TECHNISCHE UNIVERSITÄT DRESDEN

Department of Computer Science Institute for Software and Multimedia Technology, Software Technology Group

Academic Skills in Software Engineering (ASiSE)

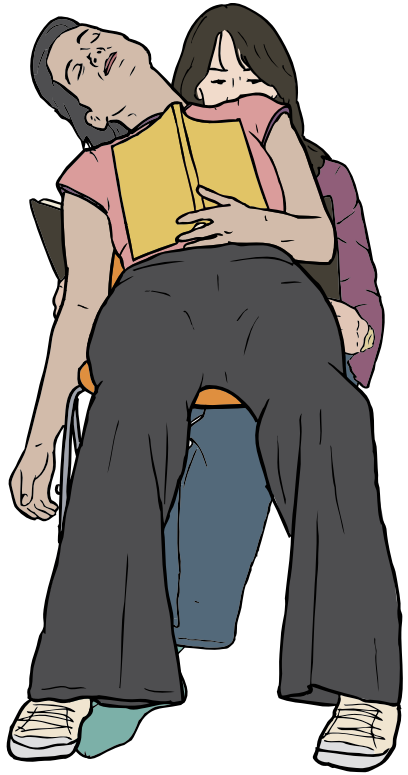
Bulk Collection, Filter, and Classification of Literature

Exercise
Tuesday, 5. DS, APB/E001

Thomas Kühn (thomas.kuehn3@tu-dresden.de)

Scientific Research
Daily Business

Reading



Writing

Organizing

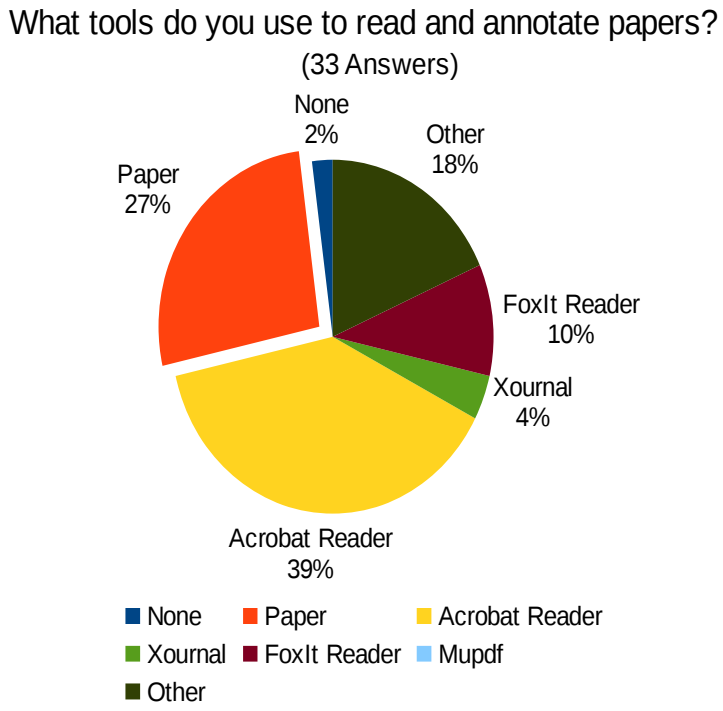
Images from OpenClipart.org (Creative Commons by Steve Lambert)

