The University of Manchester

A Snapshot of the OWL Web

Nicolas Matentzoglu
Samantha Bail
Bijan Parsia

University of Manchester, School of Computer Science, Information Management Group

The University of Manchester

The WEB Ontology Language

- Features are
  - web centered: URIs everywhere
  - web based: owl:imports
- The majority of OWL is on the Web
  - ….somewhere
- To understand OWL, we must find it

The University of Manchester

Studying ontologies

- Exemplars
  - SNOMED
  - Galen
- Curated collections
  - BioPortal
  - TONES / Oxford OL
- Large-scale repositories
  - Swoogle
  - Watson

Studying ontologies

- Exemplars
  - SNOMED
  - Galen
- Curate
  - BioPo
  - TONE
- Large-sc repositories
  - Swoogle
  - Watson

Versions
Variants
Faceted
publishing

The University of Manchester

Studying ontologies

- Exemplars
  - SNOMED
  - Galen
- Curated collections
  - BioPortal
  - TONES / Oxford OL
- Large-scale repositories
  - Swoogle
  - Watson

The University of Manchester

Studying ontologies

- Exemplars
  - SNOMED
  - Galen
- Curated collections
  - BioPortal
  - TONES / Oxford OL
- Large-scale repositories
  - Swoogle
  - Watson

Can we create a more representative collection?

The University of Manchester

Goal: a **snapshot** of ontologies

-

-

The University of Manchester

Goal: a **snapshot** of ontologies

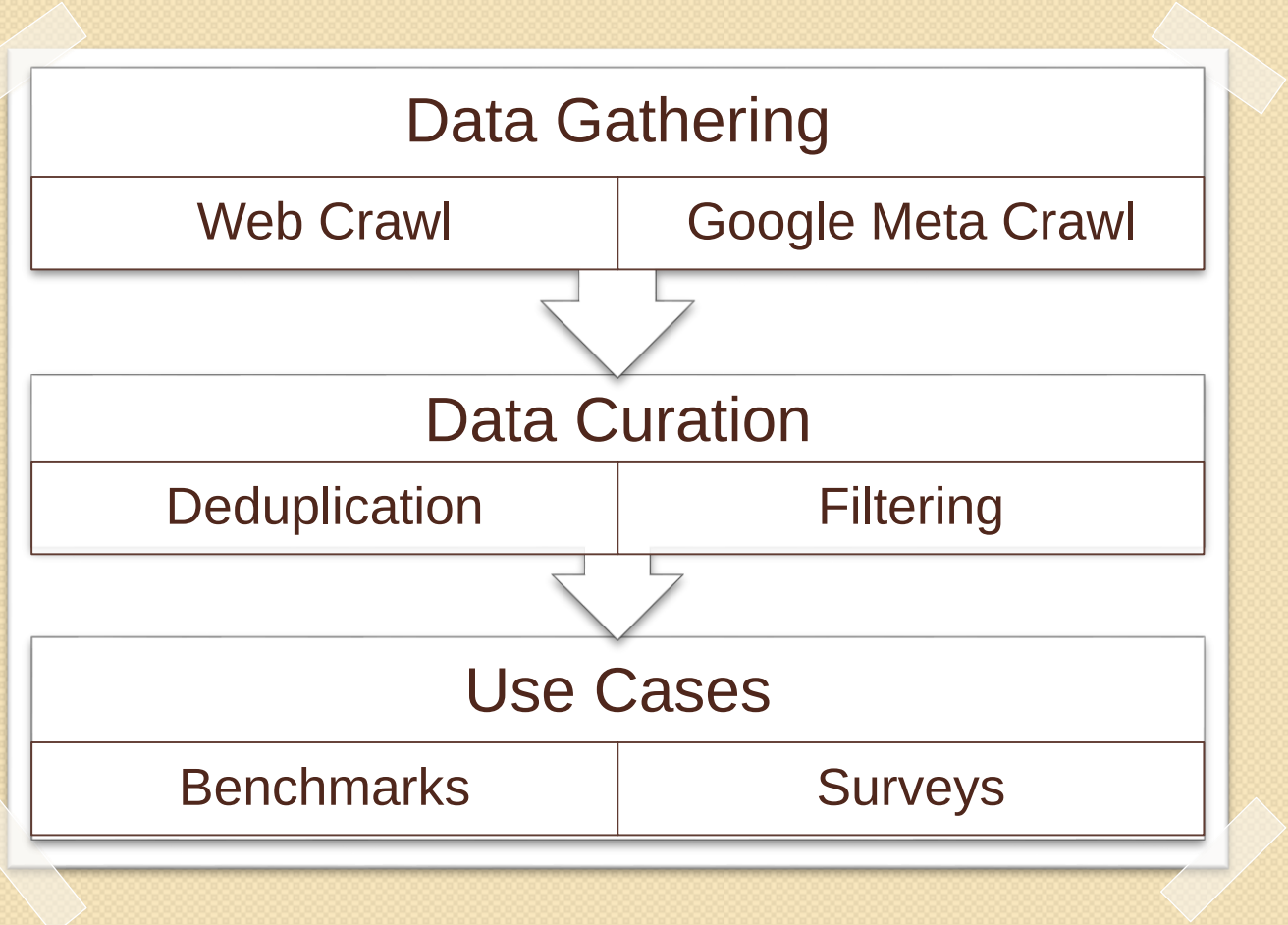
- Take snapshots of ontologies on the Web: collections…
  - of distinct ontologies
  - of arbitrary origin
  - at a given time

The University of Manchester

Goal: a **snapshot** of ontologies

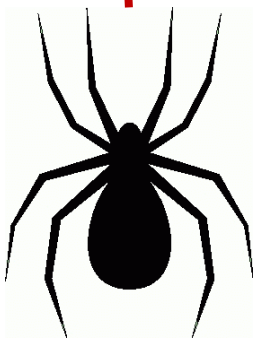
- Take snapshots of ontologies on the Web: collections…
  - of distinct ontologies
  - of arbitrary origin
  - at a given time
- …in order to support meaningful experimentation
  - make statements about ontologies
  - test our tools and techniques
  - make informed decisions

Mechanics


Data Gathering
Web Crawl
Google Meta Crawl
Data Curation
Deduplication
Filtering
Use Cases
Benchmarks
Surveys

