Clearinghouse for Financing Development Data

Data Sourcing Strategy

Technical document

19 April 2021

Preamble

The Clearinghouse for Financing Development Data aims to provide a dedicated platform for aid providers and recipients to align their priorities, optimise their decision-making and demonstrate a stronger case for data and statistics. To reach this goal, the platform prioritises user-focused, transparent and accessible information about partners, projects and results, which will build on timely, accurate and in-depth data.

This document seeks to inform the Bern Network core group members of the latest developments in relation to the technical development of the Clearinghouse, in particular the proposed way forward for the data sourcing for the platform. It has been prepared by the Clearinghouse project team, comprised of PARIS21 and Open Data Watch, on behalf of the Bern Network Secretariat.

The data sourcing strategy presented in this document has been developed using a participatory approach, including desk review of available data sources, consultation meetings with owners of key data repositories, exchanges with pioneers in this field, discussions with crucial prospective users and assessments on data gaps. While maintaining scalability in the future, this strategy specifically focuses on the development of the platform's first public version, to be launched in October 2021.

In the following, we describe the data sourcing objectives, propose a detailed strategy, and an implementation roadmap to guide us while establishing the Clearinghouse platform.

Contents

1. Objectives

Data play a crucial role in the Clearinghouse platform. The platform's success will largely depend on the exhaustiveness, quality and presentation of the data. The data sourcing activities will therefore aim to:

- Collaborate with partners to create synergies and avoid duplication of data collection efforts;
- Collect, align and harmonise existing silos of data to make them accessible in one place;
- Source data where necessary, transform it into consistent and comparable formats.

With partners playing an important role in all three objectives, their engagement on and buy-in to the concept of the Clearinghouse is crucial to achieve these objectives. Therefore, in parallel to the data-related objectives, this strategy also includes an advocacy component throughout the process to promote the platform and create a community of early-adopters. A participatory governance mechanism is also included to engage with partners and ensure the quality of the data sourcing work.

2. Strategy

The Clearinghouse team proposes the following approach to data sourcing:

Data to be sourced for the Clearinghouse can be grouped into following three data categories: 1. data that is publicly available in machine-readable formats; 2. data that is not available publicly or not available in machine-readable formats; and 3. data that is not available. Figures 1 presents these three flows visually.

Firstly, the Clearinghouse platform will gather the publicly available data offered in machine-readable formats from various sources (Data category 1: available and accessible data). The platform will ingest the publicly available and accessible data after processing.

Secondly, it will source and repurpose data that is not available publicly or not available in machine-readable formats (Data category 2: available but not accessible data). The team will get in touch with various partners to source data that is not available and, if available publicly but not in machine-readable formats, undertake the required processing before uploading to the platform.

Thirdly, where data crucial for the platform is not available (Data category 3: not available data), the Clearinghouse team will collect and create augmented data. We propose a data validation mechanism that collects data only for those data gaps not filled by the previous two processes. As part of this mechanism, initial information collected from various sources will be validated and complemented by recipient countries.

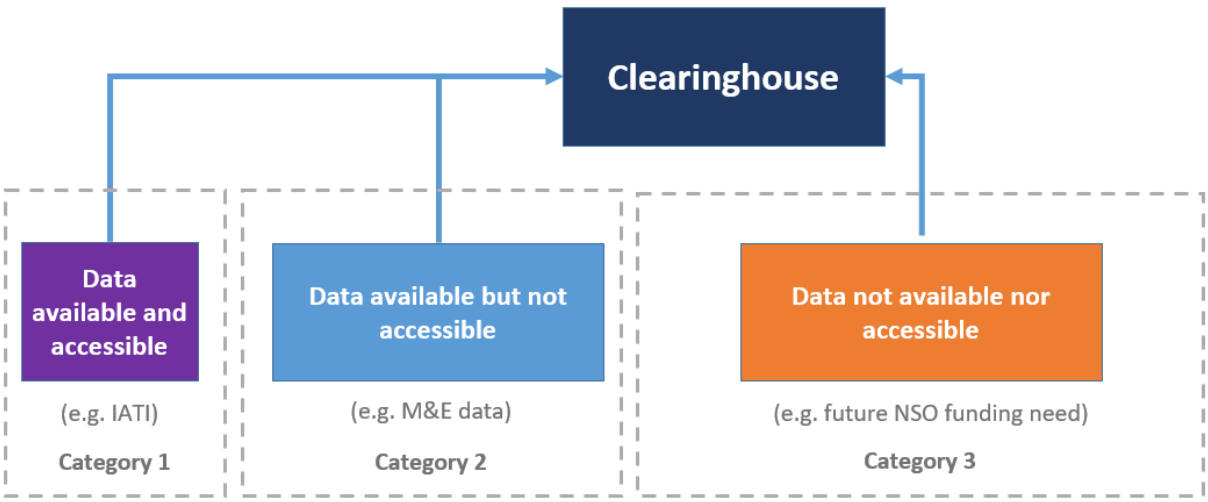

Figure 1: Data category flows; Source: PARIS21 visualisation

Further details on each of the three data categories and associated processes are presented below.

