AIDDATA
A Research Lab at William & Mary

DOCUMENTATION

JULY 2017

Geocoding Methodology (Version 2.0.2)

Prepared By: AidData Research and Evaluation Unit

AidData Geocoding Methodology (Version 2.0.2)

Contents

AidData Geocoding Methodology (Version 2.0.2)

1. Introduction

This codebook prescribes AidData's method for geocoding foreign aid information and programmatic development data that is made available through a wide range of donor, recipient, and other stakeholder-based sources. In broad terms, **geocoding** is a process by which an address is assigned a single data point with a corresponding latitude and longitude. **Georeferencing**, not to be mistaken for geocoding, is a process in which an internal coordinate system of a map, or a satellite/aerial image, is spatially referenced. Coders first "geoparse" through project documents to identify information on sub-national locations. Next, they "georeference" this location information (finding the matching location on a map). Finally, they "geocode" said locations by assigning a specific latitude and longitude coordinate. AidData's Geocoding Methodology was initially derived from the UCDP Georeferenced Event Dataset (GED) Codebook version 1.0 (Sundberg et al., 2010[1]), which covers the georeferencing of violent events. The UCDP system was adapted and complemented by additional protocols to enable the coding of aid projects, and has since been updated to align with the internationally recognized best practices of the IATI standard (Strandow et al., 2011[2]).

The current AidData methodology allows for a hierarchy of **location class** through 4 unique classifications including populated places, administrative boundaries, structures and topographical features. The specificity of the prescribed geographic location is defined through an additional code for **geographic exactness**. Locations are coded to the level of subnational precision — what we describe as *"granularity"* — that the supporting documentation confidently allows. Sources vary significantly in the quality of location information that is reported; in some cases the exact locations are named, and in other instances only a general area or a proximate location is provided. The main objective of the AidData methodology is to track all locations to which aid dollars are committed or disbursed. Locations that may be expected to benefit indirectly from development finance are not geocoded. The implied mandate for this methodology is to "follow the money."

1 Strandow, Daniel & Findley, Michael & Nielson, Daniel & Powell, Josh. (2011). The UCDP and AidData Codebook on Georeferencing Aid. Version 1.1.

2 Sundberg, Ralph & Mathilda Lindgren & Ausra Padskocimaite. (2010). UCDP Geo-referenced Event Dataset (GED) Codebook. Version 1.0.

2. Roadmap for Geocoding

The AidData geocoding methodology is designed to be flexible enough to accommodate project records from a wide range of sources. Donor-sourced geocoding is referred to as a *"top-down"* process (e.g. World Bank *Mapping for Results*), whereas recipient-sourced geocoding is referred to as a *"bottom-up"* process (e.g. Uganda Aid Management Platform). In the case of a "top-down" dataset, the geocoding process is typically single-sourced; all documentation is provided by the donor institution itself. There are some exceptions, as is the case with our Tracking Underreported Financial Flows (TUFF) geocoded datasets on development finance from non-traditional donors. These datasets are geocoded from open source documentation that is collected, curated, and standardized by AidData. A "bottom-up" dataset is most commonly sourced from the recipient in the form of a standardized, multi-source Aid Information Management System (AIMS). AIMS are typically housed in the Ministry of Finance and/or Ministry of Planning within an aid recipient country. The records within these systems serve as the project documentation used in a bottom-up geocoding process. For all source types, the time intensity of the coding process will vary significantly based on the quality and availability of project documentation.

2.1 AidData's 4-Step Methodology

For a given project record, the geocoding process will typically include the following steps:

1. **Review basic project information and relevant documentation:** coders are required to review the project title, description, and all of the most current documentation[3] available to ensure accurate location information is accessed and recorded.

2. **Augment documentation with additional sources (optional):** may include a cross-check with donor or recipient databases as well as targeted web searches using project title, donor, recipient and/or unique project ID.

3. **Code specified location(s):**

    a. Identify correct location (*must* match name and location type described in the documentation)

    b. Apply appropriate location class and exactness codes

    c. Provide source name and URL

    d. Written notes with justification for the location including source page number, as well as a description of third party documentation and references (if used). These notes are for internal use only and will not be included in the final geocoded data product.

    e. Brief project description (optional) to be included in the final geocoded data product.

