OCHA centre for humdata HDX

THE STATE OF OPEN HUMANITARIAN DATA 2021:

ASSESSING DATA AVAILABILITY ACROSS HUMANITARIAN CRISES

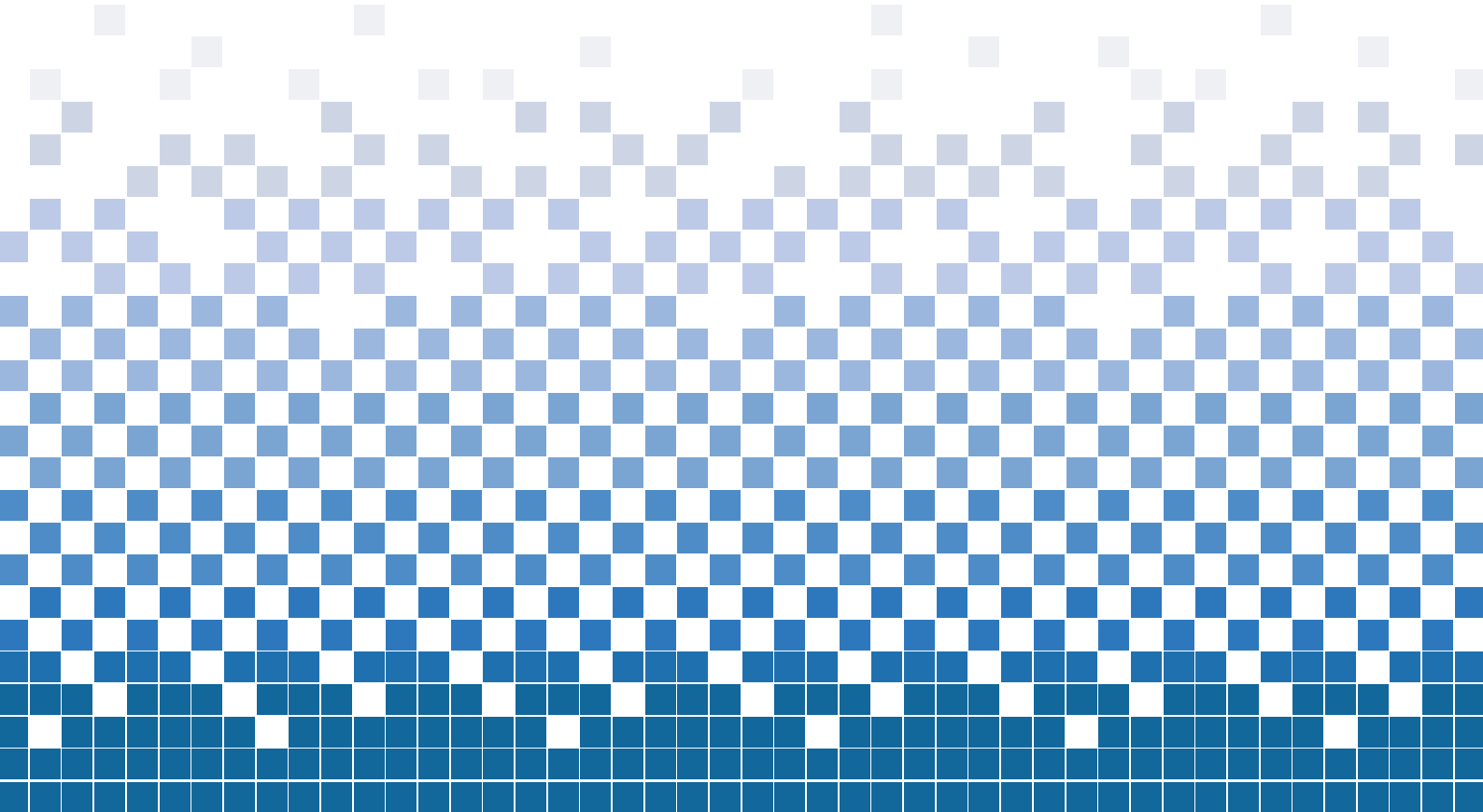


TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This report was produced in January 2021 by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) Centre for Humanitarian Data in The Hague, which manages the Humanitarian Data Exchange platform (HDX). OCHA thanks all of the organizations that have shared data through HDX, the donors who have supported this work over the years, and the HDX users who are committed to ensuring humanitarian response is data-driven. For additional information, contact the Centre for Humanitarian Data at centrehumdata@un.org.

LIST OF ABBREVIATIONS

ACLED	Armed Conflict Location & Event Data Project
DESA	United Nations Department of Economic and Social Affairs
GAM	Global Acute Malnutrition
FAO	Food and Agriculture Organization of the United Nations
HDX	Humanitarian Data Exchange
HOT	Humanitarian OpenStreetMap Team
HRP	Humanitarian Response Plan
IDP	Internally Displaced Person
IHME	Institute for Health Metrics and Evaluation
IOM	International Organization for Migration
IPC	Integrated Food Security Phase Classification
JRC	Joint Research Centre of the European Commission
NGO	Non-Governmental Organization
OCHA	United Nations Office for the Coordination of Humanitarian Affairs
SAM	Severe Acute Malnutrition
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNHCR	United Nations High Commissioner for Refugees
UNFPA	United Nations Population Fund
UNICEF	United Nations Children's Fund
WFP	World Food Programme
WHO	World Health Organization

1. INTRODUCTION

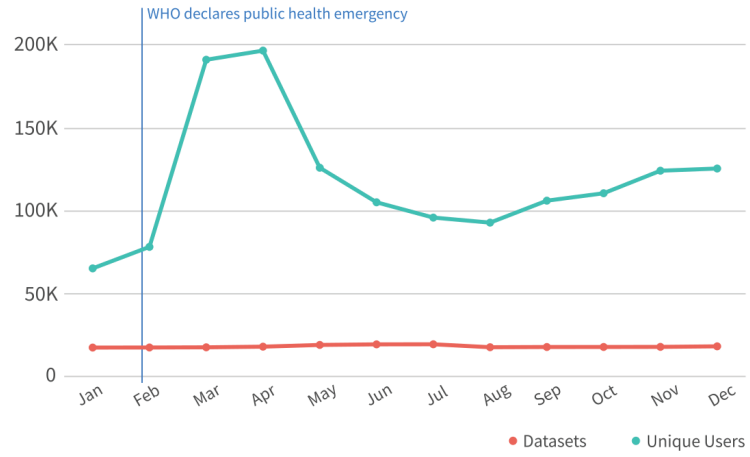
The goal of this report is to increase awareness of the data that is available to inform humanitarian operations around the world and to highlight what is missing, as measured through OCHA's Humanitarian Data Exchange (HDX) platform.[1] In a year dominated by the COVID-19 pandemic and its devastating impact on already-vulnerable populations, we saw record-breaking demand for data in the humanitarian sector coupled with persistent data gaps.

When we talk about data in humanitarian settings, we are talking about the world's most vulnerable people: 235 million people will need humanitarian assistance and protection in 2021, a near 40 percent increase from 2020.[2] The data tells us that hunger is on the rise, internal displacement is at its highest level in decades, severe weather events are more common, and disease outbreaks are increasing. All the while, the gap between humanitarian needs and the financing available to address those needs keeps growing.

> "Starkly and powerfully, the COVID-19 pandemic illustrates how critical data use, with a human face, is to protecting lives and livelihoods. The crisis is a wake-up call."[3]
>
> - United Nations Secretary-General António Guterres

In the second year of producing *The State of Open Humanitarian Data*, it is evident that relevant, complete, and timely data is essential for understanding current humanitarian crises and anticipating what comes next. We can see that the humanitarian data ecosystem is getting stronger as more organizations shift from manual to automated ways of sharing data and as awareness around the importance of responsibly collecting and sharing data increases.

As the go-to place for humanitarian data, HDX is a useful proxy for measuring data activity and availability on a global scale. When HDX was launched in 2014, it held around 800 datasets. Over the past seven years, that number has grown to over 18,200 datasets — an increase of 2,175 percent. HDX saw record growth in 2020 with over 1.3 million people using the platform (compared to 600,000 people in 2019) and over 2.2 million datasets downloaded (compared to 930,000 in 2019). The most popular datasets were related to the COVID-19 pandemic.



Monthly Unique Users and Total Datasets on HDX in 2020

[1] HDX is an open platform for finding and sharing humanitarian data across crises and organizations. The goal is to make humanitarian data easier to find and use for analysis. Learn more at **data.humdata.org**.

[2] Taken from the OCHA Global Humanitarian Overview for 2021, available online at **gho.unocha.org**.

[3] Taken from the Data Strategy of the UN Secretary-General for Action by Everyone, Everywhere (2020-2022), available online at **un.org/datastrategy**.

We created the Data Grids in 2019 as a way to help users in their quest for relevant, complete and timely data.[4] The Data Grids bring together a limited set of foundational datasets needed to understand a humanitarian context and provide a comparable way to assess data availability across locations. They place the most important crisis data into six categories and 27 sub-categories. The categories are affected people; coordination and context; food security and nutrition; geography and infrastructure; health and education; and population and socio-economy. (See Annex A for all sub-category definitions).

Data is included in the Data Grid if it is relevant to the thematic area and sub-national. Once that threshold is passed, the data is considered 'complete' if it has broad geographic coverage, is shared in a commonly-used format, and is up-to-date; if any of those criteria are not met it is considered 'incomplete'. (See Annex B for details on the Data Grid creation process and criteria).

DATA GRID CRITERIA

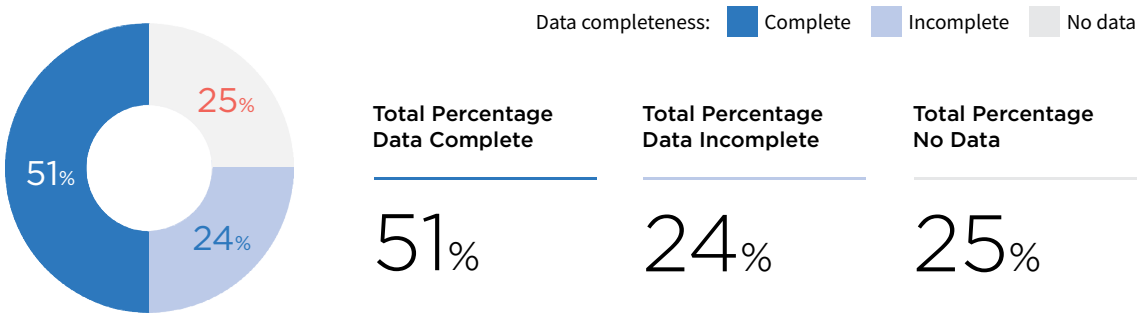
Step 1: Is it relevant and sub-national?

Step 2: Does it have broad geographic coverage? Is it in a commonly-used format? Is it timely?

Complete	Incomplete	No Data
At least one dataset in the sub-category meets all criteria.	At least one dataset in the sub-category meets some criteria.	Available data in the sub-category does not meet the criteria or does not exist on HDX.

