EVALUATING NORWEGIAN CHURCH AID INTERVENTIONS THROUGH MULTIVARIATE ANALYSIS WITH BIG DATA

EMERGING FINDINGS ON HOW TO PAIR DATA FROM CIVIL SOCIETY ORGANISATIONS WITH EXTERNAL DATA FOR BETTER EVALUATIONS

Introduction

This briefing note presents the key findings from a 2020[i] scoping study on the potential of combining Norwegian Church Aid (NCA[ii]) data with external datasets for better evaluations, and what doing that would require.

Through its work, NCA produces a significant amount of data for monitoring and evaluation purposes. This includes geo-located, household-level survey data as well as project locations (e.g. wells and other water supply points, schools). NCA also regularly conducts evaluations of a global nature (multi-country evaluations).

The evaluation type used by NCA depends on the evaluation's overall purpose and questions, the nature of the intervention to be evaluated, the existing evidence and available resources.

One area of interest is to explore the potential of combining datasets – data produced by NCA together with open databases. This includes:

- Traditional data sources (e.g. administrative data, survey data) made accessible with open data protocols, improving the possibilities for finding and reusing data files.
- Geographical data combining points (e.g. geolocated events), polygons (e.g. borders or water shores) and/or raster images (e.g. satellite photographs) that can be combined using powerful geographic information system (GIS) software.
- New data sources of increasing size, frequency and diversity (big data) collected *organically* by sensors (e.g. temperature or traffic measurements) or transaction machines (e.g. scanner or credit card data and mobile calls), with associated technologies for data storage and manipulation.

Evaluators' interest in big data has grown considerably in recent years, reaching a new peak as a result of the Covid-19 pandemic.

Requirements for datasets

In considering combining its data with open datasets, NCA has some requirements and preferences.

NCA's requirements include:

- Access to data files is granted with **minimum user access requirements**.
- Open data files are disseminated in **machine-readable formats** by widely available software.
- Data files are **accompanied by documentation** about methodological issues (i.e. metadata describing

variables, definitions and codes), quality parameters (geographical precision, sample size, reference date, etc.) and paradata about the collection process (time of data entry, identification of data entry operators, etc.).
- The data holder ensures **timely updates** of data.

Desirable characteristics include:

- **International coverage** (in the same file or in separate collection exercises), reducing the entry cost of understanding the data properties when using the information for different countries or interventions.
- Data files using **international geostatistical standards**[iii] **to define variables, classifications and breakdowns** and provide more opportunities for **inter-operability** with other files (i.e. comparing, linking and matching different datasets).[iv]
- Data content expressed as **coded variables**, minimising the use of open-ended questions and literal entries. Standard statistical techniques exist for treating numerical and categorical variables, while textual data analysis requires advanced machine learning.

Promising external datasets

Statistical sources (secondary sources)

Microdata from household or business sample surveys

The interoperability of household survey microdata and NCA data files is limited without advanced statistical techniques. However, there is high potential of using *statistical data matching* and *small area estimation* techniques with NCA household survey data, and could enrich NCA data (e.g. on gender-based violence or social cohesion) with other socio-economic variables.

Aggregated statistical data

There are two main limits to detailed NCA data: its representativeness (especially when collected via sampling) and the confidentiality of personal or sensitive information. NCA can use aggregated statistical data as part of needs assessments to identify areas and thematic priorities for interventions. However, such datasets have a limited potential for evaluation, given the difficulty of linking variables from external datasets and the impact of NCA interventions.

Open geocoded data

This might include geostatistical data files, or any other type of content with place-related information.

In addition to basic geographical layers such as administrative borders and locations, geospatial files provide information about natural and human-made environments (land use, bodies of water, obstacles, infrastructure, etc.) and events (e.g. violent outbreaks or natural disasters), including real-time data (e.g. meteorological data). Geospatial files have a high degree of international harmonisation and potential global coverage, since they are usually compiled by international agencies. Geospatial information has the advantage of easy visual interpretation, which facilitates dissemination to non-specialist users. Furthermore, file formats are increasingly usable by non-GIS software.

A special difficulty of geospatial data is the geographical coordinate system. Several standards will have to be mastered before linking NCA data and geospatial data files, as they need to use the same system to avoid representation issues.

Open administrative data

Administrative data might include lists of projects, budgets, etc. This type of data often includes textual variables, such as names, addresses and project titles, which cannot be directly subjected to statistical analysis. Textual data requires the formation of thesauri to carry out semantic analysis, or manual processing based on key word searches.

This means that high-potential administrative data for NCA interventions will need to be carefully selected. For instance, the International Aid Transparency Initiative (IATI)[v] datastore in certain locations seems to be relevant for NCA planning and evaluation, but would require preparation of a thesaurus relating to the thematic priorities (preparing a list of key words related to gender-based violence, water, sanitation and hygiene and peacebuilding, etc.) before IATA and NCA data could be usefully integrated.

Applicability in practice – an example from Somalia

NCA ran some experiments linking NCA datasets with pre-selected external sources to test the utility of the approach and answer evaluative questions, using an NCA gender-based violence (GBV) dataset from Somalia. This dataset comprised 200 individuals in nine internally displaced person (IDP) camps in two districts of Somalia (Garowe and Mogadishu), from 12–17 February 2019. It contains 108 variables, 12 of which are automatically recorded paradata.

NCA aimed to test whether it could be possible to address questions related to relevance (To what extent is NCA intervening in the areas of greatest GBV prevalence?) and coherence (To what extent is NCA's work consistent with the intervention of other actors?). External data from the following sources was used, considering recorded GBV incidents, humanitarian and development projects, and risks factors as proxies (e.g. lighting and distance to water sources):

- GSHHG Distance to Water
- Armed Conflict Location & Event Data (ACLED)
- IATI Datastore
- Global Flood Hazard Frequency and Distribution
- Estimated number of IDPs at sites assessed by CCCM
- CCCM Cluster Somalia Detailed Site Assessment (DSA)

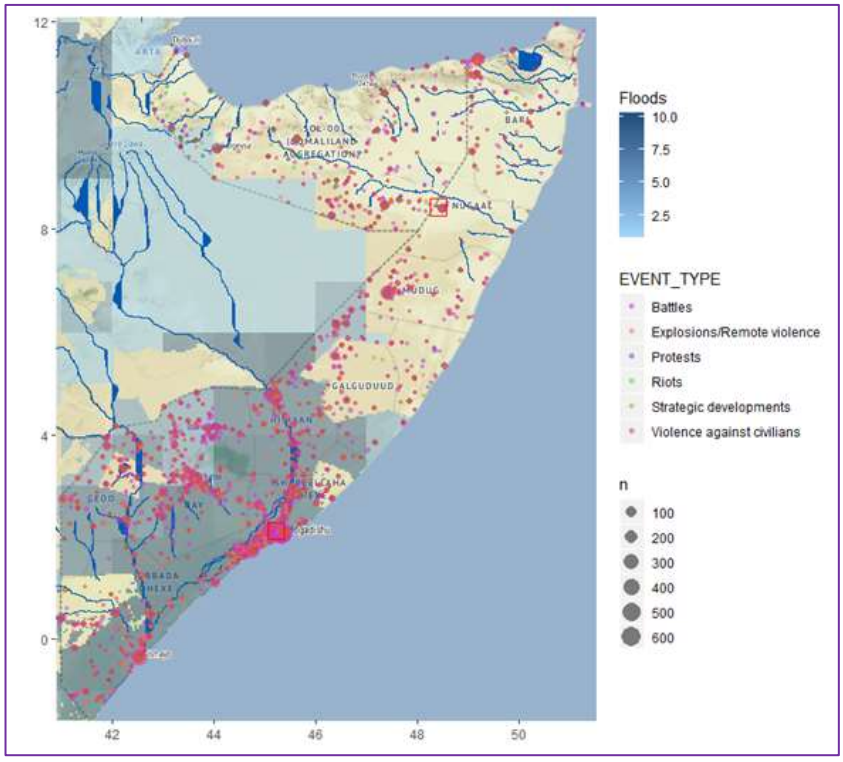


Figure 1 Combined information from ACLED, GSHHG and Global Flood Hazard Frequency and Distribution, Somalia. (NB: Large-scale GIS mapping plotting, e.g. access to safe houses, was also produced but is not included here due to sensitivities)

These analyses allowed to draw some lessons on the evaluation process, opportunities and limitations - and confirmed that NCA GBV interventions captured in that dataset from Somalia are externally relevant and coherent.

Some conclusions

The main conclusion from NCA's scoping study is clear: it is indeed possible for NCA to maximise the use of its internal data in evaluations by pairing or matching it with openly available datasets. This presents a wide range of possibilities to address OECD DAC evaluation criteria, especially in relation to relevance, coherence and impact.

