Humanitarian Data Exchange
Quality Assurance Framework

This is a descriptive report on the data quality assurance framework that will be adopted by the Humanitarian Data Exchange (HDX) platform. It does not suggest new policies or practices on quality assurance but builds on existing best practices within internationally agreed quality management frameworks. This framework may need to be adjusted as the platform evolves.

Table of Contents

1. Introduction

The goal of the HDX platform is to make humanitarian data available (i.e., easier to find) and useful for decision making (i.e., reliable and comparable for analysis). Doing so will require OCHA and its partners to invest effort into standardising and improving the humanitarian data that is collected: the HDX system will be able to store, validate, and standardise humanitarian data, and to provide data support tools for humanitarian decision making, but the system cannot function unless the community as a whole can feed good data in.

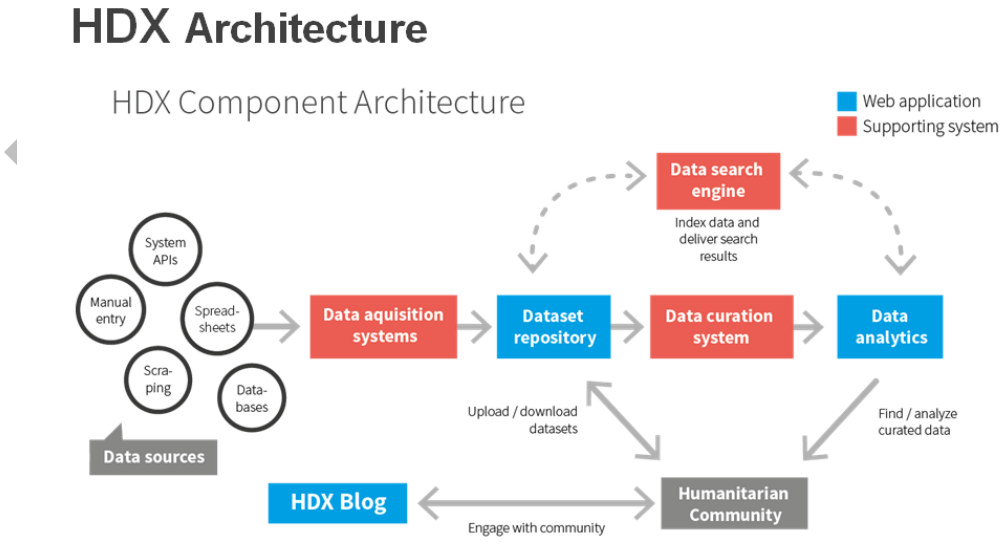
With that challenge in mind, the HDX system is designed to follow a two-tiered approach:

1. We will provide a place where the humanitarian community can upload and share raw, **uncurated** data, to encourage sharing with minimum barriers to entry.

2. We will provide a repository of **curated** data, together with web tools for advanced analysis and visualisation, to support humanitarian decision-making, reporting, and outreach.

Initially, the curated data will be relatively small compared to the uncurated data, but the project will engage in a continuous process of **progressive data enhancement**, working with humanitarian information actors (1) to share data that is not yet shared, and then (2) to improve data that is shared but not yet curated.

In line with this approach, the HDX architecture will be comprised of two main components: (a) a repository for the aggregation of uncurated community datasets, using an open source software called CKAN; and (b) an analytic interface for a curated, common humanitarian dataset of about 200+ metrics that can be compared across countries and eventually crises. For the latter, we will build a normalized, transactional database. Once the platform reaches a certain level of maturity, users will be able to access a range of statistical methods and real-time analytical tools. The visual below provides a logical progression of the two components.



The HDX system design philosophy includes four major principles:

- **Aggregation, not creation** — HDX will collect and curate information that *already exists* in the humanitarian community, and will provide tools to help the community use that information for decision support.
- **Progressive data enhancement** — HDX will enable data-sharing on multiple levels, both uncurated and curated. Over time, we will work to help and encourage information actors to share more data that is currently unshared, and to curate more data that is shared but uncurated.
- **Open data** — HDX will provide technical support for (a) sharing any data, and (b) allowing data providers to decide *not* to share some data for privacy, security, or similar reasons.
- **Open technologies** — HDX will use open source, open content, and open data as often as possible, to reduce acquisition, procurement, and compliance costs.

2. Definition of terms

The following are definitions of common terms used throughout this paper.

Term	Definition	Source
Statistics	Numerical data relating to an aggregate of individuals or events. The science of the collection, organization, analysis, interpretation and presentation of data.	OECD, Glossary of statistical terms. OECD Statistics portal
Statistical indicator	A statistical indicator is a data element that represents statistical data for a specific time, place, and other characteristics and is adjusted for at least one dimension to allow comparability. For instance, a simple aggregation of number of affected people by a disaster is not in itself an indicator as it is not comparable between populations. However, if those values are standardized (in this example, the number of affected people per thousand populations), the resulting data elements make an indicator.	Statistical Data and Metadata exchange http://www.sdmx.org
Metadata	Metadata is data that defines and describes other data and processes. Statistical metadata comprise data and other documentation that defines objects in a formalized way.	Economic Commission for Europe of the United Nations, 'Terminology on Statistical Metadata', Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva, 2000
Time series	A statistical time series is a set of ordered observations on a quantitative characteristic of an observable event taken at different points of time.	The Oxford Dictionary of Statistical Terms, Oxford University Press, 2003
Data validation	Data validation is the process of monitoring the output of data collection, compilation and dissemination, and ensuring the quality of the statistical results. Data validation describes the methods and procedures for assessing statistical data and how the assessment results are monitored to improve statistical processes.	Statistical Data and Metadata exchange http://www.sdmx.org
Data set	A data set is as a collection of similar data, sharing structure covering a fixed period of time. A data set can simply be understood as any organized collection of data.	OECD Glossary of terms for Statisticians', Geneva, 2000