4. **Arbitration:** reconciliation of two geocoded records (may be manual or automated)

AidData's geocoding methodology traditionally prescribes a **"double-blind" approach** to the geocoding process as a quality assurance (QA) measure. Each unique project record is assigned to two separate coding Research Assistants (RAs) without cooperation. The two geocoded records are then reconciled by a trained arbitrator to create one finalized project record be included in the raw dataset. In cases where the results of the two initial code rounds are identical, the project is automatically arbitrated and finalized.

3 As the priorities and geographic footprint of projects are subject to change over the course of implementation, geocoders are advised to use caution with regard to highly detailed subnational information available in antiquated documentation. Information derived from such sources must be carefully evaluated to ensure their accuracy.

3. Geocoded Data Products: Locating aid events in time and space

The data that has been produced under this coding scheme is compatible with a wide variety of data sources on development activities, including development results and indicator data, domestic budgets, and surveys. AidData's geocoded datasets are released as a suite of relational data products that capture aid projects and their corresponding locations and transactions to allow users to configure the data in a way that is most appropriate to their analysis.

It is important to note that the unit of analysis for the financial component of the geocoded data is based on **funding commitments** as opposed to disbursed aid or calendar days (PLAID, 2010). Since data on the exact dates and locations of funding disbursements is sparse, most geocoded projects can only be related to the year that a specific commitment was made. With this in mind, AidData's methodology will not capture commitment or disbursement amounts by location, as such granular financial information is rarely available. Transaction data can only be described at the project level. The final geocoded data release will include a **Transactions Table**, where each row represents an individual financial commitment for a given project. There is often a one-to-many relationship between transactions and development projects, where a single project may consist of multiple transactions.

Likewise, in many cases, a single development project is designed to reach multiple locations. This one-to-many relationship between development projects and their locations is addressed through the **Projects Table** and **Locations Table** within the geocoded data release. The projects table will provide a complete listing of all projects covered in the dataset where each row represents a single project record. When aid projects are intended for multiple locations, we include an additional row of data in the locations table for each unique location for a given project record. Under this model, a locations table for our geocoded data will often include multiple rows for a single development project.

4. Location Type, Geographic Exactness and Precision Categories

4.1 IATI Standard

For geocoded data to be useful for a wide range of applications, it is crucial to make it possible to select subsets of the data based on levels of granularity. To facilitate this, AidData has adopted the IATI standard for describing the **location class** and **geographic exactness** of a given geocoded location. Coders select one location code and one exactness code for each location of a given project record. These fields are included in our geocoded datasets to help data users select portions of the data for analysis on the basis of their precision.

IATI Standard			
Location Class		**Geographic Exactness**	
1	Administrative Regions (e.g. state, province, independent political entity)	1	Exact
2	Populated Place (e.g. city, village)	2	Approximate
3	Structure (e.g. building, bridge, road)		
4	Other topographical feature (e.g. river, mountain, national park)		

The location class specification within the IATI standard disaggregates locations into 4 discrete categories to categorize geographic features. The binary geographic exactness specification describes whether the geocoded location is the final expected destination of the financial flow, or an approximation based on the best available information.

AidData Use of Exact and Approximate Markers

The use of the IATI geographic exactness specification may follow one of two distinct approaches (Method A and B). The first prescribes exactness on the basis of the geographic feature being coded, the second on the basis of the specificity of project documentation.

Method A: An early iteration of AidData's use of the exactness specification followed the first model. Locations were coded on the basis of the presence of a clear geographic boundary. Under this model, populated places and administrative divisions would always be coded as "exact" regardless of the nature of the project or available documentation since the geographic boundary of these locations is fixed. The "approximate" code would be used only in rare cases where the geographic boundary of an area is contested or vague.

Example A: For instance, a project aims to build 3 new schools in the district of Kigali. While the precise locations of the schools are unknown, under Method A, the district of Kigali would be coded as ***exact****, since a district has a precise geographic boundary.*

Method B: In an effort to make the exactness specification more useful in analysis, we have opted to use the second model, where exact and approximate codes describe the precision of the location information provided. The "exact" specification will be used when the coder is confident that they have coded the end destination of a financial flow. Flows that can only be traced to a general area or a proximate location will be coded as "approximate."

