QUALITY MEASURES FOR HUMANITARIAN DATA

SPRINT REPORT

APRIL 2023

DATA
NUTRITION
PROJECT

QUALITY MEASURES FOR HUMANITARIAN DATA

SPRINT REPORT

APRIL 2023

DATA NUTRITION PROJECT

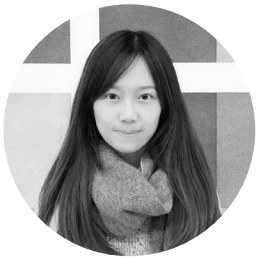
MEET THE TEAM

Lead
Kasia Chmielinski

Technology
Matt Taylor

Research
Sarah Newman

Design
Jessica Yurkofsky



Design
Chelsea Qiu

I.
BACKGROUND

Goals

The purpose of this Data Labeling project was for select members of the Data Nutrition Project team to research and prototype possible quality measures for humanitarian datasets that are hosted on the HDX platform, which is owned and managed by the UN Centre for Humanitarian Data. The scope included:

- **User and Platform Research.** We conducted user research with the Centre team (and additional stakeholders suggested by the Centre) to learn about 1) Different conceptions of data quality in the humanitarian sector; 2) How users find and select data on HDX, including priority of criteria; 3) The current DPT / HDX team QA workflow with regards to assessing data quality.
- **Quality Measurement Prototype.** Building on user research and an assessment of the state of the data and the needs in play, and using two preselected datasets as examples, we prototyped a quality measures label for HDX. The prototyping involved varying degrees of fidelity and was shaped by feedback from the Centre team.
- **Preliminary thoughts on Scalability.** Through research and prototyping, we began to explore how this effort could scale, including paths toward automatability. Our findings are discussed in this report.

Philosophy

The Data Nutrition Project is a non-profit initiative that formed in 2018 to develop tools and practices to improve transparency into datasets. Our team is interdisciplinary, and we leverage insights from a variety of fields, including product development, data science, ethics, engineering, design, and education. Our approach with our Nutrition Labels for Datasets is threefold: 1) We encourage the creation, documentation, and publishing of higher quality data; 2) We enable transparency into datasets through our legible, extensible, interactive framework; and 3) Our Labels provide education about what kinds of information a user should ascertain before using a dataset. We bring this approach into our work with clients, where we prioritize user-centered design, realistic goals, and practitioner-focused outcomes, informed by our experience working in data transparency initiatives and with the real tradeoffs and tensions faced by data practitioners. In seeking to do work that is both applied and realizable, we aim to provide not only a long-term vision but also a roadmap with recommendations for future iterations of a project.

II.
CHALLENGES

There are many challenges that can be impediments to dataset quality. This is certainly the case in the humanitarian sector, where crises unfold quickly and data capture will almost always be imperfect, often as a consequence of the need for rapid collection. Furthermore, on a more philosophical level, the assigning of rankings, scores, or grades to a dataset will always be tricky business, for the legitimacy of the scoring standards themselves can undermine the effort for scoring in the first place. We believe it is useful to explicitly enumerate these challenges before we describe our recommendations. The latter were formed in light of the former, which will be familiar to the HDX team and to others who have worked on dataset metrics, measures, and assessments.

Challenge 1 - Identifying scoring methods that are succinct while not overly simplistic

Scores are meant to provide information quickly and ease comparison, while inviting further exploration. They can, however, risk being reductive or overly simplistic. This is particularly difficult when comparing datasets whose provenances are entirely different. A score that is too simplistic will not only be useless but may also seem arbitrary. A single score to compare across inconsistent data types or domains may risk both. Depending on the scoring framework, there is the additional challenge of validating accuracy: *what is the rubric by which this score was determined? How is accuracy of evaluation defined and ensured?*

Challenge 2 - Balancing scalable (quantitative) & comprehensive (qualitative) measures

Qualitative information helps mitigate some of the concerns above, as it is often more context-aware than quantitative statistics alone. However, qualitative information (such as detailed provenance information) is also resource-intensive to collect, often is domain-specific, and is sometimes impossible to obtain, such as when provenance is simply unknown. Conversely, quantitative measures can be easier to automate and thus easier to scale, but they can miss nuance or context that is essential to understanding the particulars of a dataset. *The tension here is between usefulness and scalability; in our experience, this is the most common challenge in dataset transparency efforts.*

Challenge 3 - Communicating quality to motivate rather than disincentivize

Our hope is that quality measures will, in the short term, facilitate better data use choices, and in the long term motivate the creation and publishing of better quality data by changing user expectations and data collection habits. However, depending on how measures are disclosed and how scores are determined, they could discourage full transparency when sharing data in cases where increased transparency might negatively affect a score. The challenge here is to *motivate better quality data without penalizing or disincentivizing current dataset owners from sharing data or disclosing shortcomings*. For example, over the course of our research, a data organization voiced concern that their data was being marked "incomplete" even though it was "as good as

it possibly could have been" given a particular set of circumstances. It is important to note that in some scenarios certain information cannot be ascertained and this should not reflect negatively on the quality of the dataset.

Challenge 4 - Building a quality framework that balances flexibility with consistency

As the humanitarian sector changes over time with respect to crises and data needs, any discrete quality metrics will also likely change. For these reasons, *whatever is built will need to be adaptable. However, consistency is also important*, so that datasets from different time periods can be compared, and so that dataset owners and site visitors can develop familiarity and comfort with the site. For example, when and how do existing datasets get re-evaluated under updated scoring rubrics? What is the right approach to keeping information up to date (timely and punctual) that will be both robust and scalable across thousands of datasets? When or how will that score be altered or downgraded as time goes on?

Challenge 5 - Determining responsibilities within the data pipeline

HDX is committed to hosting good quality data, and requires QA and other data onboarding processes. However, HDX, like all organizations in the humanitarian sector, has limited resources, with respect to time and personnel for validating datasets. Furthermore, even if there were no resource constraints, *there are always knowledge gaps* between on-the-ground domain expertise and data experts looking at raw or processed data. This challenge is shared among all data validation efforts, and it might be even more drastic in the domain of humanitarian data, where data collection methods require agility, and thus context-awareness and domain knowledge is essential for accurately interpreting or validating the data. However, HDX is extremely well-positioned to build a *process*, and for that process to incorporate *shared responsibility* for data-validation, as we discuss below.

III.
KEY FINDINGS

The five-week sprint gave rise to a number of key findings, which informed our final designs and recommendations.

Finding 1 - HDX is best positioned to define rather than assess quality

Due to the range of dataset types on HDX, limitations of domain knowledge, and resource constraints, HDX is not well positioned to conduct quality assessments for all datasets on its platform, and instead, HDX should focus that attention to data types or categories that have been defined as critical, such as the Data Grids. HDX is best positioned to

1) *Define the framework* for "quality" of datasets on HDX (meaning define *what* is getting collected and assessed);

2) Facilitate the gathering of this information from data organizations;

3) Provide a display of this information to data users on HDX, designed in alignment with user needs.

Finding 2 - There is an opportunity to leverage existing quality measures

Data quality assessment is already conducted on HDX, albeit at different times and displayed or communicated in disparate regions of the site. This provides an opportunity, as a first step, to aggregate, prioritize, and organize information that has already been gathered. Further ambitions to collect and validate data that is not currently collected should come *only after* the current effort to bring transparency and legibility to the valuable information that HDX already collects. It is our belief that a phased approach, starting with the information already collected, and expanding out from there to leverage the credibility of data organizations, will be most suitable to HDX's current infrastructure.

Finding 3 - Domain experts and third-party validators can provide complementary value

Due to the variety of data types and domains on HDX, data content quality assessment is likely to require verticalization (e.g. different approaches for GIS data, CODs, etc). Considering the ranges of domain expertise required, and the limited bandwidth on the Centre team, our recommendation, as stated in Finding 1, is that the Centre define the parameters of "quality" through a set of *extended* metadata that is to be displayed HDX, and then work with data

owners and third-party assessors who will be the ones to contribute much of this extended metadata. This would follow HDX's work to aggregate the information already collected, and determine the best design approach for communicating this information to users. Many data organizations have quality frameworks or assessments for their own data, and these could be indicated on HDX to communicate things like known issues and certain strengths of a dataset within its domain. Working in collaboration with these organizations can lend institutional validation to these quality frameworks – which, either independently or alongside additional support, could motivate third parties to work with HDX – and would provide domain expertise to HDX and its users. Using these external third-party-determined metrics also encourages other organizations to consider adapting their use frameworks to include quality, which can drive cultural change around responsible data usage.

Finding 4 - Automation of QA and assessment tasks can enable scale

While there is some initial groundwork required to prepare for automating certain QA tasks, we believe that HDX investing the time to move in this direction will, in the long run, enable quality measures to be assessed and applied at scale. For example, automating the collection of certain metadata (such as restrictions around API use, required metadata through API collection, or requiring the use of HXL) can make it easier to enforce and collect technical quality information about datasets. We know that HDX has done a lot of work toward formatting datasets into HXL and advocating for others to do the same. We understand that this is no small feat, but if HXL can be more widely adopted, this would enable greater interoperability, comparability, and analysis.

