CrowData Documentation

Release 0.1

Gabriela Rodriguez & Manuel Aristaran

June 02, 2014

Contents

`CrowData` is a tool to collaborate on the verification or release of data that otherwise would be hard or impossible to get via automatic tools.

CHAPTER 1

When to use Crowdata?

- VozData is a website from La Nacion in Argentina to convert scanned PDF documents from senate spendings into an usable dataset. Collaborating to free data from PDFs.

1.1 Contents

1.1.1 Technical Introduction

The basic features for 'Crowdata' are

- Store a set of documents (PDF or other formats supported by Document Cloud)
- Define a form, via the admin, for the information that wants to be extracted from the documents.
- Present users with a document and the form to allow anybody, that is registered to the website, to send us the information they see in the document.

- Have access to download the CSV of all the information extracted from the PDFs by the users.

1.1.2 How To Install It Locally

1. Python 2.7.5
2. We recommend the use of *virtualenv <http://virtualenv.org>* — Install it.
3. Create a virtual environment and activate it:

```
$ virtualenv ~/.python-envs/crowdata
$ . ~/.python-envs/crowdata/bin/activate
```

4. Get the source code:

```
$ git clone https://github.com/crowdata/crowdata.git crowdata
$ cd crowdata
```

5. Install dependencies:

```
$ pip install -r requirements.txt
```

(If you are using Ubuntu, you may need to install *python-dev* before dependencies.)

6. Create PostgreSQL database:

```
$ createuser -s -h localhost crow_user
$ createdb -O crow_user -h localhost crowdata_development
```

7. Create extensions for doing *trigram matching <http://www.postgresql.org/docs/9.2/static/pgtrgm.html>* and *removing accents <http://www.postgresql.org/docs/9.1/static/unaccent.html>* in PostgreSQL:

```
$ psql -ucrow_user
crow_user=# \c crowdata_development
crowdata_development=# CREATE EXTENSION pg_trgm;
crowdata_development=# CREATE EXTENSION unaccent;
```

Note: In Debian/Ubuntu you need to install postgresql-contrib-9.1 and geospatial libraries.

8. We keep local settings out of GIT. You will need to copy *local_settings.py.example* to *local_settings.py*. You will need to edit the database settings there.:

```
DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.postgresql_psycopg2', # Add 'postgresql_psycopg2', 'postgr
        'NAME': 'crowdata_development',                      # Or path to database file if using
        'USER': 'crow_user',
        'PASSWORD': '',
        'HOST': '',
        'PORT': '',
    }
}
```

