Identifying Relevant Sources for Data Linking using a Semantic Web Index

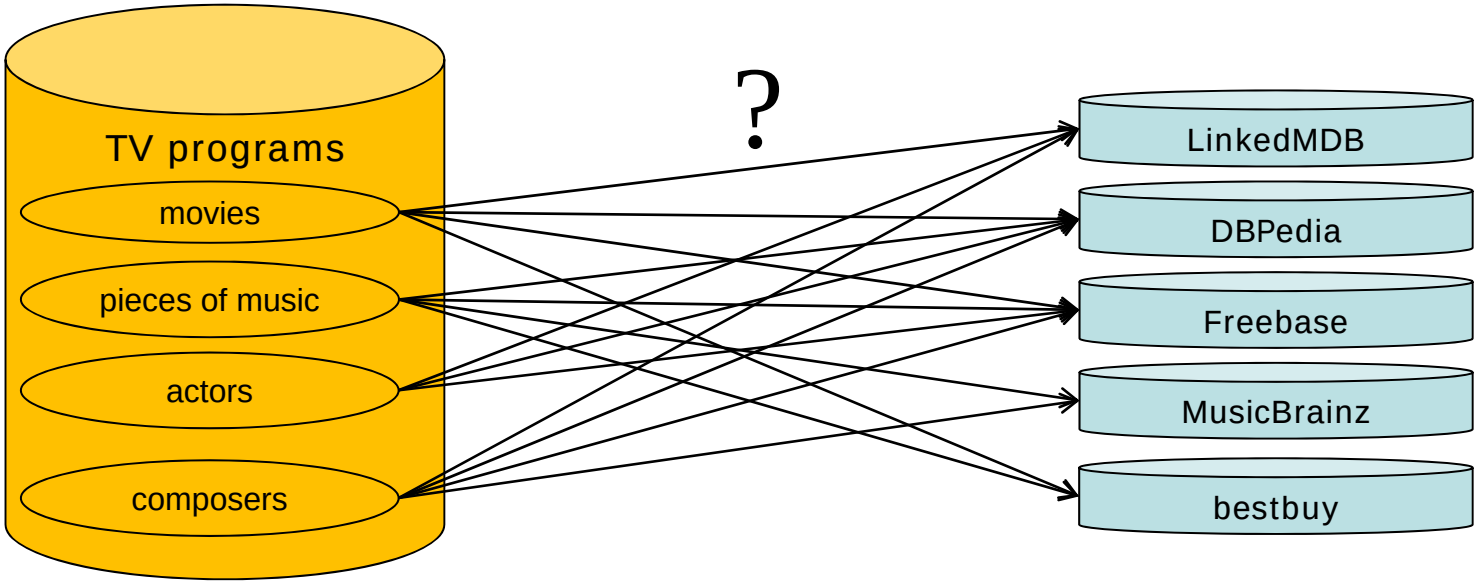
Andriy Nikolov
Mathieu d'Aquin

Knowledge Media Institute
The Open University, UK

The Open University

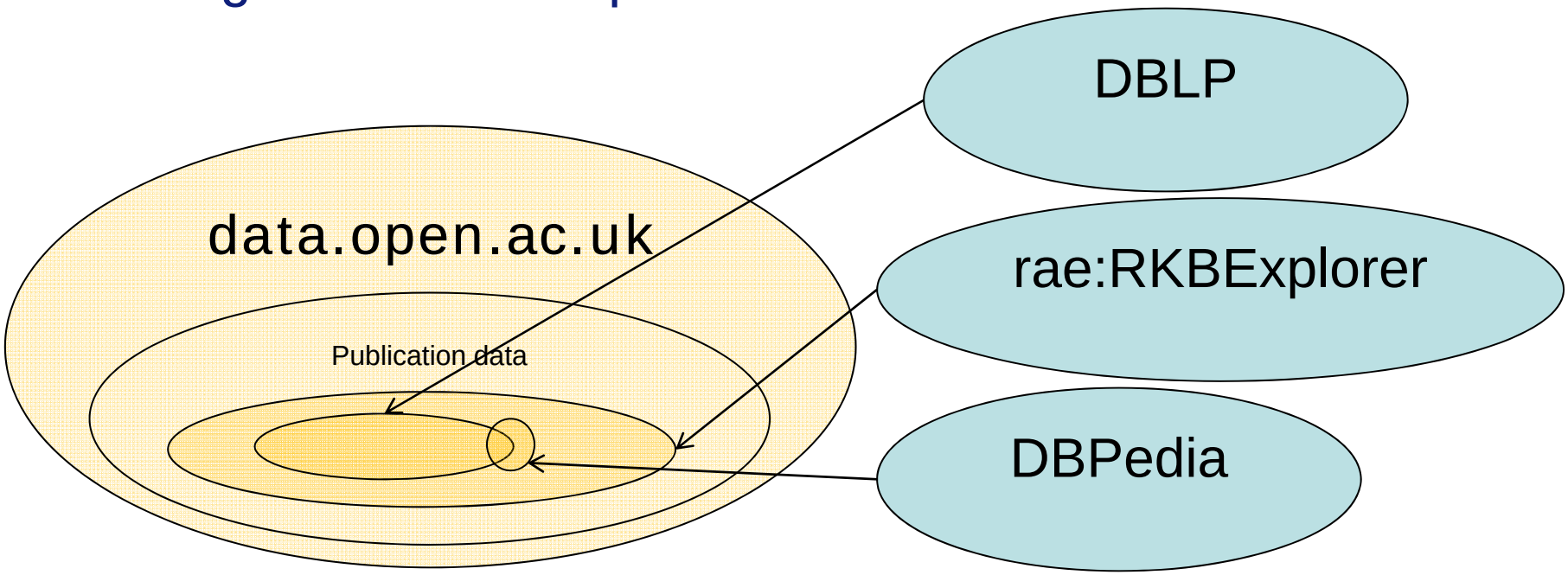
How to link a new dataset?

- What other repositories contain relevant data which I should link to?
  - *Select the **external repository***
- How to select the relevant data instances to link?
  - *Select the **relevant classes** within the chosen repository*



