



*Empowered lives.
Resilient nations.*

A GUIDE TO DATA INNOVATION FOR DEVELOPMENT

FROM IDEA TO PROOF-OF-CONCEPT



You don't have to be a data scientist to innovate with data.

Any development practitioner interested in data innovation can benefit from this guide. *You do not need expertise in data science to integrate data into your projects! If you are a forward-looking innovator excited about improving your work and the work of your institution and partners, this guide can help.*

WHAT IS DATA INNOVATION?

Data innovation is the use of new or non-traditional data sources and methods to gain a more nuanced understanding of development challenges.

Data innovation often combines non-traditional with traditional sources of data, such as household surveys, to reframe issues and shed new light on seemingly intractable problems. New, or non-traditional data sources may include digital data derived from social media, web content, transaction data, GPS devices (see pp 19 for more). Because combining data sources can provide more complete, more timely, and/or more granular information about an issue, data innovation can open opportunities for more cost- effective interventions, as well as provide entirely new insights that may have been overlooked through traditional approaches.



CASE STUDY:

Improving late-night transportation through data innovation

Low-income workers in Seoul needed a new transportation option for commuting late at night. There was no night bus service, and taxis were expensive and hard to find. To establish a new night bus route, city officials analysed aggregated and anonymized mobile phone usage patterns to understand the most common points of departure and destination for travelers. This insight allowed officials to efficiently create a targeted “Owl Bus” route map, which best served late-night travelers.

For more information: <http://www.citiesalliance.org/node/5063>

WHY DO DEVELOPMENT PRACTITIONERS NEED DATA INNOVATION?

Governments and development agencies around the world are increasingly moving away from old models of problem-solving, and searching for new, more networked models of resilience.

Data innovation is a vital element of this effort. Identifying and integrating faster, more detailed insights into development programme planning processes can lead to better-targeted responses and more efficient resource allocation. Data innovation is also part of reaching the Sustainable Development Goals (SDGs). Effective data collection, analysis, and monitoring can help policymakers to course-correct programmes and policies more quickly, leading to cost efficiencies and greater returns on investments, as well as empower communities ultimately helping to achieve the goals.

HOW CAN THIS GUIDE HELP?

This guide will provide practical guidance in designing a data innovation project, as well as tools for recruiting allies both within and outside of your organization.

It does not include specifics about data analysis; you'll collaborate with a Data Expert for those. Instead, this guide offers an understanding of the scope and scale of data innovation projects, and guidance on asking the right questions from the beginning.

This guide covers the project design phase **from the earliest hint of a data innovation idea through the creation of a proof-of-concept.**



This project design phase may take as long as the actual solution development process itself, because it requires thoughtful coordination among various stakeholders. But with this groundwork firmly in place, data innovation has a better chance of success.

ABOUT THIS PUBLICATION

Recognizing the potential of data innovation, the **United Nations Development Programme (UNDP)** innovation teams at the Regional Centre for Europe and the Commonwealth of Independent States (ECIS) and the Regional Hub for Arab States embarked on a “big data for development exploration journey,” seeking to harness new sources of data to improve services or programme implementation on topics including disaster risk management, improved welfare, migration, and poverty reduction. As part of this exercise, six UNDP Offices from the two regions developed data innovation proof-of-concept projects with governmental and civil society partners, to address urgent local challenges and gain hands-on experience in data innovation.

UN Global Pulse provided technical guidance, coaching and tools to the project managers for the duration of the exercise, and the cohort shared advice and challenges at in-person workshops. This guide is based on these experiences, and on the methodology created by Global Pulse to design and implement data innovation projects across the UN system.

The joint exercise showed how alternative sources of data can and should play a role in pursuing development outcomes and, as such, hold great promise for fulfilling the **Sustainable Development Goals**—both from the perspective of pursuing the outcomes as well as enabling (close to) real-time monitoring and evaluation.

This publication is a how-to resource for UNDP staff around the world, and is designed to help further data innovation efforts within the UN and throughout the development field.

ABOUT THE UNITED NATIONS DEVELOPMENT PROGRAMME (UNDP)

UNDP is the UN’s global development network, advocating for change and connecting countries to knowledge, experience, and resources to help people build a better life. UNDP is on the ground in 177 countries and territories, supporting their own solutions to development challenges and developing national and local capacities that will help them achieve human development and the Sustainable Development Goals.

Learn more online at www.undp.org

ABOUT UN GLOBAL PULSE

Global Pulse is an innovation initiative of the United Nations Secretary-General. Global Pulse is working to promote awareness of the opportunities big data presents for sustainable development and humanitarian action, forge public-private data sharing partnerships, generate high-impact analytical tools and approaches through its network of Pulse Labs, and drive broad adoption of useful innovations across the UN System.

Learn more at: www.unglobalpulse.org



HOW TO READ THIS GUIDE

This practical guide will help you lay the groundwork for integrating data innovation into your projects. If you’re working on your first data innovation project, we suggest you go through the guide chronologically to make sure you remember each step and keep your project on track.

Make sure to meet the requirements of Checkpoint A before progressing to Sections Two and Three.

If you have previous experience managing data innovation projects, our suggestion is that you will refer back as you progress on your journey, dipping in and out of specific modules **as they become relevant** to your work.

Table of Contents

SECTION ONE




Explore the Problem & System

SECTION TWO

Assemble the Team


SECTION THREE

Create the Work Plan

	MODULE 1 p. 9
	Define the problem you're trying to solve.
	 Problem Definition Tool p. 10
	MODULE 2 p. 11
	Inventory and understand the data gaps.
	 Data Gaps Tool p. 12
	MODULE 3 p. 13
	Map all stakeholders
	 Stakeholder Mapping Tool p. 14
	MODULE 4 p. 15
	Understand who will use your data innovation results
	 Data Journey Tool p. 17
	MODULE 5 p. 18
	Identify your data wish list and define your hypothesis
	 Project Concepting Tool Annex

CHECKPOINT A (p.22):
Official permission

	MODULE 1 p. 26
	Account for Five Key Responsibility Areas
	MODULE 2 p. 30
	Secure support from Data Holder(s)
	MODULE 3 p. 32
	Find your Data Expert

	MODULE 1 p. 36
	Develop a concept note
	MODULE 2 p. 38
	Ensure data protection and privacy
	MODULE 3 p. 41
	Plan how to measure and share your results

	ANNEX p. 48
	Full library of tools for use

CHECKPOINT B (p. 45):
Create a Proof-Of-Concept

SECTION 1:

EXPLORE THE PROBLEM & SYSTEM

Every successful data innovation project starts with **a well-defined problem** (such as a development challenge or a policy issue) as well as a deep **understanding of the institutional system** around the problem.

With many fascinating data sources available, innovators in both the public and private sectors often get carried away collecting or diving into data sources that don't really matter to the actual challenges they face. Because they start with the data solution (instead of the problem), such projects are often unsuccessful.

This section will help you avoid this common mistake, with **five tools** for reflecting on what decisions, actions, and changes you hope to enable. It is designed to encourage rigor and spark inspiration, and organized based on common exploration processes, but you may find that specific tools are more useful individually, or at different times in your own journey. In some cases, completing one tool will help you flesh out another. Hop around the tools of this section, print them out, share them with colleagues, and make them your own.

OBJECTIVES

- Identify your data gaps
- Create a stakeholder power map
- Establish your Policy Research Question



CASE STUDY:

Laying the groundwork for data innovation in Armenia

Tourism has great potential for growth in Armenia. In supporting the sector as a national and local economic development vehicle, the government could benefit from more precise empirical data. Commonly used data sources, such as hotel booking logs and border control records, provide only part of the picture; they can also be time-consuming and expensive to collect.

The UNDP Armenia office saw an opportunity to **help the tourism industry** based on up-to-date tourist preferences by analyzing the number of roaming foreign SIM cards in use.

MODULE 1:

DEFINE THE PROBLEM YOU'RE TRYING TO SOLVE.

What is the specific development challenge or policy problem you hope to address through data innovation? Articulating the problem clearly will establish important boundaries, save time and resources, and maximize project returns.

Use the **Problem Definition Tool** (*sample on the following page*) to identify the key challenge you are trying to address, and to begin thinking about what data already exists on the issue. The tool will push you to think specifically about how better measurement or more complete information can help solve this problem. What new decisions will be made? What responses can be faster?

As you complete the **Problem Definition Tool**, you should desk research previous projects in the same problem area, or around similar data sources. This can spark new ideas for potential data sources for a solution; it may also serve as one indicator of your project's chance of success.

*You can find a printable version of the **Problem Definition Tool** in the Annex.*

The team and a local telecom operator worked together to analyse aggregated records from two tourist areas for a two-month pilot, testing whether insights about tourists' countries of origin and travel patterns within Armenia could be reached. The proof-of-concept was successful, and the team is now preparing to analyse a dataset based on a full tourist season, to share with both government decision-makers and local businesses to understand and adapt to shifting trends.

This success was possible because of careful groundwork and planning, following a process similar to that detailed in this section. Throughout this section, examples from the Armenia team's work will illustrate specific elements of the process, to help you develop a similarly thorough foundation to set your project up for success.

For more information reach out to Max Perry-Wilson (max.perry-wilson@undp.org) and Marina Mikhitarian (marina.mkhitarian@undp.org) from the UNDP Armenia office.

EXAMPLE: PROBLEM DEFINITION TOOL

The **Problem Definition Tool** completed by the **Armenia team** began with the fact that tourism is a key growth sector, with great potential for economic development in a country with abundant natural and cultural assets; tourism had also been identified as a priority in the national development strategy, but with very little actionable data that would allow the Government and tourism sector to design timely new services and products to capture more of a market share during the season.

The problem definition tool showed that additional information sources could help specific actors further their response to the problem. See case study on pp 8-9 for more on this project.

//DATA INNOVATION FOR DEVELOPMENT GUIDE

PROBLEM DEFINITION TOOL

START HERE

What is the key issue you are trying to address and **why** is it important?

Limited statistics (or real-time data) exist, or are used for development planning and business strategy, including tourist flows and preferences, or emerging hotspots.

Environment: What factors contribute to the problem?

National/local governments lack technical capacity to collect data from innovative sources that would allow more real-time strategy development and implementation.

Historically, few strategies for tourism development have been driven by any data with high levels of granularity.

People: Who does it directly affect?

Private sector businesses in the tourism industry (hotels, resorts, transport agencies, tour companies, etc). These could more quickly respond to new opportunities (i.e., recruiting an Iranian-speaking employee to serve a surge of tourists from Iran)

Government and development agencies investing in the tourism industry, which could better target their investment and measure their impact

//DATA INNOVATION FOR DEVELOPMENT GUIDE

PROBLEM DEFINITION TOOL

PAGE 2

Current Problem Solvers: Who is already working on this issue?

Government

Development agencies and civil society

Private sector businesses

Tourism Associations

At what time intervals are decisions about addressing this issue made by the Problem Solvers?

Existing data: What data, relevant to this problem, exist already?

Border agency data on tourists entering the country

Booking records completed by hotels and tour agencies

Business registrations associated with the tourism sector

Tax records of businesses working in the tourism sector

2013 Armenia international visitor survey conducted by USAID

//DATA INNOVATION FOR DEVELOPMENT GUIDE

PROBLEM DEFINITION TOOL

PAGE 3

In general, what is the **timeliness** and **level of granularity** (geographic? demographic?) of the existing data on the problem?

Potential new data sources for a solution:
Based on the factors above, what potential new or additional data sources could provide insights into this problem?

Mobile network data to track population movement: can be disaggregated by roaming (foreign) SIMs, domestic SIMs, and foreign customers who purchase local SIMs (via SIM registration).

GPS and other tracking technologies

Tourism company data on the movement of assets (such as buses or staff)

Cultural attraction admission data

Using your completed **Problem Definition Tool**, establish a first iteration of one or several of your key statements: your **Policy Research Question(s)**.

Your **Policy Research Question** is a short question that connects the data sources you just identified with the problem itself. Can this data help? It is phrased as a “yes or no” question:

Can we use [X data source(s)] for insights on [Y problem insights]?

This **Policy Research Question** will be refined *throughout the problem exploration phase and the completion of additional tools*. It will lead to the construction of your Hypothesis: a short statement that establishes what aspect of the data source you will test for insights on specific problem indicators.

SAMPLE POLICY RESEARCH QUESTION

ARMENIA ROAMING MOBILE DATA

Policy Research Question: Can data about “roaming” mobile phone SIM cards provide insights into tourists’ preferences about where they travel and plan to go while in Armenia?

MODULE 2: INVENTORY AND UNDERSTAND THE DATA GAPS


In the last module, you defined the problem you are trying to solve. **Now, it’s time to think about what important data (see pp 19) you have (and lack) internally, across both short- and long-term planning cycles.**

Also think about insights you wish you had (perhaps at a different granularity, on a different dimension or more frequently), but cannot get from your current data collection methods. Bridging these “data gaps” will likely be the focus of your data innovation project.

Use the **Data Gaps Tool** to make a complete list of readily available data. The goal of this is to understand fully the data that already exists, including what you rely on for long-term planning (even if it’s imperfect). Be sure to capture all data that are easily available within your organization, including what is gathered in reports and stored in databases. Consider quantitative as well as qualitative data, such as small-scale surveys, focus groups, and ethnographic methods like observation and interviews.

Note that one data source is not usually sufficient. A thoroughly completed Data Gaps Tool can highlight the fact that multiple data sources may be necessary to fill all gaps. There is a printable tool in the Annex.

EXAMPLE: DATA GAPS TOOL



//DATA INNOVATION FOR DEVELOPMENT GUIDE

DATA GAPS TOOL

//PROJECT NOTES:

PAGE 1

STEP 1

List one kind of data or source of data in the top row of each column

Existing data on the problem

Look at any programmes that deal with this problem. What data currently exists that is used to support day-to-day operations? Long-term planning? Communicating with or persuading others? For Monitoring & Evaluation?

Border agency information on tourists entering the country

Booking records completed by hotels and tour agencies

2013 research on tourism in Armenia conducted by USAID

STEP 2

For each data source, answer the following questions

Is it openly available, or does it require special permission to access?

Available to government

Openly Available

Openly Available

Is it structured or unstructured?

Structured

Structured

Structured

How often is it collected?

Collected in real-time at the border. Shared quarterly.


Quarterly

One-off

The Data Gaps Tool completed by the Armenia team shows that, while there is a wealth of data about foreign tourist activity available in hotel booking records and other surveys, there was a gap in information about internal tourist flows (particularly in relation to domestic tourists) and about what foreign tourists spend.

Additionally, data on international tourists were limited, as hotels and tour agencies often only submit data on those tourists paying for services through credit card, which is especially limited outside of Yerevan.

See case study on pp 8-9.



