Fact Sheet Technical Tool Type

Data Sources (incl. Open Data and Big Data)

Source: istock/Laurence Dutton

Consultant comparing baseline data from various sources on a screen for a feasibility study

Relevance of this Tool Type within the Project Cycle

Data sources can be helpful to acquire and analyze existing data on a project's target population and areas at low cost. This opportunity is frequently overlooked during feasibility studies that often solely focus on conducting assessments while comparable data already exists.

Open data are of particular interest as it is usually available online at no cost of acquisition. This can lower costs considerably and afford stakeholders rapid assessment, instead of having to rely on time-consuming and costly primary data collection on the ground.

Big data (i.e. data sets that are too large or complex to be dealt with by traditional data-processing application software) have gained importance with the increasing use of smart devices, sensors, the Internet of Things, Artificial Intelligence (AI) and social media platforms, which constantly produce data without their users necessarily being aware of it. This is useful in cases where large amounts of individual user data collected by network providers – such as on user movements, communication, and payment activities – are needed. This can help to quickly identify crisis situations based on the movements and messages of large numbers of people.

Definition

Data sources can be proprietory (collected and owned by a private entity with no availability to the public), public (collected and owned by a governmental institution with limited access and editing options), or open. Open-data sources are publicly available datasets on all sectors and geographical areas – some of them without restrictions on their use and sharing, some of them restricting commercial use. For open data licensing types
> Legal aspects.

Step 1: Check the Digital Principles

Before selecting adequate data sources and designing access to and use of these, the nine Principles for Digital Development should be considered: www.digitalprinciples.org

Step 2: What Information Do I Need?

Given the wide range of existing data that can be obtained, various project tasks can be supported by secondary open data:

- ✓ Input to country and sector strategies
- ✓ Project targeting strategies
- ✓ Feasibility studies
- ✓ Socio-economic, peace and conflict,do-no-harm, environmental assessments
- ✓ Remote monitoring (e.g., comparison of project outcomes with government survey data)
- ✓ Outcome and impact assessments

a) Structured open datasets

Structured open datasets refer to organized data with clear relationships between data points that often come in the form of tables. Datasets are available on any topic, for instance demographic data, socio-economic data, geographical data, environmental data, and governance data—for example electoral data and polling data.

Step 3: Where Do I Find Structured Open-Data Sources?

✓ **Metasearch for datasets:** Google offers a free keyword search for datasets on any topic from a range of public and open sources online that allows setting filters according to the latest update, file type, user rights, subject, and acquisition costs.

✓ **Direct download from institutions:** International organizations and academic and government institutions publish their datasets online and allow free downloading. Examples for different types of data include:

✓ **Platforms for open datasets:** Another way to receive a curated overview of reliable data sources is through platforms agglomerating research and social impact data > Links to further sources

Demographic data

- World Pop by University of Southhampton (Spatial demographic data)

Socio-economic data

- Multiple Indicator Household Cluster Surveys on women and children by UNICEF
- Living Standards Measurement Study by World Bank (Household survey data)
- World Development Indicators by World Bank
- Joint Monitoring Programme by WHO and UNICEF (WASH data)

Geographical data

- ASTER Global Digital Elevation Model by NASA and Government of Japan (Elevation data)
- EarthExplorer by U.S. Geological Survey (Satellite imagery)
- ESA Sentinel Hub

Environmental data

- FAOSTAT by FAO (Agricultural and environmental data)
- EarthData by NASA (Geographical and atmosphere data)

Governance data

- Global Barometer (public opinion surveys) by academic and not-for-profit organizations
- Worldwide Governance Indicators by World Bank (Index and data on good governance)

b) Structured microdatasets

Microdata refers to individual responses to surveys by national authorities, research institutions, or nongovernmental organizations (NGO) and can give detailed information on a target population. It contains sensitive personal data and is therefore not openly available, butprovided as aggregated data for publishing.

Where Do I Find Microdatasets?

To obtain full microdatasets an application is necessary, including a description of the designated use. Access is usually granted under certain restrictions on sharing and disaggregation of information. For example, if the data is spatially aggregated up to the municipality level, further disaggregation in other dimensions, for example, gender, may be prohibited. Access to microdata can be requested from the World Bank, IPUMS International, or directly via the respective national statistical bureau. Further, **online platforms** that allow for individual evaluations based on microda-ta without accessing full datasets are also available, for example, STATcompiler.

c) Unstructured datasets (including big data)

Unstructured datasets refer to either unorganized and/or large volume datasets that need specialized software for analysis. The two main sources of big data for projects are:

- **Social data:** user data from social media platforms and the global system for mobile communication (GSM).
- **Machine data:** produced by scanning textfiles from archives or using industrial equipment, sensors, and smart meters (Internet of Things).

Where Do I Find Unstructured Data/Big Data Sources?

Social data from social media can be procured from social media providers. This data can provide more accurate data on the actual number of inhabitants of a refugee camp and their needs and problems. **GSM** data needs to be acquired through mobile network providers. The GSM Association offers some open-source insights for all world regions. Few openly accessible sources for **sensor data** within cities and universities exist
> Fact Sheet Sensors.

Step 4: Which Tools and Methods can I use for data processing?

Processing of big data and unstructured datasets requires programming expertise and knowledge of methods to make use of relevant information:

- **Data mining, machine learning and artificial intelligence** are used to find and summarize relevant information that may be unknown or hidden in large datasets. Data mining techniques can be conducted using the open-source programming languages R or Python.
- **Data visualization tools** can help to understand and structure data. Data analysis can be programmed using Python, but most tools are commercial, for example, Tableau or Power BI.
- **Social network analysis and visualization** is helpful to examine the relationships and structure among individuals, groups, and organizations within a specified network for a project. Data for this approach can be collected via surveys and from social media. Open-source tools for the visualization of connections within a given population include Cytoscape, Gephi, and Visone.