Selection criteria

- Additional information about local instances
- Popularity
- Degree of overlap

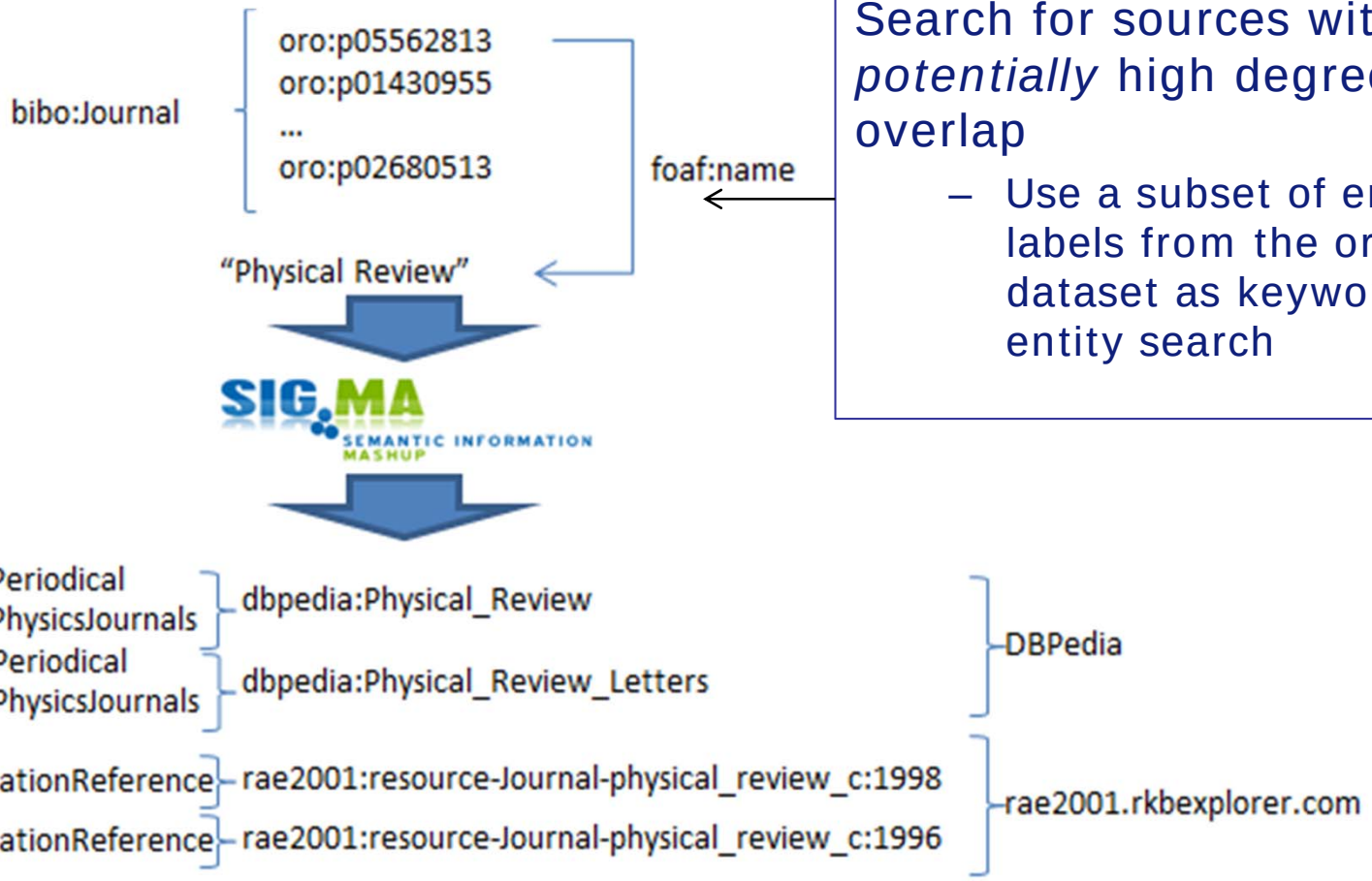


KNOWLEDGE MEDIA INSTITUTE

Available information

- Additional information about resources
  - *Schema ontology*
  - *Test examples*
- Popularity
  - *VoiD descriptors*
    - Linking repositories
  - *Catalog of repositories (CKAN)*
- Degree of overlap
  - *VoiD descriptors (only topic relevance)*
  - ***Relevant info hard to obtain on the client side***

