Document link: [Save the Children - Humanitarian Data Insights - DataDive brief](https://docs.google.com/document/d/1g7Kkfnxy_XKpF1Hn7fmCbvA6jqCQhsvM5B2iEJ1pfxc/edit?usp=share_link)

Github repo: <https://github.com/datakind/STC-DataDive-Nairobi-June-2023>

Google Drive: [Humanitarian Insights - Save the Children](https://drive.google.com/drive/folders/1SwcWN_lLJcErR8c8ndELwsWcOZdLZUGc?usp=share_link)

Main Slack General Channel: [General Channel](https://nairobidatadiveevent.slack.com/archives/C059JJZLWBF)

Slack Channel for the Save the Children Stream: [Humanitarian channel](https://nairobidatadiveevent.slack.com/archives/C059JP53153)

Contents

[**What Does DataDive Event Success Look Like? 2**](#_hd4ryiyfifbc)

[Examples of success 2](#_2pnz2mihh35a)

[**DataDive Project Details 2**](#_lr6tqpj9hm9y)

[Introduction 3](#_7km18xsf75k4)

[Problem 4](#_i4pobmguted7)

[**Data Sources 4**](#_2kotp7s8tyih)

[The Humanitarian Data Exchange 4](#_pb2auu1rx6e2)

[**Workstream #1: Data visualization 6**](#_2rywydlogw01)

[**Workstream #2: Natural Language Querying of Humanitarian Datasets 7**](#_1aqrkg46n9z9)

[**DataDive volunteer resources 8**](#_s3slpwjeffd5)

# 

# What Does DataDive Event Success Look Like?

*Success is about incremental learning and improvement*

At DataDives, we seek to increase understanding of the issue area we are partnered to work within - for our partners, for ourselves, and for the sector. We also seek to turn these new understandings into additional insights so that our partners can better shape their own actions. We provide catalyzing capacity, energy, and expertise to specific problems so that partners can take away knowledge that will help them take the next best step on their data science and data engineering journey.

| Examples of success To increase understanding of this issue area, here are a few examples of what constitutes “success” for this project.   1. Conducting an exhaustive EDA on a shared dataset and being able to summarize what is - or is not - contained within that dataset 2. Demonstrating the application of visualization and language modeling towards understanding a data ecosystem 3. Creating prototype / proof of concept tools & visualizations that demonstrate the potential utility of new ways of visualizing and understanding data 4. Creating recommendations for how partner organizations might invest their resources as they continue on a data science journey 5. Showcasing the “Art of the Possible” by creating specific case studies or examples of how data science techniques can lead to better/faster/more insightful/less resource intensive outcomes 6. Designing the workflows that would allow for repeatable data extraction from multiple sources, combination, and visualization |
| --- |

## 

# DataDive Project Details

| *“The traditional approach to dealing with humanitarian crises has been to watch disaster and tragedy – whether from famines, wars, storms or disease – build, and then gradually decide, normally driven by public and political reaction to media coverage, that we need to respond, then to mobilize money and organizations to help, and then after that to start to get help to the people who need it.*  *That is a reactive approach. It saves many lives. But it is slower, and hence less humane and more expensive, than it needs to be.*  *It would be better to take an anticipatory approach, where we plan in advance for the next crises.”*  Mark Lowcock, Under-Secretary-General for Humanitarian Affairs and Emergency Relief Coordinator - Anticipation saves lives: How data and innovative financing can help improve the world’s response to humanitarian crises ([UN OCHA, 12/2019](https://reliefweb.int/report/world/mark-lowcock-under-secretary-general-humanitarian-affairs-and-emergency-relief)) |
| --- |

| *“Our (Save the Children) primary pain points are:*   * *There is so much data available, difficult to make sense of it all at once* * *There is no easy way to slide and dice on an organization by organizational bases* * *Current map views are way too broad to be useful* * *Metrics on their own often don’t make sense (for example, areas will appear as highly problematic on current map views but there are also hardly any people living in those areas)* * *We need to be able to view trends per capita them to be practical and useful*   *”*  Feedback from Save the Children provided during analysis for the “Data for Humanitarian Action” project with DataKind DC |
| --- |

## Introduction

[Save the Children](https://www.savethechildren.net/) represents a significant amount of the humanitarian response across the globe. Through partnership with DataKind since March 2021, Save the Children has focused on developing tools that support data-forward activities (e.g., digital data collection, analytics, decision-making). As they have identified and collected data sources, the next phase in this data tool building process of creating demonstration analytics that will help in educating other leading humanitarian organizations embrace - and demonstrate - frameworks and systems that showcase the value of data.