Data gathering: KrOWLer

URLs being generated by server-side software has also made it difficult for web crawlers to avoid retrieving duplicate content. Endless combinations of HTTP GET (URL-based) parameters exist, of which only a small selection will actually return unique content. For example, a simple online photo gallery may offer three options to users, as specified through HTTP GET parameters in the URL. If there exist four ways to sort images, three choices of thumbnail size, two file formats, and an option to disable user-provided content, then the same set of content can be accessed with 48 different URLs, all

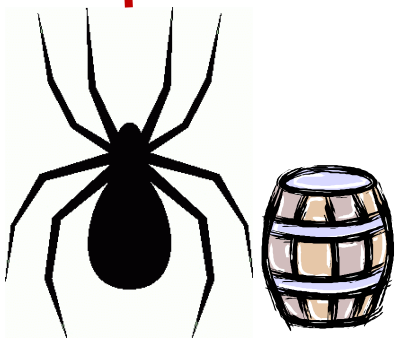


Seeds

Data gathering: KrOWLer

URLs being generated by server-side software has also made it difficult for web crawlers to avoid retrieving duplicate content. Endless combinations of HTTP GET (URL-based) parameters exist, of which only a small selection will actually return unique content. For example, a simple online photo gallery may offer three options to users, as specified through HTTP GET parameters in the URL. If there exist four ways to sort images, three choices of thumbnail size, two file formats, and an option to disable user-provided content, then the same set of content can be accessed with 48 different URLs, all

http://uri.pattern.matcher

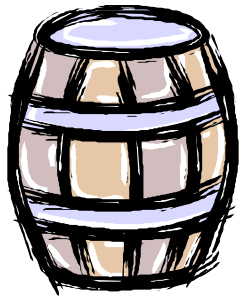


Seeds

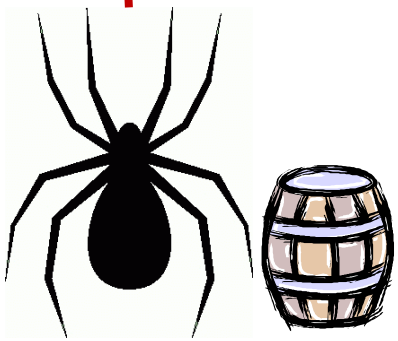
Data gathering: KrOWLer

URLs being generated by server-side software has also made it difficult for web crawlers to avoid retrieving duplicate content. Endless combinations of HTTP GET (URL-based) parameters exist, of which only a small selection will actually return unique content. For example, a simple online photo gallery may offer three options to users, as specified through HTTP GET parameters in the URL. If there exist four ways to sort images, three choices of thumbnail size, two file formats, and an option to disable user-provided content, then the same set of content can be accessed with 48 different URLs, all