//DATA INNOVATION FOR DEVELOPMENT GUIDE

DATA GAPS TOOL

//PROJECT NOTES:

PAGE 1

STEP 1

List one kind of data or source of data in the top row of each column

Existing data on the problem

Look at any programmes that deal with this problem. What data currently exists that is used to support day-to-day operations? Long-term planning? Communicating with or persuading others? For Monitoring & Evaluation?

Border agency information on tourists entering the country

Booking records completed by hotels and tour agencies

2013 research on tourism in Armenia conducted by USAID

STEP 2

For each data source, answer the following questions

Is it openly available, or does it require special permission to access?

Available to government

Openly Available

Openly Available

Is it structured or unstructured?

Structured

Structured

Structured

How often is it collected?

Collected in real-time at the border. Shared quarterly.

Quarterly

One-off

Color-code data to highlight gaps.

Looking at all of the data sources you have listed, use a marker or stickers to code. Use the following indicators, each color indicating a different kind of flow in the data source.

Red: Updated too infrequently

Blue: Not geographically specific

Green: Not disaggregated

Yellow: Obsolete source

If you find that you have an internal data source that does not need any of these color codes, that data source may already be subject to your development project and does not need to be covered by external data innovation projects.

MODULE 3: MAP ALL STAKEHOLDERS

In addition to the need to focus on the problem to be solved, another common pitfall of data innovation projects is **focusing on theoretical or technical concerns while overlooking the operational pathways necessary for real impact.**

Understanding all actors and their relative influence over the problem will help frame your plans within the existing government, civil society, academic, and other systems responsible for addressing a problem.

Your data innovation project, like all of your work, is part of a much larger system of efforts by governments, NGOs, and communities dealing with the problem. By understanding this broader network of people involved, you can better position your project to secure vital support from powerful actors and to open participation and collaboration to the people who are most affected by it.

THREE TYPES OF ACTORS TO CONSIDER:

- **COMMUNITIES:** The people who experience the problem directly, and interact with Problem Solvers.
- **PROBLEM SOLVERS:** The civil servants, NGO staff, front-line responders, and others on-the-ground.
- **POLICYMAKERS:** The people who have access to resources and control allocation.

Use the **Stakeholder Mapping Tool** to list all of the parties potentially involved in or affected by your data innovation project and organize them according to their motivations, roles, and constraints.

In this step, you should go beyond a default listing of partners you work with. Focus in on the individuals who have leverage and stake in the intervention you are planning, including members of formal groups and organizations as well as issue experts. Note that, while development projects often lump actors according to a simpler “supply vs. demand” categorization framework, this tool encourages you to understand actors according to their influence over the problem at hand.

There will probably be a few partners who have some capability to enact change. With this in mind, the **Stakeholder Mapping Tool** is designed to distill the particular roles of relevant partners in terms of your Policy Research Question.

12

13

EXAMPLE: STAKEHOLDER MAPPING TOOL

On the Stakeholder Map completed by the Armenia team, “community” and “Problem Solvers” are the same; tourism businesses will use the new data insights. The map also highlights tourist associations as valuable potential channels for information distribution.

See case study on pp 8-9 for more on this project.

//DATA INNOVATION FOR DEVELOPMENT GUIDE

STAKEHOLDER MAPPING TOOL

PAGE 1

STEP 1

Brainstorm and list all stakeholders

Include the people directly and indirectly affected by the problem; private- or public-sector institutions who have data; academic organizations with data science experience; political actors committed to addressing the problem; national statistics institutes or other involved government agencies.

Categorize each stakeholder as "Community"; "Problem Solver"; "Policymaker"; or "Other". In general, "Community" should be the first stakeholders you consider.

STAKEHOLDER

Small tourism businesses

What is their influence over the problem?

Medium, important data contributor, and directly foster tourism industry (but lack access to local or regional data)

What is their influence over the project?

Low

How might this person benefit from the project?

Receive insights on their target markets, including new growth areas

How does data support this person's decision-making now?

Where available, use information about tourist preferences to adapt business strategies

What could this person do with better data on the problem?

Adapt business strategies or investments to meet new opportunities

What could they do to undermine the project?

Dismiss data insights

What is the best way to keep them engaged?

Meetings and presentations that clearly show benefits of new data

//DATA INNOVATION FOR DEVELOPMENT GUIDE

STAKEHOLDER MAPPING TOOL

PAGE 5

STEP 2

Plot all stakeholders on the following map

Those in the center of the circle have the greatest influence, those on the perimeters have the least. Include brief notes about strategies for engagement or opportunities for support.

NOTES: _____

MODULE 4: UNDERSTAND WHO WILL USE YOUR DATA INNOVATION RESULTS

On your Stakeholder Map, there are two groups of actors who need to play a central role in the design of your data innovation solution: **The communities who experience the problem directly, and the Problem Solvers who are responsible for addressing it.**

These two groups should be the main focus of the data innovation design. If you tailor the project to their needs and constraints, you have a greater chance that they will actually use your solution effectively.

Take a moment to consider whether your Problem Solvers can be the same people and communities who experience the problem directly (i.e., the “beneficiaries” of your work). Rather than focusing on large government or NGO responders, many data innovation solutions directly empower affected communities to address problems themselves. These kinds of solutions are best practice for building resilience and sustainability.

The **Data Journey Tool** (on pp 17) explores the problem statement from a different perspective: the interactions that generate data points. You should fill out the Data Journey Tool putting yourself in the shoes of the communities you are trying to help - try to identify what data points are generated from that perspective. For example, what data points would a person living with HIV/AIDS be generating? Also think whether the people you are trying to assist and the Problem Solvers are one and the same category. If the Problem Solvers are not the same as the people experiencing the problem, you might want to fill out a second Data Journey Tool answering each question from the point of view of the Problem Solver - for example a UN agency or a government official that use a dashboard to implement policies.

[illegible]



Designing for government and community co-management

Can you include the people who experience the problem directly as part of the data innovation? The case of PetaJakarta, an open-source flood map in Indonesia, shows the potential of this approach. Whenever a person tweets about flooding, the tool automatically pulls geolocation data connected to the tweet to populate a flood map. If a person's phone has disabled geolocation, the map automatically sends them a message asking them to provide location information, which can then be uploaded to the map.

The data and map supplement the existing disaster risk management information ecosystem available to government agencies. It is also available to Jakarta residents--providing information about flooding in the city in real time, so that everyone can make informed decisions about their own safety in the face of flooding.

“Designing the platform to meet the needs of citizen-users and government agencies enables and promotes civic co-management as a strategy for climate adaptation.”
- PetaJakarta.org

Models like PetaJakarta, that include communities in the creation of data innovations, and that open their results directly to those communities in order to use them, are at the cutting edge of building resilience and sustainable change in communities. They are best practice in data innovation. (Source: PetaJakarta.org)

At this stage, the **Data Journey Tool** will help you understand the problem from the perspective of service design, highlighting the data touch points and highlighting potential opportunities for intervention. *There is a printable version of this worksheet in the Annex.*

EXAMPLE: DATA JOURNEY TOOL

The **Data Journey Tool** completed by the **UNDP Armenia team** shows the multiple potential touch-points where data is generated by tourists, and thus may yield insights about their preferences.

See case study on pp. 8-9 for more on this project.

//DATA INNOVATION FOR DEVELOPMENT GUIDE
DATA JOURNEY TOOL

STEP 1

Fill in a short description of an individual who represents the beneficiary, community or target of a certain development intervention (ideally based on ethnographic research). What are his or her goals, constraints, motivations?

Looking at your completed map, are there additional data gaps to add to the **Data Gaps Tool**?

Are there additional data sources available to help address the problem?

At what point in the journey would a data innovation most help the Problem Solver or the communities deal with the problem?

A tourist visiting Armenia on vacation

//DATA INNOVATION FOR DEVELOPMENT GUIDE
DATA JOURNEY TOOL

STEP 2

Working from ethnographic research or your knowledge of the issue, fill out the user's "starting point" below. **What is the typical journey of such an individual with the problem?**

STEP 3

Now, on each row of dots below, **plot the steps a person might take to address the problem**. Each touchpoint (where the person visits an office, fills out a form, talks to another person, or takes any other action) **should go on its own dot**.

	Search	Reserve	Travel To Armenia	Border Experience	Arrival In Armenia	Tourist experiences	
	•	•	•	•	•	•	
	Search query for places to visit in Armenia	Online accommodation booking	Drive	Iranian border	Purchase new SIM on arrival	Use hotels/hostels	
	•	•	•	•	•	•	
	Use popular travel guidebook	Contact friends of friends in Armenia	Bus	Georgian border	Leave phone at home. Use guidebook as reference point	Visit key museums/touristic spot	
Decides to travel to Armenia	•	•	•	•	•	•	
START	Ask Armenian friends/diaspora	Couchsurfing	Fly	Zvartnots Airport	Use roaming with own SIM	Public transport	END
	•	•	•	•	•	•	
	Visits Travel agency	Book Through Agency			Use SIM provided by travel agency	Organised tour	
	•	•	•	•	•	•	
			Buy a camper van		Turn phone off. Ask local people for advice	Social network interaction	
	•	•	•	•	•	•	

STEP 4

What data is gathered at each touchpoint? Write it below.

STEP 6

What is the timeline of the actions? Plot it out here.

MODULE 5:

IDENTIFY YOUR DATA WISH LIST AND DEFINE YOUR HYPOTHESIS

Once you have completely documented and analysed the sphere of the problem from all angles, **it is time to start looking for useful, innovative external data sources**. What do you need to measure or count to fulfill your task/work? What new dimensions of the problem would you like to be able to count or measure if you could?

Work with colleagues to brainstorm the external sources of data that either deal with the problem explicitly or could be a proxy indicator, providing new insights on the problem. Consider all available sources to find the best solution (i.e., don't just jump on the fanciest big data solution).

EXAMPLE: KINDS OF DATA

Consider all types of data that may pertain to the problem, and **remember that there are often multiple sources** (official statistics, UN data, or private sector data) and multiple types (crowdsourced data, open data, and big data). The list on pp 19 can help spur ideation:

What is reported

Data produced from explicit attempts to gather information from people

- Large surveys (household surveys)
- Programme data
- Mobile surveys
- Crowdsourced data

What people say

Data produced when people explicitly share something with the world (usually big data)

- Social media
- Online news
- Blogs and forum posts
- Online archives
- Radio and TV

What people do

Data produced passively when people make transactions through digital services (usually big data)

- Online searches
- Mobile phone use
- App use
- Postal traffic
- Financial transaction records
- Digital shopping records

What sensors measure

Data collected by physical sensors recording actions and physical changes (usually big data)

- Weather sensors
- Traffic cameras
- Satellite and drone imagery
- GPS records
- Ambient sensors

With your data wish list in hand, it is time to bring all of your problem exploration and planning work together. You have fully explored the problem, the system, and the data options; now, use the **Project Concepting Tool** to synthesize your ideas and refine your hypothesis. This tool will help you make sure your chosen data innovation idea fits the needs of the problem you've identified; it will also help you synthesize the project into a concept, which you can share with supervisors and potential project team members to explain not just the "what" but the "why" of your data innovation project.

In the last step of the **Project Concepting Tool**, you will refine and solidify your hypothesis. The hypothesis should be simple and clear enough that someone with no experience in data innovation can read it and understand the project you're proposing. See Annex for the Project Concepting Tool.



CASE STUDY:

Measuring poverty from outer space

Measuring poverty is a long-running challenge for development practitioners. The household survey, which is widely used, is time consuming, expensive, and often requires elaborate data collection and analysis processes. In Sudan, due to armed conflict, household surveys are obviously much, much more complicated.

One innovative proxy indicator for poverty levels is electricity consumption, as households tend to decrease their consumption when they have fewer resources. The UNDP Sudan office set out to test whether satellite data could be used to estimate poverty levels via changing night-time energy consumption. The team used data pulled from night-time satellite imagery, analyzing these illumination values over a two-year period, in conjunction with electric power consumption data provided by the national electricity authority. The proof-of-concept successfully showed that the availability of electricity can be measured from outer space and also reflected the energy poverty prevalent in the country.

Combined with desk research from similar recent studies in Kenya and Rwanda by the World Bank, the team found that night-time satellite imagery has potential to be a reasonable proxy for poverty. The proof-of-concept provides the validation needed to rally further resources and continue investigation.

For more information reach out to Anisha Thapa (anisha.thapa@undp.org) and Jorg Kuhnel (jorg.kuhnel@undp.org) from the UNDP Sudan office.



TIP: DON'T RUSH! The process of exploring the problem and its ecosystem merits deep thought. So take your time! It is common for teams to work through several potential hypotheses before finding a feasible, useful project.

Using your previously completed tools and the Project Concepting Tool, you will refine your **Policy Research Question** and establish the **Hypothesis**, which will be the basis for your project interventions.

Your **Policy Research Question** is a short question that connects the data sources you identified with the problem itself. **Can we use [X data source(s)] for insights on [Y problem insights]?**

At this point, you should be able to refine your research question(s) based on the understanding of the problem developed filling out the different tools and based on an extensive background research and literature review of previous research. If everything seems reasonable, you are ready to develop the hypothesis.

Your **Hypothesis** is a short statement that establishes what aspect of the data source you will test for insights on specific problem indicators. Since you will either verify or falsify the hypothesis during your project, it is phrased as a “true or false” statement:

Since we know [A facts] about [X data source(s)], we believe we can use: [B element of relevant data source] to see [C problem indicator]. We will validate our results by comparing [D existing data].

These are some questions that might help refine your hypothesis: What data gap would be most effective to bridge? What unexpected barriers surfaced? What do Problem Solvers need? What indicators can be used to determine whether your data innovation is successful?

The **validation** methodology is extremely important. A good practice is to validate the new data source with a retrospective case study where you can compare your results against a ground truth data set that was used in the case. If there is no established ground truth dataset, e.g. a new SDG indicator, you should think about alternative sources of measurement that could be used to cross validate a relevant sample.

***The Policy Research Question and the Hypothesis are going to shape your entire project.** You’ll refine both of them throughout the project design. Ideally, you will develop the hypothesis together with project partners, and validate it in discussion with those who will be playing a role in field-testing, and especially those who will benefit from the subsequent solution.*

KOSOVO 112 CALLS *(See the full case study, pp. 29)*

Policy Research Question: Is it possible to use the temporal and spatial distribution of emergency call records to map and predict hotspots of demand for emergency services?

Hypothesis: Since we know that all 112 calls are recorded by emergency responders, we believe we can use the time and location of these calls to build a predictive model of emerging trends in emergency response. We will validate our results by comparing them to actual recorded call patterns.

**References to Kosovo shall be understood to be in the context of UN Security Council Resolution 1244 (1999).*

EGYPT: WEATHER SENSORS

Policy Research Question: Can data collected sensors network and weather data inform irrigation planning and water management?

Hypothesis: Since we know over 30 environmental sensors distributed across the country collect data for over 10 environmental indicators on an hourly basis, we believe we can complement that data with local and international weather station data, and use quantitative modelling and analysis as well as visual analytics, to support decision-makers to make more informed water-management decisions.