## Problem

*In order for humanitarian organizations to respond rapidly to humanitarian disasters, they need to understand (i) what data exists across an ecosystem which includes tens of thousands of non-standardized datasets; (ii) how this data can be combined to create deeper insights; and (iii) whether it can be trusted.*

Collecting, analyzing, and visualizing external humanitarian data helps Save the Children organizations to plan humanitarian response activities and make decisions regarding where best to invest resources so that they can maximize their impact. However, Save the Children has recognized that while this is a resource they have invested in, the underlying environment does not promote a systematic and effective use of data collectively, so many of Save the Children’s peer organizations do not have the same access to information.

As a result the entire humanitarian relief sector often duplicates resources, serving the same people or communities rather than focusing on complementary services or geographical areas so as to optimize individual resources by collaborating to make the greatest impact possible. Specifically, Save the Children wants to understand if and how public data can be synthesized at a subnational level such that organizations can best direct resources for food security (economic and health) intervention in a way that might minimize duplicated efforts.

The aim of the work in this DataDive stream is to explore how a non-technical user at a humanitarian organization can gain a quick understanding of the humanitarian data available to them, how it might be combined and its quality and format.

# Data Sources

## The Humanitarian Data Exchange

The data used in this DataDive was extracted from the amazing [Humanitarian Data Exchange](https://data.humdata.org/) (HDX), a platform to “Find, share and use humanitarian data all in one place”. There are many other additional data sources, but to keep things simple we’ll use data from this single platform. The challenges we are tackling would apply to any other sources of humanitarian data.

In order to get you up-and-running faster, we have already extracted metadata from HDX using [this notebook](https://github.com/datakind/gpt-3-meta-data-discovery/blob/main/hdx_gpt-3_tag_prediction.ipynb). You shouldn’t need to run the notebook, but it might be a good reference to see how the data was produced.

You will come across these two terms on HDX and in the provided sample data:

* ‘Dataset’ - is a collection of files corresponding to one group of data, having its own page on HDX, for example see [here](https://data.humdata.org/dataset/faostat-prices-for-kenya)
* ‘Resource’ - Each dataset can have multiple resources, each one being a downloadable file visible on the Dataset page on HDX

Here is the data:

| Data Location | Description |
| --- | --- |
| Meta Data Sheet for all datasets with HXL tags at the dataset level  [Excel file](https://docs.google.com/spreadsheets/d/1WAibTpbjGBN5UtVaR0br5N2fygY0T1X_/edit?usp=share_link&ouid=106808949113099347741&rtpof=true&sd=true)  [Pickle file](https://drive.google.com/file/d/15KckfAZNrIYj_omGhOWtLhyDAP0VP_4x/view?usp=share_link)6 | A dataframe was extracted from [Humanitarian Data Exchange](https://data.humdata.org/) (HDX) of datasets and resources related to **Kenya** which included tabular data, and for each resource the file downloaded and the top few rows extracted as well as metrics about number of rows and columns.  The table excerpts can sometimes be misleading due to the varying format of Excel sheets, where sometimes merged and blank rows appear at the top of the file. This shouldn’t affect analysis too much for the data dive, but if interested you can find more information on this [here](https://medium.com/towards-data-science/parsing-irregular-spreadsheet-tables-in-humanitarian-datasets-with-some-help-from-gpt-3-57efb3d80d45).  Not *all* Kenyan metadata for tabular datasets on HDX are included, some were excluded due to issues with downloading data, but the majority were processed.  There is a row for each table of data in a resource (file), so for Excel sheets there can be multiple rows, one for each tab.  The columns in this file are as follows:   * **resource\_id** - HDX resource ID * **resource\_name** - HDX resource name * **resource\_format** - HDX resource format XLSX, CSV * **dataset\_id** - HDX Dataset ID * **dataset\_name** - HDX Dataset ID * **dataset\_org\_title** - Organization that uploaded the data * **dataset\_last\_modified** - Last modification date of dataset * **dataset\_tags** - Tags of the dataset, e.g. ‘population’ * **dataset\_groups** - Groups. For the most part this includes the countries the dataset applies to * **dataset\_total\_res\_downloads** - Total number of downloads for resources in dataset * **dataset\_pageviews\_last\_14\_days** - Total number of page views for dataset in last 14 days * **file** - File name when downloaded (not needed for this DataDive) * **type** - File type. Same as resource\_format above * **dataset** - Full JSON record of HDX dataset, might include a few extra fields not listed above * **sheet** - If an Excel file has multiple tabs (sheets), each one is identified * **resource** - Full JSON record of HDX resource, might include a few extra fields not listed above * **num\_rows** - Number of rows in table (can also include blank/merged rows as unprocessed) * **num\_cols** - Number of columns in table (can also include blank/merged rows as unprocessed) * **columns** - Column names of table * **first\_row** - First row * **has\_hxl\_header** - Flag to indicate if table has [HXL tags](https://hxlstandard.org/) * **first\_nrows** - An excerpt of the first 15 rows of data, including header   As noted above in field ‘has\_hxl\_header’, columns can have [Humanitarian Exchange Language](https://hxlstandard.org/) (HXL) meta data tags. These might be useful if visualizing or querying the meta data down to the column level.  To read in this data, a [sample notebook](https://github.com/datakind/STC-DataDive-Nairobi-June-2023/blob/main/analysis/getting-started.ipynb) has been provided in the [Github repo](https://github.com/datakind/STC-DataDive-Nairobi-June-2023/tree/main) for this workstream. |

