



Hasso  
Plattner  
Institut

IT Systems Engineering | Universität Potsdam

## Search Engines – Exercise

## Smart Data Extraction

Franz Goerke  
Thomas Berger

# Our guideline

2

- Requirements:
  - Extraction of tables / structured data
  - Analyze table structures
  - Enable search for tables
  - Tableindex creation
  
- Enhancements:
  - Get relation among tables
  - Link resolving
  - Statistics about table contents

# New Database structure

3

## ■ New Tables

- WikiTables
- ArticleTitles

TableID	ArticleID	Content
ArticleID	Title	

## ■ New Indices

- Inverted Table Index
- Linked Table Index
- Attribute Table Index

TableID	ArticleID	Content
TableID	LinkedTables	BacklinkedTables
Attribute	Tables	

# Relations among tables

4

## ■ Approaches

### □ Link analysis

- Tables contain links to articles
  - » Not very useful
  - » Links mostly refer to just one single term

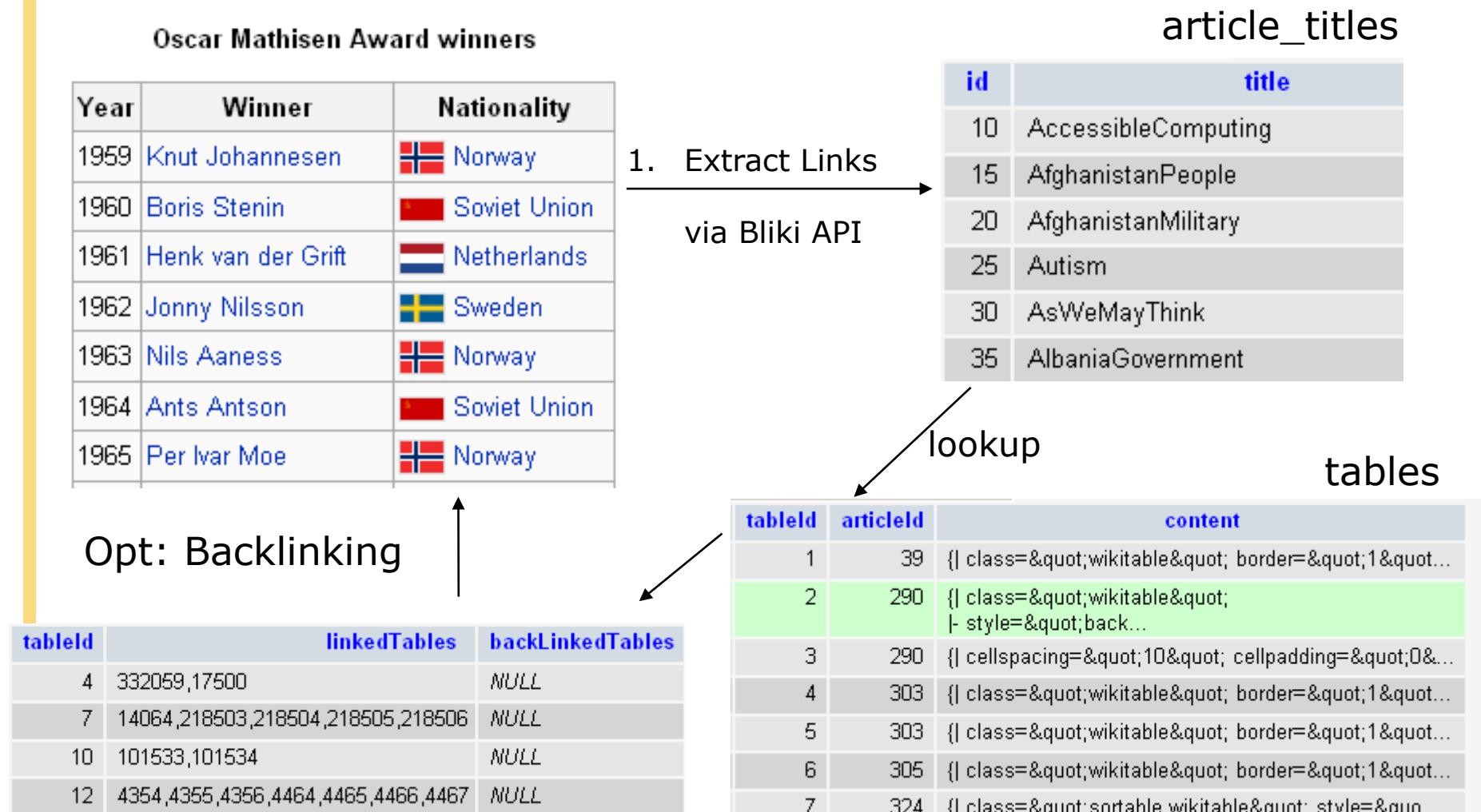
### □ Content analysis

- For every term query inverted index
- Count occurrences of tables for every word
- Use top results
- Very useful

### □ Attribute analysis

- Not useful
- Most attribute names are equal (name, year, language)

# Linked tables



Thanks for your Attention

6

# Demo & Questions