Advanced GIS and spatial statistics can combine multiple data sources to model "catchment areas" for services, "risk areas" for violence or environmental hazards, and also to produce more complex metrics (e.g. calculating spatial correlation coefficients).

There are, however, a series of prerequisites in order for this to be effective. Properly merging multiple data sources may require a data analysis specialist who can manage huge amounts of data using powerful languages/programs such as R or/and Python, and advanced GIS and spatial statistics. Merging data files is challenging. Even when all sources use standard geographic variables, file formats and the way variables are stored can make it difficult to harmonise all sources. Furthermore, working with certain databases requires skills in the statistical analysis of textual data, when external sources include literal descriptions of interventions.

Evaluators and evaluation managers also need to become familiar with the basics of both the analysis and characteristics of internal and external databases. More importantly, a certain degree of evaluation ingenuity is a must; technical requirements and skillsets are not much help if NCA cannot imagine and craft the right evaluative questions, considering the wealth of data and its full possibilities. Technical and resource challenges in combining datasets can be overcome as long as the creative mindset is there, but it is hard to get something of worth without the right evaluative thinking and analytical framework.

---

[i] The scoping study was designed by Norwegian Church Aid and carried out by DevStat in January–June 2020. Contacts: Javier Fabra-Mata, Senior Advisor for Evaluations and Research, javier.fabra-mata@nca.no; Andrej Viotti, Methods, Evaluation and Learning Unit Team Leader, andrej.viotti@nca.no . Release date: 3 August 2020.

[ii] Norwegian Church Aid is a diaconal organisation mandated by churches and Christian organisations in Norway to work with people around the world to eradicate poverty and injustice. Thematically, NCA's 2020–2030 development work comprises three global programmes and three strategic initiatives

[iii] International geostatistical standards include: geodetic system coordinates, geographical codes and place names; and defined socio-economic and environmental variables based on Sustainable Development Goal indicators, UN Recommendations for Population and Housing Censuses or sector-specific classifications (e.g. goods and services, economic sectors, occupation status, education levels, etc.).

[iv] Interoperable data can easily be reused and processed in different applications, allowing different information systems to work together. Interoperability is a key enabler for the development sector to become more data-driven.

[v] https://iatistandard.org/en/

Evaluating NCA interventions through multivariate analysis with big data

NORWEGIAN CHURCH AID
actalliance

Final Scoping Study Report

Version 1

Submitted on 12th June 2020

devstat

1. Purpose of the document

This project addresses the interest of NCA to explore the **potential of integration of internal data** (from projects) and accessible external sources (Open Data), achieving greater interoperability respond to the needs for advocating for, planning, monitoring, and evaluating interventions.

The project implementation is structured in 4 steps:

Step 0: Preparation of the Inception Report (approved on 29th April 2020).

Step 1: Development of a protocol for assessment of potential of data sources delivered on 28th April 2020. It includes a template for the assessment of access and format aspects, statistical aspects, including

- Aspects related to the nature of data collection (types of statistical sources)
- Aspects related to the process of data collection (paradata)
- Aspects related to the data contents

as well as usability aspects, described in terms of the capacity of the file to answer (by complementing internal NCA data sources) evaluation questions according to the DAC criteria (relevance, coherence, effectiveness, efficiency, impact and sustainability). It also includes the protocol (sequence of tasks) to carry out the assessment.

Step 2: Assessment of selected sources (delivered on 29th May**).** Based on the protocol established in Step 1, we assessed 11 data files from selected data sources identified in the Inception Report.

Step 3: Assessment of a particular NCA data set (Diagnosis study) regarding the microdata from the Somalia Household Survey of attitudes, behaviours and experiences around GBV in NCA's catchment areas, 2019.

Step 4: Conclusions and recommendations (this document) consolidates the findings in the **Final Scoping Study report.**

This document is structured as follows:

Section 1 (this section) recalls the objectives and phases of the project implementation.

Section 2 identifies the common characteristics of open data sets, then discusses some specific aspects based on the nature of data files.

Section 3 summarises our view on the potential impact of using external data files in NCA processes, without an assessment of NCA's internal capacities which was out of the scope of this project.

Section 4 describes several software tools used in our Scoping Study, which may prove useful for NCA staff for further statistical work and provides ex post estimates of the workload to process the external files selected.

Section 5 propose a roadmap for NCA strategy for data integration.

Annexes reproduce technical information produced in previous reports, for easier reference.

2. Identification of Open Data sets with high potential of interoperability with NCA interventions' data

Development and humanitarian actors can benefit from an extended landscape of **data** and **technology** for decision-making, including planning, monitoring and evaluating interventions (Letouzé, 2019). This includes:

- Traditional data sources (e.g. administrative data, survey data) made accessible with Open Data protocols, improving the possibilities for finding and re-using data files.
- Geographical data combining points (e.g. geolocated events), polygons (e.g. borders, water shores), raster images (e.g. satellite photographs), that can be combined using powerful Geographic Information (GIS) software.
- New data sources, of increasing size, frequency and diversity (Big Data) collected *organically* by sensors (e.g. temperature, traffic measurements), transaction machines (e.g. scanner, credit card data, mobile calls), with associated technologies for storing and manipulation.

Becoming familiar with the nature of data files, the technologies for manipulating them, and the statistical techniques to analyse the information, highly increases the potential of external information to improve the NCA analytical capacity.

This Scoping Study has reviewed diverse data files with different potential to contribute to NCA, using a standard assessment protocol which included a template for documentation (which can be re-used by NCA for other datasets), presented in **Annex 1**. This section presents the different types and requirements related to the data sets.

2.1. Common characteristics of high-potential data sets

We can highlight several desirable characteristics of data files with high potential:

- Access to data files is granted with **minimum user access requirements** such as commitment of users to *credit where credit is due*. Some interesting data files may require registering for research purpose, which could easily be justified by NCA. Data from non-profits, international organisations and national governmental agencies are easier to access than those collected by private entities (e.g. mobile phone operators, payment portals).

Open data files (see

- Box 1) disseminated in **machine-readable formats** (.csv, .xls, .shp, etc.) by widespread software. Proprietary formats from statistical software (.sps, .sas, etc.) require licensed software in general.

Box 1. Definition of Open Data.

Open Data is an expression for the dissemination of data of any type (e.g. cartographic, statistical, administrative) based on free access, facilitating the re-use of data by providing them in an adequate format. Open Data refer to the idea that certain data should be freely available for re-use for purposes foreseen or not foreseen by the original creator. By providing **easy and free access to data**, it is possible to unleash their potential and help fostering the transparency and the accountability the institutions. The international standard for Open Data is set by the "Open Data Handbook" (http://opendatahandbook.org/). A rating system ("5 stars") for open data proposed by Tim Berners-Lee, founder of the World Wide Web is the following: to score the maximum five stars, data must (1) be available on the Web under an open licence, (2) be in the form of structured data, (3) be in a non-proprietary file format, (4) use URIs as its identifiers (see also RDF), (5) include links to other data sources. To score 3 stars, it must satisfy all of (1)-(3), etc.