Data category 1: available and accessible data

The Clearinghouse will use data that is available and accessible in machine-readable formats, which will account for around 40% of all data in the platform. Where possible, the data sourcing process will use available APIs (Application Programming Interfaces) to automate data sourcing and ensure the presentation of up-to-date data on the platform. The Clearinghouse platform will channel and present this type of data without or with only minor processing. This category includes data already available to the Clearinghouse team, such as the Partner Report on Support to Statistics (PRESS) dataset. It also includes publicly available data from other and their sources, such as the World Bank Data Bank, UNSD SDG data base, UNDP Open Platform, Eurostat Donor Survey, the ODIN series of ODW, and the database published by IATI.

Data category 2: available but not accessible data

This category, accounting for around 40% of data in the clearinghouse, includes data that is not publicly available (e.g., Monitoring and Evaluation data of aid providers such as World Bank, IMF, Asian Development Bank), data that is available but not accessible in a machine-readable format (e.g. tables and charts in PDF documents). This category of data also includes policy documents, images and texts on websites of both providers and recipients. Here further steps are required to use the data for the platform. If data is not publicly available, the team will start a dialogue with partner organisations to set up data-sharing agreements. Bern Network members will be kindly invited to contribute to these efforts. If data is not accessible in a suitable format, the team will perform the necessary data extraction, cleaning and

formatting to harmonise the data. For valuable data that requires additional agreement to use and publish, the Clearinghouse team will follow other matured licenses and process and draft relevant documents.

Data category 3: not available data

After maximising the data that can be collected through synergies and based on partners' work, the Clearinghouse team also aims to collect a small share (about 20% of all data) but very crucial data. Taking into consideration of the survey fatigue and burdens on partners, the Clearinghouse team aims to collect this type of data via a "validation mechanism" (see Section 4). The mechanism collects data through prefilled and pre-researched surveys, interviews, and email correspondences, whichever minimises respondents' burden. Data under this category includes but not limited to the expected upcoming cost for data, demand for support based on country context and demonstrable results that can make the case for data and statistics.

3. Implementation

This section provides an overview of the implementation of the data sourcing strategy. It describes the different steps of data sourcing and gives an overview of the data sourcing techniques.

3.1. Data sourcing steps

The Clearinghouse team will follow a sequential procedure to create a coherent database for the Clearinghouse (see Table 1 below).

Different steps of data sourcing ↓	**Data category 1 Data available and accessible**	**Data category 2 Data available but not accessible**	**Data category 3 Data not available**
Collection	x	x	x
Processing and cleaning			x
Validation	x	x	x
Analysis & Data visualization	x	x	x

Table 1: Data pipelines per data category, Source: PARIS21

3.2. Data sourcing approaches

We suggest five different data sourcing approaches: a) Automatic data fetching (using APIs), b) Non-automatic data fetching (manual), c) Engagement to access data and process it, d) Extracting data from PDFs and other non-machine readable formats, e) Data Validation Mechanism. We propose to collaborate with international organizations, development cooperation providers and recipient countries on these approaches. The necessary metadata and documentation will accompany all data sourcing approaches to explain the process and data.

a. Automatic data fetching

Automatic data fetching can be applied to data category 1. The platform will use Application Programming Interfaces (APIs) to fetch the data from various sources automatically. Data sets that fall into this category include specific data from the IATI initiative and the newly published World Bank Performance Indicators.

b. Non-automatic data fetching

When APIs are not available, but data is available as structured data (e.g. Excel or CSV), bulk downloads can be a good technique (applicable to data category 1). The team will warrant that metadata is available with the data that describes the field formats.

c. Engagements with data providers

For data that is either not under public use license or requires strengthening insights from data publishers, the Clearinghouse team will work with the data owner to find the best pathway to develop insights, de-sensitise data and create added value through transparency. The Clearinghouse team has already started to contact relevant institutions and started dialogues to discuss the data sourcing exercise's scope. Table 2 describes the partners with whom the team is in contact with the type of data they own, the sourcing status, and the engagement objective. Bern Network core group members are invited to collaborate with the project team on this engagement.

