

Makar: A Framework for Multi-source Studies Based on Unstructured Data

Mathias Birrer, Pooja Rani, Sebastiano Panichella, Oscar Nierstrasz

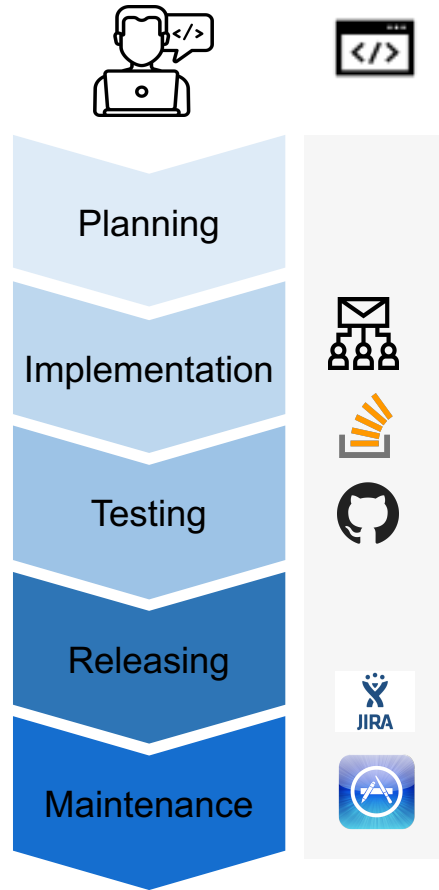