CASE STUDY:
Improving agricultural yields through data innovation

Climate change is increasingly threatening agriculture around the world, including in Egypt, where the UNDP office set out to address the problem through data innovation, in collaboration with the Faculty of Computers and Information at Cairo University.

The challenge to farmers, and the lack of data, comes from unexpected shifts in rainfall and planting seasons. Most agricultural producers rely on historic data, which is increasingly unreliable in a shifting climate. But data innovation is offering a new tool.

WORKING FROM PRECEDENT
The team knew that in Colombia, for example, a team of researchers and a local industry organization won the UN Global Pulse Big Data Climate Challenge by creating a predictive model for rice farmers.

The team used multiple data sources:

- annual rice surveys
- harvest monitoring data
- experiments on rice sowing dates (provided by the industry organization)
- weather data (provided by the National Institute of Hydrology, Meteorology, and Environmental Studies)

Using this data, the researchers identified geographically specific relationships between climate and agricultural yields. The tool they created successfully forecast a drought, and helped 170 farmers in Colombia avoid an estimated \$3.6 million loss during the pilot. *(Source: UN.org)*

INVESTIGATING LOCAL DATA SOURCES IN EGYPT
Following precedents like the Colombia research, the team in Egypt used the steps detailed in Section 1, Module 5 to uncover sources of data that could be used for a local innovation project. They established two sources: data from the sensors network Central Laboratory for Agricultural Climate, and local and international weather station data, a combination that (based on previous successes) proved to reveal insights for decision-makers in the area of water management.

For more information reach out to Sherif El Tokali (sherif.el.tokali@undp.org) and Nadine Abou Egheit (nadine.abou.elgheit@undp.org) from the UNDP Egypt office.

CHECKPOINT A: GET THE PROJECT GREEN LIGHT

Securing permission from organizational management is an early make-or-break step. Be prepared for others in your organization to resist your idea; it can sometimes take several months (or longer) to make a compelling business case for the value of this new field of data innovation.

Tailor your pitch to your supervisor, based on their motivations and goals. Personal experience and existing working relationship will be the best guide, although it is generally wise to start by introducing the problem that needs to be solved, instead of diving directly into data innovation. When making the case for data innovation, the following four reasons are the most compelling:

1. CREATING NEW INSIGHTS:

***Increasing efficiency during implementation:** Data innovation can offer a new layer of insights for existing projects. These insights can make implementation more efficient, and better position your office with governments and donors.*

FOR EXAMPLE: The UNDP Kosovo-Global Pulse project showed that analyzing the patterns of previous 112 emergency calls could be used to reduce response time, by modeling and mapping out temporal trends and spatial hotspots of demand for emergency services. (See pp. 29)

2. MOBILIZING MORE RESOURCES:

Successful data innovation projects are increasingly high-profile, and a number of governments and donors are interested in supporting new ones. Data innovation may attract new sources of funding and support to your team.

FOR EXAMPLE: Based on the success of the UNDP Armenia office’s data innovations (including the tourism case study detailed in Section 1), the European Commission engaged the UNDP to export its innovations to the government for a EUR 300,000 investment.

FOR EXAMPLE: Based on the UNDP-Global Pulse work on big data for development, the UNDP Istanbul Regional Hub pitched an idea for data for policy and governance to the Slovakian Government who invested \$1.5 million.

3. EARNING POSITIVE RECOGNITION:

Teams that do data innovation work are highlighted for their work at major conferences, events, and in high-level conversation.

FOR EXAMPLE: Tunisia, one of the countries responsible for nationalizing SDG pilots, complemented their work on SDG 16 (“peace, justice and strong institutions”) by measuring public sentiment about corruption using social media data. The real-time layer of data brought to the existing monitoring project has led the National Statistics Institute to be highlighted at major conferences. Subsequently, the team from the Tunisian National Institute of Statistics has joined the Task Team on the SDGs and Big Data as part of the Global Working Group on Big Data for Official Statistics, organized by the UN Statistical Commission.

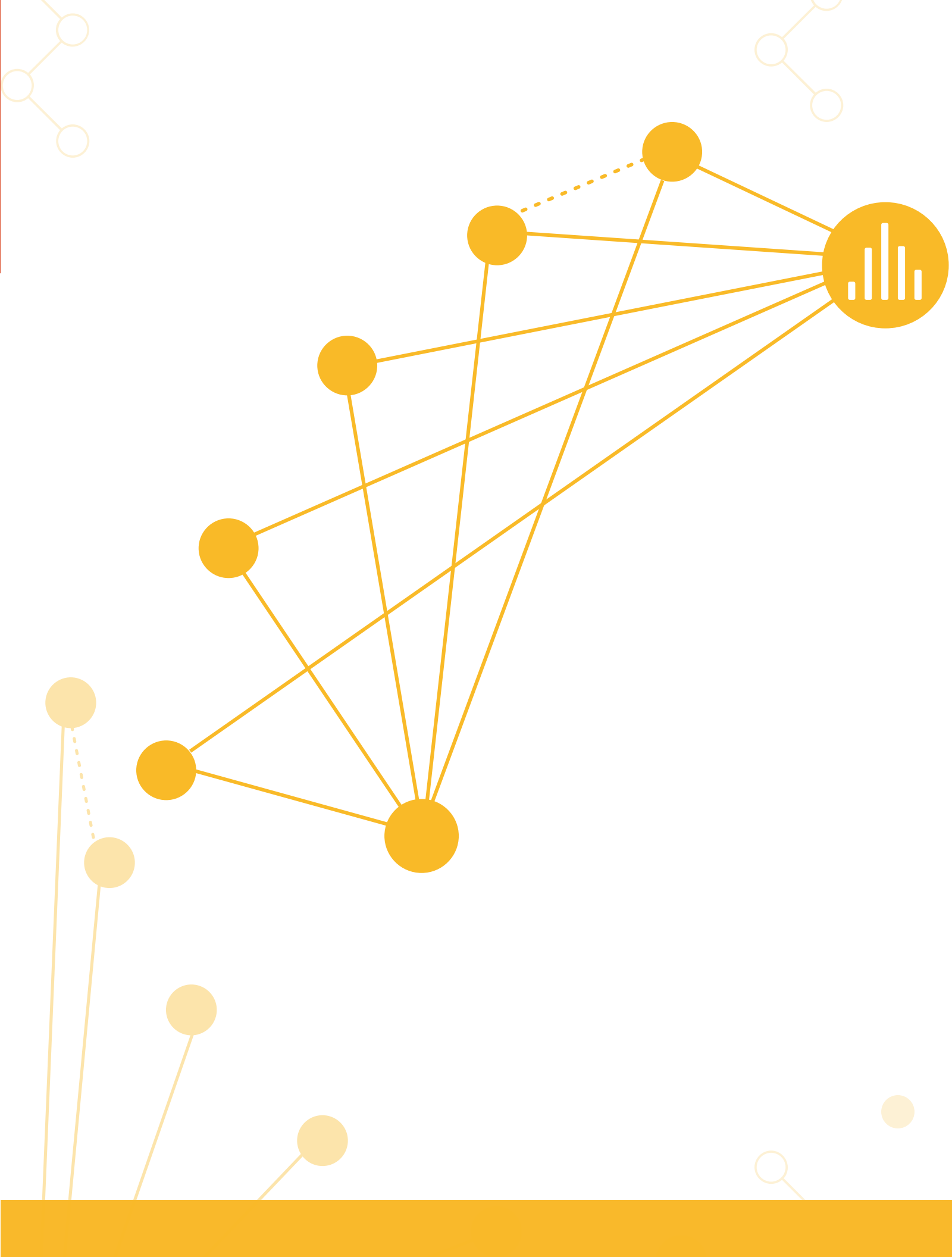
4. OFFERING NEW SERVICES TO CLIENTS:

The development industry is changing, and more start-ups are offering innovative new approaches to solving challenges. Data innovation is an important way for development practitioners to offer real value. A data innovation pilot can turn into a service offering for government partners in an increasingly competitive market.

FOR EXAMPLE: After completing a data innovation project on predicting mobility patterns for disaster response management, the UNDP former Yugoslav Republic of Macedonia office is now offering its expertise to help government partners redesign services, such as **access to services for people with disabilities, strategies for urbanization, and health and environment policy making.**

CHECKPOINT COMPLETED?

Great! It's time to move into the specifics of the work plan.



SECTION 2: ASSEMBLE THE TEAM

Most data innovation projects are deeply collaborative. External allies might be necessary to secure access to, transfer, or analyse data from one or more sources, and internal allies will be necessary to interpret and eventually act on the data. In development specifically, data innovation requires team members who have expertise in one or more areas of data science as well as in the problem domain.

This chapter will help you plan the roles and alliances you will need to negotiate in order to implement your project, and offer exercises and tips for securing talent and human resources within a budget.

OBJECTIVES

- Build your project team
- Engage with the Data Holder(s)
- Engage with affected communities

MODULE 1: ACCOUNT FOR FIVE KEY RESPONSIBILITY AREAS

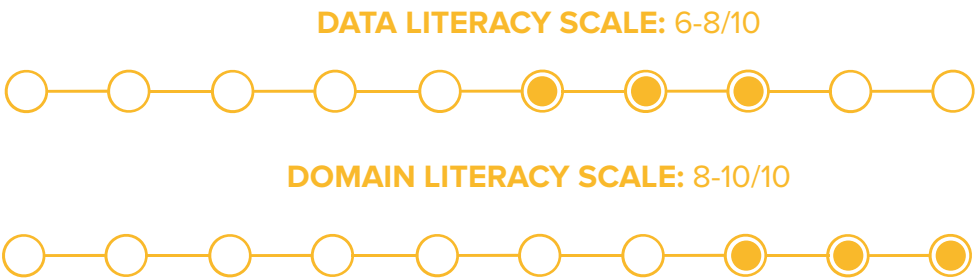
Typical projects require a team to divvy up five main areas of responsibility. Most often, one person will be able to take on two or more roles:

PROJECT MANAGER

Who will drive the project forward and ensure its success?

The Project Manager (who is probably you!) makes sure that the project is on time, on budget, and that all stakeholders are in the loop. In case you are managing a project on behalf of senior management or other departments in your organization who are actually the Problem Solvers (see below), it will be important to have good lines of communication to ensure true ownership of the project and its results as a learning process.

Key to success: *Understand the constraints and opportunities faced by all other stakeholders.*

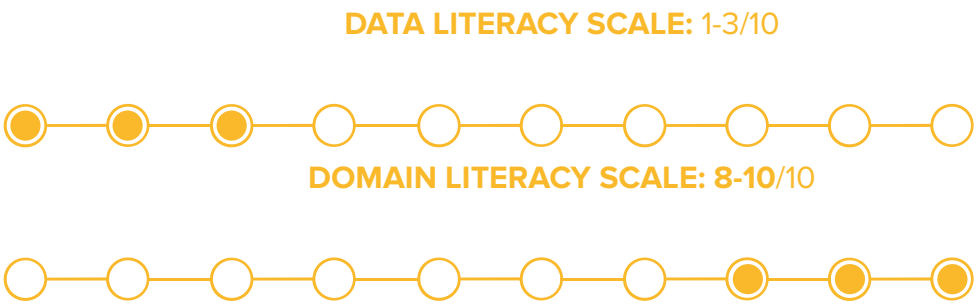


PROBLEM SOLVER (OR DOMAIN EXPERT)

Who will act on the insights derived from the analysis?

The best insights in the world are worth nothing if no-one acts on them. The Problem Solver is the person (or people) with expertise and a deep understanding of the substantive issue or challenge. If the project isn't designed for their needs, you will waste time and money. Consider involving people at all levels, and in all relevant offices and geographies. The Problem Solver, as the domain expert, should have a big role throughout the research process. The Problem Solver may not have a pre-existing understanding of "what's possible" when working with big data, so some training or orientation may be valuable.

Key to success: *Engage the Problem Solver in the design of the project and all outputs to ensure that it is targeted to the most useful application.*

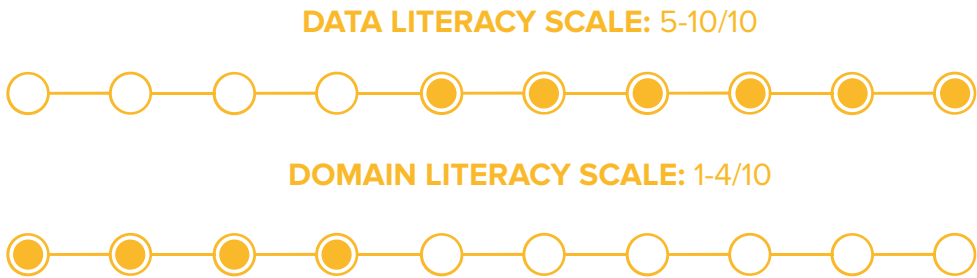


DATA HOLDER(S)

Who will provide access to the data source(s)?

Sometimes, you may need to work with "Open Data", in which case you can access the data directly through the web or other means; but other types of data may have restricted access and require direct request or negotiation with a Data Holder. The Data Holders may be another department in your own agency, a private corporation, an NGO, or a government agency; in any case, a partnership based on trust is crucial, and may take some time to develop. Some Data Holders will not be involved in a project beyond providing access to data, while others may also provide analysis or other resources, and may want to be involved in communicating (and celebrating!) the project results.

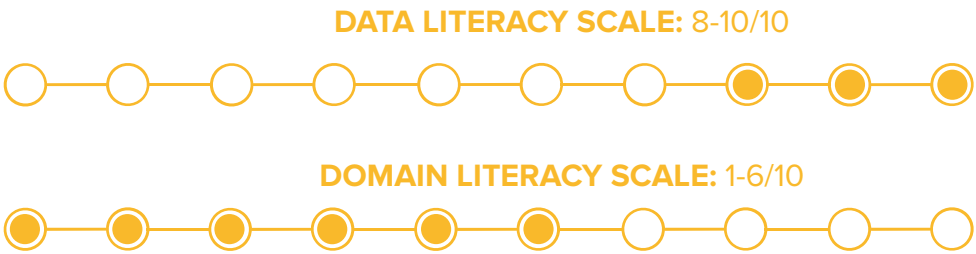
Key to success: *Nurture a relationship with an internal focal point, who understands the holder-organization's decision-making hierarchy, believes in the project, and can champion its progress.*



DATA EXPERT

Who will do the heavy lifting of analyzing the data?

Good data scientists and engineers are hard to come by. The field is relatively new and the private sector has scooped up a high number of talented professionals. Recruiting a qualified Data Expert will require some level of data literacy. Discuss with all partners if anyone has tried to work with this type of data before, who can either dedicate time to recruitment or may serve some part of the researcher role. In many cases, dividing the data roles between multiple people (often with support from the Data Holder) can be a good solution. The Data Expert need not have a strong background in the problem domain, although it is an advantage. Sometimes, you can substitute part of your data expert team with a specialized tool or software, depending on the type of data source and analysis.