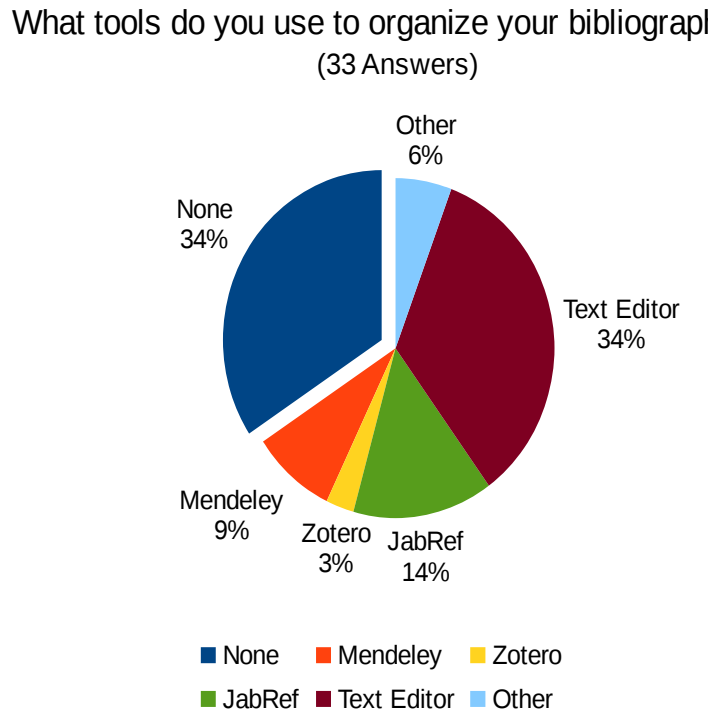
Common Tasks

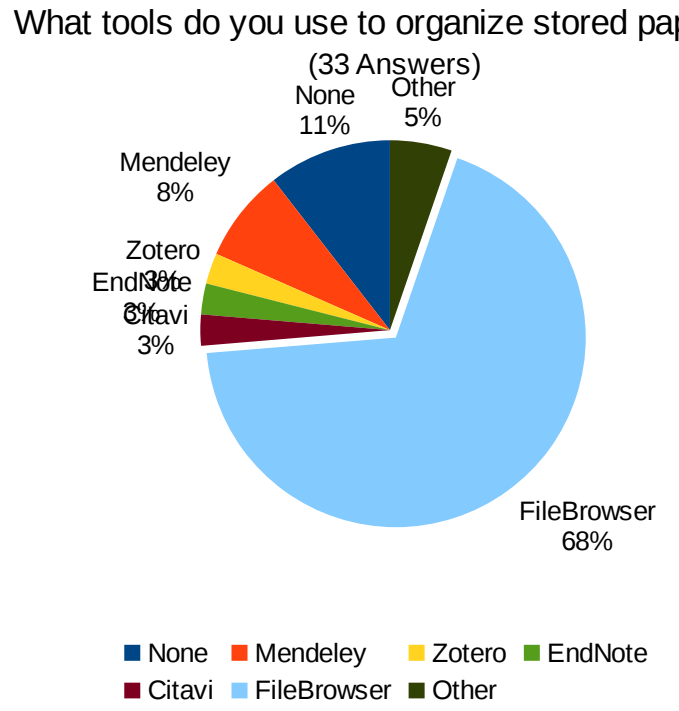
- Find relevant / related publications
  - Query scientific search engines
  - Look up *BibTex* for specific publications from the web
- Investigate found publications
  - Skim papers
  - Make notes and hints
  - Organize downloaded files
  - Maintain a corresponding **bibliography** of *BibTex* entries

A Small Survey

- Q1:What tools do you use to read and annotate papers?
- Q2:*What tools do you use to organize your bibliography?*
- Q3:*What tools do you use to organize stored papers?*

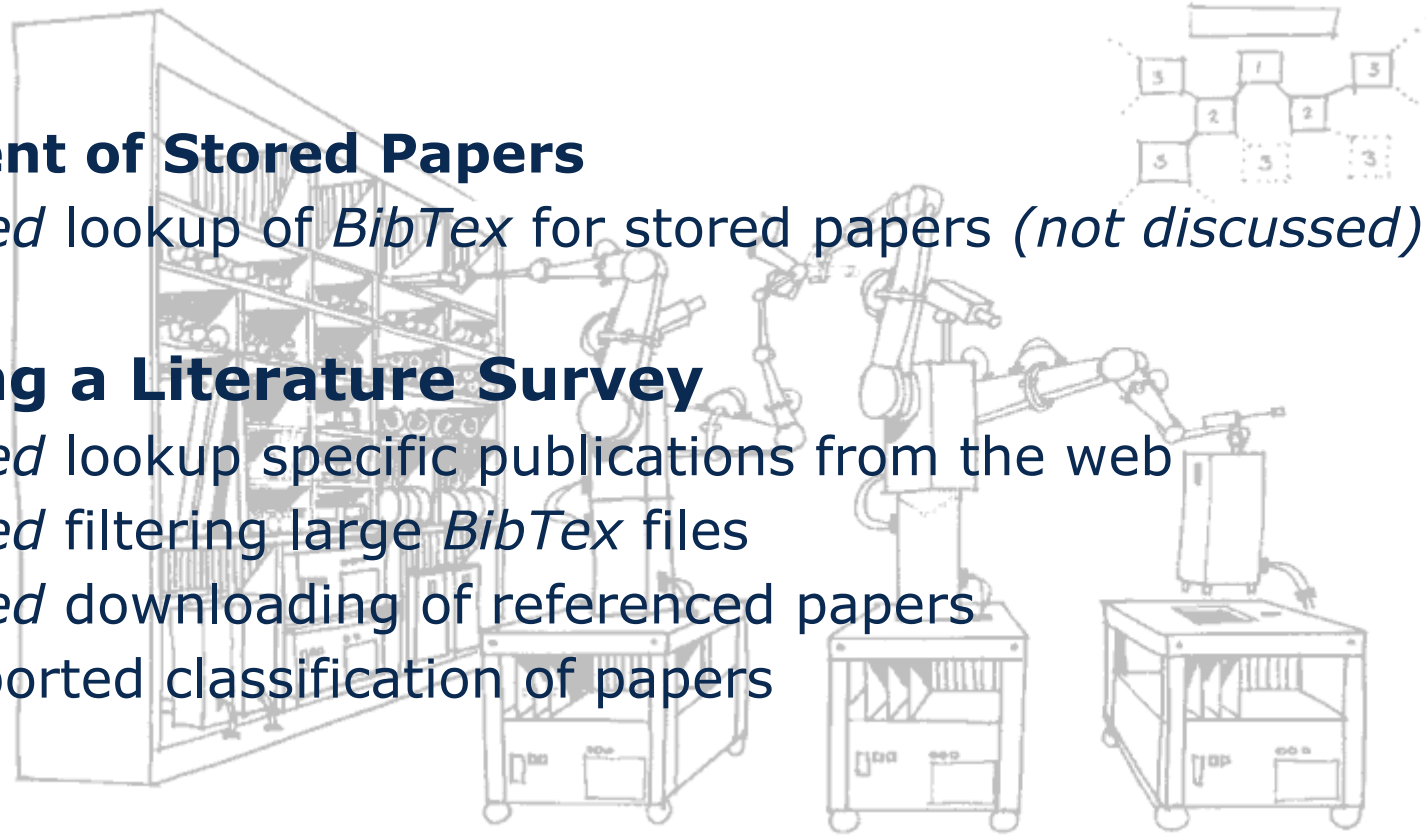






Common Tasks

- Management of stored papers
  - Search text fragments in papers
  - Look up *BibTex* for stored papers
- Conducting a literature survey
  - Look up *BibTex* for specific publications from the web
  - Filtering large *BibTex* files
  - Downloading papers
  - Classifying found papers



