RUCIO

Google
Summer of Code

Donkeybot

Vasilis Mageirakos, GSoC ‘20

Donkeybot

Premise, resources and desired outcome.

Concept :

Due to the vast amount of support requests, we are looking into methods to assist the support team in answering these requests. Ideally, the support would be provided by an intelligent bot able to process and understand the user’s requests and finally trigger appropriate action.

Gather and analyse the relevant data

Create the bot able to identify questions and answers

Generate appropriate answers on new user queries

Deliverables :

1. Fetching and parsing of emails, GitHub issues and Rucio documentation.
2. Prototype Bot creation able to utilize the above data.
3. Question-answering prototype pipeline and demo.

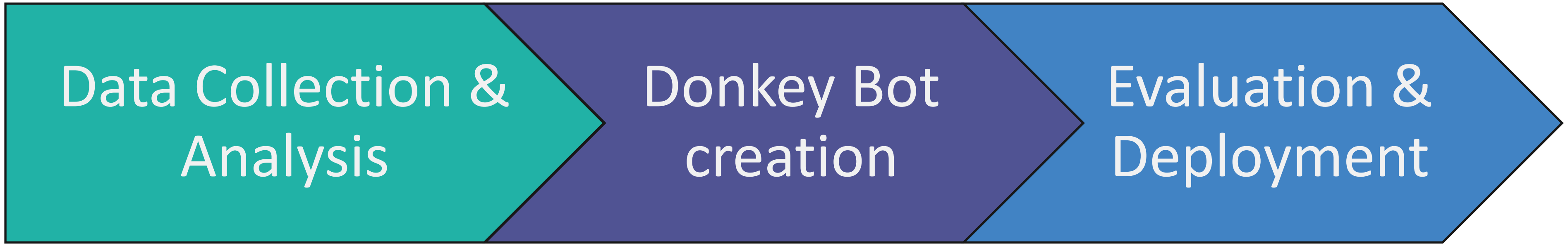
Project Timeline

Understanding key milestones.

Question detection module (Regex patterns)

Search engine module (BM25)

Answer generation module using transformers (BERT, ALBERT, RoBERTa, GPT-3)



Data Collection & Analysis

Donkey Bot creation

Evaluation & Deployment

NER tagger creation for user information hashing

Fetching, parsing and storage modules (SQLite)

Data quality analysis

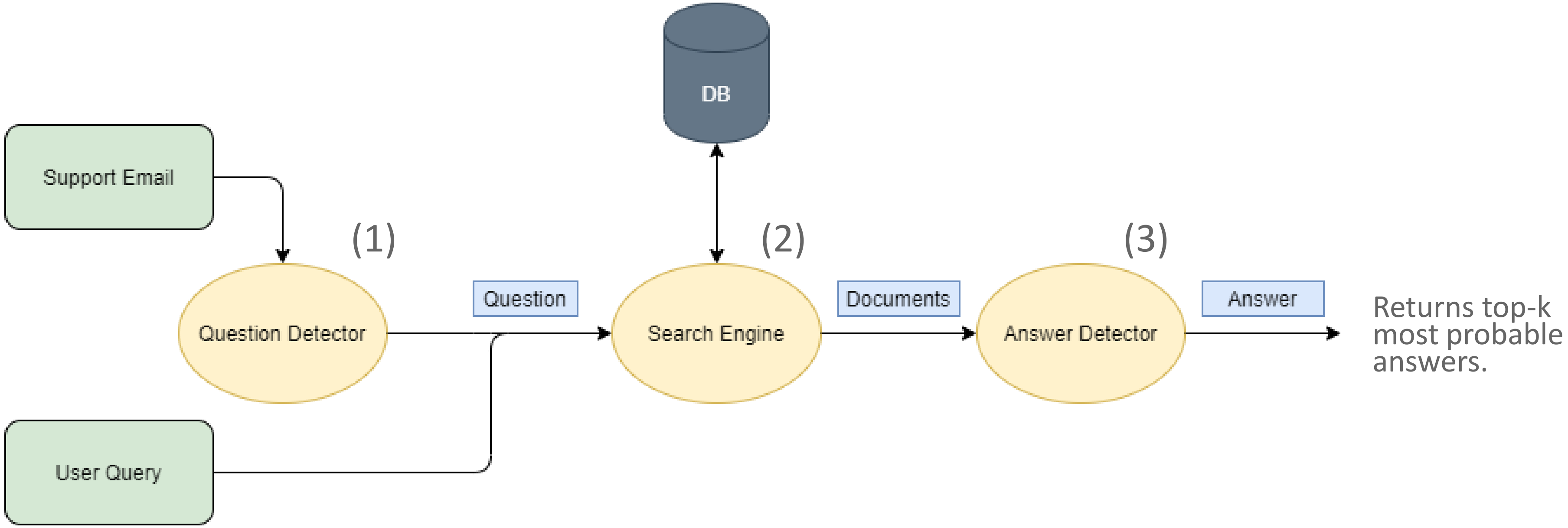
Baseline performance evaluation (validation set/live testing)