Data Scientists, Data Engineers and Data Visualization Specialists

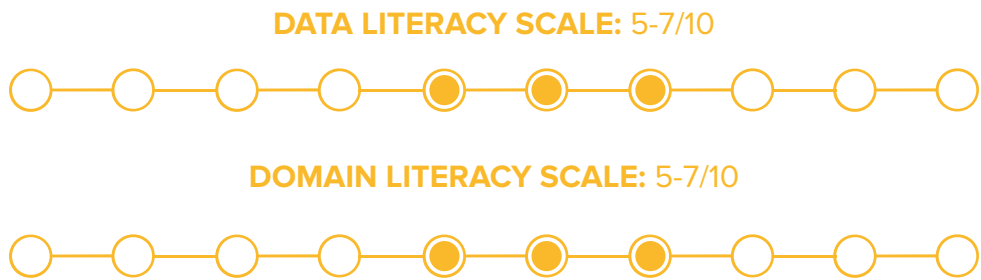
There are three types of data experts needed for data innovation. **A data scientist understands the analytical tools and methods** necessary to work with big data and can pull insights out of datasets. **A data engineer establishes the architecture and technical infrastructure** needed to transfer, store, and process the data efficiently and securely (which may require multiple, high-functioning servers). Sometimes, data science and data engineering can be done by the same person. Projects may also call for a **data visualization specialist** to map, visualize, and interact with the data and the insights.

DATA PRIVACY & LEGAL EXPERT

Who will ensure privacy protections and regulatory needs are followed?

While likely only relevant for discrete phases of a project, the guidance of a legal and data privacy expert may be necessary. Make sure to get either internal or external legal support for partnership negotiations for data access (if needed) and ensure project design compliance with any national or local laws, as well as ethical standards related to data use, handling, transfer, storage, and analysis if applicable. This person should also support you in the development of a risk mitigation plan for your project if possible.

Key to success: Plan regular check-ins to make sure the risk mitigation plan is continuing to cover all potentialities associated with data use as the project evolves or changes



CASE STUDY:



Mapping 112 calls to plan emergency response in Kosovo

Allocating resources to emergency services is a challenge for cities around the world (since fires, traffic accidents, and other security and health issues don't follow a set schedule). Many are beginning to use the data recorded during emergency calls (in this case 112) to better predict service positions and schedules.

In Kosovo, a team of data innovators worked together to not only create a predictive model, but to do it with direct input from communities.* Team members included the UNDP Kosovo office, Open Data Kosovo, and the Kosovo Emergency Management Agency (EMA). The team used data from 112 calls to find patterns in type and timing of emergency calls over the course of a 22-month period in one city; to address data gaps about specific locations of emergencies, they have created a crowdsourcing tool for people to report geographic information as well as access information about emergencies in their areas. Here's how the team accounted for the five key project responsibility areas:

PROJECT MANAGER: The UNDP Kosovo office coordinated the project, with support from Global Pulse.

PROBLEM SOLVER: The EMA provided its call records to the team in part because the project promised to bring new insights for better allocating emergency response resources.

DATA HOLDER: The EMA had the primary data in its internal 112 call records. Communities are helping create the geographic data, which is gathered and "owned" by all team members in the shared crowdsourcing tool.

DATA EXPERT: Open Data Kosovo, a well-known local CSO, brought its data expertise in part because the project, by opening more government data to the public, fit well within its mission.

DATA PRIVACY & LEGAL EXPERTISE: The team met with the Kosovo Agency for the Protection of Private Data, created to uphold Kosovo's data privacy laws, for guidance on preparing for and monitoring privacy concerns.

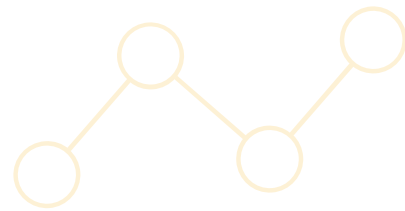
KEY TO SUCCESS:

Early in the project, the team convened, along with institutional and CSO stakeholders, for a workshop on data innovation, including a presentation by UN Global Pulse. By asking questions, seeing demonstrations of data innovation possibilities, and understanding both the risks and benefits, the team was able to kick off the project in sync, with energy and enthusiasm. Through early communication with privacy experts in Kosovo, the team was able to preempt privacy concerns and bring the Kosovo Agency for the Protection of Private Data on board. Alignment with an existing UNDP Disaster Risk Reduction (DRR) project then helped with access to and commitment from key partners.

**This approach is "best practice" in data innovation; when both communities and governments co-manage and have full access to data insights, it can exponentially increase resilience.*

<http://europeandcis.undp.org/blog/2016/04/20/can-big-data-help-us-make-emergency-services-better/>

For more information reach out to David Svab (david.svab@undp.org) and Shpend Qamili (shpend.qamili@one.un.org) from the UNDP Kosovo office.



MODULE 2:

SECURE SUPPORT FROM DATA HOLDER(S)

Engage with Data Holders as early as possible. Without access to the necessary data sources or data sets, the project will have to go back to the drawing board. In some cases, the data required may be proprietary, or held by a private sector company. Therefore, the establishment of a partnership will likely be required.

This may take some time and negotiation. As building trust takes time, there are not many things that can be done to speed these negotiations. However, a project can get a head start by thinking strategically about the best entry points for engaging a private organization, and by developing an internal champion.

DATA PHILANTHROPY:

How to convince private companies to share access to data or insights

Some of the data you need may be owned by private sector companies, like telecommunications or technology companies, whose primary goal is profit, and who have spent millions on their data systems.

However, the practice of “Data Philanthropy” - making private sector data available for the public good- is growing. There are many different potential modalities, and the practice is still in its infancy. Any potential data partnership should involve a series of meetings. Here we list common arguments that can help understand and make the case for data philanthropy:

COMMON CONCERNS:

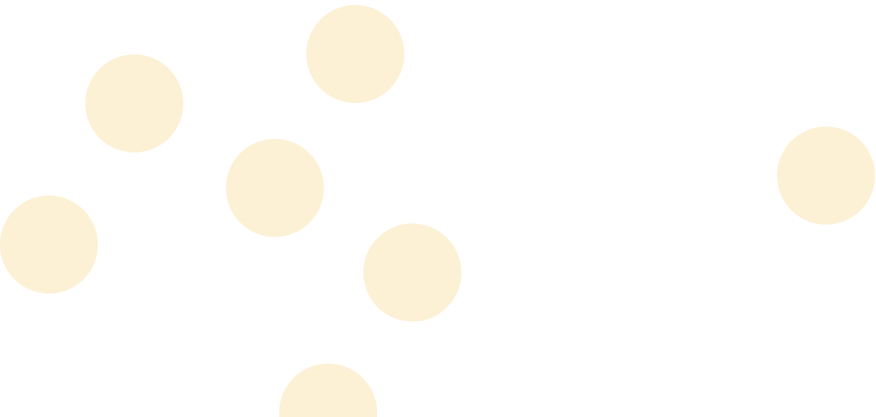
- Unclear benefits:** The value of a project may not be obvious, especially with companies less familiar with either the domain or the potential of data innovation
- Privacy concerns:** Companies fear risking private data, and may lack knowledge about how to anonymize or aggregate data to protect privacy
- Competition concerns:** In the case of a data leak, companies fear commercial competitors might gain advantage through the data

COMMON MOTIVATIONS:

- Corporate social responsibility (CSR):** Some businesses have committed entire departments to CSR efforts, and have established channels for requests
- Motivate staff:** Corporate staff might be excited at the chance to serve society; the opportunity to contribute expertise to a project may provide valuable human resource development
- Business development value:** Involvement with an innovative success might offer publicity and prestige, as well as add to internal product development

EFFECTIVE ENGAGEMENT TACTICS:

- Prepare for common concerns:** Have a detailed privacy plan to share, as well as a plan for data protection, and be open to feedback on both
- Start small:** To gain trust and show the value of a project, request a small dataset to begin for the proof-of-concept. Success depends on making the data request as precise and small as it can be and still demonstrate added value
- Engage champions:** Working through known networks can greatly speed a request; having a well-defined, trusted focal point within the company can help ensure follow-up and alignment
- Get recommendations:** When accessing private sector data, a government representative or regulator may be willing to offer a letter of endorsement for the project, building trust and prestige



MODULE 3:

FIND YOUR DATA EXPERT

While some of the responsibility areas may be straightforward to fill, chances are high that your organization lacks an available data scientist. If the budget allows, and depending on your region, you may be able to hire or contract data experts for the project.

Alternatively, you may consider working with academic institutions, or engaging an intern or volunteer. Consulting with the Data Holder may provide additional leads. Some avenues for sourcing external expertise include:

ACADEMIC PARTNERSHIPS:

Professors or graduate students at local universities may be interested in collaborating with development efforts, which can offer them many benefits, including access to interesting data sets for their own research and real-world practice. From experience, academic researchers tend to be motivated by publication; when considering structuring a partnership arrangement or internship opportunity, consider ways that your collaboration may provide fodder for their goals.

VOLUNTEERS OR INTERNS:

Certainly, identifying a talented Masters or Post-Doctoral researcher and engaging them through an internship is one way to secure skills on your team. Another alternative is to look into UN Volunteers (UNV), which hosts the Online Volunteering Service and provides connection to volunteer professionals, including data scientists and engineers who apply to work remotely on discrete projects or tasks. A volunteer may sign on as part of a project, or may just provide early guidance and support. UNDP and Global Pulse have regularly sourced expert collaborators and volunteers through the the Online Volunteering service, available at: <https://www.onlinevolunteering.org/en>.

HACKATHONS AND DATADIVES: You may meet data enthusiasts and experts in your community through attending a hackathon or data dive (events that attract data scientists, youth, and social innovators who put their skills to work to develop a rapid prototype or solution). In addition to recruitment potential, such events help develop civil society and build connections between governments and the people.

DATA ANALYTICS COMPANIES: There might be private sector companies able to fill your data expert roles—as a whole package or by providing access to specific software and tools. The field is in constant evolution with both big players and small start-ups. Some companies have data-for-good pro-bono programmes.



TIP: DON'T GIVE UP! Aligning your project allies and/or partners is one of the most time-intensive parts of the entire project. It should begin in the early planning stages, and will likely extend concurrently with the planning and even into the proof-of-concept phase. Successful partnerships are built on trust and take time.



CASE STUDY:

Engaging multiple partners to succeed

Natural disasters affect hundreds of millions of people worldwide every year. Emergency response efforts depend on the availability of timely information, such as the movement and communication behaviours of affected populations. As such, analysis of aggregated and anonymized Call Detail Records (CDRs) collected by mobile phone operators can reveal new, real-time insights about human behaviour during such critical events.

In this study, the close collaboration between the government, a privately-owned mobile carrier, academics, UN agencies on the ground, and Global Pulse—as well as the availability of public data sources—made it possible to show that the **patterns of mobile phone activity in affected locations during and after floods could be used as indicators** of:

- 1) FLOODING IMPACT ON INFRASTRUCTURE AND POPULATION
- 2) PUBLIC AWARENESS OF THE DISASTER

This was done by **combining mobile phone activity data with remote sensing data** to understand how people communicated during severe flooding in the Mexican state of Tabasco in 2009, in order to explore ways that mobile data can be used to improve disaster response. The representativeness of the research was validated by comparing the mobile data with official population census data.

For more information reach out to UN Global Pulse (info@unglobalpulse.org)



SECTION 3: CREATE THE WORK PLAN

Your data innovation project needs a strong, well-documented plan. It's a bulwark against project delays and miscommunications. It should also allow for project iteration and adjustment through the research, and will likely be a long-term work-in-progress, as stakeholders provide feedback and new partnership agreements are reached.

Specific elements of your work plan will vary, based on your personal preferences and your organizational and regulatory needs. Therefore, this section is not a comprehensive guide to creating a work plan, but rather a supporting tool. It offers a focus on the most vital elements of a work plan, reminders about elements that are commonly forgotten, and keys to success across the board.

OBJECTIVES:

- Develop a concept note
- Establish a privacy protection plan
- Finalize all of the details

MODULE 1:

DEVELOP A CONCEPT NOTE

Throughout the project, you will need a concise but detailed summary of the project so that all parties involved can stay aligned. This will build on the hypothesis you created in Section 1, and will likely be refined throughout project design, as stakeholders share expertise and work through the specifics of logistical needs.

For example, partner roles and responsibilities may not be formalized until all have signed on, and advice from a privacy specialist may be necessary for solidifying the privacy protection plan. Still, it is important for the concept note to be comprehensive from the beginning, even while noting that certain specifics are undecided.

Your **concept note** may include sections which cover some of the following:

- Problem statement and Background
- Objective
- Deliverables and Outcomes
- Scope of project (e.g. geographic, target population, languages, time scale, topics)
- Data (what data will be used in the project? List: programmatic data, new data sources etc.)
- Methodology (what analytical approach will be tested? Refer to previous reseach if applicable)
- Team (list organizations/ offices involved and partners)
- Modus Operandi (how will the project be executed? List roles and responsibilities)
- Timeline (may include phases and milestones)
- Budget / Resources

ESTIMATE A COMFORTABLE TIMELINE

Data innovation projects are not fast. Some projects can yield early positive results that can begin feeding into other programmes, but developing a working prototype of a new data solution will usually take at least between six and twelve months from the first inception of the idea. Some elements of the project can't be sped up, such as coordinating between stakeholders, and processing data. Set realistic expectations in the work plan.

DRAFT AN ITEMIZED BUDGET

The budget is a crucial piece of the concept note. Overestimate to start, and then look for ways (such as through partnerships) to secure donated or shared resources to save costs. Some typical budget line items are:

- **Personnel:** Data scientists; data engineers; domain experts; data visualization specialists; legal and data privacy expert; graphic designer; and others, depending on organizational capacity
- **Data access:** May be rented, purchased on an honorarium, or donated
- **Equipment:** Workstations; software licenses; data storage and processing; etc; much will depend on whether data analysis will be done internally
- **Office costs:** Office rent, utilities, events, printing, and communications costs
- **Travel:** Site visits to problem focus areas, meetings with distributed teams



CASE STUDY: *Timeline management for wide recognition*

SDG 16, which calls for “peace, justice, and strong institutions,” is one of the most difficult to measure. Most countries do not have a baseline for the goals, as factors like corruption just don’t have strong indicator precedents. Tunisia is part of a group of pilot countries working to find ways to track and evaluate progress on SDG 16. Within this pioneering work, the Tunisia National Statistics Institute and the UNDP Tunisia office saw an opportunity for data innovation. Having completed the first household survey on governance and democracy, the team explored complementary data sources.

