Automating AidData's Data Extraction Process

Miranda Elliott

College of William & Mary

Summer 2014

Outline

1. Introduction
   - About AidData
   - About My Project
2. Human Process
   - What is Geocoding?
3. Automation
   - Current Program
   - Future Goals
4. Conclusion
   - Is This Useful?
   - Sources
   - Discussion

About AidData

- Institutional partnership between W&M, BYU, UT-Austin, Development Gateway and ESRI that aims to make development finance information more accessible and actionable by creating tools that enable better development policy, practice, and research and allow investors to more effectively target, coordinate, deliver, and evaluate foreign aid
- Have personally contributed to building their geographic database of project-level foreign aid through geocoding

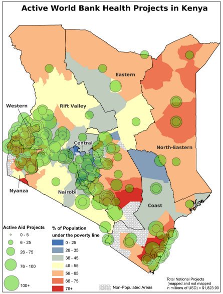


Figure 1 : Map Using Geocoded Data

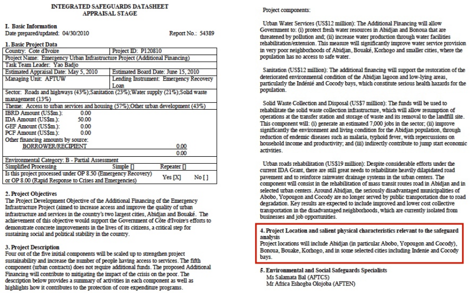
About My Project

- Ideally: Automating the geocoding process
- Realistically: Providing tools for human geocoders that will increase their efficiency

What is Geocoding?

Sample Project: P120810 - Emergency Urban Infrastructure Project (Additional Financing)

- Read the project documents and determine where the project is funding activities



INTEGRATED SAFEGUARDS DATASHEET
APPRAISAL STAGE

I. Basic Information

Date prepared/updated: 04/30/2010 Report No.: 54389

1. Basic Project Data

Country: Cote d'Ivoire	Project ID: P120810
Project Name: Emergency Urban Infrastructure Project (Additional Financing)	
Task Team Leader: Yao Badjo	
Estimated Appraisal Date: May 5, 2010	Estimated Board Date: June 15, 2010
Managing Unit: AFTUW	Lending Instrument: Emergency Recovery Loan
Sector: Roads and highways (43%);Sanitation (23%);Water supply (21%);Solid waste management (13%)	
Theme: Access to urban services and housing (57%);Other urban development (43%)	
IBRD Amount (US$m.): 0.00 IDA Amount (US$m.): 50.00 GEF Amount (US$m.): 0.00 PCF Amount (US$m.): 0.00 Other financing amounts by source: BORROWER/RECIPIENT 0.00 0.00	
Environmental Category: B - Partial Assessment	
Simplified Processing	Simple [] Repeater []
Is this project processed under OP 8.50 (Emergency Recovery) or OP 8.00 (Rapid Response to Crises and Emergencies)	Yes [X] No [ ]

2. Project Objectives

The Project Development Objective of the Additional Financing of the Emergency Infrastructure Project (aimed to increase access and improve the quality of urban infrastructure and services in the country's two largest cities, Abidjan and Bouaké. The achievement of this objective would support the Government of Côte d'Ivoire's efforts to demonstrate concrete improvements in the lives of its citizens, a critical step for sustaining social and political stability in the country.

3. Project Description

Four out of the five initial components will be scaled up to strengthen project sustainability and increase the number of people having access to services. The fifth component (urban contracts) does not require additional funds. The proposed Additional Financing will contribute to mitigating the impact of the crisis on the poor. The description below provides a summary of activities in each component as well as highlights how it contributes to the protection of core expenditure programs.

Project components:

Urban Water Services (US$12 million): The Additional Financing will allow Government to: (i) protect fresh water resources in Abidjan and Bonoua that are threatened by pollution and; (ii) increase water production through water facilities rehabilitation/extension. This measure will significantly improve water service provision in very poor neighborhoods of Abidjan, Bouaké, Korhogo and smaller cities, where the population has no access to safe water.