Example B: Under Method B, the same project described in Example A would take an ***approximate*** *code, since the precise locations of the schools -- the true target locations -- are unknown.*

4.2 AidData Precision Codes

Prior to the new IATI standard, AidData's original geocoding methodology specified levels of granularity based on an eight-point precision code system. The first six categories were adapted from UCDP's Geo-referencing Project Codebook (Sundberg et al., 2010), with only minor modifications to effectively capture development finance locations. Precision categories 7 and 8 were unique to the AidData Geocoding Methodology first introduced in the UCDP/AidData codebook (Strandow et al., 2011).

AidData Precision System	
1	The coordinates correspond to an exact location, such as a populated place or a physical structure such as a school or health center. This code may also used for locations that join other locations to create a line such as a road, power transmission line or railroad.
2	The location is mentioned in the source as being "near", in the "area" of, or up to 25 km away from an exact location. The coordinates refer to that adjacent location.
3	The location is, or is analogous to, a second-order administrative division (ADM2), such as a district, municipality or commune.
4	The location is, or is analogous to, a first-order administrative division (ADM1), such as a province, state or governorate.
5	The location can only be related to estimated coordinates (e.g. between populated places; along rivers, roads and borders; or more than 25 km away from a specific location). Also used large topographical features (greater than ADM1) such as National Parks which spans across several administrative boundaries.
6	The location can only be related to an independent political entity, but is expected to be disbursed locally. This includes aid that is intended for country-wide projects as well as larger areas that cannot be geo-referenced at a more precise level.
7	The location is unclear. The country coordinates are entered to reflect that subnational information is unavailable.**

	**This code was removed from the geocoding methodology in 2016. It will only appear in datasets published up to 2016.
8	The location can only be related to an independent political entity, but the central government will be the only direct beneficiary (e.g. capacity building, budget support, technical assistance).

4.3 The IATI-to-AidData Precision Crosswalk

To effectively accommodate the preferences of researchers while harmonizing with international standards for open development data, we retain AidData precision codes along with the IATI classification for location type and geographic exactness in our geocoded data products. To achieve this, source data is geocoded based on the IATI system and then translated to AidData precision codes using an IATI-approved crosswalk between the two systems.

IATI-Precision Crosswalk				
Location Class		Geographic Exactness		Precision Code
2/3	+	1	=	1
2/3	+	2	=	2
1	+	1	=	3
1	+	1	=	4
4	+	2	=	5
1	+	1	=	6
1	+	2	=	8

5. About the Geocoding Process

5.1 Reviewing Project Information & Documentation

Each aid project may have location information on several levels. First, a project title or brief description may contain a clear location reference. For some datasets, these represent the sum total of available source information. However, these sources may not reflect comprehensive information on all of the intended target beneficiaries, or the most granular location information. After reviewing title and description for location information, project documentation is assessed. As location information may be provided in more than one project document, coders are advised to employ reasonable due diligence and exhaust all possible current sources of information.

If there is no direct mention of any location in the sources, and the title and abstract do not indicate that aid is granted to the central government for the purpose of capacity building/technical assistance, aid is assumed to go to the country in general. The country coordinates are coded with a location class of 1 for "Administrative Region" and an exactness code of 1 (corresponds to Precision Code 6). The data user should use their discretion to determine if unclear aid locations coded at the national level should be included in their analysis.

5.2 Augmenting with Additional Sources

In this optional step, coders use established best practices and workspace-specific guidelines to identify additional sources of location information. These searches tend to be most successful with rich project-level information, including but not limited to: unique project IDs, implementation timelines, and proper names of implementing partners. Coders are required to furnish the urls to these sources and clearly identify the relationship between the new source and the project record. Locations identified through external sources are authenticated through manual arbitration.

5.3 Standardizing Location Information

The AidData geocoding methodology derives coordinates for coded locations through the Geonames[4] gazetteer. Geonames provides an online service, which contains names, administrative hierarchies and coordinates of administrative divisions, populated places, waterways, and other geographic features. The database currently consists of over 10 million records, including 2.8 million populated places and over 5 million alternate location names. The latitude and longitude coordinates are recorded with a six decimal precision. The map projection used is the standard World Geodetic System 1984 (WGS 84) (Sundberg et al, 2010). Coders may augment Geonames with additional locations as needed if they are able to identify appropriate coordinates for a specified location through Google Earth or other recognized secondary sources of geographic information. All new locations are added to the Geonames database in the process of geocoding to maintain a stable record of all location information.