Extra Resources

* The tasks for visualization of meta data outlined below can be inspired by some previous work done, for example in building the [HDX Universe](https://medium.com/@Simon_B_Johnson/the-shape-of-the-humanitarian-data-exchange-68746393a0a) (article) and [visualization](https://brcmapsteam.github.io/hdx_universe/). A great infrastructure for graph visualizations is [neo4j](https://neo4j.com/) and [GraphXR](https://www.kineviz.com/graphxr)
* For natural language querying of data, there has been some amazing work on converting natural language to SQL using Large Language Models (LLMs) and [LangChain](https://python.langchain.com/en/latest/index.html), see for example [this agent](https://python.langchain.com/en/latest/modules/agents/toolkits/examples/pandas.html) for a Pandas dataframe, [this agent](https://python.langchain.com/en/latest/modules/agents/toolkits/examples/vectorstore.html) for a document agent (though we would want something with persistent storage). Neither are perfect for our use case, but might be good starting points.

# Workstream #1: Data visualization

The intent of work in this workstream is to explore how we can provide visualizations which help organizations understand what humanitarian data is available, how recent/active it is, and if possible, how it links together. Success would be visualizations that allow the user to intuitively filter and navigate to find out key information about humanitarian datasets so they can incorporate them in their humanitarian response.

You could start by exploring the data in an Google sheet, Python notebook, or use a BI platform such as [Power BI](https://powerbi.microsoft.com/en-us/downloads/). Power BI desktop is free and once ready the DataKind team can publish your online report.

For displaying network graphs, as mentioned above, neo4j and Graph XR might be good options.

Explore how you might display the full hierarchy: Dataset > Resource > [ Sheet tabs if Excel ] > Columns. Columns would need to be parsed out of the ‘columns’ column in the data.

Key to all visualizations is to try and pretend you are a non-technical user from a humanitarian organization. Try and consider design towards making any visualization as simple as possible to use! Oh ….. And make it super pretty! An engaging visualization often situates the user, design matters.

# Workstream #2: Natural Language Querying of Humanitarian Datasets

Visualizations are a great way to help people find the data they need, but what if they could simply ask a question?

“What are the most active datasets for food security?”at

“How many population datasets have data down to the sub-county level?”

“What datasets might link together with the. It’s active food security dataset?”

“Describe the data for dataset X”

This workstream explores just that!

With the advent of Large Language Models such as ChatGPT natural language understanding has accelerated, and it is now possible to convert natural language to database queries such as SQL. This means a question in any language can be used to return data.

In our context the data is the meta data for humanitarian datasets on HDX, plus information about column heading and data excerpts.

* It would be good to first start interrogating the data with SQL to get and understanding of the types of queries that return useful information.
* Then it might be good to apply the technique mentioned in the ‘Additional Resources’ above to use LangChain with a large language model such as GPT-3 or Bloom. Note: The DataKind team can provide an API key for GPT 3.5 turbo
* It would be absolutely amazing if the technique supported Swahili and English!
* If you have time, a simple interface would be fantastic, where a user could enter their question

It’s worth noting that the task above queries meta data and excerpts, but the technique could be extended to include the actual data itself.

Good luck!

# DataDive volunteer resources

1. Document link: [Save the Children - Humanitarian Data Insights - DataDive brief](https://docs.google.com/document/d/1g7Kkfnxy_XKpF1Hn7fmCbvA6jqCQhsvM5B2iEJ1pfxc/edit?usp=share_link)
2. Github repo: <https://github.com/datakind/STC-DataDive-Nairobi-June-2023>
3. Google Drive: [Humanitarian Insights - Save the Children](https://drive.google.com/drive/folders/1SwcWN_lLJcErR8c8ndELwsWcOZdLZUGc?usp=share_link)
4. Main Slack General Channel: [General Channel](https://nairobidatadiveevent.slack.com/archives/C059JJZLWBF)
5. Slack Channel for the Save the Children Stream: [Humanitarian channel](https://nairobidatadiveevent.slack.com/archives/C059JP53153)
6. Additional resources for small-group collaboration
   1. [Slack calls](https://slack.com/help/articles/216771908-Make-calls-in-Slack) - you can call people 1-to-1 directly in slack
   2. [Google meet](https://support.google.com/meet/answer/9302870?co=GENIE.Platform%3DDesktop&hl=en) - you can schedule a meeting for now, or for later in Google meet if you have a google email account