Supervised feedback capabilities for bot continuous improvement

User Interface creation

Question-Answering Pipeline

Question Answering system’s architecture.



Returns top-k most probable answers.

(1) Uses regex patterns to extract Questions from support emails.

(2) Uses BM25 algorithm to retrieve top-k most relevant documents.

(3) Uses BERT to find answers in each document retrieved with corresponding confidence scores.

Answer Detection

Understanding input/output data and structure BERT expects.

Input :

```
{ 'question': ‘how can a question answering service produce
answers’,
   'context': ‘One such task is reading comprehension. Given a
passage of text, we can ask questions about the passage that
can be answered by referencing short excerpts from the text.
For instance, if we were to ask about this paragraph, "how
can a question be answered in a reading comprehension
task“ ‘
}
```

Output :

```
{ 'score': 0.38941961529900837,
  'start': 128,
  'end': 169,
  'answer': 'referencing short excerpts from the text.’}
```

BERT is pretrained on SQuAD 2.0

100k answerable questions and 50k unanswerable from a set of Wikipedia articles and books.

Document Types

Understanding the different documents that are retrieved and used for answer detection.

Email	Issues	Documentation
Document retrieved based on Question similarity. *	Document retrieved based on Question similarity. *	Document retrieved based on similarity.
Emails from the same conversation used for context.	Comments from the same issue used for context.	Whole document used as context.
Noisy, messy data	Less noisy more structured data	Least noisy and most structured

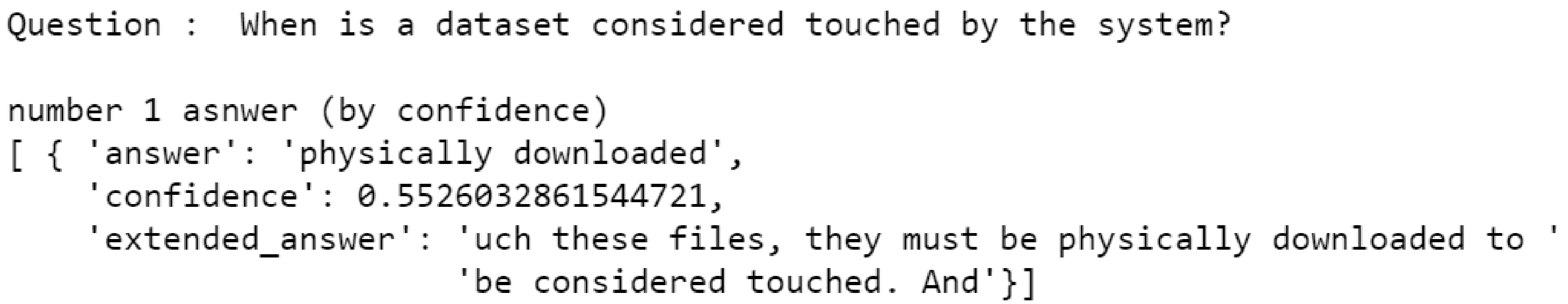
* Question detection has already been run on archived emails, issues and issue comments to populate our storage.