Table 2: Suggested list of data providers

Partner Organisation	Data	Data category	Engagement status	Objective
World Bank	M&E Data and detailed project data	Data exists but not accessible	Discussions started	Improving tracking funding flow data and provider profiles
UNDP	Insights on project	Data available and accessible	Discussions started	Improving tracking funding flow data and provider profiles

OECD DCD	Provider profiles	Data available and accessible	Discussions started	Improving provider profiles
OECD FSD	SDG Tagging Mechanism	Data exists but not accessible	Discussions planned	Improve the SDG Tagging Mechanism

d. Extracting data from PDFs and other non-machine-readable formats

Data published as PDF (Portable Document Format) documents, often used for high-quality printing, can be read by humans, but they pose a challenge for digital systems. Similarly, data in image formats are difficult to process. Extracting data from PDFs and other non-machine-readable formats would require varied techniques. We apply those techniques for data category 2.

e. New data collection through data validation mechanism

The data validation mechanism consists of four modular surveys that would, in a first step, be prefilled through desk research/open data sources (see Figure 2). In a second step, the recipient countries and providers would validate and complement information where necessary. We apply this technique to data category 3.

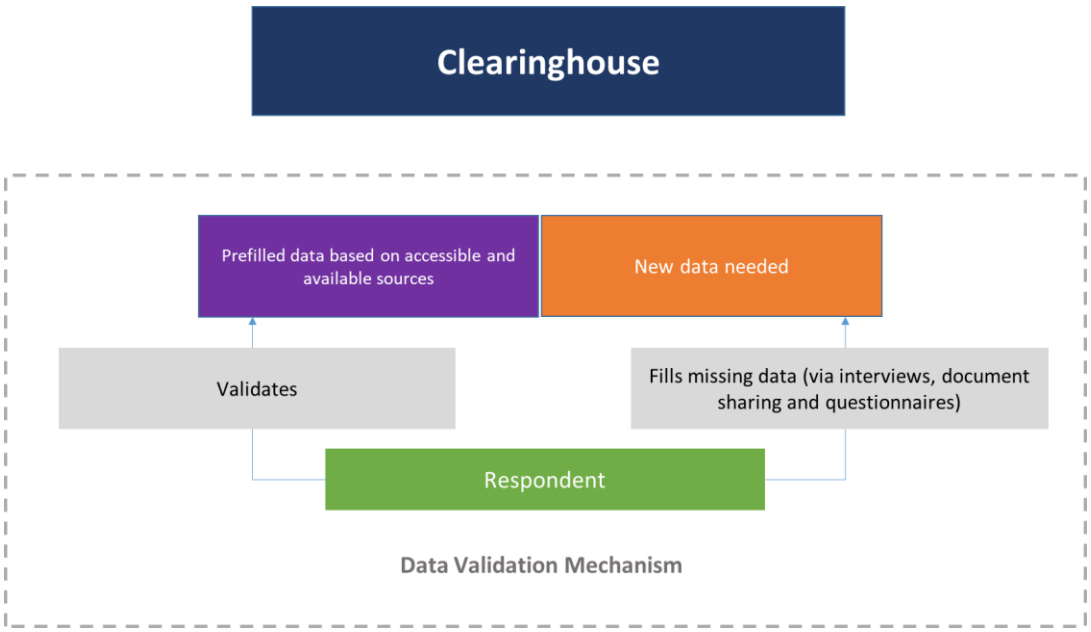

Figure 2: Data Validation mechanism, Source: PARIS21

The data validation mechanism requires careful design and piloting in collaboration with partner organisations. The team has started initial scoping interviews with national statistical offices in the African region to understand the relevance of such a mechanism for mobilising more and better funding. So far,

all countries have expressed their support for the initiative and agreed to participate in the modular surveys' initial piloting. The pilot's key objective is to reduce the reporting burden and provide a channel for respondents to voice their concerns about the validation mechanism's functions.

We suggest piloting the survey with a small number of respondents from May to June 2021 to get early feedback from partners. We also plan to complement the survey with guided interviews with World Bank country officers, FAO, UNDP and UNICEF experts in the field to validate the country responses.

provides details on the content of the different survey modules to provide sufficient data for a functioning Clearinghouse by October 2021.