Sanitation (US$12 million): The additional financing will support the restoration of the deteriorated environmental condition of the Abidjan lagoon and low-lying areas, particularly the Indénié and Cocody bays, which constitute serious health hazards for the population.

Solid Waste Collection and Disposal (US$7 million): The funds will be used to rehabilitate the solid waste collection infrastructure, which will allow resumption of operations at the transfer station and storage of waste and its removal to the landfill site. This component will: (i) generate an estimated 7,000 jobs in the sector; (ii) improve significantly the environment and living condition for the Abidjan population, through reduction of endemic diseases such as malaria, typhoid fever, with repercussions on household income and productivity; and (iii) indirectly contribute to jump start economic activities.

Urban roads rehabilitation (US$19 million): Despite considerable efforts under the current IDA Grant, there are still great needs to rehabilitate heavily dilapidated road pavement and to reinforce rainwater drainage systems in the urban centers. The component will consist in the rehabilitation of mass transit routes road in Abidjan and in selected urban centers. Around Abidjan, the seriously disadvantaged municipalities of Abobo, Yopougon and Cocody are no longer served by public transportation due to road degradation. Key results are expected to include improved and lower cost collective transportation in the disadvantaged neighborhoods, which are currently isolated from businesses and job opportunities.

4. Project Location and salient physical characteristics relevant to the safeguard analysis

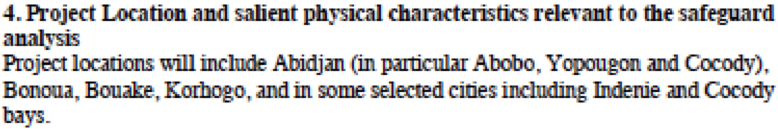
Project locations will include Abidjan (in particular Abobo, Yopougon and Cocody), Bonoua, Bouake, Korhogo, and in some selected cities including Indenie and Cocody bays.

5. Environmental and Social Safeguards Specialists

Ms Salamata Bal (AFTCS)
Mr Africa Eshogba Olojoba (AFTEN)

Figure 2 : Screenshot of World Bank Integrated Safeguards Data Sheet

What is Geocoding?



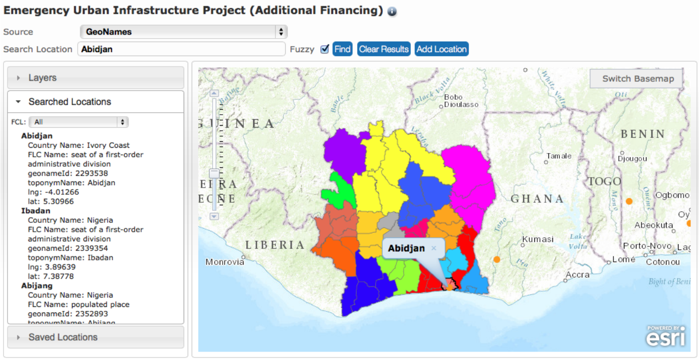
4. Project Location and salient physical characteristics relevant to the safeguard analysis

Project locations will include Abidjan (in particular Abobo, Yopougon and Cocody), Bonoua, Bouake, Korhogo, and in some selected cities including Indenie and Cocody bays.

Figure 3 : Project Locations

What is Geocoding?