University of Bern, Switzerland

```
/**  
 *  TODO  
 */
```

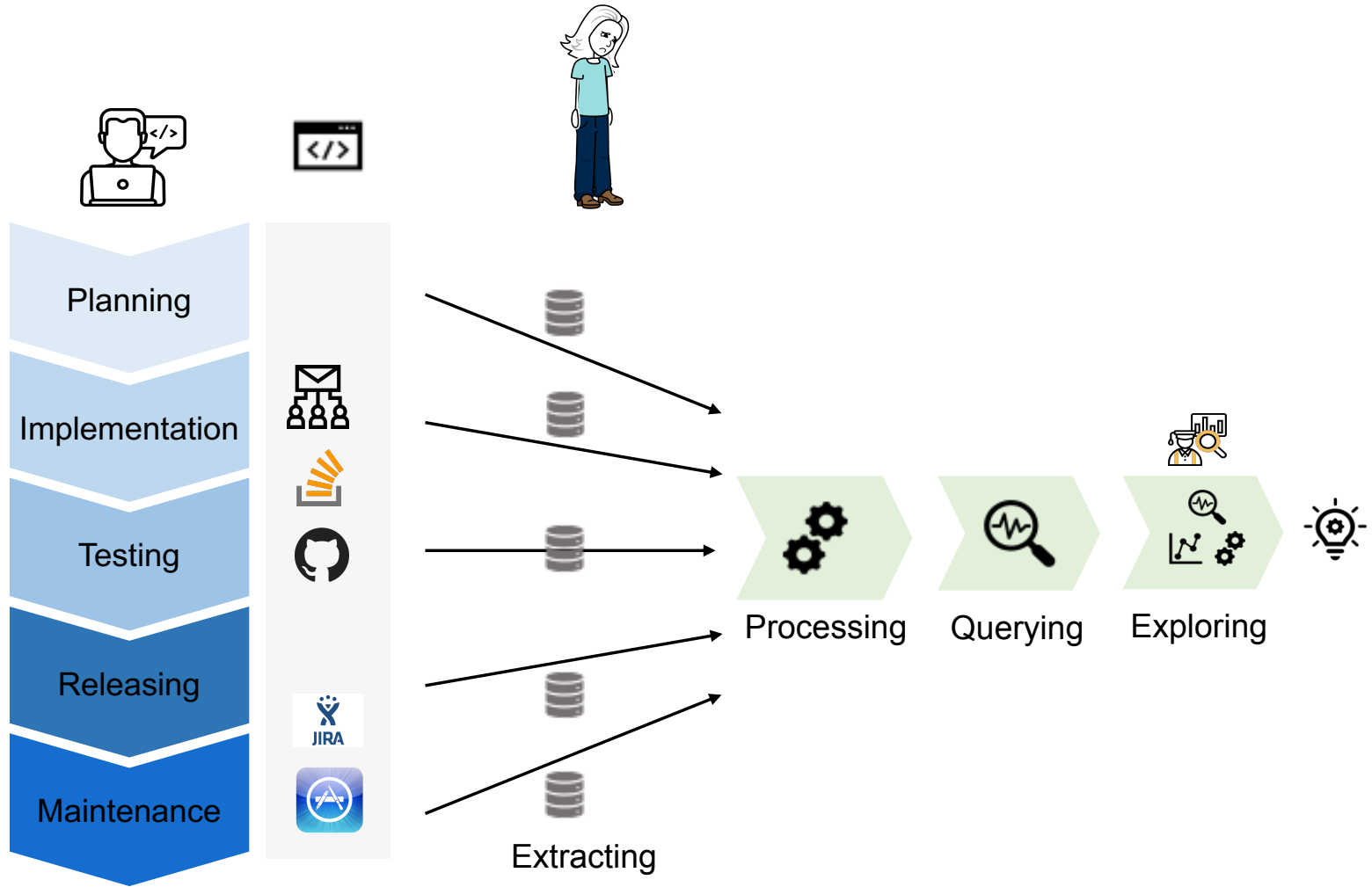
```
public void log(String s) {  
    System.out.println(s);  
}
```

