The limits of my language mean the limits of my world

- Wittgenstein

PostgreSQL's Full Text Search

Richard Huxton - dev@archonet.com

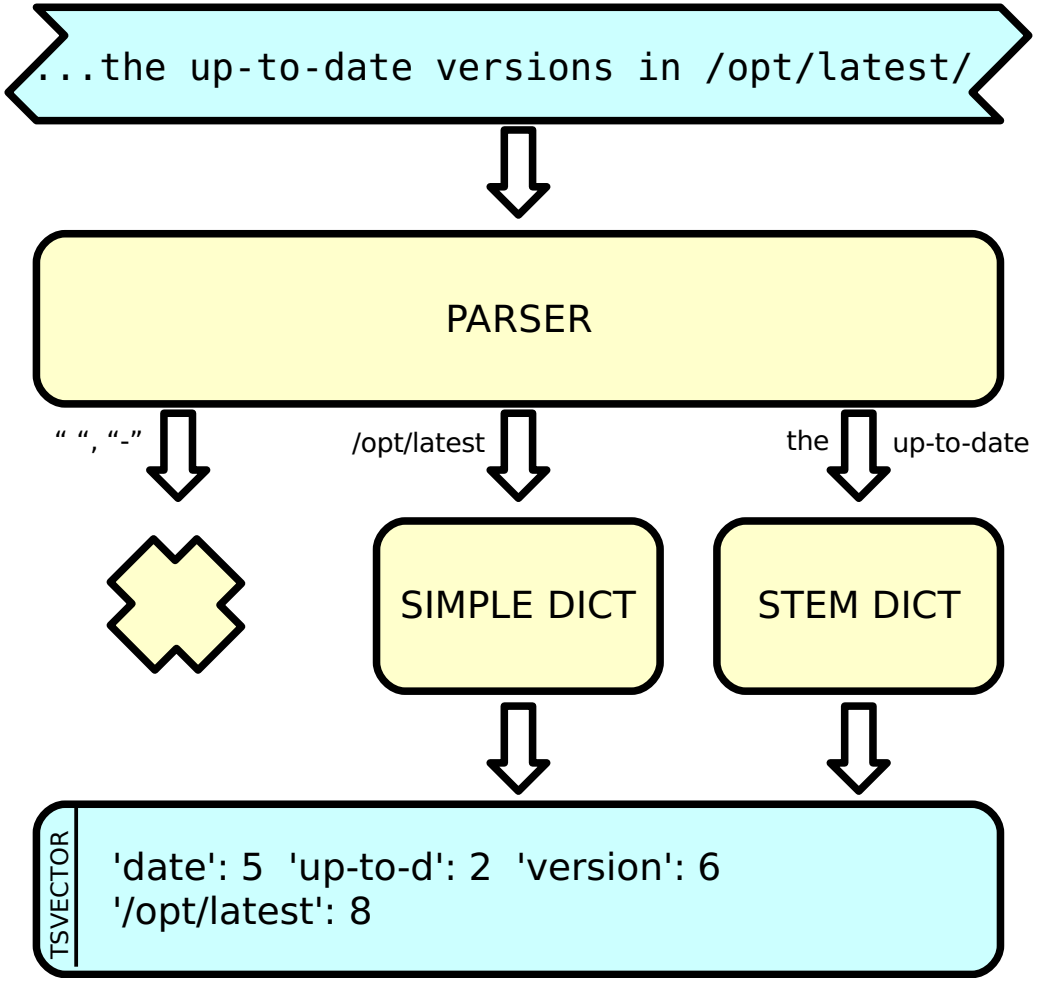
TSearch Is...

- PostgreSQL's full-text search
- Contrib module since 7.2
- In core with 8.3
- Perceived as complicated...

