THE STATE OF OPEN HUMANITARIAN DATA 2022

ASSESSING DATA AVAILABILITY ACROSS HUMANITARIAN CRISES

OCHA centre for humdata HDX

TABLE OF CONTENTS

1. INTRODUCTION 4

2. KEY MESSAGES 6

3. GLOBAL OVERVIEW 8

4. COMPLETENESS BY LOCATION, CATEGORY AND SUB-CATEGORY 10

5. COMPLETENESS BY LOCATION AND CATEGORY 11

6. COMPLETENESS BY LOCATION AND SUB-CATEGORY 15

7. COUNTRY DEEP-DIVE: SOMALIA 17

8. ORGANIZATION DEEP-DIVE: UNITED NATIONS POPULATION FUND 19

9. CONTRIBUTING ORGANIZATIONS 20

10. DATA FOR MODELLING 21

11. CONCLUSION 23

ANNEX A: DATA GRID SUB-CATEGORY DEFINITIONS 24

ANNEX B: INCLUSION OF DATA IN THE DATA GRIDS 26

ANNEX C: CHANGES TO THE DATA GRIDS 28

ACKNOWLEDGEMENTS

This report was produced in January 2022 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
CODs	Common Operational Datasets
FAO	Food and Agriculture Organization of the United Nations
GAM	Global Acute Malnutrition
HDX	Humanitarian Data Exchange
HRP	Humanitarian Response Plan
IDP	Internally Displaced Person
IOM	International Organization for Migration
IPC	Integrated Food Security Phase Classification
NGO	Non-Governmental Organization
OCHA	United Nations Office for the Coordination of Humanitarian Affairs
SAM	Severe Acute Malnutrition
UN	United Nations
UNDRR	United Nations Office for Disaster Risk Reduction
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

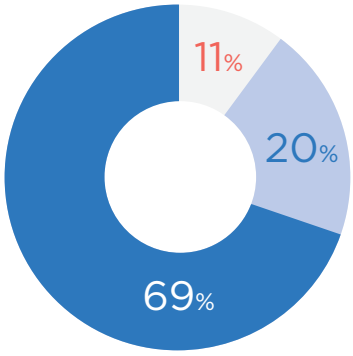
1. INTRODUCTION

In our third year of producing *The State of Open Humanitarian Data*, we can report steady progress in closing data gaps across most humanitarian operations. This momentum is the result of global advocacy and investment paired with field-level data sharing and outreach. Although the demand for data following the onset of the COVID-19 pandemic has peaked, it remains strong.

Our understanding of data availability and use comes from managing the Humanitarian Data Exchange (HDX), an open platform for finding and sharing data across crises and organizations.[1] The data on HDX reflects the reality of the world's worst humanitarian emergencies, from persistent displacement to a lack of food and shelter for vulnerable populations. The data also shows the response to these crises, from who is providing what assistance to funding levels and more. The UN estimates that 274 million people will need humanitarian assistance and protection in 2022, at a cost of US$41 billion.

In 2021, HDX was used by 1.4 million people in 236 countries and territories, maintaining the growth we saw in 2020. Organizations added 2,071 new datasets to the platform, bringing the total to almost 18,500 datasets, which were downloaded over 1.8 million times. Although HDX includes data about all countries in the world, we concentrate here on 27 locations with Humanitarian Response Plans (HRPs).[2] In 2021, 35 percent of all datasets downloaded from HDX were related to an HRP location.[3]

At the start of 2022, we estimate that 69 percent of relevant, complete crisis data is available across 27 humanitarian operations, based on the analysis of the Data Grids. If we add the data that is relevant but incomplete, the total is 89 percent. This leaves 11 percent of categories with data that does not meet the criteria or with no data. The 27 Data Grids include 482 unique datasets, with a range of 12-22 datasets per location.



Total Percentage Data Complete

69%

Total Percentage Data Incomplete

20%

Total Percentage No Data

11%

Data completeness: Complete Incomplete No Data

This report contains details on the data available for each location, category and sub-category covered in the Data Grids as of 31 December 2021. It includes a country deep-dive for Somalia and showcases the work of the United Nations Population Fund (UNFPA) on population statistics. It also has an assessment of the data required for modelling, specifically for developing anticipatory action trigger mechanisms, with examples from Ethiopia, Malawi, Nepal and the Philippines.

[1] **https://data.humdata.org/**

[2] 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 or sector.

[3] We use Mixpanel, a third-party analytics service, to understand the behavior of users on the HDX platform. We do not send identifying information to Mixpanel.

Data Grid Criteria

The HDX Data Grids narrow the focus within each HRP location to a limited set of foundational data needed to understand a humanitarian context. They provide a comparable way to assess data availability across locations and categories and are the basis for the analysis in this report.

The Data Grids include six categories: affected people; coordination and context; food security and nutrition; geography and infrastructure; health and education; and population and socio-economy. (See Annex A for definitions.)

Data may be included in a Data Grid if it is relevant to the category and sub-national. 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, then it is considered 'incomplete'. The sub-category is complete if it includes at least one dataset that is complete. If the sub-category contains only incomplete datasets, then that sub-category is considered incomplete. A sub-category is empty if no data meets the above criteria or the data does not exist on HDX. (See Annex B for the Data Grid criteria and curation process.)

In 2021, HDX maintained Data Grids for 27 locations and 21 sub-categories. The 27 locations are consistent with 2020, but the number of sub-categories decreased by six. Six were merged into three due to overlapping datasets and three were removed due to a lack of relevance across all locations. (See Annex C for changes to the Data Grids.)

We recognize the valuable contributions of all data-sharing organizations publishing data on HDX, and in particular the 20 organizations that populate the Data Grids, with reference to 125 data sources. Datasets included in the Data Grids are downloaded almost four times more than the average dataset on HDX. Trusted partnerships and focused advocacy efforts have led to many new or updated datasets in the Data Grids this year, including:

- Sex and age disaggregated population statistics for Burundi with updated projections for 2021, contributed by UNFPA and sourced to the National Statistics Office.[4] (See more on this type of data in Section 8.)
- The number of health facilities and their functionality by province in Afghanistan, contributed by OCHA Afghanistan and sourced to WHO. The location of the health facilities is not provided due to the sensitivity of this data.[5]
- The number of registered internally displaced people (IDPs) in Burkina Faso at the commune, province and regional level with gender and age breakdown, contributed by OCHA Burkina Faso and sourced to the Conseil National de Secours d'Urgence et de Réhabilitation.[6]
- The location of schools and their status by commune in Mali, contributed by OCHA Mali on behalf of the Ministry of Education.[7]
- Transportation status of roads in the Central African Republic (CAR) indicating the level of difficulty for types of vehicles, contributed by OCHA CAR and compiled by UNDP.[8]
- Joint severity analysis by sector at the district level for Zimbabwe, via OCHA ROSEA on behalf of multiple humanitarian partners.[9]

[4] **https://data.humdata.org/dataset/cod-ps-bdi**
[5] **https://data.humdata.org/dataset/afghanistan-health-facilities**
[6] **https://data.humdata.org/dataset/situation-des-personnes-deplacees-internes**
[7] **https://data.humdata.org/dataset/mali-schools**
[8] **https://data.humdata.org/dataset/republique-centraficaine-acces-physique-dans-les-communes**
[9] **https://data.humdata.org/dataset/zimbabwe-affected-areas-and-needs-data**

We hope that our analysis will increase awareness of the data available for humanitarian response activities and draw attention to what is missing within a critical subset of data. Humanitarian crises are dynamic and so is the data needed to understand them. The HDX team will continue to be strong data advocates and work in partnership with the dozens of organizations contributing and using data on the platform. We look forward to continued collaboration and to closing data gaps.

2. KEY MESSAGES

- Progress is being made in closing data gaps across most humanitarian operations. This momentum is the result of global advocacy and investment paired with field-level data sharing and outreach.

- Although the demand for data following the onset of the COVID-19 pandemic has peaked, it remains strong. In 2021, HDX was used by 1.4 million people in 236 countries and territories. Around 1.8 million datasets were downloaded.

- We estimate that 69 percent or over two-thirds of relevant, complete crisis data is available across 27 locations with humanitarian operations, based on the analysis of the HDX Data Grids. If we add the data that is relevant but incomplete, the total is 89 percent. This leaves about 11 percent of the Data Grids with no data.

- Afghanistan (86 percent) and CAR (86 percent) have the most complete Data Grids. The next most complete locations are Chad (81 percent), the Democratic Republic of the Congo (DRC) (81 percent), and Sudan (81 percent).

