MPRI Internship Defense
Advances in Holistic Ontology Alignment

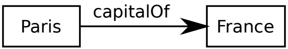
Antoine Amarilli
Supervised by Pierre Senellart

Télécom ParisTech

The Semantic Web

```
<p><b>Paris</b> is the <a href="Capital_city">
capital</a> of <a href="France">France</a></p>
```

Facts on the Web



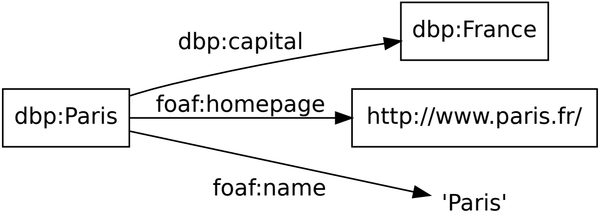
Facts on the semantic Web

The Web. Lots of information in semi-structured HTML documents.

The semantic Web. An effort to represent information in a *structured* and *semantic* way.

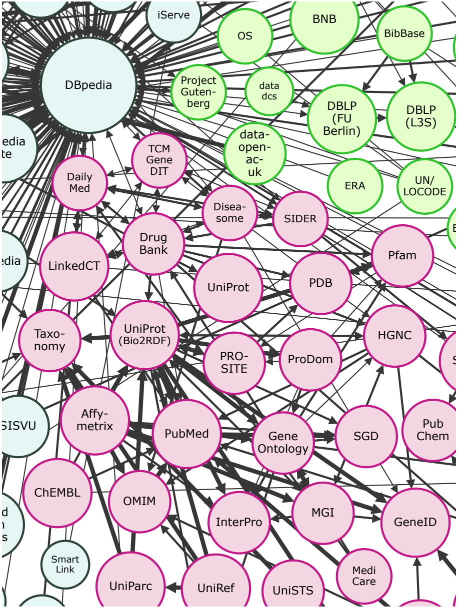
Uses. Interoperability, integration of sources, constraints, complex queries, inference.

Ontologies



- Ontologies are the information sources of the Semantic Web.
- Vertices are entities or literals.
- Edges are facts labeled with a relation.
- Sources : manual creation, existing databases, information extraction.

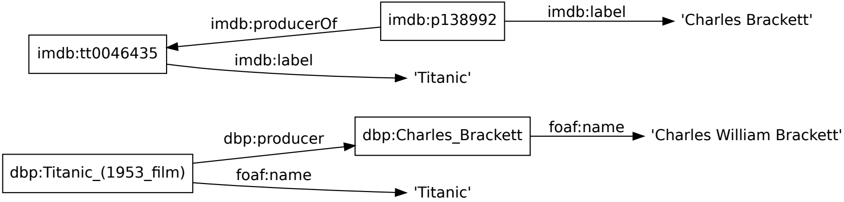
Linked Data Cloud



- Many ontologies are created independently: different entities and relations express the same things.
- Linked Data: integrate existing ontologies in a network structured by equality links between equivalent concepts.
- To automatically derive those links, we need to perform ontology alignment.

Ontology Alignment

Sometimes URIs do not help us and literals are ambiguous or have minor differences...



Sometimes the structures of the two ontologies do not match...

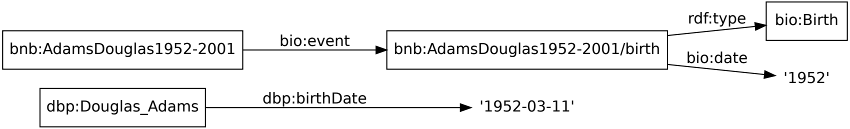
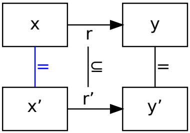
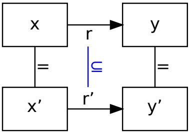


Table of Contents

Paris

- Paris: Probabilistic Alignment of Relations, Instances, and Schema.
- To bootstrap a matching, Paris uses an equality function on literals and applies propagation rules.