http://uri.pattern.matcher

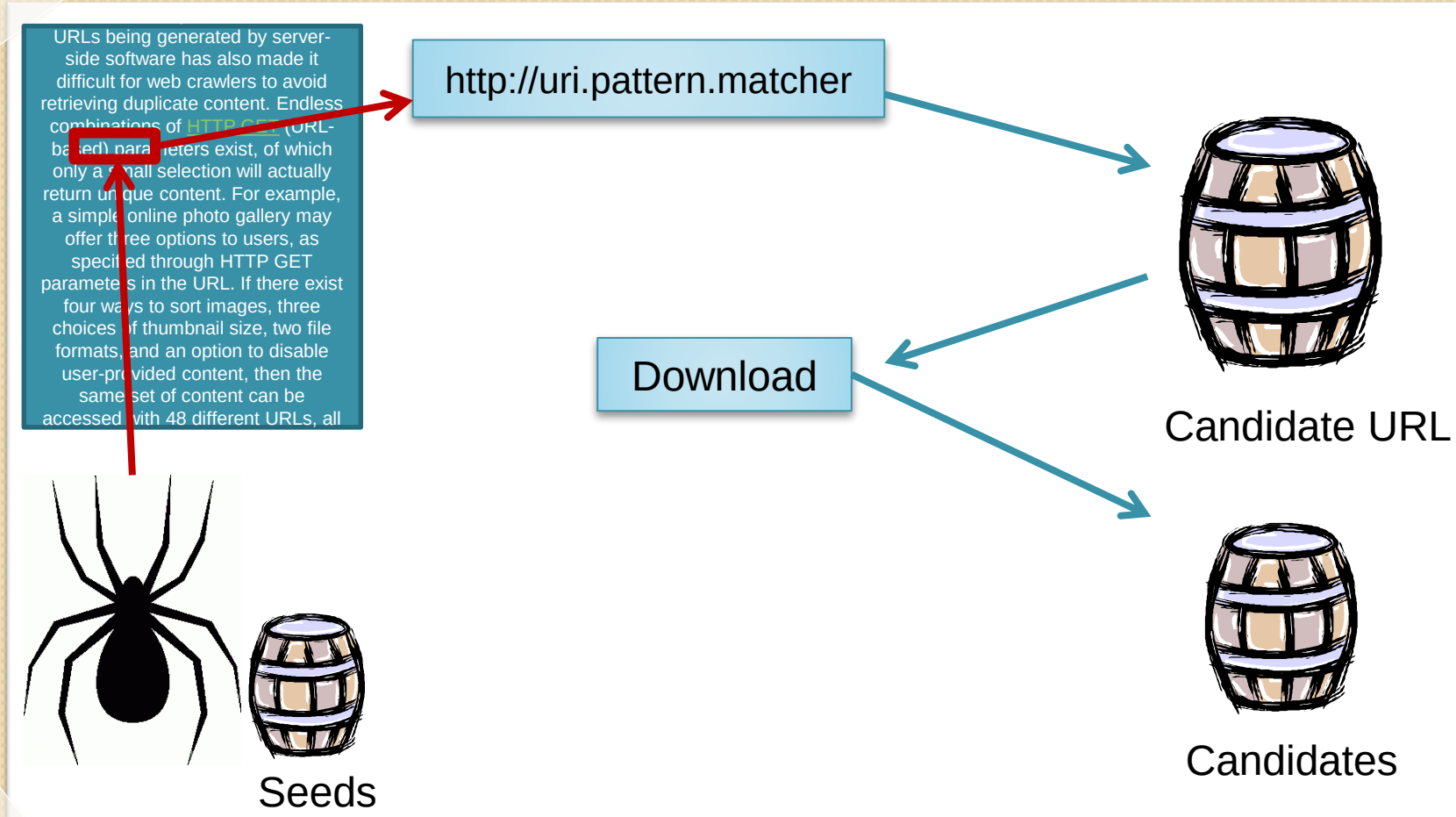


Candidate URL

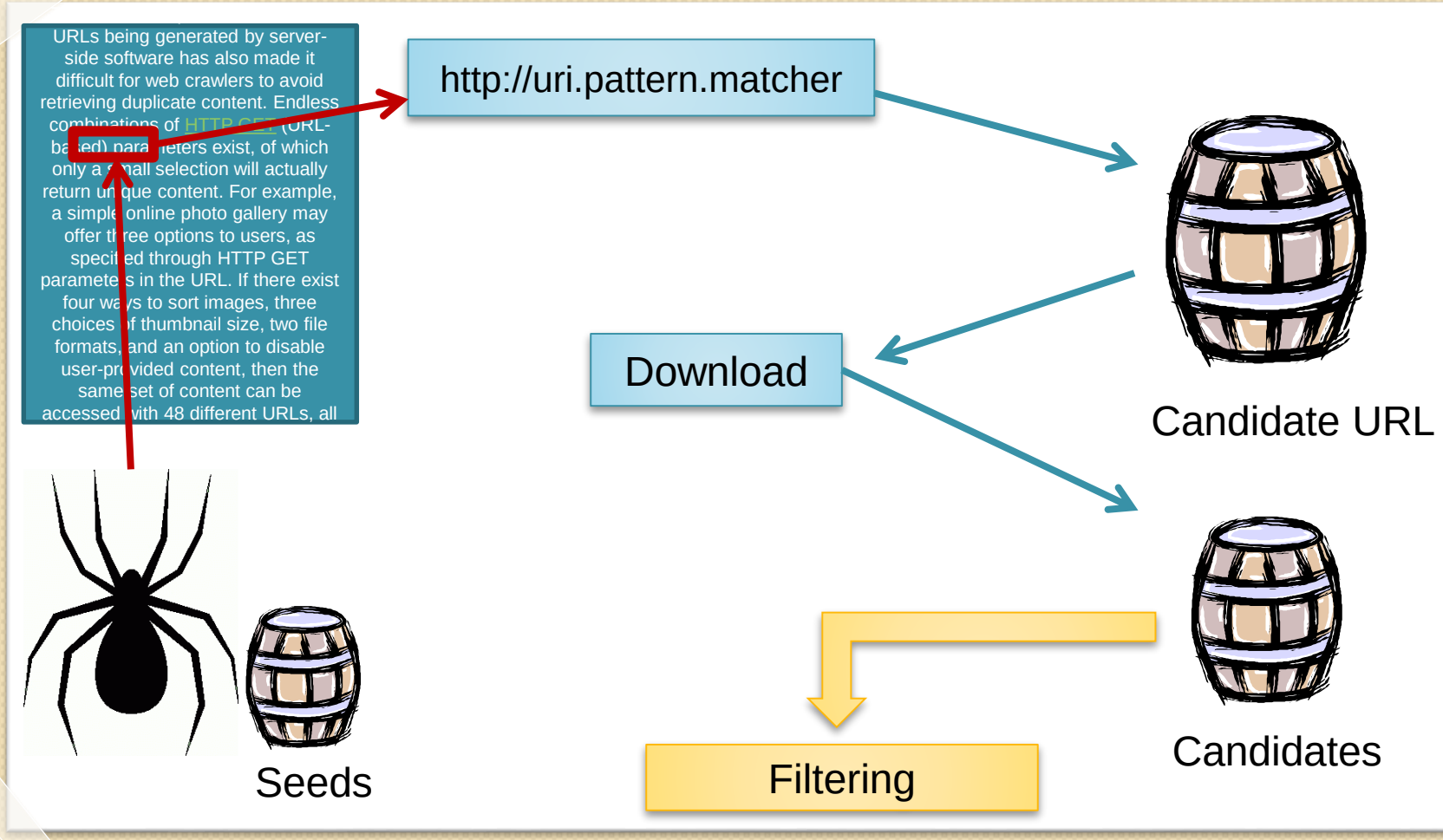


Seeds

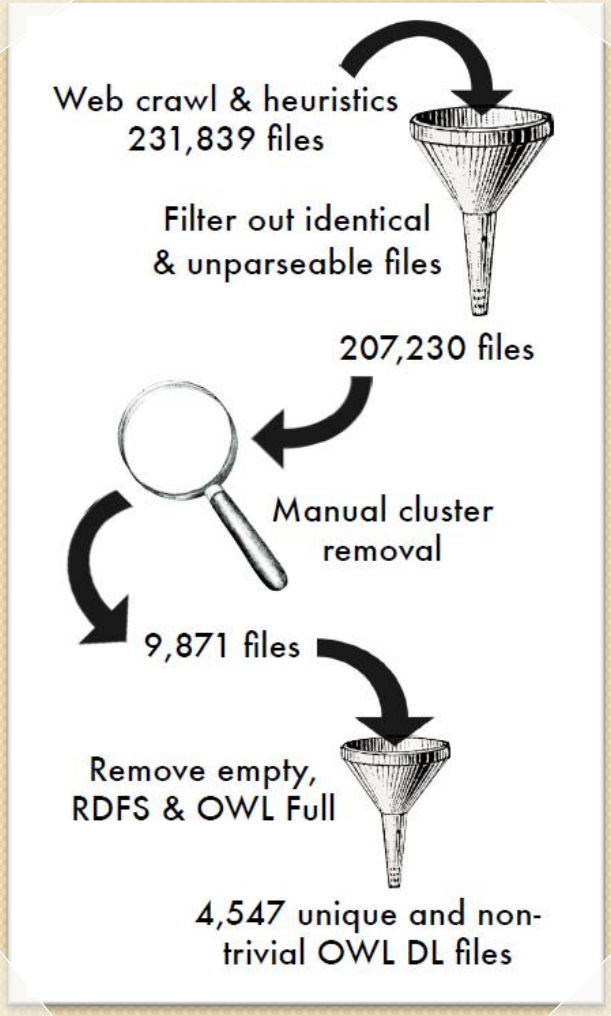
Data gathering: KrOWLer



Data gathering: KrOWLer

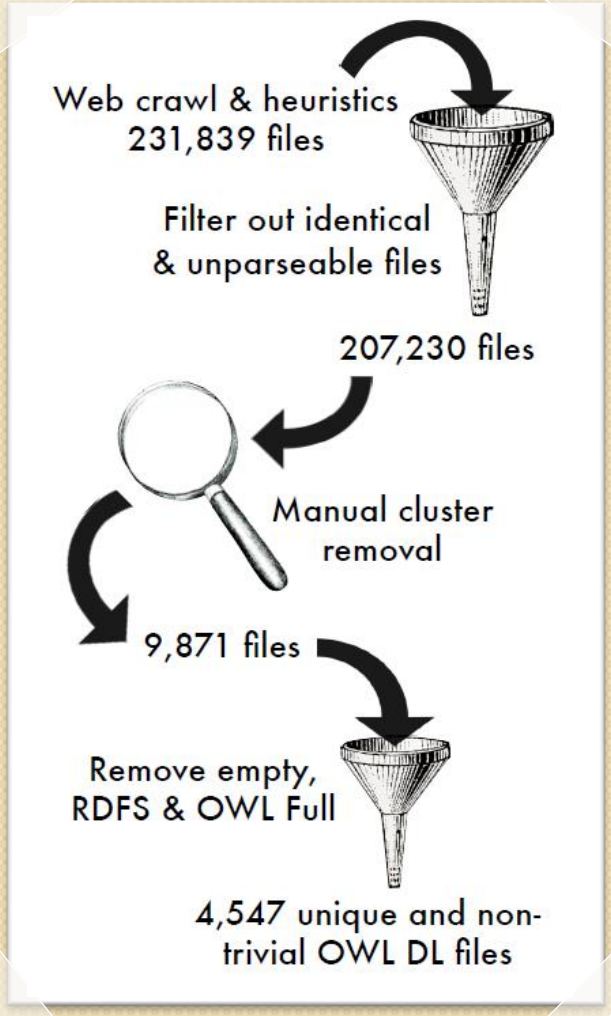


Data curation: Filtering



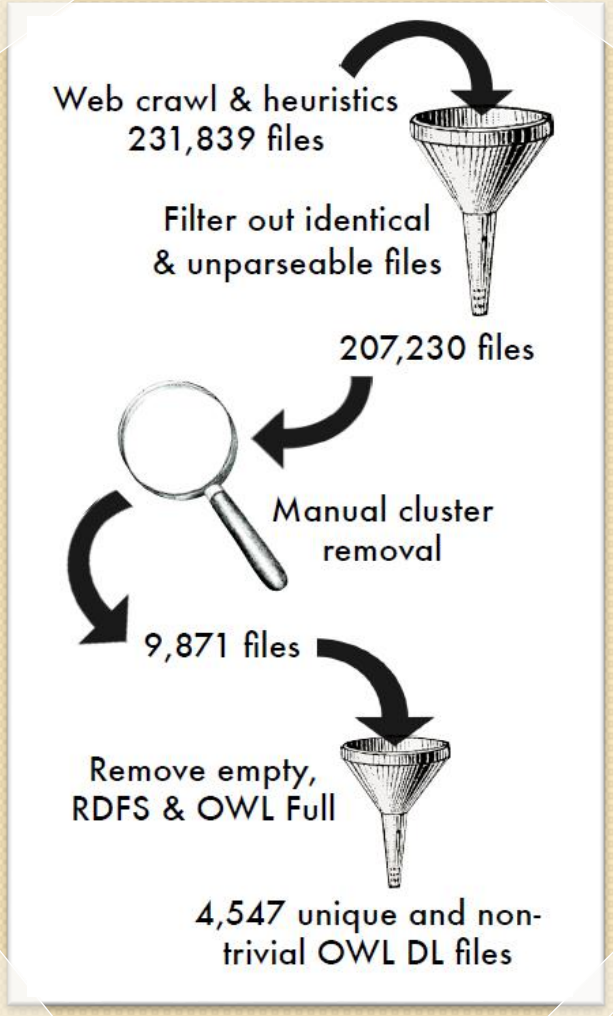
Data curation: Filtering

·Initial download contained ~270,000 candidates