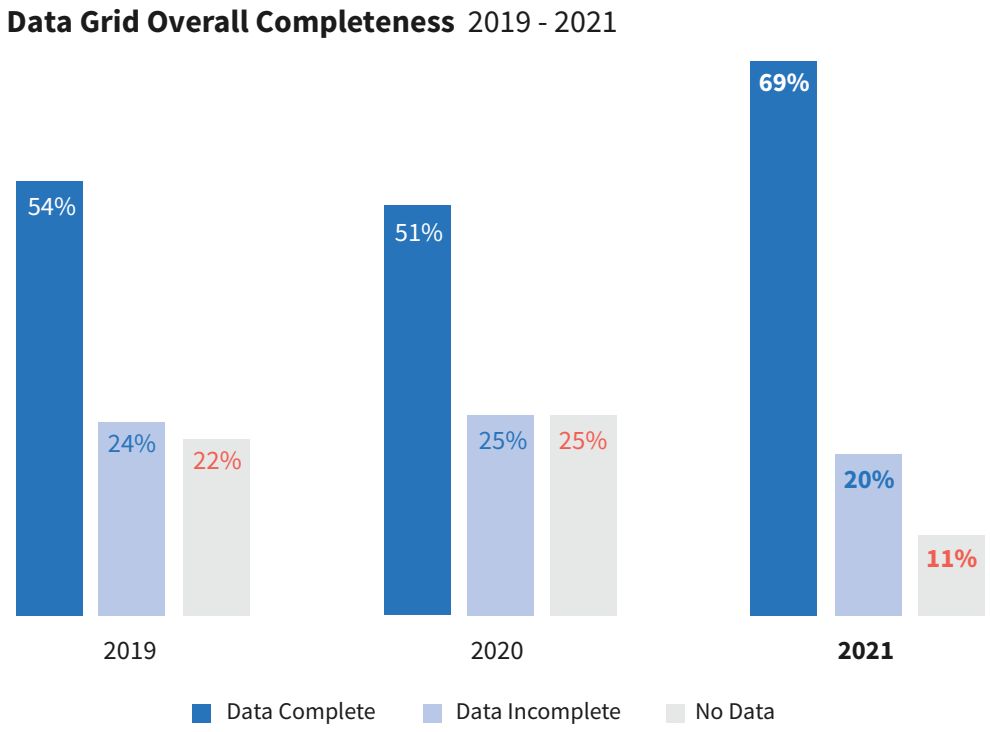
- Venezuela (43 percent of the sub-categories with no data), Colombia (24 percent of the sub-categories with no data) and Libya (19 percent of the sub-categories with no data) are the countries with the largest data gaps.

- Significant improvements were made with the availability of acute malnutrition data (from 19 to 73 percent complete). This is due to outreach with local partners by OCHA field offices.

- Datasets included in the Data Grids are downloaded almost four times more than the average dataset on HDX. Ethiopia is the most popular location and geography and infrastructure is the most popular category (as measured by total dataset downloads). The top five user locations were the United States, Pakistan, the United Kingdom, Ethiopia and Nigeria.

- Humanitarian crises are dynamic and so is the data needed to understand them. The availability and completeness of data will fluctuate year-to-year as new data requirements come into focus. Since 2019, we have seen an increase in data completeness from 54 percent to 69 percent and the categories with no data fall from 25 percent to 11 percent (see chart below).

- A priority for 2022 will be improving access to forecast, observational and impact data about hazards and shocks. This will be essential to expanding anticipatory action frameworks to more countries with humanitarian operations. To give momentum to these efforts, we will create a new category in the HDX Data Grids.



We call on partners to generate and share data that is missing or incomplete for many humanitarian crises, including:

- The location of education facilities (potential sources: national governments, UNICEF, Education Cluster).
- The location of health facilities (potential sources: national governments, Health Cluster, WHO).
- Local transportation routes with an indication of status (potential sources: national governments, WFP, Logistics Cluster).
- Reference lists of historical hazards and their impact on vulnerable populations (potential sources: national governments, humanitarian agencies, UNDRR, researchers).
- Hazard forecast and observational data (potential sources: national governments, regional climate service providers, researchers).

In cases where data is sensitive, organizations can use HDX Connect[10] to share only the metadata and make the underlying data available by request. Alternatively, an incomplete or aggregated version of the dataset that removes sensitive information (such as locations) can be shared publicly.[11]

[10] HDX Connect datasets still contribute to the completeness of a Data Grid. Learn more: **bit.ly/hdx-connect**

[11] Learn more about our process for statistical disclosure control: **https://centre.humdata.org/learning-path/disclosure-risk-assessment-overview/**

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

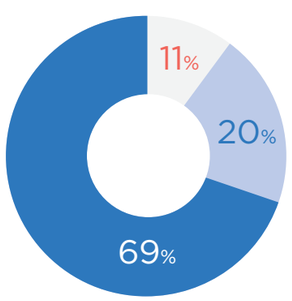
6

Number of Sub-Categories

21

Number of Contributing Organizations

20



Total Percentage Data Complete

69%

Total Percentage Data Incomplete

20%

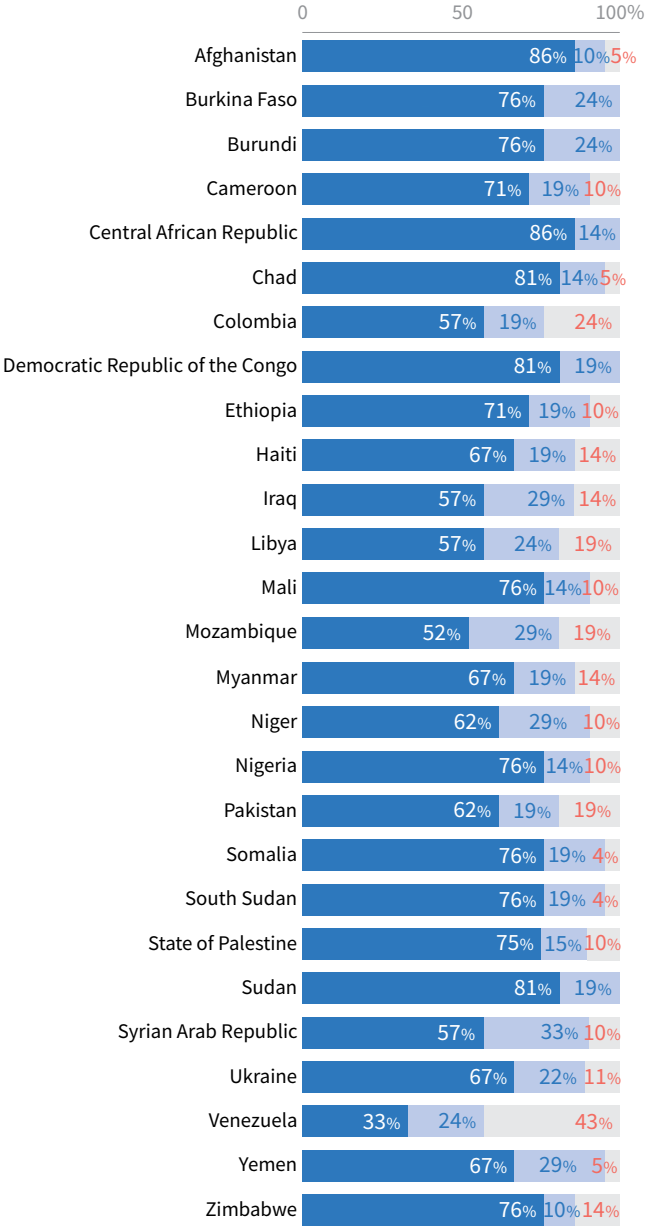
Total Percentage No Data

11%

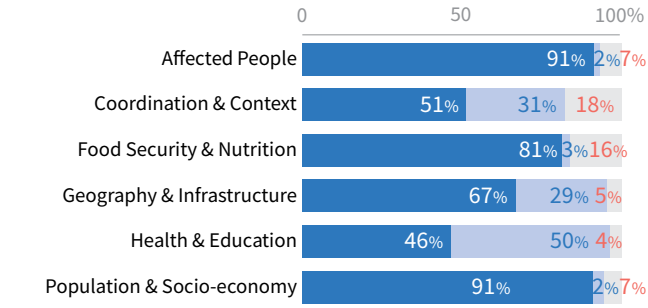
4. COMPLETENESS BY LOCATION, CATEGORY AND SUB-CATEGORY

Given that the Data Grids became more complete over the course of 2021 (from 51 percent to 69 percent), we can see improvements across location, category and sub-category. Five countries – Burkina Faso, Burundi, DRC, CAR and Sudan – no longer have any sub-categories with missing data. The categories with the most improvement are affected people and population and socio-economy. Overall, the sub-categories with no data have decreased from 25 percent to 11 percent year on year.

