Workshop on Firm-level Data

collected by Development Finance Institutions

Hosted by the *Centre for Global Development*

Tuesday, April 6th 2021 from 9:00 AM to 11:40 AM EST

Event Summary

Table of Contents

1. The Private Sector Development Research Network

The Private Sector Development Research Network (PSDRN) is a community of institutions with an active research agenda on Private Sector Development. The network's founding objectives have been to facilitate information exchange and collaboration among network participants in order to advance understanding and knowledge on Private Sector Development that is operationally relevant.

The PSDRN annual conference serves as the network's biggest event and has one theme of broad interest to its members each time. Institutions sponsoring the events are CDC Group; International Finance Corporation (IFC); International Growth Centre (IGC); London Business School Wheeler Institute; European Bank for Reconstruction and Development (EBRD); Center for Global Development (CGD); Interamerican Development Bank (IDB Invest); and Overseas Development Institute (ODI).

2. Topic and structure of the Workshop

The focus of this workshop was on private sector data collected by Development Finance Institutions (DFIs), their value-added, their use for analytical purposes and new methodologies with the potential to resolve challenges with their collection and use. Workshop participants consisted of DFI staff, users of DFI data, academics and other external experts with insights and inputs on methods.

The workshop was structured around two main sessions, including a presentation of DFI data and a panel discussion, which was followed by an optional technical session on methods (see Annex for detailed agenda). The workshop's first session, 'DFI data and applied research,' discussed data that is not commercially or publicly available for analytical purposes, such as information on owners and sponsors of private firms DFIs invest in, risk assessments of investments or end-beneficiary (e.g. SMEs) data collected. Commonalities and differences in the information collected across institutions were an important part of this discussion. Example uses of data were also shared during that session, as well as promising extensions which DFIs can work towards. The workshop closed with a plenary discussion, 'Challenges and Opportunities in making better use of DFI Data,' a high-level discussion focusing on challenges in sharing and using DFI data. The workshop also offered an optional working session, 'Methods to codify, digitize and store data, and make it usable.' This was open to any participants interested in discussing the latest methods of digitizing and storing information from various formats (e.g. project documents, internal databases, memos and the internet), as well as methods used to process large volumes of data, such as machine reading, machine learning or other artificial intelligence tools.

3. Synthesis of interventions

The workshop opened with remarks from Mark Plant, Chief Operating Officer of the European branch and Co-Director of Development Finance of the Center for Global Development, who emphasized data and evidence as the number one constraint in analytical work on private sector development. Through their operations and analysis, DFIs collect information that could generate useful knowledge not just for the institutions, but for the development community as a whole, added Plant, and he invited participants to discuss challenges and opportunities to make that happen.

3.1 Presentation session: DFI data and applied research

The objective of the first session of the workshop was to understand what data exists within DFIs and learn more about their applications. The session was moderated by Neil Gregory, Chief Thought Leadership Officer at IFC, and included presentations by Çağatay Bircan, EBRD; Aneese Lelijveld, CDC Group; Camilo Mondragon-Velez, IFC; and Patricia Yanez-Pagans, IDB Invest, followed by Q&A with workshop participants.

Neil Gregory opened the session by briefly discussing the importance of DFI data and highlighting the need to facilitate their use internally and externally. While acknowledging the challenges in accessing and using DFI data, Gregory focused on the unique opportunities they represent, including for further collaboration and research.

Çağatay Bircan started off by giving an overview of the data that EBRD collects. This included financial and non-financial data for both clients and indirect beneficiaries (e.g., investee companies of PE/VC funds EBRD invests in), as well as for cancelled investments. Harmonizing such data across clients and time can be a challenge (e.g., due to varying accounting standards). A third type of data recorded at EBRD is around internal organization and human resource aspects. Bircan shared three examples of how such data is used in their research. First, a study on quantifying value creation in EBRD's private equity funds relied on thousands of textual records reported by funds. This was linked to external datasets to measure economic outcomes and fund returns to Limited Partners. Second, a study explored whether women faced a promotion gap at EBRD. The study used personal records of EBRD staff to uncover determinants for promotions, finding that project assignments are a significant factor. These results are helpful in learning how to build optimal teams at EBRD. Third, joint research with the Turkish Central Bank sought to uncover whether banks participating in an EBRD support program increased lending to women entrepreneurs. Bircan wrapped up by outlining the advantages and disadvantages of DFI data. The former include the non-public feature of the data that is otherwise not available to academics, and the connection to the real world that can be established through in-house interaction with practitioners. The latter include difficulty in making attribution and inconsistency in internal methodologies and data which make comparisons difficult.