Data curation: Filtering

- Initial download contained ~270,000 candidates
- ~37,000 syntactic heuristics

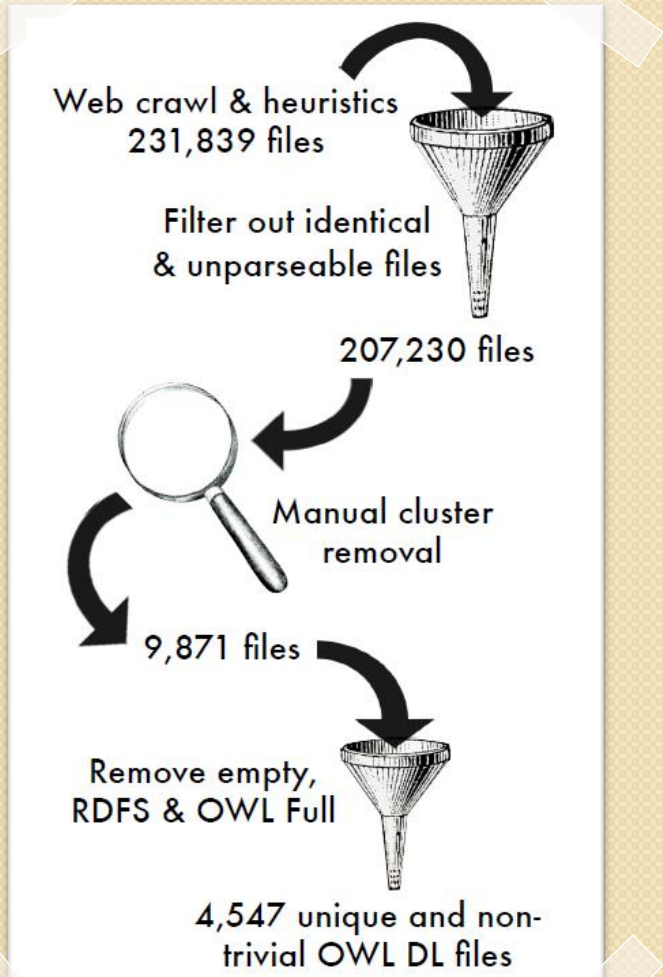


Data curation: Filtering

·Initial download contained ~270,000 candidates

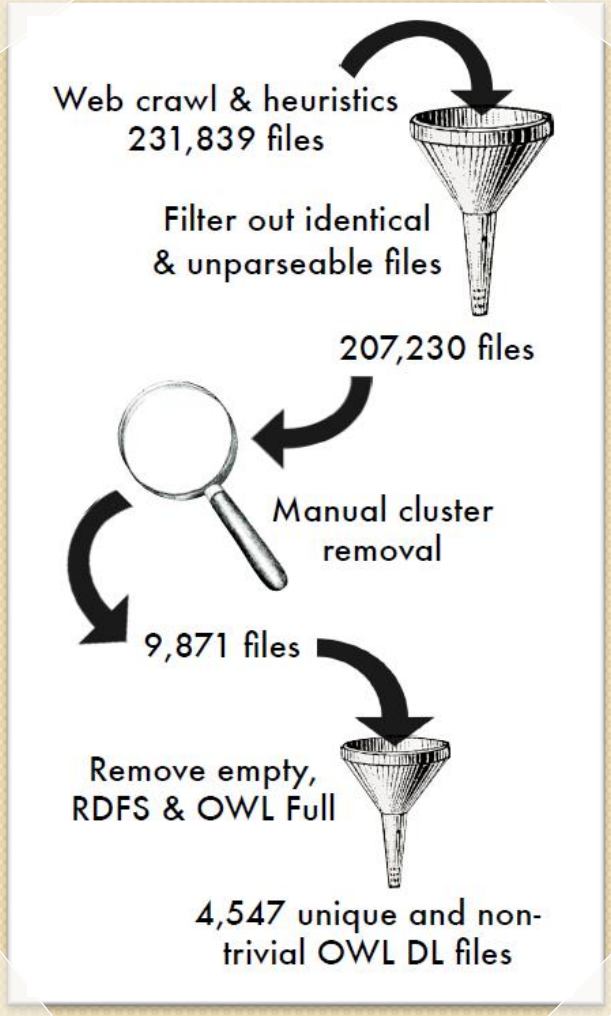
·~37,000 syntactic heuristics

·~18,000 byte identical



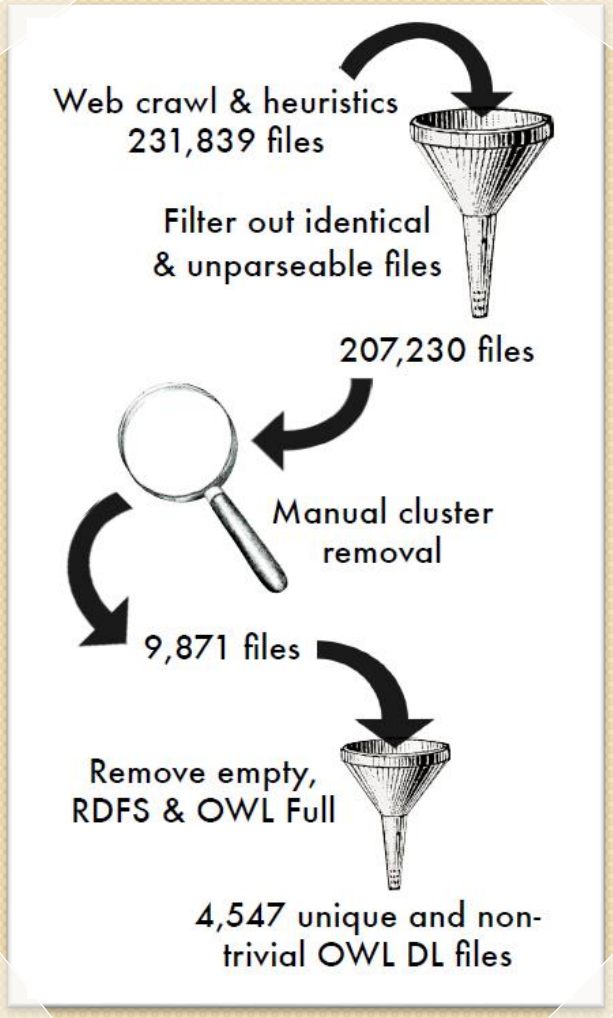