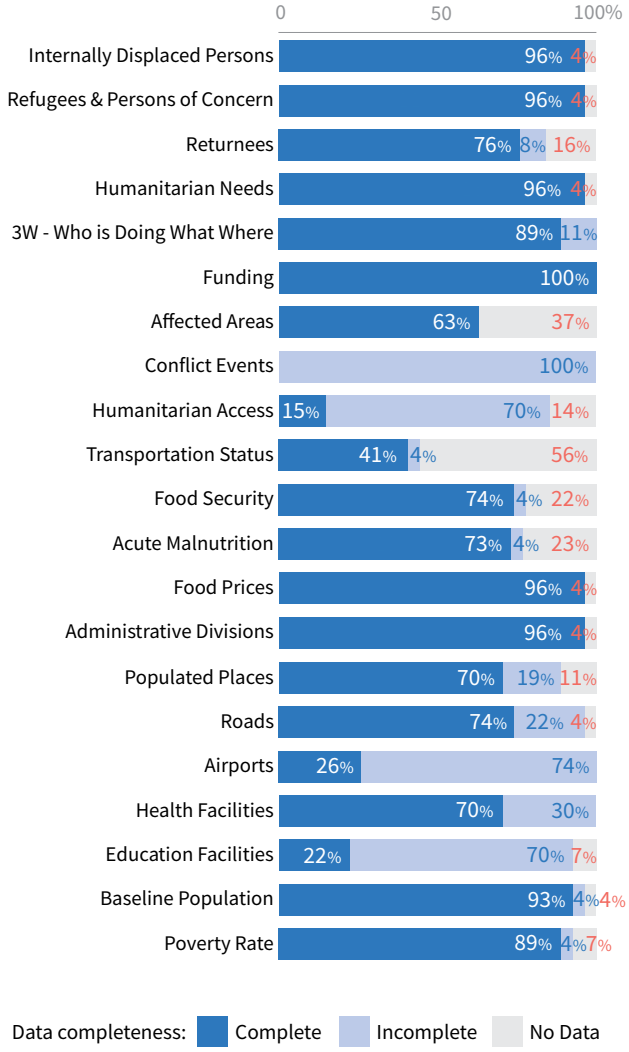
BY LOCATION



BY CATEGORY



BY SUB-CATEGORY



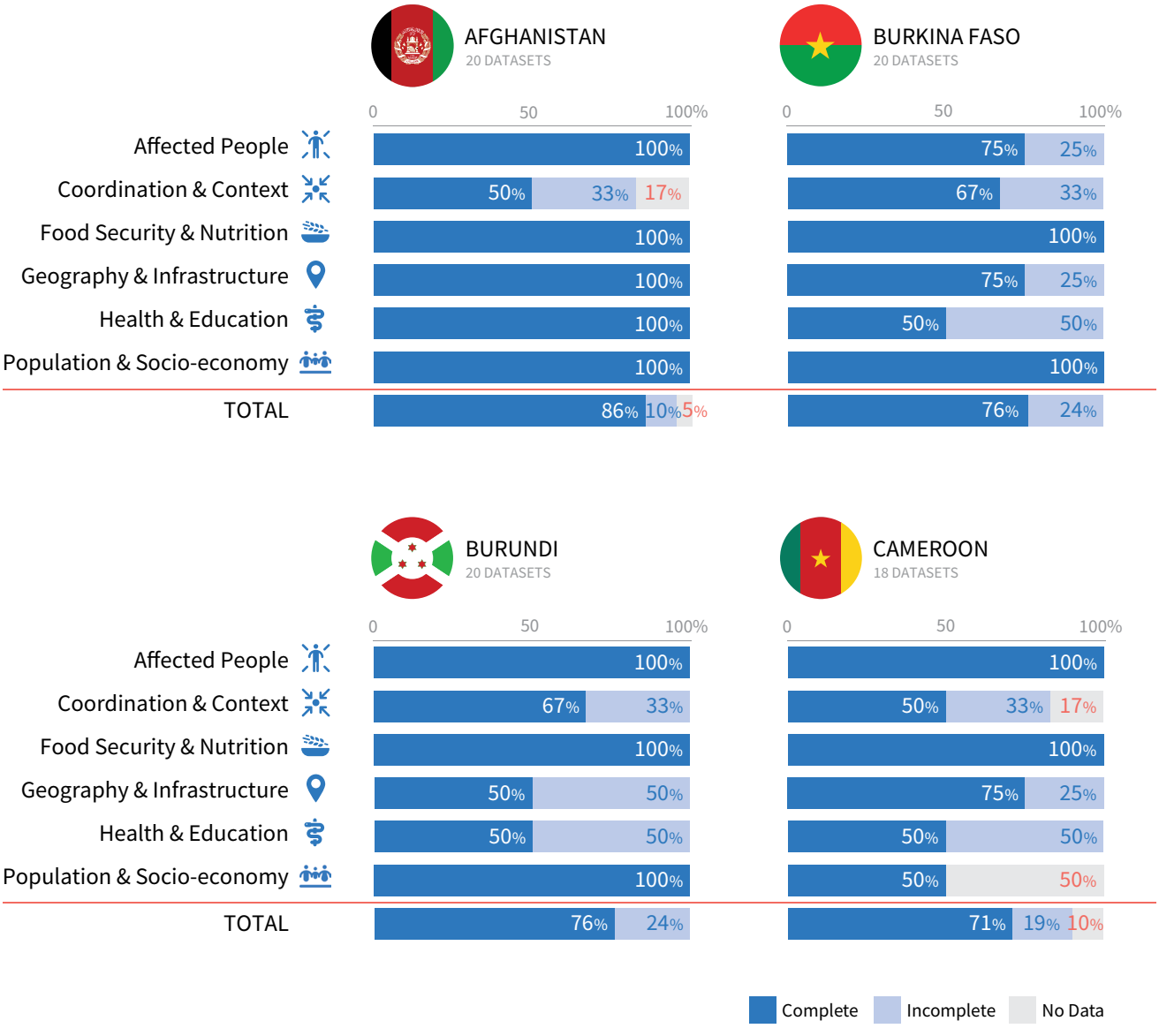
5. COMPLETENESS BY LOCATION AND CATEGORY

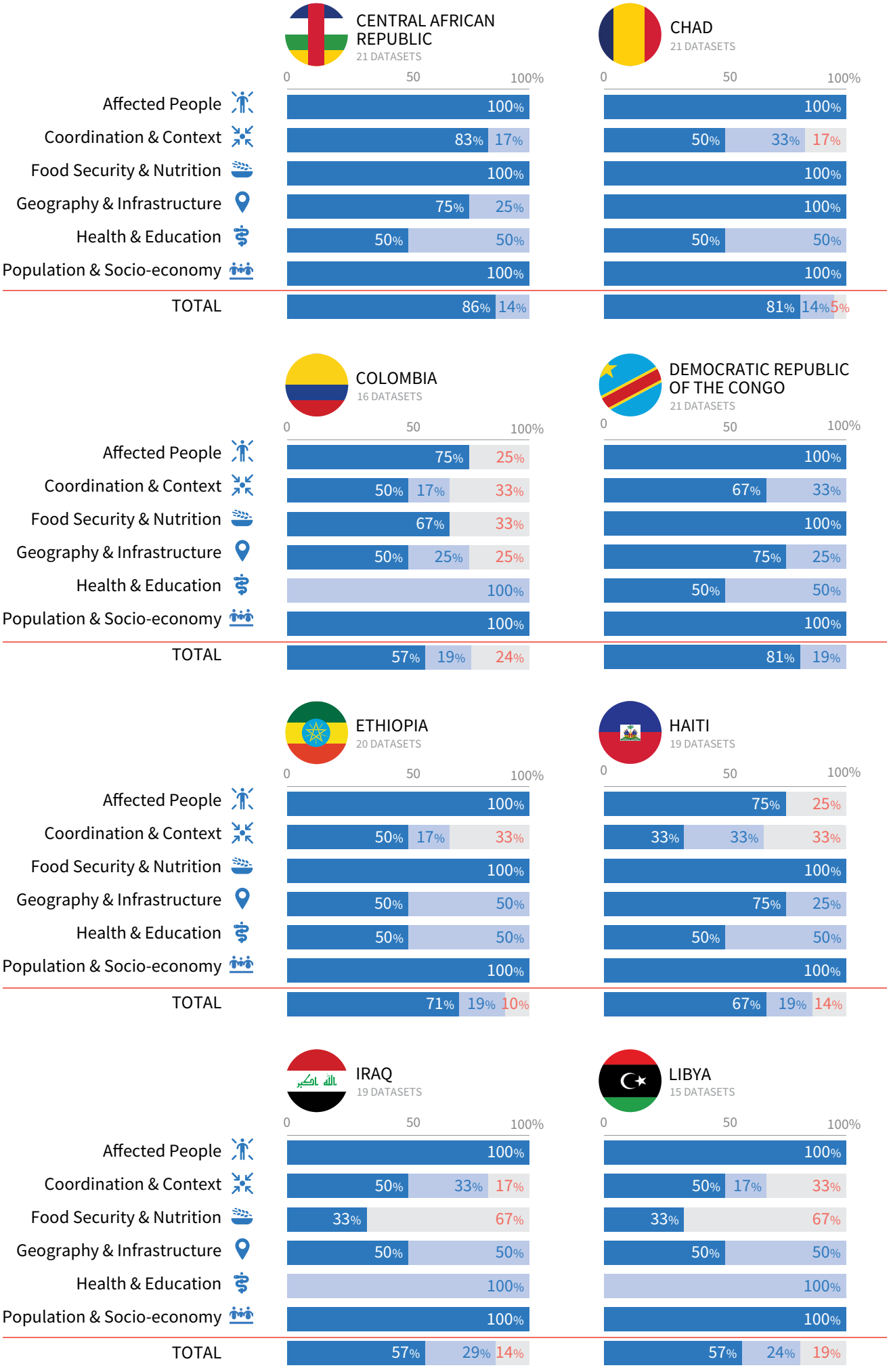
Afghanistan and CAR share the highest degree of data completeness at 86 percent. While both locations have incomplete sub-categories, Afghanistan has no data for transportation status in the coordination and context category.