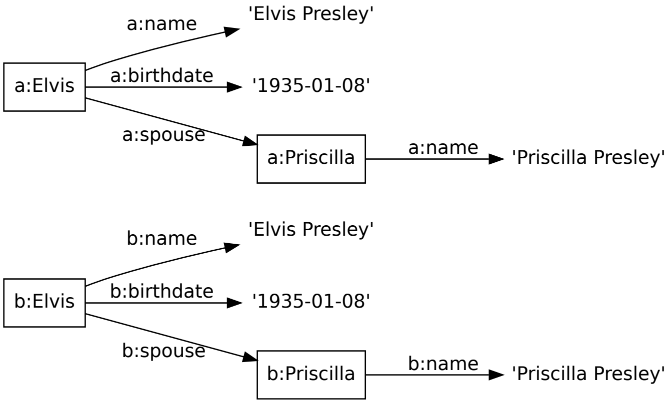


- The rules are represented as a system of equations which we iterate until a fixpoint is reached:

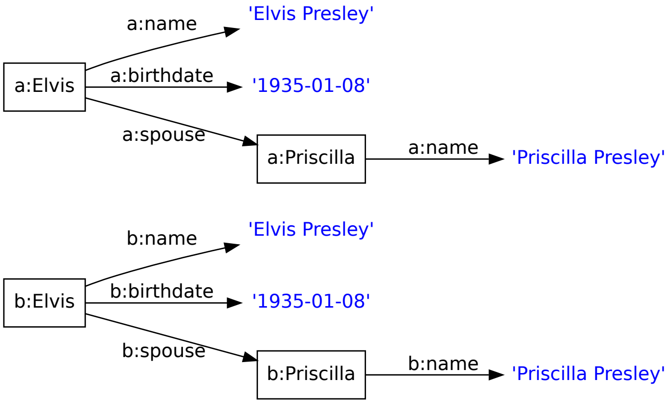
$$\Pr^{n+1}(x \equiv x') = 1 - \prod_{\substack{r(x,y)\\ r'(x',y')}} \left(1 - \Pr^n(r' \subseteq r) \times fun^{-1}(r) \times \Pr^n(y \equiv y')\right) \times \left(1 - \Pr^n(r \subseteq r') \times fun^{-1}(r') \times \Pr^n(y \equiv y')\right)$$

$$\Pr^{n+1}(r \subseteq r') = \frac{\sum_{r(x,y)} \left(1 - \prod_{r'(x',y')} \left(1 - (\Pr^n(x \equiv x') \times \Pr^n(y \equiv y'))\right)\right)}{\sum_{r(x,y)} \left(1 - \prod_{x',y'} \left(1 - \Pr^n(x \equiv x') \times \Pr^n(y \equiv y')\right)\right)}$$