- Find or add these locations to the GeoNames database



Emergency Urban Infrastructure Project (Additional Financing)

Source: GeoNames

Search Location: Abidjan

Fuzzy ☑

Abidjan
Country Name: Ivory Coast
FLC Name: seat of a first-order administrative division
geonameId: 2293538
toponymName: Abidjan
lng: -4.01266
lat: 5.30966

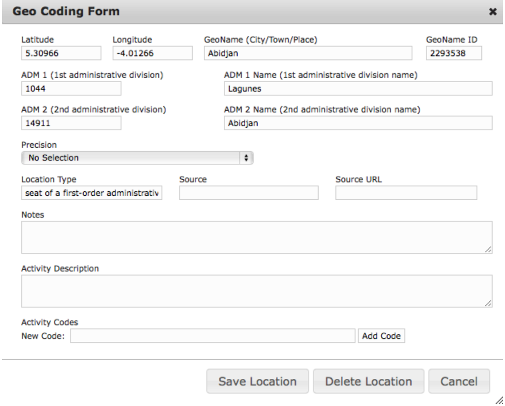
Ibadan
Country Name: Nigeria
FLC Name: seat of a first-order administrative division
geonameId: 2339354
toponymName: Ibadan
lng: 3.89639
lat: 7.38778

Abijang
Country Name: Nigeria
FLC Name: populated place
geonameId: 2352893
toponymName: Abijang

Figure 4 : Screenshot of Toolkit Location Search

What is Geocoding?

- Determine what activities are being funded at this location, the type of location (city vs. province vs. river) , and list where you found this information



Geo Coding Form

Latitude: 5.30966
Longitude: -4.01266
GeoName (City/Town/Place): Abidjan
GeoName ID: 2293538

ADM 1 (1st administrative division): 1044
ADM 1 Name (1st administrative division name): Lagunes

ADM 2 (2nd administrative division): 14911
ADM 2 Name (2nd administrative division name): Abidjan

Precision: No Selection

Location Type: seat of a first-order administrativ
Source:
Source URL:

Notes:

Activity Description:

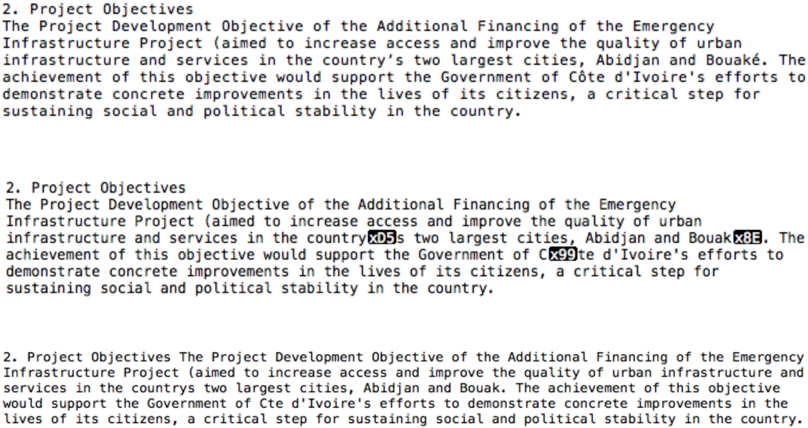
Activity Codes
New Code:

Figure 5 : Screenshot of Location Coding Form

Preparing the Document for Text Mining

- Convert document from PDF to TXT
- Strip non-ASCII characters

Preparing the Document for Text Mining



```
2. Project Objectives
The Project Development Objective of the Additional Financing of the Emergency
Infrastructure Project (aimed to increase access and improve the quality of urban
infrastructure and services in the country’s two largest cities, Abidjan and Bouaké. The
achievement of this objective would support the Government of Côte d'Ivoire's efforts to
demonstrate concrete improvements in the lives of its citizens, a critical step for
sustaining social and political stability in the country.
```