Do developers discuss code comments?

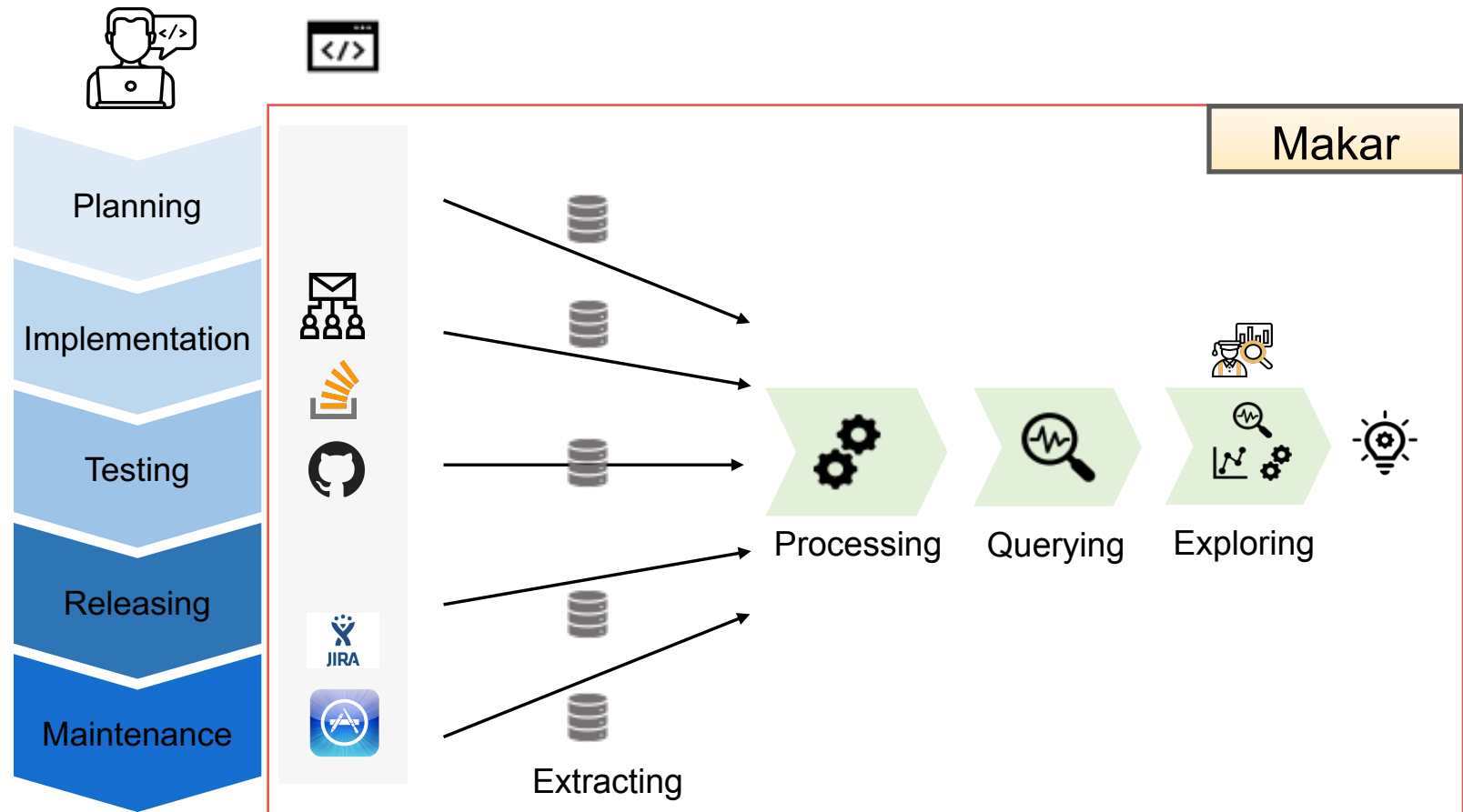
Developers use various discussion sources