For 2020, HDX maintained Data Grids for 27 locations. This reflects the expansion of the Data Grids from previous years to cover all locations with a Humanitarian Response Plan.[5] Added in 2020 were: Burkina Faso, Burundi, Cameroon, Ethiopia, Haiti, Iraq, Libya, Mali, Niger, Nigeria, Pakistan, South Sudan, Syrian Arab Republic, Ukraine and Zimbabwe. Bangladesh and the Philippines no longer have active Humanitarian Response Plans and therefore do not have Data Grids on HDX.

At the start of 2021, we estimate that just 51 percent of relevant, complete crisis data is available across 27 humanitarian operations. If we add the data that is relevant but incomplete, the total is 75 percent. This leaves 25 percent of categories with data that does not meet the criteria or with no data. The Data Grids include an average of 24 datasets per location.



With almost double the locations (from 14 to 27) covered by Data Grids compared to last year's report, the percentage of data completeness has stayed the same (from 54 percent in 2019 to 51 percent in 2020). The sub-categories with no data have increased slightly from 22 percent to 25 percent. But if we compare 2020 to early 2019, when we first created the Data Grids, we have made great progress. Back then, over 50 percent of the sub-categories had no data.

[4] View all of the HDX Data Grids online at **bit.ly/datagrids**.

[5] Humanitarian Response Plans (HRPs) are prepared by UN Humanitarian Country Teams in locations where there is an ongoing humanitarian emergency. HRPs are generally prepared annually, and outline an overall strategy and specific activities for each humanitarian cluster.

This report contains details on data availability in each of the locations, categories and sub-categories covered by the Data Grids as of January 2021. We provide a country deep dive for Mali, which shares the highest degree of data completeness with Chad. And we highlight a partnership with the Integrated Food Security Phase Classification (IPC) that helped to increase the availability of food security data across several countries.

We also show how data available on HDX and from new sources supported the development of a model that forecasts the number of COVID-19 cases, hospitalizations, and deaths in six countries with humanitarian operations. With the scale of the COVID-19 pandemic, along with the unsuitability of much COVID-19 data for the Data Grids (particularly the lack of sub-national data), we chose to create a dedicated page on HDX for all data related to the pandemic.[6]

We recognize the valuable contributions of all the data-sharing organizations publishing data on HDX, and welcome the 38 organizations sharing data on the platform for the first time in 2020. The new data they contributed helps to fill critical gaps in the data coverage, and ensures HDX continues to be a key resource for the humanitarian community and beyond. Trusted partnerships and focused advocacy efforts have led to many new, relevant datasets on HDX this year, including:

- A baseline assessment of the number of IDPs, returnees and host communities at the district level in Somalia from IOM.[7]
- The location of health facilities by district for Yemen contributed by WHO on behalf of the Ministry of Health.[8]
- Updated population statistics at the district level for Ethiopia contributed by OCHA Ethiopia on behalf of the Central Statistics Agency.[9]
- Global and Severe Acute Malnutrition rates by health zone for the Democratic Republic of the Congo from UNICEF.[10]
- Multidimensional poverty data at the sub-national level for multiple countries from the Oxford Poverty & Human Development Initiative.[11]

The HDX team will continue to update the Data Grids throughout 2021 in partnership with the dozens of organizations contributing and using data on HDX. Improving data availability takes time, focus, and resources, both for those collecting and sharing data as well as those tasked with turning data into actionable insights. We are grateful for and encourage the funding of humanitarian data initiatives across the sector, and wish to specifically acknowledge the support of the Government of the Netherlands and The Rockefeller Foundation for funding this report and HDX more broadly. We look forward to continued collaboration and to closing data gaps.

2. KEY MESSAGES

- The COVID-19 pandemic created unprecedented demand for data in the humanitarian sector, but persistent data gaps remain. With every country in the world affected by COVID-19, the disparity in data availability in countries experiencing humanitarian crises became more clear.
- The COVID-19 pandemic also brought into stark focus the value of predictive models to inform humanitarian response strategies. Significant challenges exist in relation to data gaps and data quality, limiting the viability and accuracy of model development. Model output is only as good as model input.

[6] View this dataset on HDX at **bit.ly/covid-hdx**.
[7] View this dataset on HDX at **bit.ly/somalia-disp**.
[8] View this dataset on HDX at **bit.ly/yem-herams**.
[9] View this dataset on HDX at **bit.ly/eth-csa**.
[10] View this dataset on HDX at **bit.ly/RDC-mal**.
[11] View a sample from this dataset on HDX at **bit.ly/afg-mpi**.

- HDX saw record growth in 2020 with over 1.3 million people using the platform (compared to 600,000 people in 2019) and over 2.2 million datasets downloaded (compared to 930,000 in 2019). The most popular datasets were related to the COVID-19 pandemic.

- The Data Grids bring together a limited, carefully-curated set of foundational datasets needed to understand a humanitarian context and provide a comparable way to assess data availability across locations.

- As measured through the Data Grids, just over 50 percent of relevant, complete crisis data is available across 27 humanitarian operations. This global overview figure disguises a wide variation in data that is available or missing in each crisis. Across these locations, 25 percent of all data is missing. Another 25 percent is considered incomplete.

- Chad and Mali share the highest degree of data completeness at 70 percent. Ukraine and Zimbabwe share the lowest at 26 percent each. Locations with lower completeness tend to be those where humanitarian presence is reduced or otherwise restricted, and opportunities to establish trusted relationships around data sharing are limited.

- Population and socio-economy is the most complete data category. Health and education is the least complete category, driven by mostly incomplete data for the location of health facilities and the location of education facilities. Three sub-categories are 100 percent complete across all locations: casualties, funding, and food prices, owing to the work of ACLED, OCHA, and WFP respectively.

- Of the 289 organizations currently sharing data on HDX, 32 contribute data that is included in the Data Grids. This is a 14 percent increase from the previous year and reflects the addition of eight new organizations to the Data Grids, two of which were new to HDX.

- Data completeness is often context-specific. For instance, data on affected schools is mostly missing due to the sensitivity of this data, especially in conflict settings. It is therefore difficult to collect and share. On the other hand, data on damaged and destroyed buildings is less likely to be collected or kept up-to-date in non-conflict locations.

We call on partners to share or help source the following data that is critical but often missing or incomplete for many crises:

- Global Acute Malnutrition (GAM) and Severe Acute Malnutrition (SAM) rates (potential sources: UNICEF, World Bank, WHO).

- The location of education facilities (potential sources: national governments, UNICEF, UNESCO, the Global Education Cluster).

- The location of health facilities (potential sources: national governments, Global Health Cluster, WHO).

- The agreed list of populated places with locations (potential sources: national governments, UNFPA).

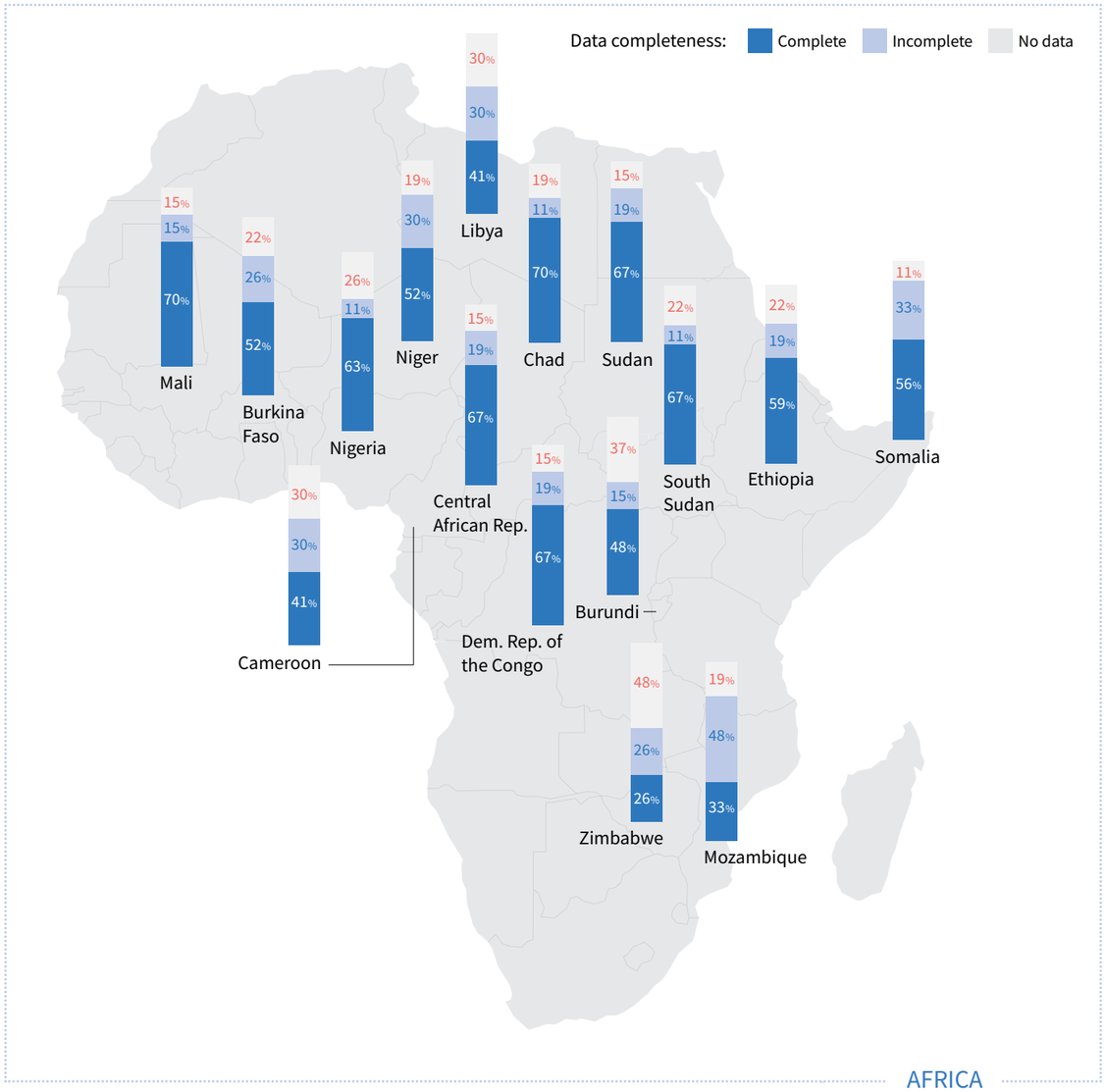
- The agreed geographic dataset of sub-national administrative divisions (source: national governments).