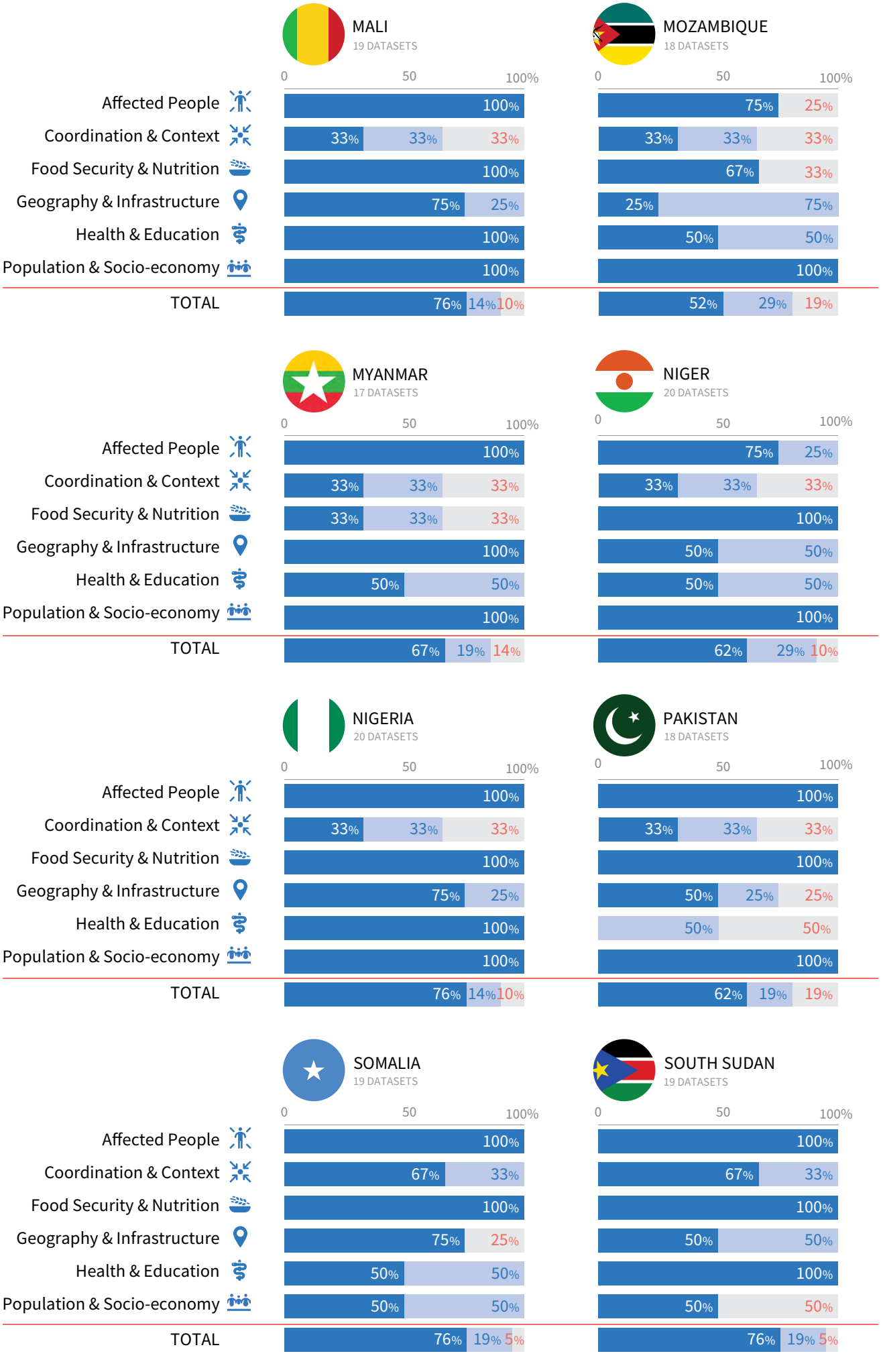
The next most-complete locations for data are Chad (81 percent), DRC (81 percent) and Sudan (81 percent), all of which saw improvements from previous years. Venezuela (43 percent of the sub-categories with no data), Colombia (24 percent of the sub-categories with no data) and Libya (19 percent of the sub-categories with no data) are the countries with the largest data gaps.

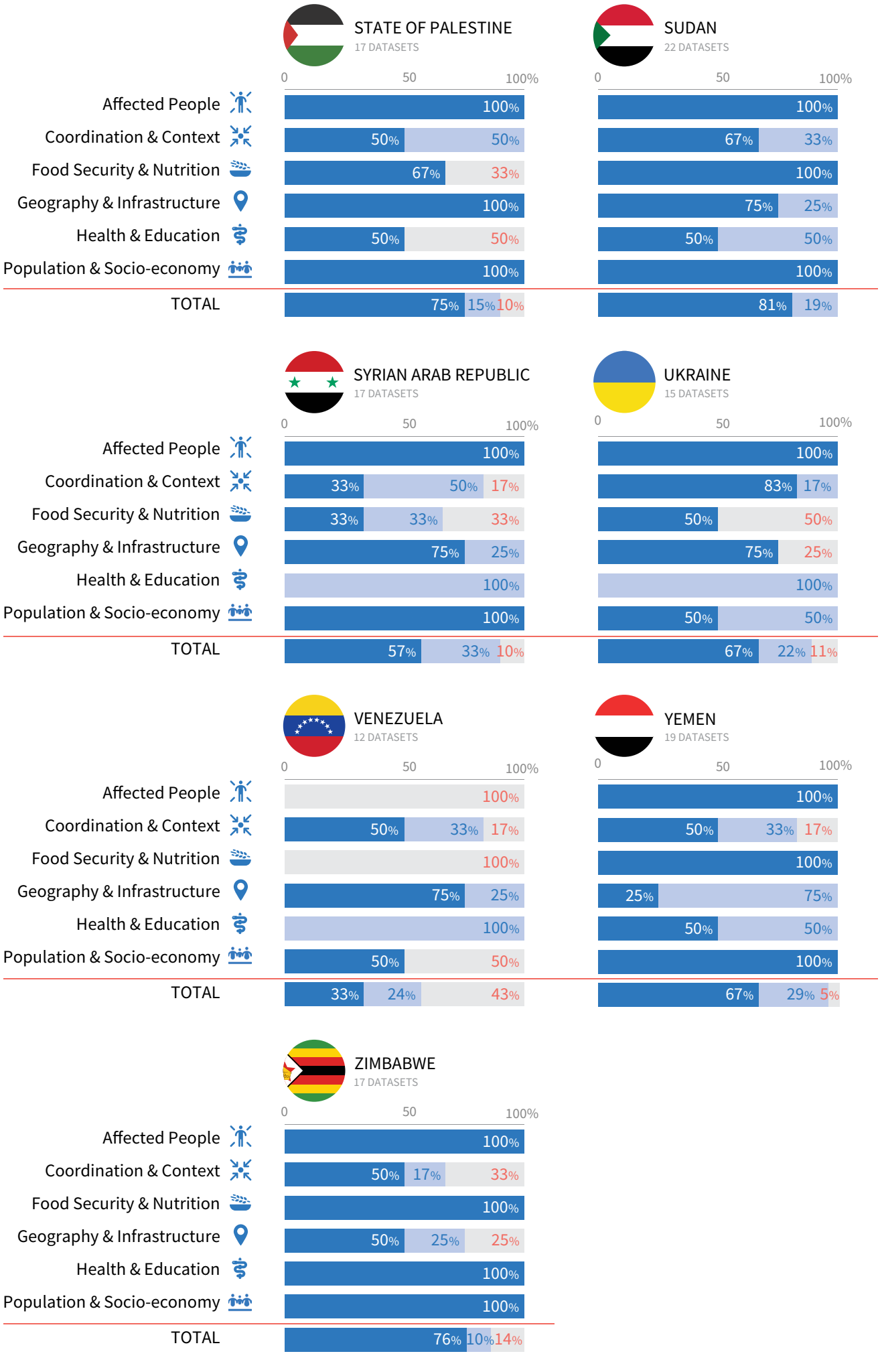
The most complete categories are affected people (91 percent), population and socio-economy (90 percent) and food security and nutrition (81 percent). Health and education is the least complete category at 46 percent, owing to incomplete data for the location of health and education facilities. The most popular category, in terms of total dataset downloads, is geography and infrastructure.

