EPSILON – European Platform for Data Science: Incubation, Learning, Operations and Network

Training Material for Teaching and Self-Learning

Best Practices

Module 5/6

Hochschule Harz
Harz University of Applied Sciences
EPSILON
Funded by the European Union

Agenda I

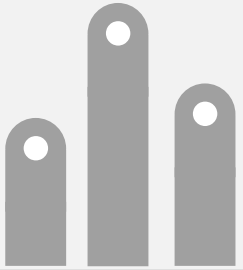
- What qualifies as a Data for Good project?
- Requirements for a successful project
- Data Maturity of an organization
- Finding Project Partners
- Characteristics of a typical Data for Good Project Team
- Project Scoping

Agenda II

- The typical Data for Good project lifecycle
- Best Practices for running a project
- Ethical and Legal concerns
- Typical Output
- Typical Organizational Structure
- Necessity of a legal entity

What qualifies as a Data for Good project?

A Guide created by the Data for Good Community CorrelAid.



What qualifies as a Data for Good project?

Partner

Legal Form
Is the legal form one we collaborate with?

Financial Situation
Does the partner have money to pay someone for this?

Purpose
Is the partner aligned with CorrelAids values?

Project

Educational Factor
Can our volunteers learn something in the project?

For-Good Factor and Ethical Considerations
Will this project have a positive impact for the partner and society as whole? Will it cause no harm?

Sustainability
Is the project sustainable for the partner?

Role of the Partner Organization
Do we develop someone else's core product/project?

Source: CorrelAid docs, n.d. a

Characteristics of the Partner I

What is the legal form of the potential partner?

- **Non-profit organisation / e.V.** ✔
  CorrelAid generally works with non-profit organizations (NPO). In Germany, these are usually associations (e.V.) with a specific charitable purpose.
- **Foundations** ❓
  Foundations are also a potential partner, depending on their size and financial resources.
- **Social entrepreneurs** ❓
  Social entrepreneurship is on the rise, leading to a whole range of "non-profit enterprises" (gGmbH etc.) that could also be considered.
- **Individuals** ❓
  Sometimes CorrelAid also supports people who are pursuing a for-good project on their own initiative, usually in addition to their job and open source.
- **Non-official initiatives** ❓
  Particularly in political contexts, there are often initiatives that are not (yet) organized in a legal form and are rather loose.

Source: CorrelAid docs, n.d. a

Characteristics of the Partner II

What is the legal form of the potential partner?

- **Political parties ❓**
  CorrelAid is non-partisan and typically **avoids working with** political parties or affiliated organizations. However, they encourage discussing ideas or connecting with like-minded people independently.
- **Academic institutions / individuals ❓**
  Collaborations with university professors, PhD students, or postdocs depend on the project's content and impact, as well as the financial situation.
- **For-profit companies ❌**
  CorrelAid **does not engage** in projects for for-profit companies unless in very special circumstances.

Source: CorrelAid docs, n.d. a

Financial Situation of the Partner

As a rule, CorrelAid's Data4Good projects do not incur any costs [...], apart from the costs for personal kick-off workshops.

This raises the question of whether projects should be carried out with organizations that have large budgets or are profit-oriented. Here are some questions to ask:

- **How many employees / how much budget do they have?**
  - NPOs (e.V.) and foundations up to ~10 – 15 full-time employees: ✔️
  - other circumstances: please check with the core team ❓
- **Could the organisation (realistically) pay someone for the work?** e.g.
  - a working student / student assistant?
  - an external consultant?
  - an employee?
  - If this is **clearly** the case and you get the impression that they are just looking for free labor, you should **not do** the project; CorrelAid, for example, does not want to replace paid work! ❌
- Another way is to ask: "**would/could this project happen without CorrelAid?**"
  - most NPOs will deny this so it's important to stay critical.

Purpose

Is the organisation aligned with your values?

- Any potential partner organizations or individuals should align with your values.
- You should do some research to find out who they are and what they do. For example, you could look at their website and social media accounts.

Characteristics of the Project I

What's a "good" CorrelAid project?

When shouldn't we do a project?