Table 3: Overview of survey modules

Module	Topic	Objective	No. of questions
A	Domestic funding and government priorities	Expressed funding needs at agency level	10
B	Statistical activity at a project level	Quantify expressed funding need for ongoing and future projects. Showcase successes and lessons learned	10 per project
C	Capacity needs NSO/NSS	Uncover latent need for funding	25

Table 4 below shows the target group of each of the four surveys and modules used. Survey 1 is targeted at the NSOs of all 74 IDA countries and collects information on the domestic funding for statistics and the government priorities in that realm. Moreover, NSOs have the opportunity to report one data and statistics flagship project that they want to showcase on the Clearinghouse platform. Survey 2 represents an in-depth study of 5 selected countries. It includes domestic funding for statistics, ongoing and future projects, and a statistical system's capacity assessment. Survey 3 targets the critical institutions of the NSS of the same 5 countries. It includes information on domestic funding and ongoing and future projects but excludes the detailed capacity assessment. Survey 4 covers all DAC and Non-DAC Donors. It focuses on funding priorities and engagement channels. Moreover, providers get the opportunity to showcase successful flagship projects and share lessons learned.

Table 4: Overview of surveys

Survey	Coverage	Modules used
1	NSO of 74 IDA countries	A + B (Flagship project only)

2	NSO of 5 selected countries	A + B + C
3	NSS of 5 selected countries	A + B

4. Long-term outlook

The Clearinghouse will deliver new and unique analytical insights on data financing and be the first platform to provide this information globally, representing a significant advance for the statistical and development communities. The platform shall be scalable for potential expansion in geographic coverage and future development in thematic areas such as gender data, environmental statistics and administrative data.

4.1. Update data with flexibility

While content on the Clearinghouse will be updated to retain its time-relevance, this strategy will ensure that partners do not need to take the same reporting burden to renew the information every year. Once the first round of information is collected and country/provider/project profiles are established, future updates will be much easier using the developed data structure and parsing methodology.

Once the first round of data is collected, the Clearinghouse will allow partners directly funnelling updated data in a flexible manner based on their decision making process. Updates of this data can be carried out by a national or agency-level focal point who coordinates the data flow in its country/agency and act as the de-facto data owner. If desired by partners, the clearinghouse team can also facilitate the process or update the information for partners. A quality assurance mechanism will be integrated with the updating process to ensure the integrity of updated data.

To achieve the decentralised and partner-led update process, the Clearinghouse team will develop independent “partner hubs” for countries and agencies to allow direct funnelling of data in a flexible manner to closely connect to their respective decision-making processes. The hubs will be developed based on other aid-management platform and other clearinghouses, with simple and intuitive design that requires minimum training. Technically, the hubs will transfer granular data into the main Clearinghouse platform through built-in data exchanges between databases.

4.2. Sequencing in-depth country studies

The Clearinghouse will add further value by pioneering analyses and create augmented data to enrich the information garnered. In-depth information will be collected and generated for countries through dedicated data sourcing exercises explicitly developed for the Clearinghouse, including questionnaires, country consultations and analytical work to gather inputs from partners directly (as explained above). In

light of the workload and quality standards for these exercises, the in-depth information will be provided on the platform in a sequenced way, beginning with five countries in 2021 and then scaling up to reach all 74 countries eligible to receive IDA resources by 2025 (subject to resourcing and partner interest).

4.3. Participatory governance and technical guidance

A technical sub-group will be established to oversee the data sourcing process and provide key guidance, comprised of members of the Bern Network core group as well as other key and relevant partners. The sub-group will also work together to ensure data is collected through the validation mechanisms and the results are validated. While keeping geographic and sectoral representation of different partners in the group, membership of the sub-group willshall be on a voluntary basis for all Bern Network members and beyond as relevant. The Clearinghouse team will convene with the sub-group every two months to report on the process and results, beginning in spring of 2021. It will also solicit specific consultation and support from sub-group members bilaterally.

5. Timeline and milestones

Milestone	Date
Review of proposed data validation mechanism	April 2021
Validation by Bern Network core group	April 2021
Creation of sub-group on data sourcing and content development	May 2021
Module A, B and C developed and reviewed by experts	May 2021
Pretesting of the data validation mechanism for Module A, B and C	May - June 2021
Data collection through data validation mechanism using revised Module A, B and C	July – August 2021
Sourcing of publicly available and accessible data in collaboration with project partners and data providers	March – September 2021
Processing and analysing collected data before uploading to the Clearinghouse	August – September 2021
The official launch of the Clearinghouse Platform	October 2021