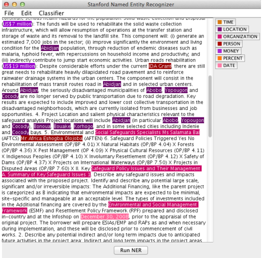
```
2. Project Objectives
The Project Development Objective of the Additional Financing of the Emergency
Infrastructure Project (aimed to increase access and improve the quality of urban
infrastructure and services in the countryxD5s two largest cities, Abidjan and Bouakx8E. The
achievement of this objective would support the Government of Cx99te d'Ivoire's efforts to
demonstrate concrete improvements in the lives of its citizens, a critical step for
sustaining social and political stability in the country.
```

```
2. Project Objectives The Project Development Objective of the Additional Financing of the Emergency
Infrastructure Project (aimed to increase access and improve the quality of urban infrastructure and
services in the countrys two largest cities, Abidjan and Bouak. The achievement of this objective
would support the Government of Cte d'Ivoire's efforts to demonstrate concrete improvements in the
lives of its citizens, a critical step for sustaining social and political stability in the country.
```

Figure 6 : Before Conversion to TXT; After Conversion to TXT & Before Non-ASCII Strip; After Non-ASCII Strip

Stanford Named Entity Recognizer

- CRF classifier
- Type of discriminative undirected probabilistic graphical model
- Predicts sequences of labels for sequences of input samples by encoding known relationships between observations from a training set of documents
- Using 7-class model trained on MUC



constitute serious health hazards for the population. Solid waste Collection and Disposal (US$7 million): The funds will be used to rehabilitate the solid waste collection infrastructure, which will allow resumption of operations at the transfer station and storage of waste and its removal to the landfill site. This component will: (i) generate an estimated 7,000 jobs in the sector; (ii) improve significantly the environment and living condition for the Abidjan population, through reduction of endemic diseases such as malaria, typhoid fever, with repercussions on household income and productivity; and (iii) indirectly contribute to jump start economic activities. Urban roads rehabilitation (US$19 million): Despite considerable efforts under the current IDA Grant, there are still great needs to rehabilitate heavily dilapidated road pavement and to reinforce rainwater drainage systems in the urban centers. The component will consist in the rehabilitation of mass transit routes road in Abidjan and in selected urban centers. Around Abidjan, the seriously disadvantaged municipalities of Abobo, Yopougon and Cocody are no longer served by public transportation due to road degradation. Key results are expected to include improved and lower cost collective transportation in the disadvantaged neighborhoods, which are currently isolated from businesses and job opportunities. 4. Project Location and salient physical characteristics relevant to the safeguard analysis Project locations will include Abidjan (in particular Abobo, Yopougon and Cocody), Bonoua, Bouake, Korhogo, and in some selected cities including Indenie and Cocody bays. 5. Environmental and Social Safeguards Specialists Ms Salamata Bal (AFTCS) Mr Africa Eshogba Olojoba (AFTEN) 6. Safeguard Policies Triggered Yes No Environmental Assessment (OP/BP 4.01) X Natural Habitats (OP/BP 4.04) X Forests (OP/BP 4.36) X Pest Management (OP 4.09) X Physical Cultural Resources (OP/BP 4.11) X Indigenous Peoples (OP/BP 4.10) X Involuntary Resettlement (OP/BP 4.12) X Safety of Dams (OP/BP 4.37) X Projects on International Waterways (OP/BP 7.50) X Projects in Disputed Areas (OP/BP 7.60) X II. Key Safeguard Policy Issues and Their Management A. Summary of Key Safeguard Issues 1. Describe any safeguard issues and impacts associated with the proposed project. Identify and describe any potential large scale, significant and/or irreversible impacts: The Additional Financing, like the parent project is categorized as B indicating that environmental impacts are expected to be minimal, site-specific and manageable at an acceptable level. The types of investments included in the Additional financing are covered by the Environmental and Social Management Framework (ESMF) and Resettlement Policy Framework (RPF) prepared and disclosed in-country and at the Infoshop on December 30, 2008, prior to the appraisal of the original project. The borrower will prepare ESIAs/EMP and RAPs as and when necessary during implementation, and these will be disclosed prior to commencement of civil works. 2. Describe any potential indirect and/or long term impacts due to anticipated future activities in the project area: Indirect and long term impacts in the project areas