3. Defining quality

Data-driven decision making requires confidence in the quality of the data and a commitment to maintaining a comprehensive data quality framework. Inconsistent, inaccurate, incomplete and out-of-date data are very often the root cause of poor decisions, biased statistical analysis, unsatisfied customers and waste of human and economic resources. Maintaining high quality standards in the overall management of data and keeping a reputation of objectivity and impartiality is essential for the sustainability of the project.

It is important to note that this quality assurance framework is situated in the context of a larger effort that would encompass best practices on community data collection and the development of humanitarian data exchange standards (e.g. HXL). Good data quality depends on sound data collection methods. Data quality is maintained when there is less human manipulation of the data – the more automated the data sharing is (through APIs, data standards, etc.), the less room for errors and bias. The project team will advocate for all data producers to align with the fundamental principles of statistics[1] but realize this will take time to take hold within the humanitarian community.

The most commonly accepted way of defining data quality is in terms of the broad notion of fitness for purpose. Quality is a multi-dimensional entity that combines the relevance of the data for the users and the basic characteristics of the data (accuracy, timeliness, accessibility, interpretability and comparability). Every characteristic of the data and its interactions must be reviewed and managed in order to increase the usability of the data.

The five dimensions of quality are commonly[2] defined as follows:

1. ***Relevance***. The relevance of data refers to the degree to which it meets the current and potential future needs of the clients. This characteristic refers to whether the available information sheds light on the most important issues of the users of information.
2. ***Accuracy***. The accuracy of the data is the degree to which the information correctly describes the phenomenon it was designed to measure. Generally, it could be described in terms of the bias of the estimates.
3. ***Timeliness***. The timeliness of the data refers to the delay between the reference data point to which the information represents and the date on which the information becomes available. Timeliness is a factor that directly affects both relevance and accuracy.
4. ***Accessibility and interpretability***. The accessibility of data refers to the ease with which it can be obtained from the data sources. Cost associated with obtaining and disseminating the data is often mentioned as another aspect of accessibility. The interpretability of the data reflects the availability of the supplementary information (metadata) needed to utilize and understand the data effectively.
5. ***Comparability***. The comparability of data refers to the degree to which it can be successfully brought together with other statistical data within an analytical framework. In this respect, the use of standard methods and concepts promotes coherence across sources of data.

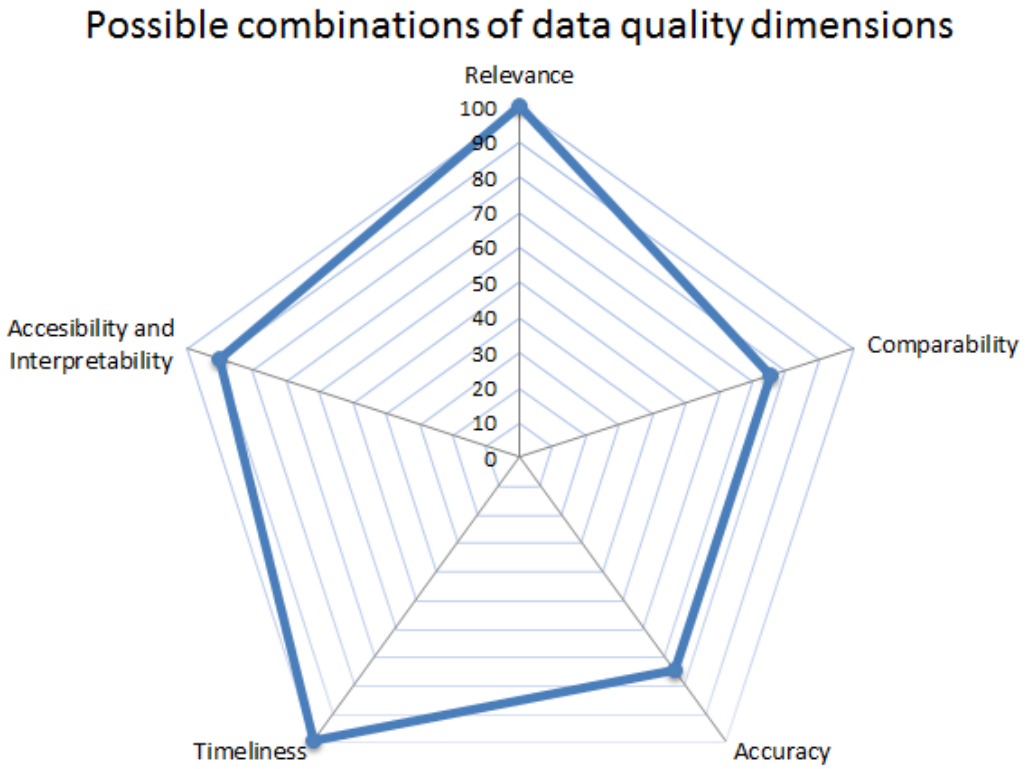
There is not a single model to assess all characteristics of the data with one single quality indicator. Quality is not absolute. Therefore, the assessment of data quality is always a trade-off between the various dimensions of the data, the importance of the data and the complex relationships between them. Actions

[1] http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.asp

[2] Agencies such as Statistics New Zealand, Statistics Canada, The Australian Bureau of Statistics, Statistics Sweden, Statistics Norway, Statistics Netherlands, UNECE, Eurostat, OECD and several other agencies also define quality in these terms.

taken to address one dimension could affect other dimensions. For instance, in humanitarian response work, timeliness may be deemed more important and weighted accordingly in relation to other quality dimensions.