Data curation: Filtering

- Initial download contained ~270,000 candidates
  - ~37,000 syntactic heuristics
  - ~18,000 byte identical
  - ~5,000 not parseable



Data curation: Filtering

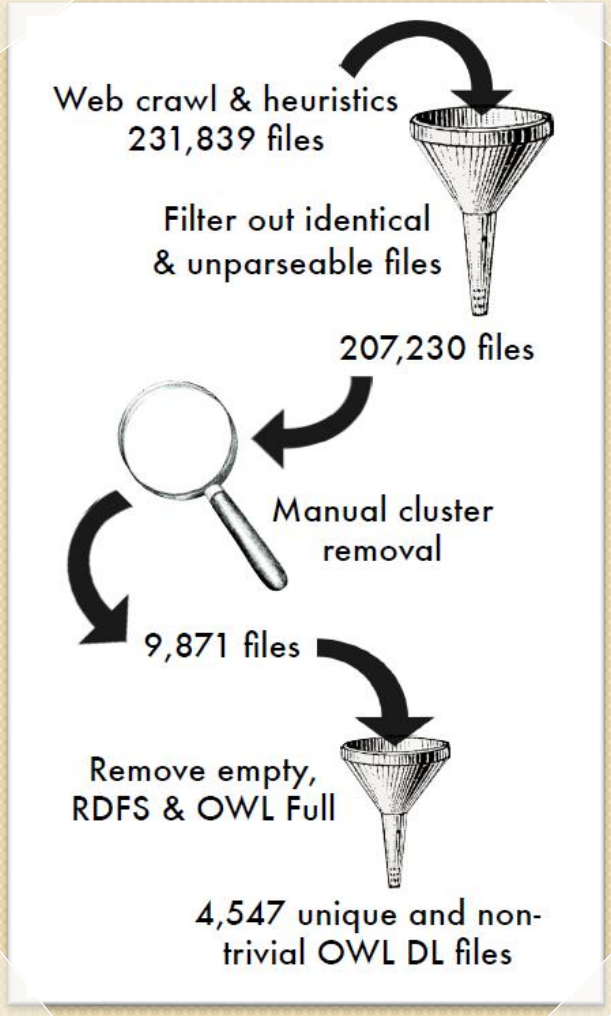
·Initial download contained ~270,000 candidates
·~37,000 syntactic heuristics
·~18,000 byte identical
·~5,000 not parseable
·~6,000 byte identical after OWL/XML serialisation