Challenges



Makar: A tool for Multi-source Studies



Features

Extract data from different sources

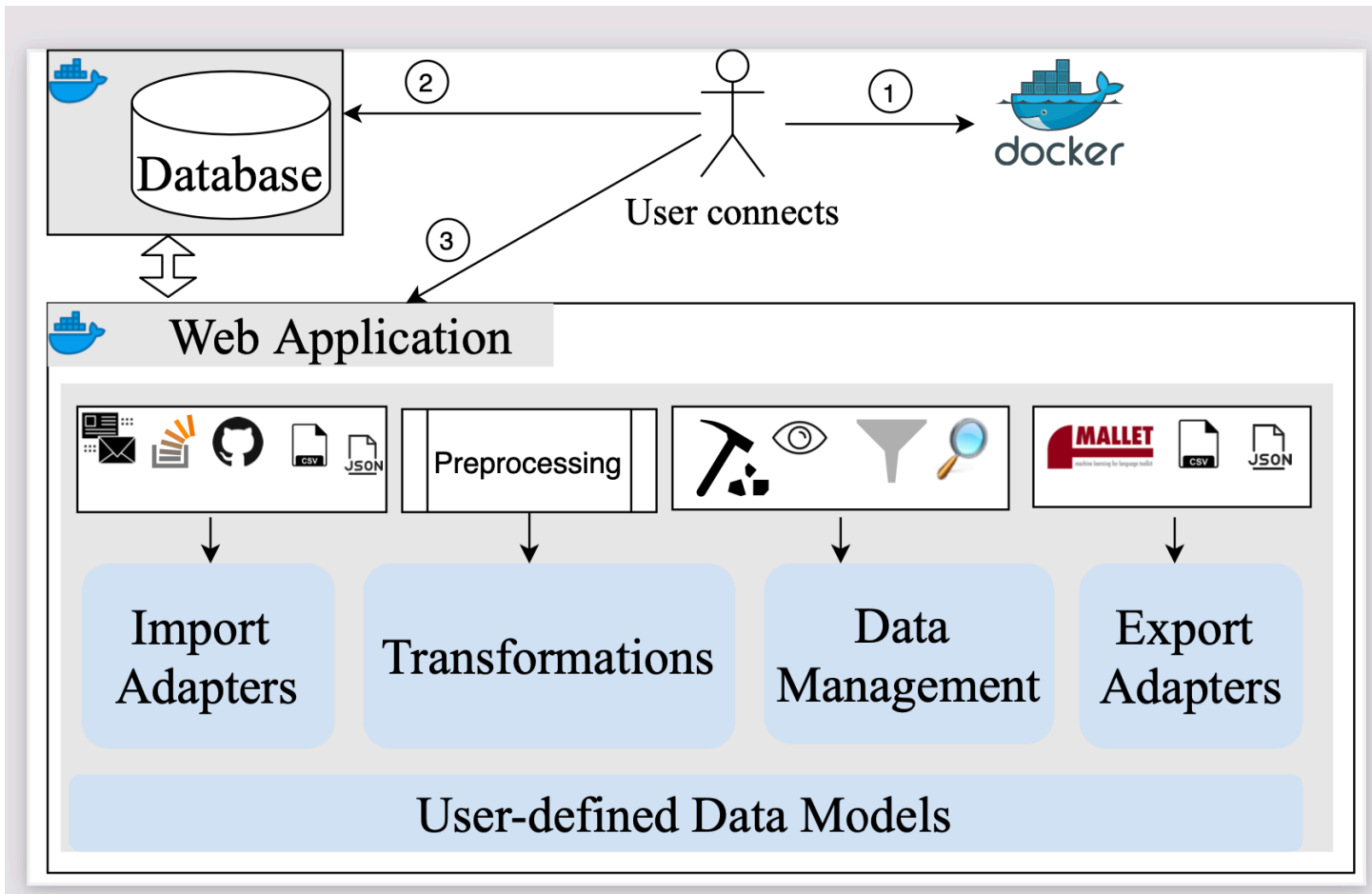
e.g., Stack Overflow, Github, Mailing Lists