Finding 5 - Primary use case is data selection, which can ultimately be supported through metrics comparison

Through our conversations and interviews, we learned that the primary reason for having quality measures included on the platform is to improve dataset selection, with additional use cases including dataset comparison (choosing among several on HDX, wanting to quickly ascertain which is better for a certain need) or dataset combination (merging several to cover a bigger geographic region or to otherwise extend a particular dataset). The availability of legible, digestible measures, such as format, update frequency, and uses and restrictions, permits users to quickly scan for their particular needs. We recommend prioritizing information that is already available, and in later phases collecting (and collaborating with others to collect) additional information that can bring even more value to dataset users.

IV.
APPROACH & DIRECTIONS

We explored quality measures on the HDX platform through a tailored 5-week discovery sprint in a three-part inquiry: 1) Researching data quality principles in the context of humanitarian data and the existing platform; 2) Developing several prototype directions building on this principles-based foundation drawing on our expertise in data quality "labels"; 3) Refining our direction and recommended implementation strategy based on feedback from the Centre.

Research

Our team spent the first two weeks researching background materials and interviewing diverse stakeholders. We read and analyzed approximately a dozen reports about data quality principles in the humanitarian sector (published by HDX, UNICEF, UK AID, IOM, OCHA, ICRC, GAHI, DSEG, GSQAF, including others) in order to build an understanding of quality measures on the HDX platform. We sketched a matrix aligning and comparing principles across several major organizations (compared primarily to GDQAF principles, which were most commonly cited as baseline) in order to better understand these concepts in the context of HDX (*fig. 1*). Notably, we saw a gap in the literature published by HDX around "credibility," so we sketched this into the matrix (in light blue).

In parallel, we conducted interviews with stakeholders that represented key points along the data collection, processing, hosting, and use timeline, including data partners (IOM, Humanitarian OpenStreetMap Team) and several within the Centre (Data Partnerships, Data Responsibility, organization onboarding, product development, quality assessment process, and others), in order to understand how the Centre thinks about quality, to learn about existing mechanisms for identifying and surfacing quality issues, and to explore future scenarios for expanding or adjusting quality assessment practices.

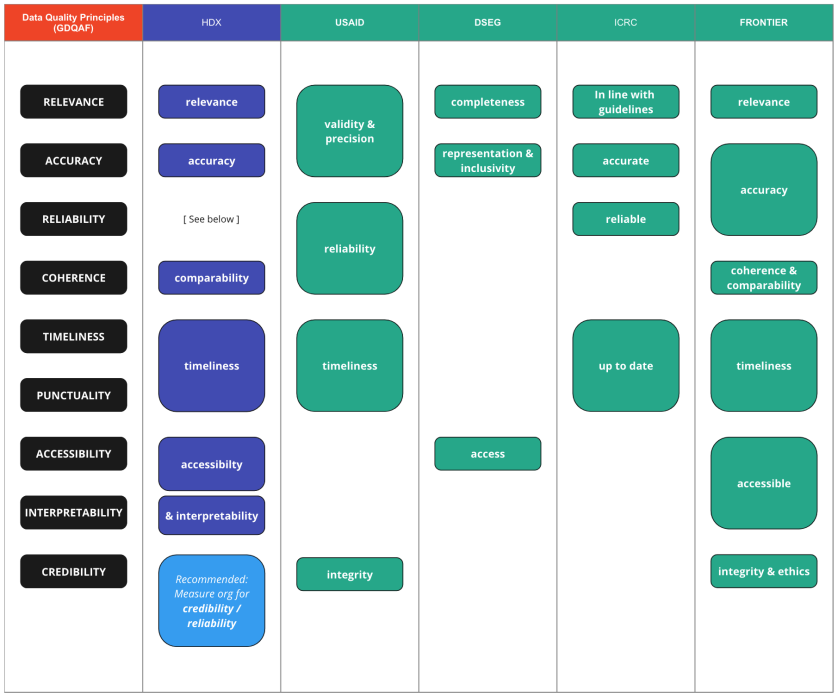


Figure 1. Matrix aligning data quality principles across several organizations, with HDX in blue. Full size matrix included in appendix.

Prototype directions

Based on the research and interviews conducted in the first two weeks of our sprint, the DNP team developed four potential paths forward to highlight specific data quality principles. These were:

1. Comparability: features to support dataset selection

Throughout the interviews, we heard that a primary use case for data quality assessment on HDX (and more broadly in the sector) was the enabling of *dataset selection*, either for a particular need at hand, to join with other external or proprietary data (e.g. Combine several datasets about a particular geography), or to compare against similar datasets to assess which is best aligned for a particular use (e.g. Identifying which administrative boundary dataset is appropriate if there are several to choose from). To support this particular use case of dataset selection, which we felt was best aligned to the data principle of **comparability**, we proposed features that support the direct comparing and contrasting of datasets based on metadata comparison (*fig. 2*)

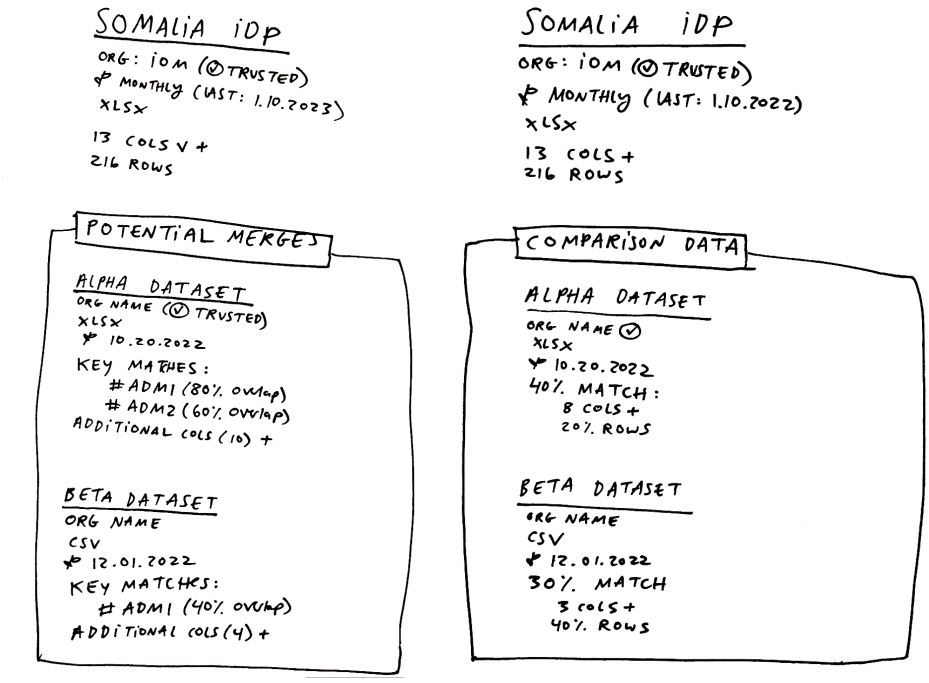


Figure 2. Prototype sketches of features to support the comparing and contrasting of metadata on similar datasets.

2. Credibility: leveraging trust in organizations

From the earliest conversations, data organizations and the Centre teams stressed the critical dependency between data quality and data collection and processing practices. In fact, although there were several measures and principles of data quality that came "after" the processing of the data – such as freshness, accessibility, and interpretability – some of the most salient measures could only be assessed by those most familiar with the origin of the data itself (*fig. 3*). This highlighted not only the importance of but also the reliance upon data organizations to help surface data quality on HDX.

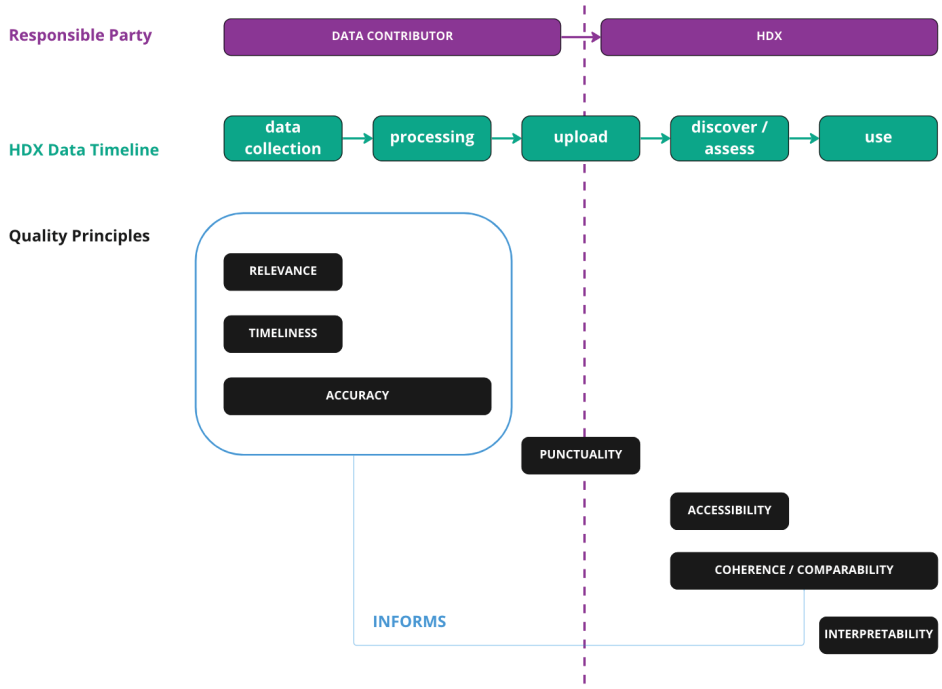


Figure 3. Aligning the data pipeline to quality principles and responsible parties highlights the critical relationship between the data organization and HDX for the assessment and communication of data quality.