Data curation: Filtering

·Initial download contained ~270,000 candidates
·~37,000 syntactic heuristics
·~18,000 byte identical
·~5,000 not parseable
·~6,000 byte identical after OWL/XML serialisation

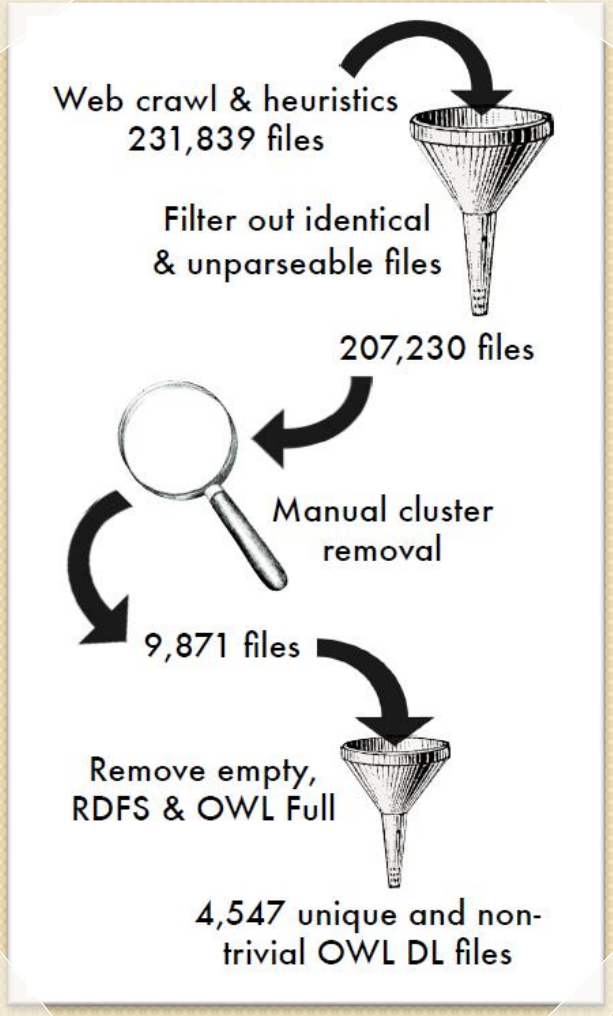
·~200,000 filtered out semi-automatically



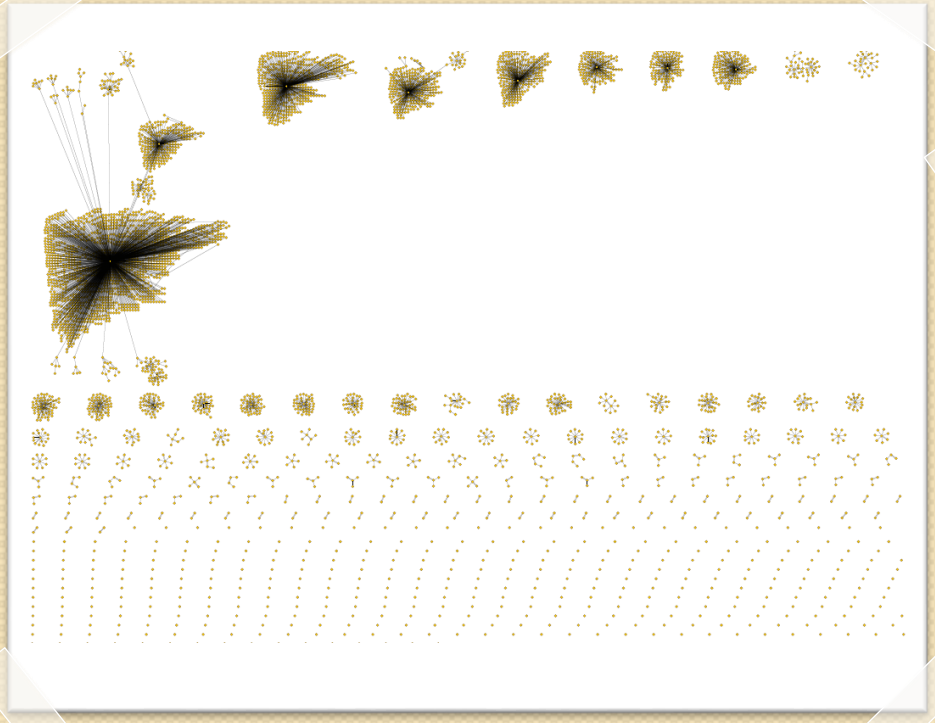
Data curation: Filtering

·Initial download contained ~270,000 candidates
·~37,000 syntactic heuristics
·~18,000 byte identical
·~5,000 not parseable
·~6,000 byte identical after OWL/XML serialisation