The following visual shows a five-axis radar chart to represent a potential desired end-state for each of the five dimensions as described above.



The ultimate goal of this quality assurance framework is to ensure that data is adequately verified, validated and cleaned for humanitarian use and analysis. This following data quality framework relies on experts knowledge and experience, consultation with data users, interaction with users of social networks, the humanitarian community, and ultimately on judgment.

4. Assessing quality of uncurated datasets

Later in this document, we will define how quality will be assessed in the curated database. For the data that is not curated, the platform will still disseminate the data but it will request users to provide some information about the data they are uploading (user's metadata). This action will take place at the time data is being upload into the platform.

Recently, the HDX conducted a user experience research with a number of participants in selected locations around the world. One to one interviews were conducted and questions on topics like accessing data methodologies, data quality, data security and data sharing practices were asked. The research[3] helped us identifying key users insights. One of them was that people do not want to share data mostly because they are worried about the quality. Having a highly visible quality rating system may discourage data sharing and without the initial sharing data, the HDX platform will not work.

We believe that by providing metadata of the information being uploaded and by avoiding OCHA playing a direct role in assessing the quality of the data, we could bring up the elements for users to directly determine the strengths and limitations of the data sets. The metadata that we are requesting to the users is intended to provide features that could highlight strengths or potential weaknesses or defects in the data. Considering the subjective nature of data quality, it should be understood that there is not system that can identify all problems with the data and users should exercise responsibility in the treatment of data and they should disclose any limitations in the data.

The metadata to be requested at the time of uploading data will cover the following items:

Source: This item captures the information related to where the data was gathered from? Who is the data owner? Who is the person that knows more about this data?

Reference date: This item captures the information related to when the data was collected or about the date the data refers to.

Contributor: This item collects the information about who is contributing with this data into HDX.

Caveats/Defects: This item provides the room to describe potential known defects on the data

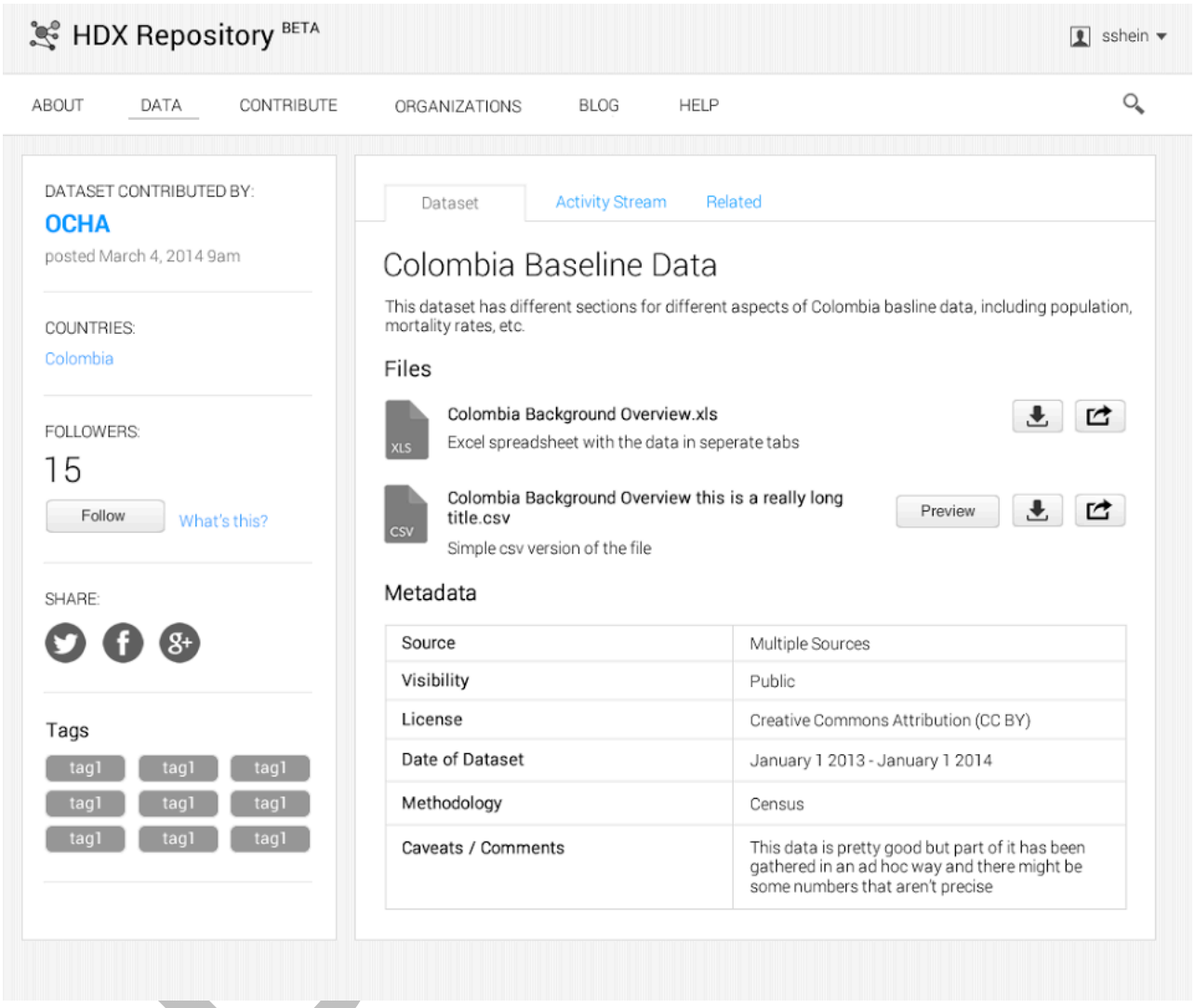
Methodology: This item is intended to provide how data was gathered. It will be presented initially as select all that applies:

- *Census data*. A census is a study that obtains data from every single member of a population of reference. It requires a questionnaire and a clear definition of the population frame.
- *Sample survey data*. A sample survey is a study that obtains data from a subset of a population to estimate attributes of the population. It requires a questionnaire and a sample of individuals selected at random from the entire population.
- *Direct observational data*. Observation data can be obtained from key informants, focus groups, or site visits. Generally speaking, observational data cannot be generalized.
- *Registry data*. Registry data is generally obtained by reviewing or accessing agency records.
- *Other type, specify*. Other methodologies used to obtain the data.

[3] Full report can be seen at https://docs.google.com/file/d/0B4J8rDnUw80gdE54QSWxiTzJKNU0/edit

Terms of use/License: This item defines who data can be shared with. Users will have the ability to share their data publicly or privately. At the moment, the HDX team is also working on the overall data policy and the ways OCHA will manage sensitive data.

Initial wireframe with metadata fields



5. Quality management

The following attributes – relevance, accuracy, timeliness, accessibility and interpretability, and comparability – will be applied internally for assessing the quality of data before it becomes available through the curated database. In order to achieve an acceptable level of quality, all data series included will have to satisfy some degree of the aforementioned quality characteristics. When data fails a quality assurance test, the data manager will have to handle errors, balancing the various factors listed in this

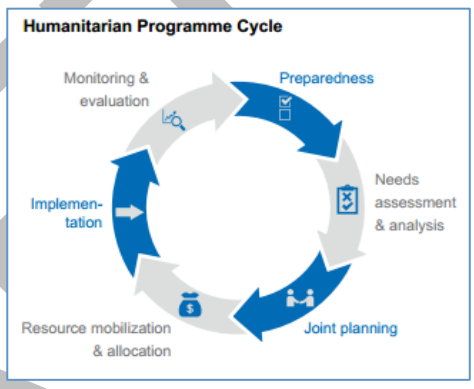
document until the data reaches an acceptable level of quality. Only at that point will a data series will be ready for dissemination through the analytic interface of the platform.

5.1 Relevance

Relevance relates to the degree to which statistical information meets the needs of users. Managing relevance involves liaising with users and clients, reviewing user needs frequently, setting priorities and ensuring that the disseminated statistics conform to international statistical standards.

5.1.1 Relevant humanitarian data

Within the framework of the HDX project, relevant humanitarian data is defined as the 'common humanitarian dataset' (CHD). The CHD is a composite of existing, sometimes siloed, data groupings. To get to the first version of the CHD, the team reviewed existing systems and established data schemas for humanitarian response data across the programme cycle (see related visual).



This included data from the IASC-defined common operational dataset (mainly admin boundaries, populated places and population statistics), data often found in appeals and humanitarian reports (CAPs, SitReps, humanitarian bulletins), data from OCHA's Financial Tracking System, as well as common administrative data (such as surge deployments, vacancies). There was nothing defined for what constitutes useful country context and pre-crisis data so this grouping was based on what was commonly used in various reports and indexes.

After compiling the first version of the CHD metrics in early 2013, the team shared it with various internal and external users to get feedback. These consultations led to some adjustments. For instance, the list of pre-crisis data has grown in size and sensitive data on access has been removed for the time being. Also, in May 2013, supported by the Economist Intelligence Unit, OCHA conducted 45 hours of interviews with humanitarian managers to understand their data needs. Interviewees were asked about the type of questions they need data to answer about humanitarian situations. This helped inform the basic data building blocks required for humanitarian analysis.

The CHD is divided into five sections, each of which allocates a series of indicators. A subset of them is presented in the table below.