Our second prototype aimed to *leverage trust in organizations* based on their data practices and commitment to data quality as a proxy for the data quality principle of **credibility**. For this approach, our recommendation (which follows the initial work performed by DNP team member and 2021 Strategic Communications Data Fellow Kasia Chmielinski), HDX would build a framework for organization trust "levels" that signal an active, relational approach towards quality: one that depends on the organization that produces or publishes the dataset. Datasets would then qualitatively inherit the credibility from the organization that produced them, with the organization serving as a proxy for responsible data collection and processing practices (*fig. 4*).

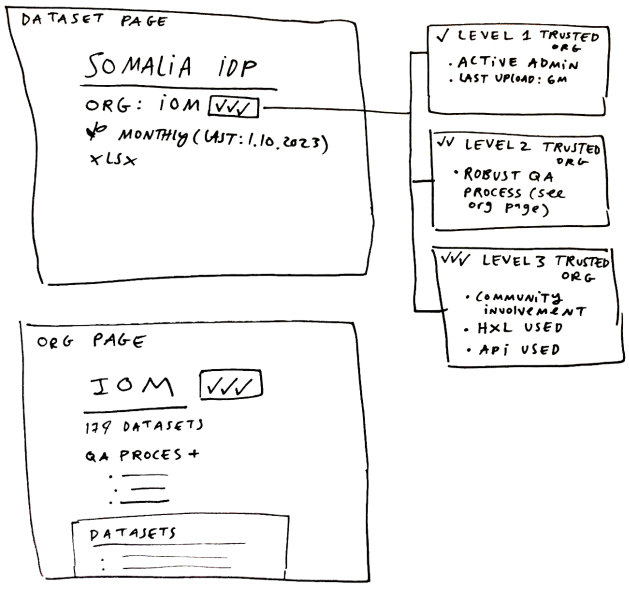


Figure 4. Prototype sketches of a dataset with a "level 3" trust organization indicator as a proxy for the data quality principle of credibility.

3. Completeness, timeliness: assessing metadata completeness

The last two prototype directions are related to fitness for purpose - assessing whether what is represented in the dataset is appropriate for use (**complete, timely, relevant, accurate**). We found fitness for purpose metrics to be the most challenging due to the distribution of responsibility in dataset management, and the realities of data collection in humanitarian situations. Stated simply, it is very hard to assess quality without an ideal "ground truth" dataset against which to compare, or without access to information about the collection and processing practices of the data owner. To facilitate our analysis, we approached these measures along two axes: the assessor (data owner vs. third party), and the type of assessment (qualitative vs. quantitative) (*fig. 5*). The resulting matrix helps clarify four directions: qualitative (not easily scorable) assessment at two levels of granularity, where the data owner can be much more detailed than a third party, and two quantitative (more easily scored or ranked) approaches, one domain-specific (e.g. GIS data quality, ranked by the data owner) and the other focused on adherence to a metadata or technical standard (e.g. metadata completeness, in this case conducted by HDX).

Building off this analysis, our third prototype direction falls within the third-party quantitative assessment for **completeness** and **timeliness**: measuring and reporting *adherence to standards* of metadata completeness. This version, along with the former (Credibility of Trusted Orgs) and the one that follows, combine for our recommendation to the Centre team. In this version, HDX builds a framework of expectations of metadata standards and provides a score or indicator on whether a dataset meets that standard. For example, HDX could incentivize higher data quality through a measurement that assigns a higher score (or ranking) for the use of automatic data submission, the use of standard data elements (such as P-codes or HXL), or attestation of a third party review. For simplicity of communication, these metadata could be categorized into sections (such as "Trust," "Content Quality," "Uses," etc.) (*fig. 6*).

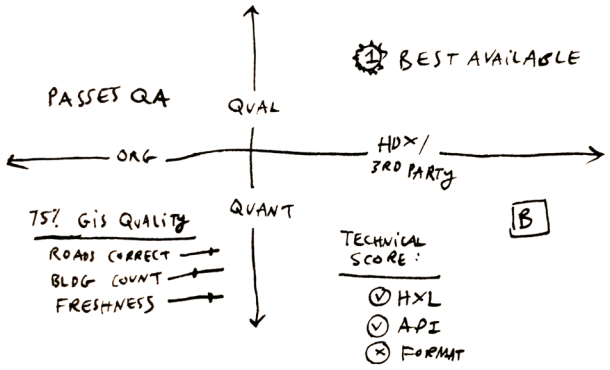


Figure 5 (left). Quadrant analysis of fitness for purpose measures along two axes: responsible party (data owner / third-party) and type of measure (qualitative / quantitative).

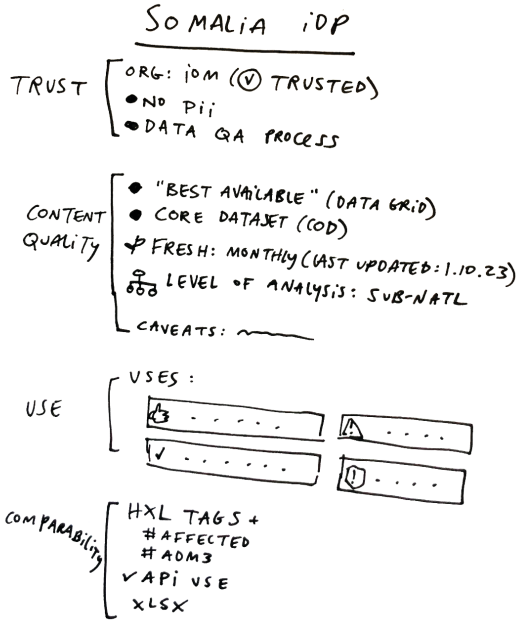


Figure 6 (right). Prototype for a "quality label" that assesses and highlights metadata completeness across a number of categories and against a common framework.

4. Relevance, accuracy: assessing fitness for purpose

The final prototype approach, assessing fitness for purpose in alignment with data principles **relevance** and **accuracy**, most likely requires direct input from the data owner or a party that is familiar with the entire lifecycle of data collection and processing (see the left hand side of the quadrant matrix *fig. 5*). This is because it is extremely difficult, if not impossible, to understand the quality of content contained within a dataset without understanding the context in which it was gathered, processed, and how it will be used. This requires both knowledge of the data collection process as well as significant domain expertise.

Our suggestion for this approach is thus to acknowledge the dependency between HDX and the data owner (organization) and leverage structured frameworks for the communication of dataset quality information. For example, HDX could partner with third party organizations that certify certain data quality domains (e.g. GIS data, CODs, boundary data, education or health data) and when datasets achieve that certification, HDX could communicate that information on the platform. In some instances, HDX may also be a certifier – the Data Grids are a good example of this – but for the sake of scaling, expanding to third party certifications will be less resource-intensive and more applicable to the breadth of datasets within the HDX platform. An additional option might be for HDX to work with particular data organizations to surface their internal quality measures through standardized metadata fields that, while not consistent with respect to *content* across data organizations, would be a consistent *field* that appears on all datasets regardless of organization or domain (e.g. a "domain-specific quality measure" field that could be defined differently across organizations and domains). For example, in the figure below (*fig. 7*), HDX has defined a portion of the metadata structure to include both a "self-reported QA adherence" from the Data Organization as well as a "domain-specific metric" built by a third party.

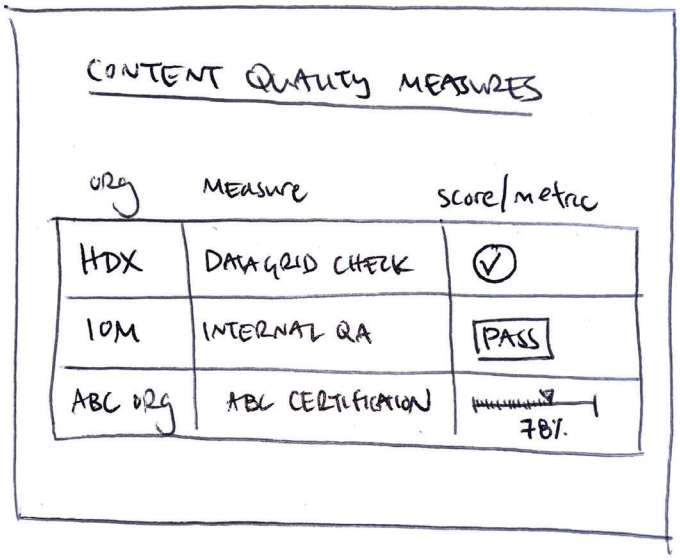


Figure 7. Prototype sketch of Content Quality Measures that include quality certifications from HDX, Data Org / Owner and Third-Party expert organizations. These could be binary (pass / fail) or quantitative (e.g. a score).