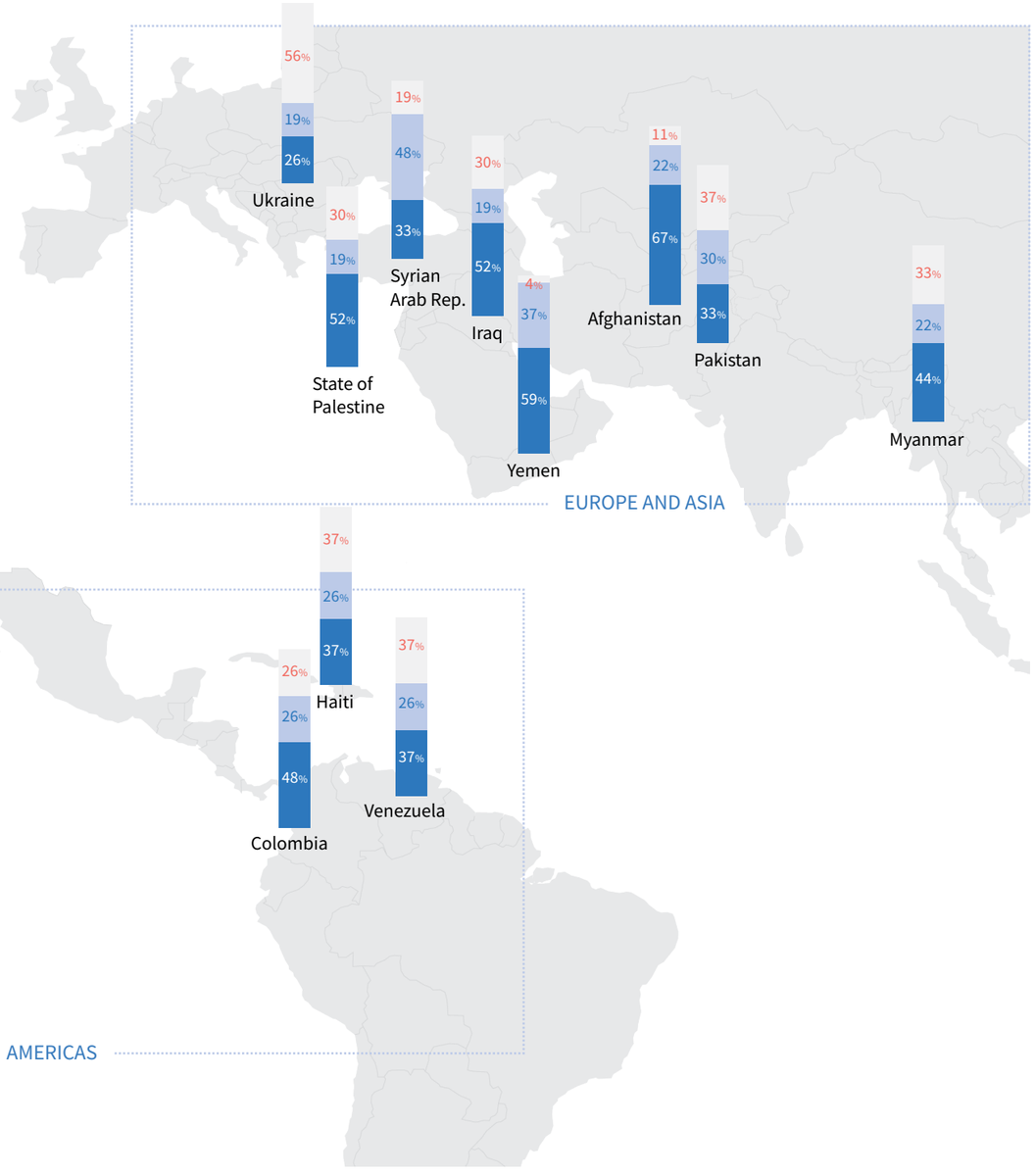
- Local transportation routes with an indication of status (potential sources: national governments, WFP, the Global Logistics Cluster).

Where this data is not available from authoritative sources, we call on partners to develop proxy indicators using innovative methods for data collection and analysis. An example of this could be using anonymized social media data or call detail records to understand movement patterns in a location. Where these efforts are underway, we ask that the data is made available responsibly through HDX or as metadata only through HDX Connect.[12]

[12] HDX Connect allows for sharing of metadata only, with the underlying data only made available upon request. HDX Connect datasets still contribute to the completeness of a Data Grid. Learn more: **bit.ly/hdx-connect**.

3. GLOBAL OVERVIEW





The boundaries and names shown and the designations used on this map do not imply official endorsement or acceptance by the United Nations.
Percentages may not total 100 due to rounding.

Number of Locations

27

Number of Categories

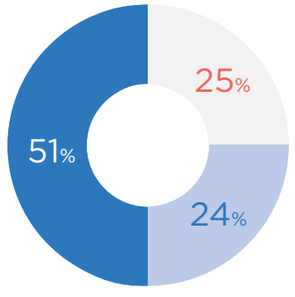
6

Number of Sub-Categories

27

Number of Contributing Organizations

32



Total Percentage Data Complete

51%

Total Percentage Data Incomplete

24%

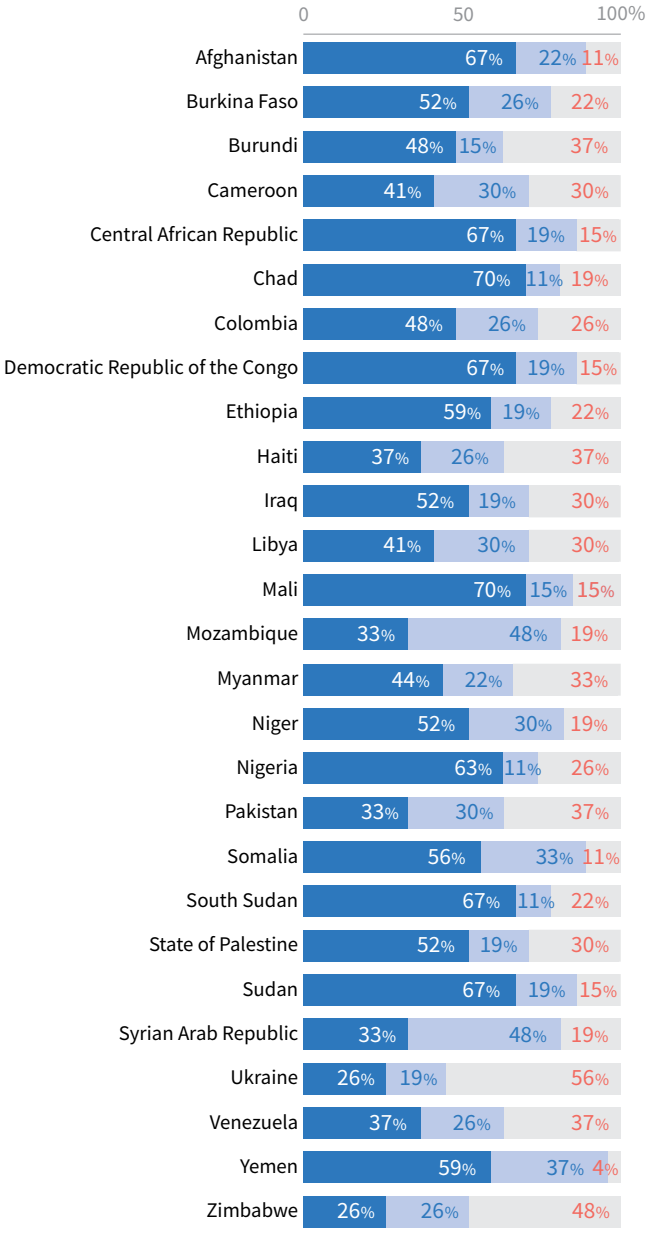
Total Percentage No Data

25%

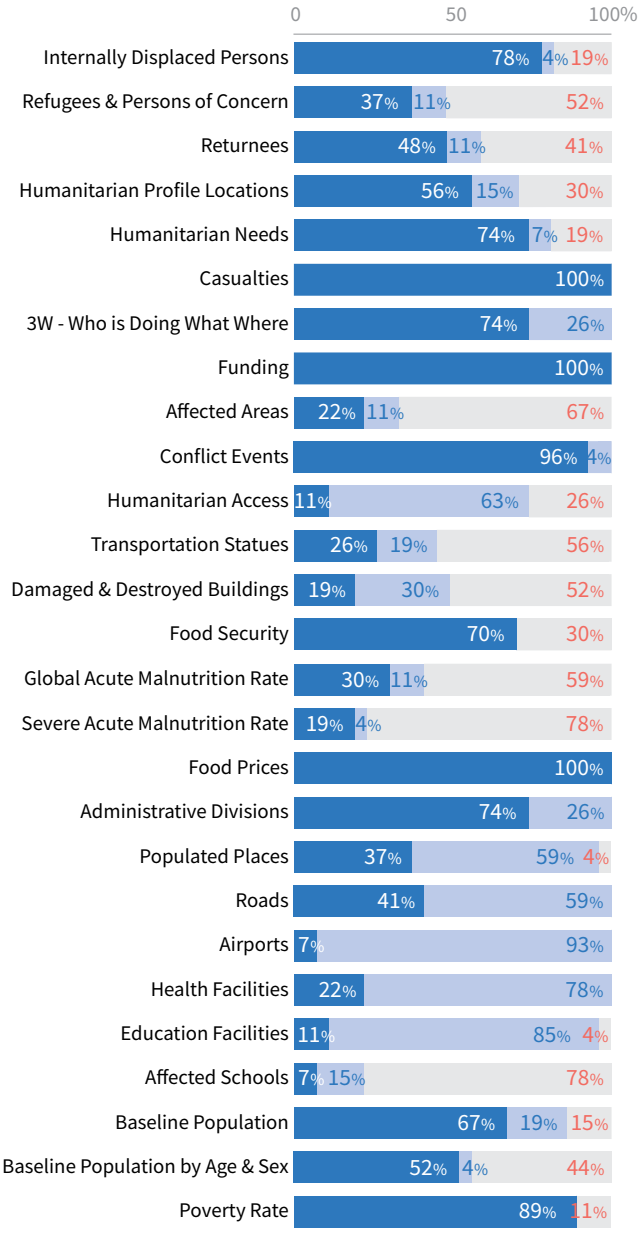
4. COMPLETENESS BY LOCATION, CATEGORY, AND SUB-CATEGORY

The global overview figure of 51 percent complete across 27 locations disguises a wide variation in data that is available or missing in each location, category, and sub-category. Across these locations, 25 percent of all data is missing. Chad and Mali share the highest degree of data completeness at 70 percent. Ukraine and Zimbabwe share the lowest degree of completeness at 26 percent, as well as the highest number of sub-categories with no data (56 percent and 48 percent respectively). This reflects the difficulty of sourcing relevant and timely data in those locations.