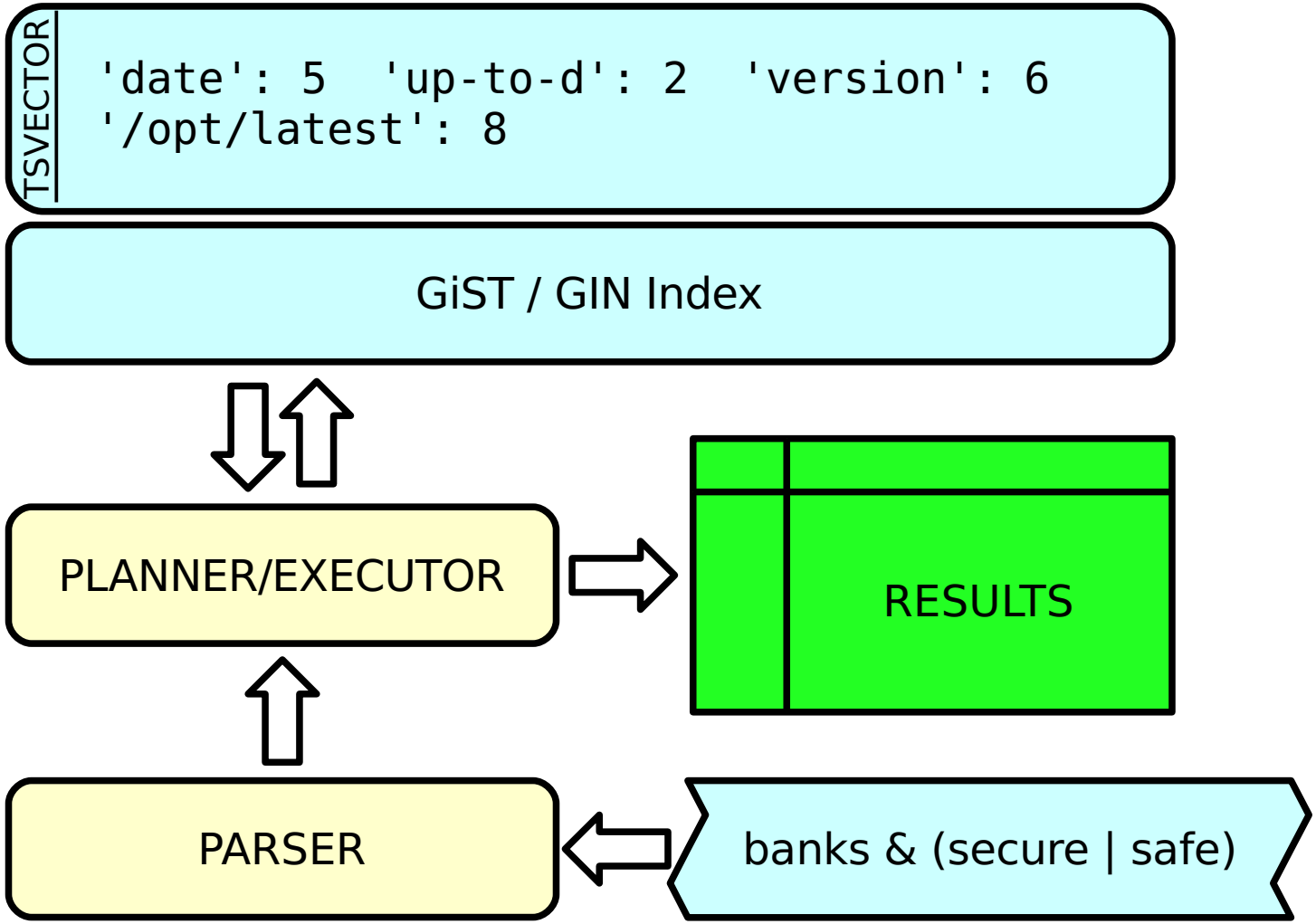
Elements

1. Configurations
2. Parsers
3. Dictionaries
4. tsvector type
5. GiST / GIN indexes
6. tsquery type

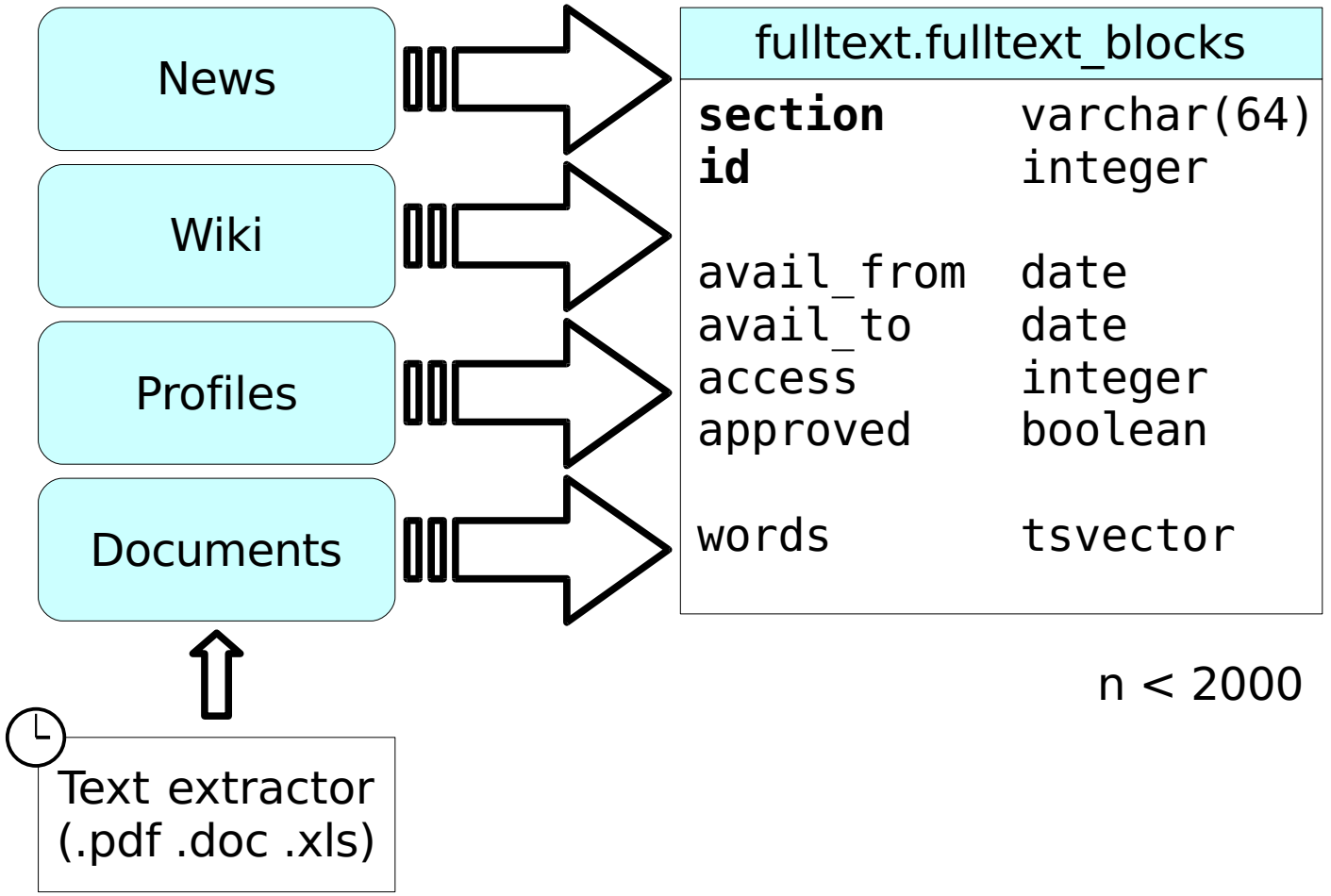
Building a tsvector



Querying a tsvector



Industry Assoc.n Website



The Query

```
SELECT
    section,
    id,
    ts_rank_cd('{0.1, 0.2, 0.4, 1.0}',words,query) AS rank
FROM
    ftext.ftext_blocks,
    to_tsquery('cust', 'safety & electrical') AS query
WHERE
    CURRENT_DATE BETWEEN avail_from AND avail_to
    AND approved
    AND access >= 0
    AND words @@ query
ORDER BY
    rank DESC
```

The Query

```
SELECT
    section,
    id,
    ts_rank_cd('{0.1, 0.2, 0.4, 1.0}',words,query) AS rank
FROM
    ftext.ftext_blocks,
    to_tsquery('cust', 'safety & electrical') AS query
WHERE
    CURRENT_DATE BETWEEN avail_from AND avail_to
    AND approved
    AND access >= 0
    AND words @@ query
ORDER BY
    rank DESC
```