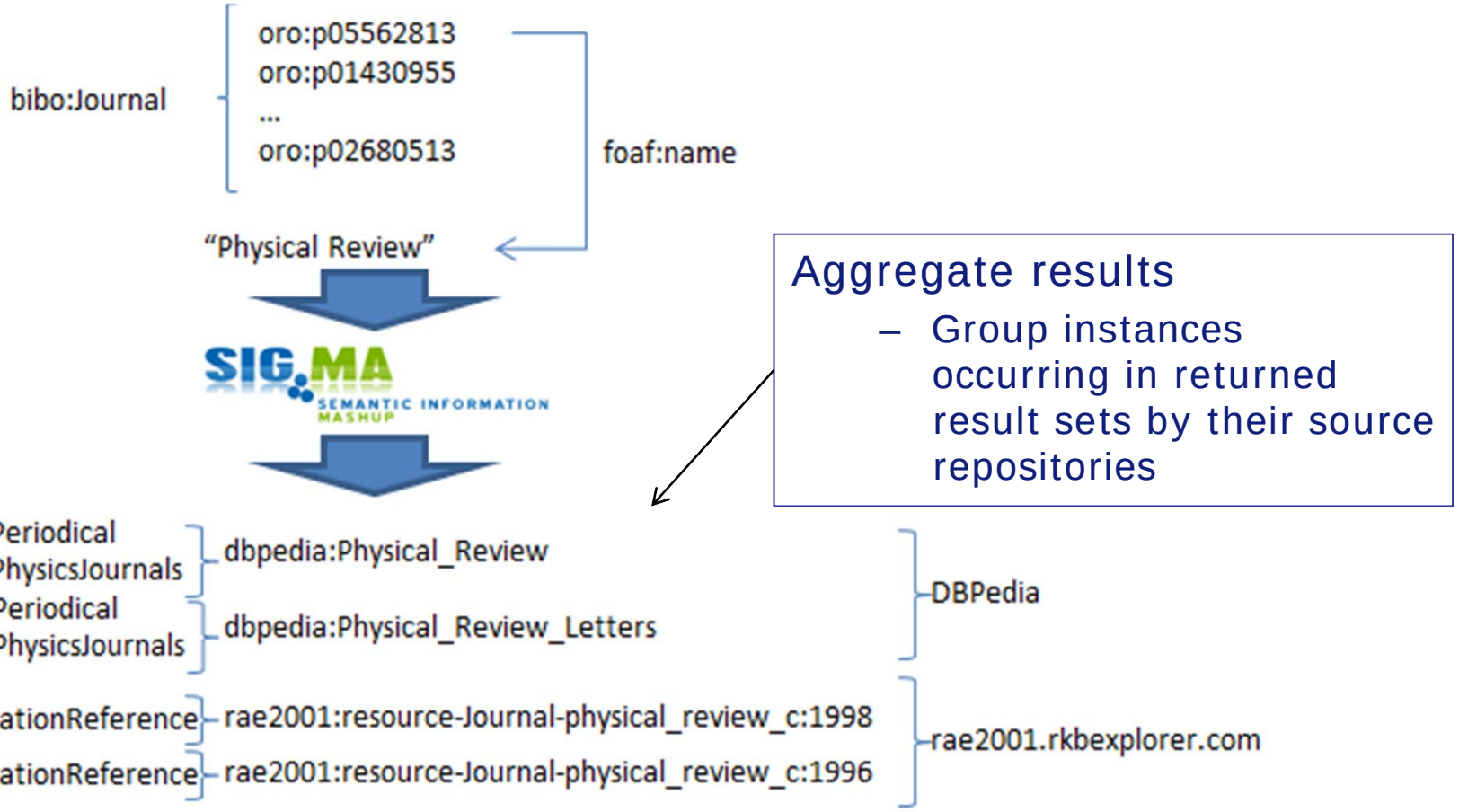
Approach



Search for sources with *potentially* high degree of overlap

- Use a subset of entity labels from the original dataset as keywords for entity search