V.
RECOMMENDATIONS & DESIGNS

Recommendations

From our four prototype directions outlined above, and integrating feedback from the Centre team, we recommend a *combination* of prototype directions 2, 3, and 4, organized into three phases of work. Detailed work plans and sketches for Phase 1 are included in this report.

We also share recommendations for how to approach the continuation of these explorations in Phases 2 and 3, and include a preview (low fidelity sketch) of how extended information that might be included in Phase 3 could appear on HDX. The phased approach enables HDX to prioritize the information that is already available, while beginning to build an on-ramp for further work that will require additional resources and infrastructure. A summary of the phases is included below, and discussed in more detail in *Section VI*.

Phase 1 - Fully scoped, ready to implement

Aggregate and make easily accessible the content that HDX already collects with a specialized Quality Measures pane on the HDX site. This Quality Measures pane is divided into four parts: Use, Trust & Safety, Content Quality, and Technical Specs. The measures in each section are summed, not as a score, but as an indicator for comparability, and an indication of the value of transparency into dataset information. These sums are visible in the search view on HDX as at-a-glance indicators about dataset quality measures.

Phase 2 - High-level spec, requires additional research and design

Create an organization review and vetting process that allows for *trusted orgs to serve as a proxy* – or at least an additional indicator – *for dataset credibility*, and introduce org badges into the Quality Measures pane; automate QA processes where possible; begin research into domain-specific quality measures.

Phase 3 - High-level spec, requires additional research & design

Collect *additional quality measures through third-party* organization QA processes; introduce self-reporting assessment for *domain-specific quality measures*; implement dataset comparison suggestions of "similar datasets," and consider adding further measures pending user feedback from Phases 1 and 2.

Additional notes on our recommendations:

- Our approach involved prototyping on real datasets. This proved to be an essential part of our process; it enabled our work to be grounded in the particulars of the HDX context. We will follow this model for any future work with HDX.
- Based on our research and interviews, completeness, accuracy, and relevance were the quality measures that were most useful, and HDX already collects certain information that has indicators for these areas. The work in Phase 1 involves organizing this information into a legible knowledge structure, and designing it in a visible and easily accessible way. The chart in *Section VI* below (*fig. 12*) Indicates where and when this information is collected, and possible answers for each field.
- Based on the needs of HDX users and the circumstances – 20K+ disparate datasets, limited resources for additional dataset security – we determined that HDX ought to build *frameworks for measuring metadata completeness* and, additionally, seek out quality certifications (likely conducted by third parties, but could also be conducted internally by HDX).

Designs

DNP created three Phase 1 designs (a template + two dataset-specific versions), a "search" view, and a Phase 3 preview of the dataset comparison section. Full resolution designs are included in the Appendix.

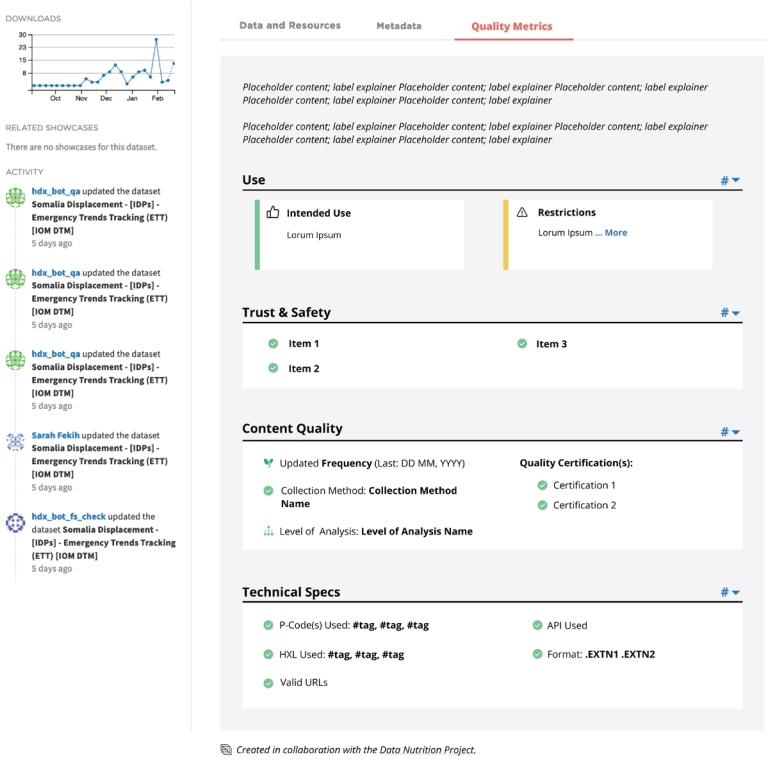


Figure 8. Phase 1 content in Quality Measures Pane on HDX site.

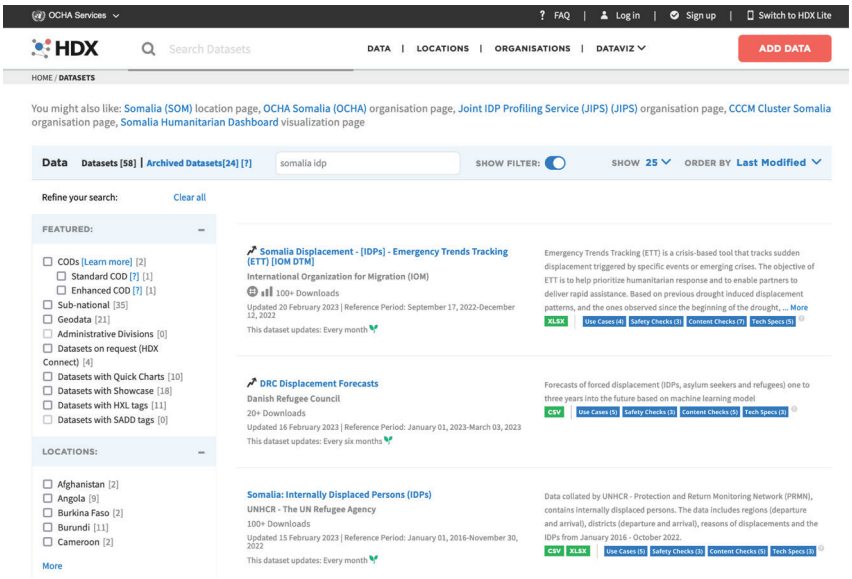


Figure 9. HDX search view with Quality Measure counts.