Category	Description	Type of Variables and indicators
Country context	Country description	Country name, location, time zone, geography, language, religion, ethnic groups
Pre-crisis indicators	Economic status, crisis vulnerability, health, food security, nutritional status	Total population, GDP per capita, under-5 mortality, life expectancy, literacy rates, access to sanitation, access to electricity
Operational indicators	Needs assessments, response activities,	Crisis name, disaster type, affected people, houses

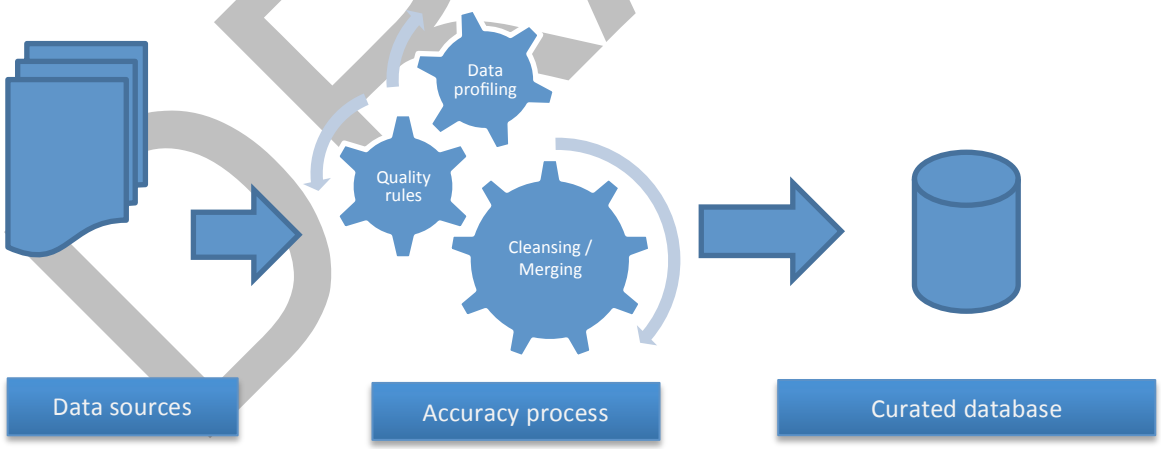
	gaps	damaged, displaced people, people in need, who is responding with what
Funding indicator	Humanitarian financing	Appeal status, CERF, common humanitarian fund, emergency response fund
Administration indicators	Support to the crisis	Budget, number of staff, surge deployments, vacancies, information products issued

The relevance of the data within the CHD will need to be assessed and governed by the HDX team. A process for adding new indicators will be defined as we work through the first iteration of the platform. We are seeing new data series being developed – such as the data from the humanitarian needs overview product – that we will want to add to the CHD. We also did not include data on 'who is doing what where' (3W/4W) in the initial CHD definition but will want to do that given that this data will be a focus on the HXL working group.

The HDX team will carry out an annual online survey, starting in July 2014, to assess the current and future data needs of clients and stakeholders. The team will also continuously interact with humanitarian experts that work with data through the HDX blog space and through face-to-face meetings and events. We will also pay close attention to the user metrics on data downloads coming from the HDX site. By monitoring the current and future needs, we plan to stay abreast of the developments and interests of the humanitarian community.

5.2 Accuracy

Accuracy is the degree to which the data correctly describes the event/phenomenon it was primarily designed for. Assessing data for accuracy is not a one-time activity, but an ongoing practice that ensures trusted data over time (see visual below). The initial activity to assess accuracy is to conduct data profiling.



5.2.1. Data profiling

Data profiling means reviewing different factors that could affect the accuracy of the statistics by detecting patterns and outliers in the data. Understanding the data allows us to add value to the data and helps with data interpretation.

With the support of a volunteer data scientist[4] an automated data validation process is being developed for the curated data. Automation allows data managers to increase the scope and number of quality checks that can be performed. Instead of spending significant resources on data validation, resources can be redirected towards activities with higher pay-off such as data analysis and error handling analysis. While some manual intervention may be needed, generalized rules and reusable software are useful to accomplish this purpose.

The main task of the automated quality assessment is to identify extreme data values in a period or across countries. The presence of outliers is a warning sign of potential errors that could lead to biased analysis and incorrect decisions.

The following are basic validation rules that are under development:

1) Data ranges
Ensuring data is within the data ranges as defined in the indicator definition. That is, every figure included in a data series referred as percentage (of a total) should be bounded by 0-100%; For example, the World Bank's data series on percentage of population with access to electricity, calculated as the ratio of people with access to electricity divided by total population, should be a number between 0% and 100%. Data series deviating from this rule should be flagged for error handling.

Access to electricity (% of population)		
(Source: World Bank at http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS)		
Country	2009	2010
Afghanistan	15.60	30.00
Algeria	99.30	99.30
Angola	26.20	40.20
Argentina	97.20	97.20
Bahrain	99.40	99.40
Bangladesh	41.00	46.50
Benin	24.80	27.90
Bolivia (Plurinational State of)	77.50	80.20

2) Extent of change (value deltas)
Ranges on changes in values are unlikely to change dramatically from one year to another. Significant changes[5] should be flagged for error handling. For example, FAO's data series on land area, calculated as the total area of a specific country in squared kilometers is very unlikely to change from one year to another, unless the country is newly created (Sudan/South Sudan figures were expected to be revised after the split of Sudan on July 9, 2011).

[4] Andrew Rosenfeld
[5] Significant changes can be determined by using the global average and by building a confidence interval around the average change for all countries in a specific year.

Land Area Sq. Km (Source: FAO Statistics)									
Country	2001	2002	2003	2004	2005	2006	2007	2008	2009
Afghanistan	652,230	652,230	652,230	652,230	652,230	652,230	652,230	652,230	652,230
Albania	27,400	27,400	27,400	27,400	27,400	27,400	27,400	27,400	27,400
Algeria	2,381,740	2,381,740	2,381,740	2,381,740	2,381,740	2,381,740	2,381,740	2,381,740	2,381,740
Somalia	627,340	627,340	627,340	627,340	627,340	627,340	627,340	627,340	627,340
State of Palestine	6,020	6,020	6,020	6,020	6,020	6,020	6,020	6,020	6,020
Sudan	2,376,000	2,376,000	2,376,000	2,376,000	2,376,000	2,376,000	2,376,000	2,376,000	2,376,000

3) Data type validation

Data series defined to be of the type of an integer value, such as some sort of code (e.g. 1,2,3). For example, EM-DAT series on the total number of people affected by disasters should not have a fractional value (e.g. 2.5 people).