Aneese Lelijveld outlined the three types of data that CDC collects: investment data (e.g., returns, ex-ante impact scores); firm-level data (e.g., financial statements, impact data) for both direct and indirect beneficiaries; and thematic data round gender and climate change (e.g., commitments, carbon footprints). Data at CDC is not centralized, which render access difficult across teams. Non-sensitive data (e.g. basic project data) is shared publicly but firm-level confidential data (e.g. financial statements) can only be shared with non-disclosure requirements with trusted advisors with an established relationship with CDC. Sensitive investment data (e.g., investment returns) require higher level permissions and need to be anonymized before sharing. CDC data is typically used for reporting and internal learning, evaluations and insights, and to contribute to wider research. Potential extensions around CDC data include: (i) modelling to estimate gaps in data (e.g., in carbon footprints); (ii) creating standardized legals (e.g. on gender standards) to improve data accessibility; and (iii) building existing tools to better understand the limitations and demands of DFI data.

Camilo Mondragon-Velez listed the features of firm-level data captured at IFC as: (i) Client (and PE investee) firm characteristics; (ii) Financial statements; (iii) Operational data; (iv) Development impact indicators; and (v) ESG and ownership variables. This data is used internally to analyze financials (risk measurement, valuation, credit performance, etc.), ESG measures (risk identification and monitoring) and development impact (benchmarking, ex-ante assessment, monitoring) of IFC investments. Furthermore, this data is also leveraged for applied research outputs, both on sector specific (e.g., SME finance) and cross-sector themes (e.g., impact investing). To expand data collection and application efforts, IFC is currently undertaking efforts to expand its Anticipated Impact Measurement and Monitoring (AIMM) indicators; to understand the trade-off between development impact and financial returns to optimize its portfolio approach; and to leverage artificial intelligence to extract more information around development impact risks.

Patricia Yanez-Pagans highlighted that IDB Invest collects firm-level data around four dimensions: financial and economic indicators, development outcomes, ESG sustainability, and additionality (both financial and non-financial). Internal use of this data for operational purposes include conducting a textual analysis on project documents to automate lessons learned, classifications and uncovering their relationship to project characteristics. These findings are then subsequently shared with staff leading new projects to highlight potential concerns. Data is also used to conduct impact evaluations which are often published in collaboration with academics. Examples of such work include studies on loan defaults, benefits of branchless banking and impact of digital credit on SMEs. Challenges around firm-level data collected at IDB Invest include the need to improve internal capacity to anonymize data to facilitate its sharing. On the other hand, opportunities exist in supporting clients to strengthen their data systems and in providing non-financial additionality through in-depth evaluations, with both leading to better data collection.

A lively discussion followed the presentations, with participants focusing on additional features data collected as well as the broader approach taken to collection and applications. Participants were particularly interested in understanding data collected for non-client firms for comparison purposes. The four presenting DFIs clarified that they do not collect non-client firm data — though other means available are used to produce counterfactuals. EBRD also examines data on indirect beneficiaries and cancelled investments, while IFC relies on existing databases (such as Orbis and WB databases) for that purpose. Participants were also interested in understanding how gender disaggregated data was collected. Such data was mostly collected around board composition and employment at client firms (and sometimes for investee firms), although gender data collection efforts were primarily project specific. Participants also queried about approaches taken in deciding which projects are chosen for in-depth impact evaluations. Speakers mentioned a combination of top-down and bottom-up approaches, with IDB Invest placing more emphasis on projects that are innovative, large or representative, while EBRD and IFC indicated greater influence of the practicality when conducting such exercises and demand from operational teams.

3.2 *Panel Discussion: Challenges and Opportunities in making better use of DFI Data*

The panel discussion, moderated by Nancy Lee, Senior Policy Fellow, Center for Global Development, included interventions from David Atkin, the Abdul Latif Jameel Poverty Action Lab (J-PAL) and Massachusetts Institute of Technology; Shaida Badiee, Open Data Watch; Erik Berglöf, Asian Infrastructure Investment Bank; and Anastasia Gekis, IFC

Nancy Lee introduced the panel and the topic, emphasizing at the outset that DFI data are a public good. Lee argued that it is therefore imperative for DFIs to make their data available, as publicly funded institutions. Such action would be to their benefit too, as DFIs could extract greater development returns from these existing assets by further understanding the characteristics of firms with greater development impact, as well their needs. The long-term goal, according to Lee, should be a universal database jointly held by DFIs, although the realization of such a collective project would require solutions for anonymizing data and respecting confidentiality.