Social media presented an exciting opportunity. While “corruption” may not be specifically quantifiable, public sentiment can be one indicator—and people are vocal on social media about their perceptions of corruption. The team researched whether public sentiment could be effectively measured by analyzing keywords in Twitter messages. By pulling tweets from the same time frame in which they had conducted the household survey, they were able to reach a clear correlation between the two data sources, indicating the value in further investigation.

COMPLEMENTARY DATA SOURCES OFFER THE FULL PICTURE

The household survey offers a static snapshot of a particular time and place; the team is continuing to investigate whether social media can be regarded as a “heart-rate monitor,” providing a real-time diagnosis of ongoing changes and variability.

KEYS TO SUCCESS

The project enjoyed a thoughtful timeline, with clear expectations for results that led to further opportunities. Clearly, the team is working toward long-term goals—the establishment of SDG baselines. But the team wisely set up short-term results to share, based on the results of a three-month period analysis, taking advantage of existing software tools and coinciding with the household survey period. While work continues, this early result from the proof-of-concept has allowed the team to communicate about the project for the benefit of a global field. Because of the initiative, Tunisia has been widely recognized as a pioneer in using big data for SDG measurement, and was invited to be part of a working group on big data and statistics within the UN. This project has led the National Statistical Institute of Tunisia to foresee the establishment of a big data team.

<http://europeandcis.undp.org/blog/2015/11/25/diagnose-and-treat-measuring-a-countrys-pulse-with-social-media/>

For more information reach out to Eduardo Lopez-Mancisidor (eduardo.lopez-mancisidor@undp.org) from the UNDP Tunisia office.

MODULE 2:

ENSURE DATA PROTECTION AND PRIVACY

The complexities of many data innovation projects can put fundamental human rights, including the right to privacy, at risk, sometimes in unexpected ways; as such, ensuring that you have a comprehensive data protection plan that includes guidance or policy on data handling, risk assessment, and risk management must be a primary concern of any project from the beginning through to the end, including when you publish the outcomes of your projects.

It can be frustrating to wait until proper legal data protection measures are put in place when a dataset is available to use, and holds insights that can address a development challenge. Yet privacy and ethical concerns must always come first.

It is important to have your data privacy and data protection plan in place for your project. Some stakeholders and potential partners will also want to see your plan as they are evaluating participation in your project; consider this plan one of your valuable assets. Your data protection plan must highlight the core principles for an ethical and privacy protective analysis of data. Engage with a legal and privacy expert as soon as possible in the planning process, and throughout the project.



TIP: Ensure compliance!

Check if your organization has any of the following instruments, and ensure that your project is compliant:

- Data Privacy and Data Protection Policy or Guidelines
- Data Sensitivity and Classification Scheme
- Risk, Harms and Benefits Assessment and Risk Mitigation Plan
- Data Security Policy

TIP: Conduct a Risks, Harms and Benefits Assessment before the launch of a new or substantially changed project.

It is better not to do a big data project than to do one that could cause significant harm. Remember that harms can never be disproportionate to the expected positive impact. A Risks, Harms and Benefits Assessment should be an ironclad tool and a key step in your Data Protection Plan. The Assessment should be used through the life cycle of the project. Remember that an assessment of risks, harms, and benefits should be done taking into account the interests of the beneficiaries, the context, and any possible impacts on known or unknown individuals and groups of individuals. It should at minimum include considerations of such contextual factors as geography, culture, and societal and policy norms. An assessment should draw from conversations with the stakeholders, domain experts, and individuals who understand the context first hand.

What could go wrong?

When developing a Risks, Harms and Benefits Assessment, consider legal, political, and data security risks. Carefully think of the harms that can be caused to individuals or groups as a result of data misuse.

A few generic examples:

- The data analysis is used by third parties (including governments) to reach objectives that violate human rights, including privacy rights
- An internal server is hacked; this breach compromises and leaks sensitive data
- Insufficient communication about the project, including data protection practices, can lead to loss of trust and public criticism
- Misunderstanding of applicable data protection laws can lead to liability and challenge a central element of the project

Don't be deterred! These risks are not warnings, but inspiration for mitigation planning!



TIP: Complete the **Data Innovation Risk Assessment Tool** to help you understand whether you have considered the minimum steps in your data privacy and data protection thinking process. You can find a printable version of the checklist in the Annex (pp 73). Keep in mind that some questions may require you to consult with a privacy, legal or data security expert.

SPECIAL CONSIDERATIONS:

RE-IDENTIFICATION RISK

De-identification(Anonymization) of data is one of the most important and also widely used techniques to protect data. Depending on the sensitivity level, data can be anonymized in various ways (e.g., by removing direct identifiers from a data set (such as name, phone number, id number etc.), aggregating data to an approximate location or average demographic information).

However, some datasets can become easily re-identifiable, especially when combined with others. In many instances, an individual or groups of individuals can be re-identified by a virtue of unique characteristics. Unauthorized re-identification of data violates privacy and can also lead to serious harms (e.g., human rights abuses, legal action, fines etc.)

It is crucial to protect anonymity of data by using proper security and administrative safeguards. Assessing a likelihood of re-identification is also a key component of risk management. Moreover, in order to prevent a possibility of re-identification, data experts must commit to not re-identifying an anonymized data set.

Furthermore, those experts who worked on anonymizing a dataset should not be analyzing the same data set for the purposes of your data innovation project.

OPTIONS FOR DATA TRANSFER & STORAGE OF DATA

Planning for appropriate transfer and storage of data is critical. It is also part of the risk management process and is the key element to ensure proper data security. Managing and processing even an average-size dataset can require ten or twelve servers, so remember that your project will likely require some new resources. Specific details on data storage are beyond the scope of this guide, but here is a brief overview of some ways to access data, and some of the considerations involved. Depending on many factors such as data sensitivity, infrastructure, cost efficiency, the Data Expert and Data Holder should collaboratively agree on the best method of transfer.

OFF-SITE TRANSFER OF A DATA COPY: The Data Holder may provide a data copy for the re-search team to manage. This offers a lot of flexibility, but the cost and risk are both high, as the research team must cover infrastructure costs (servers, networks, software) and establish their own security protocols.

REMOTE ACCESS TO A FULL, ON-SITE DATA SET: The Data Holder could provide access to the research team through a secure communication channel (such as VPN). A Data Holder and Data Expert may agree to use a third party cloud storage provider. While slightly less flexible than receipt of an off-site copy, this arrangement keeps the cost of servers and infrastructure with the Data Holder. When selecting a cloud provider it is important to ensure they have appropriate data security. Make sure to check where a data server is located (including in a case of a third party cloud storage) and what local laws may apply to your data. For example, there may be restrictions on where the data can or cannot be transferred or data can be subject to investigations by governments under national laws.

IN-PERSON, ON-SITE ACCESS TO DATA: The Data Holder may allow members of the research team to work with the data from the physical location belonging to the Data Holder. While this requires additional logistics, it provides the highest data security, as there are no networks or channels for potential data breaches.

MODULE 3: PLAN HOW TO MEASURE AND SHARE YOUR RESULTS

Data innovation is an exciting and growing practice in the development field; the efforts of project managers worldwide are immensely valuable for the global community as we seek to understand and reach the potential of this innovative new science.

*Beyond the specific solution you develop, a project can create positive outcomes by **inspiring others through positive examples**, or by **contributing to the growing body of research around using big data**. Sharing specific analysis tools developed through Open Source code can create direct impact for other efforts.*

ESTABLISH METRICS & AN EVALUATION PLAN

Robust metrics and evaluation protocols ensure that programmes are working as planned. The work plan should include a clear method for evaluating success, both because it will help the project team make better decisions and because it will help the results resonate with stakeholders and high-level decision makers.

Plan to ask the hard questions: Could the same impact be generated through well-established methods? Is the method sustainable, effective, and efficient? These questions should not only come at the end; plan to evaluate and iterate through the entire course of the project.

REMEMBER THAT WORK IN ANY INNOVATION FIELD ALWAYS CARRIES SOME RISK. BE AWARE AND HONEST:

Even with everything perfectly planned and all stakeholders aligned, the project may still not work out. Some hypotheses are false, in which case no amount of data can help. Even if that is the case, there is great value in publishing the results and lessons learned—you’ll be helping save others from making the same mistake, and contributing to the broader field.



TIP: PLAN TO ITERATE! The best development projects are heavily iterative, and one of the great strengths of data innovation projects are the real-time information they provide. Use this resource. Plan to monitor and adjust the programme through the course of implementation based on new information.

DECIDE ON, AND ITEMIZE, USEFUL PROJECT OUTPUTS

As the project gets close to actual data gathering and analysis, it is important to plan for how the results will be gathered and shared. Think also about whether and how often these outputs will need to be updated; if so, will more funding be needed down the road? If the project implementation is leading to successful results, it might be useful to have prepared a funding proposal for project follow-up or scale.

To define outputs, start back at the problem identified at the beginning. Who are the people who are going to act on the insights developed through the project? What is most helpful for them?

It is likely that all stakeholders are imagining different outputs, at different levels of sophistication, depending on their own needs and perspectives. If expectations are not carefully managed, the outputs will disappoint, putting project funding and support at risk.

WHAT DO PROBLEM SOLVERS WANT?

The most important outputs will be those that best meet the needs of the Problem Solvers, those who are working on the problem. Work with Problem Solvers on the shape and format of outputs. Ask questions like:

- What technologies do you use on a normal day?
- How and when do you normally update practics and decision-making?
- How often will decisions be made if updates are automatically shown in a dashboard? Daily? Weekly? Monthly?

EXAMPLE TYPES OF OUTPUTS:

- A brief policy recommendation with visualizations to inform decision-making
- A report to inform decision-making and / or advance the debate on big data
- An operational dashboard to inform programmes
- An infographic to inform decision-making to communicate the findings to a larger public
- An academic paper with new analysis methodologies

ADDITIONAL POTENTIAL OUTCOMES:

- New or improved partnerships
- New in-house capabilities
- Open Source code (so that others can do similar analyses)
- Software tools
- The establishment of a data lab in your office or programme country
- New resource mobilization opportunities
- TORs and templates that can be used in future projects

CREATE A COMMUNICATIONS PLAN

Think carefully about the objectives of communicating about your data innovation project, and the framing.

If your project was the first of its kind for your organization, perhaps it could be beneficial to frame the experience as an experimental learning opportunity. Alternatively, if your data innovation project was part of a bigger and more established programme of work, then make sure the stakeholders of that programme are aware of your communications plan, and have the opportunity to weigh in. Your communication plan may be part of mitigating the risks of a data innovation project (for example, by pre-empting negative media coverage by clearly explaining the objectives).

The communications plan can have strategic benefits for your organization. Communicating about success can help attract attention and resources to an organization or project team, and may fulfill one of the goals pitched to secure permission for the project in the first place.

Remember that communication is an expertise, and may require additional team members or input.



TIP: COMMUNICATION AFFECTS PRIVACY! Communicating your insights and results may come with privacy implications. Include your communications plan in privacy conversations, and establish standards for what framing language you will use to ensure that you’re protecting the privacy of data subjects.



CASE STUDY:

Using data to save lives in natural disasters

When natural disasters strike, rescue services need to know where people are most at risk. The last population census in the former Yugoslav Republic of Macedonia was carried out in 2002, leaving a major gap in population data. This gap can create serious problems during emergencies. For example, during a winter storm in 2012, rescue helicopters were dispatched to remote villages only to discover that nobody was there. Having real-time tools in place could help capture the population’s mobility and ensure that those most-at risk are reached on time.

UNDP and the former Yugoslav Republic of Macedonia Crisis Management Center have recently completed integrated risk and hazard assessments in each of the 81 municipalities in the country. Using mobile data to complement all this work seemed the natural next step. The way people use mobile phones provides insights into patterns of behaviour on the ground that can be life-saving.

The team set out to study whether mobile phone calls could serve as proxy indicators for people’s mobility, as well as to understand how men and women use mobile phones differently (for clues into their diverse levels of vulnerability based on sex-disaggregated data).

Luckily, there is no lack of such data in the country, as the number of active SIM cards exceeds the number of citizens.

Soon, a Memorandum of Understanding will be signed between UNDP and all the major mobile operators in the country, paving the way to explore what the data can show about mobility trends and many other important issues (such as air pollution levels) that will support more informed decision-making to help meet SDG targets in the years to come.

For more information reach out to Vasko Popovski (vasko.popovski@undp.org) or Jasmina Belcovska Tasevska (jasmina.belcovska@undp.org) from the UNDP former Yugoslav Republic of Macedonia office.

CHECKPOINT B:
CREATE A PROOF-OF-CONCEPT

Once the work plan is fully established, your team will create a small proof-of-concept experiment to ensure that all of the pieces of your project can work and that your hypothesis is workable. This experiment should be the smallest effort possible, but that is significant enough to give you confidence in the usefulness of the proposed plan. An early milestone could be validating results against other historic data sources.

The proof-of-concept may be part of a longer process of securing funding and support for the project itself. But, aside from the occasional outstanding opportunities for partnerships, you should have your work plan fully established and in place before completing the proof-of-concept, so that you can move forward toward developing your data innovation solution.



CASE STUDY:

Understanding public perceptions on biofuels

Global Pulse and the Packard Foundation worked together on a study to analyse **how public perceptions of and attitudes towards biofuels in the UK and Germany evolved** over a period of three years from 2013-2015.

The team used Twitter data to explore changes in the balance between statements for and against the use of biofuels to understand whether the stated reasons for being against biofuels shifted with the emergence of advanced biofuels. A customized keyword taxonomy was developed to retrieve tweets referencing biofuels: 350,000 tweets from the UK and 35,000 from Germany were identified. However, findings proved mostly inconclusive, as there did not seem to be a trend with regard to people’s perceptions on the topic.

However, other insights emerged from the study such as the fact that in the UK, tweets pointed towards an increased focus on environmental and climate change related opposition towards biofuels, with less focus on the “food” versus “fuel” debate. Although findings did not yield the expected results of shedding light into the attitudes of people on biofuels, the project created a rich source of insights and replicable tools that others might use when exploring different datasets and trying to develop a proof-of-concept experiment.

For more information reach out to UN Global Pulse (info@unglobalpulse.org).

WHAT IF MY PROOF-OF-CONCEPT DOESN'T WORK?

No matter how carefully you’ve planned, there are always times when something goes wrong. It’s important to set expectations with your team and stakeholders throughout the process, especially before the proof-of-concept, for exactly this reason. Sometimes you need to access more or different data, sometimes you need to use different analysis techniques, and sometimes you need to go back to the drawing board. Usually, you will be able to at least retrace your steps to find what went wrong, rather than starting over from scratch. Work closely with your partners to identify how to iterate and keep moving forward.

CHECKPOINT COMPLETED?

Great! You’re ready to implement your **data innovation solution**.