Number of people made homeless by disasters (Source: EM-DAT, The International Disaster Database)								
Country	2006	2007	2008	2009	2010	2011	2012	2013
								2130
Afghanistan	8210	3480	180	3250	1000	9700	2680	
Albania		75		150				
Algeria	150			2500				
Angola	225	6000		5065	78875	100		
Argentina		5000					2000	
Australia	7141	90						

4) Part/Whole relationships

When separate indicators exist for overall and by type, the sum over type should equal overall. For example, total funding for an emergency should match the sum of per-cluster funding indicators (including "no cluster assigned") for that emergency.

5) Consistency among different sources

If a data series can be collected from more than one source, or there are closely related data series, deviations should be minimal or explainable, otherwise the series should be flagged for error handling. For example, under five mortality rate per 1,000 live births is a data series originally developed by the Inter-agency Group for Child Mortality Estimation (UNICEF, WHO, UN Population Division and World Bank) and can be sourced at UNICEF'S Childinfo[6] portal. However, the data series is referenced at several data portal such as the World Bank, the United Nations Statistics Division Millennium Development Goals Database, and the UNDP Human Development Reports database. The HDX team will obtain data from original sources.

6) Sparsity (ie, completeness)

Data series having only few countries with data should be flagged for error handling. Data series with missing years/months/days in the middle of a time range should be flagged for error handling.

7) Correlation

Data series with correlations close to unity should be reviewed to assess whether there are duplicated series.

8) Logs

All logs of the data validation are being kept and made public to provide data users the means to reproduce the data as received by the original source.

[6] http://childinfo.org

Other factors being reviewed for accuracy of the statistics include:

- Coverage deficiencies (population frame problems, missing metadata);
- Input data errors (coding errors, missing data, wrong periods);
- Errors introduced during the production of the data (duplicated records);
- Methodological deficiencies (incomplete metadata, failures in methodologies);
- Output data errors (editing errors, transformation errors, transmitting errors)

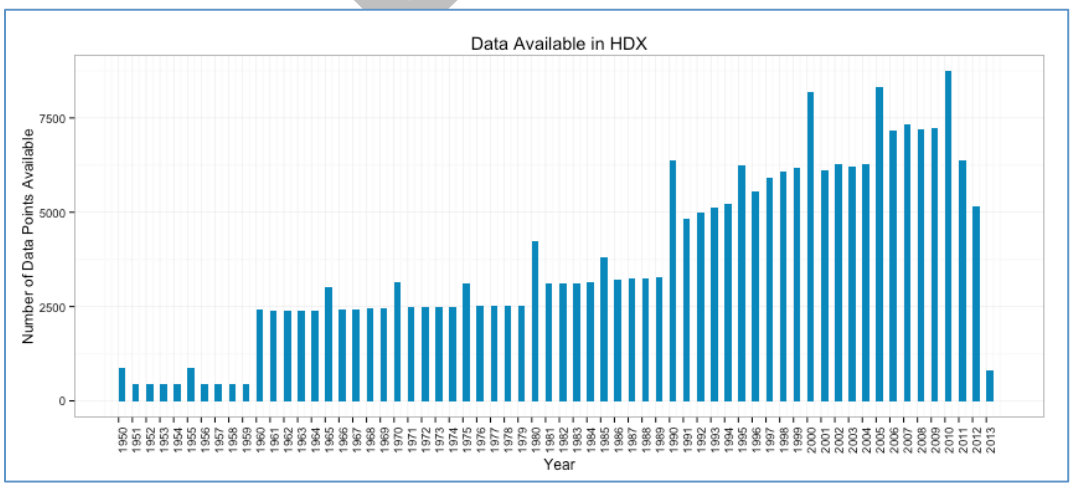
These validation rules are currently applied using Pandas[7], a Python data analysis library, but may be re-implemented in Java to fit more directly into the HDX platform. In order to address the quality of the statistics on the above mentioned factors, a detailed revision of the metadata for each series is being conducted.

5.3 Timeliness

The timeliness of statistical information refers to the delay between the reference period to which the information pertains and the date on which the information is made available. The timeliness of information also affects its relevance.

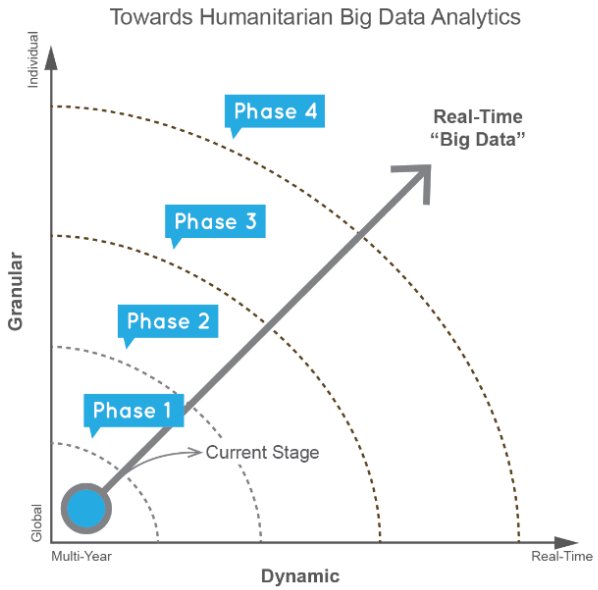
The humanitarian community requires data as contemporary as possible. To reduce the delay in the availability of information, the HDX team has created a number of automated methods that take advantage of APIs[8] and data scraping technologies to ensure that our system has the newest version of data as soon as it becomes available.