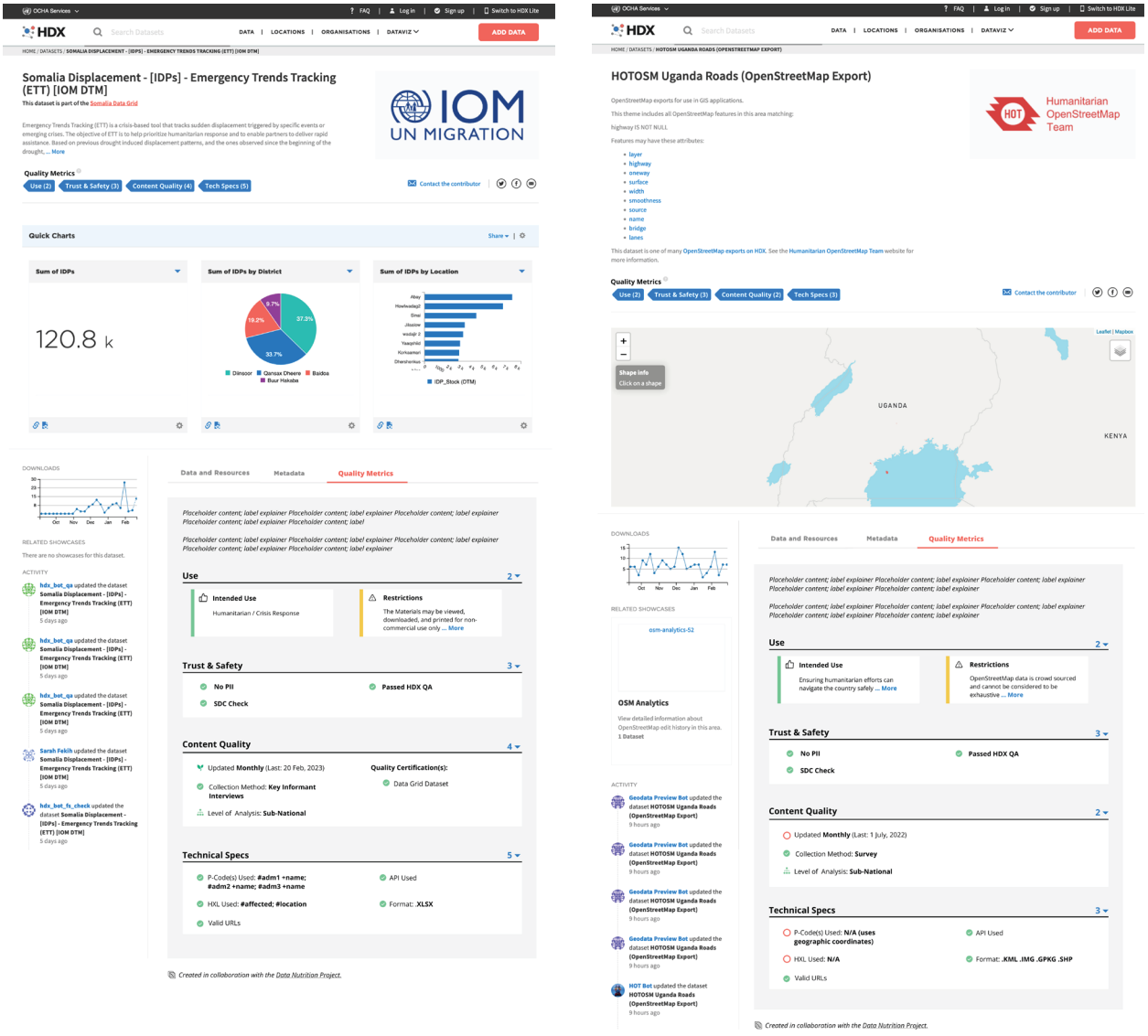
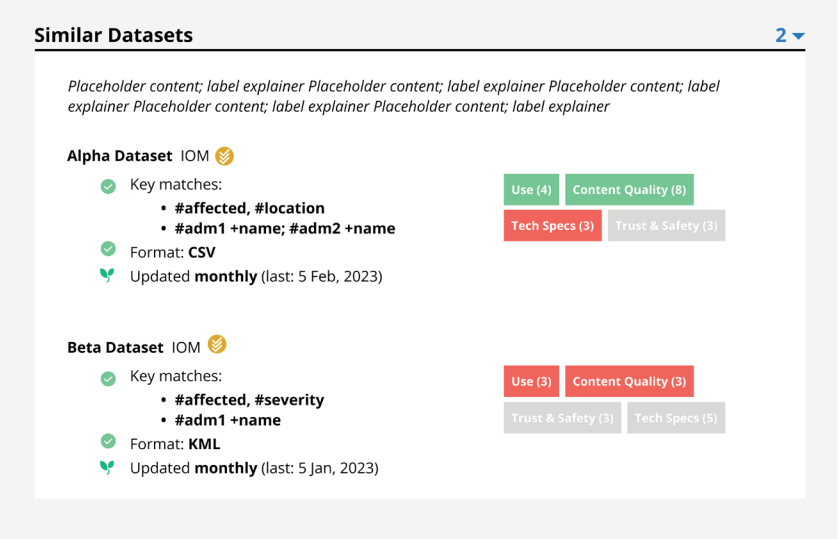


Figure 10. Views of Quality Measures Panes (Phase 1) for two example datasets.



Created in collaboration with the Data Nutrition Project.

Figure 11. "Similar Datasets" comparison sketch, to be refined in Phase 3.

VI.
ROADMAP & IMPLEMENTATION

Phase 1

Approach

We recommend creating a new *Quality Measures* pane (alongside the *Data & Resources* and *Metadata panes*) for the HDX dataset landing page. This view will consolidate existing measures from across HDX into one view that gives users a quick overview of dataset information and proxies for quality, organized into four sections:

Use | Trust & Safety | Content Quality | Technical Specs

The technical needs for this should be minimal, and implementation process could include: 1) Implementing the prototype on HDX's staging server; 2) Soliciting feedback on the prototype from key stakeholders, integrating feedback where possible; and 3) Rolling out the measures design to HDX. Because all of the information present in the Phase 1 prototype is already collected, it would be a matter of consolidating information rather than creating new information or processes. Below (*fig. 12*)is a chart of the information contained in the Phase 1 designs, and where that information comes from in the HDX workflow.

Technical Considerations

Phase 1 involves no database changes, and minimal back-end implementation. It will focus primarily on front-end implementation for HDX, and soliciting feedback from stakeholders.

- **Database/DevOps.** The data needed for the Phase 1 measures view exists within HDX already, meaning there should be no database migration/change considerations.
- **Back-end.** Implementation would be focused on making sure all of the necessary quality measures are available to the front-end website. While many of these are already available to the front-end, such as last-updated date and caveats, some things may not be readily available, such as whether a QA check has been conducted. This would involve an audit of existing data access endpoints, with the potential need to implement a few new back-end endpoints for existing information. It may also require reformatting some of the back-end endpoints to ensure the information needed by the front-end is in the right format.
- **Front-end.** Front-end work would be the most significant portion of this phase. It would include implementing back-end requests and an accessible, localizable UX.
- **Feedback.** It will be important to have a clearly defined list of stakeholders from whom to get feedback on Phase 1 designs, and clear reasons for why they are being chosen, and the kind of feedback sought. It will be important to set limitations for the scope of their feedback, so that they understand what is possible in Phase 1.

Phase 1 is intentionally a consolidation and "surfacing and organizing what is already there" phase and thus should not require substantial technical work.

Quality Section	Measure	HDX Process/source	Response Parameters
Use	Intended Use	**NEW**: Dataset Upload - Intended Use Field	Open Text
	Restrictions	Dataset Upload - Caveats	Open Text
Trust & Safety	PII	HDX QA	Yes, No
	SDC Check	HDX QA	Yes, No
	Passed HDX QA	HDX QA	Yes, No
Content Quality	Update Frequency & Last Updated date* **When these two pieces of information conflict, it should be noted as conflicting information, and the more recent one gets prioritized.*	Dataset Upload - Expected Frequency Update, Upload timestamp	Frequency - Multiple choice (Every day, every week, every two weeks, etc.) Last uploaded- Date-time
	Collection Method	Dataset Upload - Methodology	Multiple Choice (Census, Sample Survey, Direct Observational Data, Registry, Other)
	Level of Analysis	HDX QA	Multiple Choice (national or sub-national)
	Quality Certifications	Review by HDX Data Team	Badge (i.e. it is present if the review has been done, absent if not). Possible badges include: Datagrid Dataset (**NEW**, would have to be imported through a script), COD
Technical Specs	P-Code	Dataset Upload - Automatically reviewed	Badge (i.e. it is present if P-codes are used, absent if not).
	HXL	Dataset Upload - Automatically reviewed	Badge (i.e. it is present if HXL is used, absent if not).
	Valid URLs	HDX QA	Yes, No
	API Used	**NEW**: Dataset Upload - Automatically reviewed	Yes, No
	Format	Dataset Upload - Resource Upload	Select All that Apply (.csv, .kxl, .xlsx, etc. [Centre to generate comprehensive list])

Figure 12. Phase 1 Quality Measures table including source of information and response parameters.

Further recommendations

Phase 2

Phases 2 and 3 research and designs could be explored in future engagements.

Approach

Phase 2 has three specific foci: 1) determining measures for and assessing organizational trust, 2) creating capacity for longer-term quality measures development, and 3) research on domain-specific measures. For organizational trust, the goal would be to come up with an architecture for measuring an organization's data processes, which would be used as a proxy for the credibility of the organization's datasets. This would then be easily viewable as a quality indicator. The element of creating capacity for QA automation would include identifying the components of the QA process that are automatable, and implementing changes in technical workflows to create this automation. The third component of this phase involves researching domain-specific quality measures. For example, how is quality evaluated for GIS infrastructure datasets, or food security datasets? Domain-specific quality measures will add value to dataset review and may reveal potential avenues for automated assessment.

Technical Considerations

For each of the three components of Phase 2, technical considerations would vary based on the capacity of HDX. A narrative exploring the range of technical considerations (from limited capacity to high capacity) would be included after conducting further research.

Phase 3

Approach

Following on the recommendations from Phases 1 & 2, Phase 3 would enable a more comprehensive quality metrics interface. This would include implementing the research from Phase 2 on domain-specific quality measures and either 1) soliciting third party organizations to build certification processes for these metrics, or 2) proposing a self-assessment for agreed-upon domain-specific measures.

With more comprehensive quality measures, HDX will have the information available for meaningful dataset comparison. This comparison could be displayed in a section called "Similar Datasets," enabling users to quickly compare across HDX datasets along the metrics that are most important for them, and thus choose the best data for their use cases.