** On final prototype we might end up retrieving based on document similarity since Named Entity extraction won’t yet be implemented to improve retrieval.

Examples 1/3

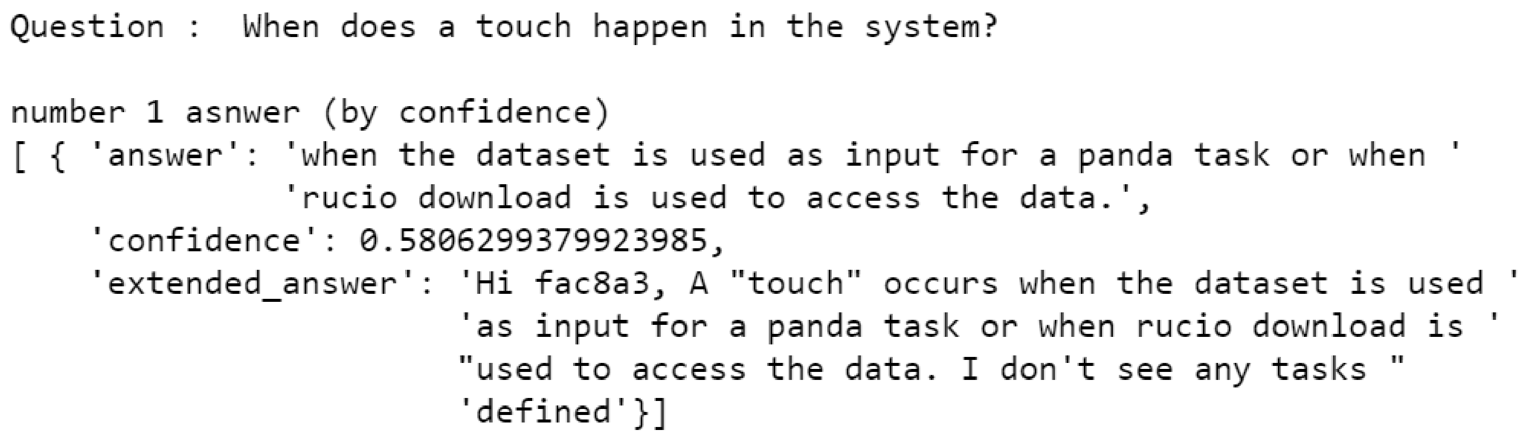
Let’s look at some Rucio specific questions.

Source : Rucio support emails



```
Question :  When is a dataset considered touched by the system?

number 1 asnwer (by confidence)
[ { 'answer': 'physically downloaded',
    'confidence': 0.5526032861544721,
    'extended_answer': 'uch these files, they must be physically downloaded to '
                       'be considered touched. And'}]
```



