Concept note: VIDA data mapping

Executive summary:

The VIDA project has complex data compatibility challenges. Data need to be compared from different sites at different times, but these data are diverse not only in terms of quality/cleanliness, but also in terms of format, language, and even digitalization status. In order to carry out meaningful comparative analysis on the project's data, three steps must be taken: (i) all data must be digitized; (ii) all data must be "mapped" so that equivalent variables (questions) and levels (categorical responses) in different datasets can be joined; (iii) all data must be "translated" (using the mapping utilities created in step 2) into one standard format.. This document lays out a basic plan, timeline and budget for achieving the above three steps.

Plan:

Step 1: Android OpenDataKit survey form digitization for Mali

Mali's VA data is in paper format and needs to be digitized by data clerks on site. In order to enter data, a program needs to be engineered that matches the format and language of the paper surveys. There are two separate forms which require separate programs:

1. For deaths of infants < 4 weeks of age (here)
2. For deaths of infants 4 weeks to 59 months of age (here)

The data entry programs for each of the above will be written by DataBrew, and delivered in .XML format to the VIDA team, whereafter Uma Onwuchekwa will oversee the deployment of the programs to android tablets, and the data entry process using the programs.

Deliverables:

- Two .XML format data entry programs in in OpenDataKit compatible language.

Step 2: Mapping of all meta-data in a software package

The task of mapping data consists of translating all questions (variables) and responses (levels) to one lexicon (that of the InterVA). Doing so requires a great deal of "manual" work: reading and understanding non-machine-readable data dictionaries (pdf and word documents which explain coded data), using subject area knowledge to "match" similar but not identical questions, and abstracting into computer code the rules for that matching. Additionally, given the diversity in sites' data management practices and personnel, a non-negligible amount of effort is foreseen in simply acquiring the correct and complete relevant datasets and accompanying meta-data (information, dictionaries, etc.). This step is likely to be the most time and labor-intensive of the collaboration.

Deliverables:

- A visual mapping of all datasets and their components for explanatory and reproducibility purposes.
- Algorithms (written in the form a software library in the R programming language) which take any site's data from any time and map all elements (variables and responses) to the standard lexicon.

Step 3: Translation of all data in a web application

Having fully abstracted the mapping of meta-data (Step 2), the software package will be used to “translate” all data elements. For the purposes of user ease and accessibility, the mapping tool will be in the form of a public web application. On the front-end of the web application, the user will specify an input format, input file, and output format; on the back-end, the software written in Step 2 will map the input files data elements to the standard lexicon, and then translate (if applicable) to the output format. The web application will be useful for translating specific datasets to a specific format (for miscellaneous analysis); more importantly, it will be useful for aggregating datasets by first translating to the same lexicon.

Deliverables:

- A web application (hosted by DataBrew, but with source code publicly available).

Timeline:

February	March	April	May	June	July	August
Step 1						
	Step 2					
				Step 3		
					Iterative feedback/ tweaks	

Step 1: Android OpenDataKit survey form digitization for Mali

- February 12-16. Deliverable on Friday, February 16th at COB.

Step 2: Mapping of all meta-data in a software package

- February 19-May 31. Deliverables on Thursday, May 31st at COB.

Step 3: Translation of all data in a web application

- June 1-August 30. Initial prototype app delivered by Friday, June 29th. Two month iterative improvement phase thereafter. Project completion by Friday, August 31st.

Budget:

Step	Price	Details
Android OpenDataKit survey form digitization for Mali	$800	2 forms x estimated 20 hours each x 20 USD hourly
Mapping of all meta-data in a software package	$12,000	Estimated 3 months half-time software developer / data scientist
Translation of all data in a web application	$8,000	Estimated 1 month half-time web developer / data scientist, followed by 2 months quarter-time.
Total	**$20,800**	Market price

Comment on pricing and timing:

DataBrew generally charges market price for data science consulting services. However, inspired by our roots in the Univ. of Chicago's "Data Science for Social Good", we strive to discount our fees for organizations which are involved in delivering a "social good". In other words, some underfunded "social good" collaborations are effectively subsidized by fully funded (or non social good) projects.

The discount is a function of the client's ability to pay, the type of work being undertaken, the landscape of alternative partners available to the client, and the capacity of DataBrew at the time of the project. This discounting is made possible by organizations paying market price when able, and by both DataBrew and clients being transparent and clear about capacity and expectations when full funding is not available.

The VIDA project meets every definition of "social good." If budget-constrained, we should discuss a discount. When DataBrew embarks on discounted projects, we carry out the work with all the rigor, quality and attention to detail of full budget projects; however, since discounting can affect team capacity (in the sense that it limits our ability to outsource tasks through hiring) and priorities (when capacity-constrained, we're forced to prioritize obligations to paying clients), it can affect timelines. To the extent that VIDA wishes to negotiate a discounted rate, we should consider revising timelines, particularly for Steps 2 and 3.