CrowData Documentation

Release 0.1

Gabriela Rodriguez & Manuel Aristaran

Sep 27, 2017

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.

But the outcome of Crowdata is more than only to extract data. With Crowdata you can work with your community on a data set. They can navigate the data, help to extract it via a game and make comments on information that may be interesting to look at by journalists.

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.

Contents

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.

Installation

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', 'postgresql', 'mysql', 'sqlite3' or 'oracle'.
```

```
        'NAME': 'crowdata_development',                      # Or path to␣
↪database file if using sqlite3.
        'USER': 'crow_user',
        'PASSWORD': '',
        'HOST': '',
        'PORT': '',
    }
}
```

9. Initialize the database:

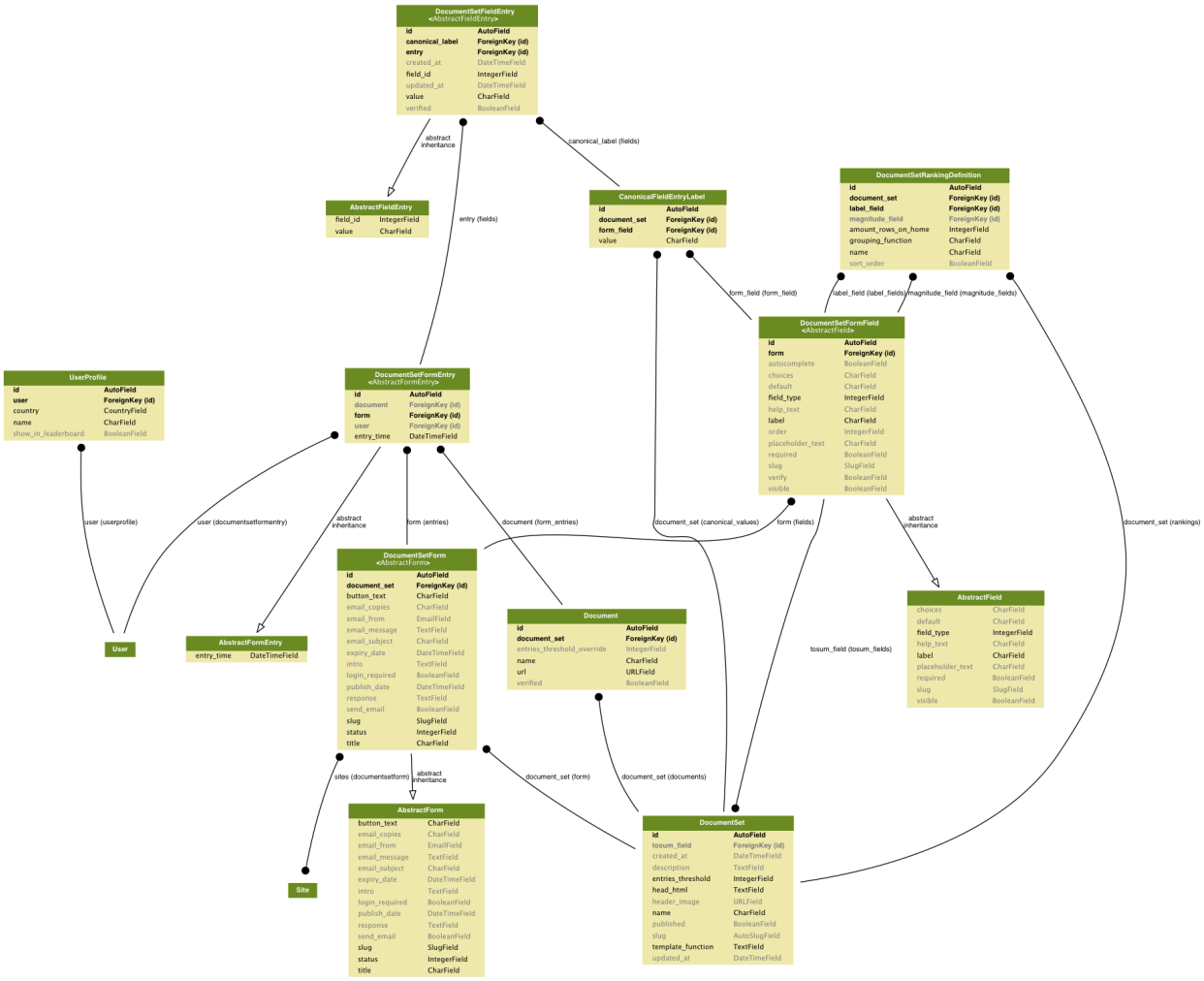
```
$ python manage.py syncdb
$ python manage.py migrate --all
```

10. Start the development server:

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

Schema

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


DocumentSetFieldEntry
<AbstractFieldEntry>
id AutoField
canonical_label ForeignKey (id)
entry ForeignKey (id)
created_at DateTimeField
field_id IntegerField
updated_at DateTimeField
value CharField
verified BooleanField
abstract inheritance
canonical_label (fields)
AbstractFieldEntry
field_id IntegerField
value CharField
entry (fields)
CanonicalFieldEntryLabel
id AutoField
document_set ForeignKey (id)
form_field ForeignKey (id)
value CharField
DocumentSetRankingDefinition
id AutoField
document_set ForeignKey (id)
label_field ForeignKey (id)
magnitude_field ForeignKey (id)
amount_rows_on_home IntegerField
grouping_function CharField
name CharField
sort_order BooleanField
form_field (form_field)
label_field (label_fields)
magnitude_field (magnitude_fields)
DocumentSetFormField
<AbstractField>
id AutoField
form ForeignKey (id)
autocomplete BooleanField
choices CharField
default CharField
field_type IntegerField
help_text CharField
label CharField
order IntegerField
placeholder_text CharField
required BooleanField
slug SlugField
verify BooleanField
visible BooleanField
UserProfile
id AutoField
user ForeignKey (id)
country CountryField
name CharField
show_in_leaderboard BooleanField