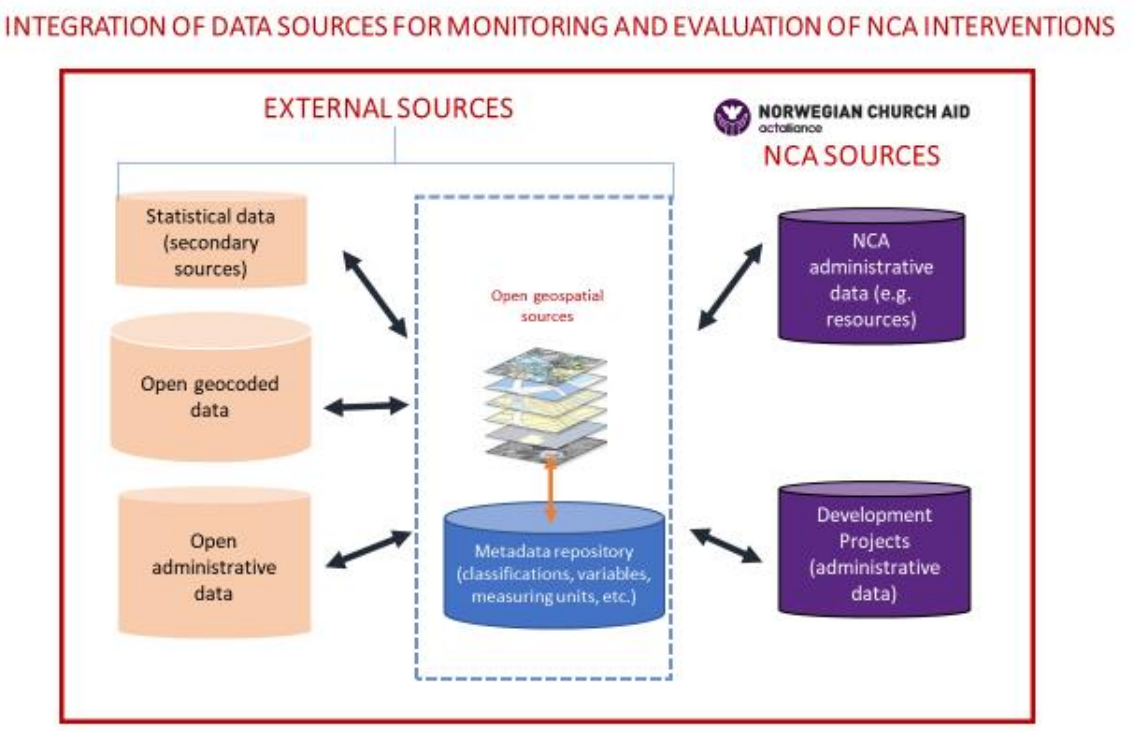
- Data files **accompanied by documentation** about methodological issues (i.e. metadata describing variables, definitions, codes), quality parameters (geographical precision, sample size, reference date, etc.) and paradata about the collection process (time of data entry, identification of data entry operators, etc.). Data curated by international organisations are often documented according to standards for metadata.
- **International coverage** (in the same file or in separate collection exercises) decreasing the *entry cost* of understanding the data properties when using the information for different countries/interventions. Data files originating in internationally coordinated survey programs, global satellite observations, and in general, compiled under sponsorship of international organisations are more likely to have an international coverage.
- **Timely update** of contents is ensured by the data holder. While geographical features (borders, infrastructure, climate) are often stable, political, social, economic and environmental conditions can quickly change in the form of violent outbreaks, social turmoil, natural hazards (floods, fires, extreme events) which have to be recorded in order to keep the information relevant for NCA interventions.
- Data files using **international geostatistical standards for definition of variables, classifications and breakdowns** provide more opportunities for **inter-operability** with other files, i.e. comparing, linking and matching different sets of data. Data is said to be interoperable when it can be easily re-used and processed in different applications, allowing different information systems to work together. Interoperability is a key enabler for the development sector to become more data driven. International geostatistical standards are for example:
    - geodetic system coordinates, geographical codes and place names
    - definition of socio-economic and environmental variables based on SDG indicators, UN Recommendations for Population and Housing Censuses, sector-specific classifications (e.g. goods and services, economic sectors, occupation status, education levels…)
- Data contents expressed as **coded variables**, minimizing the use of open-ended questions and literal entries are easier to treat. Standard statistical techniques exist for the treatment of numerical and categorical variables, while textual data analysis requires advanced machine learning treatment.

2.2. Specific characteristics of high-potential data sets

One feature of the "data revolution" is the diversity of data sources with regard to purpose, collection method and format. To classify the external sources that NCA may use to enrich its information basis, we will follow the ones depicted in Figure 1. These are:

- Statistical sources (secondary sources):
  - Microdata from household or business sample surveys
  - Aggregated statistical data
- Open geocoded data, such as geostatistical data files, or any other type of contents with a place-related information.
- Open administrative data, such as lists of projects, budgets, etc.

Figure 1. Conceptual model for integrating external and internal data sources.



2.2.1. Survey microdata

Several international organisations promote global survey programmes to households and economic units, using common methodological tools (questionnaires, definitions, breakdowns, tabulation programs, interview manuals, etc.). Very often, data are provided in open formats for research purposes at the (anonymized) individual record level.

A quick look at the International Household Survey (IHSN) catalog[1] lists thousands of openly accessible data files (upon registration) which can be used by NCA in the planning phase to identify target beneficiaries,

[1] https://catalog.ihsn.org/index.php/catalog

establish base lines, etc. This is the case of the Multiple Indicator Cluster Surveys (MICS), promoted by UNICEF, the Demographic and Health Surveys (DHS), funded by USAID or agricultural surveys promoted by FAO. Their international coverage and degree of standardization decrease the cost to researchers when using the information for several periods or countries.

An important drawback is the almost impossibility to link the data at individual level due to (1) sample selection and (2) anonymization. Sample selection implies that the data set records cannot be extrapolated to populations for which the sample was not deemed to be representative. For example, extracting from the MICS survey records corresponding to a particular location and extrapolating them to the location of interest may not be possible if the sample design did not consider that location as a stratum for which estimates have to be representative. In addition, the precision of such estimates would likely be very low.

Statistical techniques for *small area estimation* can help provide estimates for such locations, with the use of complementary, exhaustive data (see Box 2).

Box 2. Small area Estimation techniques.

Small Area Estimates (SAE) are statistical techniques to estimate parameters of interest (an average, a total, a proportion, etc.) and adjusting models (e.g. to measure the explanatory power of certain factors on a response variable), for small sub-populations (in geographical terms, or in terms of another characteristic). SAE make intensive use of auxiliary information, and thus, are a good example of how external information can be used to improve the potential of internal data.

SAE methods include: design-based models (only using the internal data source), model-assisted and model-based models (using also auxiliary information). The choice of estimation method is conditioned to the properties of auxiliary information.

An excellent review of methods is given in (Eurostat, 2019).

Techniques for *data matching* creating pseudo-records based on similar records (households or individuals) enhance the potential for combining information (see Box 3).

Box 3. Statistical data matching.

Statistical matching (also known as data fusion, data merging or synthetic matching) is a model-based approach for providing joint statistical information based on variables and indicators collected through two or more sources. The potential benefits of this approach lie in the possibility to enhance the complementary use and analysis of existing data sources (e.g. cross-cutting statistical information that encompasses a broad range of socio economic aspects), without further increasing costs and response burden. However, statistical matching is a complex operation which requires specific technical expertise and raises several methodological issues. Two main approaches can be delineated in terms of outputs that can be obtained through matching:

(1) the *macro approach* refers to the identification of any statistical structure that describes relationships among the variables not jointly observed of the data sets, such as joint distributions, marginal distributions or correlation matrices

(2) the *micro approach* refers to the creation of a complete micro-data file where data on all the variables from different data sets is available for every unit. This is achieved by means of the generation of a new data set ("pseudo-records") from two data sets that are based on an informative set of common variables.

An essential feature of statistical matching is that, although the units in the concerned data sets should come from the same population, they are usually *not overlapping*. You identify and link records from different sources that correspond to similar units. This is the basic difference compared with record linkage, where units included in the data sets overlap that allows to link records from the different data sets that correspond to the same unit. Therefore, record linkage deals with identical units, while statistical matching, or synthetic linkage, deals with 'similar' units.

Source: Eurostat (2013). *Statistical matching: a model-based approach for data integration*[2] (prepared by DevStat).

Interoperability of household survey microdata and NCA data files is limited without advanced statistical techniques. However, the potential of using *statistical data matching* and *small area estimation* techniques with NCA GBV survey data is high, since it would allow enriching GBV-related data with other socio-economic variables.

Given the thematic areas of NCA's interventions, we recommend that NCA research and monitoring staff gets familiar with the MICS and DHS surveys, its questionnaires and variable definitions, and the structure of its datasets. With regard to statistical standards, we recommend that the UN Recommendations for Population and Housing Censuses, which are coherent with sector-specific standards (housing, education, occupation, health status, etc.) are used in NCA's own surveys to increase the possibility of matching internal to external data.

2.2.2. Aggregated statistical data

This is the most common type of available datasets. Repositories with global coverage and a very wide range of topics include those maintained by international organisations. At the national level, most national statistical institutes (NSIs) disseminate online statistical indicators.

Aggregated data can be however very detailed: economic data are available by activities, type of enterprise and location; social data can be broken down by gender, age, education and occupation status, location. The limit for detail is twofold: the representativeness of data (especially when collected via sampling) and the confidentiality of personal or sensitive information.

Aggregated statistical data can be used by NCA to provide baselines, identify areas and thematic priorities for interventions, but have a limited potential for evaluation, given the difficulty of linking the evolution of variables from external datasets and the impact of NCA interventions.

We recommend that NCA staff gets familiar with the production and dissemination of official statistics in countries of interventions. Usually, the production is based on a National Strategy for the Development of Statistics and a multi-annual statistical programme. The dissemination is increasingly made online. It is very important to decrease the response burden of informants, and to avoid duplication of efforts in the data collection. Thus, we recommend that NCA carefully studies the statistical programme of countries of intervention before carrying out additional surveys. An excellent inventory of available statistical data for the development and humanitarian field is (Wright, 2016).