4 http://www.geonames.org/v3/

6. Basic Rules for Geocoding

6.1 Finding Locations

Once location information has been identified in the documentation, the coder will consult the Geonames gazetteer via AidData's internal coding platform or Geonames.org to find a matching location record. If no exact matches are available within the correct country, coders will perform a "fuzzy" search to capture a broader range of location matches and alternate naming conventions. If the correct location is not currently available via Geonames, the coder may use other third party resources to identify appropriate coordinates. Third party resources may include Google Earth, Google Maps, Wikipedia (with appropriate references), and GADM.org (database of administrative boundaries for countries all over the world). Once the correct location is identified, it is entered into the coding interface along with the source information and notes to guide manual arbitration of the new location.

6.2 Verifying Locations

Many place names are not unique, even within a specified country. For example, there are at least 35 unique populated places called Washington within the United States. In cases such as these, a location name may not be sufficient to identify the correct location. Coders must use additional information regarding the administrative boundaries, location classification, and nearby landmarks to triangulate the correct location with confidence. This issue is covered in greater depth in Section 7.

6.3 Coding Points

Locations such as cities and villages (Location Class 2), hills (Location Class 4), farms and buildings (Location Class 3) will be represented as a single point location. Geographic exactness will be specified on the basis of the project activity. If aid flows are expected to be disbursed in that precise location, the point is specified as exact. Geocoders work to identify coordinates for the most granular location possible. However, coordinates for specific structures and buildings are often scarce outside major landmarks such as government buildings, universities and hospitals in primary cities. In cases where the coordinates for a given structure are unavailable, the corresponding city/town/village is coded as approximate. Likewise, in cases where sources refer to a unit within an institution or structure, such as a specific department within a university, the coder will use the most specific geographic point available -- often the parent location. This is also true for intervention sites that may not be attached to a precise coordinate pair (such as a microfinance community group) or would benefit a community as a whole or are frequently impossible to locate (such as a borehole for fresh water). In these cases, the populated place may be coded as exact since it is the most granular known location.

In cases where multiple components of a project are expected to be executed in the same location, the location is only coded once for the whole project.

Example: For instance, if there are funds going to farms somewhere in the location Bengo (location class 2), as well as to support hospitals in the same area, then Bengo would be coded once as approximate with notes for each component.

Suburbs are borderline cases. Suburbs to cities should be considered to be locations in their own right and are coded accordingly if the coordinates are available, even if the city itself is also coded. If the coordinates of a major suburb are not available, the city (location class 2) is specified as approximate.

As a rule, if a point location cannot be identified via the search functions in Geonames or by using other resources, then the coordinates of the broader administrative boundary are coded as approximate. For example, if a dam is being built as part of the project and the exact location of the dam cannot be identified, the most granular known geographic boundary will be coded as approximate rather than estimating a point in or near the corresponding body of water. However, if the dam itself can be visually located, then the location should be added to the geonames database and be used to geocode.

6.4 Coding Areas

There may be some cases where project documentation will refer to a broader area rather than a specific point location. The most common cases will be with regard to Administrative Divisions (ADM1 and ADM2) and Political Entities (location class 1), but there may also be instances that involve large geographic features such as mountains, lakes, and protected areas (location class 4). In AidData's geocoded data products, the latitude and longitude representation of such areas are estimated as the centroid point of that area. Geographic exactness will be specified on the basis of the project activity. Projects that are expected to be disbursed comprehensively throughout a given area will be specified as exact. Projects that are disbursed to the local government or to a subset of locations within a given area will be specified as approximate. In instances where the administrative boundaries noted in the documentation have been changed, coders will code boundaries that represent the historic area as closely as possible.[5]

Islands present a somewhat unique case for the IATI location class specification. In cases where the Island as a whole conforms to the administrative system, the administrative boundary is coded with a location class 1. In cases where the island does not conform to the administrative system, or is a part of a larger administrative boundary, the island is coded with a location class 4. Geographic exactness is specified based on the same rules as prescribed above.