In his intervention, **Erik Berglöf** stressed the importance of data as a tool to enable development progress, and highlighted the Global Emerging Markets (GEMs) Database as a successful illustration of making valuable data publicly available, whilst respecting confidentiality and ensuring sufficient anonymization. A recent project at the Asian Development Bank aims to build a database with geospatial information, which the bank aims to make widely available.

David Atkin introduced fruitful research questions that could be addressed using DFI data and discussed the type of data required from DFIs for this purpose. In terms of research questions, Atkin highlighted opportunities to measure the impact of credit provision on firms with respect to performance and balance sheet metrics, but also in terms of other impacts that are not easily captured by balance sheet information or other financial statements. In addition, research could assess the indirect impacts of interventions on other firms that do not receive funding but engage with DFI clients. Suppliers of DFI funded firms or other market participants might benefit from DFI intervention, though currently little is known about this channel of impact. In the absence of RCTs, granular information on firm-level characteristics could improve the matching of similar firms to assess interventions against a counterfactual. Internal scoring systems, as well information on how exactly credit decision are taken, could be leveraged for Regression Discontinuity experiments or Instrumental Variable designs.

During the discussion that followed, Atkin encouraged DFIs to draw on the support of JPAL in designing identification strategies and RCTs for their operations and highlighted lessons learnt from engaging with client firms. Two strategies have proved successful in his experience: emphasizing the benefits to the firm from sharing data, for instance in terms of learning and improving operations or with respect to attracting impact funding; and highlighting the overall benefits of research, as firms, in Atkin's experience, tend to be open to such conversations.

Shaida Badiee highlighted that open data is both the responsibility of public institutions and a benefit to institutions themselves. The Open Data Project that Badiee initiated and supervised at the World Bank, was aligned with the Bank's transparency policy and had a positive impact on its reputation. Importantly, the project prompted governments and other organizations to follow suit and led to new technical assistance projects in supporting client countries. Common barriers to the development of open data initiatives, according to Badiee, include the high upfront costs of dealing with legacy systems, or the loss of revenue where data was sold. The most crucial success factor for realizing open data policies at the country and organizational level, however, is buy-in and commitment by leadership. The strategic importance of data in development is not reflected in current ODA numbers, where a doubling of investment in open data support would be warranted.

In the discussion that followed, Badiee highlighted specific steps that have been helpful for often reluctant, large organizations to operationalize an open data policy. 'Open by default' models set negative lists of data that should not be publicly available, which often is operationally easier for organizations deciding what should be published. To create support within organizations, institutions need to generate external demand through advocacy with stakeholders. Additionally, a review and revision of the terms of use for the data held, is essential to allow for expansion of their use.

In her intervention, **Anastasia Gekis** described the ongoing data strategy process at IFC, highlighting that the basic premise for the initiative is to leverage data for better decision making, to generate higher development impact. The common goal of the development community, to realize prosperity for all, can only be realized if the value of data is unlocked and the resulting insights are shared across organizations with the proper safeguards in place. The current strategy process at IFC aims for a balanced approach – disciplined and growth-oriented – to serve its ability to support internal decision making with the broadest range of data and analytical tools available. To support this broader goal, IFC is also strengthening its sourcing and governance processes, as well as further developing the skills and analytical capabilities that are characteristic of a data driven organization. IFC experience acknowledges that operational hurdles in realizing a data strategy need to be addressed and a coordinated approach will be important for success.

3.3 *Technical Presentation Session: Methods to codify, digitize and store data, and make it usable*

The objective of this working session was to discuss the latest methods of digitizing and storing information from various formats (e.g. project documents, internal databases, memos and the internet), as well as methods used to process large volumes of data, such as machine reading, machine learning or other artificial intelligence tools. Moderated by Imtiaz Ul Haq, Economist at IFC, the session included presentations by Daniel Björkegren, Brown University; Atiyah Curmally, IFC; and Jonathan Hersh, Chapman University.

Daniel Björkegren presented a study on the Digital Adoption Program that took place in Rwanda in 2008 involving a subsidy for rural households to buy handsets on credit. The focus of the presentation was on data needed to assess progress on what happens in a program, what its effects are, and ultimately how to design programs with a view to monitoring impact. The analytical approach included matching national development bank data statistics on target areas, with data on mobile phone adoptions from household surveys and transaction data to identify where the devices were activated. Björkegren demonstrated that, conditional on the availability of rich data sources on intensity of use and spending, one can also evaluate changes in use with variation in prices and variation in network penetration. Projections can give answers to a wide range of actionable questions on target populations for such programs, their effects and spillovers. Rapid feedback from digital data can be particularly useful in adjusting programs on the ground.