Management of Stored Papers

- *Automated* lookup of *BibTex* for stored papers *(not discussed)*

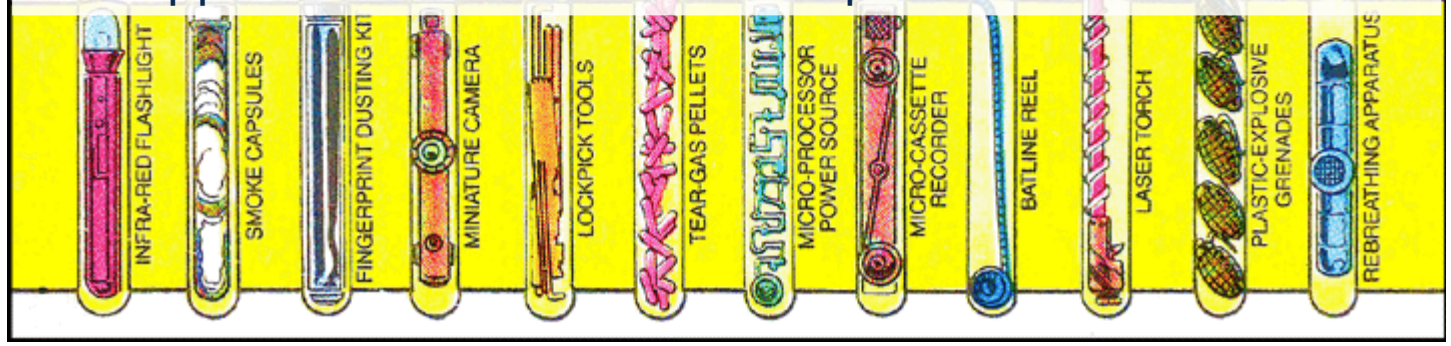
Conducting a Literature Survey

- *Automated* lookup specific publications from the web
- *Automated* filtering large *BibTex* files
- *Automated* downloading of referenced papers
- Tool supported classification of papers

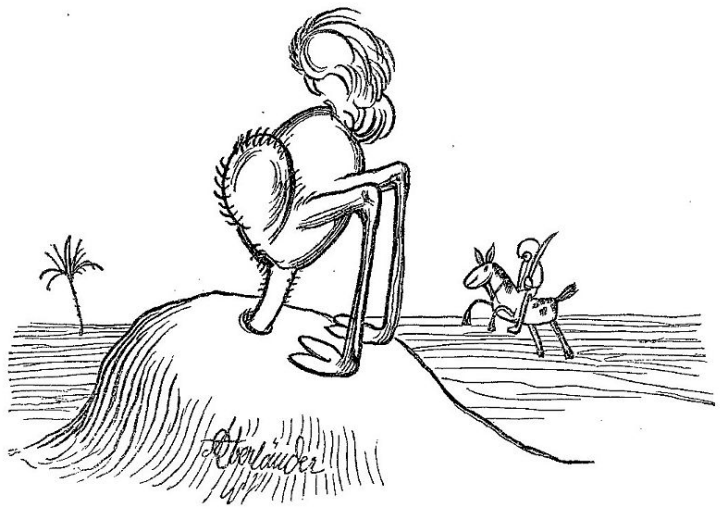
Picture by Nasa (public domain)

Organizing Research
What we Need?

- ***getbibtex.rb***†
  Fetches bibtex entries for stored papers
- ***gsresearch.rb***†
  Collects bibtex entries from Google Scholar
- ***bibfilter.rb***
  Filters large BibTex files by various criteria
- ***gsdownload.rb***†
  Downloads all files referenced by a BibTex files
- ***SLR-toolkit***
  Supports classification of multiple BibTex items



Organizing Research
† Disclaimer



Adolf Oberländer (public domain)

- Never use these scripts in jurisdictions, which prohibit automated use of Google Scholar
  - See Google's terms of Use
- Do not use these scripts to attack google services
- These tools are only for research purpose
- *„I would pay for using a Google Scholar API“*

Automated Managment

- Find naming schema for stored publication

  <Full Name of First Author>_<Full Title>.pdf
  (e.g.: *Charles W Bachman_Data Structure Diagrams.pdf*)
- Keep all documents in one folder (e.g.: *library/*)
- Use author's last name for subfolder (e.g.: *library/Bachman/*)

Steps

1. Automated sorting of new files into subfolders

   ```
   $ ./mvtodir.sh
   ```
2. Generating the file list for **getbibtex**

   ```
   $ ./gettitles.sh > titles.txt
   ```
3. Initializing / Updating the bibliography