A total of 482 unique datasets are included on the Data Grids. With 21 sub-categories, we would expect a complete Data Grid to include between 21-28 datasets. For the 27 locations covered in this report, the Data Grids include between 12 to 22 datasets.




CENTRAL AFRICAN REPUBLIC
21 DATASETS
CHAD
21 DATASETS
COLOMBIA
16 DATASETS
DEMOCRATIC REPUBLIC OF THE CONGO
21 DATASETS
ETHIOPIA
20 DATASETS
HAITI
19 DATASETS
IRAQ
19 DATASETS
LIBYA
15 DATASETS
0
50
100%
Affected People
Coordination & Context
Food Security & Nutrition
Geography & Infrastructure
Health & Education
Population & Socio-economy
TOTAL
100%
83% 17%
50% 33% 17%
75% 25%
50% 50%
86% 14%
81% 14% 5%
75% 25%
50% 17% 33%
67% 33%
50% 25% 25%
57% 19% 24%
67% 33%
75% 25%
81% 19%
75% 25%
33% 33% 33%
71% 19% 10%
67% 19% 14%
50% 33% 17%
33% 67%
57% 29% 14%
50% 17% 33%
57% 24% 19%


MALI
19 DATASETS
MOZAMBIQUE
18 DATASETS
MYANMAR
17 DATASETS
NIGER
20 DATASETS
NIGERIA
20 DATASETS
PAKISTAN
18 DATASETS
SOMALIA
19 DATASETS
SOUTH SUDAN
19 DATASETS
0
50
100%
Affected People
Coordination & Context
Food Security & Nutrition
Geography & Infrastructure
Health & Education
Population & Socio-economy
TOTAL
100%
33% 33% 33%
100%
75% 25%
100%
50% 50%
100%
76% 14% 10%
75% 25%
33% 33% 33%
67% 33%
25% 75%
50% 50%
100%
52% 29% 19%
100%
33% 33% 33%
33% 33% 33%
100%
50% 50%
50% 50%
100%
67% 19% 14%
75% 25%
33% 33% 33%
100%
50% 50%
50% 50%
100%
62% 29% 10%
100%
33% 33% 33%
100%
75% 25%
50% 50%
100%
76% 14% 10%
100%
33% 33% 33%
100%
50% 25% 25%
50% 50%
100%
62% 19% 19%
100%
67% 33%
100%
75% 25%
50% 50%
50% 50%
100%
76% 19% 5%
100%
67% 33%
100%
50% 50%
100%
50% 50%
100%
76% 19% 5%


STATE OF PALESTINE
17 DATASETS
SUDAN
22 DATASETS
SYRIAN ARAB REPUBLIC
17 DATASETS
UKRAINE
15 DATASETS
VENEZUELA
12 DATASETS
YEMEN
19 DATASETS
ZIMBABWE
17 DATASETS
0
50
100%
Affected People
Coordination & Context
Food Security & Nutrition
Geography & Infrastructure
Health & Education
Population & Socio-economy
TOTAL
100%
50% 50%
67% 33%
100%
75% 25%
50% 50%
100%
75% 15% 10%
81% 19%
100%
33% 50% 17%
33% 33% 33%
75% 25%
100%
100%
57% 33% 10%
67% 22% 11%
100%
50% 33% 17%
100%
75% 25%
100%
50% 50%
33% 24% 43%
100%
50% 33% 17%
100%
25% 75%
50% 50%
100%
67% 29% 5%
100%
50% 17% 33%
100%
50% 25% 25%
100%
100%
76% 10% 14%

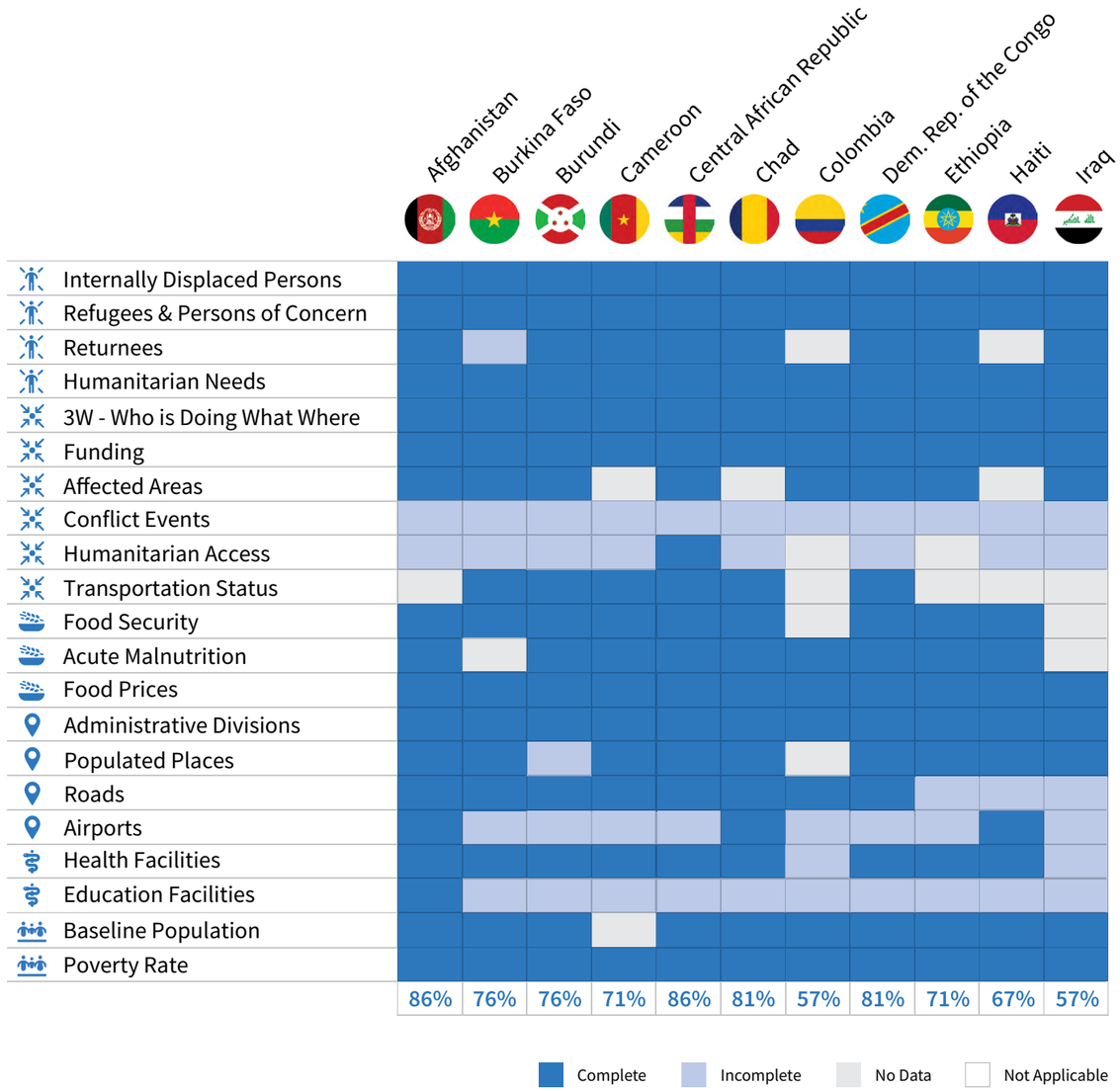
6. COMPLETENESS BY LOCATION AND SUB-CATEGORY