Atiyah Curmally presented an extensive dataset on ESG analytics that IFC has built using historical data the organization has tracked in more than 11,700 documents going back 13 years. The organization strategically used artificial intelligence in order to develop analytics that support sustainability due diligence, in addition to financial monitoring of investments.

MALENA – the project's name which is short for Machine Learning ESG Analyst – made use of a Natural Language Processing Algorithm, a machine learning tool trained to detect institutional labels for nearly 600 risk terms, using sentiment analysis. Information from this analysis feeds to a user interface that aggregates it in an easy-to-use platform for staff. The technology powering MALENA includes artificial intelligence (Google BERT), a Data Science Platform that can be used for other applications/objectives, MS Azure Cloud, and standardized training tools for ESG term identification. The project continues to expand with the objective of extracting inferences for IFC operations, while addressing challenges with client confidentiality, training data bias and interpretation.

Jonathan Hersh opened his presentation with a discussion on why development data projects fail, arguing that the answer often lies in inconsistent strategies. Hersh stressed that models can only be as good as the data pipeline used to deliver analytics. To succeed, one needs to build a consistent 'data lake' first, and then all data and models should be accessed using Applied Programming Interfaces (API). An organizational culture that fosters innovation and openness is also important to facilitate adoption of consistent strategies in data projects. In an effort to identify promising AI tools for development organizations, Hersh pointed to deep learning, a recent development from the last 10 years. Modern data pipelines need to move from raw data to analysis and output through successive transformation – a number of programming languages and open sources can be used for that purpose. To illustrate, Hersh presented an application to analyze information on ongoing violence in Syria, where a deep learning model was trained to recognize the destruction of buildings from Google Maps imagery, using the data augmentation approach to expand training labels, over 6 Syrian cities. Using satellite pictures, the program can gather information on war impact over time and, in principle, over many other places.

In the discussion that followed, Curmally highlighted the increasing returns of building an infrastructure for related analysis, which is ongoing at the World Bank Group, acknowledging that many smaller organizations may find it costly. There is a role for a large development institution to make this infrastructure more widely available. Hersh highlighted the labor-saving potential and the view of these technologies as investments that save for the future. Making better use of promising new tools, such as label smoothing by weighing information based on quality and trust of data, has also been highlighted by presenters as promising avenues to improve the accuracy of results from these analyses. Björkegren added that, the more one knows about the context and the features represented in the data, the better the design of algorithms. Context allows us to extract more information from less data. A conversation followed, on overcoming the confidentiality constraints leveraging new methods where presenters highlighted the importance of sound administration and processes used in the data and classifications of data in layers, as well as tracking usage.

Detailed Agenda

9:00 AM –
10:30 AM

Opening remarks

Mark Plant, Director of the CGD Sustainable Development Finance Program and COO of CGD Europe

Presentation session:
DFI data and applied research

Moderator: Neil Gregory, Chief Thought Leadership Officer, IFC

Presenter: Çağatay Bircan, Senior Research Economist, EBRD

Presenter: Aneese Lelijveld, Evaluation Executive, Impact Group, CDC Group

Presenter: Camilo Mondragon-Velez, Principal Research Economist, Modelling & Analytics - Sector Economics & Development Impact, IFC

Presenter: Patricia Yanez-Pagans, Lead Economist, Development Effectiveness Division, IDB Invest

Q & A session

10:30 AM –
10:40 AM

10-minute break

10:40 AM –
11:40 AM

Plenary session:
Challenges and Opportunities in making better use of DFI Data

Moderator: Nancy Lee, Senior Policy Fellow, Center For Global Development

Panelist: David Atkin, Co-Chair, Firms, Abdul Latif Jameel Poverty Action Lab (J-PAL); Professor of Economics, Massachusetts Institute of Technology

Panelist: Shaida Badiee, Managing Director and Co-founder, Open Data Watch

Panelist: Erik Berglöf, Chief Economist, Asian Infrastructure Investment Bank

Panelist: Anastasia Gekis, Manager, Operations Management, IFC

Q & A session

WORKSHOP CLOSE
OPTIONAL:

12:30 PM -
2:00 PM

Working session:
Methods to codify, digitize and store data, and make it usable

Moderator: **Imtiaz Ul Haq**, Economist, IFC

Presenter: **Daniel Björkegren**, Assistant Professor of Economics, Brown University

Presenter: **Atiyah Curmally**, Principal Environmental Specialist, IFC

Presenter: **Jonathan Hersh**, Assistant Professor of Economics and Management Science, the Argyros School of Business, Chapman University

Q & A session