Rationale

In our research on Common Operational Datasets (CODs) and GIS road-mapping datasets, we discovered that there were some specific methods for quality analysis that applied to all datasets in a given domain, but not beyond that domain. CODs have their own quality metrics framework and third party certification; GIS road-mapping datasets

tend to have a standardized structure that most dataset creators use. Consequently, it seems feasible that the findings from Phase 2 on domain-specific datasets could be rolled into third party certification in some cases, or automated analysis in others (where data is already informally standardized).

And, while the measures in Phase 1 are a helpful starting point to assess quality, more analysis is needed for robust dataset comparison. The addition of third party certifications and automated analyses by domain will provide the content for such comparison. Technical considerations for Phase 3 would be enumerated after the research, development, and design for Phases 2 and 3.

VII.
ADDITIONAL CONSIDERATIONS

Impact on existing data processes and systems

Currently, data organizations upload their datasets to HDX either in bulk (using the HDX / CKAN APIs) or manually (using the upload form process). Many dataset quality measures are already collected during these processes, and *we recommend that additional information could be gathered through updating the API and the form*. Critically, our Phase 1 recommendation does not require this additional infrastructure, and instead utilizes only information that HDX already collects. In Phase 2 and beyond, there are additional automated and manual processes in which more metadata is gathered, some of which could be leveraged as quality measures. The particulars of this information - what is gathered and when, and what opportunities there are to collect more data automatically or otherwise – require further exploration.

Inclusion of quality measures in future technical projects

Over the course of our engagement with the Centre, we learned that there are several additional projects in flight that could overlap with the further collection and surfacing of quality metrics on HDX. These projects include the Data-as-a-Service work (ArgoDesign) and shifts towards workflow management of the QA process. It is an open question as to how and whether these projects could address the quality measurements initiative. As these projects get scoped further, we recommend surfacing any possible overlaps early and identifying what, if anything, can be added to more quickly enable the collection and surfacing of quality measures on HDX.

Trusted Org & Third Party Certification program definition

As outlined above, for Phase 2 would entail research for two initiatives that leverage external organizations as proxies for credibility: 1) A trusted organization program, which results in a trust score or level that is inherited by all datasets from that organization; and 2) Beginning research on a third party certification program, to be implemented in Phase 3, that enables third party validators of content quality (most likely domain expertise) to report whether a dataset meets assessment criteria. Although we have drafted some potential metadata options, including a trust level or score for organizations and an area in the UI to hold "certifications" (which could sit within the quality measurements pane described in Phase 1), further research is required to define the set of metadata and the processes of collecting that information; this would include stakeholder interviews with current contributing orgs, some of the (informally) trusted orgs, and additional third party organizations.

"Compare" feature dependencies

Phase 3 considerations include the addition of domain-specific metadata, third-party certification metrics, and the ability to compare and see additional, related datasets

on HDX. Multiple conversations with the Centre and its users highlighted the critical importance of data selection. However, the notion of comparing and contrasting datasets requires standardized metadata, some of which is already collected, but much of which is not programmatically accessible. In particular, the usefulness of the "compare" feature, which requires additional research but could appear, for example, within the "Quality Measures" tab on the dataset page or the search results returned after submitting a query – increases significantly with the inclusion of technical information about the dataset that would be made available through the HXL-ation process. This is no doubt a challenge, considering that the majority of datasets are not yet HXLated. There is an open question about how much metadata must be available on HDX datasets for the "compare" feature to be useful.

Resource identification

Each of the recommendations made in this report will require resources from the Centre and, in some cases, beyond the Centre. This echoes a common refrain in technology processes and projects about resource management, and thus requires consideration within the context of roadmap prioritization across the larger team. Phase 1 will require design and technical resources for implementation, though we have tried to scope this phase to be relatively small with respect to back-end changes. Phases 2 and 3 will require additional resources, especially for the trusted org and third party validation programs and UX / Product Management / Development resources for the "compare datasets" feature set. Additionally, any work with third parties requires not only building but also maintaining relationships over time.

VIII.
CONCLUSION

This report outlines the findings from a five-week research and design sprint, undertaken between February into early March 2023, by the DNP team, to deliver a set of implementable prototypes for quality metrics indicators on the HDX site (Phase 1). While the process included contending with some known – and some new – challenges, we are pleased to share that this sprint has concluded in a number of valuable findings, a concrete set of designs based on content that is already available on or collected by HDX, a determination of the path forward of a summation "score" that indicates *whether* the information is available, rather than an normative grade on the information itself, and prospects for future work (Phases 2 and 3). We are also delivering complete prototypes of two distinct datasets as an illustration of the information that can be conveyed in the Phase 1 quality measures pane of the HDX site.

In summary, HDX already attends to dataset quality. We found that a lot could be gained simply by consolidating disparate elements from the HDX upload and review process, and making this information readily available to dataset users. We also found that summing the information that is provided within four discrete sections (with a numerator but no denominator) enables for a gentle indicator of quantity of information without penalizing others for not having that information.

Many fields that involve data-driven decision making are only now starting to ask questions about dataset quality – questions that HDX has already answered and started to build into its systems. With a concrete, phased approach, HDX can implement quality measures for data in a way that meets its users' needs and sets an example for many other fields. DNP looks forward to continued collaboration in this process.

IX.
APPENDIX

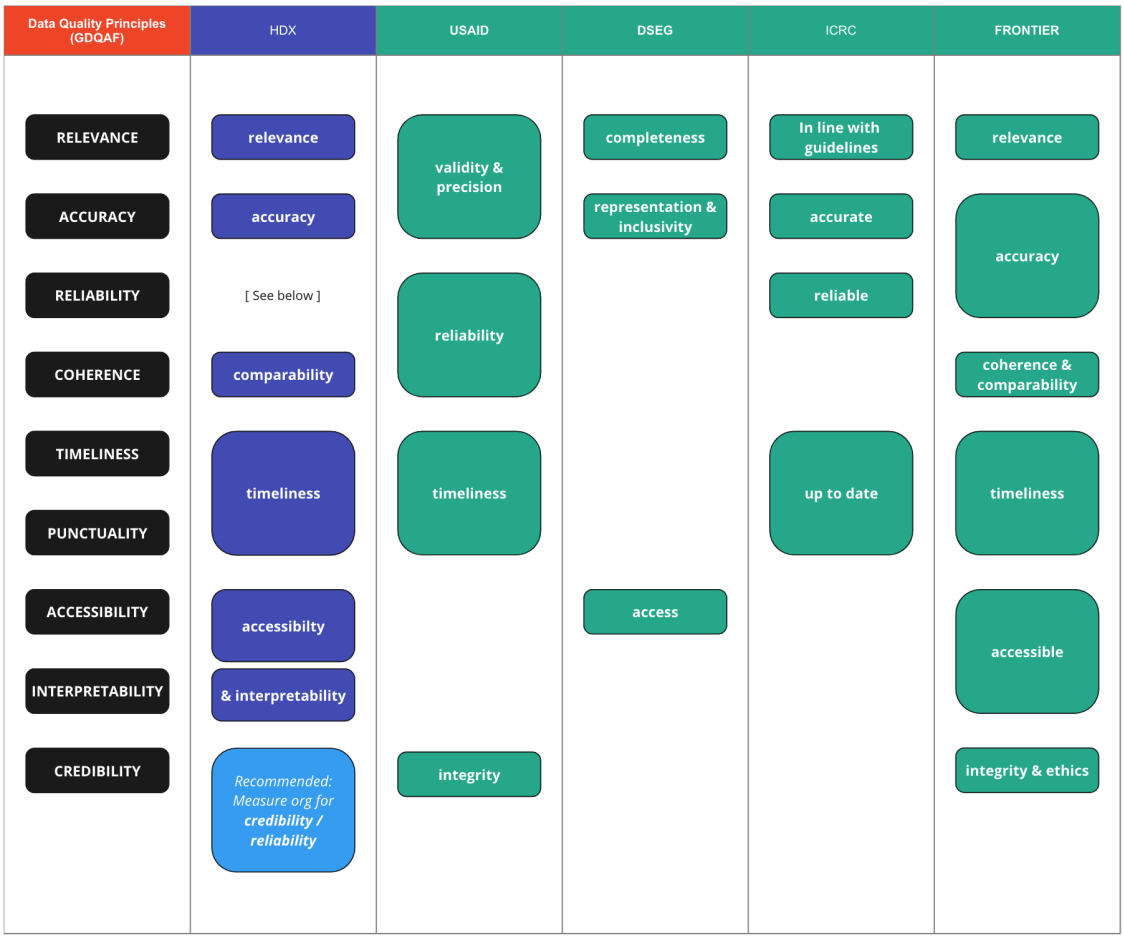


Figure 1. Matrix aligning data quality principles across several organizations, with HDX in blue. Documents cited:

a. *USAID, Democratic Republic of Congo, "How to conduct a data quality assessment (DQA): An Aid Memoir for a COR/AOR" (March 2012)*

b. *Frontier Technologies Hub, "releasing the power of digital data for development: a guide to new opportunities" (June 2019)*

c. *Data Science & Ethics Group, "A Framework for the Ethical Use of Advanced Data Science Methods in the Humanitarian Sector" (April 2020) - https://www.hum-dseg.org/dseg-ethical-framework*

d. *International Committee of the Red Cross, "Handbook on Data Protection in Humanitarian Action. Second Edition" (2020) - https://missingpersons.icrc.org/library/handbook-data-protection-humanitarian-action-second-edition*

e. *The United Nations Statistics Division, "Generic Data Quality Assurance Framework for a UN Agency" (September 2015) - https://unstats.un.org/unsd/unsystem/Documents-Sept2015/GSQAF-GenericData-Sept2015.pdf*