Support mapping and processing the data

Explore and perform ad-hoc searches

Extending the dataset easily

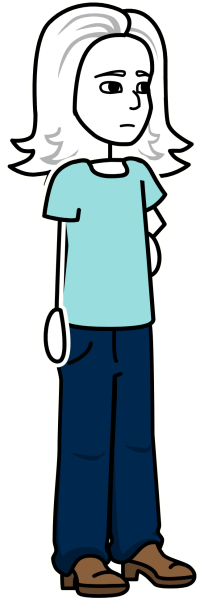
Makar Architecture



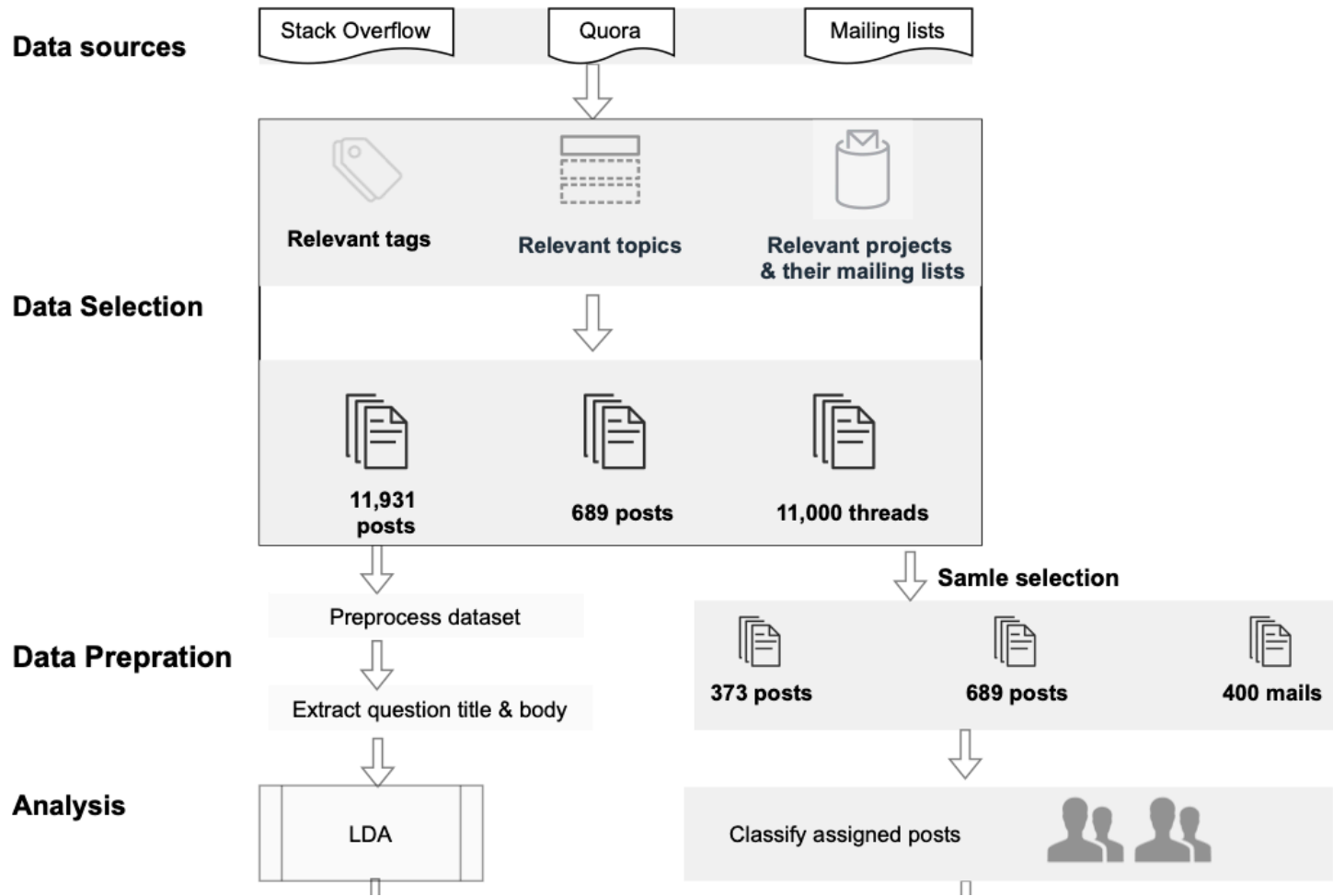
Demo

Do developers ask questions about comments?

Do the comment related questions contain code snippets?