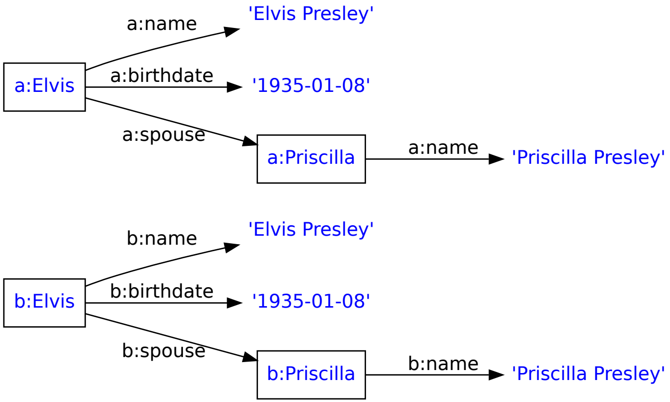
Paris by Example



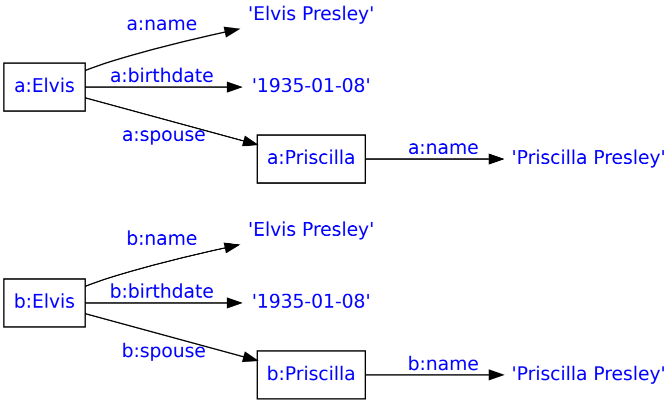
Paris by Example



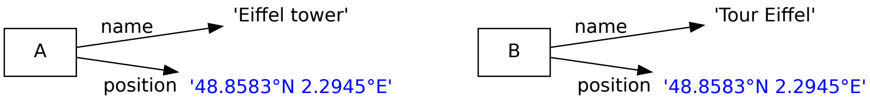
PARIS by Example



Paris by Example



Relation Functionalities



- Two instances should be aligned when they share the same values for aligned functional relations.
- In theory, the ontology schema should indicate which relations are functional.
- In practice, no schema, and no "strict" functionality: compute a fuzzy functionality in $[0, 1]$ from the data.

Existing Implementation and Previous Results

- PARIS is implemented in Java.
- PARIS was evaluated on:
  - toy datasets from the OAEI,
  - DBpedia and YAGO (two ontologies extracted from Wikipedia)
  - YAGO and IMDb
- The evaluation is done in terms of precision, recall and F-measure.

	Instances			Classes			Relations		
	Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
OAEI person	100%	100%	100%	100%	100%	100%	100%	100%	100%
OAEI restaurant	95%	88%	91%	100%	100%	100%	100%	66%	88%
DBpedia–Yago	90%	73%	81%	94%	-	-	93%	-	-
IMDb–Yago	94%	90%	92%	28%	-	-	100%	80%	89%

Table of Contents

Table of Contents

Performance Improvements

- The original PARIS takes a few hours per iteration.
- Ways to improve this:
  - Replace BerkeleyDB by an in-memory representation of the ontologies.
  - Parallelize the propagation of entity alignment scores over all entities. Aggregate results at the end to avoid races.
  - Change the hardware (now that the computation is CPU-bound).

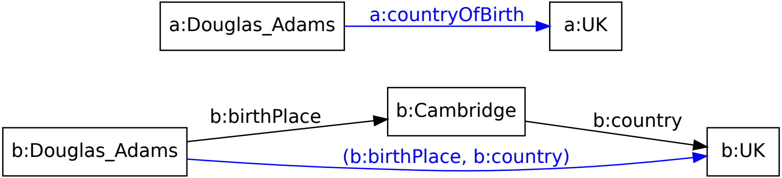
Performance Improvement Results

Iteration	Original PARIS	New PARIS (1 thread)	New PARIS (4 threads)
Startup	0h00	0h27	0h10
1	4h04	0h40	0h27
2	5h06	3h00	1h02
3	5h00	0h34	0h24
4	5h30	0h29	0h16
Total	20h	5h	2h

Table: Running times for the DBpedia–YAGO alignment task. The original PARIS was run on an Intel Xeon E5620 CPU clocked at 2.40 Ghz on a machine with 12 GB of RAM. The new PARIS was run on an Intel Core i7-3820 CPU clocked at 3.60 Ghz with 48 GB of RAM.

Table of Contents

Join Relations



- The simplest possible difference in structure between ontologies: relations of one ontology correspond to join relations in the other ontology.
- The terminology is motivated by the “join” operator of relational algebra.
- We see the join as a binary predicate: the intermediate nodes are existentially quantified but projected away.

Support in PARIS

- We must keep the representation of joins implicit in PARIS (memory constraints).
- We must recursively enumerate all possible join facts instead of enumerating all possible facts.
- We must avoid duplicate facts caused by multiple possible choices for the intermediate nodes.
- We cannot afford to enumerate all possible relations anymore (many possible joins).

$\Rightarrow$ New algorithm to compute the entity and relation alignments simultaneously.

Practical Issues

- How to determine the functionality of join relations?
- How to select interesting joins to perform without exploring all joins?
- How to achieve acceptable running time on large ontologies?

