Harvest Research Data with Smart Harvesting II

Using OXPath for Web Data Extraction

Web Data is Research Data

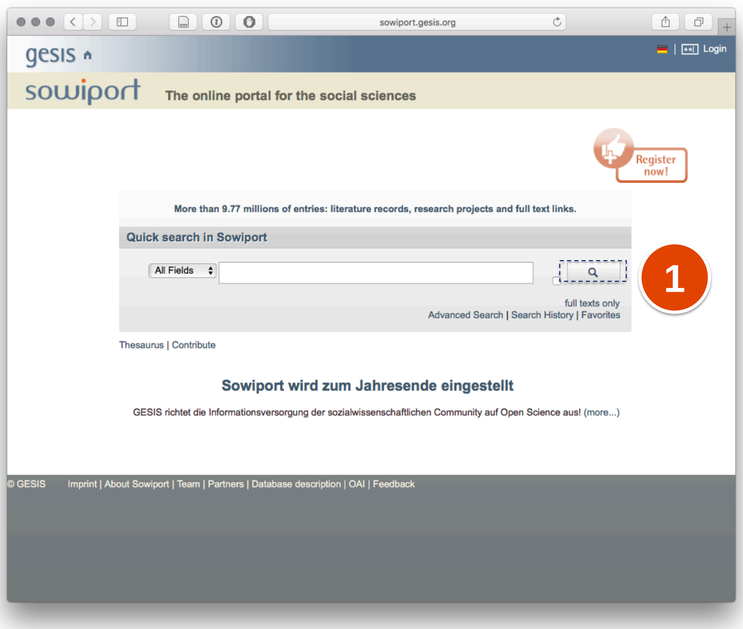
- **Original use case:** publication metadata harvesting (dblp, SSOAR)
- Other data sources:
  - **Social Media Data** – Twitter (beyond API access), Blogs, ...
  - **Discussion forums** and **comment sections**
- **Potential usage scenarios:**
  - Natural Language Processing
  - Machine Learning
  - Information Retrieval
- Our own research: **IR test collections**