- A good Data for Good project should be beneficial for both parties:
  Volunteers get to apply and expand their knowledge and skills, and the nonprofit organization gets help with its data challenge.
- If it becomes apparent during the brainstorming process that the project would be extremely one-sided in either direction, you should politely decline.
- Examples of biased projects would be:
  - a project that would only be of interest to the volunteers but would not provide any real positive benefit to the NPO.
  - a project where CorrelAid volunteers are used for free work that is not rewarding in any way, e.g. a project that only consists of mundane data entry tasks. Every CorrelAid project should provide the opportunity to apply and learn useful data science skills.

Source: CorrelAid docs, n.d. a

Characteristics of the Project II

Educational Factor

- The most important question is: Can our volunteers really learn something in the project?
- Even if you find a project with a lot of data cleaning and a little bit of data visualization boring, it can be a very good learning experience for the less experienced data scientists.
- However, a project should not only consist of mundane tasks, such as data entry or organizing data in spreadsheets.

Source: CorrelAid docs, n.d. a

Characteristics of the Project III

For-Good Factor and Ethical Considerations

Does the project have a for-good factor?

- Will it positively influence and improve the work of the organization?
  In most cases, this will be the case for non-profit organizations. Most NPOs do not collect data for its own sake but have a clear goal in mind.
- Internal projects also should have a for-good factor, i.e. beyond learning about technologies, the project should also have a benefit. For example:
  - Open-source development
  - Deriving interesting insights from open data to answer social questions
  - Create data visualizations to draw attention to social issues

Source: CorrelAid docs, n.d. a

Are there any ethical considerations?

Will the project reinforce existing inequalities? e.g. racial or gender inequalities?

- yes; the project is not in line with our values, hence we don't do the project ❌
- not sure ❓
- no ✔️

Is it a technical, data-driven solution in the best interest of the affected people?

- no; the project is not in line with our values, hence we don't do the project ❌
- not sure ❓
- yes ✔️

Will the technical solution make moral/ethical decisions?

- yes / not sure ❓
- no ✔️

Will the project entail data from voice recordings, videos or photos of individuals?

- yes ❓
- no ✔️

Source: CorrelAid docs, n.d. a

GDPR (General Data Protection Regulation) related Questions:

- Does the organization have the permission of the data subjects to use and analyse the data?
  - no; organisations need the permission of the data subjects ❌
  - not sure ❓
  - yes ✔️
- Is the data content sensitive?
  The GDPR defines data which are particularly sensitive and are not to be analysed except in very certain circumstances.
  - yes ❓
  - no ✔️

Source: CorrelAid docs, n.d. a

CorrelAid Guide

Sustainability

Will the project be **sustainable** for the partner organization?

Will the project have an **impact beyond** the time the volunteer is available?

- Issues such as hosting, maintenance and handover should definitely be discussed. In the case of one-off and non-complex projects, such as the creation of a one-off report, this may not be necessary.
- **Example:** If a non-profit organization approaches you with a need for a dashboard to monitor their internal data processes, and a Shiny dashboard seems fitting, consider evaluating with them whether it is feasible for them to a) host it themselves, b) host it on shinyapps.io, or c) if it could be used only locally on the laptops of 1-2 people.

Source: CorrelAid docs, n.d. a

Scope

Is the project do-able with volunteers and in a limited amount of time?

- If the project idea seems quite extensive, it may make sense to divide it into smaller sub-projects or to propose an initial project to explore the idea and then a follow-up project.
- CorrelAid projects, for example, are characterized by the fact that they are time-limited, require a maximum commitment of 5 hours per week and generally do not last longer than 6 months.

Source: CorrelAid docs, n.d. a

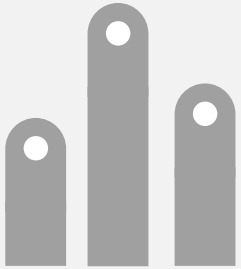
Role of the Project for Partner Organization

CorrelAid supports partner organizations with their data challenges. They like to work on challenges that have a real impact either on the work of their partners (e.g. improving their internal processes) or on the general public (e.g. developing meaningful data visualizations).

Source: CorrelAid docs, n.d. a

Requirements for a successful project

Blog post by the DSSG Summer Fellowship Team



Requirements for a successful project.

