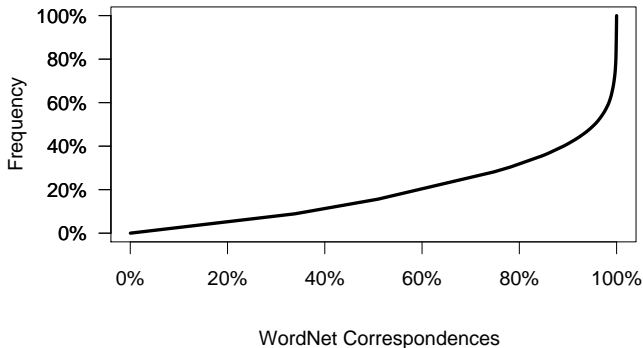
Semantic-distance-evaluation

Applying the evaluation methods

- Aria
- WordNet
- Case 3: SVCN

Results and Conclusions

- Evaluation methods
- Alignments



The total number of mappings is 4101.

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluationSemantic-
distance-
evaluationApplying the
evaluation
methods

Aria

WordNet

Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

AAT-WordNet results

Evaluation Type	Precision	Recall
Judging of individual mappings	0.71 ± 0.05	
Variation: Weighted by frequency of use	0.68 ± 0.04	
Comparison to a reference alignm.(RA)	0.62 ± 0.10	0.45 ± 0.10
Variation: semantic distance to an RA	0.64 ± 0.09	0.47 ± 0.10

Evaluation Type	Precision
After random correction	0.72 ± 0.04
After correction of frequent stratum	0.81 ± 0.04

Correcting the most frequent mappings gives better results than correcting random mappings.

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluationSemantic-
distance-
evaluationApplying the
evaluation
methods

Aria

WordNet

Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

AAT-WordNet results

Evaluation Type	Precision	Recall
Judging of individual mappings	0.71 ± 0.05	
Variation: Weighted by frequency of use	0.68 ± 0.04	
Comparison to a reference alignm.(RA)	0.62 ± 0.10	0.45 ± 0.10
Variation: semantic distance to an RA	0.64 ± 0.09	0.47 ± 0.10

Evaluation Type	Precision
After random correction	0.72 ± 0.04
After correction of frequent stratum	0.81 ± 0.04

Correcting the most frequent mappings gives better results than correcting random mappings.

Context

Ontology Alignments

Evaluation of Alignments

Frequency- evaluation

Semantic- distance- evaluation

Applying the evaluation methods

Aria

WordNet

Case 3: SVCN

Results and Conclusions

Evaluation methods

Alignments

- SVCN is a thesaurus developed and used by several Dutch ethnographic museums.
- the Object facet has 4,200 concepts.
- Originally derived from AAT, but concepts added and removed over time.
- Maximum hierarchy depth is 13.
- The broader/narrower hierarchy is well-designed, but contains more errors than AAT's.

AAT-SVCN frequency of use of mappings

Context

Ontology Alignments
Evaluation of Alignments

Frequency-evaluation

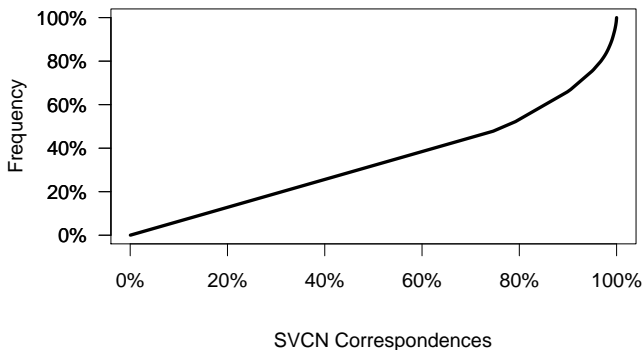
Semantic-distance-evaluation

Applying the evaluation methods

Aria
WordNet
Case 3: SVCN

Results and Conclusions

Evaluation methods
Alignments



The total number of mappings is 2748.

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluationSemantic-
distance-
evaluationApplying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

AAT-SVCN results

Evaluation Type	Precision	Recall
Judging of individual mappings	0.89±0.03	
Variation: Weighted by frequency of use	0.89±0.03	
Comparison to a reference alignm.(RA)	0.84±0.07	0.80±0.08
Variation: semantic distance to an RA	0.90±0.06	0.86±0.07

Evaluation Type	Precision
After random correction	0.93±0.03
After correction of frequent stratum	0.91±0.03

No significant differences between the evaluation methods!

End-to-end evaluation

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluationSemantic-
distance-
evaluationApplying the
evaluation
methods

Aria
WordNet

Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

- Reformulation of query from one vocabulary to another.
- How many correct results are returned for the query?
- Results rated on a binary (yes/no) and 6-point scale.

Vocabulary	Precision		Recall	
	Binary	6-point	Binary	6-point
ARIA	0.27	0.37	0.83	0.88
SVCN	0.46	0.48	0.93	0.96
WordNet	0.46	0.48	0.80	0.81

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

Results and Conclusions: frequency-evaluation

- + If there are relatively few mappings in an alignment (WordNet, ARIA), some will be more central than others. In that case, a frequency based evaluation will be different from a normal evaluation.
- * *Frequency based evaluation will give a realistic estimation of application performance for sparse alignments.*

Results and Conclusions: semantic-distance evaluation

- + Semantic-distance-evaluation behaves the same as 6-point scale ratings in an end-to-end evaluation.
- * *Semantic-distance-evaluation is a good alternative for end-to-end evaluations using a rating scale instead of dichotomous ratings.*
 - *e.g. in applications in which users expect to see also moderately relevant results*

Results and Conclusions: alignments

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluationSemantic-
distance-
evaluationApplying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

- + Alignment of SVCN to AAT scores very high regardless of evaluation method. Manual correction of mappings does not add significantly to the results.
- * *Alignment of vocabularies with a reasonably clean hierarchy and high similarity to the target vocabulary is so good that there is no need for manual correction.*
- + Precision is similar for WordNet and ARIA: around 0.70.
- * *A weakly structured, small vocabulary can be aligned with approximately the same precision as a large, richly structured vocabulary.*

Results and Conclusions: alignments

Context

Ontology Alignments
Evaluation of Alignments

Frequency-
evaluation

Semantic-
distance-
evaluation

Applying the
evaluation
methods

Aria
WordNet
Case 3: SVCN

Results and
Conclusions

Evaluation methods
Alignments

- + For WordNet and ARIA, manual correction of frequent mappings improves the results significantly, while correction of random mappings does not.
- * *If frequency curve is steep, it pays off to first determine the most frequent mappings before manually correcting mappings.*

Outline

Context

- Ontology Alignments
- Evaluation of Alignments

Frequency-evaluation

Semantic-distance-evaluation

Applying the evaluation methods

- Aria
- WordNet
- Case 3: SVCN

Results and Conclusions

- Evaluation methods

Alignments

Questions?

Spare slide: Semantic-distance evaluation: how II

- assessments are no longer dichotomous but are measured on an interval level.
- Common recall and precision measures are not suited for this scale.
- Therefore, we use *Generalised Precision* and *Generalized Recall* as proposed by Kekalainen et al.:

$$gP = \sum_{a \in A} \frac{r(a)}{|A|} \qquad gR = \frac{\sum_{a \in A} r(a)}{\sum_{a \in R} r(a)} \quad (2)$$