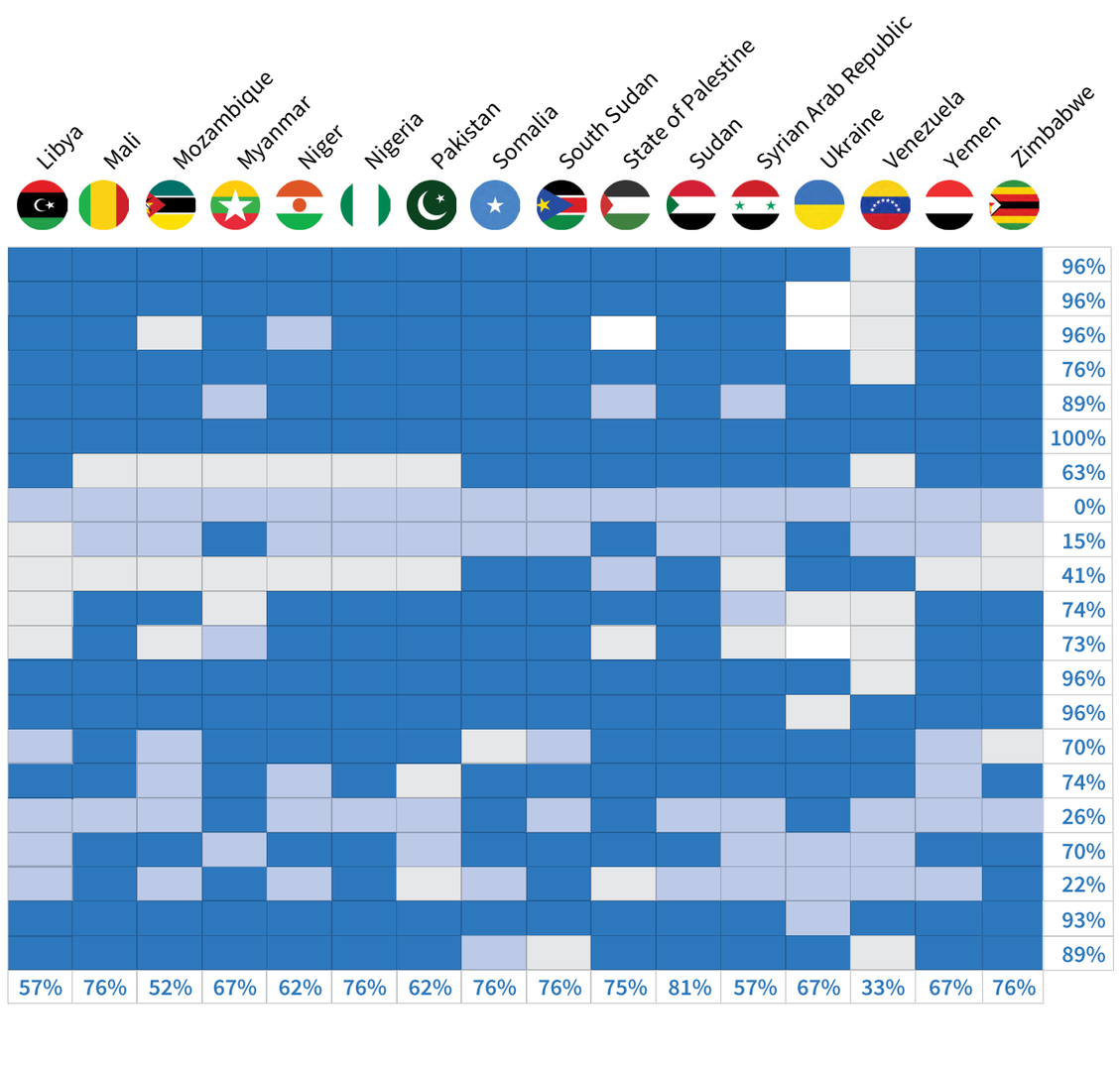
The most complete sub-category is funding, followed by IDPs, refugees and persons of concern, humanitarian needs, food prices, and administrative divisions (all 96 percent complete), owing to the contributions of OCHA, IOM, UNHCR and WFP.

Transportation status (empty in 15 locations) and affected areas (empty in 10 locations) are the most challenging sub-categories to fill. For transportation status, the data should include local transportation routes (e.g., roads, railways, ports) and their current accessibility but the data shared is mostly limited to roads. The affected areas sub-category is filled when there is data on the impact of a rapid onset weather event or on the severity analysis of a Humanitarian Needs Overview.

There was significant improvement in the geography and infrastructure category which moved from 40 percent to 67 percent complete due to better maintained data for administrative divisions and populated places.

This view of the Data Grids also shows where data is consistently (and centrally) shared across all locations but is considered incomplete (e.g., airports, education facilities, humanitarian access and conflict events). This is the case for secondary or crowd-sourced data that may not be the best available or is limited in coverage.





7. COUNTRY DEEP DIVE: SOMALIA

There has been steady progress with the Somalia Data Grid over the past year. Overall data completeness improved from 56 percent to 76 percent, with only one sub-category missing data (populated places). The Somalia Data Grid includes 19 datasets shared by 10 organizations. Together, these datasets were downloaded 6,411 times in 2021. The most popular dataset is the Somalia administrative boundaries.

Somalia has two complete categories: affected people, and food security and nutrition. The Food Security and Nutrition Analysis Unit at FAO newly shared data (via the Somalia Nutrition Cluster) on acute malnutrition by district. Other new additions to this Data Grid are the location of airstrips from the Logistics Cluster and transportation status from WFP. The OCHA Somalia office has played an important role in engaging local partners on data sharing and maintenance, with positive results.

AFFECTED PEOPLE

3 Datasets

Internally Displaced Persons

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

Refugees & Persons of Concern

- The UNHCR Data on forcibly Displaced Populations and Stateless Persons UNHCR - The UN Refugee Agency

Returnees

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

Humanitarian Needs

- Somlia: Humanitarian Needs Overview OCHA Somalia

COORDINATION & CONTEXT

7 Datasets

3W - Who Is Doing What Where

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

Funding

- Somalia - Requirements and Funding Data
  OCHA FTS

Affected Areas

- Somalia Humanitarian Needs Overview OCHA Somalia

Conflict Events

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

Humanitarian Access

- Somalia: Violent and Threatening Incidents Against Healthcare Insecurity Insight

Transportation Status

- Somalia - Transportation Status OCHA Somalia

FOOD SECURITY & NUTRITION

3 Datasets

Food Security

- Somalia: Acute Food Insecurity Country Data
  Integrated Food Security Phase Classification

Acute Malnutrition

- Somalia - Acute Malnutrition Burden and Prevalence
  Somalia Nutrition Cluster

Food Prices

- Somalia - Food Prices World Food Programme

GEOGRAPHY & INFRASTRUCTURE

3 Datasets

Administrative Divisions

- Somalia - Subnational Administrative Boundaries
  OCHA Somalia

Populated Places

No Data

Roads

- Somalia - Roads OCHA Somalia

Airports

- Somalia - Airstrip OCHA Somalia

HEALTH & EDUCATION

2 Datasets

Health Facilities

- Somalia - Health Facilities OCHA Somalia

Education Facilities

- Somalia Schools and Enrollment OCHA Somalia

POPULATON & SOCIO-ECONOMY

2 Datasets

Baseline Population

- Somalia - Subnational Population Statistics
  OCHA Somalia

Poverty Rate

- Somalia - Population Living in Poverty
  OCHA Somalia

8. ORGANIZATION DEEP DIVE: UNITED NATIONS POPULATION FUND

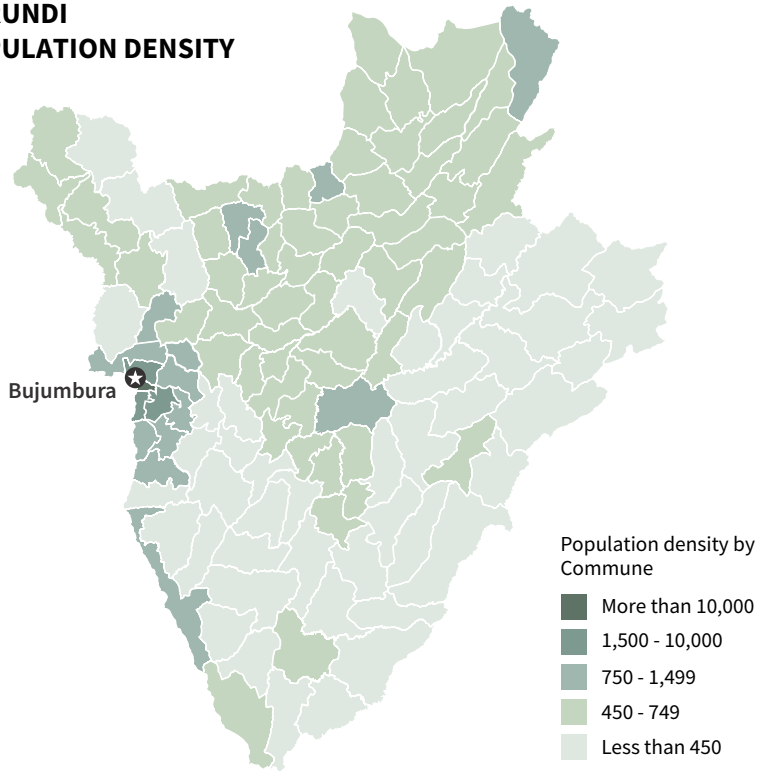
Population statistics are the starting point for estimating the number of people affected by a humanitarian crisis. As part of its reproductive health mission, UNFPA works to strengthen national population data systems. Over the last several years, UNFPA and OCHA have partnered to ensure that sex and age disaggregated population data is accessible to the humanitarian community, and we can see the progress by looking at HDX.