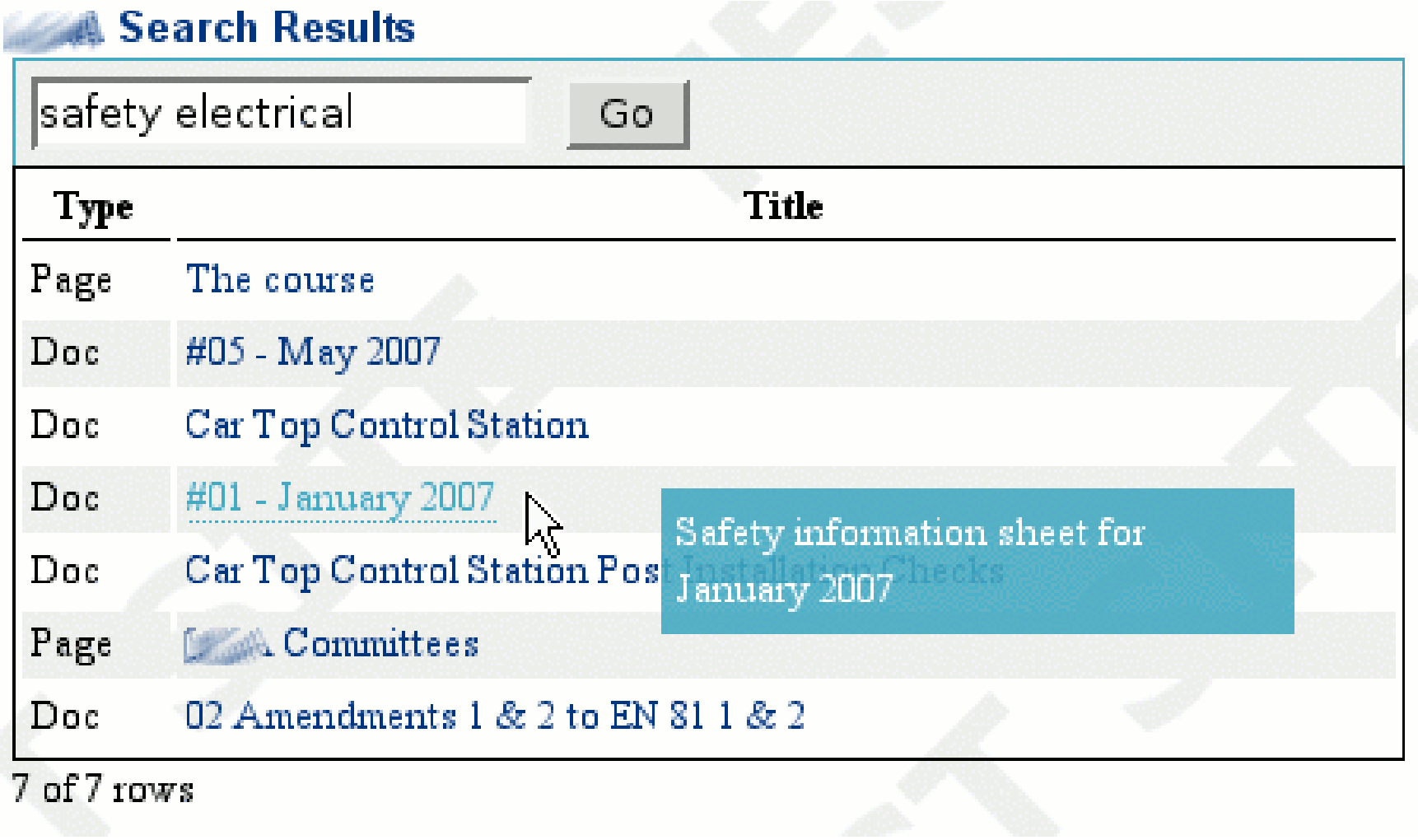
The Query

```
SELECT
    section,
    id,
    ts_rank_cd('{0.1, 0.2, 0.4, 1.0}',words,query) AS rank
FROM
    ftext.ftext_blocks,
    to_tsquery('cust', 'safety & electrical') AS query
WHERE
    CURRENT_DATE BETWEEN avail_from AND avail_to
    AND approved
    AND access >= 0
    AND words @@ query
ORDER BY
    rank DESC
```

The Query

```
SELECT
    section,
    id,
    ts_rank_cd('{0.1, 0.2, 0.4, 1.0}',words,query) AS rank
FROM
    ftext.ftext_blocks,
    to_tsquery('cust', 'safety & electrical') AS query
WHERE
    CURRENT_DATE BETWEEN avail_from AND avail_to
    AND approved
    AND access >= 0
    AND words @@ query
ORDER BY
    rank DESC
```

Results


Search Results
safety electrical
Go
Type
Title
Page
The course
Doc
#05 - May 2007
Doc
Car Top Control Station
Doc
#01 - January 2007
Doc
Car Top Control Station Post Installation Checks
Page
Committees
Doc
02 Amendments 1 & 2 to EN 81 1 & 2
Safety information sheet for January 2007
7 of 7 rows

Gotchas – Config/Parser

- Default vs non-default configs
- Configuration changes
- Word boundaries
  - “123,000” => “123” “000”
  - "U.K." => "u.k"
  - "either/or" = file-path

Gotchas – Dictionaries

- Custom stopwords, use accept=false
- No “oddities” dictionary
  - custom regexp dict available
- Files are external
- Files must be UTF-8

Gotchas - Indexes

- GiST is smaller but slower
- GIN is larger but faster
- GiST updates faster
- GIN + weights = @@@

Gotchas – Searches

- No phrase-search
- Selectivity estimates - fixed
- ts_rank - score is per-doc

Debugging

- ts_debug()
  - configuration
- ts_stat()
  - word frequencies

Summary

- Know which config you are using
- Prefer GIN over GiST
- Except for rapid updates / weight search
- “Scrub” messy inputs
- Move filtering to subquery
  - ts_rank, ts_headline are expensive

Questions?

?