$\Rightarrow$ We only perform the join alignment on small ontologies.

Table of Contents

Log-transformation and Product Graph

$$\Pr^{n+1}(x \equiv x') = 1 - \prod_{\substack{r(x,y)\\ r'(x',y')}} \Big(1 - \Pr^n(r' \subseteq r) \times \mathit{fun}^{\text{-}1}(r) \times \Pr^n(y \equiv y')\Big) \times \Big(1 - \Pr^n(r \subseteq r') \times \mathit{fun}^{\text{-}1}(r') \times \Pr^n(y \equiv y')\Big)$$

- The entity alignment equation is justified by a probabilistic model (independent choices).
- If the relation functionalities and alignments are in $\{0, 1\}$, we can apply a log-transformation:

$$\mathrm{LPr}^n(x \equiv x') := -\log(1 - \Pr^n(x \equiv x'))$$

- By looking at propagation in the product graph, we get a nicer equation, for some matrix M and a constant literal alignment vector L:

$$\mathrm{LPr}^{n+1} = M\,\mathrm{LPr}^n + L$$

Green Measures

$$\mathrm{LPr}^{n+1} = M\,\mathrm{LPr}^{n} + L$$

- This equation is similar to PageRank ($\mathrm{LPr}^{n+1} = M\,\mathrm{LPr}^{n}$ where M is a stochastic matrix) except:
  1. The matrix is not stochastic.
  2. Diverging to $+\infty$ means convergence (because of the log-transformation).
  3. L is pouring alignment weight to the aligned couples of literals.
- This last point can be linked to the use of Green measures to focus the PageRank computation.
- This interpretation suggests possible changes to the entity alignment equation (but we lose the probabilistic interpretation).

Table of Contents

Literal Similarity Functions

`Edgar R. Burroughs`	`Douglas Adams`	`and Constance Garnett`
`Edgar Rice Burroughs`	`Adams, Douglas`	`Constance Garnett`

- The original Paris uses an exact literal equality function.
- Possible refinements: adjust for case, strip special characters, etc.
- Yet, we would need a better equality function giving > 0 weight to the alignment of similar literals.
- Approximate dictionary searching problem: given a literal, to find quickly all similar literals in the other ontology.

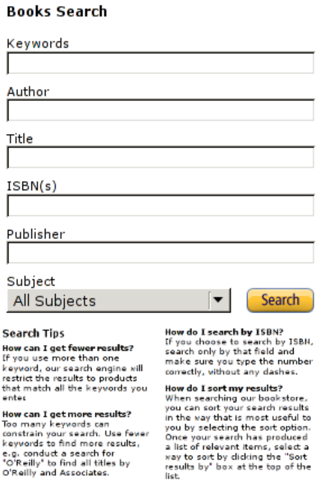
Results

- We use a shingling technique which was implemented by Mayur Garg (who interned in the team from IIT Delhi).
- I interfaced his code with PARIS.
- The performance of the shingling technique matches ad-hoc normalization on the OAEI restaurants dataset.

	Precision	**Recall**	**F-measure**
Paris with exact equality	95%	88%	91%
Paris with shingling	96%	95%	96%
Paris with normalization	98%	96%	97%