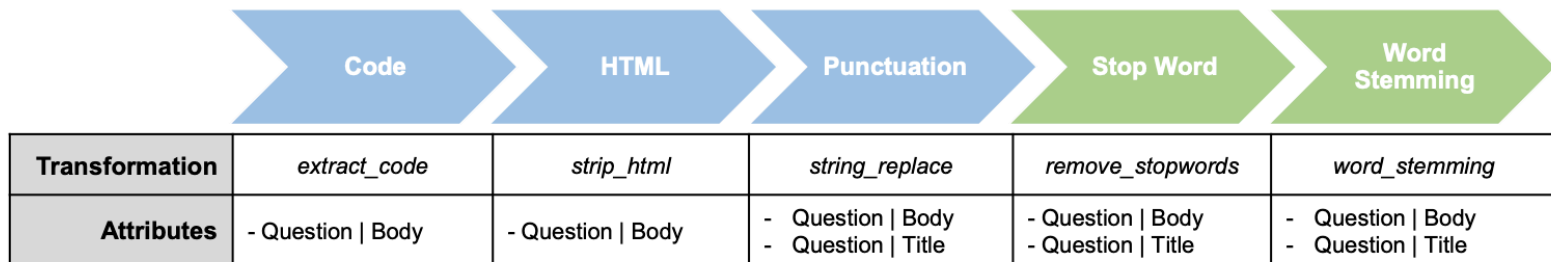
Case Study



Case Study

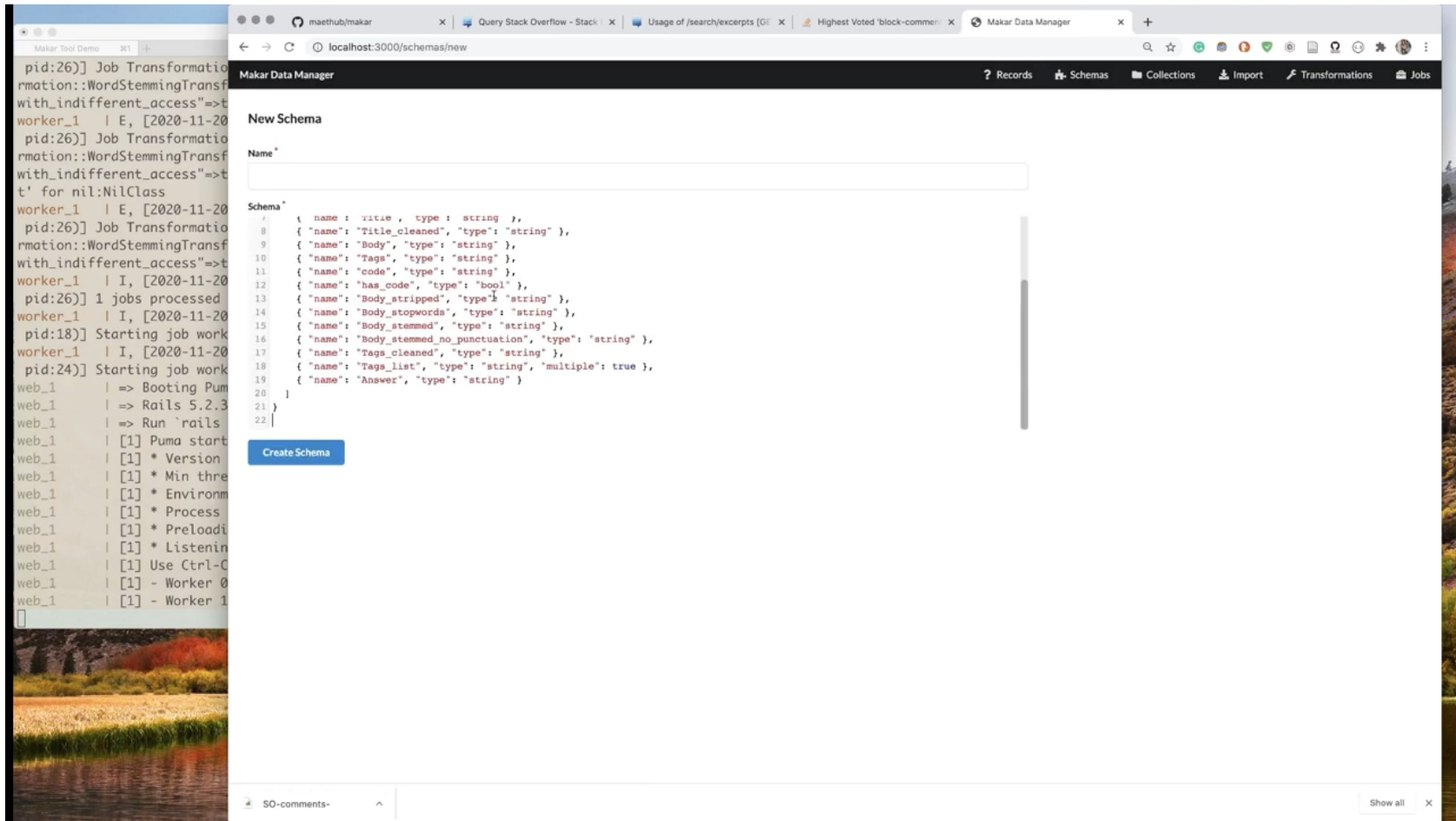
Import adapters: Stack overflow, CSV, Apache mailing list