Figure 7 : Screenshot of Stanford NER software ran on World Bank project document

Stanford Named Entity Recognizer

- Connect to remote Stanford NER server with pyner module
- Create list of all tokens tagged as locations
- *In sample project:* found no incorrect locations (aside from 'bouak' which is misspelling of 'bouake') and every correct location but 'indenie'

Output

```
['abidjan', 'abobo', 'bonoua', 'bouak', 'bouake', 'cocody', 'korhogo',
'yopougon']
```

Creating the GeoNames Dictionary

- Connect to GeoNames with geonamescache module
- Create dictionary with GeoNames location names as keys and tuples of their corresponding latitude and longitude, country code, and Geoname ID as values

Sample Dictionary Entry

```
{retiro :  [((40.41317, -3.68307), 'ES', '6544495'), ((-34.58333,
-58.38333), 'AR', '3429576')]}
```

Fuzzy Search

- Use NLTK's Levenshtein distance calculator to search for location matches with a distance less than or equal to 1 from the Geonames dictionary
- Levenshtein distance is the minimum number of single-character edits required to change one word into the other
- Create list of GeoNames location matches
- *In sample project:* eliminated 'bouak' (misspelling of 'bouake'), added incorrect location 'bouar'

Output

```
['abidjan', 'abobo', 'bonoua', 'bouake', 'bouar', 'korhogo']
```

Searching for Missed Locations

- Search through sentences containing identified locations for other locations in GeoNames
- Add missed locations to list of original GeoNames location matches
- *In sample project:* no missed locations found

Output

```
['abidjan', 'abobo', 'bonoua', 'bouake', 'bouar', 'korhogo']
```

Eliminating Geographic Outliers

- Determine country that majority of locations lie within and eliminate locations not within this country
- Calculate $Q1$, $Q3$, and IQR for latitude and longitude of each remaining location
- Eliminate locations with either coordinate value outside of their respective range of $Q1 - 1.5 * IQR$ to $Q3 + 1.5 * IQR$
- *In sample project:* eliminated incorrect location 'bouar'

Output

```
['abidjan', 'abobo', 'bonoua', 'bouake', 'korhogo']
```

Current Final Products

- Output list of determined locations
- Output list of sentences containing any location in that list for human geocoders to read and manually determine project activities happening there and other locations the program potentially missed

Current Final Products

Locations

```
[’abidjan’, ’abobo’, ’bonoua’, ’bouake’, ’korhogo’]
```

Sentences

```
[’ project objectives the project development objective of the additional financing of the emergency infrastructure
project (aimed to increase access and improve the quality of urban infrastructure and services in the countrys two
largest cities, abidjan and bouak’, ’ (i) protect fresh water resources in abidjan and bonoua that are threatened
by pollution and’, ’ this measure will significantly improve water service provision in very poor neighborhoods of
abidjan, bouak, korhogo and smaller cities, where the population has no access to safe water’, ’ the additional
financing will support the restoration of the deteriorated environmental condition of the abidjan lagoon and
low-lying areas, particularly the indni and cocody bays, which constitute serious health hazards for the
population’, ’ (ii) improve significantly the environment and living condition for the abidjan population, through
reduction of endemic diseases such as malaria, typhoid fever, with repercussions on household income and
productivity’, ’ the component will consist in the rehabilitation of mass transit routes road in abidjan and in
selected urban centers’, ’ around abidjan, the seriously disadvantaged municipalities of abobo, yopougon and cocody
are no longer served by public transportation due to road degradation’, ’ project location and salient physical
characteristics relevant to the safeguard analysis project locations will include abidjan (in particular abobo,
yopougon and cocody), bonoua, bouake, korhogo, and in some selected cities including indenie and cocody bays’, ’
municipalities of abidjan, bouake, and korhogo’, ’ disadvantaged areas of abobo, yopougon and cocody’, ’ as a
result, urban poverty and overcrowding has dramatically increased, including in the two largest cities of the
country abidjan and bouak, where nearly half of the population of 20 million resides today’, ’ helping the
government to ensure the delivery of basic infrastructure and social services to urban populations living in
difficult and unsanitary conditions, and expanding these services to more people in abidjan and other cities,
notably in the cnw, is a key part of the banks strategy to support crisis recovery and sustainable peace and
development’, ’ to increase access and improve the quality of urban infrastructure and services in the cities of
abidjan, bouak, korhogo and selected smaller cities’, ’ (i) protect fresh water resources in abidjan and bonoua
that are threatened by pollution’, ’ this measure will significantly improve water service provision in very poor
neighborhoods of abidjan, bouak, korhogo and smaller cities, where the population has no access to safe drinking
water’, ’ $15 million of the crisis response widow (crw) stage i and $4 million from ida resources will finance
rehabilitation of mass transit routes road in abidjan and in selected urban centers’, ’ around abidjan, the
seriously disadvantaged municipalities of abobo, yopougon, and cocody are no longer served by public transportation
due to road degradation’, ’ increase access to and improve the quality of urban infrastructure and services in the
countrys two largest cities, abidjan and bouak’]
```