In the final AidData geocoded data product, the names of ADM1 and ADM2 boundaries associated with each geocoded location are saved in the data as text/strings in the "ADM1" and "ADM2" columns. In cases where there is not an administrative boundary system in a given country (e.g. small island nations), these fields are left blank in the finalized dataset.

6.5 Coding Lines

In cases where project documentation specifies activities along a linear path, as in the case of large-scale infrastructure projects such as roadways and power lines (location class 3), the locations may be captured using one of two methods. When the path of the line is available via OpenStreetMap or other sources, the line may be traced to create a line vector location (location class 3) to be coded as exact. Alternatively, if the path is not documented or line vector functionality is unavailable, then the point of origin, all identified through-points, and the endpoint are coded (location class 2) as exact, as well as the corresponding ADM2s that the line goes through (location class 1) as approximate.

In cases where the path of the line is too vague to confidently identify all corresponding ADM2s, the ADM1 (location class 1) is coded as approximate. The ADM2s are coded in addition to the point locations to reflect the fact that funding is allocated across the entire road rather than simply to the point locations along the linear path. In cases where the documentation refers to a section of a linear feature as the intended destination for funding, that portion alone is coded rather than the entire feature. In cases where points of origin, through-points, and/or endpoints are not specified, and the path of the linear feature remains unclear, the known administrative divisions are coded as approximate.

Example: An unknown road running from Nairobi, Kenya to Mombasa, Kenya would necessitate five geocoded locations: (1) Nairobi (location class 2) specified as exact; (2) Mombassa (location class 2) specified as exact; (3) Nairobi Area Province (location class 1) specified as approximate; (4) Eastern Province -- the sole administrative division between Nairobi and Mombassa (location class 1) specified as approximate; and (5) Coast Province (location class 1) specified as approximate. These geocoded locations represent the fact that funding is allocated across the entire road through each province, rather than simply to the endpoints.

5 There may be some cases where recipient nations have revised the boundaries of their administrative divisions or have created entirely new administrative divisions by carving areas out of larger areas. The priority in such instances is to best approximate the area that is intended in the source. Thus, if a province is divided into several new provinces, each of the new provinces within the boundaries of the defunct province is coded. A more difficult case occurs when countries decrease the number of provinces. In this case, the current province which contains the territory of the defunct province is coded and a note is made of the defunct province as the intended recipient.

7. Advanced Rules for Geocoding

The advanced rules are designed to support the coder when source documentation provides only vague indications of locations that receive funding. In some cases, these issues can be addressed through desk research and information provided through trusted third-party sources. When that is not possible due to time or other constraints, or when additional sources have been exhausted, these advanced rules are used to code ambiguous locations. The advanced rules of the AidData methodology are derived from two key guiding criteria:

- *Be Conservative:* assign aid to larger geographic areas when needed in an effort to capture the correct area.
- *Strive for Granularity:* work to locate coordinates that reflect meaningful locations rather than artificial points (such as centroids of administrative divisions).

7.1 Locations with Ambiguous Names

Some sources provide the name of a location, but do not include a clear indication of the location type. As alluded to in section 6.2, this can be especially problematic when the location is a common name with multiple matches in the Geonames gazetteer. In cases such as these, coders will employ a number of methods (described below) to triangulate the correct location with the highest possible confidence. Coding vague locations will typically involve a search for secondary documentation in an effort to verify and augment existing location information.

One form of ambiguity emerges when there are several options in Geonames, but all available locations are a slight variation on the name mentioned in the project's source documents. In this case, the coder will undertake the following steps to mitigate the ambiguity:

1. Eliminate options that do not match the appropriate feature class and expected administrative boundaries
2. Review alternate naming conventions included in the Geonames record
3. Perform desk research to determine if the location name has multiple spellings due to translation etc.
4. Use contextual information regarding the project or target area to triangulate

Coders are expected to use reasonable due diligence to identify the correct location with the highest possible confidence. In the event that the correct location is identified with an alternate spelling, the location is coded as exact. In cases where only a proximate landmark can be identified, such as a hill or well of the same name, that location may be coded as approximate. If the coder opts to code a location name that is not an exact match, they must make a clear argument for the use of the alternate location in the "notes" field. As a general guideline, in the absence of additional information to verify an alternate spelling, locations that differ by more than 1 character should not be coded.