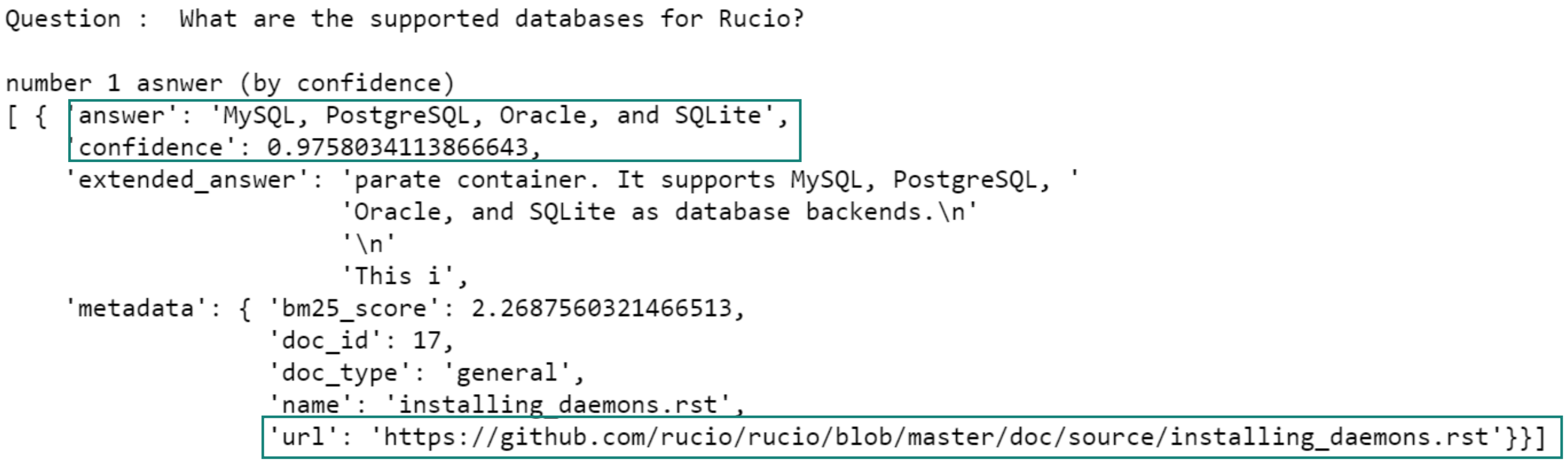
```
Question :  When does a touch happen in the system?

number 1 asnwer (by confidence)
[ { 'answer': 'when the dataset is used as input for a panda task or when '
              'rucio download is used to access the data.',
    'confidence': 0.5806299379923985,
    'extended_answer': 'Hi fac8a3, A "touch" occurs when the dataset is used '
                       'as input for a panda task or when rucio download is '
                       "used to access the data. I don't see any tasks "
                       'defined'}]
```

Examples 2/3

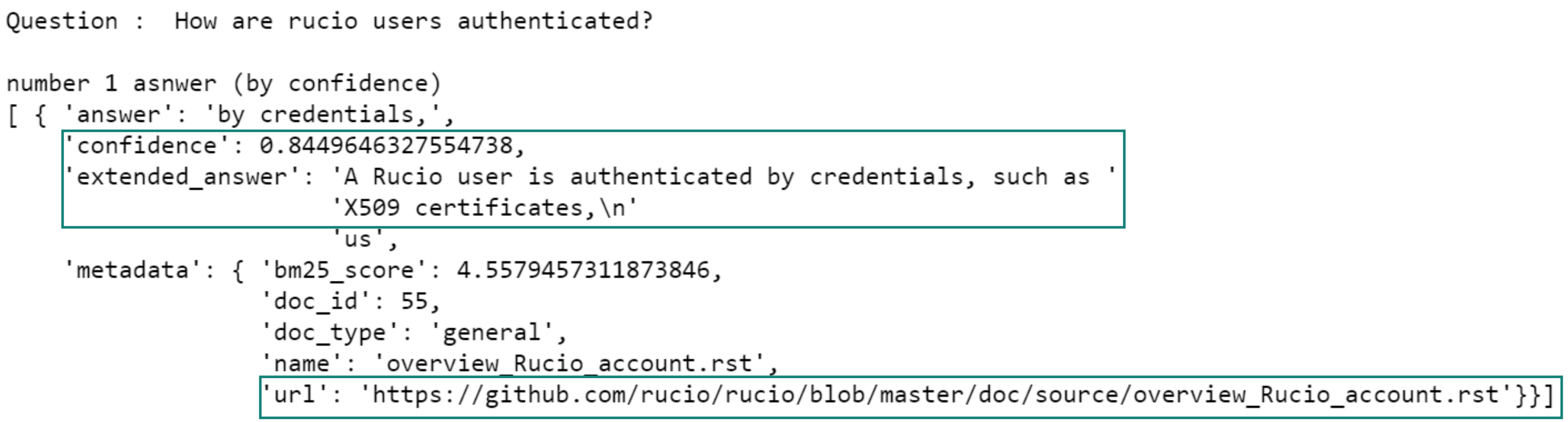
Let’s look at some more Rucio specific questions.