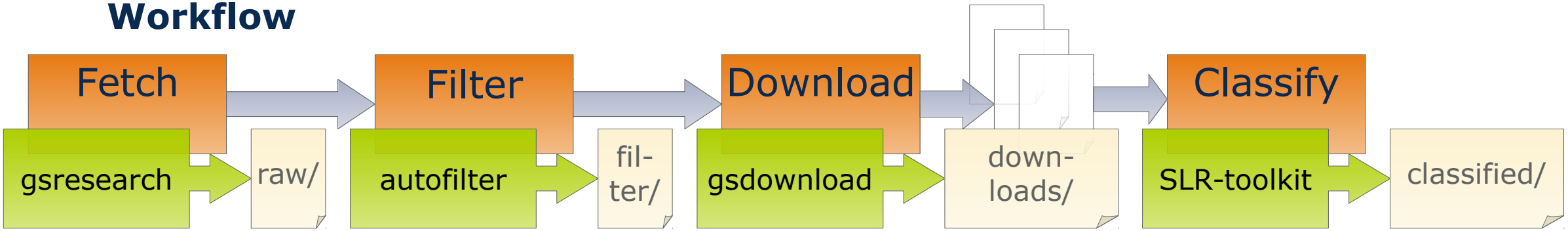
   ```
   $ ruby getbibtex.rb titles.txt my.bib 1>> my.bib
   ```

Task

- Fetch all publications matching a query string

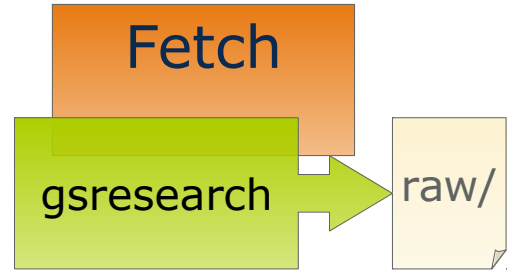
  ***With**: ospp, workflow*
  ***Exact**: sebastian richly*
- Sort out irrelevant publications
- Download PDF files for all relevant publications
- Collect statistics about survey process

Workflow



Automatic Querying

- Defining a search query
  - Exact, With, Any, and Without
  - Time span (from year to year)
- Directly supported by **gsresearch**



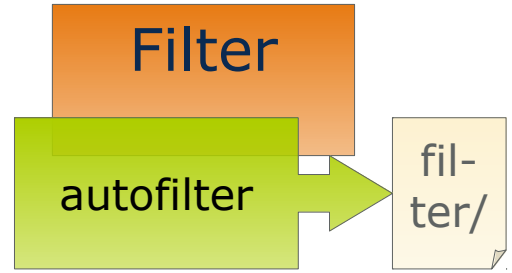
Steps

1. Test your query with Google Scholar[1)] (advanced search)
2. Change the **gsresearch.sh** accordingly
3. Run the script with
   ```
   $ ./gsresearch.sh
   ```
4. Be patient, very patient

1) https://scholar.google.com

Automatic Filtering

- Further filter the inital dataset
- Using **bibfilter** to select items by
  - document class, publisher, citation count, …
- Two automatic filtering steps in **autofilter**
  - Select items by publisher
    ACM, IEEE, Springer, ScienceDirect
  - Filter items with low impact ← **DEEMED EVIL**
    Citation Count < Log(Age)



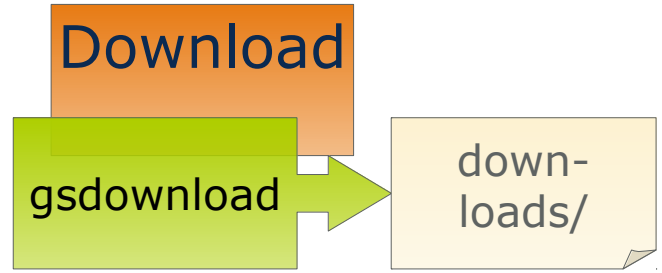
Human Filtering

- Check the title of the paper and (abstract, content)

```
$ mkdir filter_human
$ for f in `ls filter_rel/`; do
      ruby bibfilter.rb 'filter_rel/$f' > 'filter_human/$f' ;
  done
```

Automatic Download

- Download final set of relevant
- Access files via the publisher's site
- Support for the big four:
  ACM, IEEE, Springer, ScienceDirect
- Extensible towards other publishers
- Downloaded files are referenced within bibtex items