**CONCLUSION:
DEVELOPING THE SOLUTION AND
CREATING A NEW NORMAL**

In a data innovation project, design and planning is a full half of the battle. The work described in the previous chapters represents months of effort and discussion. While some early results may have impact, the greatest rewards are still in the future, as the actual research and solution development phase begins. *This new phase requires practical expertise and contextual adjustment at a level of specificity that is beyond the scope of this guide.*

Whether your experiment with harnessing new data sources for development has been successful or not, you are most likely at a point where the implications of your work can affect both the programme team you are a part of, as well as various other organizational units. With the proof-of-concept, your contribution to your team is a highlight of a “new normal”—a new way to complement the established means of doing development with data innovation, in order to make better sense of policy issues, to generate solutions that may not have been otherwise intuitive and obvious, and to create a more efficient impact on the ground. The ways in which those learnings and implications are further integrated and established in an organization surpass the intentions and ambitions of this guide; that’s a topic to fill another guide entirely.

We do hope, though, that this guide serves its purpose as a starting point, equipping innovators with the perspective and knowledge needed to bring the power of data to their work. The “data revolution” is just beginning; how it takes shape will depend on the practitioners on the ground—the readers of this guide—whose work will set new precedents and will convince policymakers to adopt and act on new insights.

ANNEX

These pages contain the following tools for project development:

PROBLEM DEFINITION TOOL P.49
The example for this tool is located on p. 10

DATA GAPS TOOL P.52
The example for this tool is located on p. 12

STAKEHOLDER MAPPING TOOL P.57
The example for this tool is located on p. 14

DATA JOURNEY TOOL P.62
The example for this tool is located on p. 17

PROBLEM SOLVER'S WORKFLOW P.65

PROJECT CONCEPTING TOOL P.69

DATA INNOVATION RISK ASSESSMENT TOOL P.73

PROBLEM DEFINITION TOOL



START HERE

 **What** is the key issue you are trying to address and **why** is it important?

 **Environment:** What factors contribute to the problem?

 **People:** Who does it directly affect?



Current Problem Solvers: Who is already working on this issue?



At what time intervals are decisions about addressing this issue made by the Problem Solvers?



Existing data: What data, relevant to this problem, exist already?



In general, what is the **timeliness** and **level of granularity** (geographic? demographic?) of the existing data on the problem?



Potential new data sources for a solution:

Based on the factors above, what potential new or additional data sources could provide insights into this problem?

How granular or detailed is the data **geographically**? *(FILL IN CORRESPONDING CIRCLE)*

HIGH



MODERATE



LOW



NOTES:



How granular or detailed is the data **demographically**?*

HIGH



MEDIUM



LOW



NOTES:



***A reminder on sex-disaggregated data:** Achieving and monitoring the SDGs will require sex-disaggregated data, and many development organizations place a high priority on gender focus. If you end up using a dataset that can not be disaggregated by gender, explicitly specify this gap in the work plan, and account for potential bias in results.

How long is it **retained**?

Do the current Problem Solvers use it for decision-making, evaluation, or other?



//DATA INNOVATION FOR DEVELOPMENT GUIDE
DATA GAPS TOOL

//PROJECT NOTES:

STEP 1

List one kind of data or source of data in the top row of each column

Existing data on the problem

Look at any programmes that deal with this problem. What data currently exists that is used to support day-to-day operations? Long-term planning? Communicating with or persuading others? For Monitoring & Evaluation?

STEP 2

For each data source, answer the following questions

Is it **openly available**, or does it require **special permission** to access?

Is it **structured** or **unstructured**?

How often is it collected?



How granular or detailed is the data **geographically**? *(FILL IN CORRESPONDING CIRCLE)*

HIGH

MODERATE

LOW

NOTES:



How granular or detailed is the data **demographically**?*

HIGH

MEDIUM

LOW

NOTES:



***A reminder on sex-disaggregated data:** Achieving and monitoring the SDGs will require sex-disaggregated data, and many development organizations place a high priority on gender focus. If you end up using a dataset that can not be disaggregated by gender, explicitly specify this gap in the work plan, and account for potential bias in results.

How long is it **retained**?

Do the current Problem Solvers use it for decision-making, evaluation, or other?



STEP 3

Other organizational data:
What other data sources does your organization use to support day-to-day operations and long-term planning?

Is it **openly available**, or does it require **special permission** to access?

Is it **structured** or **unstructured**?

How often is it collected?



//DATA INNOVATION FOR DEVELOPMENT GUIDE
STAKEHOLDER MAPPING TOOL

STEP 1

Brainstorm and list all stakeholders

Include the people directly and indirectly affected by the problem; private- or public-sector institutions who have data; academic organizations with data science experience; political actors committed to addressing the problem; national statistics institutes or other involved government agencies.

Categorize each stakeholder as "Community"; "Problem Solver"; "Policymaker"; or "Other". In general, "Community" should be the first stakeholders you consider.

STAKEHOLDER	
<div></div>	
What is their influence over the problem?	What is their influence over the project?
How might this person benefit from the project?	How does data support this person's decision-making now?
What could this person do with better data on the problem?	What could they do to undermine the project?
What is the best way to keep them engaged?	



//DATA INNOVATION FOR DEVELOPMENT GUIDE
DATA GAPS TOOL

STEP 4

Color-code data to highlight gaps.

Looking at all of the data sources you have listed, use a marker or stickers to code, by color, the specific gaps in the data you need. Use the following indicators, each color indicating a different kind of flaw in the data source:



Red: Updated too infrequently

How often does your organization make decisions about a problem? If the data is updated less frequently than your decision-making cycle, circle it in red.



Blue: Not geographically specific

How geographically detailed do you need your visibility on the problem to be? If a data source is too broad (i.e., if it can not be localized to the needed level of detail), circle it in blue.



Green: Not disaggregated

What specific demographics or characteristics (i.e., sex, ethnicity) do you need to be able to understand about the people who experience the problem? If the data source can not be disaggregated to reveal this specificity, circle it in green.



Yellow: Otherwise flawed

Are there other reasons why you cannot use this data to address the problem? Highlight any other gaps in yellow.

If you find that you have an internal data source that does not need any of these color codes, that data source may already be suited to your decision-making cycle, and you may not need to reinvent the wheel with a new data innovation project!



STEP 1

Brainstorm and list all stakeholders

STAKEHOLDER

What is their influence over the problem?

What is their influence over the project?

How might this person benefit from the project?

How does data support this person's decision-making now?

What could this person do with better data on the problem?

What could they do to undermine the project?

What is the best way to keep them engaged?



STEP 1

Brainstorm and list all stakeholders

STAKEHOLDER

What is their influence over the problem?

What is their influence over the project?

How might this person benefit from the project?

How does data support this person's decision-making now?

What could this person do with better data on the problem?

What could they do to undermine the project?

What is the best way to keep them engaged?



STEP 1

Brainstorm and list all stakeholders

STAKEHOLDER

What is their influence over the problem?

What is their influence over the project?

How might this person benefit from the project?

How does data support this person's decision-making now?

What could this person do with better data on the problem?

What could they do to undermine the project?

What is the best way to keep them engaged?



STEP 2

Plot all stakeholders on the following map

Those in the center of the circle have the greatest influence, those on the perimeters have the least. Include brief notes about strategies for engagement or opportunities for support.



NOTES:



STEP 2

Working from ethnographic research or your knowledge of the issue, fill out the user's "starting point" below. **What is the typical journey of such an individual with the problem?**

STEP 3

Now, on each row of dots below, **plot the steps a person might take to address the problem**. Each touchpoint (where the person visits an office, fills out a form, talks to another person, or takes any other action) **should go on its own dot**.

•	•	•	•	•	•	•
•	•	•	•	•	•	•
•	•	•	•	•	•	•
•	•	•	•	•	•	•
•	•	•	•	•	•	•
•	•	•	•	•	•	•

STEP 4
What data is gathered at each touchpoint? Write it below.

STEP 6

What is the timeline of the actions? Plot it out here.

DATA JOURNEY TOOL



STEP 1



Fill in a short description of an individual who represents the beneficiary, community or target of a certain development intervention (ideally based on ethnographic research). What are his or her goals, constraints, motivations?

User: _____



Looking at your completed map, are there additional data gaps to add to the **Data Gaps Tool**?

Are there additional data sources available to help address the problem?

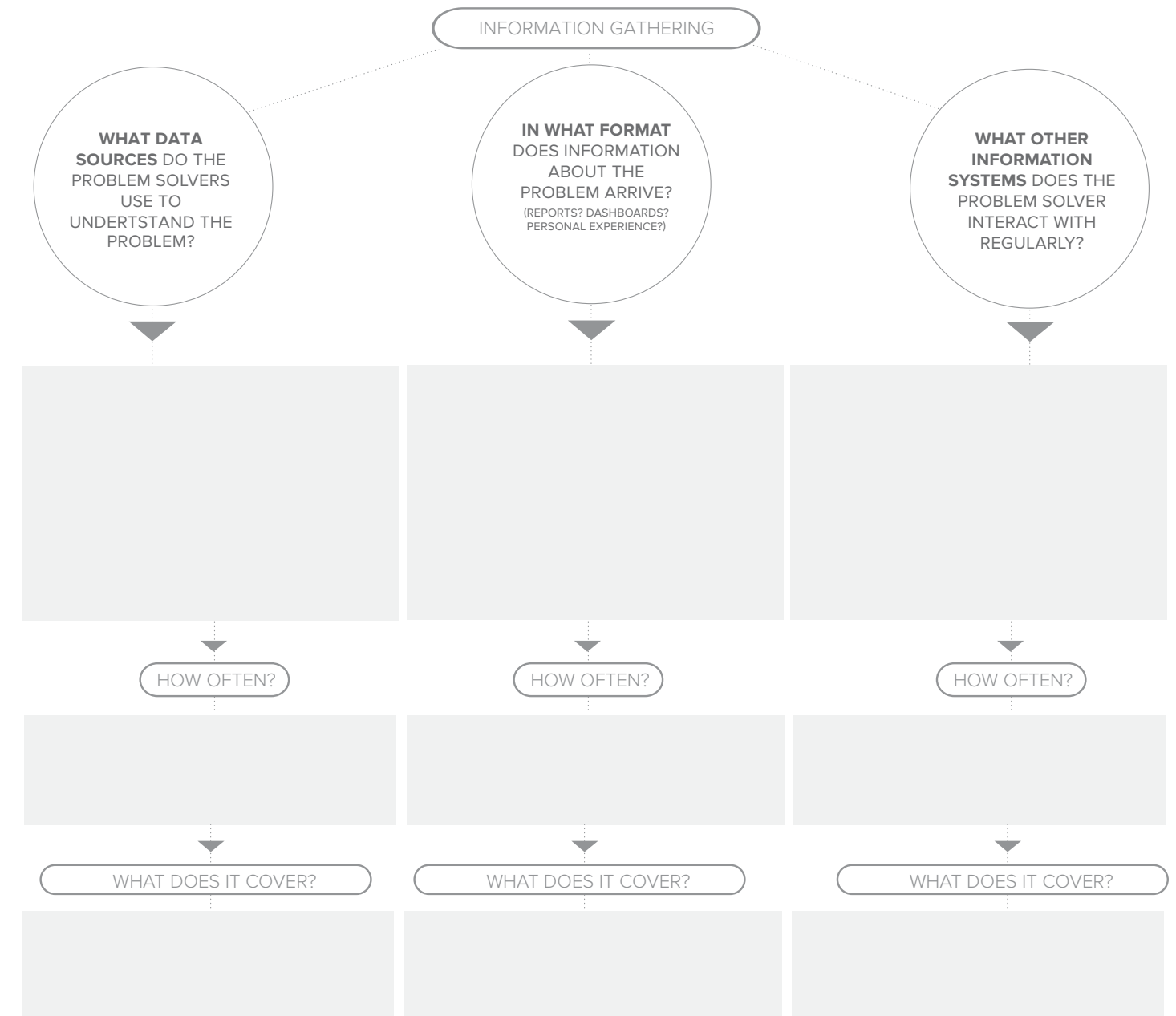
At what point in the journey would a data innovation **most help the Problem Solver or the communities deal with the problem?**