Source : Rucio documentation



```
Question :  What are the supported databases for Rucio?

number 1 asnwer (by confidence)
[ { 'answer': 'MySQL, PostgreSQL, Oracle, and SQLite',
    'confidence': 0.9758034113866643,
    'extended_answer': 'parate container. It supports MySQL, PostgreSQL, '
                       'Oracle, and SQLite as database backends.\n'
                       '\n'
                       'This i',
    'metadata': { 'bm25_score': 2.2687560321466513,
                  'doc_id': 17,
                  'doc_type': 'general',
                  'name': 'installing_daemons.rst',
                  'url': 'https://github.com/rucio/rucio/blob/master/doc/source/installing_daemons.rst'}}]
```



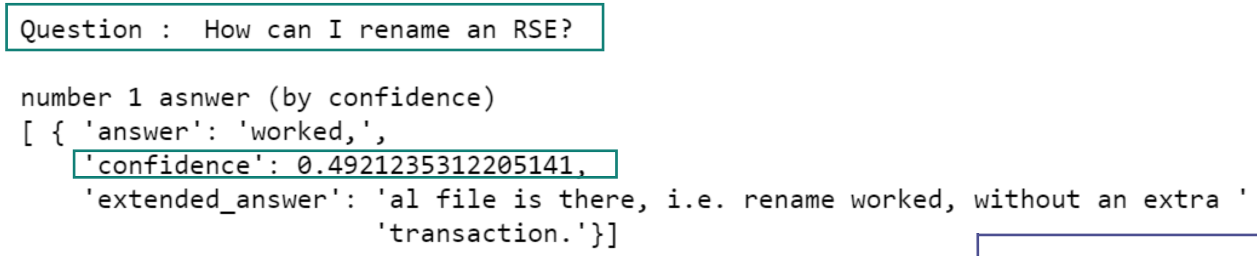
```
Question :  How are rucio users authenticated?

number 1 asnwer (by confidence)
[ { 'answer': 'by credentials,',
    'confidence': 0.8449646327554738,
    'extended_answer': 'A Rucio user is authenticated by credentials, such as '
                       'X509 certificates,\n'
                       'us',
    'metadata': { 'bm25_score': 4.5579457311873846,
                  'doc_id': 55,
                  'doc_type': 'general',
                  'name': 'overview_Rucio_account.rst',
                  'url': 'https://github.com/rucio/rucio/blob/master/doc/source/overview_Rucio_account.rst'}}]
```

Examples 3/3

Donkeybot can even answer more general questions!

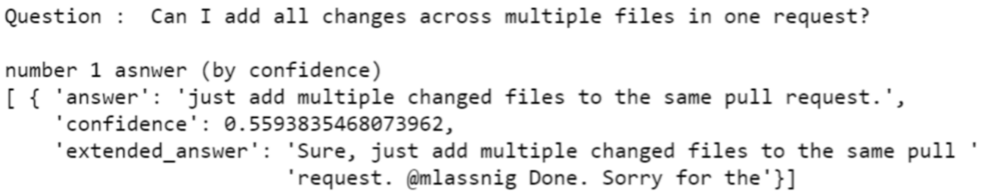
Source : Rucio support emails



```
Question :  How can I rename an RSE?

number 1 asnwer (by confidence)
[ { 'answer': 'worked,',
    'confidence': 0.4921235312205141,
    'extended_answer': 'al file is there, i.e. rename worked, without an extra '
                       'transaction.'}]
```