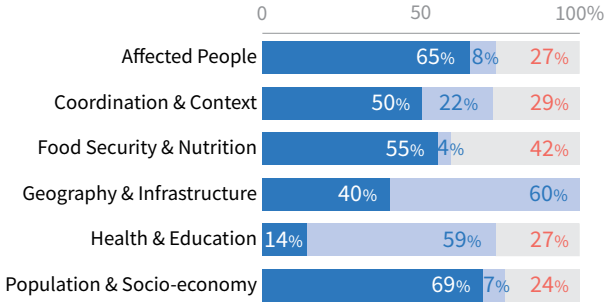
BY LOCATION



BY SUB-CATEGORY



BY CATEGORY



Data completeness: Complete | Incomplete | No data

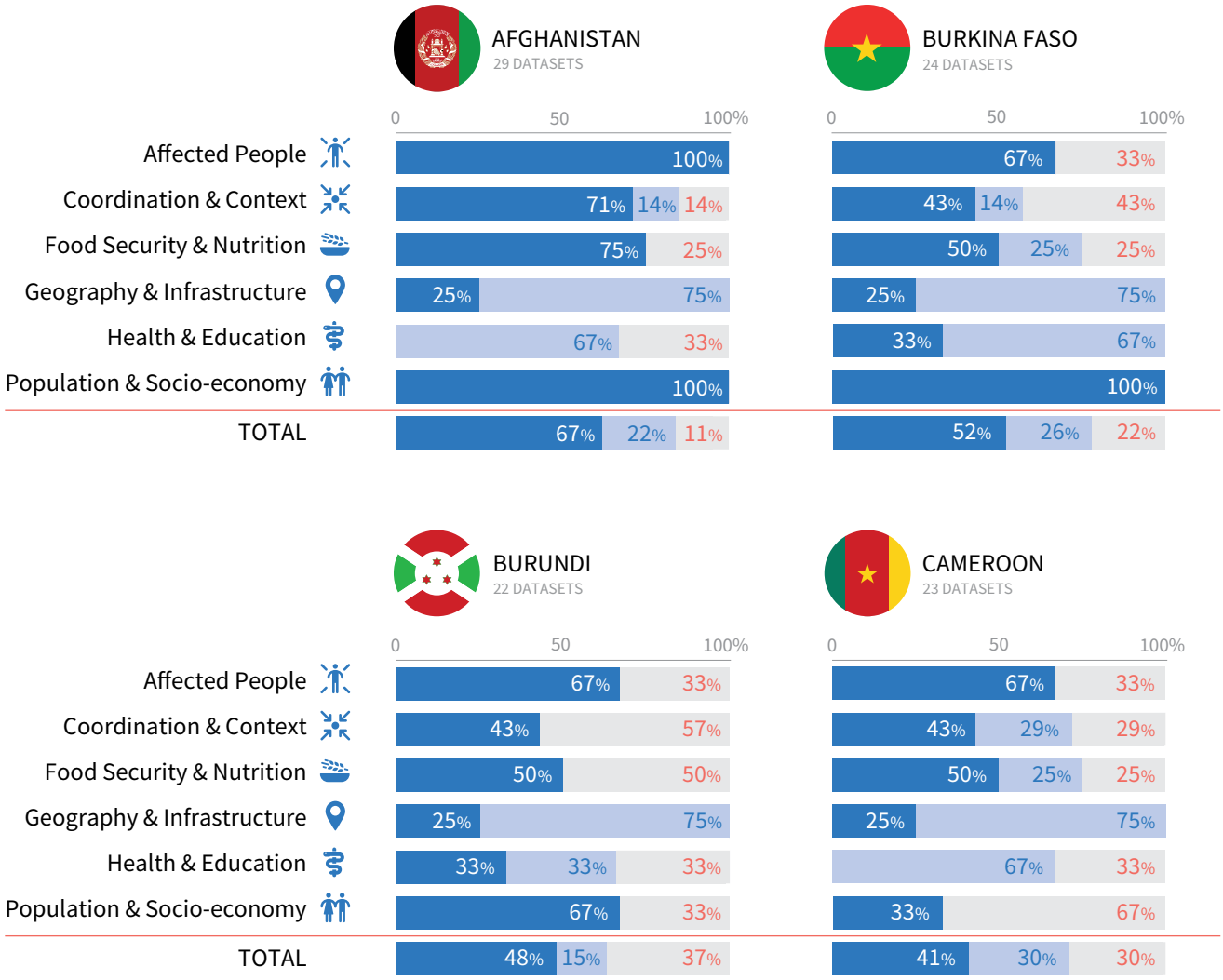
5. COMPLETENESS BY LOCATION AND CATEGORY

After Chad and Mali, the next most complete locations for data are Afghanistan, the Central African Republic, the Democratic Republic of the Congo, South Sudan and Sudan.

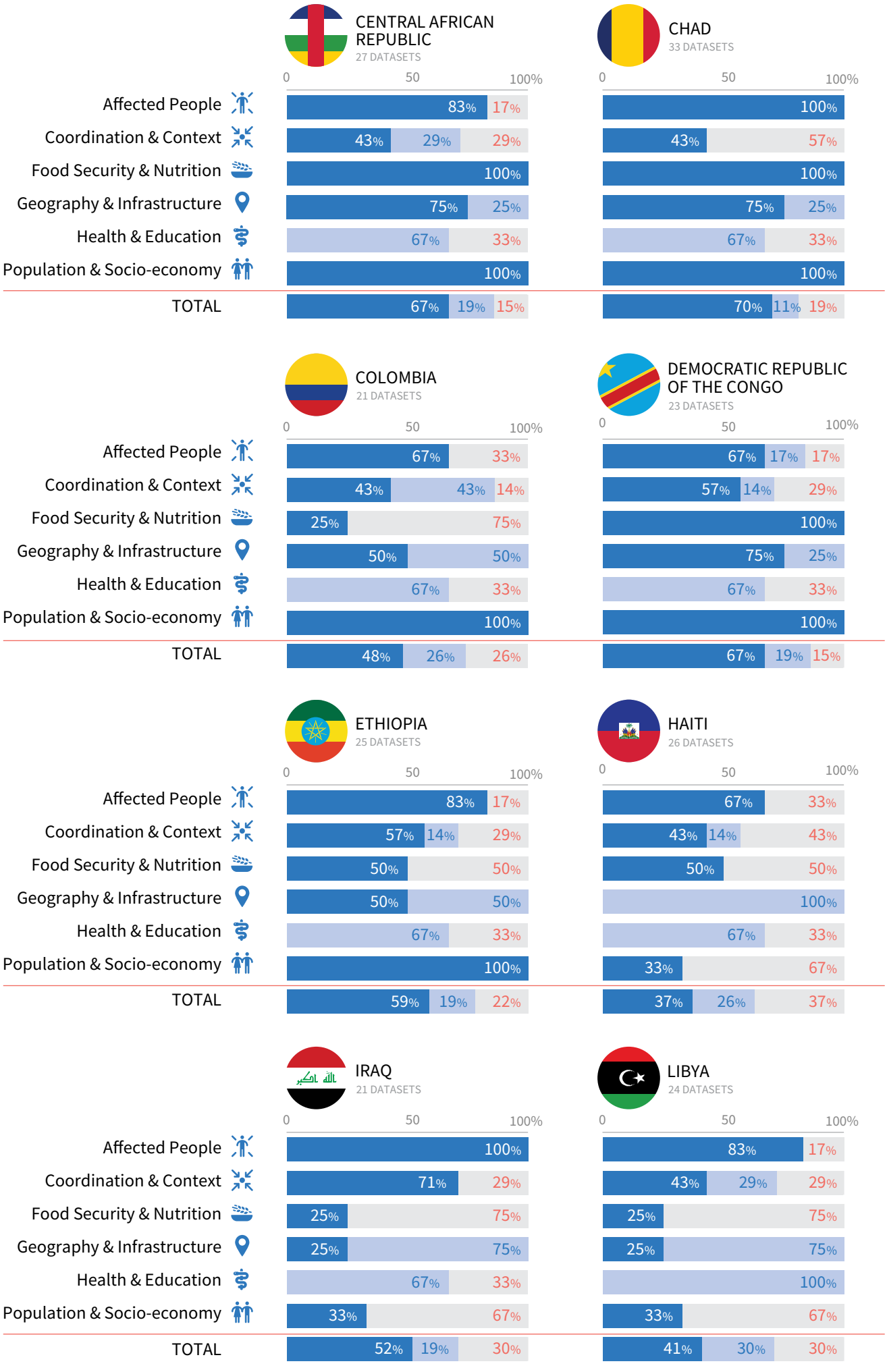
A closer look at Mozambique shows that it has much higher completeness within the affected people category than Zimbabwe, despite being its neighbour and experiencing similar impacts of catastrophic weather events like Cyclone Idai. Several complete datasets for Mozambique come from international organizations like IOM and UNHCR, but the same data is not available for Zimbabwe.

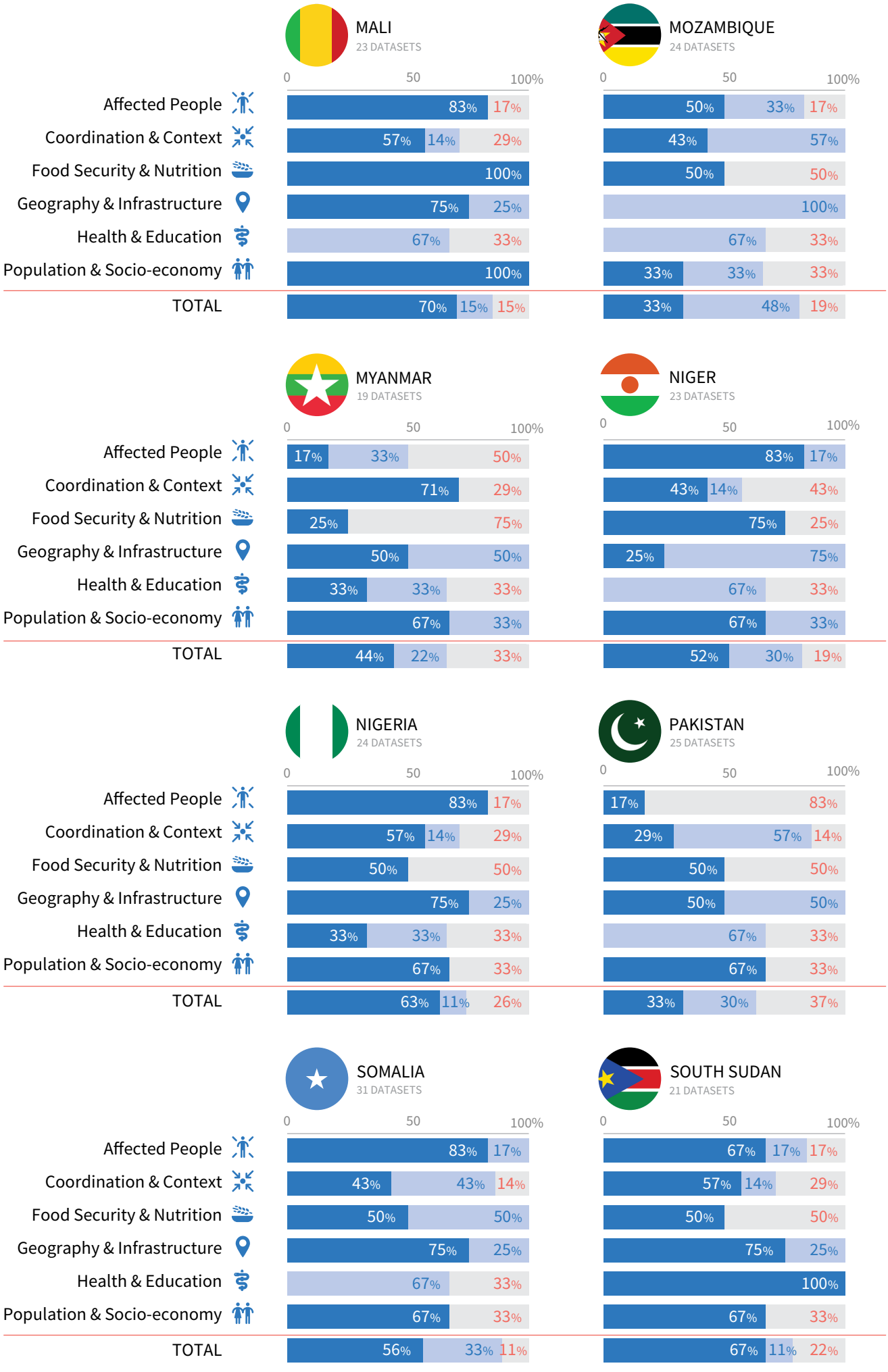
Health and education remains the least complete category at 14 percent, driven by mostly incomplete data for two sub-categories: the location of health facilities and the location of education facilities. Most of this data comes from the Humanitarian OpenStreetMap Team, which provides a crowdsourced dataset for which completeness is difficult to determine.