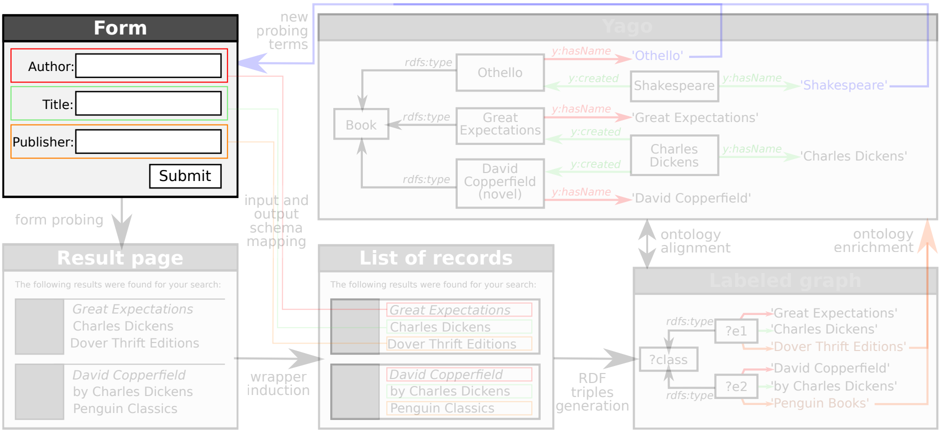
Table of Contents

The Deep Web

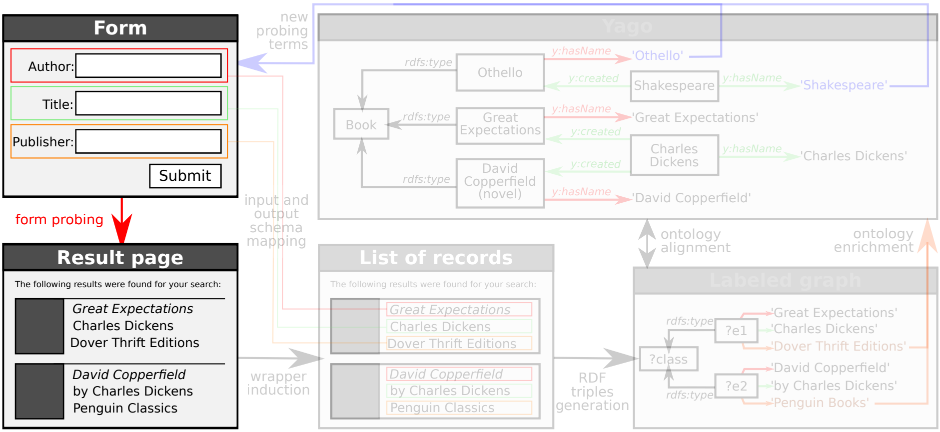


- Many structured databases can only be queried through interfaces designed for humans (Web forms and HTML result pages).
- To access this structured information, an automated agent must probe the form and perform wrapper induction on the result pages.
- To understand the meaning of the extracted records and attributes, we can use Paris (with a reference ontology).