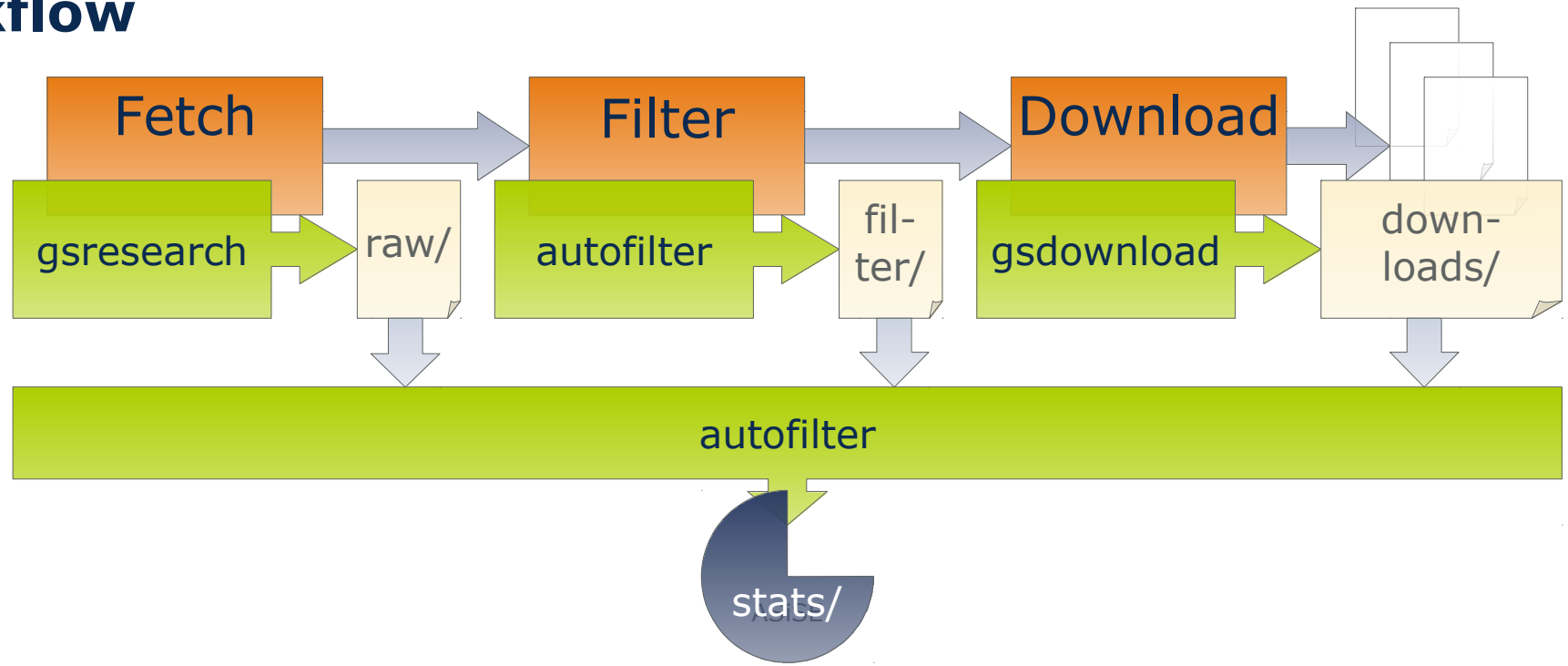
Steps

1. Run the script with
   ```
   $ ./gsdownload.sh
   ```
2. Be patient
3. Rerun
   ```
   $ ./autofilter.sh
   ```

Collecting Statistics

- Crucial to explain selection method of survey
- Generated automatically by **autofilter**
- Stored as csv files in *stats_*/* folder

Workflow

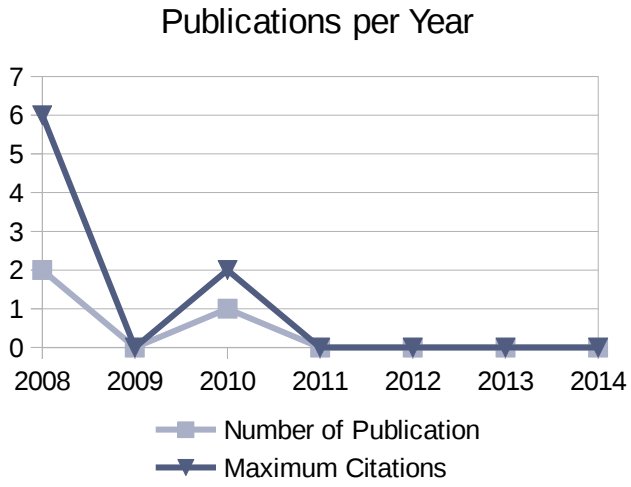


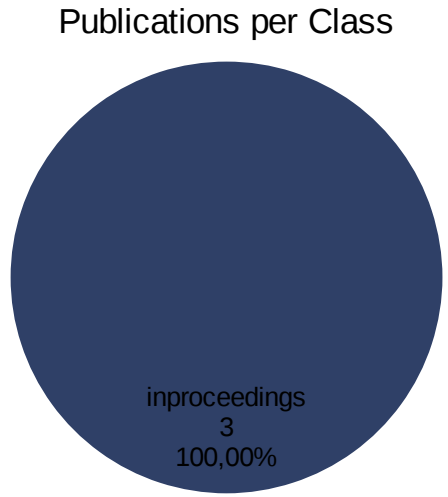
Example

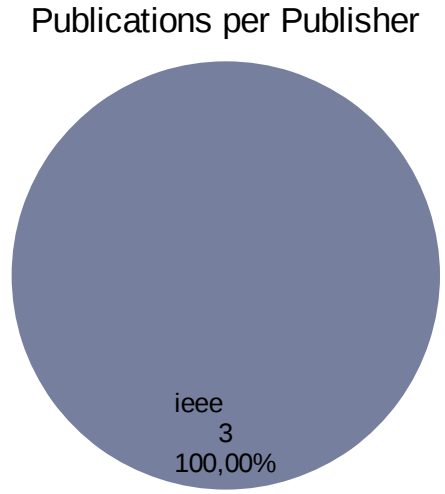
- Query for publications from 2008 to 2014

  ***With**: ospp, workflow*
  ***Exact**: sebastian richly*

- Inital dataset: 9 entries
- Automatic Filter: 4 entries
- Human Filter: 3 entries
- Download: 3 pdf files