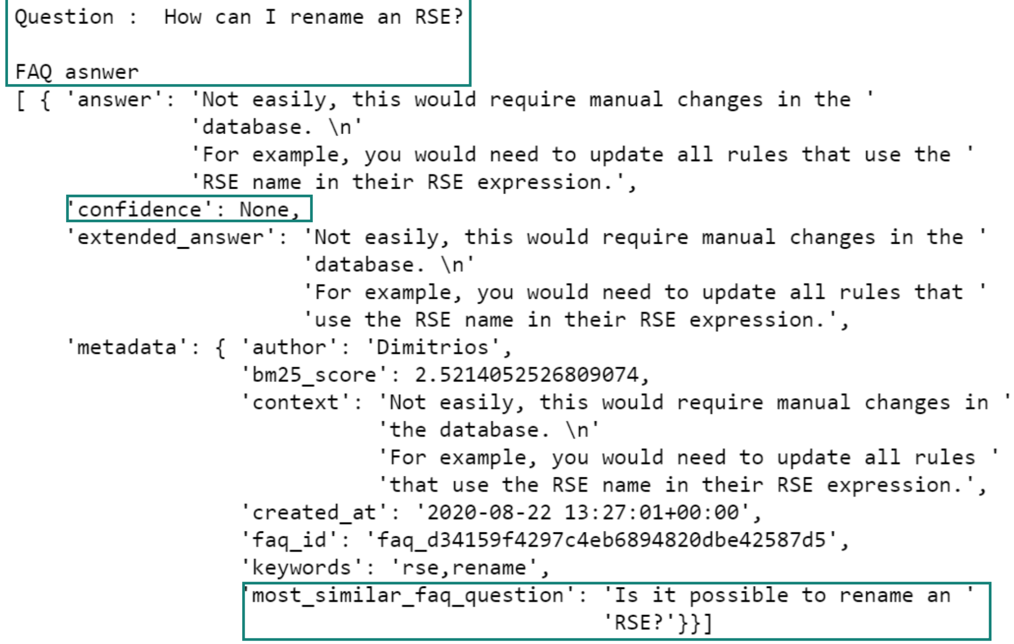
Source : GitHub Issues



```
Question :  Can I add all changes across multiple files in one request?

number 1 asnwer (by confidence)
[ { 'answer': 'just add multiple changed files to the same pull request.',
    'confidence': 0.5593835468073962,
    'extended_answer': 'Sure, just add multiple changed files to the same pull '
                       'request. @mlassnig Done. Sorry for the'}]
```

Source : FAQs



```
Question :  How can I rename an RSE?

FAQ asnwer
[ { 'answer': 'Not easily, this would require manual changes in the '
              'database. \n'
              'For example, you would need to update all rules that use the '
              'RSE name in their RSE expression.',
    'confidence': None,
    'extended_answer': 'Not easily, this would require manual changes in the '
                       'database. \n'
                       'For example, you would need to update all rules that '
                       'use the RSE name in their RSE expression.',
    'metadata': { 'author': 'Dimitrios',
                  'bm25_score': 2.5214052526809074,
                  'context': 'Not easily, this would require manual changes in '
                             'the database. \n'
                             'For example, you would need to update all rules '
                             'that use the RSE name in their RSE expression.',
                  'created_at': '2020-08-22 13:27:01+00:00',
                  'faq_id': 'faq_d34159f4297c4eb6894820dbe42587d5',
                  'keywords': 'rse,rename',
                  'most_similar_faq_question': 'Is it possible to rename an '
                                               'RSE?'}}]
```

Next Steps

What needs to happen in the future?

1. Creation of User Interface.
   - Allow us to deploy on a test server, gather usage data and evaluate the bot.
   - Allow us to create a dataset to later improve the bot.

2. Creation of custom Named Entity Recognizer.
   - Dynamically detect Rucio specific entities (DIDs, RSEs, Operations …).
   - Improve performance and generate dynamic answers.

3. Iterate over, refactor and improve the current prototype.

Questions?

Donkeybot repository : https://github.com/rucio/donkeybot

Developer contact : b.mageirakos@gmail.com (or Slack)

GitHub : https://github.com/mageirakos

Google Summer of Code : https://summerofcode.withgoogle.com/

GSoC project : https://bit.ly/3i9we8H