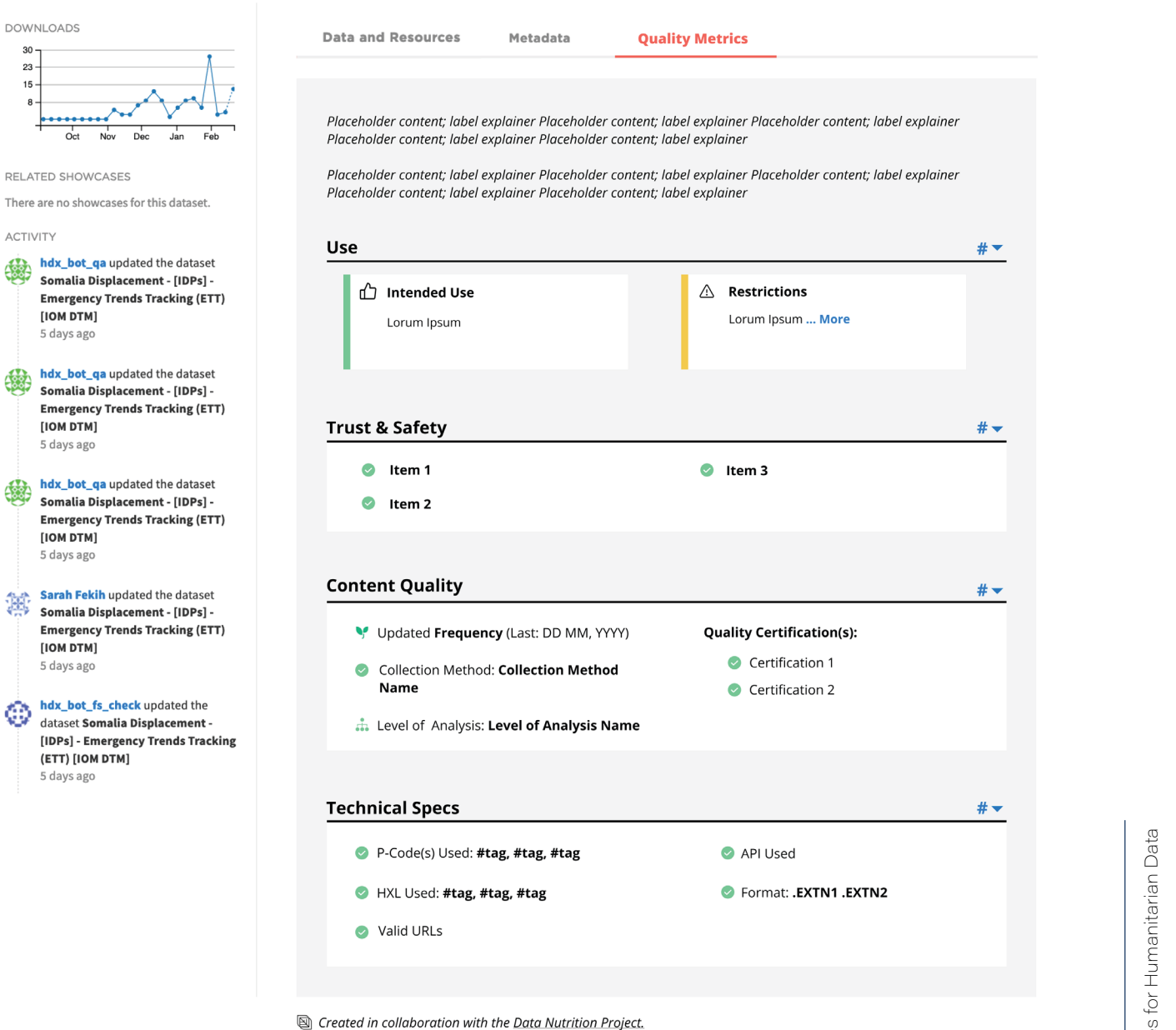


Figure 8. Phase 1 content in Quality Measures Pane on HDX site.

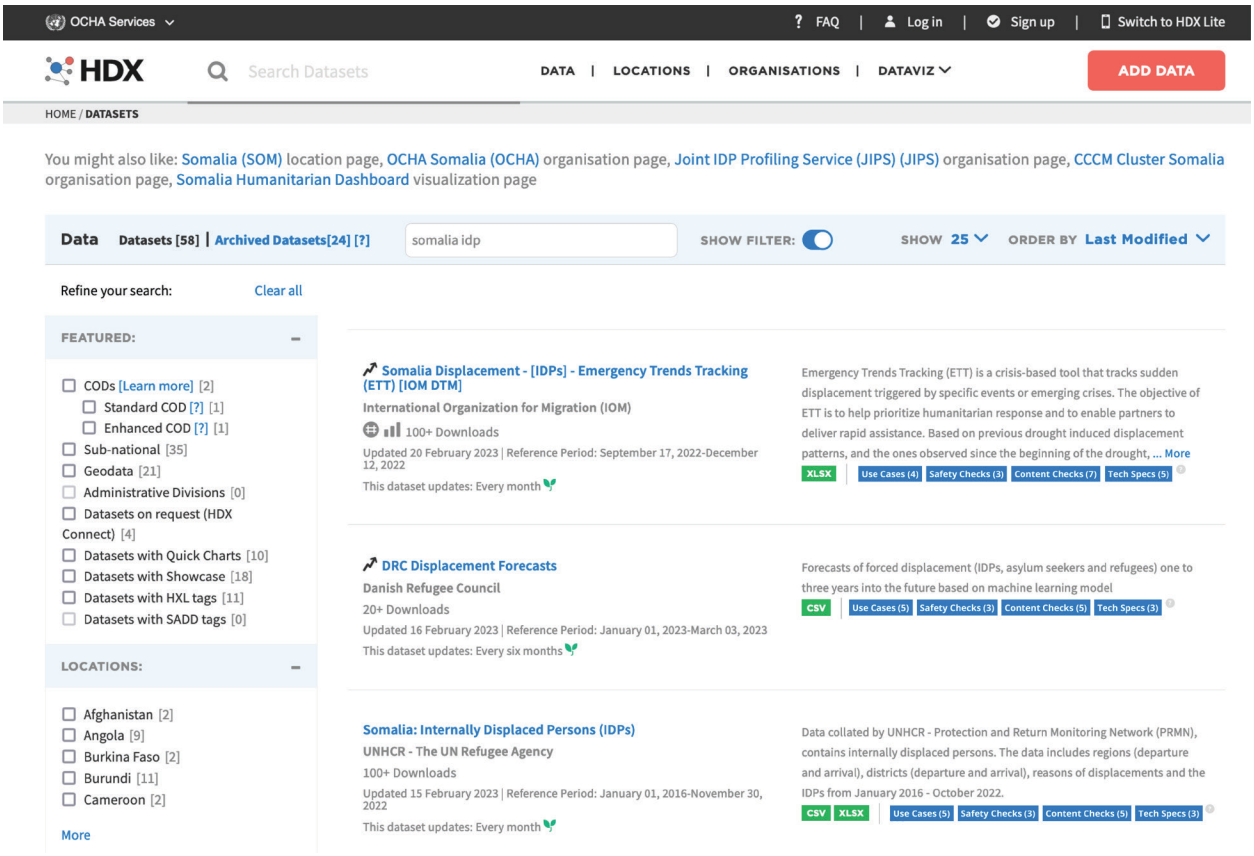


Figure 9. HDX Search view with Quality Measure counts.