We have built the basic infrastructure of the data collection process for pre-crisis data together with ScraperWiki[9], an organization specialized in the scraping of data from the web. As of February 2014, we have about 300,000 data points in our system that cover a period of approximately 60 years, 240 territories and 100 different indicators (see visual below).



[7] Pandas is an open source data analysis tool for the Python programming language. http://pandas.pydata.org/
[8] http://en.wikipedia.org/wiki/Application_programming_interface
[9] https://scraperwiki.com/

The project platform has at the moment data from 1950 to 2013. As the project makes progress, the team will include data that is more complex (i.e. field level, granular and dynamic). The visual below illustrates the project's vision for bringing in more data over time. As data becomes more granular and lower-latency, there will be compromises with data accuracy, which the team will monitor for each series.



5.4 Accessibility and interpretability

Accessibility deals with the ease with which statistical data can be obtained including the suitability of dissemination and costs. The interpretability of statistical data reflects the availability of metadata necessary to interpret and utilize the data properly.

5.4.1. User interfaces

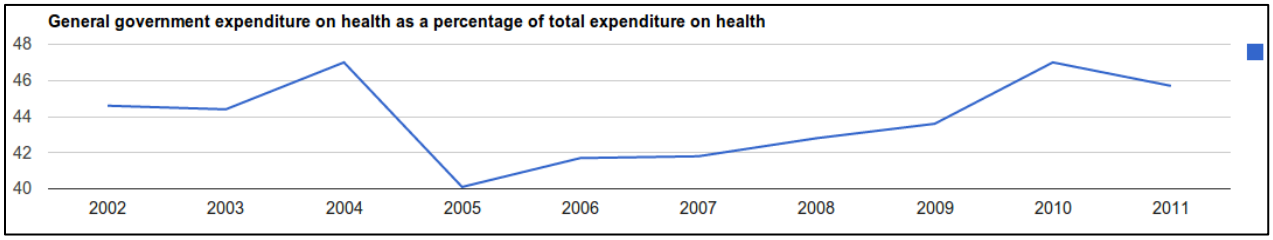
We will be designing a user interface for the uncurated and curated data through the Reliefweb site (URL likely to be data.reliefweb.int). Users will be able to search for datasets or specific indicators, visualize the data over time, and download it into a spreadsheet and take it with them.

For instance, in a user interface, we might present a time-series indicator as a single row of data:

Brazil: Health spending as % of GDP (WHO)

2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
44.6	44.4	47.0	40.1	41.7	41.8	42.8	43.6	47.0	45.7

Or we might present it as a graph:



At this time we are not able to be too specific about what the user interfaces will entail except that it will be engaging and well designed. We plan to carry out extensive user research to understand how users want to access data and how they intend to use it.

5.4.2 Open data APIs

The second gateway into HDX data will be open data APIs. These are places that computer systems can download machine-readable versions of the data for their own processes. The APIs will be designed following the principles of Representational State Transfer (REST), allowing the operations to be those of the web's Hypertext Transfer Protocol (HTTP), essentially simply downloading open data CSV or XML files. Future phases may include more-sophisticated APIs if demand warrants.

5.4.3. Statistical metadata

A vital factor in a quality assurance framework is the use of structural metadata. *Metadata* can be defined as data that define and describe other data, whereas *Statistical metadata* are defined as data about statistical data, and comprise data and other documentation that describe objects in a formalized way[10].

Statistical metadata aims to promote a consistent interpretation of statistics and thus the benefits of its use and implementation are an improvement of the interpretability of the data; better discovery, retrieval and exchange of data and metadata; common terminology to improve communication; and as a consequence of this a better quality of information.

The statistical metadata may include descriptions of statistical indicators, term definitions, units of measurements, classifications, aggregation methods, data limitations and constrains, source of potential discrepancies, comparison with other series, data collection methods, data availability and expected releases, websites and other metadata related to the statistical data.

In HDX, the statistical metadata has three main objectives: (1) to assist the humanitarian community to identify the most suitable information that meet their needs; (2) to assist them to interpret its content and (3) in case data is downloaded from the HDX platform, to assist them in the data post-analysis and potential comparability with other data.