A solvable Problem

A challenging Problem

An Important problem with social impact

A motivated, capable and committed partner

Appropriate relevant Data

Source: Data Science for Social Good Foundation, 2021

Requirements for a successful project.

A solvable Problem

- Some problems are too big, too difficult or too complex to solve in a limited time frame.
- According to the DSSG, when potential partners bring up unsolvable problems, you can usually get around it by focusing on one aspect of the problem.

A challenging Problem

- A problem that is challenging gives the volunteers an opportunity to learn.
- Challenging problems encourage teamwork, spawn creative solutions and help data scientists to develop strong skills in solving real-world problems.
- It also creates an understanding and a passion for solving problems with social impact among the volunteers.

An important problem with social impact

- DSSG uses its limited resources to tackle significant and impactful problems.
- Projects should allign with the partner organization's operational needs and contribute to social welfare.
- DSSG prioritizes projects that benefit a larger number of people and address chronic issues

Requirements for a successful project.

A motivated, capable and committed partner

- A successful project requires a committed partner with expertise who decides how results are used.
- Partners provide valuable insights and support DSSG in developing solutions, which often requires significant effort and time.

Appropriate, relevant Data

- Obtaining required data is often the biggest challenge, especially when it involves sensitive information.
- Data sharing can be complex and time-consuming, sometimes taking months to arrange.
- DSSG adapts by working with anonymized data, conducting background checks, or using partners' internal systems.

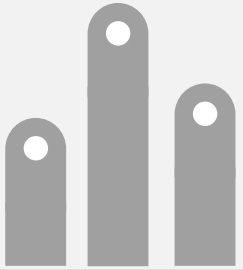
Fellowship Considerations

- Diversity of projects: DSSG selects a variety of projects to showcase the value of data-driven solutions, attracts diverse fellows, and enriches ideas through peer interaction.
- Long-term relationships: DSSG prefers to continue working with existing partner to enhance project impact and streamline cooperation, but they also seek strong new partners.

Source: Data Science for Social Good Foundation, 2021

The data maturity of the Organization

How to assess whether an organization has enough data to carry out Data for Good Project



Data Maturity

Measure if an organization has enough data to carry out a DSSG project.

- While sometimes a Data for Good project may be a consultancy for organizing data collection, a useful first step in understanding how mature an organization is, with respect to data, could be to apply some of the Data Maturity frameworks.
  - One such Data Maturity framework is available from **Data Orchard**: https://www.dataorchard.org.uk/resources/data-maturity-framework
  - Another such framework is available from the **DSSG Foundation**: http://www.datasciencepublicpolicy.org/our-work/tools-guides/datamaturity/

These frameworks can help you build an idea of an organization's technological and data readiness, but always check with a senior data scientist individual cases of an organization.

Data Maturity Framework

What does Data Orchard mean by data?

- Data Maturity is an organization's ability to improve and increase its use of data.
- "Data" encompasses all types of information collected, stored, analysed and used by an organization, including
  - Information about people served (clients, customers, residents)
  - Services used and activities participated in
  - Financial information (costs, income, grants)
  - Employee and volunteer data
  - Outcomes/ impact measures
  - External data (population/ environmental needs)

