

## Architecture of an IR Engine

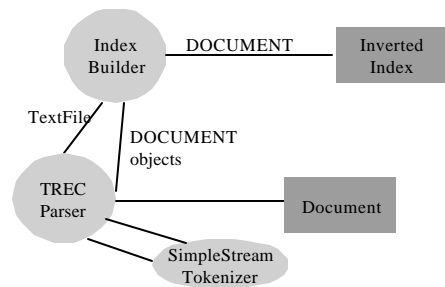
## Overview

- IR engine takes documents, indexes them, then accepts queries.
- Most IR systems use a structure called an *inverted index* to index documents.

## Requirements

- Scalability
  - Must handle large document collections
- Index Efficiency
  - Must build indexes in a reasonable amount of time
- Query Efficiency
  - Queries must run fast
- Query Effectiveness
  - Result set must be relevant

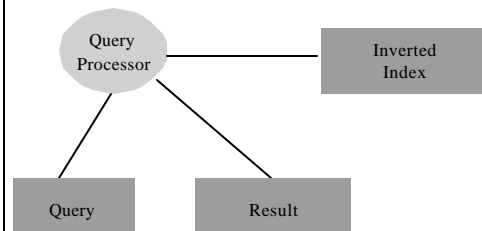
## Simple Indexing Architecture



## Indexing Objects

- IndexBuilder
  - drives the indexing process
- Parser
  - takes an input file, returns a list of document objects in the file
- Tokenizer
  - returns the next token and the token type in the input file
- Document
  - object that stores all the relevant information about a document
- Inverted Index
  - stores the entire inverted index

## Query Objects



## Object Descriptions

- **Query Processor**
  - drives the execution of a query
- **Query**
  - terms in the query
- **Result**
  - a single element of the answer set

## Sample Document

```
<DOC>
<DOCNO> FT911-3 </DOCNO>
<HEADLINE>
FT 14 MAY 91 / International Company News: Contigas plans DM900m east
German project
</HEADLINE>
<DATELINE> BONN </DATELINE>
<TEXT>
CONTIGAS, the German gas group 81 per cent owned by the utility Bayernwerk,
said yesterday that it intends to invest DM900m in the next
four years to build a new gas distribution system in the east German state
of Thuringia.
Reporting on its results for 1989-1990 the company said that the dividend
would remain unchanged at DM8.
Sales rose 9.4 per cent to DM3.37bn, but post-tax profit fell slightly from
DM31.3m to DM30.7m.
</TEXT>
</DOC>
```

## Index Building Sample Document

```
- Adding document
• about to add these terms to index:

• slightly 1
  unchanged 1
  invest 1
  intends 1
  3m 1
  gas 2
  81 1
  dm8 1
  dm31 1
  yesterday 1
  dm30 1
  remain 1
  company 1
  dm3 1
  reporting 1
  rose 1
  contigas 1

-
• group 1
• 37bn 1
• 4 1
• utility 1
• east 1
• sales 1
• dividend 1
• years 1
• owned 1
• build 1
• post-tax 1
• fall 1
• dm900m 1
• cent 2
• distribution 1
• profit 1
• german 2
• thuringia 1
• 7m 1
• system 1
• bayernwerk 1
• 1989-1990 1
• state 1
```