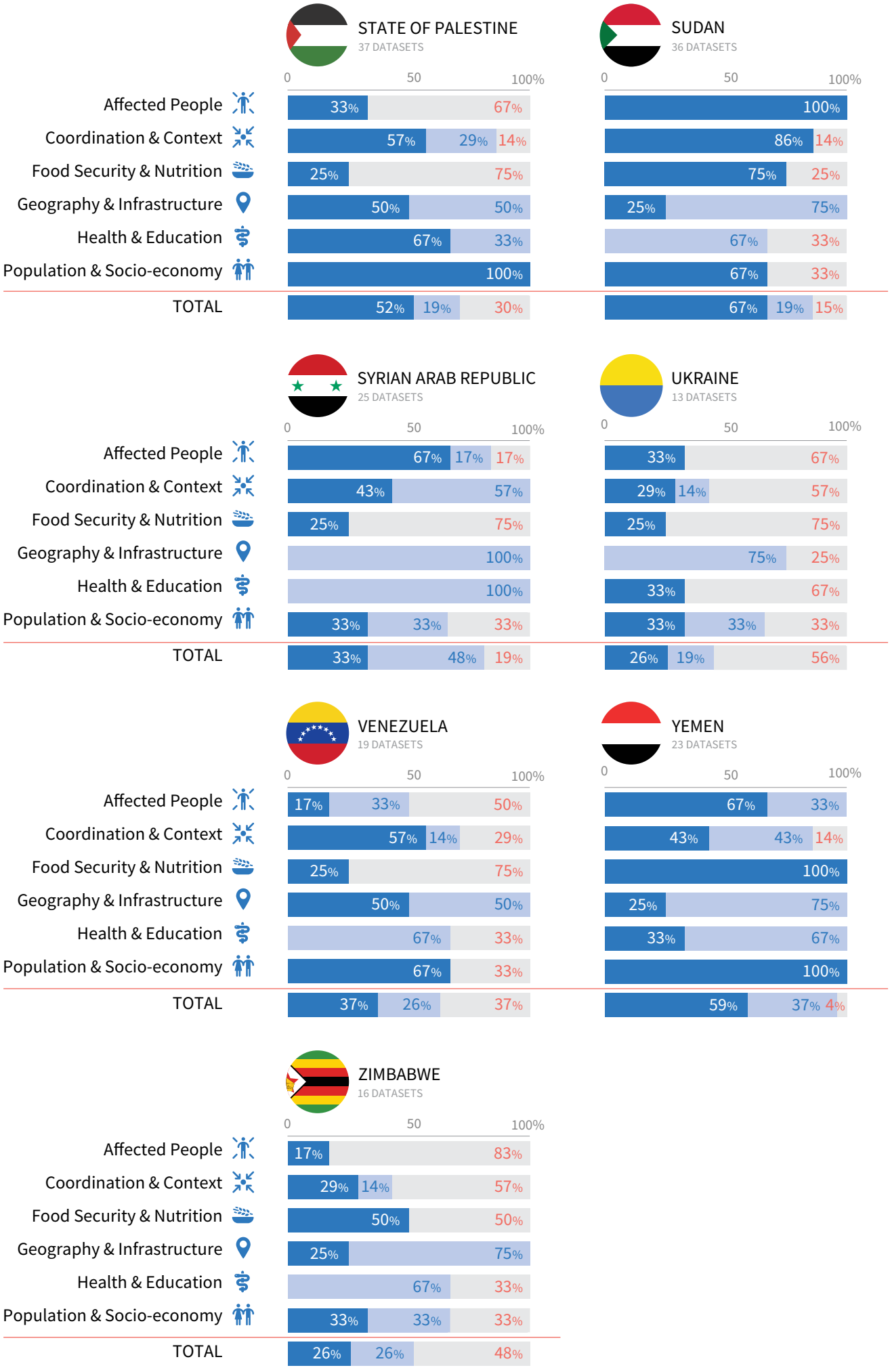
Overall, there are 639 unique datasets across all Data Grids. With 27 sub-categories, we would expect a complete Data Grid to include between 30-40 datasets.[13] For the locations covered in this report, the number ranges from 13 to 37 datasets.



[13] The number of datasets included in each Data Grid depends on how the data is organized and what it represents. In general, each sub-category includes one to two datasets. One dataset can cover multiple sub-categories, such as an Afghanistan displacement dataset from IOM that covers internally displaced persons, refugees and persons of concern, returnees, and humanitarian profile locations. There may also be several incomplete datasets under one sub-category.


CENTRAL AFRICAN REPUBLIC
27 DATASETS
CHAD
33 DATASETS
COLOMBIA
21 DATASETS
DEMOCRATIC REPUBLIC OF THE CONGO
23 DATASETS
ETHIOPIA
25 DATASETS
HAITI
26 DATASETS
IRAQ
21 DATASETS
LIBYA
24 DATASETS
0
50
100%
Affected People
Coordination & Context
Food Security & Nutrition
Geography & Infrastructure
Health & Education
Population & Socio-economy
TOTAL

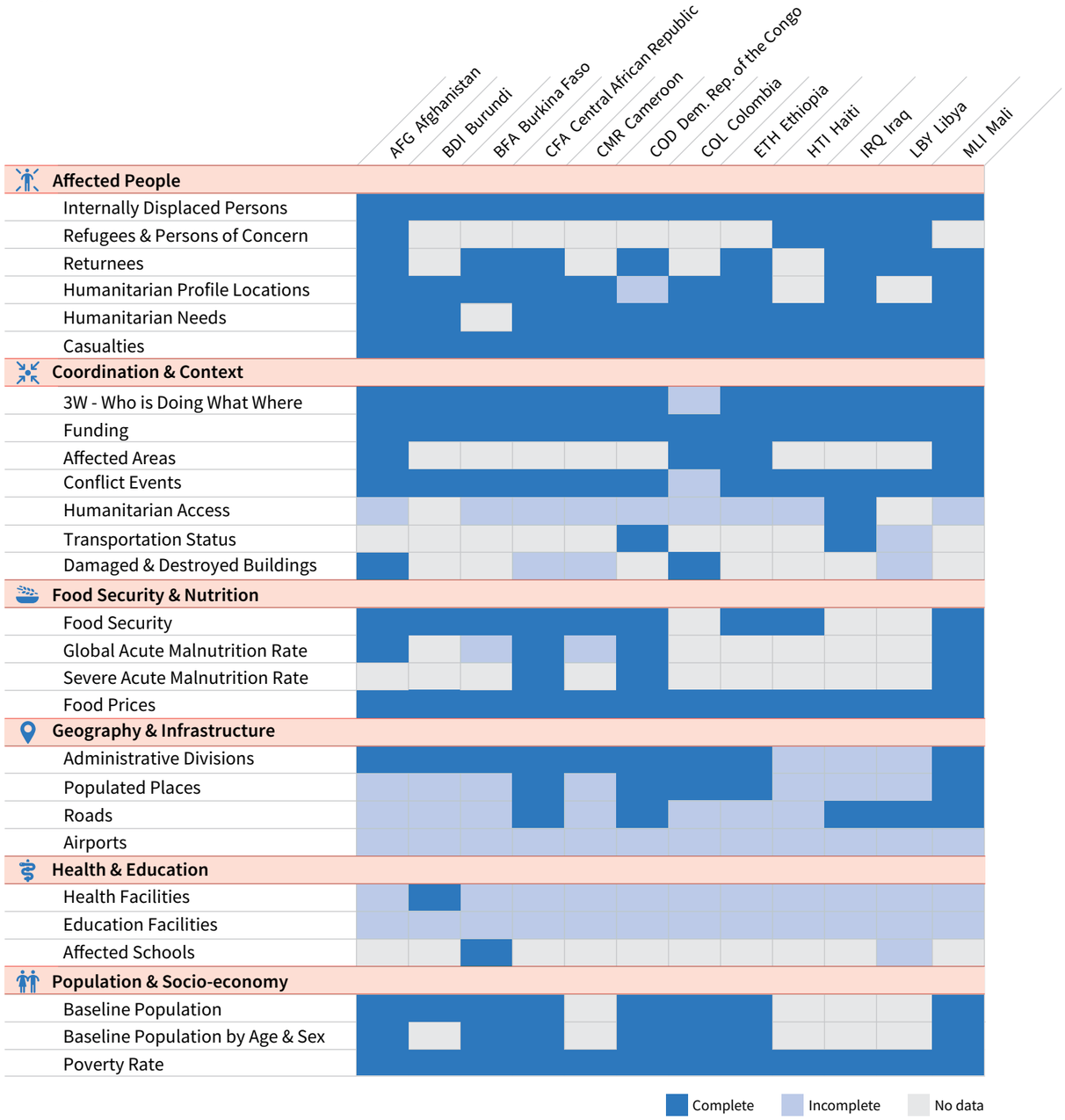

MALI
23 DATASETS
MOZAMBIQUE
24 DATASETS
MYANMAR
19 DATASETS
NIGER
23 DATASETS
NIGERIA
24 DATASETS
PAKISTAN
25 DATASETS
SOMALIA
31 DATASETS
SOUTH SUDAN
21 DATASETS
0
50
100%
Affected People
Coordination & Context
Food Security & Nutrition
Geography & Infrastructure
Health & Education
Population & Socio-economy
TOTAL
83% 17%
57% 14% 29%
100%
75% 25%
67% 33%
100%
70% 15% 15%
50% 33% 17%
43% 57%
50% 50%
100%
67% 33%
33% 33% 33%
33% 48% 19%
17% 33% 50%
71% 29%
25% 75%
50% 50%
33% 33% 33%
67% 33%
44% 22% 33%
83% 17%
43% 14% 43%
75% 25%
25% 75%
67% 33%
67% 33%
52% 30% 19%
83% 17%
57% 14% 29%
50% 50%
75% 25%
33% 33% 33%
67% 33%
63% 11% 26%
17% 83%
29% 57% 14%
50% 50%
50% 50%
67% 33%
67% 33%
33% 30% 37%
83% 17%
43% 43% 14%
50% 50%
75% 25%
67% 33%
67% 33%
56% 33% 11%
67% 17% 17%
57% 14% 29%
50% 50%
75% 25%
100%
67% 11% 22%