Interoperability requirements

Most available structured data sources and microdatasets provide valuable insights alone and can be downloaded or are provided in several data types, including common formats such as CSV/XLSX and XML/JSON files. These can be analyzed in every analytical software, including Excel, SPSS, STATA, R, or any database or programming environment. In cases where a combination of multiple sources and/or a comparison with project data is necessary—to evaluate the impact of a certain project—the importing of both datasets (or more) into the same tool is required and facilitated by using compatible data file types (see above). Sometimes the data requires some adjustments like the unification of categories (m/f/d vs. male/female/diverse). For unstructured and big data, interoperability is less of a concern since no direct relationship to project data exists in most use cases. If this is necessary, expertise in methods to standardize data formats and the use of specialized software as mentioned before is required.

Complexity of Use Cases for External Data Sources

The various data source use cases possess four levels of complexity:

1. Basic information based on **easily consumable data**
2. More complex information based on **moderately processed data**
3. Complex evaluations and application **of statistical/AI models**
4. **Advanced AI models** and data pipelines

While the higher levels usually offer the more impressive results, the lower levels are comparably easy to apply and provide benefits within a short period of time. The first group addresses average practitioners capable of using standard office software. They may benefit from information directly provided as XSLX or CSV tables or that is displayed in intuitively designed dashboards. The second group encompasses data that is provided in more complex formats (XML, JSON, APIs, query-languages, geodata) or that requires further processing (transformation or computation of statistical figures, e.g., performed in Python, R, STATA, SPSS, or QGIS). Although the skill profile of designated users is a bit more restrictive, required experience usually exists within KfW. Hence, data use cases from the first two categories may be carried out by practitioners without the explicit need for external support. Moreover, they may be integrated into terms of reference for consultants.

Complexity level three includes tasks like classification of a topic (security, health, etc.) or sentiment (positive, negative) of Twitter messages based on natural language processing methods or the evaluation of cellular connections (mobile communications), and requires pertinent skills concerning the use of Python/R and general knowledge of common data science techniques. The use cases may be carried out as (internal) projects by data scientists within KfW or be given to specialized consultancies. The development of complex AI models (e.g., for the classification of objects based on satellite imagery) or the integration of big data require very profound knowledge of special methods and usually a large amount of labor. This special expertise must be acquired externally.

Legal Aspects

Data protection: The data sources must be managed in a manner that is in line with the principles of data minimization and proportionality. Any data (structured or unstructured) may contain or reveal personal information of individuals and hence harm their privacy rights if not managed adequately. No individual data should be collected without prior consent and no data should be published to outsiders without a level of aggregation that allows for anonymization of the provided information. An agreement about usage and publication rights should be always obtained with the data providers. Thus, only personal data strictly relevant for the project should be collected and processed. If initial data minimization is not possible, data must be anonymized (e.g., by redaction or pixelation). The collected data must be **securely stored and protected**. Flawed and inadequate data security puts the rights of individuals to enjoy robust data protection at risk. > RMMV Guidebook, section 2.3.3.

Data security requirements can also arise from data protection regulations like the GDPR, which stipulate basic security requirements. Controllers of personal data must also have **appropriate technical and organizational measures** in place to satisfy data protection law. Business processes that handle personal data must be designed and implemented to meet security principles and provide adequate safeguards to protect personal data. Entities may be required under those rules to ensure the ongoing confidentiality, integrity, availability, and resilience of processing systems and services > RMMV Guidebook, Section 2.3.

In case KfW (or persons acting on behalf of it) are (also) processing personal data, the privacy check in > RMMV Guidebook Section 2.3.1 must be followed.

Before (re-)publishing information based on open data, you need to check its respective **licence type:** https://opendatacommons.org/

Project Examples / Use Cases

- The decision support system OSCAR uses open data sources as a baseline for the information platform strengthening health crisis response
- In the Investment Program Renewable Energies Eletrobras (BMZ: 2000 66 324) in Brazil, open data on annual discharge values of a small hydropower plant was used to evaluate the project.
- In the off-grid electrification program Green People's Energy for Africa (PN: 43770) in Mozambique, Open Data (e.g. on educational and health facilities) was used to identify potential mini-grid sites.

Links to Further Sources

- Digital rights check on AI: https://digitalrights-check.toolkit-digitalisierung.de
- Google metasearch of datasets: https://datasetsearch.research.google.com/
- List of national statistical offices: https://unstats.un.org/home/nso_sites/
- UN Datamarts: http://data.un.org/Explorer.aspx
- Harvard Dataverse: https://dataverse.harvard.edu/
- Open Data Impact Map: https://opendataimpactmap.org/
- List of free satellite imagery sources: https://eos.com/blog/free-satellite-imagery-sources/
- World Bank Open Data: https://data.worldbank.org/
- Open data for Africa by AfDB: https://dataportal.opendataforafrica.org/
- ASEANstats: https://www.aseanstats.org/
- World Bank Microdata: https://microdata.worldbank.org/
- International Household Survey Network: https://www.ihsn.org/
- Integrated Public Use Microdata Series: https://ipums.org/
- STATcompiler by DHS Programme: https://www.statcompiler.com/en/index.html
- GSMA: https://www.gsma.com/
- DEEP – a collaborative platform for effective aid response: https://thedeep.io/

Fact Sheet Technical Tool Type Data Sources (incl. Open Data and Big Data): May 2022

Linkages to other tool types

- (Remote) Management Information Systems
- Geospatial Tools
- Mobile Data Collection Tools
- Sensors / SmartMeters
- Cameras
- Data Sources
- Building Information Modeling

Further information on how to use this tool type in an RMMV context can be found here: