FINDWISE
SEARCH DRIVEN SOLUTIONS

Hydra

An Open Source Document Processing Framework

Joel Westberg

About Findwise

- Founded in 2005
- Offices in Sweden, Denma
  Norway and Poland
- 82 employees (May 2012)
- Our objective is to be a leading provider of **Findability** solutions utilising the full potential of search technology to create customer business value

FINDWISE
SEARCH DRIVEN SOLUTIONS

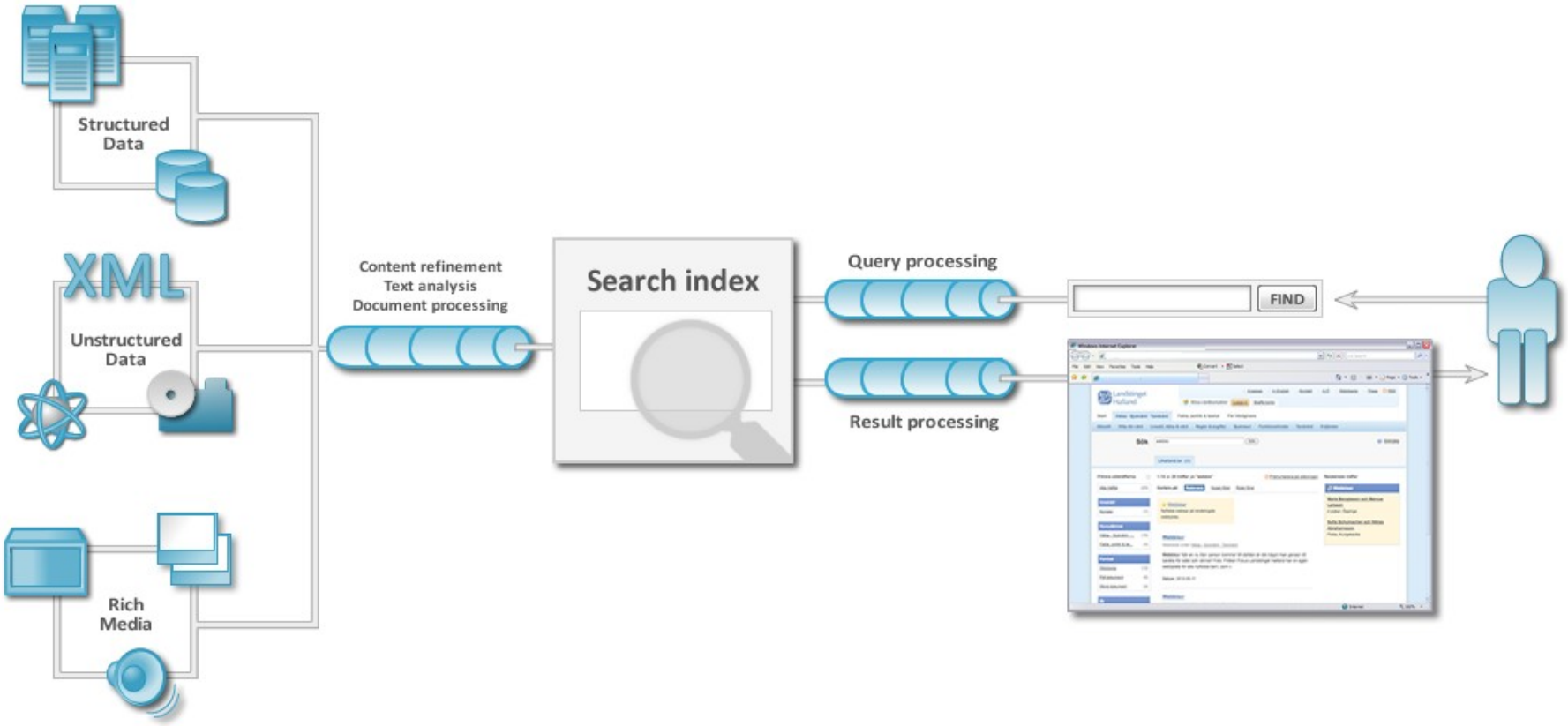

Stena
TINE
INFOMEDIA
VÄSTRA GÖTALANDSREGIONEN
protects what's good
Tetra Pak
SECURITAS
mktmedia
SVENSKA SPEL
FÖRSVARSMAKTEN
Johnston Press plc
UPPSALA UNIVERSITET
SVERIGES RIKSDAG
infonor systems
SAAB
SEB
AstraZeneca
Vetenskapsrådet
Sida
AMIRA
rushes
NORDJYSKE.DK
Swedbank

Creating search-driven Findability solutions based on market-leading commercial and open source search technology platforms:

- ü Autonomy IDOL
- ü Microsoft (SharePoint and FAST Search products)
- ü Google GSA
- ü IBM ICA/OmniFind
- ü LucidWorks
- ü Apache Lucene/Solr
- ü Ela
- ü and more...

FINDWISE
SEARCH DRIVEN SOLUTIONS

Generic Search Architecture



FINDWISE
SEARCH DRIVEN SOLUTIONS

Connecting source to search

Garbage in, garbage out. But what about unstructured data in?

- Flat data is richer than it appears
- Don’t discard information too soon!

The unstructured structured data paradox

Example: News articles

Plain text that contains invaluable metadata for search, such as:

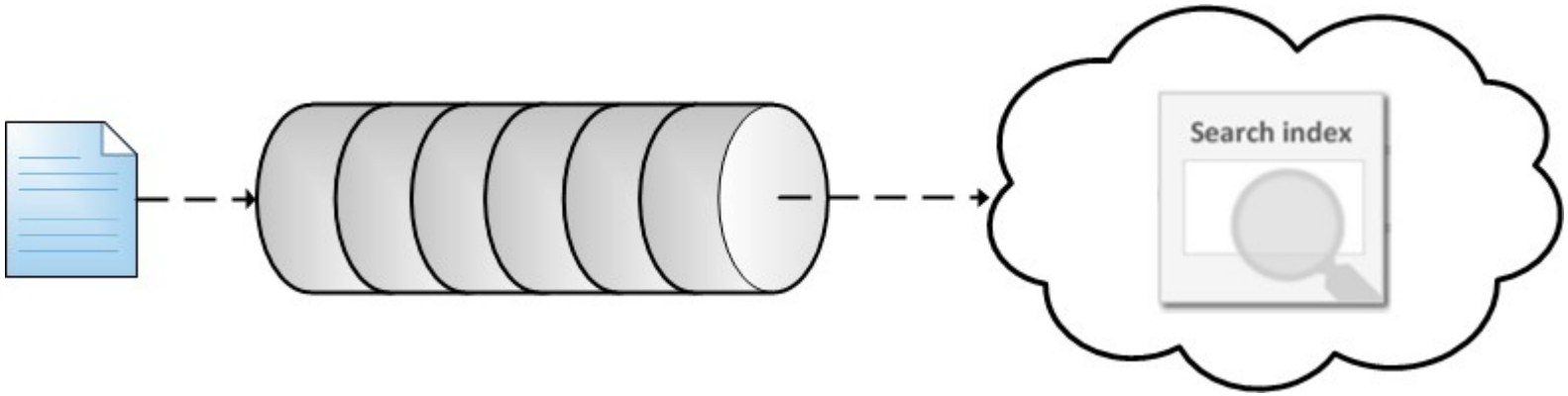
- Title
- Author byline
- Lead paragraph