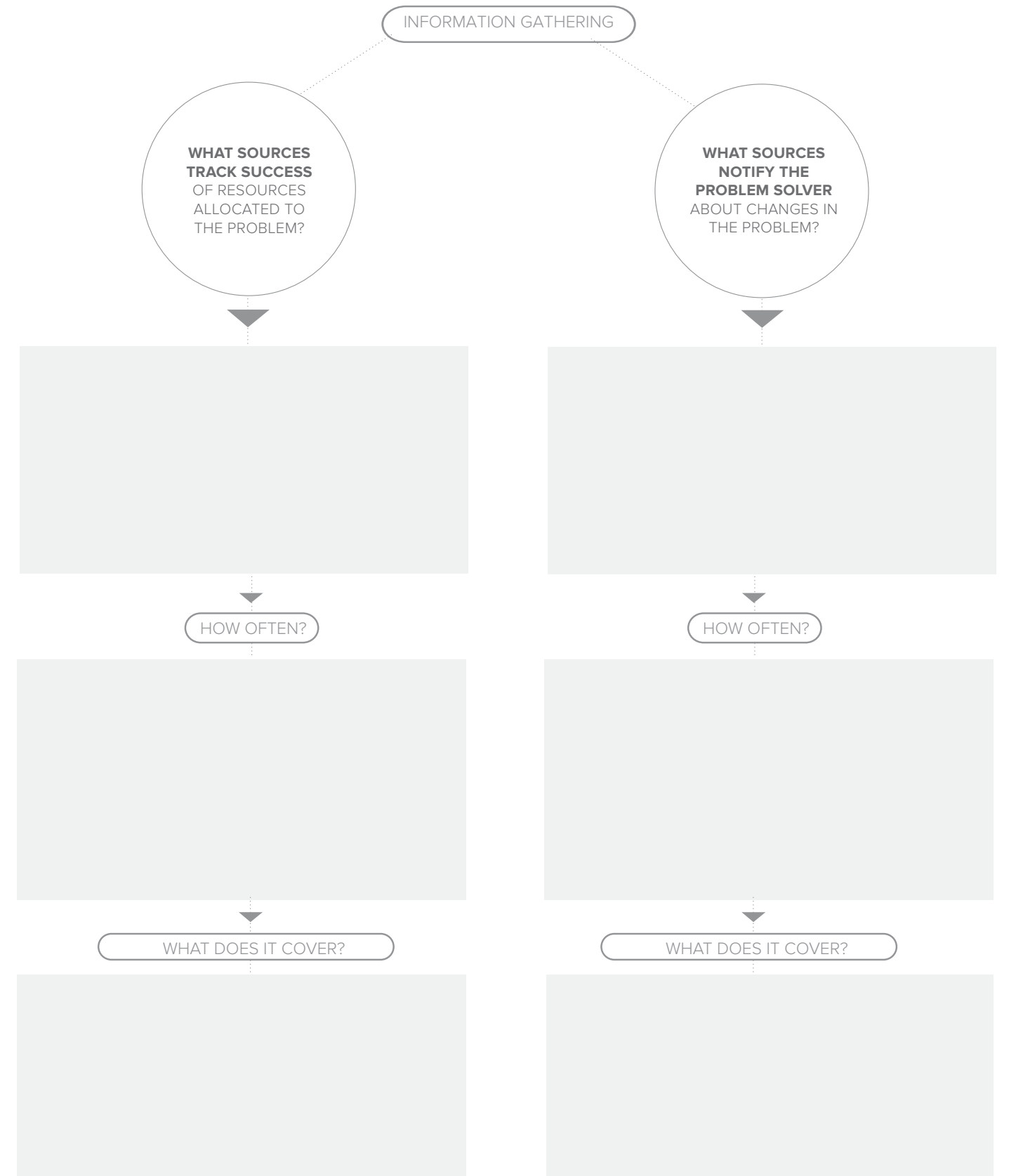


START HERE



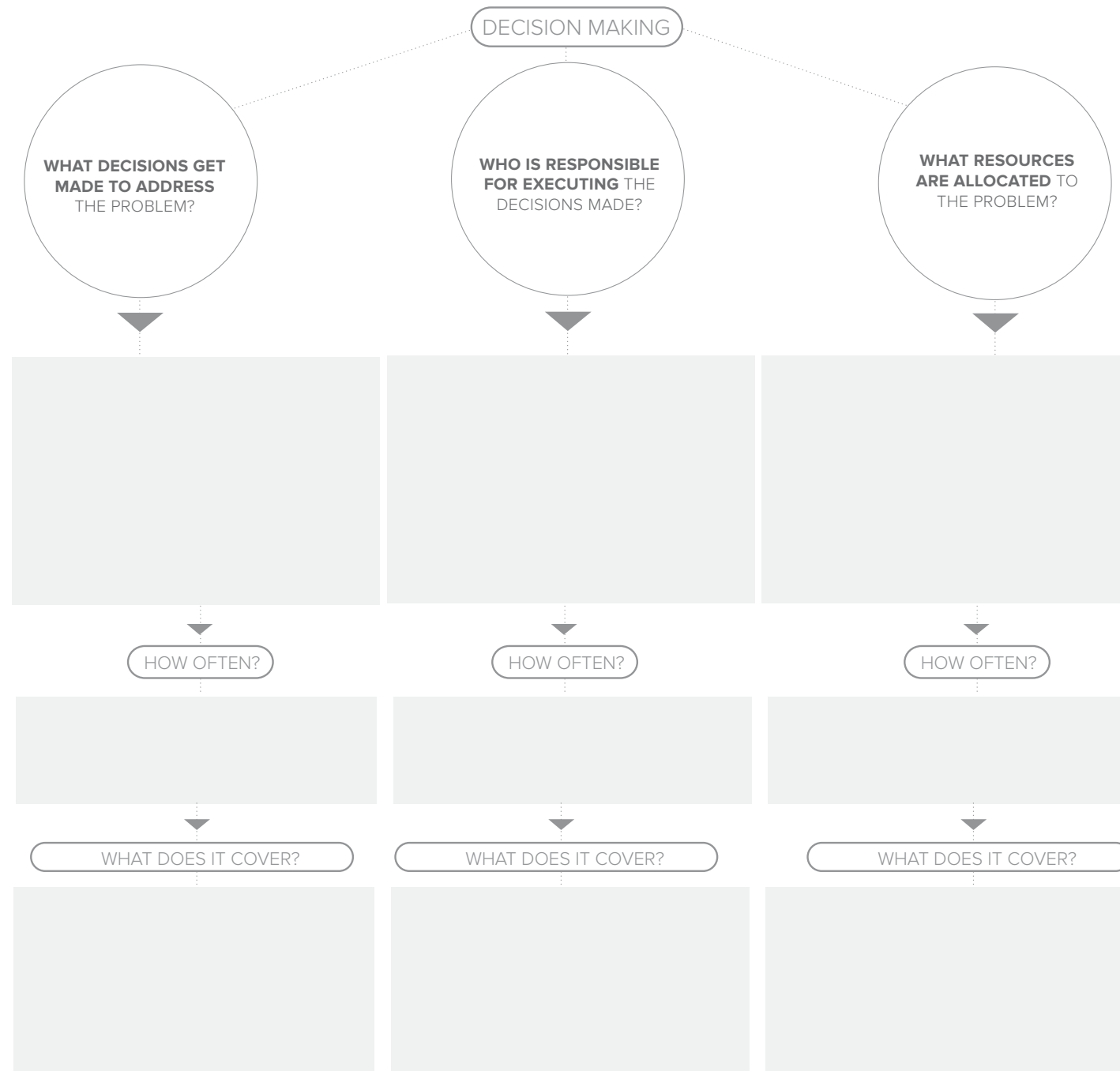
If you have time and it seems valuable, you may also want to **map how the Problem Solvers make decisions, use data, and deal with the problem in their day-to-day work**. This can help uncover further pain-points or useful leverage points, to integrate with the Data Journey Tool.







PROBLEM SOLVER'S WORKFLOW



LOOK AT YOUR ANSWERS: At what point in the process above would a data innovation most help the Problem Solver deal with the problem?

PROJECT CONCEPTING TOOL



START HERE



Your Aims: Reflect on Section 1, Modules 3 & 4

Who are you going to support through data innovation, and what do you hope they'll be able to do? **Write a concise and clear statement of the problem here.**

STEP 2



Look at your **Data Gaps Tool**, and at the Data Wish List you created in **Section 1, Module 5**.

Which data source will you pursue to reach your aims?
(You may have one or more)

Is the data **internal** or **external**?

--	--	--

What do you need, or **who** do you need to engage with, to access this data?

--	--	--

Briefly note other important data characteristics: Frequency, retention, granularity*, etc.

--	--	--

***Granularity refers to:** The scale or level of detail in a set of data



STEP 3

Refine your Policy Research Question:



Reminder: Connect the data sources you want to investigate to the indicators needed to address the problem:

Can we use **x** **data source(s)** for insights on **y** **problem indicator**?

x **data source(s):**

y **problem indicator**

--

--

Can we use _____ for insights on _____ ?



NOTES:

--

STEP 4

Refine your hypothesis:



Reminder: Specify how the data source will address the problem, and how you'll know if it works:

Since we know **[A facts]** about **[X data source(s)]**, we believe we can use **[B specific data]** to see **[C problem indicator]**. We will validate our results by comparing **[D existing data]**.

A Facts:

X Data Sources:

B Specific Data:

C Problem Indicator:

D Existing Data:

Since we know _____ about _____, we believe we can use _____ to see _____. We will validate our results by comparing _____.



NOTES:

//DATA INNOVATION FOR DEVELOPMENT GUIDE
DATA INNOVATION RISK ASSESSMENT TOOL

CHECKLIST



Rationale for the checklist: Large-scale social or behavioural data may not always contain directly identifiable personal data and/or may be derived from public sources. Nevertheless, its use could potentially cause harm to individuals.

Data use should be always assessed in light of its impact (negative or positive) on individual rights. This risk assessment tool (or checklist) outlines a set of minimum checkpoints, intended to help you to understand and minimize the risks of harms and maximize the positive impacts of a data innovation project (and is intended primarily for projects implemented within international development and humanitarian organizations).

How to use the checklist: The checklist should be considered before a new project is launched, when new sources of data or technology are being incorporated into an existing project, or when an existing project is substantially changed. In particular, this assessment should consider every stage of the project's data life cycle: data collection, data transmission, data analysis, data storage, and publication of results. If possible, the questions raised by the checklist should be considered by a diverse team comprised of the project leader as well as other subject matter experts, including – where reasonably practical – a representative of the individuals or groups of individuals who could be potentially affected. Consider consulting with data experts, data privacy experts, and legal experts so that they can assist with answering these questions and help to further mitigate potential risks, where necessary.

Note that the checklist was developed by Global Pulse as part of a more comprehensive Risk, Harms and Benefits Assessment, consisting of Two Steps: (I) Initial Assessment and (II) Comprehensive Risks, Harms and Benefits Assessment. This checklist is an Initial Assessment that should help to determine whether a more comprehensive Risk, Harms and Benefits Assessment should be conducted.

Nature of the checklist: This checklist is not a legal document, and is not based on any specific national law. It draws inspiration from international and regional frameworks concerning data privacy and data protection. The document provides only a minimum set of questions and guiding comments. The checklist and guiding comments are designed primarily as a general example for internal self-regulation. As this checklist offers only minimum guidance, you are encouraged to expand the list depending on the project's needs, risks, or specific context, or in response to the evolving data landscape.

Depending on the implementing organization (its legal status/nature) and applicable laws, the guiding principles, standards and basis for answering these questions may need to be changed.

The latest version of this checklist and the full version of the comprehensive assessment will be made available at a later stage (independently of this publication) and will be available at www.unglobalpulse.org/privacy. For more information or to provide input on the checklist, please contact dataprivacy@unglobalpulse.org. This checklist is a living document and will change over time in response to the evolving data landscape.

Instructions for completion

Please be sure to answer all of the questions by choosing at least one of the following answers: "Yes," "No," "Don't Know," or "Not Applicable." Please use the comments column to explain your decision where necessary.

For every "Not Applicable" answer, please provide an explanation in the comments. Every "Don't Know" answer should be automatically considered a risk factor that requires further consultation with a domain expert before a project is undertaken. Once you have properly consulted with an expert regarding the issue, please be sure to go back to the checklist and change your answer in the form to finalize your checklist.

A final decision based on the checklist should not be made if there is any answer marked "Don't Know."



Part 1: Type of Data

Personal Data: For the purposes of this document, personal data means any data relating to an identified or identifiable individual, who can be identified, directly or indirectly, by means reasonably likely to be used related to that data, including where an individual can be identified from linking the data to other data or information reasonably available in any form or medium. If you are using publicly available data, note that this data can also be personal, and therefore may involve some of the same considerations as non-public personal data.

1.1 Will you use (e.g. collect, store, transmit, analyse etc.) data that directly identifies individuals?

Personal data directly relating to an identified or identifiable individual may include, for example, name, date of birth, gender, age, location, user name, phone number, email address, ID/social security number, IP address, device identifiers, account numbers etc.

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

1.2 Will you use data that does not directly identify an individual, but that could be used to single out a unique individual by applying existing and readily accessible means and technologies?

Keep in mind that de-identified data (e.g., where all personal identifiers - such as name, date of birth, exact location, etc. - are removed), while not directly linked to an individual(s) or group(s) of individuals, can still single out an individual(s) or group(s) of individuals with the use of adequate technology, skills, and intent, and thus may require the same level of protection as explicit personal data. To determine whether an individual(s) or group(s) of individuals is identifiable, consider all of the means reasonably likely to be used to single out an individual or group(s) of individuals. Factors that influence a likelihood of re-identification include availability of expertise, costs, amount of time required for re-identification and reasonably and commercially available technology.

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

1.3 Will you use sensitive data?

Any data related to (i) racial or ethnic origin, (ii) political opinions, (iii) trade union association, (iv) religious beliefs or other beliefs of a similar nature, (v) physical or mental health or condition (or any genetic data), (vi) sexual orientation; (vii) the commission or alleged commission of any offence, (viii) any information regarding judicial proceedings, (ix) any financial data, or any information concerning (x) children; (xi) individual(s) or group(s) of individuals, who face any risks of harm (physical, emotional, economical etc.) should be considered as sensitive data. Consider that the risk of harm is much higher for sensitive data and stricter measures for protection should apply if such data is explicit personal data or is reasonably likely to identify an individual(s) or a group of individuals.

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

NEXT STEP:

As you go through the remaining sets of questions, please keep the data type you identified in the section above in mind. If you answered "YES" to at least one of the question above, the risk of harms is increased.

Part 2: Data Access

2.1 Means for data access

This question aims to help you understand the way in which you have obtained your data, to ensure that there is a legitimate and lawful basis for you to have access to the data in the first place. It is important to understand that whether directly or through a third party contract, data should be obtained, collected, analyzed or otherwise used in conformity with the purposes and principles of the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights and other applicable laws, including privacy laws.



How was the data obtained? (Choose from one of the answers below)

- ☐ A: Directly from individual(s) (e.g., survey)
- ☐ B: Through a data provider
(e.g. website, social media platform, telecom operator)
- ☐ C: Don't know

Comments:

NEXT STEP:

If you answered "A," please proceed to section 2.2, "Legitimacy, lawfulness and fairness of data access and use." If you selected "B," you can skip 2.2 and proceed to point 2.3, "Due Diligence on third party data providers." If you selected "C", consult with your legal expert before proceeding further.

2.2 Legitimacy, lawfulness and fairness of data access and use

Lawfulness, legitimacy, and fairness. Any personal data must be collected and otherwise used through lawful, legitimate, and fair means.

Personal data use may be based, for example, on one or more of the following legitimate bases, subject to applicable law: i) consent of the individual whose data is used; ii) authority of law; iii) the furtherance of international (intergovernmental) organizational mandates (e.g. in case where an international intergovernmental organization is the holder of the mandate and is the implementer of a data project); iv) other legitimate needs to protect the vital interest of an individual(s) or group(s) of individuals. Keep in mind that the legitimacy and lawfulness of your right to use the data must be carefully assessed, taking into account applicable law, the context, legal status of your organization; and the above bases (i- iv) are only included as examples for the purposes of this document.

Data should always be accessed, analyzed, or otherwise used taking into account the legitimate interests of those individuals whose data is being used. Specifically, to ensure that data use is fair, data should not be used in a way that violates human rights, or in any other ways that are likely to cause unjustified or adverse effects on any individual(s) or group(s) of individuals. It is recommended that the legitimacy and fairness of data use always be assessed taking into account the risks, harms, and benefits of data use.

Informed consent should be obtained prior to data collection or when the purpose of data re-use falls outside of the purpose for which consent was originally obtained. Keep in mind that in many instances consent may not be adequately informed. Thus, it is important to consider assessing the proportionality of risks, harms and benefits of data use even if consent has been obtained.

While there may be an opportunity to obtain consent at the time of data collection, re-use of data often presents difficulties for obtaining consent (e.g., in emergencies where you may no longer be in contact with the individuals concerned). In situations where it is not possible or reasonably practical to obtain informed consent, as a last resort, data experts may still consider using such data for the best or vital interest of an individual(s) or group(s) of individuals (e.g., to save their life, reunite families etc.). In such instances, any decision to proceed without consent must be based on an additional detailed assessment of risks, harms and benefits to justify such action and must be found fair, lawful, legitimate and in accordance with the principle of proportionality (e.g., any potential risks and harms should not be excessive in relation to the expected benefits of data use).

Do you have a legitimate basis for your data access and use?