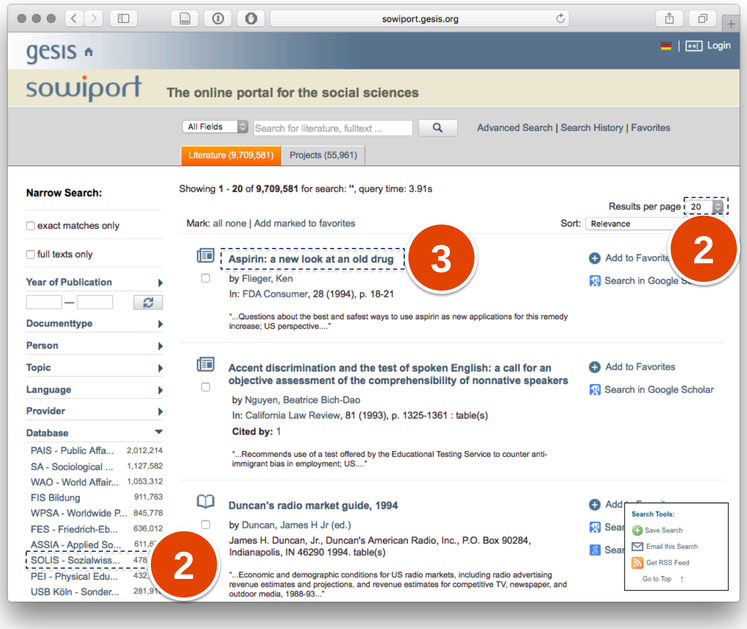
Web Scraping for Non-Programmers

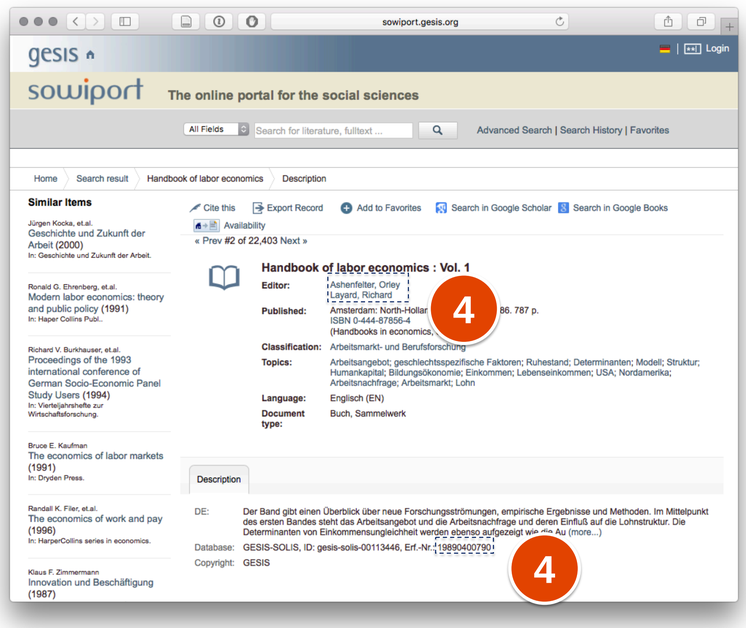
- Researchers shouldn't need **software development skills** to get web data, but:
  - Manual work is **tedious** and **time-consuming**
  - Available tools are **expensive** or **limited**
- **Basic HTML/XML knowledge** is easier to acquire

Scraping Web Data with OXPath

- **OXPath: Free** and **open source** web data extraction language
  - Extends XPath
  - Java-based library
  - Developed in Oxford, by Furche et al.
- **Adds additional functionalities** to XPath, like**:**
  - Actions → clicking, typing, mouse movements
  - Iterations → e.g. for navigating through paginated content
  - Extraction markers → tree-like output of extracted data
- **Light-weight and easy to use:**
  - In contrast to Scrapy et al. no programming knowledge is required
  - Pretest with librarians and LIS students showed that non-programmers are able to successfully write their first OXPath wrapper after a 2-hour tutorial
  - Domain knowledge and LIS background are more helpful to write wrappers than a deep understanding of programming languages…







Visit start page and click on search button to get full result list (1)

Narrow down result list by facet and set size to 100 results per page (2)

Iteratively click 'next' button to get all results

Extract a record's title and click to go to details page (3)

Open a new <record> tag in the output for each result record

Extract information from table cells (4)

```
doc('http://sowiport.gesis.org/')
  /descendant::field()[2]/{click /}
    //div[@id='facets']//a[contains(.,'SOLIS')]/{click /}
    //*[@id='limit']/option[contains(., '100')]/{click /}
    /(//*[contains(@title, 'next')]/{click /})*
      //*[contains(@class,'record')]:<record>
        [.//a[@class~='title'][./b:<title=normalize-space(.)>]/{click /}
          /. [? .//*[@id='detailed_view_metadata']//table//td
            [ ./preceding-sibling::td[contains(.,'Editor:')]]//a
              :<editor=normalize-space(.)>]
          [? .//*[@class='recordsubcontent']//table//td
            [ ./preceding-sibling::th[contains(.,'Database:')]]
              :<id=substring-after(normalize-space(.), 'Acquis. id: ')>]
        ]
```

Example Output: Result

```
<?xml version="1.1" encoding="UTF-8"?>
<results>
    <record>
        <title>Ausbildung in Betriebsinformatik</title>
        <editor>Pressmar, Dieter B.</editor>
        <id>19890400793</id>
    </record>
    <record>
        <title>Handbook of household surveys</title>
        <id>19890400792</id>
    </record>
    <record>
        <title>Hochschule - Beruf - Gesellschaft</title>
        <editor>Gorzka, Gabriele</editor>
        <editor>Heipcke, Klaus</editor>
        <editor>Teichler, Ulrich</editor>
        <id>19890400830</id>
    </record>
    ...
</results>
```

Future Use Cases and Outlook

- Expand our approach to support other **DL systems**
- Create useful **tool set** around OXPath, e.g. **scheduling** and **monitoring** of harvesting processes
- Generate **structured web data sets**
- Using hand-crafted scrapers **maintained by non-programmers**

Mandy Neumann, Philipp Schaer: TH Köln (University of Applied Sciences), Cologne, Germany
Christopher Michels, Michael Ley: Trier University, Trier, Germany
Nadine Dulisch, Brigitte Mathiak: GESIS – Leibniz-Institute for the Social Sciences, Cologne, Germany