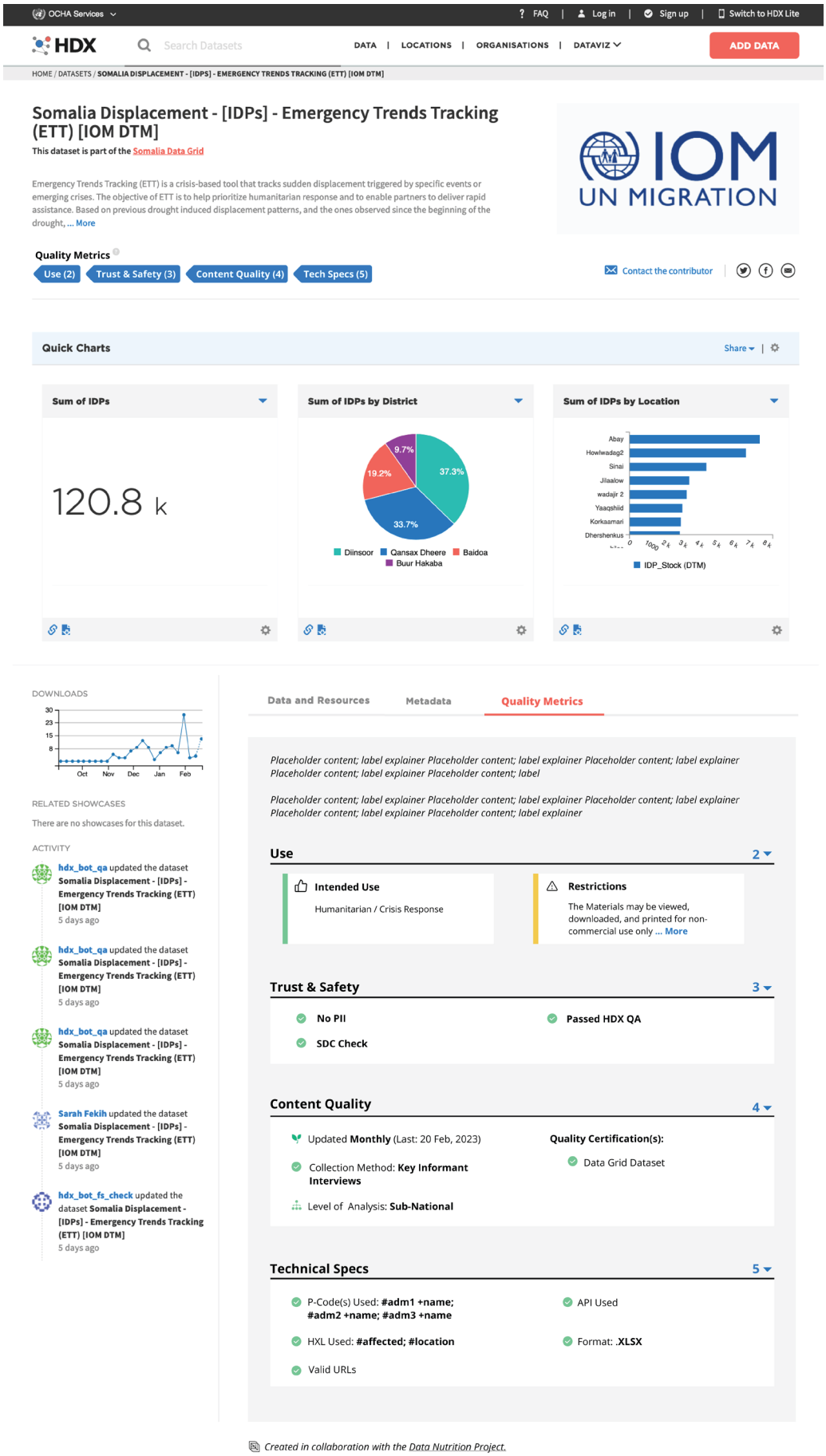


Figure 10a. Views of Quality Measures Panes (Phase 1) for two example datasets - dataset 1 of 2.

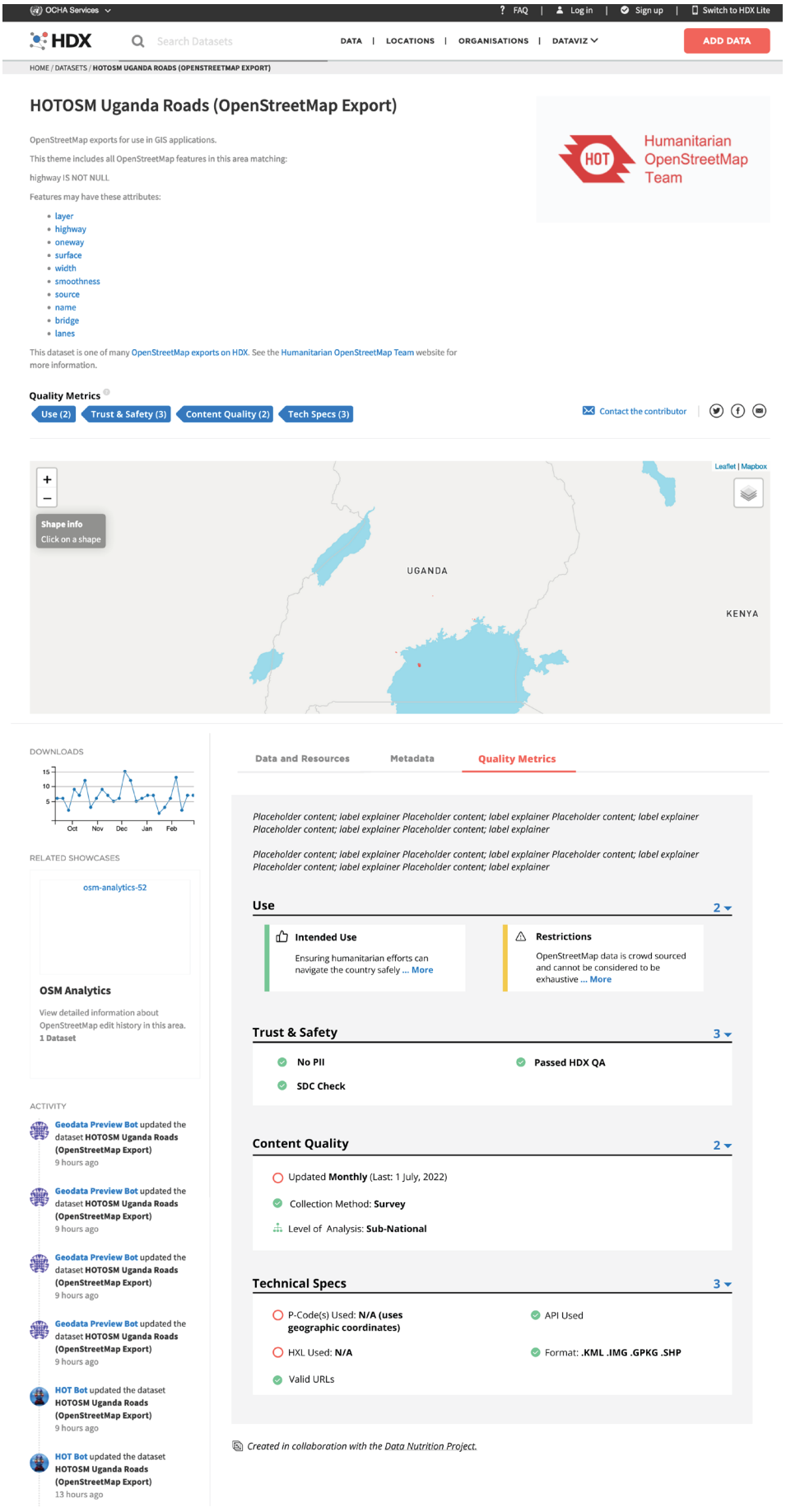
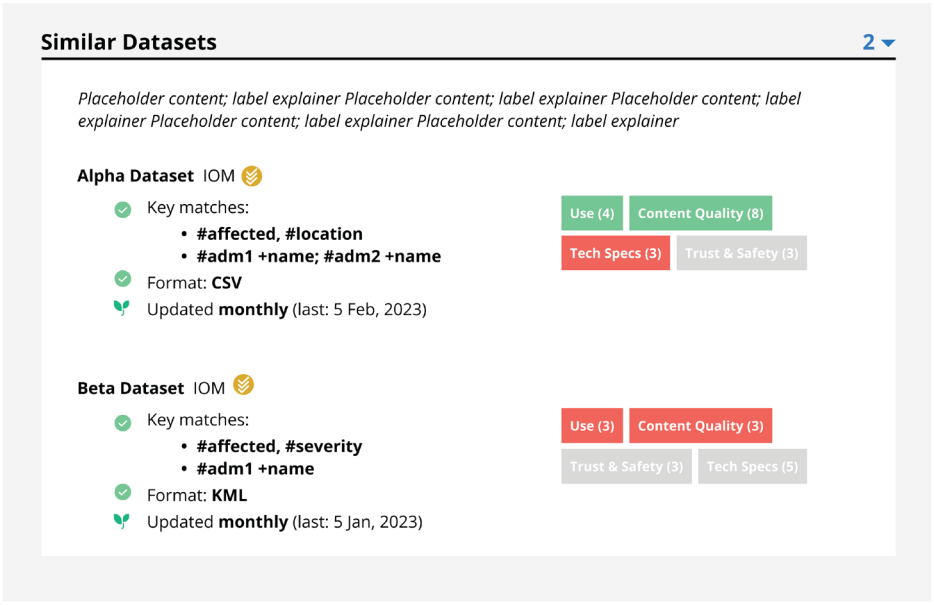


Figure 10b. Views of Quality Measures Panes (Phase 1) for two example datasets - dataset 2 of 2.



Created in collaboration with the Data Nutrition Project.

Figure 11. "Similar Datasets" comparison sketch, to be refined in Phase 3.

© 2023 Data Nutrition Project