Enrichment and structuring possibilities

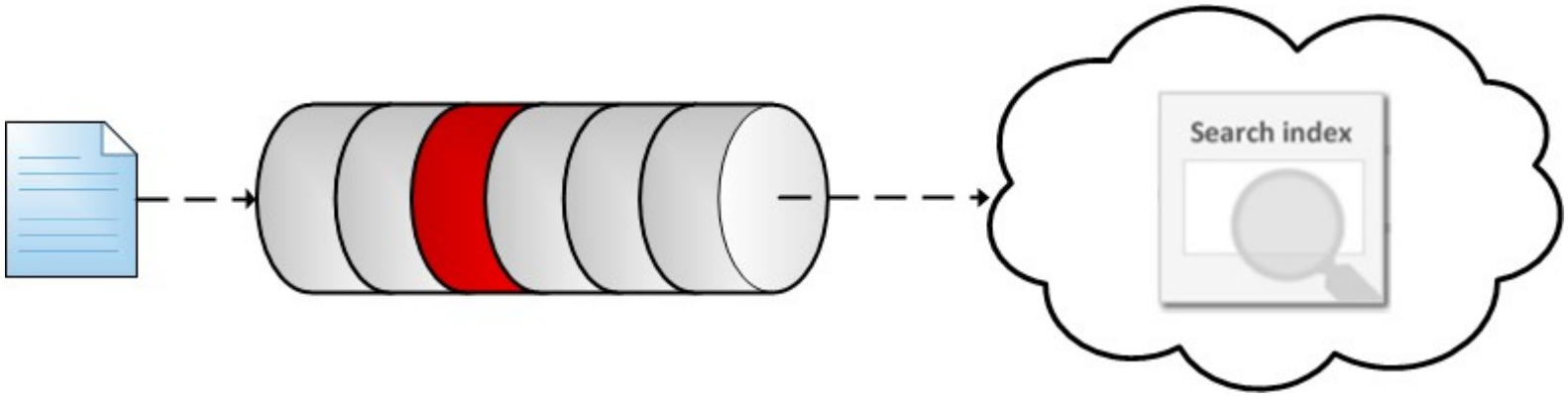
- Enrich your documents with metadata, to power your search
  - Language detection
  - Sentiment analysis
  - Headline extraction
  - Regular expression matching and extraction
- Filter out unwanted documents
- Collect statistics
- Export to Staging environments

FINDWISE
SEARCH DRIVEN SOLUTIONS

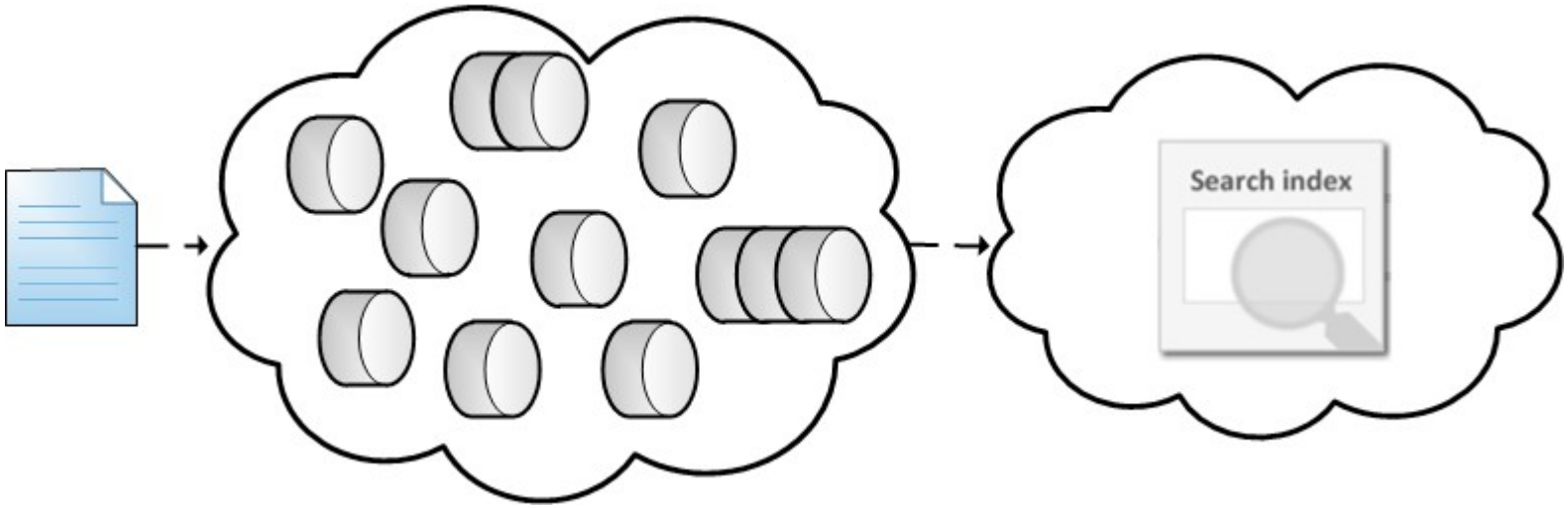
Classic Architecture



Classic Architecture



The Hydra Architecture



Main Design Objectives

Scalability

- Horizontally scalable central repository
- Independent processing nodes

Failiure tolerant

- Failiure of a stage affects only a single document
- Failiure of a node affects at most ***n*** documents
- Failiures can be automaticly detected