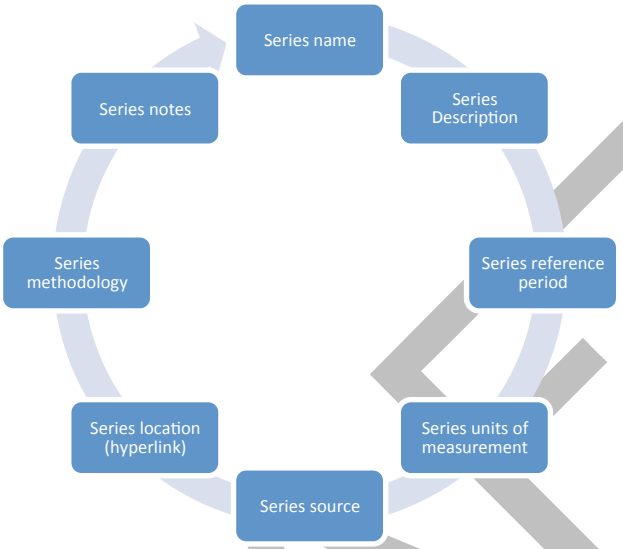
The minimum set of metadata required for the adequate interpretation of indicators will be:
- Name of the indicator
- Labels in rows and columns and definition of those labels
- Measurement Unit
- Time reference
- Source of data

[10] SDMX Metadata Common Vocabulary, January 2009 version

- Disclaimers, copyrights and restrictions of usage

Country code and area codes for statistical use will be based on those defined by the United Nations Statistics Division.

Metadata contained in the HDX platform will include the following:



5.5 Comparability

Comparability (also referred to as coherence) reflects the degree to which information can be successfully brought together with other statistical information within a broad analytical framework. This characteristic of the data covers the internal consistency of the data and the comparability with other data sources.

The development and use of standard international frameworks, concepts, classification and methodologies contribute to data coherence. In order to contribute to this aspect of data quality, OCHA is leading an initiative to develop standards to share operational data across agencies, non-governmental organizations, and governments. The HXL working group will also be reviewing existing data exchange standards such the International Aid Transparency Initiative (IATI) and the Statistical Data and Metadata Exchange (SDMX)[11] to align with already agreed data sharing standards among international organizations.

Another aspect of comparability relates to the internal consistency of the data. The project team has developed partnerships with data suppliers that support government data collection efforts, such as the World Bank Open Data for Resilience Initiative (OpenDRI). This aims to ensure that common methodologies and systems are used across humanitarian and development cycles and internal consistency is ensured.

[11] http://sdmx.org/?page_id=14

6. Disclosure control

In December 2013, OCHA[12] endorsed the Principles Governing International Statistical Activities. The sixth principle is explicit on statistical confidentiality: "Individual data collected about natural persons and legal entities, or about small aggregates that are subject to national confidentiality rules, are to be kept strictly confidential and are to be used exclusively for statistical purposes or for purposes mandated by legislation". Based on the guidelines of good practices from the UN Statistics Division, the HDX team will use the following principles for managing the confidentiality of data elements within a dataset (ie, microdata)[13]:

(a) It is appropriate for microdata collected for statistical purposes to be used for statistical analysis to support research as long as confidentiality is protected;
(a) Microdata should only be made available for statistical purposes;
(b) Provision of microdata should be consistent with legal and other necessary arrangements that ensure that confidentiality of the released microdata is protected; and
(c) The procedures for researcher access to microdata, as well as the uses and users of microdata, should be transparent and publicly available.

The HDX will also be developing terms of use for the data shared through the platform and will clarify how OCHA will handle sensitive and secure data.

[12] http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/endorse.asp
[13] http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

7. Conclusion

The following table provides a summary of the activities that will be carried out by the HDX team to ensure data quality.

Pillar	Target	Activities
Relevance	Stakeholder and user feedback	Online survey, consultations, communications, HDX blog, face to face meetings, and events
Accuracy	Scraped data, country-specific data, data to be curated, community data	Data profiling, statistical analysis, promotion of self-assessments, metadata provisions
Timeliness	Data providers, data collection	Scraperwiki, APIs, data sharing agreements
Accessibility and interpretability	Dissemination platform, CKAN	User interfaces, open data APIs, Metadata
Comparability	Data partners	HXL, IATI, SDMX, and other data standards

This quality assurance framework aims to define an effective control mechanism to assess the quality of statistics for HDX. The effectiveness of this framework depends not only on the application of each and all of its components but also on the mutual cooperation of all players involved in the implementation of this project. The evolutionary process of implementing business intelligence within the landscape of humanitarian operations should not be underestimated.

This framework also intends to create a culture of quality by providing tools needed to assess quality responsibly. Creating an environment of sustainable quality requires that all dimensions of quality are constantly being evaluated and interacting with each other.

8. References

1. Statistics Canada (2003). Methods and practices. Statistics Canada Catalogue No. 12-587 XPE. Ottawa, Canada
2. United Nations Statistics Division (2004). Handbook of Statistical Organization, Third Edition: The organization and operation of a Statistical Agency. New York. USA
3. Brackstone, G.J. (1993). Data relevance: keeping pace with user needs. Journal of official statistics Vol 9, No.1. Statistics Sweden. Stockholm, Sweden.
4. Eurostat (2003). Minutes from the working group on Assessment of quality in Statistics. Luxembourg, Luxembourg.
5. United Nations Statistics Division (2005). Household sample surveys in Developing and Transition countries. New York, USA.
6. United Nations Statistics Division (2005). Principles governing international statistical activities. New York. USA
7. Cochran, W.G. (1977). Sampling techniques, New York, USA.
8. Hiridoglou, M.A. (1944). Sampling and estimation for establishment surveys: stumbling blocks and progress. American Statistical Association. USA
9. United Nations Statistics Division (2010). Minutes from the National quality assurance frameworks workshop. New York. USA
10. Hiridoglou, M.A. and Berthelot, J.M. (1992). Statistical editing and imputation for business surveys. Survey Methodology. Ottawa, Canada.