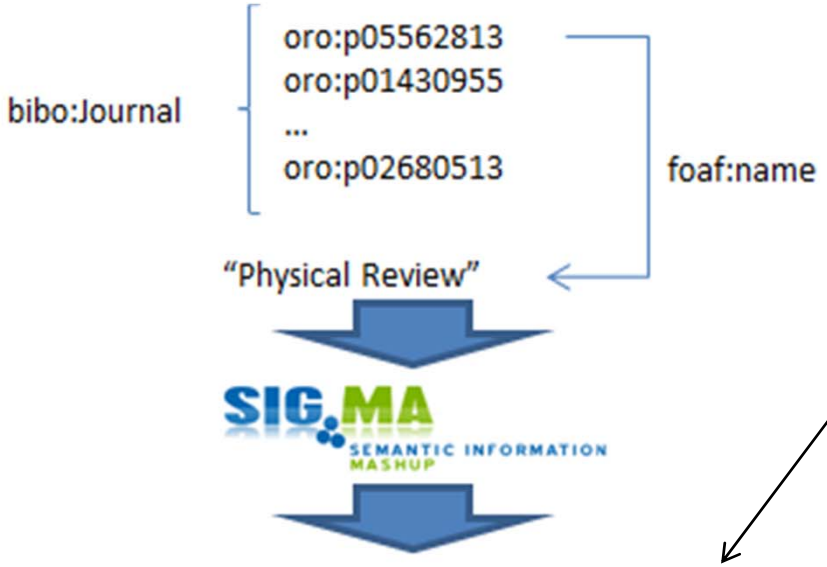
Approach



Aggregate results

- Group instances occurring in returned result sets by their source repositories

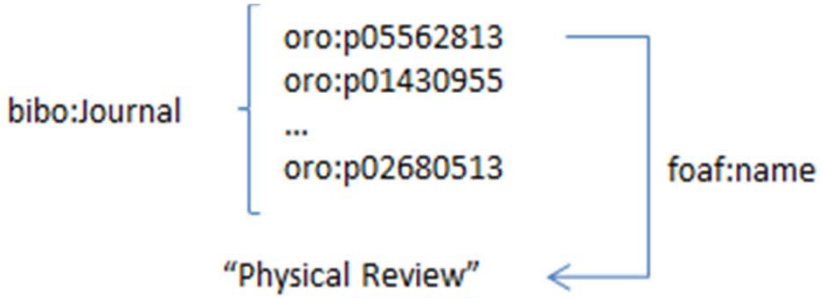
Approach



Rank sources

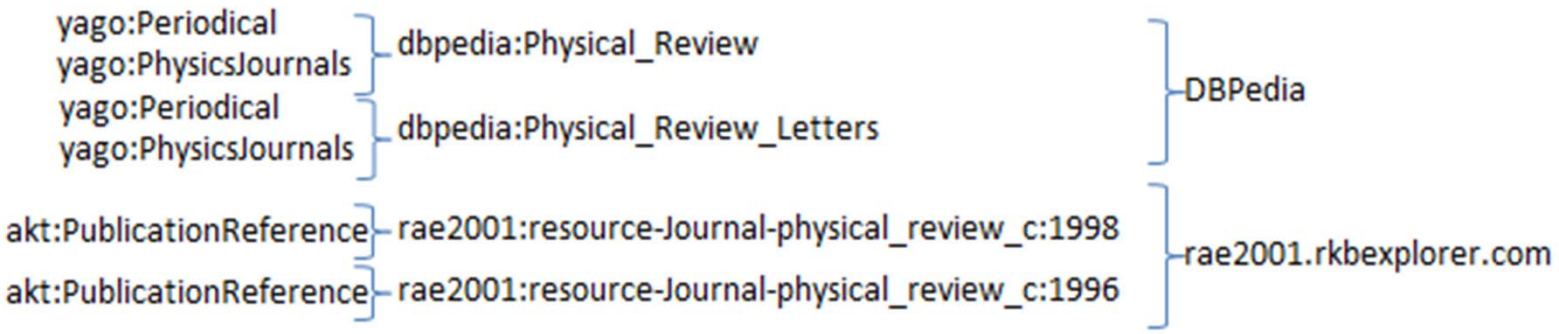
- Sort by number of individuals returned in search results

Approach



Select "most relevant" class

- Select the class in each source, which covers most of instances