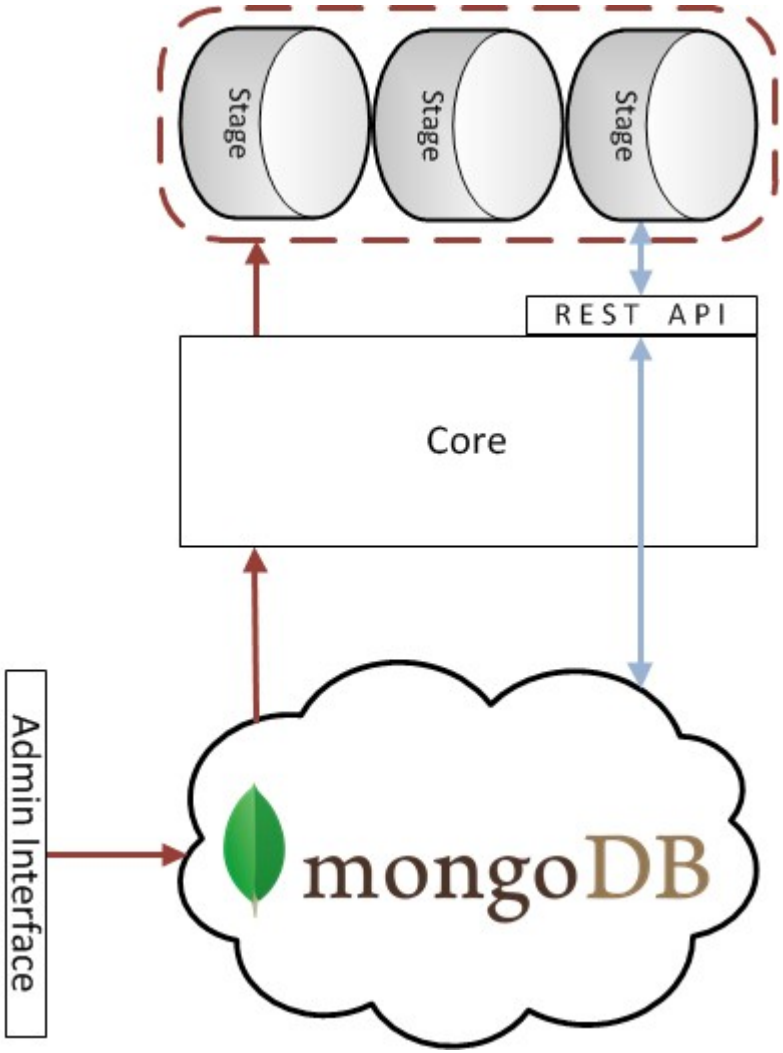
Robustness

- Independent stages

Development ease

- Debug stages from IDE against actual data
- Allow test driven pipeline development

The Hydra Architecture



Writing a Stage - Example

```
@Stage(description="This is a Simple Writer")
public class SimpleWriter extends AbstractProcessStage {
    @Parameter(description="Name of field to write value to")
    private String field;
    @Parameter(description="Value to write")
    private Object value;

    @Override
    public void process(LocalDocument doc) throws ProcessException {
        doc.putContentField(field, value);
    }

    @Override
    public void init() throws RequiredArgumentMissingException {
        if(field==null) throw new RequiredArgumentMissingException("field is missing");
    }
}
```