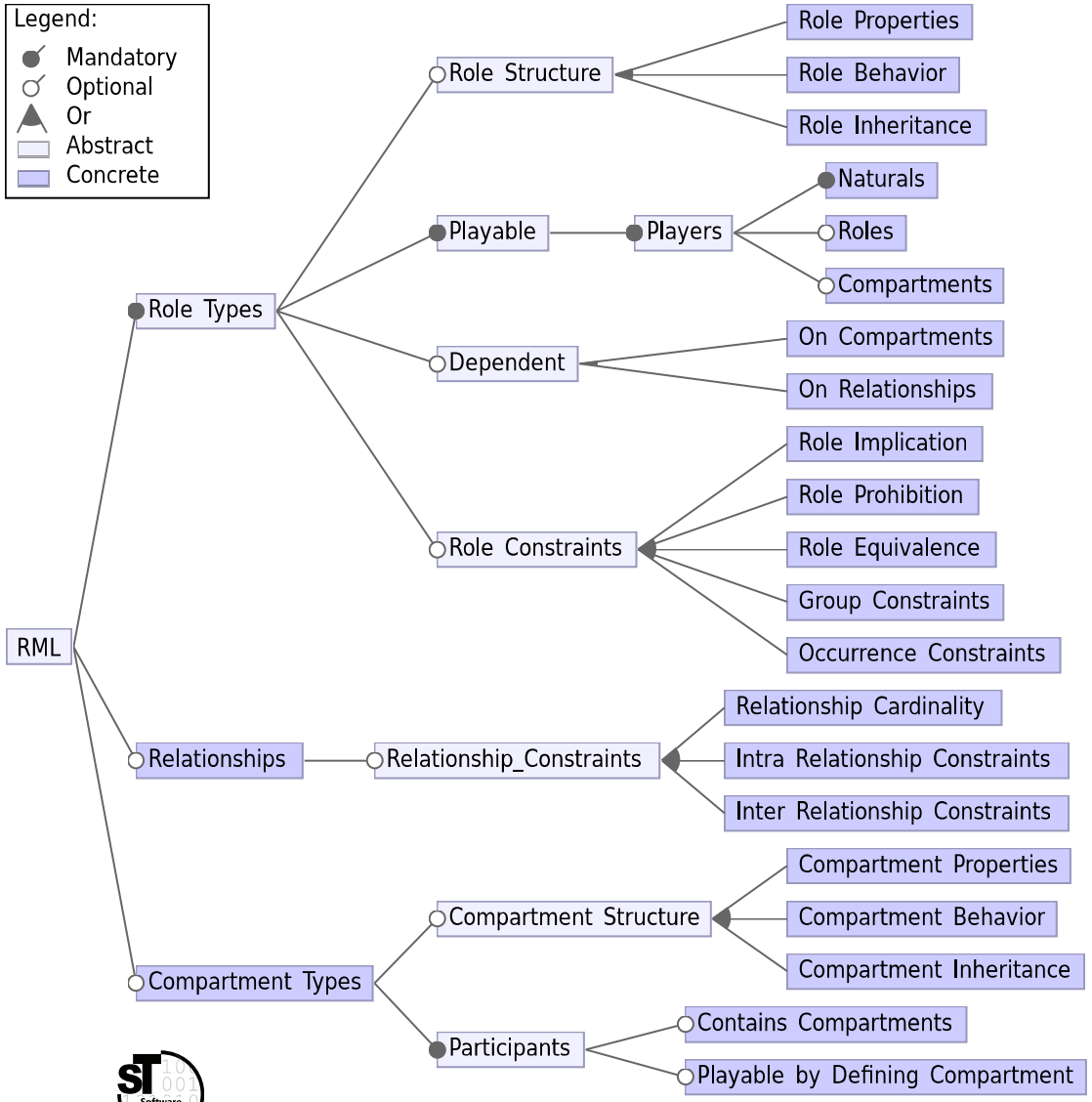




Common Tasks

- Create a classification scheme
  - Identify classifying criteria
  - Set up list, tree, or map of terms, features, requirements, or classes
- Classify papers and approaches
  - Indicate found criteria in papers
  - Maintain classification for each paper or approach
  - Produce diagrams for comparison *tables*, *bubble charts*, or *kiviat graphs*

Conducting a Literature Survey
Classification Scheme



For Papers

- Taxonomy of terms
- General classification of research papers by Shaw
- Orthogonal dimensions of classes

For Approaches

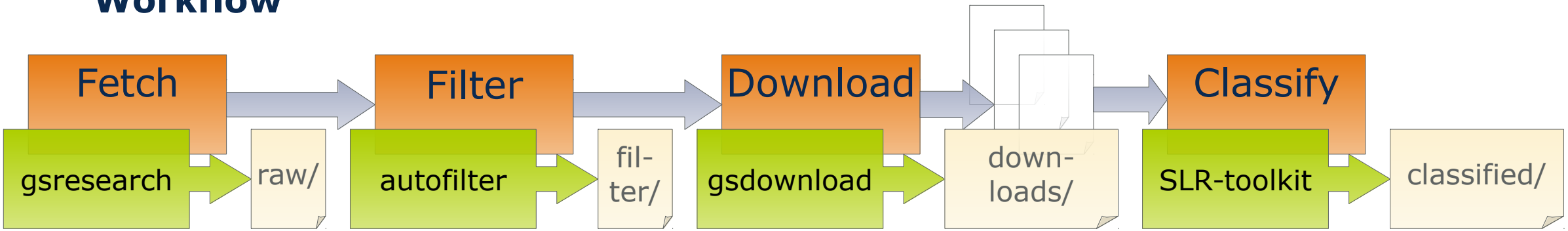
- List of (non-)functional requirements
- List of qualitative and quantitative properties
- Feature model consisting of features and dependencies