KNOWLEDGE MEDIA INSTITUTE

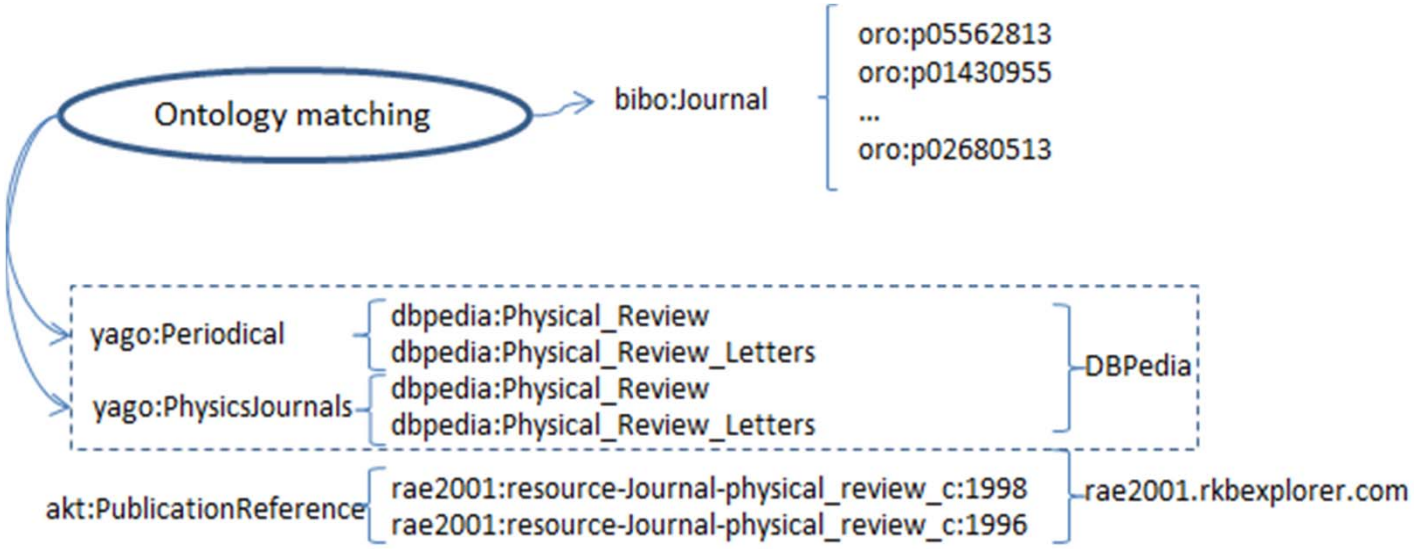
Issues: imprecise results

- Main cause: ambiguous instance labels
- Inclusion of irrelevant sources
  - *E.g., DBLP for movie score composers*
- Selection of inappropriate classes within the selected source
  - *Too generic: e.g., dbpedia:Person vs dbpedia:MusicArtist*
  - *Irrelevant: e.g., akt:Publication-Reference (journal volume) vs akt:Journal*