STATE OF PALESTINE
37 DATASETS
SUDAN
36 DATASETS
SYRIAN ARAB REPUBLIC
25 DATASETS
UKRAINE
13 DATASETS
VENEZUELA
19 DATASETS
YEMEN
23 DATASETS
ZIMBABWE
16 DATASETS
0
50
100%
Affected People
Coordination & Context
Food Security & Nutrition
Geography & Infrastructure
Health & Education
Population & Socio-economy
TOTAL

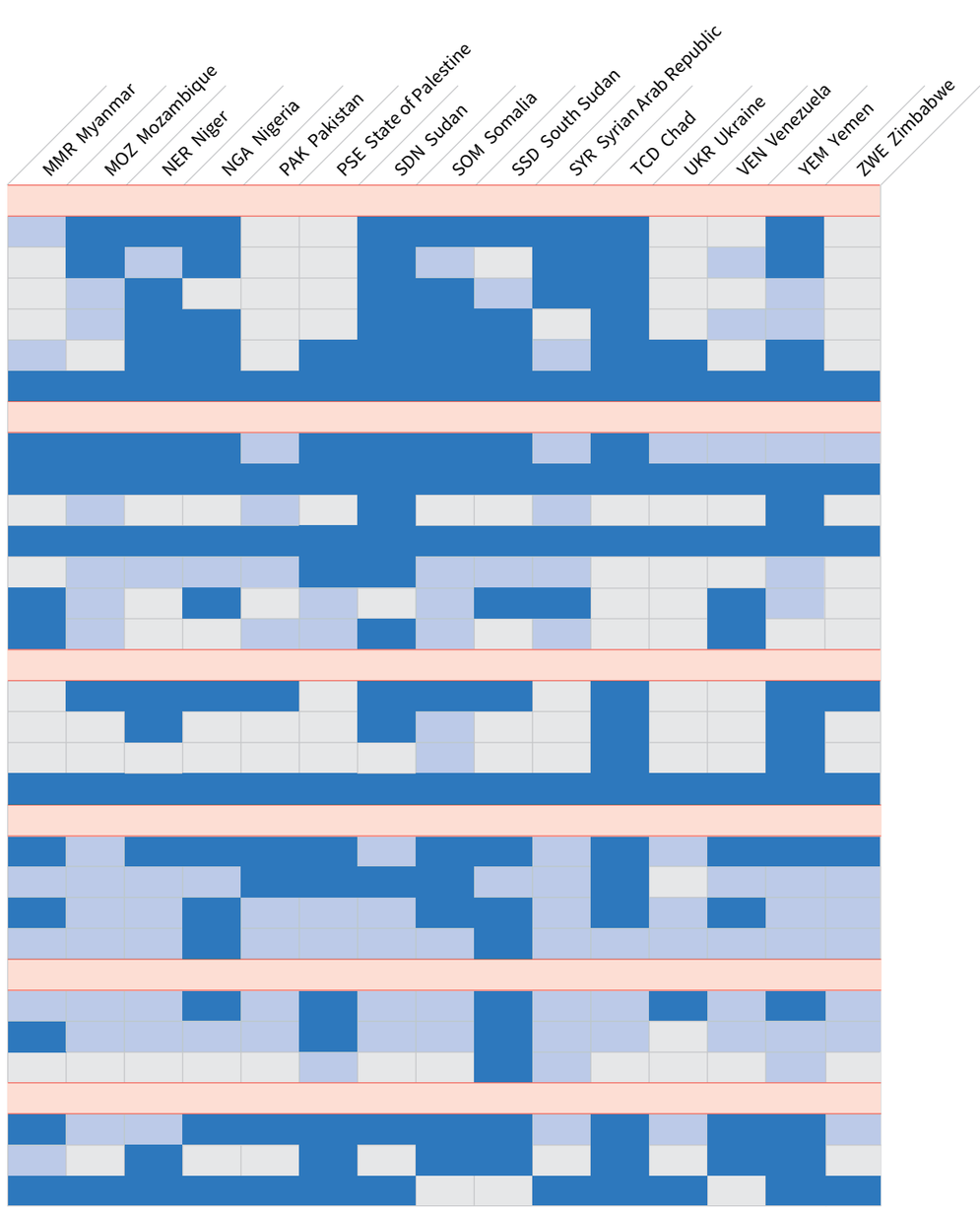
6. COMPLETENESS BY LOCATION AND SUB-CATEGORY

Three sub-categories are 100 percent complete across all locations: casualties, funding, and food prices, owing to the work of ACLED, OCHA, and WFP respectively. The other sub-category approaching total completeness is conflict events at 96 percent; Colombia is the only location without complete data in this sub-category. All other sub-categories have at least some missing or incomplete data.

The GAM and SAM rates pose a particular challenge: GAM is empty across 59 percent of the 27 locations, with SAM even emptier at 78 percent missing. This is largely because this type of data is usually made available only at the national level and therefore does not qualify for inclusion. That said, there are several locations with both sub-national GAM and SAM data, including a notable new contribution from UNICEF in the Democratic Republic of the Congo.



The sub-category for affected schools is also challenging (78 percent empty). This type of data can be sensitive, especially in conflict settings, and therefore difficult to collect and share. On the other hand, data on damaged and destroyed buildings is less likely to be collected or kept up-to-date in non-conflict locations.



7. COUNTRY DEEP DIVE: MALI

Mali has the highest degree of data completeness at 70 percent (a level shared by Chad). Mali has two complete categories: population and socio-economy, and food security and nutrition. Four sub-categories are empty: refugees and persons of concern, transportation status, damaged and destroyed buildings, and affected schools. While there are refugees in Mali, the data available on HDX is not sub-national and therefore excluded.

Datasets on the Mali Data Grid were downloaded more often in 2020 than datasets that were not, showing the particular relevance of this data to HDX users. Almost 50 percent (11 of 23) of Mali's unique datasets were contributed by the OCHA office in Mali in partnership with a variety of sources. HDX also has a two-person team based in Dakar, Senegal that focuses on West Africa. The data for Mali demonstrates how regional knowledge of data availability and trusted local partnerships can drive increased rates of data completeness.

AFFECTED PEOPLE

4 Datasets

Internally Displaced Persons

- Mali Humanitarian Response Plan OCHA Mali

Refugees & Persons of Concern

No Data

Returnees

- Mali Displacement - [IDPs, Returnees] - Baseline Assessment [IOM DTM] International Organization for Migration

Humanitarian Profile Locations

- Afghanistan Displacement Data - Baseline Assessment [IOM DTM] International Organization for Migration

Humanitarian Needs

- Mali : Humanitarian Needs Overview OCHA Mali

Casualties

- Mali - Conflict Data
  Armed Conflict Location & Event Data Project

COORDINATION & CONTEXT

7 Datasets

3W - Who Is Doing What Where

- Mali : Operational Presence
  OCHA Mali
- Current IATI aid activities in Mali
  International Aid Transparency Initiative

Funding

- Mali - Requirements and Funding Data
  OCHA FTS

Affected Areas

- Mali Humanitarian Response Plan OCHA Mali

Conflict Events

- Mali - Conflict Data
  Armed Conflict Location & Event Data Project

Humanitarian Access

- Mali: Attacks on Civilians and Vital Civilian Facilities
  Insecurity Insight
- Mali Access data March 2019 OCHA Mali

Transportation Status

No Data

Damaged & Destroyed Buildings

No Data

FOOD SECURITY & NUTRITION

3 Datasets

Food Security

- Food Security Data in West & Central Africa: Cadre Harmonise (CH) and Integrated Food Security Phase Classification (IPC) data Food Security and Nutrition Working Group, West and Central Africa

Global Acute Malnutrition Rate

- 2019 Nutrition SMART Survey results and 2020 trends OCHA Mali

Severe Acute Malnutrition Rate

- 2019 Nutrition SMART Survey results and 2020 trends OCHA Mali

Food Prices

- Mali - Food Prices World Food Programme

GEOGRAPHY & INFRASTRUCTURE

5 Datasets

Administrative Divisions

- Mali - Subnational Administrative Boundaries OCHA Mali

Populated Places

- Mali - Settlements OCHA Mali
- Mali: High Resolution Population Density Maps + Demographic Estimates Facebook

Roads

- Mali - Roads OCHA Mali

Airports

- Airports in Mali OurAirports

HEALTH & EDUCATION

3 Datasets

Health Facilities

- Mali-healthsites Global Healthsites Mapping Project
- Mali Health Districts OCHA Mali

Education Facilities

- HOTOSM Mali Education Facilities (OpenStreetMap Export) Humanitarian OpenStreetMap Team

Affected Schools

No Data

POPULATON & SOCIO-ECONOMY

3 Datasets

Baseline Population

- Mali - Subnational Population Statistics OCHA Mali

Baseline Population by Age & Sex

- Population of Mali disaggregated by age and by region (2010 - 2035 official projections) OCHA Mali

Affected Areas

- Mali: Global Multidimensional Poverty Index (MPI) Oxford Poverty & Human Development Initiative

8. ORGANIZATION DEEP DIVE: INTEGRATED FOOD SECURITY PHASE CLASSIFICATION

Food security data is critical but often hard to access for many crises. To address this gap, the HDX team worked with colleagues at the IPC Global Support Unit to make this data easier to find and use. As a result of this collaboration, by the end of 2020 IPC had shared 28 datasets on HDX, helping to complete the food security sub-category for 13 locations with Data Grids. In addition to these locations, IPC provides data for ten countries in Africa and three in Central America that do not have Data Grids.

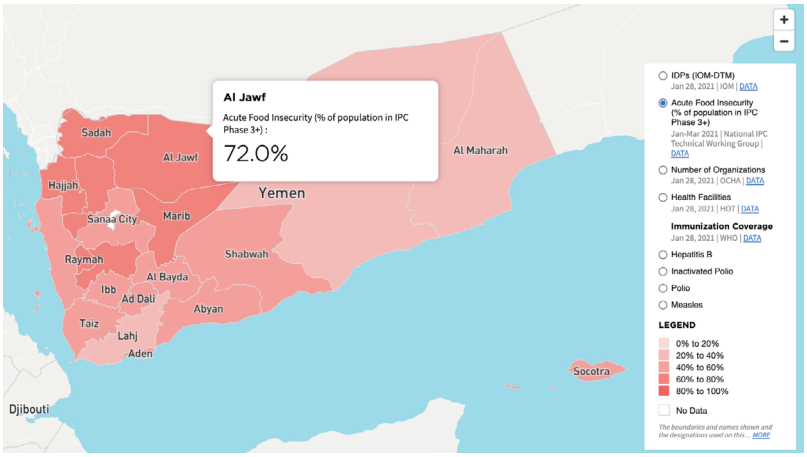
IPC is a common global scale for classifying the severity and magnitude of food insecurity and malnutrition. Originally developed in 2004 in Somalia by FAO's Food Security and Nutrition Analysis Unit, IPC is now a partnership of 15 organizations at the global level, and many more at the regional and country level, dedicated to producing and maintaining the highest possible quality in food security and nutrition analysis.