Conducting a Literature Survey
Mining a Classification Scheme

- Existence of general classification schemata, *e.g.,*
  Shaw's classification of research [Shaw2002]
- Utilize existing classifications from related **surveys** or **PhD theses**, *e.g.,*
  Feature model for language workbenches [Erdweg et al.2015]
- Creating new classification scheme
  - Start from existing schemata; extend missing dimension
  - Retrieve requirements, goals, or features from publications

Never use made up classification schemata

Workflow



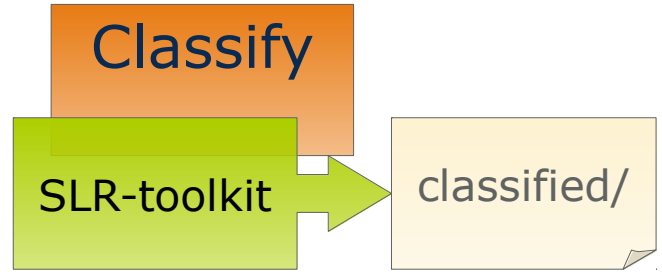
- After selecting relevant papers or approaches
- Investigate each paper annotate mentioned requirements and features
- Use tool support to track annotations for each paper or approach, *e.g.*,

  SLR-Toolkit[1] *uses BibTex annotation and supports arbitrary hierarchical classification schemes*

1) https://github.com/sebastiangoetz/slr-toolkit

Conducting a Literature Survey
Tool Support

- After selecting relevant papers or approaches
- Investigate each paper annotate mentioned requirements and features
- Use tool support for classifying papers

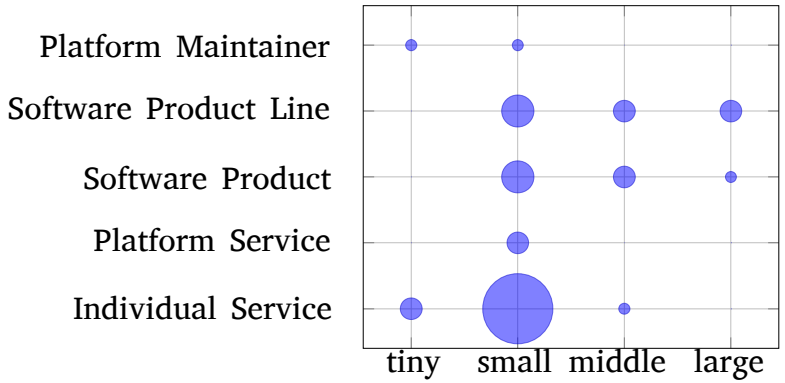


SLR-Toolkit[1]

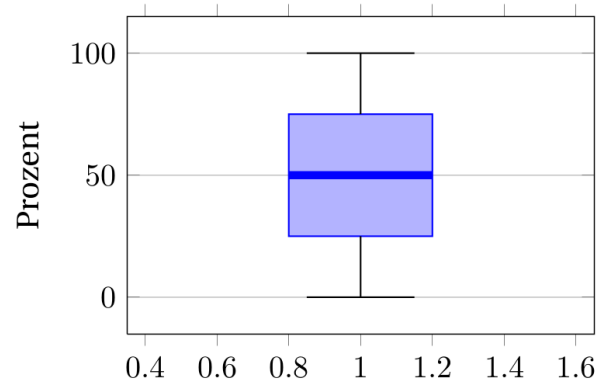
- Supports arbitrary hierarchical classification schemes
- Classification of papers per *BibTex* annotation
- Synchronization with *Mendeley*

1) https://github.com/sebastiangoetz/slr-toolkit

Conducting a Literature Survey
Showing Comparison Results



Bubble Chart



Box Plot

Qualitative Evaluation

- Comparison tables
  Terms, Icons (○ ◐ ● ◌), …
- Diagrams for detailed comparison
  (2D) *Pie charts, Histograms, …*
  (3D) *Bubble charts, 3D Plots, …*
  (nD) *Kiviatgraphs, Parallel Hierarchies, ...*

Quantitative Evaluation

- Tables for basic analysis
  Standard deviation (+/-), Mean, ...
- Plots for more complex analyses
  (2D) *Plots, Box plots, Line chart, ...*
  (3D) *Heat Maps, 3D Plots, ...*
  (nD) *Parallel Koordinates, ...*