Source: CorrelAid docs, n.d. a

Data Orchard

Data Maturity Framework

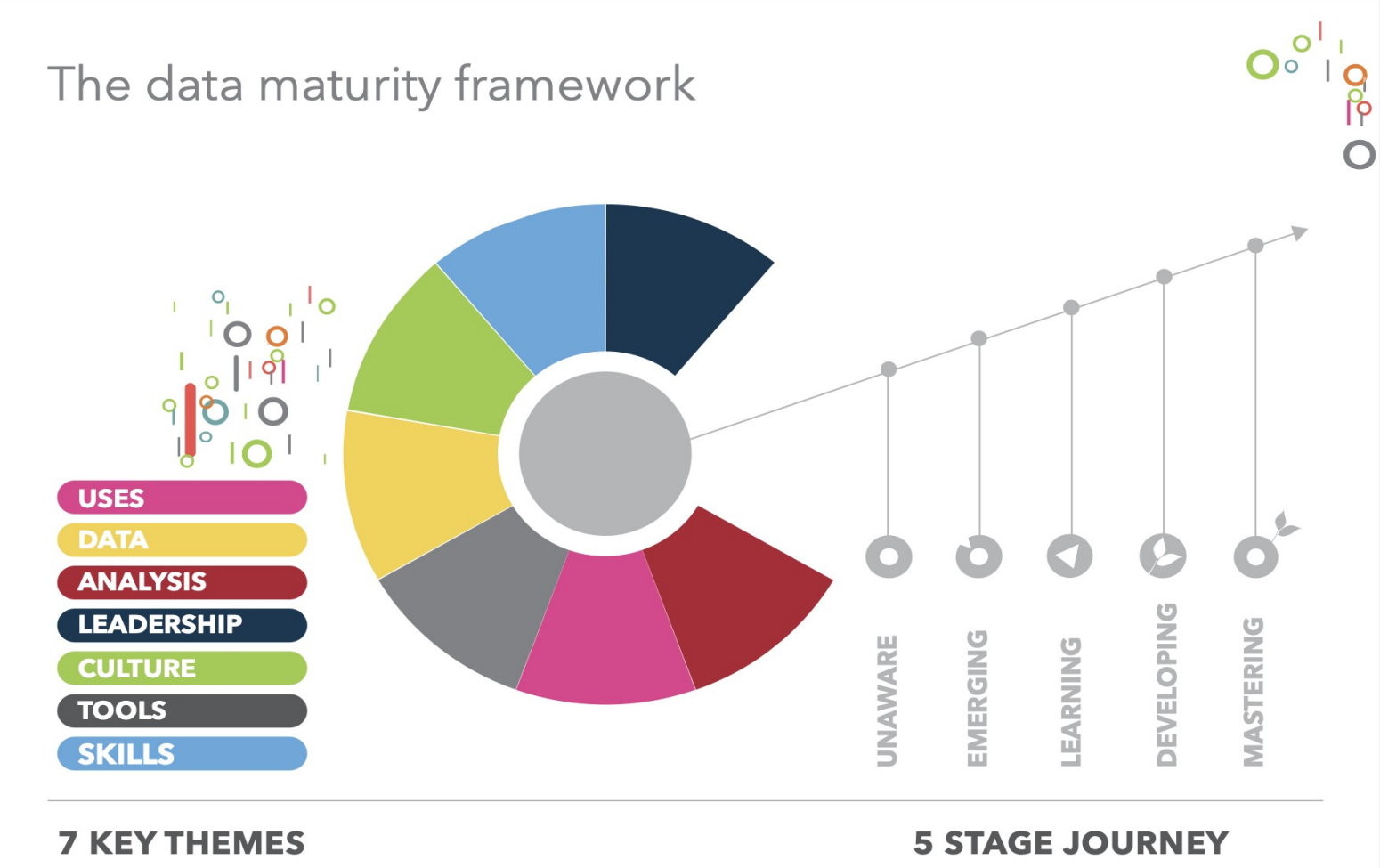
What is a Maturity Framework?

- A structured description of the characteristics
- Maturity frameworks aim to simplify complexity and make organizations measurable and comparable
- Most of them, like the one from Data Orchard, are used to promote best practise and ultimately serve the purpose of learning and improvement

Source: CorrelAid docs, n.d. a

Data Orchard

Data Maturity Framework



Source: Data Orchard, 2022

Data Orchard

Data Maturity Framework

The seven data maturity themes

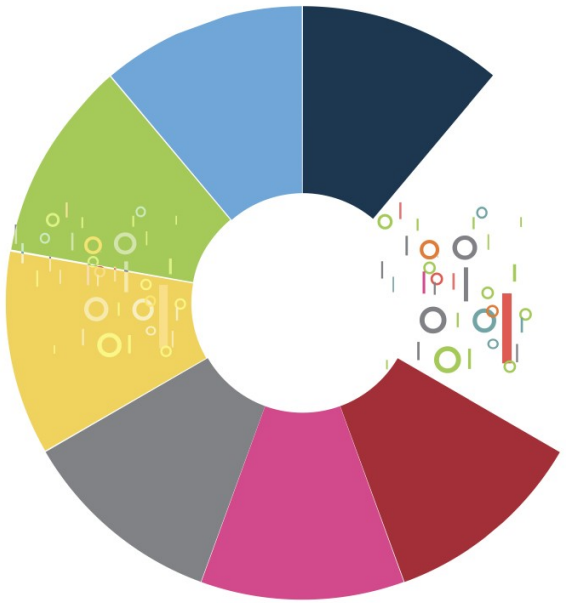
Each of the seven themes described in the framework cover a number of sub-themes.