We started working with the IPC team in June 2020 to establish familiarity with the HDX platform, assess the most efficient way to share their data, and identify which country-level datasets would be most valuable to the HDX community. Although IPC is the contributor of the data, the source for each country dataset is that country's IPC Technical Working Group, normally composed of representatives from the government, UN agencies, NGOs and other relevant stakeholders.

The IPC data that has been shared on HDX is related to their Acute Food Insecurity scale.[14] This data includes the number and percentage of people who are acutely food insecure within a geographic area and each area's phase within the scale. Phase 1 is no or minimal acute food insecurity; Phase 2 is Stressed; Phase 3 is Crisis; Phase 4 is Emergency; and Phase 5 is Catastrophe/Famine. The data covers a current period and a projection for what is expected or likely to happen in the future.

Of the 28 datasets shared under the IPC organization page on HDX, 26 are country-specific, one is regional and one is global. This data is updated periodically throughout the year and shared automatically through IPC's system. The global dataset includes data for all countries and is among the most downloaded, along with country-level data for Afghanistan, Ethiopia and Yemen.

Because the data is comparable, it has been valuable for understanding levels of acute food insecurity as an underlying vulnerability related to the COVID-19 pandemic. For example, the HDX team created a COVID-19 Data Explorer that brings together dozens of datasets from over twenty sources across 63 countries. The IPC data is included in the global map, with the ability to compare across countries, and at the country level, to see variations across states. The image below shows alarming levels of acute food insecurity in Yemen: 72 percent of the population in Al Jawf Governorate are in Crisis or worse (IPC Phase 3 or above).



Acute food insecurity data from IPC for Yemen on the HDX COVID-19 Data Explorer, with a link to the underlying data.

Our productive partnership with the IPC team has proven the value of focused engagement around a specific type of data. Overall, the food security sub-category is complete for 19 of the 27 locations with a Data Grid. Data for 13 of the 19 locations comes from IPC.[15] The remaining six locations — Burkina Faso, Cameroon, Chad, Mali, Niger and Nigeria — are completed by datasets from the Food Security and Nutrition Working Group for West and Central Africa, which uses the Cadre Harmonisé analytical framework.[16]

[14] There are three IPC scales: Acute Food Insecurity, Chronic Food Insecurity and Acute Malnutrition. Only the data for acute food insecurity has been shared on HDX so far. Learn more at **ipcinfo.org**, or watch the November 2020 HDX Data Deep-Dive, Understanding and Communicating IPC Processes and Analyses, at **youtu.be/VcrVqYQLpos**.

[15] Afghanistan, Burundi, Central African Republic, Democratic Republic of the Congo, Ethiopia, Haiti, Mozambique, Pakistan, Somalia, South Sudan, Sudan, Yemen, and Zimbabwe.

[16] Learn more about IPC and Cadre Harmonisé at **bit.ly/the-ch-ipc**.

9. CONTRIBUTING ORGANIZATIONS

Of the 289 organizations currently sharing data on HDX, 32 contribute data that is included in the Data Grids.[17] This is a 14 percent increase from the previous year and reflects the addition of eight organizations.[18] Two of these eight organizations were new to HDX in 2020: IPC and the UNICEF office in the Democratic Republic of the Congo.[19] The Oxford Poverty & Human Development Initiative also began to share their flagship multi-country Multidimensional Poverty Index on HDX in 2020.

Organizations such as WFP and OCHA's Financial Tracking Service share data (food prices and funding, respectively) across all Data Grid locations. Other organizations share data across many but not all of the locations, such as Insecurity Insight which contributes data on violent and threatening incidents against healthcare for 15 of the 27 locations. A few organizations only share a single, valuable dataset related to a specific crisis, such as the baseline population dataset from the Palestinian Central Bureau of Statistics.

Armed Conflict Location & Event Data Project

Assistance Coordination Unit

Demographic and Health Surveys Program

Drew University

Education Cluster, Central African Republic

Education Cluster, Yemen

Explosive Weapons in Populated Areas Community

Facebook Data for Good

Food and Agricultural Organization of the United Nations, Somalia Water and Land Information Management

Food Security and Nutrition Working Group, West and Central Africa

Global Healthsites Mapping Project

Humanitarian OpenStreetMap Team

Insecurity Insight

Integrated Food Security Phase Classification

InterAction

International Aid Transparency Initiative

International Federation of Red Cross and Red Crescent Societies

International Organization for Migration

Myanmar Information Management Unit

OpenStreetMap, Democratic Republic of the Congo

OurAirports

Oxford Poverty & Human Development Initiative

Palestinian Central Bureau of Statistics

REACH Initiative

United Nations Children's Fund, Democratic Republic of the Congo

United Nations High Commissioner for Refugees

United Nations Office for the Coordination of Humanitarian Affairs

UNITAR Operational Satellite Applications Programme

Venezuelan Laboratory of Social Sciences (Laboratorio de Ciencias Sociales de Venezuela)

World Food Programme

World Health Organization

WorldPop

[17] An organization on HDX can be a legal entity or an informal group and may be listed as the source or the contributor of the dataset. The entities listed in this section have created organizations on HDX and manage their data directly. Although most organizations are both the source and contributor for the data, there are cases where this varies. For instance, as part of its coordination role, OCHA aggregates data on humanitarian needs but the data is collected by multiple partners. For modeled data, the dataset may include multiple sources but the organization that has done the analysis contributes the data. This is often the case with datasets on population estimates, food insecurity, and poverty rates.

[18] An organization's dataset may be removed from the Data Grids if it is no longer timely or has been superseded by a more relevant dataset. See Annex B for more information.

[19] Some organizations on HDX have one entry for the entire organization, such as IOM. Others have multiple entries for individual field offices, such as OCHA Sudan and OCHA Afghanistan. See **data.humdata.org/organization**.

10. DATA FOR MODELLING

Predictive modelling refers to the use of statistics and machine learning to analyse current and historical data in order to make predictions about future or unknown events. Modelling represents an opportunity for organizations in the humanitarian sector to turn data into insights that enable decision makers to understand where things might be headed and make interventions accordingly.

However, it is only in recent years that predictive models have begun to be applied in humanitarian response, driven by increased momentum for anticipatory action, or responding before a crisis escalates, by piloting the use of model projections to trigger the early release of funds in places like Bangladesh, Somalia and Ethiopia.[20] The global scale of the COVID-19 pandemic has only accelerated the demand for this type of analysis to inform response strategies.

In researching some of the pitfalls that humanitarian organizations face in using predictive models for anticipatory action, we found that significant challenges persist in relation to data gaps and data quality, limiting the viability and accuracy of model development.[21] Model output is only as good as model input.

The Data Grids include the basic data inputs needed for most humanitarian models, such as the administrative boundaries of a country and population figures within each geographic area. They also include datasets that point to the underlying vulnerabilities in a specific context, such as the number of people who are acutely food insecure or the number of health facilities in a district. However, as the previous sections of this report show, this foundational data can often be missing or incomplete in a given location, which has direct implications on our ability to predict events.

By way of illustration, we have created the below table listing the specific data inputs used for the OCHA-Bucky COVID-19 model.[22] As we have seen since the start of the COVID-19 pandemic, epidemic forecasting is one tool through which we can gain an understanding of the final outbreak size and an indication of when the epidemic may peak in a country. OCHA-Bucky is a model developed by the Johns Hopkins University Applied Physics Laboratory in partnership with OCHA that forecasts the number of COVID-19 cases, hospitalizations, and deaths over two or four weeks at the sub-national and national levels. The model provides humanitarian decision-makers with the capability to plan and manage resources, and is being used by OCHA field offices in six countries: Afghanistan, the Democratic Republic of the Congo, Iraq, Somalia, South Sudan and Sudan.

The effort to develop the OCHA-Bucky model gave us a clear understanding of the COVID-19 data landscape and the broader availability of inputs for modelling. We found that we had to confront several data-related challenges:

1. Spatial granularity: For OCHA-Bucky to be useful, it needed to produce sub-national projections that could inform COVID-19 response planning. We found that only a small fraction of the data inputs were available at the desired spatial resolution (admin 2); most inputs were only available at the admin 1 level.[23] This meant that projections had to be limited to the admin 1 level, obscuring some of the sub-national trends such as differences in transmission rates between cities and rural areas.

2. Timeliness of data: Many of the national-level datasets were not updated regularly. For instance, data on medical comorbidities for the six countries where OCHA-Bucky is currently being implemented was between four and seven years old. In addition, as of January 2021, some of the sub-national data was no longer being maintained. This included the sub-national COVID-19 cases and deaths data from the ministries of health and the ACAPS government measures dataset which was used to account for non-pharmaceutical interventions in place in these locations.[24]

[20] Learn more about our support for the first-ever anticipatory release of funds by the UN's Central Emergency Response Fund at **bit.ly/aa-impact**.

[21] For more, see our *Research Findings on Predictive Analytics for Anticipatory Action* at **bit.ly/parfindings**.

[22] Visit **bit.ly/ocha-bucky** for links to the model source code, documentation, and output reports.

[23] The largest administrative division of a country is called the "first-level administrative division" or "admin 1". Examples of admin 1 divisions are 'states' (United States of America); 'provinces' (Canada); oblasts (Russia); governorates (Syria); departments (France); and cantons (Switzerland). The next-smaller level is the "second-level administrative division" or "admin 2", and so on.

[24] Non-pharmaceutical interventions are measures taken by governments and humanitarian actors to counter a pandemic without the use of medicines. Examples include school closures, physical distancing, self-isolation of symptomatic individuals, etc.

3. Data access and format: Sub-national COVID-19 cases data was available in a machine readable format for two of the six focus countries. Data for the other four countries was only available in PDF reports issued by the ministries of health. The HDX team manually extracted this data on a weekly basis and created machine-readable datasets, a time-consuming process.

4. Data quality: We often received divergent data from the WHO and national ministries of health on COVID-19 cases and deaths. Officially-reported data was missing for days at a time and then batched into a single report, limiting our understanding of trends. This has been exacerbated by a recent deterioration of the quality of COVID-19 data, with several countries now only reporting intermittently.

5. Missing data: We were unable to find sub-national data on the number of COVID-19 deaths for three countries. For all countries, we looked for COVID-19 testing data to assess the quality of the COVID-19 data reported by the authorities but it was largely missing or patchy, making it difficult to estimate the level of underreporting. This could potentially lead projections to provide a more optimistic picture than the actual situation. Similarly, we are working to include the impact of vaccination strategies in the OCHA-Bucky model, and expect data on vaccination rates will be difficult to source for countries experiencing humanitarian crises.