Performance Analysis

- Correct locations found: 5
- Correct locations not found: 3
- Incorrect locations found: 0
- All missed correct locations contained in output sentences

Correct Locations

```
['abidjan', 'abobo', 'bonoua', 'bouake', 'cocody', 'indenie', 'korhogo',
'yopougon']
```

Found Locations

```
['abidjan', 'abobo', 'bonoua', 'bouake', 'korhogo']
```

Future Goals

- Find a more comprehensive and regularly updated Geonames connection
- Be able to distinguish between paragraphs and tables, as Stanford NER is only effective in identifying locations in sentences, and find a method for identifying locations in tables
- Determine the most accurate parameters for labeling locations as outliers and whether any other trends exist regarding locations that can generally be eliminated (ex. countries, capital cities)
- Improve overall efficiency of implementation within program

Is This Useful?

- Currently, up to 3 people are working on geocoding each project - 2 interns code seperately, and if their codes don't match a research assistant corrects them
- For the sample project: without this program a human geocoder would read 20 pages, with this program they would read under 500 words and come to the same conclusions
- In its current state, this program can speed up the geocoding process, and when its results are deemed reliable and consistent it could reduce the need for such significant manpower per project

Sources

- AidData Brings $25 Million Award to W&M to Establish AidData Center for Development Policy. (2012, November 8). College of William & Mary. Retrieved July 8, 2014, from http://www.wm.edu/offices/itpir/news/aiddata-brings-25-million-award-to-wm-to-establish-aiddata-center-for-development-policy.php
- Conditional Random Field. (n.d.). Wikipedia. Retrieved July 8, 2014, from http://en.wikipedia.org/wiki/Conditional-random-field
- Custer, S. (2010, August 8). Mapping For Results. AidData. Retrieved July 8, 2014, from http://aiddata.org/blog/mapping-for-results
- Levenshtein Distance. (n.d.). Wikipedia. Retrieved July 8, 2014, from http://en.wikipedia.org/wiki/Levenshtein-distance
- NLTK 3.0 Documentation. (2014, July 8). Natural Language Toolkit. Retrieved July 8, 2014, from http://www.nltk.org
- Our Story. (n.d.). AidData. Retrieved July 8, 2014, from http://aiddata.org/our-story
- Stanford Named Entity Recognizer. (n.d.). Stanford Natural Language Processing Group. Retrieved July 8, 2014, from http://nlp.stanford.edu/software/CRF-NER.shtml
- The AidData Center for Development Policy. (n.d.). College of William & Mary. Retrieved July 8, 2014, from http://www.wm.edu/offices/itpir/aiddata/aiddata-center-for-development-policy/index.php

Discussion

Questions or Suggestions?