Filtering results

Determine potentially irrelevant classes

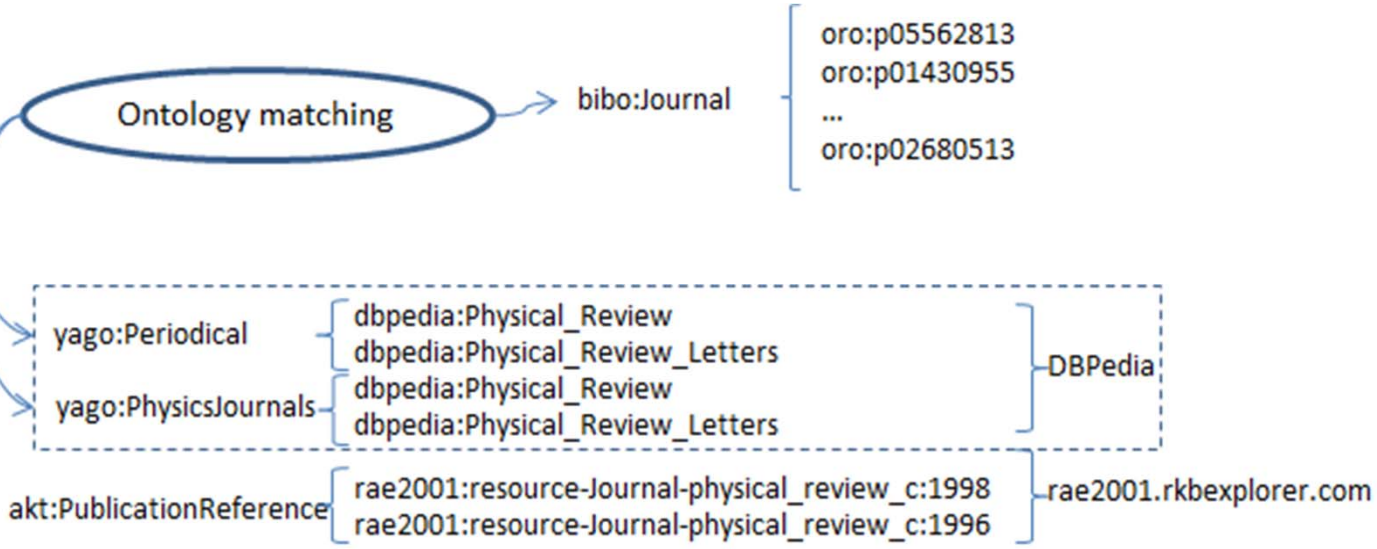
- Use state-of-the-art schema matching to select relevant classes



Filtering results

Filter out irrelevant search results

- Only consider search result instances belonging to “approved” classes



KNOWLEDGE MEDIA INSTITUTE

Preliminary experiments

- Datasets
  - *ORO journals (data.open.ac.uk): 3110 instances*
  - *LinkedMDB films: 400 instances*
  - *LinkedMDB music contributors: 400 instances*
- External components
  - *Semantic index: Sig.ma*
  - *Ontology matching techniques: CIDER, instance-based schema mappings retrieved from BTC2009 dataset*

Preliminary experiments

- Performance measure:
  - *Proportion of relevant sources among the top-10 returned results*

Before filtering	+ / -	After filtering	+ / -
rae2001 (RKB)	+	rae2001 (RKB)	+
dotac (RKB)	+	DBPedia	+
DBPedia	+	dblp.l3s.de	+
oai (RKB)	+	Freebase	+
dblp.l3s.de	+	DBLP (RKB)	+
wordnet (RKB)	-	eprints (RKB)	+
bibsonomy	-		
eprints (RKB)	+		
Freebase	+		
www.examiner.com	-		

Preliminary experiments

- Summary:
  - *Top-ranked returned repositories are largely relevant from the point of view of linking*
  - *Filtering using schema matching techniques greatly improves precision (all remaining sources are relevant)*
  - *… but at the expense of some recall*

KNOWLEDGE MEDIA INSTITUTE

Future work

- Improving the quality of results
  - *E.g., estimating the potential loss of precision/recall for different filtering decisions*
- Integrating with the data linking workflow
  - *Automatically pre-configuring the data linking algorithm*
- Repository search as a potentially useful semantic search use case (in addition to entity and document search)

KNOWLEDGE MEDIA INSTITUTE

Questions?

Thanks for your attention