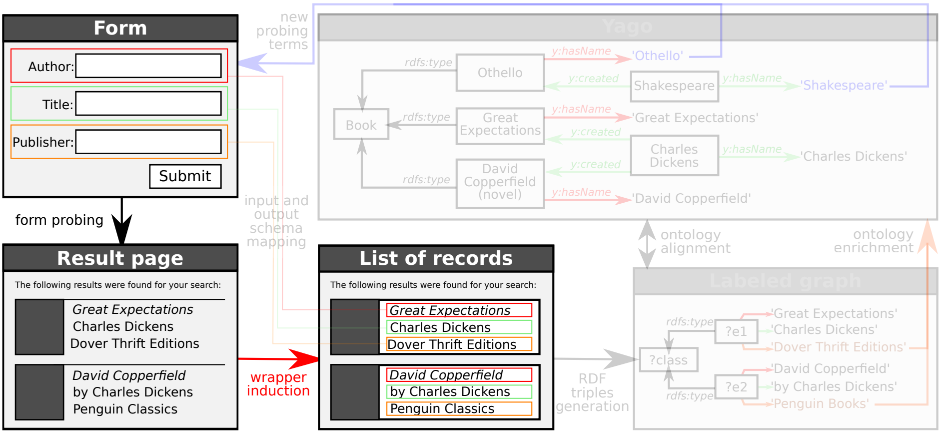
Application to Form Understanding



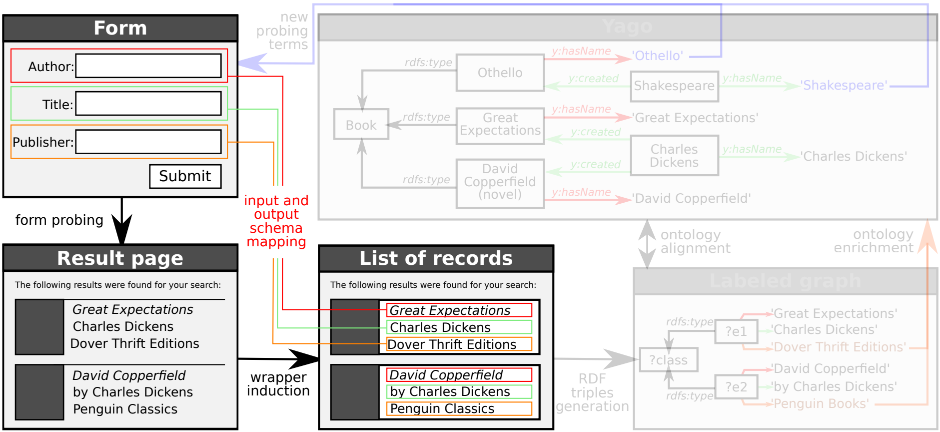
Application to Form Understanding



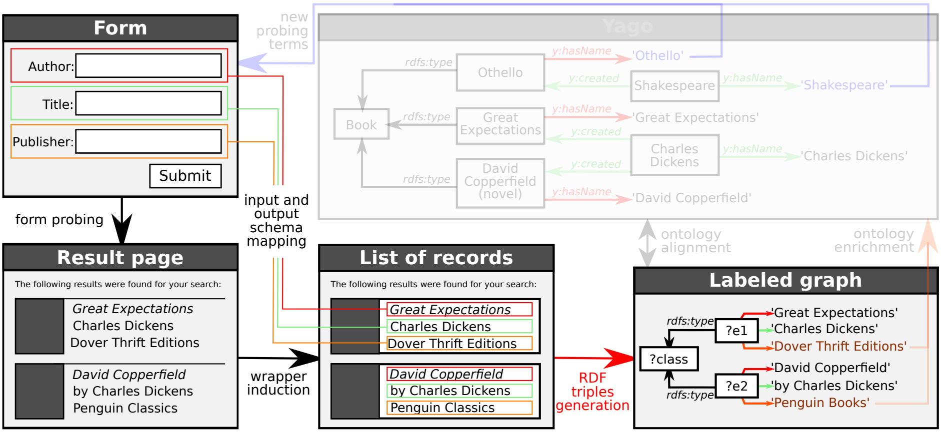
Application to Form Understanding



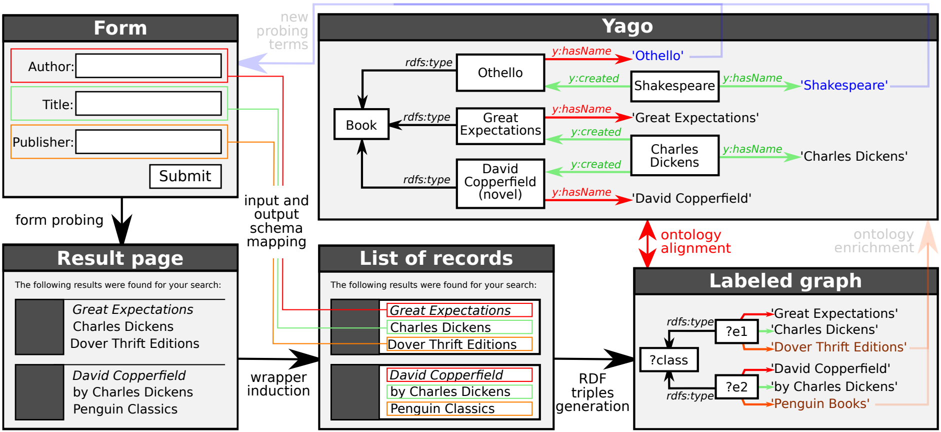
Application to Form Understanding



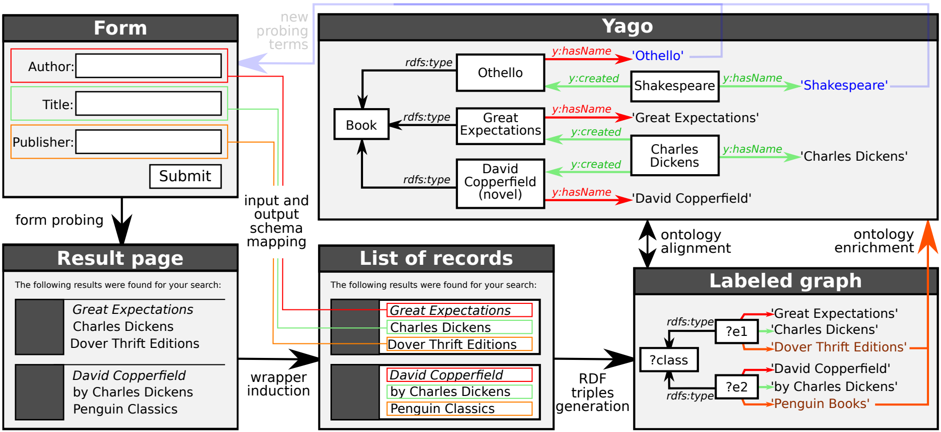
Application to Form Understanding



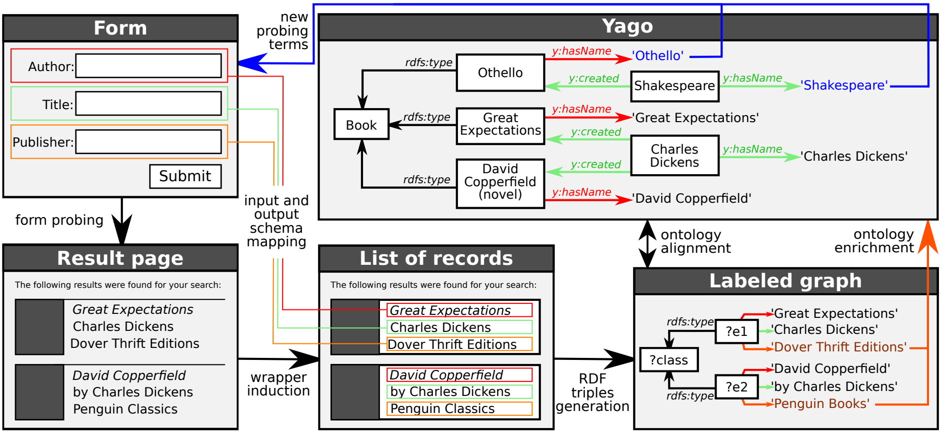
Application to Form Understanding



Application to Form Understanding



Application to Form Understanding



Results

- We experimented the approach on the Amazon book search form.
- The entity alignments with the best confidence were indeed books aligned through their title and author.
- The system identified relations: `y:hasPreferredName` and (`y:created`, `y:hasPreferredName`).
- It linked them to the result page DOM paths and form fields.
- The support for join relations and approximate string matching is required in this setting.
- The approach was presented as a vision paper in the VLDS workshop of VLDB.

Table of Contents

Summary of Contributions

- Performance improvements resulting in an 10-fold speedup over the original implementation.
- Support of join relation alignments on small ontologies.
- Insights on the relation between Paris and PageRank-inspired techniques.
- Integration of approximate string matching to improve the literal alignment.
- Application of Paris for deep Web analysis.

Further Work

Performance. Further gains to be made, perform more complete benchmarks.

Join relations. Performance improvements, especially ways to only select interesting joins. Arbitrary patterns?

Theory. Study the possible alternative choices and benchmark them. Understand the full model (we *still* have no proof of overall convergence!) and the effects of implementation tweaks. Find links with Max-SAT or Markov Logic Networks?

Literal matching. Support of various datatypes such as numbers and dates (engineering work). Fix performance issues to perform larger experiments.

Information Extraction. Try with more sources. Find links with named entity disambiguation techniques such as AIDA? Intensional use for large-scale integration.

Thanks!

Thanks for your attention!

Questions ?

The research has been funded by the European Union's seventh framework programme, in the setting of the European Research Council grant Webdam, agreement 226513, and the FP7 grant ARCOMEM, agreement 270239.

Frame 4: *Linking Open Data cloud diagram*, by Richard Cyganiak and Anja Jentzsch. `http://lod-cloud.net/`