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

2.3 Due diligence on third party data providers

This question usually applies when you are not a data collector, but rather obtained data from a third party (e.g. telecom operator, social media platform, web site). It is important that you verify, to the extent reasonably practical, whether your data provider has a legitimate basis to collect and share the data with you for the purposes of your project. For example, have you checked whether your data provider has obtained adequate consent (e.g. directly or indirectly through the online terms of use) or has another legitimate basis for sharing the data with you for the purposes compatible with your project? (See notes on "Lawfulness, legitimacy, and fairness" above)

Does your data provider have a legitimate basis to provide access to the data for the purpose of the project?

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:



Part 3: Data Use

3.1 Purpose specification

The purpose of data use should be legitimate and as narrowly defined as practically possible. Furthermore, requests or proposals for data access (or collection where applicable) should also be narrowly tailored to a specific purpose. The purpose of data access (or collection where applicable) should be articulated no later than the time of data access (or collection where applicable). In answering this question, concentrate on the reason why you need the data. Also, think about articulating your answer prior to or at the time of request for data.

Have you defined the purpose for which you will be using the data as narrowly, reasonably and practically as possible?

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

3.2 Purpose compatibility

Any data use must be compatible to the purposes for which it was obtained. Mere difference in purpose does not make your purpose incompatible. In determining compatibility consider, for example, how deviation from your original purpose may affect an individual(s) or group(s) of individuals; the type of data you are working with (e.g. public, sensitive or non-sensitive); measures taken to safeguard the identity of individuals whose data is used (e.g. anonymization, encryption). There must be a legitimate and fair basis for an incompatible deviation from the purpose for which the data was obtained. (See notes on "Lawfulness, legitimacy, and fairness" above)

Is the purpose for which you will be using the data compatible with the purpose for which you obtained the data?

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

3.3 Data minimization

Data access, analysis, or other use should be kept to the minimum amount necessary (to fulfill its legitimate purpose of use as noted in points 3.1 and 3.2). Data access, collection, analysis or other use should be necessary, adequate, and relevant in relation to the purposes for which the data has been obtained. Data should only be stored for as long as necessary, and any retention of data should be lawful, legitimate, and fair. The data should be deleted and destroyed at the conclusion of the necessary period. In answering this question, consider if at any point in time in your project cycle you have the minimum data necessary to fulfill the purpose of intended use.

Are all the data that you will be using (including its storage) necessary and not excessive?

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

3.4 Regulation and legal compliance

Make sure that you have obtained all regulatory and other required authorizations to proceed with the Project. (For example, the use of telecom data may be restricted under telecommunication laws, and additional authorizations may be needed from a telecommunication regulator; or the transfer of data from one country to another may need to comply with rules concerning trans-border data flows). Furthermore, to ensure that you have complied with the terms under which you have obtained the data, you should check existing agreements, licenses, terms of use on social media platforms or terms of consent. If you are uncertain about this question, you should consult with your privacy and legal expert.



Is your use of the data compliant with (a) applicable laws and (b) the terms under which you obtained the data?

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

3.5 Data quality

Data experts as well as domain experts should be consulted, if necessary, to determine the relevance and quality of data sets. Data accuracy must be checked for biases to avoid any adverse effects, including giving rise to unlawful and arbitrary discrimination.

Is your data adequate, accurate, up to date, reliable and relevant to the purpose of the project?

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

3.6 Data Security

Taking into account the available technology, cost of implementation and data type, robust technical, organizational safeguards and procedures, including efficient monitoring of data access and data breach notification procedures, should be implemented to prevent any unauthorized use, disclosure or breach of data. Embedding principles of privacy by design and employing privacy enhancing technologies during every stage of the data life cycle is recommended as a measure to ensure robust data protection. Note that proper security is necessary in every stage of your data use.

In considering security, special attention should be paid when data analysis is outsourced to subcontractors. Data access should be limited to authorized personnel, based on the need-to-know principle. Personnel should undergo regular and systematic data privacy and data security trainings. Prior to data use, the vulnerabilities of the security system (including data storage, way of transfer etc.) should be assessed.

When considering the vulnerability of your security, consider the factors that can help you identify "weaknesses" - such as intentional or unintentional unauthorized data leakage: (a) by a member of the project team; (b) by known third parties who have requested or may have access, or may be motivated to get access to misuse the data and information; or (c) by unknown third parties (e.g., due to the data or information release or publication strategy).

It is generally encouraged that personal data should be de-identified, where practically possible, including using such methods as aggregation, pseudonymization or masking, to help minimize any potential risks to privacy. To minimize the possibility of re-identification, de-identified data should not be analyzed or otherwise used by the same individuals who originally de-identified the data. It is important to ensure that the measures taken to protect the data do not compromise the data quality, including its accuracy and overall value for the intended use.

Have you employed appropriate and reasonable technical and administrative safeguards (e.g. strong security procedures, vulnerability assessments, encryption, de-identification of data, retention policies, confidentiality/non-disclosure, data handling agreements) to protect your data from intentional or unintentional disclosure, leakage or misuse?

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:



Part 4: Communication about your project

4.1 Transparency

Transparency is a key factor in helping to ensure accountability, and is generally encouraged. Transparency can be achieved via communication about your project (including providing adequate notice about the data use, as well as the principles and policies governing the data use). Making the outcomes of your data innovation project public can also be important for innovation.

Note, that making data (produced as an output of your project) open is an element of transparency. If you decide to make a data set open, you must conduct a separate assessment of risks, harms and benefits. In this case, you may also want to provide transparent notices on the process and applicable procedures for making the data set open.

Did or will you communicate about the data use (publicly or to other appropriate stakeholders)?

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

4.2 Level of transparency

Being transparent about data use (e.g., publishing data sets, publishing an organization’s data use practices, publishing the results of a data project, etc.) is generally encouraged when the benefits of being transparent are higher than the risks and possible harms. Also note, that level of detail (e.g., the level of aggregation) in a data set that is being made open should be determined after a proper assessment of risks and harms.

Particular attention should be paid to whether, for example, publishing non-sensitive details about a project or making non-identifiable datasets open can cause a mosaic effect with another open datasets. Accidental data linking or mosaic effect can make an individual(s) or group(s) of individuals identifiable or visible, thus exposing the individual(s) or group(s) of individuals to potential risks of harms.

Are there any risks and harms associated with the publication of the collected data or resulting reports and are they proportionately high compared to the benefits?

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

Part 5: Third Parties

5.1 Due diligence in selecting partner third parties (e.g., research partners and service providers, including cloud computing providers, etc.).

Frequently, data related initiatives require collaboration with third parties-data providers (to obtain data); data analytics companies (to assist with data analysis); and cloud or hosting companies (for computing and storage). It is therefore important that such potential collaborators are carefully chosen, through a proper due diligence vetting process that also includes minimum check points for data protection compliance, the presence of privacy policies, and fair and transparent data-related activities.

It is also important to ensure that third party collaborators are bound by necessary legal terms relating to data protection. These may include: non-disclosure agreements and other agreements containing appropriate terms on data handling; data incident history; adequate insurance, data transfer and data security conditions among other matters.

Cloud hosting. Many projects may use cloud or other hosting services, meaning that your organization does not maintain security of the hardware. It is important to ensure that your chosen cloud or hosting provider, and the data center in which they operate, have appropriate standards of security. Security certifications could be good evidence of your cloud provider's security compliance. When considering cloud storage and computing, take into account where the data will be actually located to understand potential vulnerabilities, compliance with laws, the special status of an implementing organization, including their privileges and immunities, where applicable, or rules concerning trans-border data flows.

Are your partners, if any, compliant with at least as strict standards and basic principles regarding data privacy and data protection as outlined in this checklist?

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

Part 6: Risks and Harms

Any risks and harms assessment should take into consideration the context of data use, including social, geographic, political, and religious factors. For example, analysis of the movement of vulnerable groups during humanitarian emergencies in conflict-affected zones could also be used by non-intend-ed users of data to target them with discrimination or persecution.

Any Risk, Harms and Benefits Assessment should consider the impact that data use may have on an individual(s) and/or group(s) of individuals, whether legally visible or not, and whether known or unknown at the time of data use.

When assessing your data use, consider how it affects individual rights. Rather than taking rights in opposition to each other, assessing the effect of data on individual rights in conjunction is recommended wherever possible. Use of data should be based on the principle of proportionality. In particular, any potential risks and harms should not be excessive in relation to the positive impacts (expected benefits) of data use. In answering questions 6.1 and 6.2 below also consider any potential risks and harms associated with (or that could result from) every "No" answer or "Don't Know" answer that you selected in the Sections above.

6.1 Risks: Does your use of data pose any risks of harms to individuals or groups of individuals, whether or not they can be directly identified, visible or known?

Risks should be assessed separately from harms. Note that not all risks may lead to harms. In answering this question, it is important to concentrate on the likely risks. Types of risks may vary depending on the context. For example, some of the risks that should be considered include data leakage, breach, unauthorized disclosure (intentional or unintentional), intentional data misuse beyond the purposes for which the data was obtained/or intended to be used by your organization, risk of re-identification or singling out, data not being complete or of good quality, etc.

Note that typically data analytics result in the production of a new data set. Such an outcome should be considered as a risk as well, and must be separately assessed for risks, harms and benefits before any further use/disclosure. Also, consider bias as a risk that can be produced as a result of data use. (In many cases, bias can negatively affect an individual(s) or group(s) of individuals and lead to harms).

If you have identified potential risks, please ensure to employ the necessary mitigation measures to reduce such risks to a minimum. Ensuring proper data security is one of many strong mitigation measures (see Section 3.6). If you do not know what kind of risks exist or whether the risks are likely, it is recommended that you perform a more comprehensive Risk, Harms and Benefits Assessment (as a Step 2).

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

6.2 Harms: Is your project unlikely to cause harm to individuals or groups of individuals, whether or not the individuals can be identified or known?

No one should be exposed to harm or undignified or discriminatory treatment as a consequence of data use. An assessment of harms should consider such key factors as i) the likelihood of occurrence of harms; ii) the potential magnitude of harms; iii) the potential severity of harms. The assessments should account for potential physical, emotional, or economic harms, as well as any harms that could result from infringement of individuals' rights.

Note that the risks of harms may be higher for sensitive data. Decisions concerning use of sensitive data may involve consultation with the individual(s) or a group(s) of individuals concerned (or their representative), where reasonably practical, to mitigate any risks. If you do not know what kind of harms exist or you have identified significant harms, try to perform a more comprehensive Risk, Harms and Benefits Assessment (as a Step 2 mentioned in the introduction section).

- ☐ Yes
- ☐ No
- ☐ Don't Know
- ☐ Not Applicable

Comments:

Part 7: Decision and rationale for decision

Final Assessment

Based on your answers in Sections 1-7, explain if the risks and resulting harms are disproportionately high compared to the expected positive impacts of this project.

Questions 1.1 – 1.3; 4.2; 6.1-6.2 answered as “Yes” mean that the risk is present.
Questions 2.1 – 2.3; 3.4-3.6; 4.1; 5.1 answered as “No” mean that the risk is present.

If you answered “Don’t Know” to any of the questions, consider it as a “risk factor”. You should not complete this assessment unless all questions are answered “Yes”, “No” or “Not Applicable”.

If you have answered "Not applicable", you should make sure that you explained why it is not applicable in the Comments column.

If you found any risks, you should assess the likelihood of the risks and likelihood, magnitude and severity of the resulting harms and make sure to mitigate them before the project is undertaken.

If you identify that some of the risks or harms are unclear, or high, then you should perform a more comprehensive Risk, Harms, Benefits Assessment as a Step 2 (as mentioned in the Introduction) and engage data security, privacy and legal experts.

If you have found that the likelihood of risks and harms is very low (or non-existent) in comparison to the probability of the positive impact, you should now proceed with your project. Always bear in mind you should implement as many mitigation measures for the identified risks (even if low).

Review team

Person who performed the assessment

This should be filled out and signed by the lead person responsible for conducting the assessment

Name:

Title:

Sign:

Comments:

People who participated in or reviewed the assessment (e.g. Project Lead, Data Security, Privacy, Legal Expert)

This should be filled out by those who assisted the lead person in making the decision or who have been consulted on specific questions raised above, if any (add additional reviewers, if necessary). You can indicate the specific questions that this person answered. If this person also helped to determine the final outcome of the overall assessment, please indicate in the comments section.

Name:

Title:

Sign:

Comments:

ACKNOWLEDGEMENTS

This publication would not be possible without the leadership, commitment, and investment of many colleagues:

The **UNDP team**, including Vasko Popovski of UNDP former Yugoslav Republic of Macedonia office who led the coordination of the joint “big data for development exploration journey,” under the supervision of ECIS Regional Hub’s Knowledge and Innovation team leader Milica Begovic, and RBAS Regional Hub’s Innovation Team Leader Jennifer Colville, with the support of Benjamin Kumpf from the UNDP Innovation Facility.

The “big data for development exploration journey” was made possible through funding support from the **UNDP Innovation Facility and the Government of Denmark**.

Colleagues from UNDP Offices, who shared successes and practical lessons learned from their experience initiating proof-of-concept projects, including David Svab (UNDP Kosovo), Vasko Popovski (UNDP former Yugoslav Republic of Macedonia), Max Perry-Wilson and Marina Mkhitarian (UNDP Armenia), Sherif El-Tokali and Nadine Abou El-Gheit (UNDP Egypt), Anisha Thapa and Jorg Kuhnel (UNDP Sudan), and Eduardo López-Mancisidor (UNDP Tunisia). And, Elise Bouvet from **UN Volunteers**, who has supported the UNDP country teams in finding suitable volunteers and given general guidance on how to utilize UN Volunteers’ services.

The **UN Global Pulse** colleagues who designed the methodology, conceptualized the guide, and provided expertise and solutions: Miguel Luengo-Oroz, René Clausen Nielsen, Anoush Tatevossian, Mila Romanoff, Felicia Vacarelu, Jeremy Boy, Olivia de Backer, Sara Cornish, George Hodge, Lodi Andrian, Robert Kirkpatrick, Makena Walker and colleagues from across the three Pulse Labs. Special thanks to Richard Coen, and members of Global Pulse’s Data Privacy Advisory Group, who contributed to the development of the risk assesment checklist.

Global Pulse also wishes to thank **the governments of Australia, Denmark, Indonesia, the Netherlands, and Sweden, as well as the David and Lucile Packard Foundation and the William and Flora Hewlett Foundation**, for supporting its network of Pulse Labs and various projects of the initiative which have contributed to the growing body of knowledge and community around big data for development innovation.

The **editorial and design team** who shepherded this guide into production: Kate Reed Petty, Dana Reinert.

How to cite this document:

UNDP, UN Global Pulse, 'A Guide to Data Innovation for Development: From Idea to Proof of Concept,' 2016

How to contact us:

UNDP Innovation Facility: innovator.support@undp.org

UN Global Pulse: info@unglobalpulse.org



*Empowered lives.
Resilient nations.*

The views expressed in this publication are from the authors and do not necessarily represent the view of the United Nations, UNDP or any of its affiliated organizations.