Example: A project is intended for "Lang Port", and the options in the gazetteer that are closest to that location name are "Lange" and "Langa" (populated places). If Langa and Lange both lie within the expected administrative boundary, but Langa is closer to the body of water where the port is located, then Langa (location class 2) would be coded as approximate.

In cases where multiple location options are of the same name and are of the same feature class, coders should use contextual information from the project documentation to corroborate one of the points. Factors to consider in this process include the relative distance to other specified landmarks, the nature of project activities, and other target locations of the project.

Example: For instance, there are three towns (location class 1) named Sabon Birnin in Nigeria, one each in the provinces of Kaduna, Sokoto, and Kebbi. If all of the other locations in the documentation are found in Kaduna province, the Kaduna location would be the most likely target, and should be coded as exact.

When all source documentation fails to provide sufficient information to mitigate ambiguity or otherwise triangulate the specified location, the coder will default to the next order administrative boundary. Under this model, an unidentified populated place (location class 2) would be captured within the corresponding ADM2 (location class 1) coded as approximate. An unidentified ADM2 would be captured within the ADM1 and so on.

7.2 Vague Area References

In some instances the source will refer to project activities in "most provinces" or a similarly vague reference to target locations. In these cases, there may be sufficient information to identify a general area, but not a specific subnational location. When a sufficient point of reference is available -- a regional boundary of some kind or a directional reference such as "northern provinces" -- coders may use those references as a guide for coding the area. In such instances, if the area does not conform to the current administrative system, it will be assigned a location class of 4 and specified as approximate. In the absence of a clear point of reference, coders will default to the next order administrative boundary as described above.

7.3 Unclear Locations

In some cases, it may be impossible to identify a subnational location within available documentation, or there may not be a subnational component -- the project may include a "country-wide" activity. In these cases, the project will be coded at the national level as an independent political entity (location class 1). If the country is coded due to a lack of information on subnational locations, it will be specified as approximate. In cases where the project includes "country-wide" activities, it will be specified as exact. In cases where the project involves technical support or capacity building within the central government[6], the country is specified as approximate. Even if there is no direct mention of any location in the source material, a project is still geo-referenced. It is at the discretion of the data user to assess the appropriate use of unclear locations in analyses.

Example: A World Bank local government support project in Tanzania requires local governments to apply for access to the program. With the documentation provided, it is impossible to determine which local governments actually received funding through the project. The solution is to code the country Tanzania (location class 1) as approximate.

7.4 Cases with Clear Locations and Unclear Locations

There may be some projects for which some locations are specified, and others are referred to more generally. In such instances, the coder will code all the clearly specified subnational locations and default to the next order administrative division to capture the ambiguous areas. This rule continues to apply in cases where locations within the same administrative boundary have already been coded. As described previously, it is at the discretion of the data user to determine whether vague area locations should be included in their analysis.

6 Note that that "central government ministries or financial institutions" only include those that are fully controlled by the government. Government "programs" or similar cooperative arrangements or organisations, which are composed of a wide variety of NGOs, a combination of central and local government agencies, and/or the private sector do not conform to the government technical assistance/capacity building model. The specification of geographic exactness will vary on the basis of the scope and nature of the project activities.

Example: A World Bank local government support project to be implemented in the city of Dar Es Salaam (location class 2) and communities in the Dar Es Salaam Region and other primary Tanzanian cities. In this case, the city of Dar Es Salaam also the "capital of a political entity" (location class 2) would be specified as exact. The region of Dar (location class 1) would also be coded as approximate to capture the "communities" described in the documentation. If the coder is unable to clearly identify a definitive list of "other primary Tanzanian cities," the country itself would be coded as approximate in order to capture these areas.[7]

7 It is important to weigh the need to capture the full area of the project against the relative value of a vague country location reference. In cases where the clear location(s) are considered to capture an overwhelming majority of project activities, then the country should not be coded. If the country was coded each time an unclear location was mentioned, such a rule would complicate the use of geocoded data beyond its usefulness. However, if it is entirely clear that the funding is disbursed to a set of unnamed populated places throughout the country, or a significant portion of it, then the country level is also coded.