The data collaboration with UNFPA is focused on countries experiencing or at risk of a humanitarian crisis. For these locations, OCHA oversees the common operational datasets (CODs), which are authoritative reference data used to support operations and decision-making in a humanitarian emergency.[12] The CODs include administrative boundaries, which are part of the geography and infrastructure category of the HDX Data Grids, and population statistics, which are part of the population and socio-economy category.[13]

In 2021, UNFPA became the lead agency for the COD population statistics datasets. They combine demographic expertise with working relationships with national statistical agencies to identify, validate, and if necessary project the best available statistics. These datasets are given a 'reference year' referring to the year for which the data was originally collected (normally a census year) or to which it has been projected.

UNFPA began sharing COD population statistics on HDX in late 2020.[14] Over the past year, 50 datasets have been added covering 49 countries. Five of these datasets are included in the Data Grids for Burundi, Chad, Myanmar, Nigeria and Zimbabwe. The rest of UNFPA's current population datasets on HDX are for non-HRP locations that may be at risk of a humanitarian crisis. In 2022, UNFPA plans to update population datasets for sudden onset crises (as they did for Tonga following the recent volcanic eruption[15]) and will continue projecting the oldest census data to the current year.

BURUNDI
POPULATION DENSITY



All five population statistics datasets included in the Data Grids are sex and age disaggregated. This granularity is essential for conducting needs assessments and for developing a humanitarian response plan.

Overall, the UNFPA data on HDX was downloaded over 7,000 times in 2021. The Nigeria population statistics dataset was the most popular, followed by Burundi and Zimbabwe, both of which were recently updated with projections for 2021. The datasets have been designed to join easily with the corresponding COD administrative boundaries, using established P-codes.

Sub-national population statistics from UNFPA
Sub-national administrative boundaries by IGEBU and OCHA Burundi

The boundaries and names shown and the designations used on this map do not imply official endorsement or acceptance by the United Nations.

[12] See the IASC Guidelines for Common Operational Datasets in Disaster Preparedness and Response (2010) for more details. The CODs are managed at the global level by OCHA's Field Information Services Section and at the field level by OCHA Information Management Officers in consultation with Information Management Working Groups.

[13] CODs also include humanitarian profile datasets, which are established after the onset of a crisis and include the location of displaced people, and country-specific CODs which have local relevance.

[14] **https://data.humdata.org/organization/unfpa**

[15] **https://data.humdata.org/dataset/cod-ps-ton**

9. CONTRIBUTING ORGANIZATIONS

Out of the 298 organizations sharing data on HDX, 20 contribute data that is included in the Data Grids.[16] This is a reduction in the number of organizations from past years, mainly due to a decrease in the number of Data Grid sub-categories (from 27 to 21). An organization's dataset may also be removed from the Data Grids if it is no longer timely or has been superseded by a more relevant dataset. The 20 contributing organizations are sharing data from about 125 sources.

OCHA is counted as a single organization even though 31 field offices[17] and two entities at Headquarters contribute data to HDX. In total, OCHA field offices contributed 185 datasets or 38 percent of all data in the Data Grids. For each HRP location, the OCHA office almost always contributes the top downloaded data. For total downloads across all 27 locations, the data from OCHA's Financial Tracking Service (FTS) and IOM are the most popular.

Since 2015, the Armed Conflict Location & Event Data Project (ACLED) has shared important data on conflict events for almost 200 locations around the world. In 2021, ACLED changed the frequency of the data they make available on HDX from daily to weekly. The data now has the total number of reported political violence, civilian-targeting, and demonstration events per week, per country. ACLED makes the real-time datasets available through a registration process on their website.

Armed Conflict Location & Event Data Project

Food Security and Nutrition Working Group, West and Central Africa

Global Healthsites Mapping Project

Humanitarian OpenStreetMap Team

Insecurity Insight

Integrated Food Security Phase Classification

International Aid Transparency Initiative

International Organization for Migration

Myanmar Information Management Unit

United Nations Office for the Coordination of Humanitarian Affairs

OpenStreetMap Democratic Republic of the Congo

OurAirports

Oxford Poverty & Human Development Initiative

REACH Initiative

Somalia Nutrition Cluster

United Nations Population Fund

United Nations High Commissioner for Refugees

United Nations Children's Fund Democratic Republic of the Congo

World Food Programme

World Health Organization

[16] 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 and shares data on humanitarian needs but the data is collected by multiple partners.

[17] **https://www.unocha.org/where-we-work/ocha-presence**

10. DATA FOR MODELLING: ANTICIPATORY ACTION

Anticipatory action enables humanitarian organizations to provide assistance before a shock occurs, reducing the impact of a crisis on the lives of affected people. Since early 2020, OCHA has been testing the practicalities of anticipatory action through pilots in 12 countries.[18] Each pilot has an anticipatory action framework with three elements: a forecast-based trigger mechanism embedded in a decision-making process (the model); a pre-agreed action plan (the delivery); and pre-arranged finance (the money, in this case from the UN Central Emergency Response Fund).

The Centre's predictive analytics team has been supporting this work through the development of the trigger mechanisms. A trigger mechanism translates the characteristics of a hazard or a shock and its impact (food insecurity or damaged housing) into technical specifications.[19] These specifications are the basis for the action plan's preemptive response.

Triggers can include both observational data, such as water levels, and predictive elements, such as flood forecasts. A trigger is designed to provide partners with a window of opportunity to take action while balancing the uncertainty of the forecasts.

Designing a trigger requires access to three types of data:

1. Current and historical data about the hazard;
2. Data about the historical and expected impact of the hazard and shock; and
3. Hazard forecast data.

Despite progress in filling data gaps in 2021, the availability and usability of this type of data remains a major barrier to implementing anticipatory action in many humanitarian contexts. There has been sufficient data for the existing OCHA pilots but expanding anticipatory action to new HRP locations will be challenging without improvements to the data landscape.

Below we provide a closer look at the data needed for developing triggers based on our experience with several countries, including Ethiopia, Malawi, Nepal and the Philippines. While we touch on the DRC and Mozambique, the work to anticipate cholera is ongoing so we have not included these countries in the data availability table.

1. Hazard Data

Current and historical data about a hazard is required to model the likelihood of the event occurring. This data includes weather and climate observations and disease surveillance, as well as lists of recorded historical events, which are often incomplete or hard to access.

In Ethiopia, for example, despite the number of actors producing early warning analysis for drought in the Horn of Africa, there is no consensus on the exact definition of what should be considered a drought. As a result, there was no agreed-upon reference list of historical drought years, which was needed to establish a trigger threshold and assess its performance.[20]

Similarly, in Malawi, there was no reference dataset to understand the frequency and patterns of dry spells affecting the southern region of the country, primarily due to a lack of consensus around the definition of the shock. To fill this gap, we have analyzed and made available on HDX[21] over 20 years of historical sub-national rainfall data to better understand when and where dry spells took place.

[18] The 12 pilots are: Bangladesh (floods), Burkina Faso (drought), Chad (drought), DRC (cholera), Ethiopia (drought), Madagascar (plague), Malawi (dry spells), Mozambique (cholera), Nepal (floods), Niger (drought), Somalia (drought), and The Philippines (typhoons).

[19] In this report we refer to 'hazard' as the potential source of harm (e.g., a flood) which generates a 'shock' (e.g., water level above a certain threshold) that has an impact on vulnerable people (e.g., crop losses or housing destruction).

[20] To evaluate the trigger we used historical rainfall data and food security data. The Ethiopia trigger has been independently evaluated by the Red Cross Red Crescent Climate Centre **https://drive.google.com/file/d/1gR1NrqO9v0ktvTiuwC3seSqPOfjXdxUC/view**

[21] **https://data.humdata.org/dataset/historical-dry-spells-in-malawi**

For anticipating cholera in the DRC and Mozambique, it was challenging to find datasets listing all past outbreaks. When the data was available, it was provided in multiple PDF reports, requiring significant effort to make it usable for analysis.

2. Impact Data

Data linking previous shocks to their impact on vulnerable populations is essential for the design of an anticipatory action trigger. With this data in hand, it is possible to model the expected impact of a future shock.

In the Philippines, historical data on housing damage from previous typhoons was foundational for developing an impact model based on forecasted typhoon tracks. In Nepal, the list of historical floods was estimated using data on water levels but it was difficult to estimate the severity of these events (i.e., the number of people affected or the houses destroyed by the flood). In Malawi, OCHA conducted a survey and interviews to better understand the impact of past dry spells.

Access to impact data is one of the main limitations in implementing anticipatory action in countries with humanitarian operations. There are consistent gaps in data on the impact of hazards on vulnerable populations. Even when this data is available, it is unlikely to be machine readable, sub-national and/or include relevant geographic coverage.