USES
- Purposes for collecting and analysing
- Benefits and rewards

DATA
- Collection
- Quality
- Sources
- Assets

ANALYSIS
- Type
- Technique
- Joining
- Presenting

LEADERSHIP
- Attitudes
- Plans
- Capability
- Investment



CULTURE
- Team approach
- Self-questioning
- Openness
- Protection

TOOLS
- Collection
- Storage
- Organising and managing
- Analysis and reporting
- Integration and architecture

SKILLS
- Capacity
- Skills
- Training
- Access to knowledge and expertise

Source: Data Orchard, 2022

Data Orchard

Data Maturity Framework

Data Maturity Assessment Tool

- The six remaining data maturity themes can be viewed here:
- https://www.dataorchard.org.uk/resources/data-maturity-framework
- Data Orchard also offers a „**Data Maturity Assessment Tool**“, which measures the maturity of your organization
- The tool is free to use for very small organization or for individuals

Source: Data Orchard, 2022

Data Maturity Framework

In the data maturity framework from DSSG Foundation, you can see where you stand and how you can improve your organizational, technical and data-related readiness.

The Data Maturity Framework has three content areas:

- Problem Definition
- Data and Technology Readiness
- Organizational Readiness

The Data Maturity Frameworks consists of:

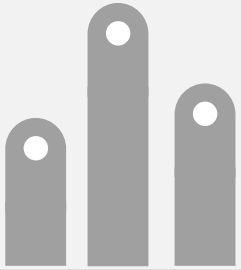
- A questionnaire and survey to assess readiness
- Data and Technology Readiness Matrix
- Organizational Readiness Matrix

The framework materials can be downloaded here:

http://www.datasciencepublicpolicy.org/wp-content/uploads/2018/05/Data_Maturity_Framework_4.28.16.pdf

Finding project partners

Blog post by the DSSG Summer Fellowship Team



Finding project partners

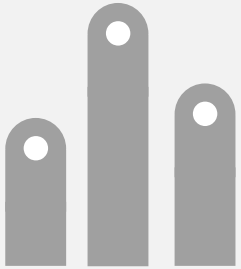
If you are an NGO, an intergovernmental organization, or a public agency, you can partner with:

- Volunteer organizations, either formal as associations or informal, that are accepting projects on an ongoing basis, like CorrelAid and their X organizations as well as DSSG Berlin or DSSG Portugal
- University affiliated programs, like DSSG Summer Fellowships, that have specific timings, but are also usually intense programs
  - There are many data scientists that are happy to invest their time to help good causes

Source: CorrelAid docs, n.d. a

Characteristics of a typical Data for Good project team

Blog post by the DSSG Summer Fellowship Team



Characteristics of a typical Data for Good project team

Data for Good projects generally have **two goals:**
Helping a partner organization apply data science to improve operations and educating aspiring data scientists

Examples:

DSSG Summer Fellowship: A typical team includes 1 project manager, 1 senior technical mentor, and 3 – 4 junior data scientists

DSSG Chapters: Volunteer-based projects have a project lead and up to 5 makers. Teams have flat, informal hierarchies with the project manager acting more as a moderator

CorrelAid: Teams include a Project Coordinator, Team Lead, and Team members, with spots reserved for trainees. The teams usually consist of 2 to 8 data scientists selected from applicants.

Source: DSSG PT, n.d. a; CorrelAid docs, n.d. b

Characteristics of a typical Data for Good project team

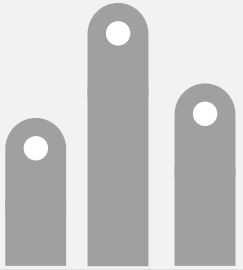
Other types of volunteer-based projects may feature **hackathons** where an objective is given and the volunteers collaborate or compete to create a solution, where teams are self-organized and the interaction with partner organizations is transactional and very short.