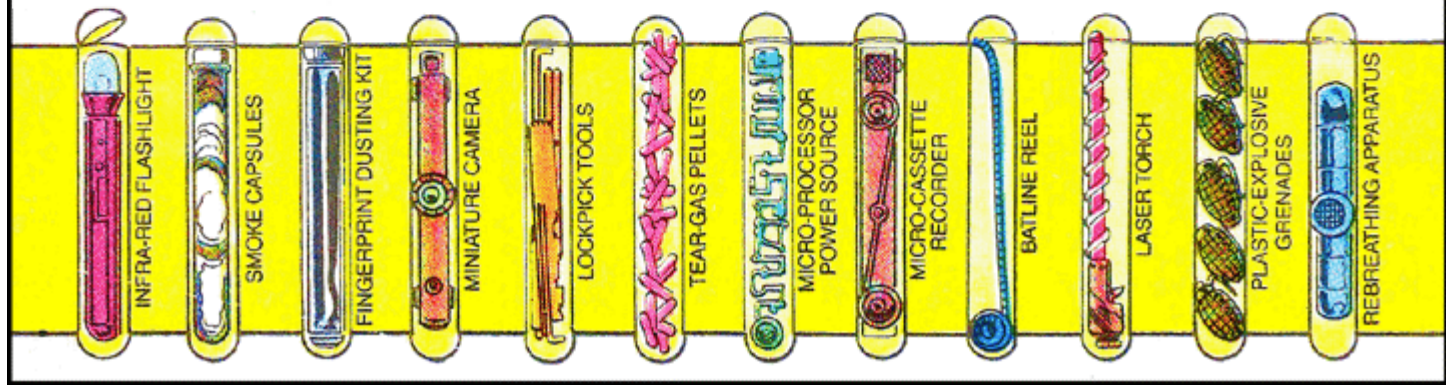
Automated Tasks

- Automated BibTex lookup for stored papers
- Automated BibTex lookup of specific Publications from web
- Automated filtering of large BibTex files
- Automated download of papers referenced by a BibTex file
- Semi-automatic literature survey

How to Get the Utility Belt?

GitHub

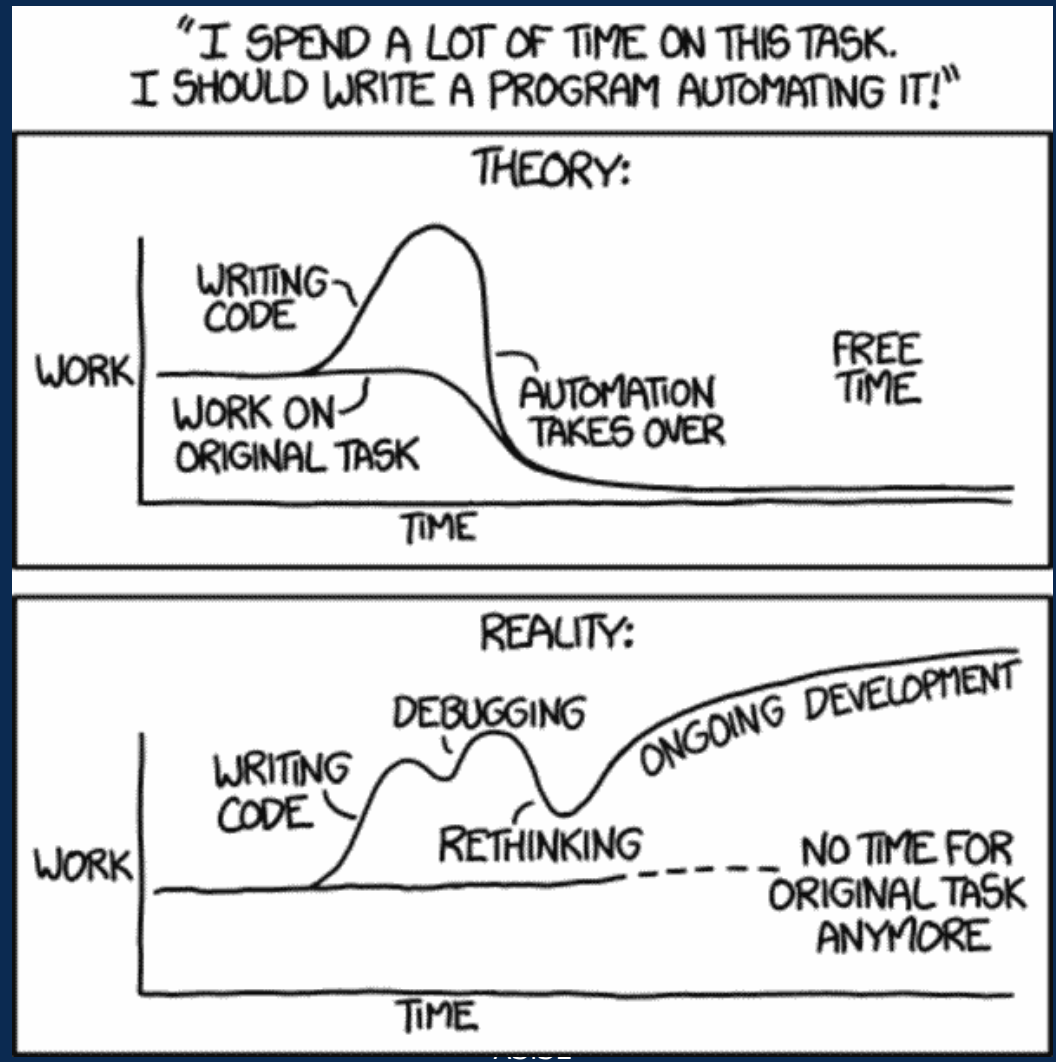
- ***bibfilter*** (https://github.com/Eden-06/bibfilter)
  contains the *bibfilter.rb* script as independent tool
- ***gsresearch*** (https://github.com/Eden-06/gsresearch)
  contains the various Ruby scripts
  - *getbibtex.rb,*
  - *gsresearch.rb, and*
  - *gsdownload.rb*
- ***SLR-toolkit*** (https://github.com/sebastiangoetz/slr-toolkit)



Exercise

1) Create a **Classification Scheme** for your related work.
2) Classify at least 5 papers wrt. this **Classification Scheme**.
3) Create a **comparison table** and add it to the *Related Work* section.

ASiSE
Collection, Filtering, and Classification



Automation (By Randell Munroe from xkcd.com)