[2] https://ec.europa.eu/eurostat/documents/3888793/5855821/KS-RA-13-020-EN.PDF/477dd541-92ee-4259-95d4-1c42fcf2ef34?version=1.0

2.2.3. Geospatial data

Geospatial data sets provide the place-related background to all kind of NCA internal data. In particular, when collected with CAPI (such as Magpie), geolocating survey data is immediate (subject to correct application of interviewer instructions). Besides basic geographical layers such as administrative borders and locations, geospatial files provide information about natural and man-made environment (land use, water bodies, obstacles, infrastructure, etc.), events (violent outbreaks, natural hazards, etc.) including real-time data (e.g. meteorological data). They have a high degree of international harmonization, since they are usually compiled by international agencies which have the potential of global coverage. Formats are increasingly usable by non-GIS software (for instance, *R* can read .shp files).

Geospatial information has the advantage of easy visual interpretation, which improves the dissemination to non-specialised users.

Some of the general-purpose geospatial data such as those with political/administrative borders or other stable geographical features can be used as common sources for all projects (central block in Figure 1).

There is an increasing integration of statistical and geospatial data, as recommend by UN Statistical Division, UN-GGIM and other UN-system entities (see Box 4). Thus, it is possible to obtain estimates of population density and other socio-economic variables, estimates through models (or in the cases of population censuses by direct enumeration), in gridded maps with very detailed resolution. This kind of information can be used to plan NCA interventions and estimate the size of impacted population and its characteristics.

Box 4. Global Statistical Geospatial Framework.

The UN Economic and Social Council (ECOSOC) on 27 July 2011 recognized the need to promote international cooperation in the field of global geospatial information and therefore set up the Committee of Experts on Global Geospatial Information Management (UNGGIM). UN-GGIM adopted decision 3/107 (see E/C.20/2013/17) at its third session, held in the United Kingdom in July 2013, which "acknowledged the critical importance of integrating geospatial information with statistics and socio-economic data and the development of a geospatial statistical framework". The Global Statistical Geospatial Framework (GSGF) currently being developed under the auspices of the UN-GGIM will provide an integrated and interoperable common method for geospatially enabling statistics and managing geospatial information at all stages of statistical production. It connects spatial information that describes the physical man-made and natural environment, and statistics that describe their socioeconomic and environmental attributes. This framework has already proven useful for the 2030 Agenda for Sustainable Development and the 2020 Round of Population censuses.

Source: (INSEE and Eurostat, 2018)

A special difficulty of geospatial data is the geographical coordinate system. There are several standards which have to be mastered before linking NCA data and geospatial data files, because they should be using the same system to avoid representation issues.

Thus, we recommend that NCA explores different GIS software solutions (e.g. QGIS, gvasig) which are open access software, before investing in commercial software. We also recommend that NCA research and monitoring staff gets familiar with geospatial files such as:

- OCHA data files on administrative borders and settlements
- Environmental information datasets (on water bodies and flood hazards, on land use)
- Socio-economic geolocated datasets (gridded population datasets, ACLED data on violence)

For more advanced spatial analysis, such as estimating spatial correlation (e.g. between GBV events and other violent outbreaks; between economic empowerment of farmers and presence of water infrastructure), it is possible to use existing *R* libraries, in general very well documented. See for instance (INSEE and Eurostat, 2018).

2.2.4. Administrative data

Administrative data are recorded as by-products of administrative activities, such as providing social services, planning budgets, granting activity licenses, etc. or in the case of development /humanitarian projects, recording the activities carried out, registering beneficiaries, etc. Usually, holders of administrative data keep them for internal purposes, but making them accessible with the relevant data protection measures is increasingly promoted under the "Open Government" initiatives[3].

Very often administrative data include textual variables, such as names, addresses, project titles, etc. This type of data cannot be directly subject to statistical analysis. Textual data require the formation of thesauri to carry out semantic analysis, or manual processing based on searches of keywords. Thus, high-potential administrative data for NCA interventions should be carefully selected. For example, the IATI datastore informing of project in certain locations seems to be relevant for NCA planning and evaluation, but would require the preparation in advance of a thesaurus related to the thematic priorities (for example, preparing a list of keywords related to GBV, to WASH, to peace-building, etc.) that would facilitate later the integration of IATI data with NCA's.

2.3. How to identify high-potential data sets

Searching for open data files can be a burdensome activity. Besides the use of catalogues such as (Wright, 2016), we recommend that NCA staff gets familiar with international repositories with standardized documentation about the listed data files.

In particular, we recommend using the following two repositories:

- **International Household Survey Network** (https://catalog.ihsn.org/index.php/catalog), which lists thousands of data files and related metadata. It is maintained by the World Bank and is base don the Accelerated Data Programme (APD) documentation initiative, which can be adapted as well to document internal NCA files.
- **Humanitarian Data Exchange (HDX)** maintained by UN OCHA (https://data.humdata.org/) to share data across crises and organisations, with a data curating process to ensure quality.

[3] See for instance https://www.opengovpartnership.org/.

Some interesting initiatives are worth following by NCA to get aware of new data sources, demonstrative applications and advocacy for "data for development". The following are described in a previous project report:

- UN Global Pulse
- Data-Pop Alliance
- Global Partnership for Sustainable Development Data
- The Open Data for Development Network (OD4D)
- Global Open Data for Agriculture and Nutrition (GODAN).

3. Potential impact on NCA processes of the use of combined information from internal and external sources

Whereas this evaluation has not considered the existing NCA program design, implementation or evaluation processes (which was not in the scope of this project), we elaborate below on the potential impact that combining existing internal and external data sources could have.

The potential of better evidence-based findings: the process of information discovery and knowledge creation is significantly enhanced by the ability to automatically establish meaningful links between independently produced and managed information resources. This is particularly important in the field of data for development, as the indivisible and inter-linked nature of the SDGs makes it more urgent than ever to join-up a vast amount of information resources and data assets independently owned and managed by many different sectors and communities[4].

Data and evidence can **lead to evidence-based interventions** and advocates a more rational, rigorous, and systematic approach. The pursuit of evidence-based interventions is based on the premise that decisions are better informed and include rational analysis and are seen to produce better outcomes. If the scope and reach of data is enhanced through the combination with other open data sources, it will also include the potential of project outcomes. Better utilization of evidence in policy and practice can help save lives, reduce poverty, and improve development performance in developing countries[5].

NCA, like other development agencies, possesses a wealth of project related data. The potential exists to leverage the current program data internally and externally.

Internally, the integration of internal data (from projects) and accessible external sources (Open Data) has the potential across the complete project cycle, with particular emphasis on design, monitoring and evaluation and learning. Through this evaluation we have also learned that greater areas of potential along the DAC criteria include relevance, (external) coherence, and, assuming sustained commitment and capacity, impact and sustainability. The potential consider different data sources can shed light into the combined approach of to the complete NCA portfolio renders possible conclusions and recommendations beyond project level, at an organizational level.