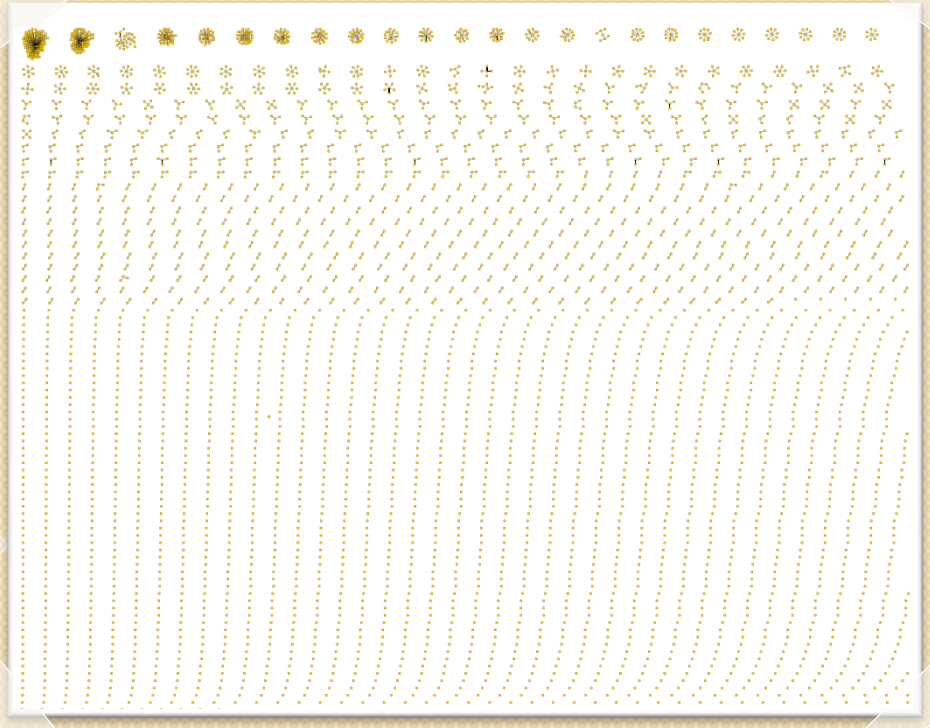
·~200,000 filtered out semi-automatically
·~5324 RDFS/OWL Full



Evaluation!?



- Evaluating not straight forward
- Cluster analysis?



The University of Manchester

Filtering is (a bit) painful

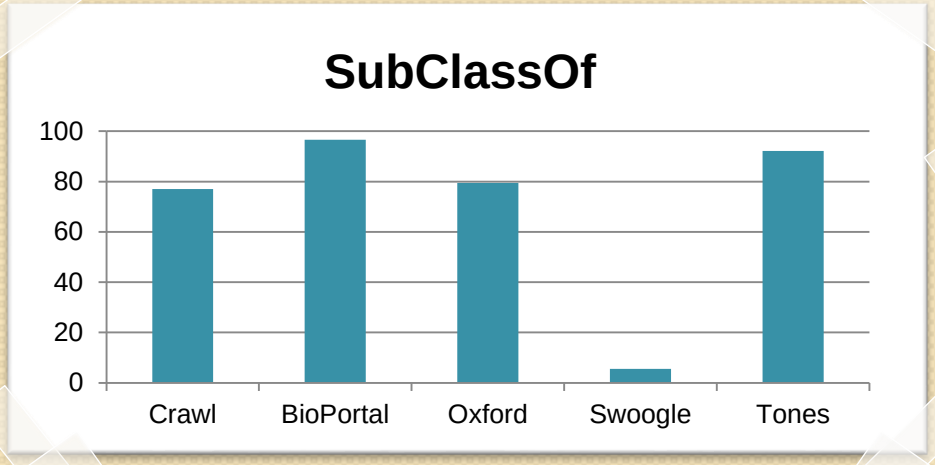
- Pairwise comparisons take a long time
- Manual inspections are
  - time consuming
  - hard to evaluate
  - …and (frustratingly) incomplete
- Automation still unsolved problem

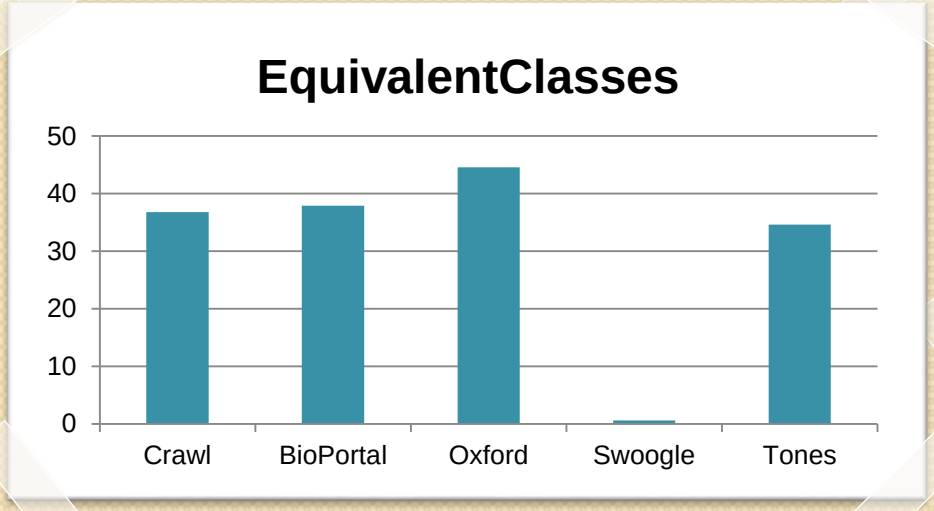
The University of Manchester

Corpus Comparison

- How does the new corpus compare to existing collections?
  - Our Crawl [Nov 2012, 4547 items]
  - BioPortal [Nov 2012, 292 items]
  - Oxford Ontology Library [April 2013, 793 items]
  - Swoogle Sample [May 2012 biased dump, 1,757 items]
  - TONES [Nov 2012, 205 items]