But most Data for Good projects require **significant time dedication** to reach a point where a positive effect is likely to occur.

Note that for each of the project teams approaches, Data for Good organizations emphasize deliberation about ensuring **diverse and inclusive teams.**

Project Scoping

CorrelAid – Scoping Guide



Project Scoping

Project scoping is a crucial task. CorrelAid estimates scoping to last up to 4 weeks.

Throughout the scoping phase, various topics require discussion with the non-profit organization (NPO):

1. Content and scope of the project
2. Expectation management & organization commitments
3. Data security / privacy & data access
4. Timeline
5. Team size & composition

Source: CorrelAid docs, n.d. c

Project Scoping

Communication & Note Taking

- Maintain ongoing dialogue with the NPO to ensure mutual interest and alignment on project goals
- It typically requires 2-4 iterations over 1-4 weeks to finalize a project suitable for our newsletter
- Use e-mail, phone, video calls, or in-person meetings. Ensure at least one personal conversation
- Initiate note-sharing early in the scoping process to create a shared record of plans and ideas

Source: CorrelAid docs, n.d. c

Project Scoping

Project Content and Scope

- Collaborate with the NPO to define project aspects and explore volunteer support opportunities
- Use CorrelAid's ehtics questionnaire to identify potential ethical concerns: Ethic Questionnaire
- Understand the situation deeply with resources like the Data Maturity Framework and a helpful question catalog: Question Catalog
- Focus on the organization's needs, set broad technology preferences, and allow flexibility for goal adjustments later in the project. Avoid excessive technical detail in the initial description

Project Scoping

Expectation Management & Organizational Commitments

- CorrelAid projects are volunteer-driven, with team members typically contributing 3 – 5 hours per week
- Volunteers may occasionally need to withdraw due to personal challenges
- CorrelAid is not a service provider; specific outcomes or tight timelines cannot be guaranteed
- If strict quality or deadlines are required, hiring a freelancer may be more appropriate
- The project should benefit both the NPO and volunteers, offering learning opportunities, including roles for less experienced data scientists

Source: CorrelAid docs, n.d. c

Project Scoping

Data Privacy & Access

- Early identification of data types is essential to ensure proper storage and processing measures
- Determine, if personal data will be involved:
  - **YES:** Typically, a pseudonymized version of the data is provided, where individuals could still be identified with additional information
  - Even if data is claimed to be anonymized, CorrelAid teams should not solely rely on this
- Consistent adherence to strict data privacy protocols is necessary when handling any form of personal data.

Source: CorrelAid docs, n.d. a

Project Scoping

Data Privacy & Access

Personal Data Involvement - NO:

- **No Personal Data:** Data protection measures depend on the NPO's preferences.
- **Key Questions:**
  - **Local Encryption Needed?**
    - Ensure the NPO is aware of risks like laptop theft.
    - If not required, the process may be simpler for team members, especially those in training.
  - GitHub for Data Storage?
  - Inform the NPO that GitHub servers are in the US, and CorrelAid does not self-host.
  - Private GitHub repositories restrict access to authorized team members and allow version control for easy collaboration.

Source: CorrelAid docs, n.d. a

Project Scoping

Data Privacy & Access

Public GitHub Repository - Open Source

- If the NPO agrees to publish code and data publicly, it will be accessible to everyone, benefiting the open-source community.
- Assist the NPO in choosing appropriate licenses for code and data. Use resources like choosealicense.com.

Get written confirmation from the NPO!
Always validate discussions held in person or over calls by sending a clearly formulated email to the NPO. Consider including the following table with the NPO's answers and request their written confirmation of the agreed privacy rules.