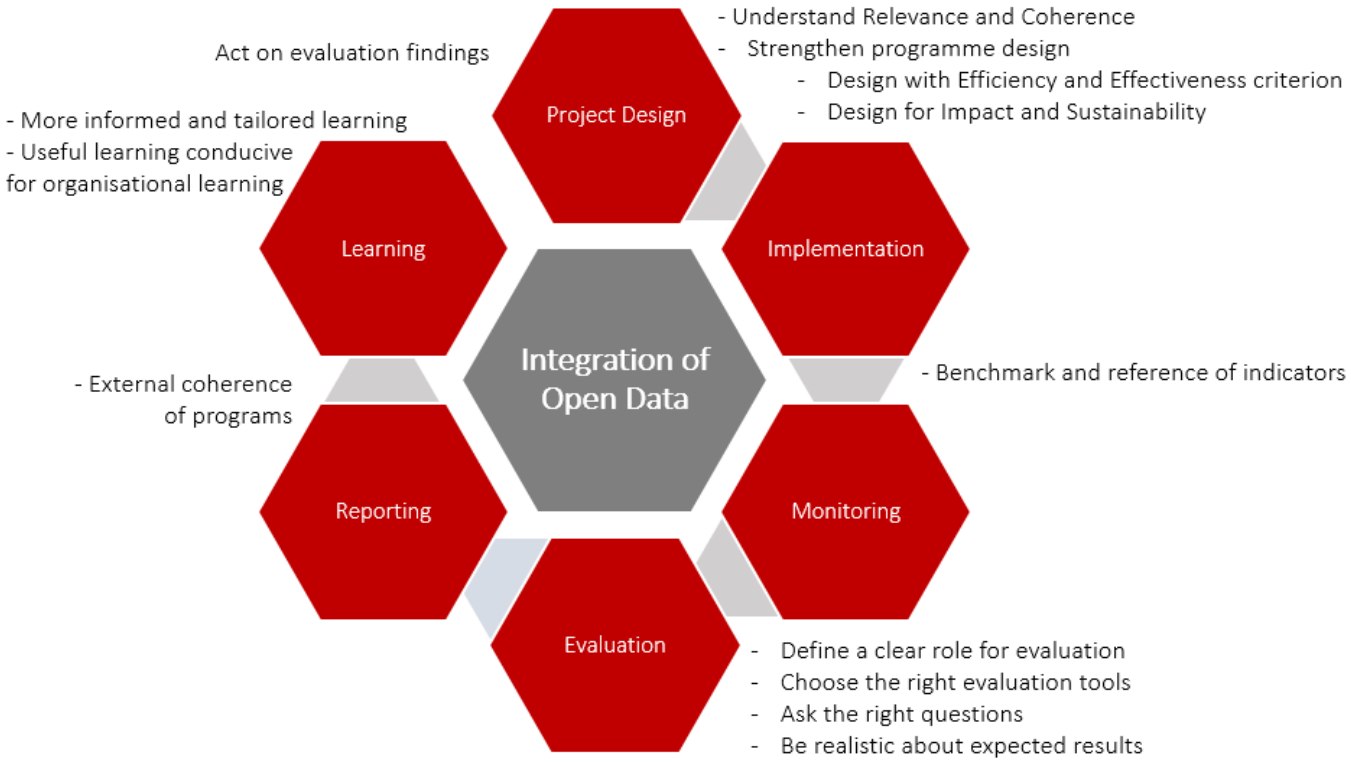
Externally, the combination and integration with other global open data sources will open the opportunity to assess the programs in their global context and more significantly assess aspects of impact and sustainability of interventions along the vision ("Together for a Just World") and long-term goals.

The potential use of external data in NCA processes is exemplified in Figure 2.

[4] https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/3683.pdf

[5] Data Interoperability: A Practitioner's Guide to Joining Up Data in the Development Sector

Figure 2. Potential use of external data in NCA processes.



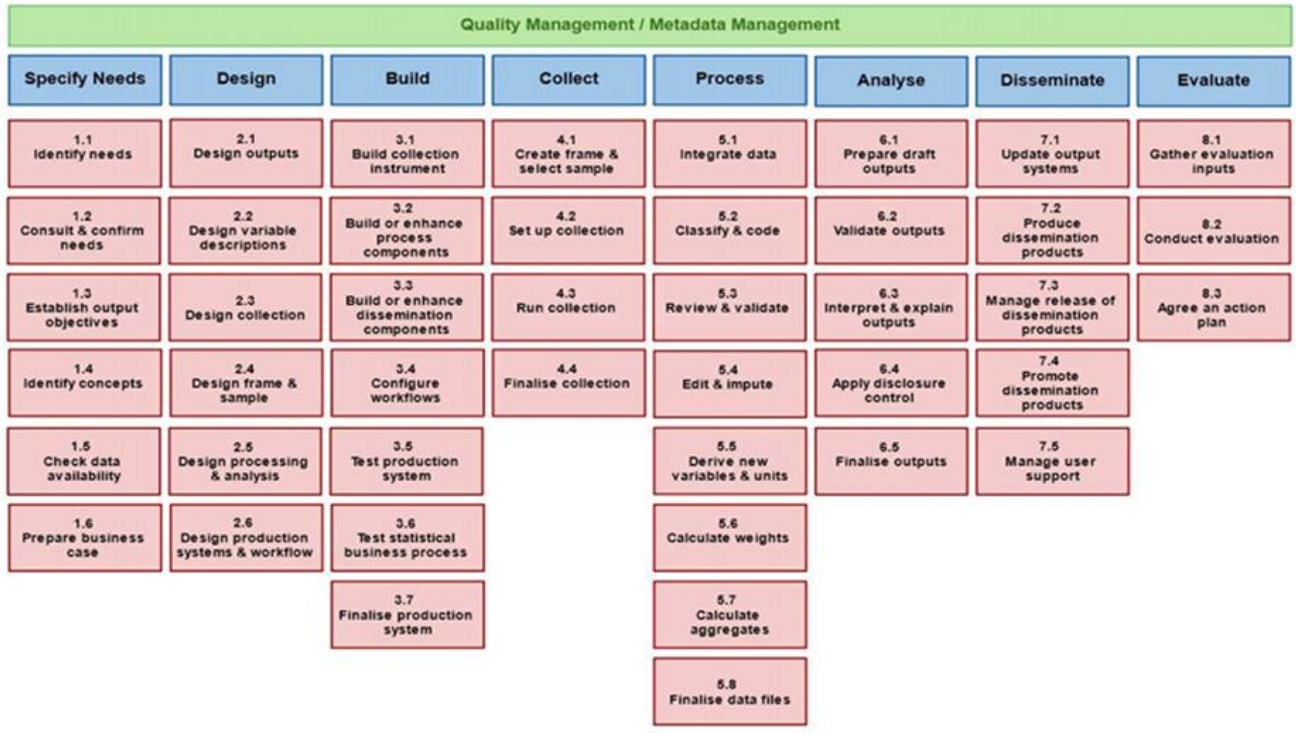
4. Technical requirements for the use of the identified datasets

This section describes technical requirements in terms of data processing activities and knowledge of software. It is recommended that NCA self-assesses its capacities for carrying out independently such activities. Alternatively, NCA can focus on the strategic aspects of *how to use the information* while outsourced collaborators can focus on *how to produce the information.*

4.1. Phases of the statistical process

A useful model for considering all the phases of a statistical operation is given by the UNECE Generic Statistical Business Process Model (GSBPM, see Figure 3). The GSBPM is one of the cornerstones of the High-Level Group for the Modernization of Statistics (HLG-MOS[6]). Its first full version was released in 2009 and has since been adopted by the statistical offices of the most advanced countries in the field. It has proven to be very useful for laying out and describing all the phases to produce statistical information. The GSBPM is intended to guide the planning of surveys and other statistical operations by systematically considering all processes and the workflow from initial preparatory steps to dissemination, documentation and archiving. The model includes preparatory activities starting from the identification of information needs, to final activities such as the dissemination of statistics and evaluation of specifics parts of the process whenever necessary. Most importantly, it allows to create a fully detailed strategy to produce such information.

Figure 3 Scheme of the Generic Statistical Business Process Model (GSBPM)



[6] https://statswiki.unece.org/display/hlgbas

The general business processes identified by the GSBPM are:

- **Specify needs**: used when new statistics are identified or when feedback from current statistics requires a review of them. Its activities are related to precise identification of statistical information needs, preparation of solutions for them and proposals of business cases to meet those needs;
- **Design**: the statistical processes are related to development and design as well as research work to define outputs, methodologies and such. It includes all the design elements needed to define or redefine the metrics that the business case asks for. The metadata and procedures to be used in the following phases are specified at this point.
- **Build**: the outputs from the "Design" processes are assembled and configured in this case to create the complete operational environment to run the process. New services are also created in response to gaps in the existing catalogue of services sourced from within the organisation and externally. These new services are constructed to that they can be reused when necessary or possible.
- **Collect**: gathering all the necessary information and load it to the proper environment for further processing. This process may include validation of data set formats, but never the transformation of data, which is done in the process phase.
- **Process**: processing of input data and preparation of it for analysis. The processing of the data makes it so it can be not only analysed, but also disseminated as statistical outputs. The activities can be parallel to those carried out in the "Analyse" process and may commence before the "Collect" one.
- **Analyse**: statistical outputs are produced and examined in detail. Statistical content for publications, reports, etc. is prepared, and it ensures that outputs are adequate before the dissemination is done. It includes sub-processes and activities that enable statistical analysts to understand the data and the statistics produced.
- **Disseminate**: manages the release of statistical products to users. Activities related with assembling and releasing the products via different channels so that users can access them. This can include presentation of the results of analysis to decision makers.
- **Evaluate**: in this last process, the purpose is to evaluate specific instances in the statistical process. It can be done at the end of it or it can be ongoing during the statistical production process. Once the evaluation of the specific instance is done, a range of qualitative and quantitative inputs is drawn as well as the identification and prioritising of potential improvements.

Applying this methodology to the statistical business processes in the field of development and humanitarian statistics (and other domains) has several benefits that cannot be overlooked. The first one is that the standardisation of terminology creates efficiency savings as well as it makes comparisons internationally much easier. Secondly, its implementation allows the adherence to the standard framework for benchmarking in statistics and hence, it facilitates the use of common tools and methods that again, result in more efficiency savings. Also, the GSBPM includes tools to manage the quality of the process much better, making for better and more reliable data. Finally, it provides a clear and easy way to understand information for data producers and users.