Preprocess the data:



Export adapters: CSV adapters

Features: Import



A complete demo is available at [Youtube](#)

Features: Search

The screenshot displays the Makar Data Manager web application. On the left, a sidebar shows a terminal-like log with various system messages. The main area is titled 'Makar Data Manager' and features a search interface. A dropdown menu is open, showing options like 'Record', 'Id', 'Schema', 'Deactivated', 'Created at', 'Updated at', 'Record value', 'Name', 'Value type', 'Data', 'Value hash', 'Created at', 'Updated at', 'Record', 'Index', and 'Data'. The search results table at the bottom shows a single record with ID 837, Schema 'SO Questions and Answers by Tag', and a detailed JSON body. The interface also includes buttons for 'Add to collection', 'Remove from collection', and 'Found 1034 records'.

pid:26]) Job Transformation: WordStemmingTransformation with_indifferent_access=>true
worker_1 | E, [2020-11-20 14:26:26] Job Transformation: WordStemmingTransformation with_indifferent_access=>true
t' for nil:NilClass
worker_1 | E, [2020-11-20 14:26:26] Job Transformation: WordStemmingTransformation with_indifferent_access=>true
pid:26]) 1 jobs processed
worker_1 | I, [2020-11-20 14:26:26] Starting job work
worker_1 | I, [2020-11-20 14:26:26] Starting job work
web_1 | => Booting Puma
web_1 | => Rails 5.2.3
web_1 | => Run 'rails
web_1 | [1] Puma start
web_1 | [1] * Version
web_1 | [1] * Min three
web_1 | [1] * Environment
web_1 | [1] * Process
web_1 | [1] * Preload
web_1 | [1] * Listening
web_1 | [1] Use Ctrl-C
web_1 | [1] - Worker 0
web_1 | [1] - Worker 1

Makar Data Manager

Records Schemas Collections Import Transformations Jobs

Data Records

Schema *
SO Questions and Answers by Tag

Collection *

Advanced Search

Matches all conditions Remove

Attribute Predicate contains Value * Remove

Add Condition Add Condition Group

Save Query Filter Name Load Filter

Collection *

Add to collection Remove from collection Found 1034 records

ID	Schema	Data
837	SO Questions and Answers by Tag	{:id=>17636764,:CreationDate=>2013-07-14,:ViewCount=>2234,:Title=>Get all comments on a post by an array of users,:Title_cleaned=>nil,:Body=><p>I'm trying to get all the comments from a post by an array of users.</p><p>This is what I'd like to be able to do:</p><pre><code>User_ids = array(10, 22, 41, 80);\n\$post_id = 57;\n\n\$args = { :tags => 'wordpress', :code => nil, :has_code => nil, :body_stripped => nil, :body_stopwords => nil, :body_stemmed => nil, :body_stemmed_no_punctuation => nil,

SO-comments-

Show all

A complete demo is available at [Youtube](#)

Features: Collection

The screenshot displays the Makar Data Manager web application. The interface includes a top navigation bar with tabs for 'Records', 'Schemas', 'Collections', 'Import', 'Transformations', and 'Jobs'. The main content area is titled 'SO Questions and Answers by Tag' and shows a 'Collection' dropdown. Below this is an 'Advanced Search' section with two search conditions: 'Attribute: Name, Predicate: contains, Value: Title' and 'Attribute: Data, Predicate: contains, Value: comment'. The 'Save Query' section is visible, along with a 'Filter Name' dropdown and a 'Load Filter' button. The 'Collection' section shows 'Found 1130 records'. At the bottom, a table displays the first record with columns for ID, Schema, and Data. The table has a search icon and a close icon in the bottom right corner.