To address these data limitations, we had to adapt our model to use the datasets available only at the national level, and we assumed that the data collected several years ago still represented the situation in the country today. We also used similar or neighbouring countries as a proxy to fill data gaps whenever possible. For instance, data on handwashing facilities was missing for South Sudan, so we used the values for Sudan as a proxy. Ultimately, the poor quality of the inputs impacted the overall quality of the projections and resulted in greater uncertainty with the results.

On a more positive note, our modelling work did create a positive feedback loop in that as we identified inconsistencies or missing data, we worked with local OCHA offices and partners to make corrections or find the required data. We hope this dynamic can drive improvements over time and unlock new datasets as stakeholders see the value of sharing data in return for model insights.

DATA AVAILABILITY FOR THE OCHA-BUCKY COVID-19 MODEL

DATA INPUT	ORGANIZATION	AFGHANISTAN	DEMOCRATIC REPUBLIC OF THE CONGO	IRAQ	SOMALIA	SOUTH SUDAN	SUDAN
Population							
Gender and age disaggregation	WorldPop						
Urban/rural disaggregation	European Commission JRC	O	O	O	O	O	O
Administrative boundaries	OCHA						
Vulnerability							
Acute food insecurity	IPC						
Indoor air pollution	WHO	N and O	N and O	N and O	N and O	N and O	N and O
Access to handwashing facilities	WHO, UNICEF	N and O	N and O	N and O	N and O		N and O
Medical co-morbidities							
Prevalence of cardio vascular diseases	WHO	N and O	N and O	N and O	N and O		
Prevalence of diabetes	IHME	N and O	N and O	N and O	N and O	N and O	N and O
Prevalence of smoking	IHME	N and O	N and O	N and O	N and O	N and O	N and O
Mobility							
Road networks	HOT						
Fraction of households owning a car	WHO	N and O	N and O	N and O			N and O
Household size	UN DESA	N and O	N and O	N and O		N and O	N and O
COVID-19							
Cumulative cases and deaths	WHO	N	N	N	N	N	N
Cumulative cases (subnational)	Ministries of Health						
Cumulative deaths (subnational)	Ministries of Health						
Contact matrix							
Interaction between age groups	Scientific Literature[25]			N			
Non-pharmaceutical interventions							
NPIs in place	ACAPS						

Data availability:

	Complete	Sub-national, up-to-date, machine readable
O	Out of date	Data is more than one year old
N	National only	One figure per country and up-to-date
N and O	National only and out of date	One figure per country and out of date
	No data	Data is missing

11. CONCLUSION

We will continue to update the Data Grids throughout the year as organizations share new, relevant data. The current status for each location is always available on HDX, both on the relevant location page and on the Overview of Data Grids page.[26] We will periodically review the categories and sub-categories to see if they should be removed or expanded and would welcome feedback on possible improvements. Please be in touch with questions or comments at centrehumdata@un.org.

[25] Prem K, Cook AR, Jit M (2017) Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS Comput Biol 13(9): e1005697. https://doi.org/10.1371/journal.pcbi.1005697.

[26] https://data.humdata.org/dashboards/overview-of-data-grids

ANNEX A: DATA GRID SUB-CATEGORY DEFINITIONS

CATEGORY	SUB-CATEGORY/DEFINITION
Affected People 	**Internally-Displaced Persons** Tabular data of the number of displaced people by location. Locations can be administrative divisions or other locations (such as camps) if an additional dataset defining those locations is also available.
	Refugees and Persons of Concern Tabular data of the number of refugees and persons of concern either in the country or originating from the country disaggregated by their current location. Locations can be administrative divisions or other locations (such as camps) if an additional dataset defining those locations is also available or if the locations' coordinates are defined in the tabular data.
	Returnees Tabular data of the number of displaced people who have returned.
	Humanitarian Profile Locations Vector or tabular data with coordinates representing the locations at which displaced people are gathered.
	Humanitarian Needs Tabular data of the number of people in need of humanitarian assistance by location and humanitarian cluster/sector.
	Casualties Number of deaths and/or persons injured, disaggregated by location. Values can be cumulative totals or a time series of new deaths and/or injured persons.
Coordination & Context 	**3W - Who is doing what where** List of organizations working on humanitarian issues, by humanitarian cluster/sector and disaggregated by administrative division.
	Funding Tabular data listing the amount of funding provided by humanitarian cluster/sector.
	Affected Areas Vector data or tabular data by administrative division which describe the type and/or severity of impacts geographically.
	Conflict Events Vector data or tabular data with coordinates describing the location, date, and type of conflict event.
	Humanitarian Access Tabular or vector data describing the location of natural hazards, permissions, active fighting, or other access constraints that impact the delivery of humanitarian interventions.
	Transportation Status Vector or tabular data representing local transportation routes with an indication of status or current practicability.
	Damaged and Destroyed Buildings Vector data with locations of damaged/destroyed buildings and an indication of damage level or tabular data indicating percentage or total number of buildings in each damage category by administrative divisions.

CATEGORY	SUB-CATEGORY/DEFINITION
Food Security & Nutrition	**Food Security** Vector data representing the IPC/CH acute food insecurity phase classification or tabular data representing population or percentage of population by IPC/CH phase and administrative division. **Global Acute Malnutrition Rate** Tabular data specifying the global acute malnutrition (GAM) rate by administrative division. **Severe Acute Malnutrition Rate** Tabular data specifying the severe acute malnutrition (SAM) rate by administrative division. **Food Prices** Time series prices for common food commodities at a set of locations.
Geography & Infrastructure	**Administrative Divisions** Vector geographic data describing the sub-national administrative divisions of a location, usually a country, including the names and unique identifiers, usually p-codes, of each administrative division. To be considered "complete", and included here, the humanitarian community working in the location has to have endorsed a preferred set of administrative boundaries as the Common Operational Dataset (COD). **Populated Places** Vector data or tabular data with coordinates representing the location of populated places (cities, towns, villages). **Roads** Geographic data describing the location of roads with some indication of the importance of each road segment in the transportation network. The data should exclude or indicate roads that are not usable by typical four-wheel-drive vehicles (footpaths, etc.). **Airports** Geographic data representing all operational airports including a name or other unique identifier and an indication of what types of aircraft can use each.
Health & Education	**Health Facilities** Vector data or tabular data with coordinates representing health facilities with some indication of the type of facility (clinic, hospital, etc.). **Education Facilities** Vector data or tabular data with coordinates representing education facilities with some indication of the type of facility (school, university, etc.). **Affected Schools** Vector data or tabular data with coordinates representing education facilities that have been affected by a crisis with some indication of the nature of the effect and the operational status of each facility.
Population and Socio-economic Indicators	**Population** Total population aggregated by administrative division. **Population by Age and Sex** Total population disaggregated age and sex categories, aggregated by administrative division. **Poverty Rate** Population living under a defined poverty threshold, aggregated by administrative division and represented as a percentage of total population or as an absolute number.

ANNEX B: INCLUSION OF DATA IN THE DATA GRIDS

The Data Grids list datasets according to a standard set of criteria, with an evaluation of their completeness telling users in advance what to expect.

This evaluation is done when a new dataset is uploaded to HDX, when a request is received by the HDX team to find data for a particular location or category, or as a periodic check by the HDX team to move Data Grids closer to completeness. The team searches for datasets to add to the Data Grids in three ways:

1. Reviewing whether any datasets already on HDX could be added to a Data Grid;
2. Communicating with the broader humanitarian community, such as OCHA Information Management Officers, to learn about potentially relevant data sources; and
3. Searching for data to see whether other sources outside the immediate humanitarian community's knowledge might exist (for example, in academia).

Each of the 27 sub-categories on the Data Grid contains data relating to a unique humanitarian theme as listed in Annex A. The first step in determining whether a dataset is to be included in a Data Grid is to check whether the dataset meets this thematic requirement. Datasets that are not considered relevant are excluded. The next step is to determine if the dataset is sub-national (i.e. contains detail for different parts of a country, typically by administrative division). If the data is at the national level only, it is excluded.

There are then three main criteria for whether a relevant, sub-national dataset is included in the Data Grid as 'complete' or 'incomplete': 1) geographically completeness; 2) in commonly-used formats; and 3) timely (full definitions are given below). If all three criteria are met, the dataset will be considered 'complete'. If not, it will be considered 'incomplete'. The dataset will then be compared against any existing datasets for a location. If the sub-category is empty, or if the data would complement other datasets in a sub-category, the HDX team will add it to the Data Grid.

The sub-category is 'complete' if it has at least one 'complete' dataset. Otherwise, if the sub-category contains only 'incomplete' datasets, then that sub-category will also be 'incomplete'. Overall category completeness refers to the proportion of sub-categories in the category that are complete. Similarly, completeness for a location refers to the proportion of sub-categories that are complete in the location.

Sub-categories are considered 'empty' if no datasets on HDX meet the above-mentioned criteria. In general, data can be missing for three main reasons:

1. It is not collected (e.g. because nobody is present to do so, because it is unsafe to access areas to collect it, because it requires investment and resources that are not available, or because nobody prioritizes it as a gap to fill);
2. It is collected but not publicly shared (e.g. because the collecting organization does not have an open data policy, because the data is sensitive and should not be shared, because the collecting entity fears sharing the data with actors they do not trust, or because it requires investment and resources to clean and share it that are not available);
3. It is collected and shared but is not shared to HDX, or known about by the HDX team.

DETAILED COMPLETENESS CRITERIA

The criteria for evaluating completeness for relevant, sub-national data are detailed below:

Is the data geographically complete (or as complete as possible) with explicit location data?

- Is the dataset geographically complete, or as complete as possible? If the dataset is disaggregated by administrative divisions, does it cover all of them? If it does not, is the meaning of a missing administrative division defined in the metadata? If there is no comprehensive list to compare against (for example, with spontaneous displacement locations), does the dataset make it clear if it attempts to be comprehensive or not? This comprehensiveness requirement means that crowd-sourced datasets, like

those derived from HOT, cannot always be considered complete, even though they may be the most complete dataset available.

- Are location references defined? The dataset should contain explicit geographic data (i.e. Geographic Information System data or tabular data with latitude and longitude fields)? If not, the dataset should be joinable to an available dataset that defines those locations.
- If the dataset is disaggregated using administrative divisions, does it use the lowest-used administrative level? Locations often have several administrative division levels with varying levels of governance, and for the dataset to be considered complete it should have data down to the lowest-used level in order to be fully disaggregated.

Is the data in commonly-used formats?

- Is it stored in a common file format? We include CSV, XLS, XLSX, SHP, etc. Formats like JSON, GPKG and others that are more difficult for the typical humanitarian data specialist would be marked 'incomplete.'
- Is the data tidy? Field names and data rows should be easy to determine. There should not be subtotal rows interspersed with data rows. The required data for the category should be in a single table on the same tab. For tabular data with coordinates, the x and y columns (usually longitude and latitude) should be in decimal degree format and separated into two columns.

Is the data timely?

- Has the dataset become out of data? Depending on how frequently the dataset is expected to be updated, the HDX team considers the age of the data and whether the dataset should have been superseded.