Non-statistical organisations, such as development and humanitarian agencies (like NCA) hardly cover all phases of the statistical process. Their contribution is crucial in the phases of **Specify Needs, Analyse, Disseminate and Evaluate.** However, the level of technical skills required for the intermediary phases **Build, Collect, Process** indicates that collaboration with specialised partners should be considered.

4.2. Data collection

This project has not included primary data collection activities (from households, individuals or businesses). However, the analysis of one exemplary case provides insights about the behaviour of interviewers, showing discrepancies in paradata (duration of interviews, routes followed, etc.). Non-uniform interviewer behaviour may lead to biases. This is best prevented with specialised training to interviewers, via manuals and in some cases, face-to-face training on how to select the sample, how to address respondents, how to use the software, etc. The analysis of the quality of the data collection, when outsources, is key to improving the quality of the inference. This is phase **Evaluate** of the GSBPM.

4.3. Data processing

Data process is the practice of cleaning and transforming raw data prior to analysis. It is an important step often involves reformatting data, making corrections to data and the combining of data sets to enrich the informational basis. Data process is essential as a prerequisite to put data in context in order to turn it into insights and eliminate bias resulting from poor data quality.

There are eight steps in processing data: *integrate data, classify and code, review, validate and edit, impute, derive new variables and statistical units, calculate weights, calculate aggregates* and *finalize data files.*

Integrate data

Integrate data is the process of grouping all input data forms. The input data can be from a mixture of external or internal data sources or variety of collection modes and the result are a harmonized data set. This process may take place at any point in this phase, before or after any of the other seven steps. When the data are integrated, depending on data protection requirements, data may be anonymized.

Classify and code

This process classifies and codes the input data. This can be automated in the process of data collection, if adequate software such as CATI or CAPI tools are used.

Review, validate and edit

In this process looks at each record to try to identify potential problems, errors and discrepancies such as outliers, item non-response and miscoding. In certain cases, imputation may be used as a form of editing. This usually requires statistical processing to detect invalid or extreme values, biases in responses, patterns in non-response.

Impute

The imputation is the process for estimate the data are missing or unreliable often using rule-based approach. The steps for the imputation are the following:

- Identification of potential errors and gaps
- Selection of data to include or exclude from imputation routines
- Imputation using one or more pre-defined methods
- Writing the imputed data back to the data set, and flagging them as imputed
- The production of metadata on the imputation process

Derive new variables and statistical units

Derives variables and statistical units are not explicitly provided in the collection but are needed to deliver the required outputs. Some derived variables may themselves be based on other derived variables. It is therefore important to ensure that variables are derived in the correct order. New statistical units may be derived by aggregating or splitting data for collection units or by various other estimation methods.

Calculate weights

This process creates weights for unit records according to the methodology. These weights can be used to "gross-up" sample survey results to make them representative of the target population, or to adjust for none-response in total enumerations. This is a very specialised skill, which usually conditions the validity of the inference.

Calculate aggregates

Creates aggregate data and population totals from micro-data. It includes summing data for records sharing certain characteristics, determining measures of average and dispersion, and applying weights to sample survey data to derive population totals.

Finalize data files

Finally brings together the results of the other process in this phase and results in a data file, which is used as the input to analyze. Sometimes this data set may be an intermediate rather than a final.

4.3.1. Preparation of internal NCA data

The datasets received for this project are the following:

- Ethiopia: Water infrastructure (wells, pipe, water supply in schools, etc.) constructed with support from NCA, between 2017 and 2019. On NCA WASH in Ethiopia, see for example this (contextual, background information only). The data was collected using Magpi.
- Tanzania: This survey is meant to monitor and capture the performance of project customers. "Customer" refers to farmers who have invested (bought from NCA) 10 dollar in a kit comprising of drip irrigation, small amount of fertilizer and in some areas seeds. Kit(s) allow the customer to establish vegetable beds (1x8m) with high profitability. The intervention also provides customers with insight into (modern) good agricultural practices (GAP).
- Somalia: Household survey of attitudes, behaviours and experiences around GBV in NCA's catchment areas, collected in February 2019. The data was collected using Magpi.

We describe in this section the technical steps that were carried out in order to analyse the information and combine it with external files, to provide insights about the type of skills that may be required.

Ethiopia

This dataset contains information about 118 records, each one of them representing a WASH infrastructure constructed with support from NCA, between 2017 and 2019 and covering two regions in the north (Amhara and Tigray) and two regions in the south (SNNPR, Oromia). 19 variables of each record are collected, 12 of

them automatically, 2 are selectable from fixed categories and 5 are open fields. It includes geolocation of the infrastructure recorded as geographical coordinates. All the variables are easily understandable and well defined.

Although no quality assessment information is available, the availability of metadata (codebook) and of some paradata (record unit, record date, submission record date, gps stamp) can be considered a good point, as it is always important to have information about descriptions of variables, layers, classifications, and to know in detail how data have been collected and entered.

The dataset format file is ".xlsx" with a negligible size of 35KB. This format is easily open by multiple analysis software tools. Since the coordinates are collected by more than one variable depending on if the interviewer was in the place of the infrastructure or not, additional variables are required for its GIS representation. Table 1 shows the suggestions for the dataset processing (this was not carried out as part of the project, being the main interest in demonstrating the possibility of integrating with external files)

Table 1. Ethiopia variables and suggestion for its preparation to be analysed.

#	Variable	Data type	Question type	Missing values	Suggestions
1	Created.By	Text	Automatic	0.0%	
2	Last.Submitter	Text	Automatic	0.0%	
3	Record.Uuid	Text	Automatic	0.0%	
4	Start.Record.Date	Date	Automatic	0.0%	
5	End.Record.Date	Date	Automatic	0.0%	
6	Last.Submission.Date	Date	Automatic	0.0%	
7	gps_stamp_latitude	GPS	Automatic	5.9%	- Check missing values, are they following a pattern?
8	gps_stamp_longitude	GPS	Automatic	5.9%	- Check missing values, are they following a pattern?
9	gps_stamp_accuracy	GPS	Automatic	98.3%	- Check missing values, are they following a pattern?
10	Date_construction_started	Date	Date picker	0.0%	- Check that is a feasible year
11	IRN	Text	Open	0.0%	- Check that the IRF follows the fulfil guide: · 3-letter country code: ETH · programme: WASH · year · number that year - Check that the year coincides with the year of the date of construction - Check the number that year - Check that there is a unique IRF by infrastructure
12	adm1_region	Text	Dropdown	0.0%	- Check that the GPS coordinates are in this region - Check that the categories coincides with the admin 1 regions
13	Zone	Text	Open	0.0%	- Check misspellings - Check that the names of the zones are the standard ones and unify them - Check that the GPS coordinates are in this zone
14	Woreda	Text	Open	0.0%	- Check misspellings - Check that the names of the woreda are the standard ones and unify them - Check that the GPS coordinates are in this woreda
15	Location_type	Text	Dropdown	0.0%	

#	Variable	Data type	Question type	Missing values	Suggestions
16	GPS_coordinates_latitude	GPS	Automatic	1.7%	- Check missing values, are they following a pattern?
17	GPS_coordinates_longitude	GPS	Automatic	1.7%	- Check missing values, are they following a pattern?
18	GPS_coordinates_accuracy	GPS	Automatic	98.3%	- Check missing values, are they following a pattern?
19	GPS_coordinates_typed	Text	Open	96.6%	- Check that the coordinates follows the fulfil guide: · Two digits numbers (decimal separator ".") split by "-" · First latitude, second longitude - Check missing values: · Do they coincide when the CPS coordinates are not missing? - Split this variable in two columns (latitude and longitude) - Compare with coordinates variables to create a unique and final position of the infrastructure

Tanzania

This dataset contains information about 1,180 records, each one corresponding to a farmer benefitting of NCA interventions. This is a weekly voluntary survey by beneficiary's farmers carried out during 2018-2020. 34 variables of each record are collected, 12 of them automatically, 19 are selectable from fixed categories and 3 are open fields. Many dichotomic ("yes/no") variables such as "Mulching done?", "Watering done?", are not referred to a temporal indication and no information about possible international classifications applied is indicate for the closed questions. Using fixed categories from international classifications and define as precisely as possible the variables will help in combining this dataset with external data sources.

Like for the Ethiopia dataset, no quality assessment information is available, however some metadata (codebook) and some paradata (QR code ID, site QR, date and time of creation and date and time of upload, cellphone) are accessible.