ID	Schema	Data
1	SO Questions and Answers by Tag	[{"id":244777,"CreationDate":"2008-10-28","ViewCount":2476219,"Title":"Can comments be used in JSON?","Title_cleaned":null,"Body":"<p>Can I use comments inside a JSON file? If so, how?</p>\n";:Tags:<json><comments>;:code=<nil>;:has_code=<nil>;:Body_stripped=<nil>;:Body_stopwords=<nil>;:Body_stemmed=<nil>;:Body_stemmed_no_punctuation=<nil>;:Tags_cleaned=<nil>;:Tags_list=<nil>;:Answer:<p>No.</p>\n<p>The JSON is data only, and if you include a comment, then it will be data too.</p>\n<p>You could have a designated data element called <code>"comment"</code> (or something) ..."]

A complete demo is available at [Youtube](#)

Features: Transform

The screenshot displays the Makar Data Manager web application. The interface is divided into several sections:

- Left Sidebar:** A file explorer showing a directory structure with files like 'pid:26]', 'Job Transformation::WordStemmingTransf', 'with_indifferent_access'=>t', 'worker_1 | E, [2020-11-20', 'pid:26]', 'Job Transformation::WordStemmingTransf', 'with_indifferent_access'=>t', 't' for nil:NilClass', 'worker_1 | E, [2020-11-20', 'pid:26]', 'Job Transformation::WordStemmingTransf', 'with_indifferent_access'=>t', 'worker_1 | I, [2020-11-20', 'pid:26]', '1 jobs processed', 'worker_1 | I, [2020-11-20', 'pid:18]', 'Starting job work', 'worker_1 | I, [2020-11-20', 'pid:24]', 'Starting job work', 'web_1 | => Booting Pum', 'web_1 | => Rails 5.2.3', 'web_1 | => Run `rails', 'web_1 | [1] Puma start', 'web_1 | [1] * Version', 'web_1 | [1] * Min thre', 'web_1 | [1] * Environm', 'web_1 | [1] * Process', 'web_1 | [1] * Preloadi', 'web_1 | [1] * Listenin', 'web_1 | [1] Use Ctrl-C', 'web_1 | [1] - Worker 0', 'web_1 | [1] - Worker 1'.
- Main Panel:**
 - Collection Questions containing comment in title:** A search filter.
 - Description:** Contains all the questions with comment in their title.
 - Auto-Filter:** All questions with comment in title.
 - Export:** A dropdown menu for 'Attributes' and 'Export format' (set to 'json').
 - Buttons:** Edit, Delete, Drop ALL Records, Table View, Custom Exports.
 - Records:** A table with 1887 records. The table has columns: ID, Schema, and Data. The data column contains JSON snippets of comments.
- Bottom:** A status bar showing 'Found 1887 records' and a 'Show all' button.

A complete demo is available at [Youtube](#)

Tool Comparison

Tool	Extract	Process	Manage
Octoparse	✓	✗	✗
Knime	Extension	✓	✓
Rapidminer	✗	Limited	Limited
ELKI	✗	✓	✗
Keel	✗	✗	✗
WEKA	✗	✗	✗
TrifactaWrangler	✗	✓	✓
Boa	Limited	✓	✓
OpenRefine	✓	✓	✓
Makar	✓	✓	✓

Future work

Extension of data source adapters

Building a UI of the study pipeline

Development of analysis and visualisation components

Facilitation of more multi-source studies

Makar: A Framework for Multi-source Studies Based on Unstructured Data

Hosted on Github

<https://github.com/maethub/makar>

Demo at YouTube

<https://youtu.be/Yqj1b4Bv-58>

Replication Package at Zenodo

<https://doi.org/10.5281/zenodo.4434822>

Contact us



<https://twitter.com/poojaruhal>

u^b

<http://scg.unibe.ch/staff/Pooja-Rani>