9. Initialize the database:

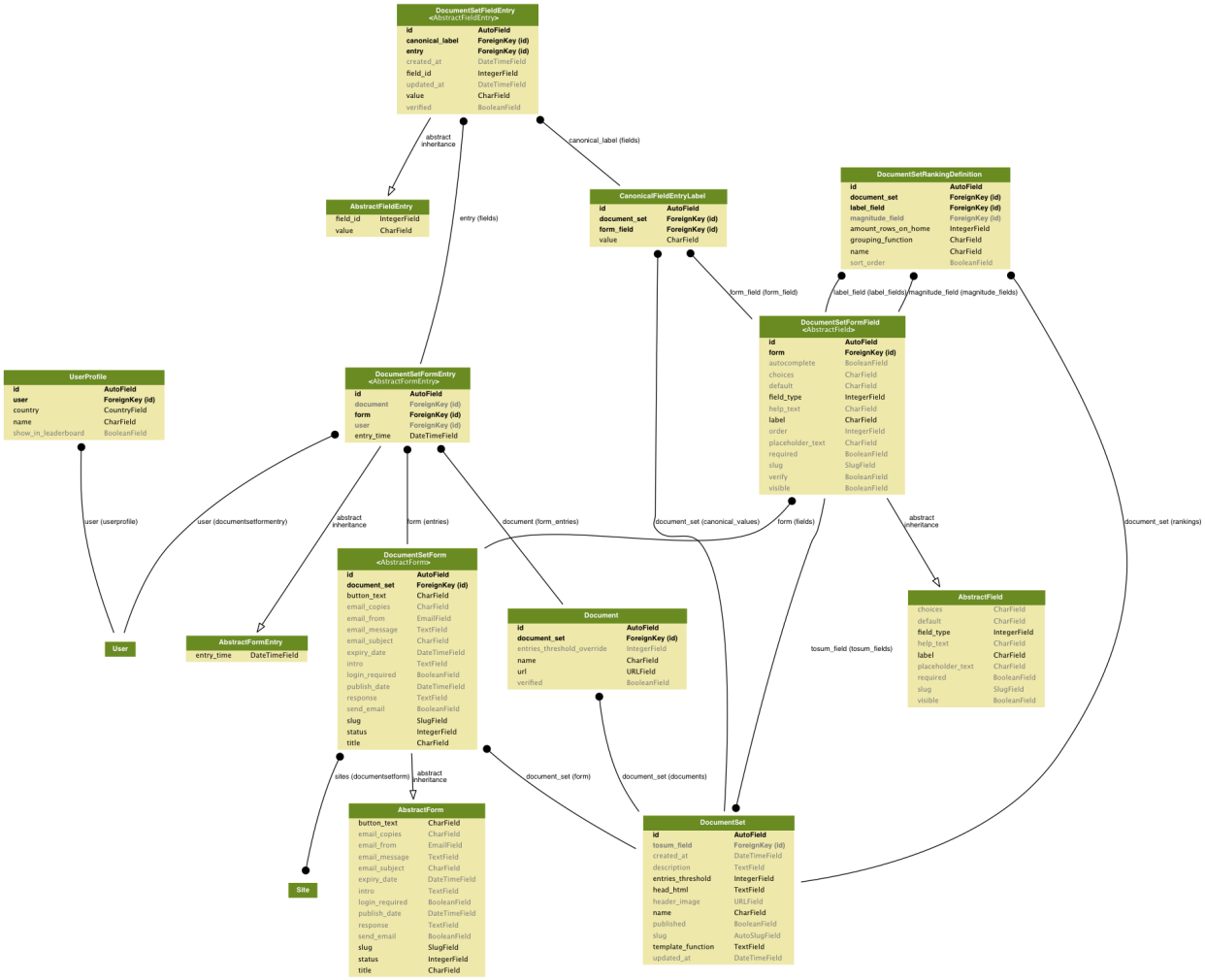
```
$ python manage.py syncdb
$ python manage.py migrate --all
```

10. Start the development server:

```
$ python manage.py runserver_plus
```

1.1.3 Schema

The arquitecture for Crowdata is generic to make it better for admin users to add new document sets and define new forms dinamically.



CHAPTER 2

Similar projects

Crowdata was inspired in the project from 'ProPublica <http://www.propublica.org>'_ called 'Free the Files <https://projects.p

It was born from a need that La Nacion had to transform scanned image PDFs into a comprehensible and structured dataset, and also to ask for community's help to catalog those spendings that call their attention.

Here some of the projects that do the same for some specific cases.

- Free the Files
- Yanukovych Leaks
- How to crowdsource MPs' expenses

CHAPTER 3

Credits

'Crowdata' is an open source project that was born when Manuel Aristaran was an Open News fellow at La Nacion in 2013. It was finally released as free software when Gabriela Rodriguez continued it for VozData in 2014. Thanks to Cristian Bertelegni and La Nacion for contributing to the code.

Now it relies on contributions from people and organizations. Please, use it, comment on it and make improvements by pull requests in GitHub.

Contributions

- Fork the repo
- Clone your fork
- Make a branch of your changes
- Make a pull request through GitHub, and clearly describe your changes

CHAPTER 5

Indices and tables

- *genindex*
- *modindex*
- *search*