The dataset format file is ".csv" with a small size of 551KB. This format is easily open by multiple analysis software tools. The automatic latitude is always missing or with a non-valid value, therefore the Universal Transverse Mercator (UTM) coordinates have to be used for its GIS representation. Depending on the software selected this representation may be not possible (in addition the UTM zone differs between registers), for the linking process it would be better transform UTM coordinates into latitude and longitude.

Table 2 shows the suggestions for the dataset preparation.

Table 2. Tanzania variables and suggestion for its preparation to be analysed.

#	Variable	Data type	Question type	Missings	Comments
1	ec5_uuid	Text	Automatic	0.00%	
2	created_at	Date	Automatic	0.00%	
3	uploaded_at	Date	Automatic	0.00%	
4	title	Text	Automatic	0.00%	
5	X1_YOUR_QR_code_ID	Barcode	Automatic	0.00%	
6	X2_SITE_QR_code_ID	Barcode	Automatic	0.00%	
7	X3_Cellphone	Phone	Open	64.00%	- Check missing values, are they following a pattern?
8	lat_4_GPS_location	GPS	Automatic	9.90%	- Check missing values, are they following a pattern? - The no missing values do not contain numbers
9	long_4_GPS_location	GPS	Automatic	9.90%	- Check missing values, are they following a pattern?
10	accuracy_4_GPS_location	GPS	Automatic	9.90%	- Check missing values, are they following a pattern?
11	UTM_Northing_4_GPS_location	GPS	Automatic	9.90%	- Check missing values, are they following a pattern?
12	UTM_Easting_4_GPS_location	GPS	Automatic	9.90%	- Check missing values, are they following a pattern?
13	UTM_Zone_4_GPS_location	GPS	Automatic	9.90%	- Check missing values, are they following a pattern?
14	X5_Status_of_site	Text	Radio	0.00%	
15	X6_Watering_done	Text	Radio	0.70%	- Check missing values, are they following a pattern?
16	X7_Mulching_done	Text	Radio	0.70%	- Check missing values, are they following a pattern?
17	X8_Weeding_done	Text	Radio	0.70%	- Check missing values, are they following a pattern?
18	X9_Trellising_done	Text	Radio	0.70%	- Check missing values, are they following a pattern?
19	X10_Fertilizer_applie	Text	Radio	0.70%	- Check missing values, are they following a pattern?
20	X11_Spraying_done_las	Text	Radio	0.70%	- Check missing values, are they following a pattern? - Revise "Organically growing" and its coherence with other questions
21	X12_Any_pests_observe	Text	Radio	0.70%	- Check missing values, are they following a pattern?
22	X13_Any_diseases_obse	Text	Radio	0.70%	- Check missing values, are they following a pattern?
23	X14_Any_nutrient_or_m	Text	Radio	0.70%	- Check missing values, are they following a pattern?
24	X15_Equipment_usage	Text	Radio	0.70%	- Check missing values, are they following a pattern?
25	X16_Filter_and_cap_us	Text	Radio	0.70%	- Check missing values, are they following a pattern?
26	X17_Equipment_well_ha	Text	Radio	0.70%	- Check missing values, are they following a pattern?
27	X18_Farmer_keeps_reco	Text	Radio	0.70%	- Check missing values, are they following a pattern?
28	X19_Any_damage_by_ani	Text	Radio	1.00%	- Check missing values, are they following a pattern?
29	X20_Previous_actionsr	Text	Radio	0.70%	- Check missing values, are they following a pattern? - Check if there were previous actions

#	Variable	Data type	Question type	Missings	Comments
30	X21_New_actionsrecomm	Text	Open	0.80%	- Check missing values, are they following a pattern? - Check misspellings - Looking for categories
31	X22_Any_investment_in	Text	Radio	0.70%	- Check missing values, are they following a pattern?
32	X23_Time_spent_in_thi	Text	Dropdown	6.60%	- Check missing values, are they following a pattern?
33	X24_Picture_of_all_be	Photo	File	7.20%	- Check missing values, are they following a pattern?
34	X25_Number_of_new_bed	Integer	Open	94.10%	- Check missing values, are they following a pattern? - Revise with question X22 "investment in new beds"

Somalia

This dataset contains information about 200 records. 108 variables of each record are collected, 12 of them automatically, 76 are selectable from fixed categories and 20 are open fields.

The dataset format file is ".csv" with a small size of 123KB. This format is easily open by multiple analysis software. The data processing was done following the GSBPM steps, see Annex 2 for further details. The requirements for data processing for this file were the following:

- A selection of external open geolocated data sources, preferably with global coverage (see recommendations in STEP 2 report), covering topics of relevance for NCA interventions
- Merging the data files is challenging. Despite of all the sources use standard geographic variables, the format of files and the way these variables were stored made it difficult to harmonize all the sources in a same format. User-friendly GIS software (as for example, QGIS) was used to visualize multiple geographic files and combine them in different layers. As a conclusion, for the proper merging of multiple sources, it may be required a specialist in data analysis who is able to manage huge amounts of data using powerful languages/programs as R or/and Python.
- Basic skills in the use of GIS at least for visualisation purposes
- Advanced GIS and spatial statistics (using statistical software such as R) skills to model and calculate analytical constructs such as "catchment areas" for services, "risk areas" for violence or environmental hazards, etc. as well as to produce more complex studies (e.g. calculation of spatial correlation coefficient).
- Skills in the statistical analysis of textual data, when external sources include literal descriptions of interventions.

4.3.2. Preparation of auxiliary data sources for linkage

The Inception Report proposed the selection of 10-15 files, considering a sample of the different types. After revising initiatives (as part of STEP 0) and carrying out research on repositories of files, we have selected 13 data files, of which some exist for several countries. Table 3 summarises the main characteristics and the technical requirements for the identified databases. The technical requirements have been coded as Low, Medium, High following the below rules:

- **Low**: The file can be opened directly by a user-friendly tool and easily visualized as a map layer. Each row contains a geographic point and some calculations could be done directly from the main variables.
- **Medium**: The file can be opened directly by a user-friendly tool and visualized as a map layer. Group of rows contains geographic shapes and the analysis required some data manipulation.
- **High**: The file cannot be opened directly by a user-friendly tool. Transformations are required to visualize the data.

Table 3. Technical requirement for external data sources

Data source	Format file	Size	Records	Dates	Location	Number of variables	Technical requirement
OCHA	.xls	4,051 KB	10,736	Yes	Region Zone Woreda	24	Low
IATI	.csv .json .xml	9.3 GB	97,6052	Yes	Region Country Sub-national locations GPS coordinates	752 (400 of them with more than 99% of missing values)	High
ACLED (Africa)	.xlsx	41.1 MB	210,085	Yes	Country Admin1 Admin2 Admin3 Location GPS coordinates	29	Low
GSHHG	.b .nc .dbj .prj .shx	113-142 MB (depending on the format file)	2,994,042 lines (binary file)	No	GPS coordinates and shapes	Geospatial layer with different resolutions and 10 levels of water forms	Medium
WFP Obstacles	.cst .dbf .prj .shp .shx	1-1943 KB (depending on the format file)	918	Yes	GPS coordinates Country Iso3	15	Medium
DHS Ethiopia 2016	.dbf .mdb .ascii .stata .spss .sas	118 - 16,200 KB (depending on the format file)	28,371	Yes	GPS coordinates	< 30	-
High-res. Pop	.csv .geotiff	21.6-21.4 MB (depending on the format file)	3,531,151	No	GPS coordinates	3	Easy
OCHA Admin Boundaries	.xlsx .csv .txt				ADM1 ADM2 ADM3 Country Region		Easy
Global Flood Hazard	.asc .prj	648 KB	3,432		GPS coordinates	8,640	Medium

Data source	Format file	Size	Records	Dates	Location	Number of variables	Technical requirement
Agri	.sav	241.6 KB-6.4 MB (depending on the file)	93,562	Yes	Region code Region name Code of district Code of wars Code of village	406	
MICS			33,901	Yes	GPS coordinates State Name State Code		-
CCCM - # of IDPs (Somalia)	.json	47 KB		Yes	GPS coordinates	22	Easy
CCCM - DSA (Somalia)	.xlsx	3,031 KB	1,890	Yes	Region District GPS coordinates	350	Medium

4.4. Software requirements for the use of auxiliary data sources

There are multiple options when you choose a software to analyse de data. Depending of the objective of the analysis and the format file of the dataset it is required a different software. For the presented databases we mainly use three programs Tableau Public, QGIS and R.

Tableau and QGIS required the dataset in multiple formats, but do not allow the manipulation of the data although they learning curve is not really high, whereas R, although required a higher technical knowledge, allows the edition of data and change the format in the required way.