Hadoop/Big Data integration

Usecases for document enrichment

- Pagerank
- Analytics

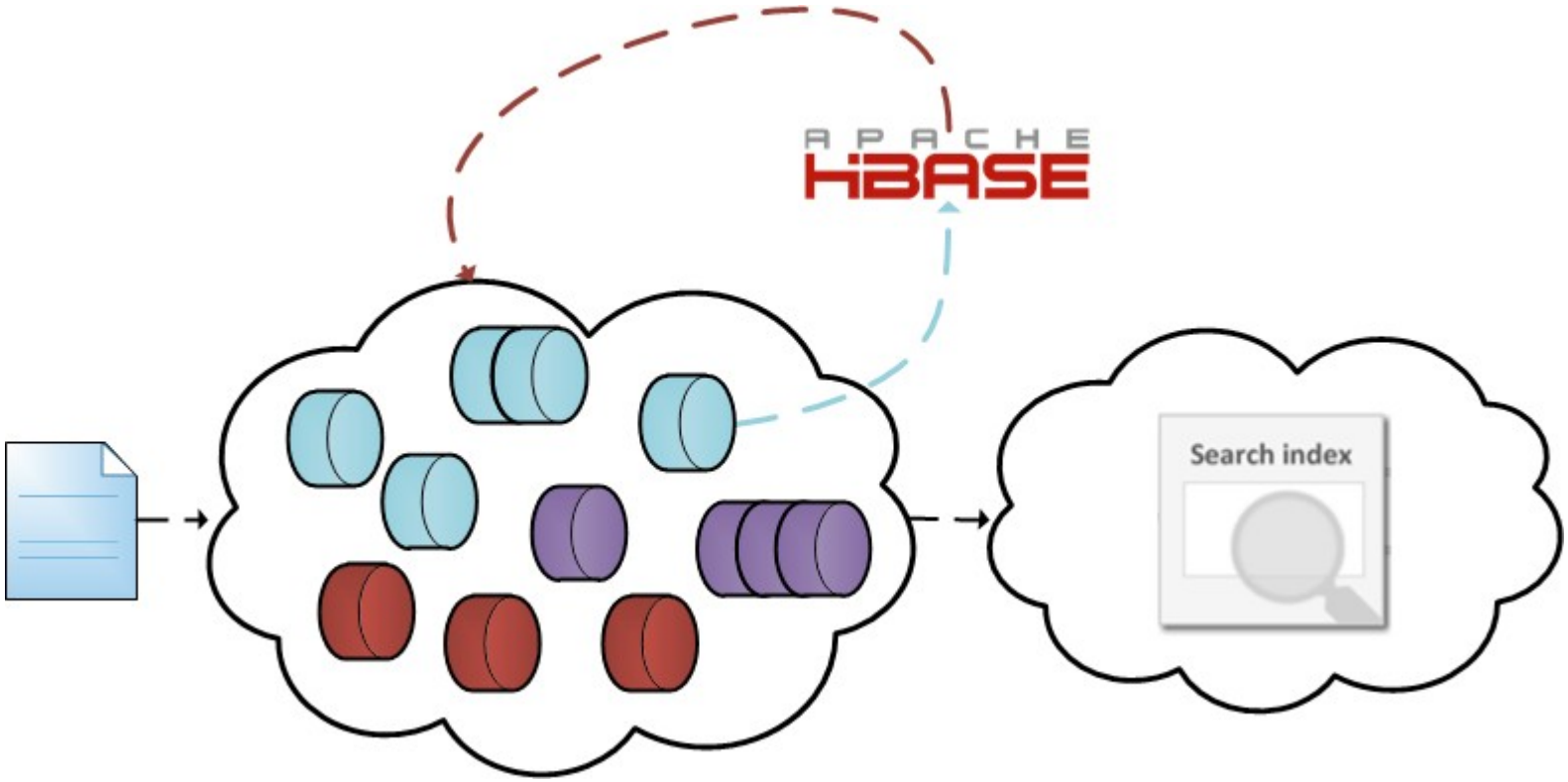
Hadoop & Map/Reduce advantages

- Huge scalability
- Ability to work on entire document set at once

Hadoop & Map/Reduce drawbacks

- Batch processing
- Time-to-index

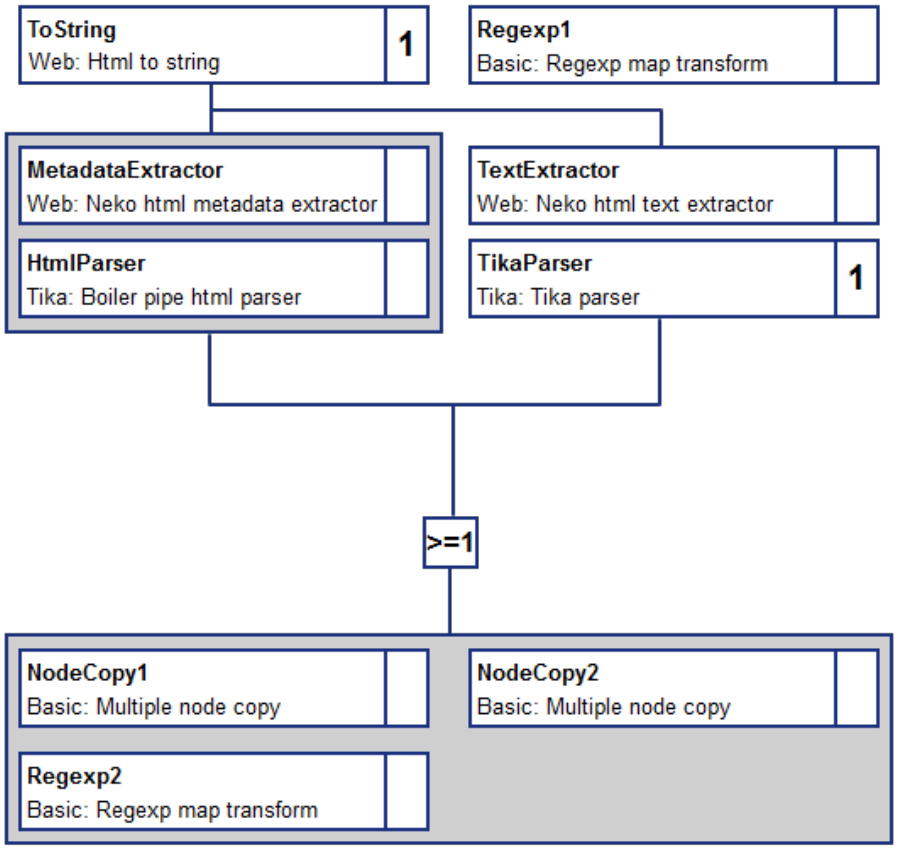
Hadoop/Big Data integration

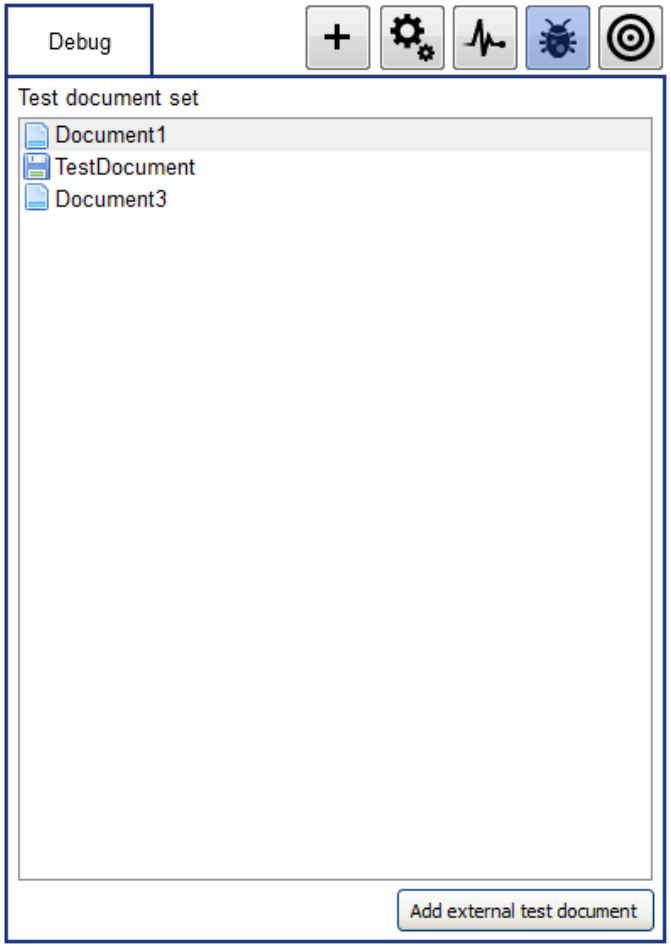


Blue – First round of indexing only
Red – Second round of indexing
Purple – All documents

FINDWISE
SEARCH DRIVEN SOLUTIONS

Future Configuration UI





FINDWISE
SEARCH DRIVEN SOLUTIONS

Open Source initiative

- Other committers
- The role of Findwise

For more information:

- http://www.findwise.com/hydra
- http://findwise.github.com/Hydra

- Email: joel.westberg@findwise.com

FINDWISE
SEARCH DRIVEN SOLUTIONS

FINDWISE
SEARCH DRIVEN SOLUTIONS

Questions?

FINDWISE
SEARCH DRIVEN SOLUTIONS

Thankyou!

Joel Westberg
joel.westberg@findwise.com