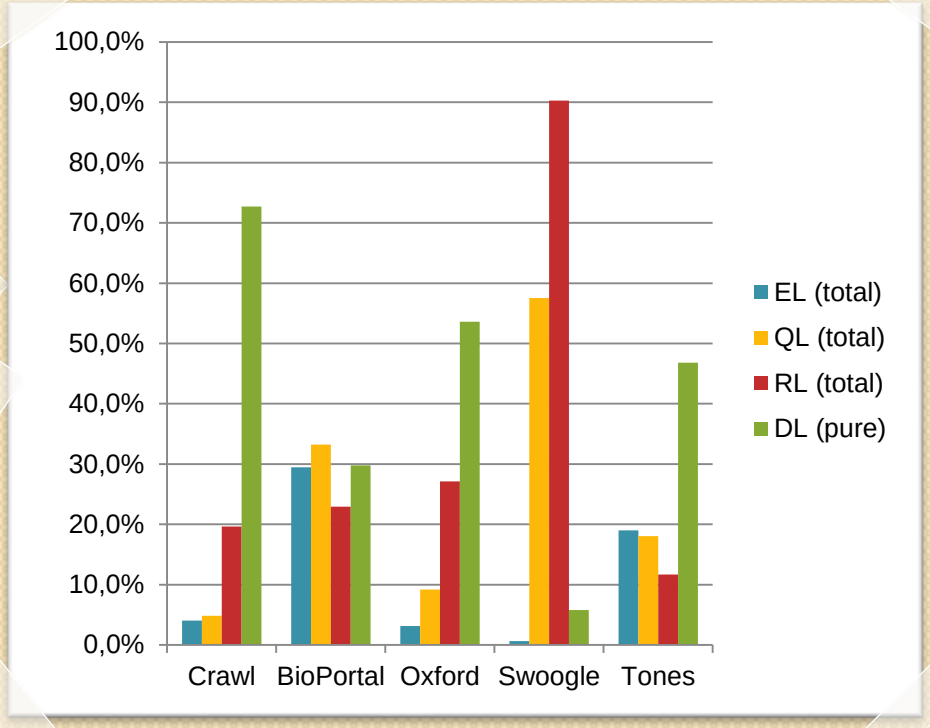
Axiom Types





More metrics in paper

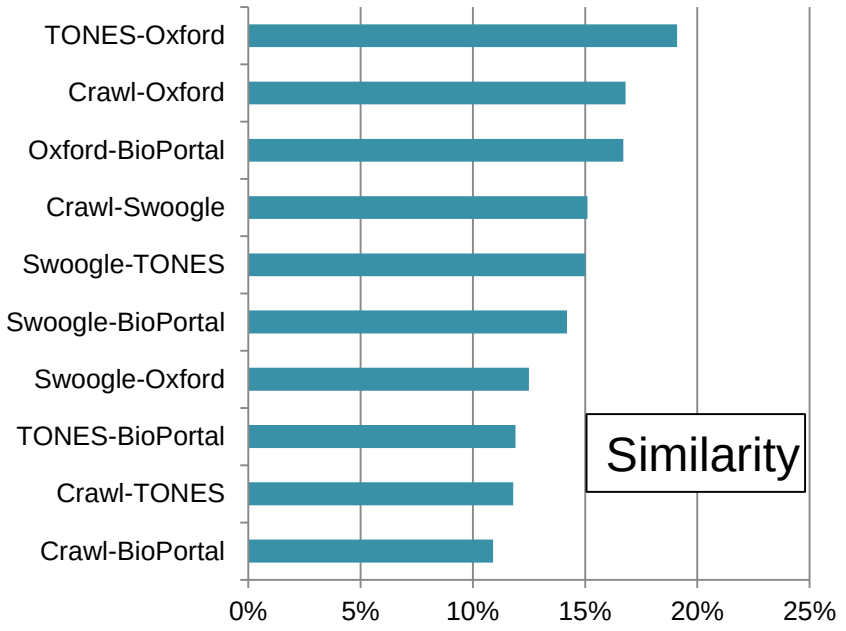
- Datatypes
- Profile Violations
- Logical axioms / entities

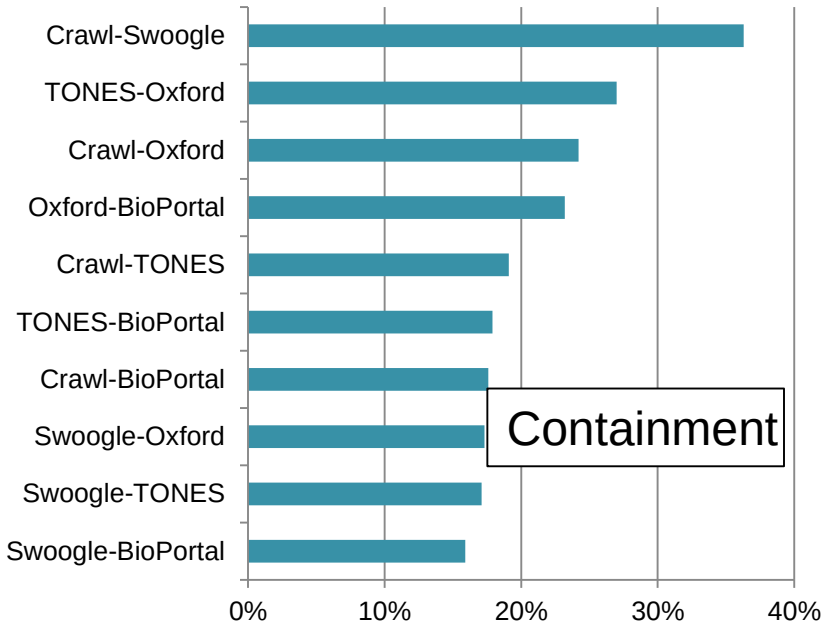


OWL Profiles

Overlap Analysis

- Similarity:
  - Overlap(Sig(O1),Sig(O2)) >=90%
- Containment:
  - Sig(O1)⊆ Sig(O2), vice versa





The University of Manchester

Current use cases

- ORE 2013 (Gonçalves et al 2013a)
  - DL 2013, fun experience and good feedback
- Benchmarking / Evaluation
  - reasoner robustness (Gonçalves et al 2013b)
  - benefits of rule refinement (Khodadadi et al 2013)
  - performance prediction using ontology features (Sazonau, 2013, MSc)
- Web surveys
  - Characterisation (Unsatisfiable Classes, Moodley 2013, ongoing)

The University of Manchester

Outlook

- Create a well *indexed* repository of ontologies
  - for benchmarking and surveys
  - dataset sharing
  - ongoing crawl
- Automate the manual cleaning
  - and/or enable cluster based sampling

Thanks

matentzn@cs.man.ac.uk
bparsia@cs.man.ac.uk
bails@cs.man.ac.uk

http://owl.cs.manchester.ac.uk/owlcorpus/

SWSA
SEMANTIC WEB SCIENCE ASSOCIATION

CDT CS
Centre for Doctoral Training in Computer Science

MANCHESTER
1824