4.4.1. Tableau Public

Tableau Public is a free service that lets publish interactive data visualizations to the web. Visualizations that have been published to Tableau Public can be embedded into web pages and blogs, they can be shared via social media or email, and they can be made available for download to other users. These visualizations are created in the accompanying app Tableau Desktop Public Edition. This app runs under Windows 7 or newer (x64), Microsoft Server 2008 R2 or newer and macOS High Sierra 10.13, macOS Majove 10.14, macOS Catalina 10.15 and IMac/MackBook computer 2009 or newer.

For spatial questions, the use of maps is required to understand the trends or patterns in your data. Tableau Public is a great tool for creating maps and other types of charts. Tableau Public has supported maps since Tableau Public 4.0. With Tableau, you can create the following common map types:

- Proportional symbol maps
- Choropleth maps (filled maps)
- Point distribution maps
- Heatmaps (density maps)
- Flow maps (path maps)
- Spider maps (origin-destination maps)

After you set up your data source, you might need to prepare your geographic data for use in Tableau. Not all of these procedures will always be necessary to create a map view, but it's important information to know when it comes to preparing geographic data for use in Tableau. Depending on the type of map you want to create, you must assign certain data types, data roles, and geographic roles to your fields (or columns).

In addition, Tableau Public is able to create other kind of graphics allowing the visual analysis of other kind of questions. Moreover, it is possible to create a dashboard with a combination of different kinds of graph and tables.

4.4.2. QGIS

QGIS is a user-friendly Open Source Geographic Information System (GIS) licensed under the GNU General Public License. It runs on Linux, Unix, Mac OSX, Windows and Android and supports numerous vector, raster, and database formats and functionalities. Further, QGIS permit you make operations with spatial data such as calculate the distance between two or more points or also you make animations with your data among other things. Finally, QGIS is a program that offer endless of functionalities for spatial files or data bases but have a medium learning curve and do not support may files with a lot information.

4.4.3. R Software

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. R programming language has shown remarkable growth in the last five years becoming in one of the languages most popular in data science.

R Packages are collections of R functions, data, and compiled code. While R comes with a set of packages by default, there are over 10,000 user-created libraries that were built to enhance R functionality. R has a huge and growing number of spatial data packages.

R is not only use for the graphic visualization. The use of R requires a high technical knowledge, however, with R, you are able to manipulate de dataset and make all the transformation needed to the proper visualization or to compute complex statistical models to analyses your data.

Apart from this, one of the great advantages, is the capability of R. R is able to manage huge databases with millions of records whereas that Tableau or QGIS could present some problems with massive data.

Other possibility is to use R to prepare the dataset and save the file is a proper format to be represented with other visualization tool.

4.5. Estimated technical resources

As part of this project, we carried out a detailed assessment of a particular NCA data set, specifically, Somalia GBV NCA data set (Step 3, see Annex 2). This allowed us to explore the technical possibility of linking such dataset with those selected in Step 2. For the technical part, two DevStat in-house specialists with technical skills were involved: a data scientist and a computer technician. To develop this task, four part-time (50%) weeks were required. Half of this time was spent in the data processing of selected external resources. This leads to an estimate of 4 person-weeks.

The big size of some files and the different formats of the geographic information were the main challenge. To be able to link multiple external resources, R programming language was required and more than 20 specialised packages were used. Although the majority of these packages was well known by the team, some specific ones on geospatial analysis had to be studied for the harmonization of geographic information.

5. Suggested roadmap for a strategy for data interoperability, aligned to the strategic mission and vision of NCA and NCA's 2020 – 2030 development

During the course of this scoping study the team has had access to organizational documents including NCA's 2020-2030 Programme Framework and NCA's Global Strategy. Evaluating NCA interventions though multivariate analyses with big data can inform Global Programs and Strategic Initiatives in the pursuit of the long -erm goals. Data interoperability therefore speaks to *how* NCA works, and the roadmap here suggested is aligned and benchmarked to NCA's organisational priorities as outlined in the referred documents.

5.1. Technical activities: Statistical and IT capacity building required for the sustainable exploitation of combined internal and external data

Short-term

- Registering in repositories of microdata for research, establishing partnerships with providers of relevant data sources
- Get basic knowledge about the capacities of specific software (R, Tableau, QGIS)
- Get familiar with the contents and formats of recommended external sources for routine use in the planning, monitoring and evaluation phases

✓Lean and effective ✓Flexible and competent ✓ Technology oriented

✓Accountable ✓Able to manage security and risks

Medium-term:

- Improve the data collection package in intervention countries by adding training materials/sessions for interviewers to increase harmonization in data collection and decrease non-sampling errors

✓Innovative and learning ✓Accountable ✓Able to manage security and risks

5.2. Organisational activities

Short-term:

- Develop an inventory of internal files (using for instance the NADA[7] tool and the template proposed in STEP 1)
- Assess the statistical and IT skills in internal staff at NCA

✓Lean and effective ✓Technology oriented

Medium-term

- Develop competencies in use of evidence (see technical activities above)
- Identify external partners to support data processing activities (sampling design, questionnaire preparation, data processing, data visualization)

✓Lean and effective✓Technology oriented

Able to manage security and risks

Long-term:

- Advocate, with the demonstration of NCA's experience, for using evidence in Faith-Based organisations (develop showcases)

✓A bold voice for justice ✓Cooperating closely with the ACT alliance

[7] https://nada.ihsn.org/

Annex 1: Template for assessment of external data sources

Characteristics	
Access and format	
Source name	
URL, link to data	
Scope/description	
Data provider	
Legal basis for the data collection	
User access rights	
Availability of metadata	Yes…. Provide link / No
Availability of quality assessment	Yes…. Provide link / No
File format(s)	human-readable (download of files in Excel, CSV, etc.) of machine-readable (XML, JSON, API, etc.)
Version	
• Date of check	
• Frequency of updates	
Statistical aspects	
Type of data source	Administrative records/ Survey microdata /Aggregate statistics / Geospatial /Social media / Transaction data (CDR, M2M, etc.) etc.
Statistical unit (level of detail)	Event / Individual /Household /Enterprise /Spatial unit
Data collection mode	Face-to-face interview, CATI, CAPI, CAWI, administrative data transmission, satellite image, etc.
Identification of units	Case/interview ID, Enumeration area unit number, Cluster number, household number, geo-positioning variables, etc.
Time of unit data collection	Day of interview/data collection, Date of interview/data entry, starting time, ending time, duration, Number of visits (for surveys)

Characteristics	
Identification of data entry operator	Interviewer/ data entry operator ID, interviewer/ data entry operator name and contact, etc.
Language	Language(s) of data source, language of interview/data collection process, mother tongue of respondent, translator used (yes/no
Non-response, missing units	Reasons for non-response, codes for non-response or missing values
Geographic coverage	
Time coverage	
Population coverage	
Thematic coverage	
Main variables and breakdowns available	
Geographical variables	Yes /No
(for geospatial files) **Main geographical layers/ features**	Point data (e.g. GPS coordinates), closed polygons (e.g. borders), open polygons (e.g. roads, water courses), raster (e.g. aerial photographs)
Coherence with other sources:	
• definition of main variables	
• classifications used	e.g. ISCED, ISCO or other statistical classifications, administrative unit codes

6. Annex 2: Detailed assessment of Somalia HH data

Attached as separate file

Annex 3: Links to the datasets with higher potential for linkage

Multiple Indicator Cluster Surveys (MICS)	https://mics.unicef.org/surveys
Demographic and Health Surveys (DHS)	https://dhsprogram.com/what-we-do/survey/survey-display-478.cfm
FAO agricultural surveys	https://microdata.fao.org/index.php/catalog/956
OCHE operational presence	https://data.humdata.org/dataset/wash-3w-operational-presence-in-ethiopia-as-of-april-2017
OCHA data files on administrative borders and settlements	https://data.humdata.org/organization/ocha-rosa
High resolution population density data	https://data.humdata.org/search?q=High%20resolution%20population%20density %20data&ext_page_size=25
World Global obstacles	https://data.humdata.org/dataset/global-obstacles
IATI datastore	https://iatistandard.org/en/iati-tools-and-resources/iati-datastore/
GSHHG Distance to Water	http://www.soest.hawaii.edu/pwessel/gshhg/
ACLED: Armed Conflict Location & Event Data	https://acleddata.com/curated-data-files/
Global Flood Hazard Frequency and Distribution	https://sedac.ciesin.columbia.edu/data/set/ndh-flood-hazard-frequency-distribution
Estimated number of IDPs at sites assessed by CCCM	https://data2.unhcr.org/en/situations/cccm_somalia
CCCM Cluster Somalia DSA (Detailed Site Assessment)	https://data2.unhcr.org/en/documents/details/63132