These findings correspond with the completeness levels of the affected areas sub-category in the Data Grids, where 10 out of the 27 HRP locations are missing data. Where this data does exist, it may only refer to a recent shock (i.e., a flood in South Sudan[22]) or be part of a broader severity analysis conducted for the Humanitarian Needs Overview process.

3. Forecast Data

Current forecast data is provided globally, with projections for riverine floods, seasonal rainfall forecasts, or major storms; or at the regional and national level, with projections for specific countries or areas. In addition, historical forecasts are used to better understand model performance and the lead time that these models could provide as part of an anticipatory action plan.

Accessing forecast data in a machine-readable format is a challenge. In the Philippines, we relied on global data providers. In Nepal, national authorities provided qualitative flood early warning information, which had to be extracted and translated from social media and the Government's website. The challenge can be even greater when looking for historical forecast data. It is either not available or has very limited geographic or temporal coverage. For Nepal, we found that national authorities did not keep a record of historical flood forecasts, which was needed to evaluate the forecasting skills and lead time of the model.

Type of Data		Ethiopia (drought)	Malawi (dry spells)	Nepal (floods)	The Philippines (typhoons)
Hazard / Shock	Current and historical data about a hazard or shock			Not machine readable	
Impact	Data about the historical and expected impact of the hazard or shock	Not sub-national	Partial-geographic coverage		Not machine readable
Forecast	Global forecasts				
	National and regional forecasts			Not machine readable	Not machine readable

Data Complete Data Incomplete No Data

[22] https://data.humdata.org/dataset/south-sudan-flood-data

Improving Access to Hazard Data

The anticipatory action pilots have demonstrated the promise of getting ahead of crises before they escalate by using different analytical approaches to predict out-of-the-ordinary events and their projected impact on vulnerable people. This approach is enabled by the growing availability and quality of forecast data as well as the increasing awareness of humanitarian actors of the opportunities it provides.

Access to forecast, observational and impact data about hazards and shocks will be essential to expanding anticipatory action frameworks to more countries with humanitarian operations. To help humanitarian actors better assess the feasibility of anticipatory action in different contexts, the Centre is developing a checklist for trigger mechanisms. This will include data requirements in addition to other factors such as local capacity and monitoring.

Beyond anticipatory action, more robust weather and climate data is needed to understand the near and medium-term effects of the climate crisis on humanitarian operations. Access to relevant data and evidence about the humanitarian impacts of climate change is key to helping extend the annual time horizon humanitarians typically focus on.

To give momentum to these efforts, we will create a new category in the HDX Data Grids that includes the types of data detailed above. We will also work with researchers to fill known data gaps by developing reference lists of historical hazards and their impact for priority countries.

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.[23] As part of our internal Data Grid Governance Group, we will regularly review the categories and sub-categories to see if they should be removed or expanded. A priority for 2022 will be improving access to forecast, observational and impact data about hazards and shocks for anticipatory action. We welcome feedback on possible improvements. Please be in touch with questions or comments at centrehumdata@un.org.

[23] 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 or vector or tabular data with coordinates representing the locations at which displaced people are gathered. 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 Needs Tabular data of the number of people in need of humanitarian assistance by location and humanitarian cluster/sector.
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 describes 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.
Food Security & Nutrition	**Food Security** Vector data representing the IPC phase classification or tabular data representing population or percentage of population by IPC phase and administrative division.
	Acute Malnutrition Tabular data specifying the global acute malnutrition (GAM) rate or severe acute malnutrition (SAM) rate by administrative division.
	Food Prices Time series prices for common food commodities at a set of locations.

CATEGORY	SUB-CATEGORY/DEFINITION
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.).
Population & Socio-economic Indicators	**Population** 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

Data Grid curation involves the careful evaluation of datasets on HDX for inclusion in one or more of the sub-categories in accordance with a standard criteria. The HDX team conducts this evaluation when a dataset is newly added or updated on HDX.

The HDX team supports the Data Grids in three ways:

1. Reviewing whether any dataset on HDX could be added to a Data Grid;
2. Conducting targeted outreach with humanitarian organizations to find relevant datasets; and
3. Identifying potential datasets with partners in the research community, academia and/or government agencies.

The first step in determining whether a dataset should be included in a Data Grid is to check whether the dataset meets this thematic requirement defined in Annex A. Datasets that are not considered relevant are automatically excluded.

The second step is to determine if the dataset is sub-national. To fulfil this requirement the data must be disaggregated to at least the first administrative division. If the data is only available at the national level, it is typically excluded. The sub-national requirement has been partially or completely waived for three datasets: UNHCR's data on forcibly displaced populations and stateless persons; OCHA FTS's requirements and funding data; and IATI activities data. For information about the decision-making process for sub-national waivers, see Annex C.

The third step is to assess whether a thematically relevant, sub-national data set is complete or incomplete. A dataset is considered complete if it satisfies all of the following criteria:

1. Broad geographic coverage;
2. Available in commonly used formats; and
3. Timely (full definitions are given below).

If the dataset does not satisfy one or more of these criteria, it is considered incomplete.

Finally, the dataset is compared against existing datasets for that location to determine if it should be added to the Data Grid. 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. A complementary dataset is one that adds unique and useful information to the sub-category.

The IATI datasets are a good example of a complementary dataset as they add unique and useful information to the 3W sub-category. Another example is a dataset that provides data for a geographic area not covered by the existing dataset. If there is already a complete dataset available for the sub-category, the HDX team will reach out to in-country partners to determine which dataset is being used operationally and that dataset will be included in the Data Grid.

The sub-category is considered complete if it has at least one complete dataset available for the sub-category. If the sub-category contains only incomplete datasets, then that sub-category is considered 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 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 know and trust, or because of a lack of resources to clean and share it);

3. It is collected and shared but is not shared on HDX, or known about by the HDX team.

DETAILED COMPLETENESS CRITERIA

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

Does the data have broad geographic coverage with explicit location information?

- Is the dataset geographically comprehensive, or as comprehensive 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?

- 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.

Is the data in commonly-used formats?

- Is it stored in a common file format? We include CSV, XLS, XLSX, SHP, GEOJSON, etc. Formats like 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 sub-total 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 date? 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.

ANNEX C: CHANGES TO THE DATA GRIDS

A consistent focus of our work over the last year has been soliciting feedback from OCHA field offices and partner organizations to ensure that the data requirements and criteria used to curate the Data Grids reflect data needs in humanitarian operations.

We established an internal Data Grid Governance Group to discuss feedback and improvements and make decisions around changes to the Data Grids. In 2021, the Data Grid Governance Group met three times and approved the following changes:

1. Crowdsourced data may be considered complete on a case-by-case basis.
Previously, all crowdsourced datasets were automatically considered incomplete because this data collection methodology makes it difficult to determine how up-to-date and comprehensive the data is. The Data Grid Governance Group determined that crowdsourced data would be considered complete if an OCHA field office confirms that the data is the best available and is being used operationally.

2. HDX Connect datasets can be included in the Data Grids and considered complete if they have been reviewed by the HDX team.
HDX Connect is a feature that enables organizations to share metadata only and make the underlying data available by request. This ensures that organizations can verify who can access the data in sensitive contexts. The Data Grid Governance Group determined that as long as the HDX team has reviewed the dataset and assessed it against the criteria for inclusion in the Data Grids, that data can be considered complete.

3. Sub-categories were merged or removed.
The Data Grid Governance Group reviewed all sub-categories and the data within them, and agreed to the following changes:

a. The humanitarian profile locations sub-category was merged into the internally displaced persons sub-category and the definition was updated to encompass elements of both.

b. The baseline population by sex and age was merged into the baseline population sub-category and the definition was updated to require sex and age dissaggregation.

c. GAM and SAM were merged into a new acute malnutrition sub-category. This new sub-category can be satisfied with data specifying either GAM or SAM rates.

d. The sub-categories for casualties, affected schools, and damaged and destroyed buildings were removed due to a lack of relevance across all locations. They were therefore not a consistent measure for humanitarian data availability.

4. Waiver of sub-national requirement for inclusion in the Data Grid.
The UNHCR data on forcibly displaced populations and stateless persons was previously excluded from the refugees and persons of concern sub-category because it was not sub-national. We received feedback that because refugee status is based on crossing national borders, refugee statistics are typically available and used at the national rather than sub-national level. The Data Grid Governance Group determined that this data is the best available and granted a waiver to include it as complete for the sub-category. A similar sub-national waiver is in place for the OCHA FTS requirements and funding data, since it is considered the best available, and for the IATI aid activities datasets, which are assessed as incomplete but considered a complimentary dataset to the 3W.

5. Creation of a 'Not Applicable' assessment.
The Data Grid Governance Group agreed to allow 'Not Applicable' to be used when the sub-category is not relevant in a given context. This was applied to the following locations and sub-categories: Ukraine - acute malnutrition; Ukraine - refugees and persons of concern; Ukraine - returnees; and State of Palestine - returnees.