DocumentSetFormEntry
<AbstractFormEntry>
id AutoField
document ForeignKey (id)
form ForeignKey (id)
user ForeignKey (id)
entry_time DateTimeField
user (userprofile)
user (documentsetformentry)
abstract inheritance
form (entries)
document (form_entries)
document_set (canonical_values)
form (fields)
abstract inheritance
document_set (rankings)
DocumentSetForm
<AbstractForm>
id AutoField
document_set ForeignKey (id)
button_text CharField
email_copies CharField
email_from EmailField
email_message TextField
email_subject CharField
expiry_date DateTimeField
intro TextField
login_required BooleanField
publish_date DateTimeField
response TextField
send_email BooleanField
slug SlugField
status IntegerField
title CharField
Document
id AutoField
document_set ForeignKey (id)
entries_threshold_override IntegerField
name CharField
url URLField
verified BooleanField
AbstractField
choices CharField
default CharField
field_type IntegerField
help_text CharField
label CharField
placeholder_text CharField
required BooleanField
slug SlugField
visible BooleanField
tosum_field (tosum_fields)
User
AbstractFormEntry
entry_time DateTimeField
sites (documentsetform)
abstract inheritance
document_set (form)
document_set (documents)
Site
AbstractForm
button_text CharField
email_copies CharField
email_from EmailField
email_message TextField
email_subject CharField
expiry_date DateTimeField
intro TextField
login_required BooleanField
publish_date DateTimeField
response TextField
send_email BooleanField
slug SlugField
status IntegerField
title CharField
DocumentSet
id AutoField
tosum_field ForeignKey (id)
created_at DateTimeField
description TextField
entries_threshold IntegerField
head_html TextField
header_image URLField
name CharField
published BooleanField
slug AutoSlugField
template_function TextField
updated_at DateTimeField

CHAPTER 2

Similar projects

Crowdata was inspired by the project from ProPublica called Free the Files and The Guardian MP's Expenses and Sarah Palin's Emails. It was born from a need that La Nacion had to transform scanned image PDFs into a comprehensible and structured dataset, and ask for their community's help to catalog those expenses that call their attention.

Here are 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.

CHAPTER 4

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