Source: CorrelAid docs, n.d. a

Project Scoping

Question	Answer	Consequence
Can the data be stored unencryped on the local machines?	✔	Team members **do not need** to use VeraCrypt or encrypt their home folder.
Can the data be stored unencrypted on the local machines?	✘	Team members **need** to use VeraCrypt or encrypt their home folder.
Can the data be uploaded to a private GitHub repository?	✔	Team members **can upload** raw and all kinds of processed data to GitHub. The initial data transfer to the project team can be done using GitHub.
Can the data be uploaded to a private GitHub repository?	✘	Team members **cannot upload**raw and processed data to GitHub. Instead, they should document relevant folder structures in the README of the repository and put the data folder in .gitignore. The initial data transfer to the project team **needs**to be done via the CorrelCloud.
Can the code and data be published to a public GitHub repository?	✔	The repository **can be public**. Appropriate licences for code and data **need** to be chosen.
Can the code and data be published to a public GitHub repository?	✘	The repository **cannot be public.**

Source: CorrelAid docs, n.d. c

Project Scoping

Timeline: Lastly, it's essential to agree on a general timeline. A typical project may follow a structure like:

Phase	Approximate Duration
Send out call for applications	
Collecting applications	1.5-2 weeks
Team selection	1 week
Onboarding + coordination of kickoff	1-5 weeks
Kickoff workshop	online event (unless otherwise organized)
Project work	1-6 months
Handover workshop	either online (1-3 hours) or a in-person meeting (1-3 hours)
Final closing event	1 hour (CorrelAid) public event where the team and the NPO present their result to interested CorrelAiders (or even the public)
Follow-up	immediately after handover workshop and after several months

Project Scoping

Timeline & Team Composition

Timeline Planning:

- Include buffer time for holidays and disruptions; not all volunteers may be available every week
- Align with key NPO deadlines (e.g., website launch, annual meeting) to provide a clear project endpoint
- Start projects in autumn or spring for better momentum; summer projects may face vacation-related challenges

Team Composition:

- Outline the desired team structure **before** calling for applications to align with project goals
- Typical team size: 4-6 people, but can vary based on project scope
- It's better to have more staff than too few

Source: CorrelAid docs, n.d. c

Project Scoping

Team Roles & Structure

Project Lead:

- Coordinates the team, serves as the main contact, and reports to the project coordinator
- Typically more experienced, but not always required

Team Member:

- Can range from beginner to experienced data scientist or analyst.

Team Trainee:

- Reserved for less experienced data scientists.

Typical Team Structure:

- 1 Project Lead
- 2-4 Team Members
- 1 Team Trainee

Source: CorrelAid docs, n.d. c

Project Scoping

Team Roles & Structure

Scoping Overview:

- Crucial to define project scope well at the **outset**, based on learnings from Data Science for Social Good.
- Begins **after** initial screening; iterative and refined throughout the project.

Key Participants:

- Business/Policy Experts
- Data Specialists
- End-Users/Decision Makers
- Those Affected by the System

Source: Center for Data Science and Public Policy, 2018b

Project Scoping

Step	Questions
Step 0: Problem Understanding	What is the problem? Who does it impact and how much? How is it being solved today and what are some of the gaps?
Step 1: Goals	What are the goals of the project? How will we know if our project is successful?
Step 2: Actions	What actions or interventions will this work inform?
Step 3: Data	What data do you have access to internally? What data do you need? What can you augment from external and/or public sources?
Step 4: Analysis	What analysis needs to be done? Does it involve description, detection, prediction, or behavior change? How will the analysis be validated?
Ethical Considerations	What are the privacy, transparency, discrimination/equity, and accountability issues around this project and how will you tackle them?
Additional Considerations	How will you deploy your analysis as a new system so that it can be updated and integrated into the organization's operations? How will you evaluate the new system in the field to make sure it accomplishes your goals? How will you monitor your system to make sure it continues to perform well over time?

Source: Center for Data Science and Public Policy, 2018b

Project Scoping

Iterative Scoping & Ethical Integration

Iterative Scoping Process:

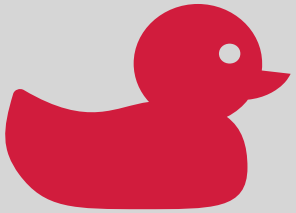
- The scoping process is iterative; steps like defining objectives and identifying measures may prompt reevaluation of previous steps
- Reviewing data may lead to redefining problems, goals, and actions, possibly restarting the scoping process
- Each step is an opportunity to reassess and refine the project scope

Ethical Considerations:

- Ethics should be integrated into every phase of scoping and project execution
- Ethical considerations are not an afterthought; they require continuous focus and involvement from all stakeholders, particularly those affected by the system

Source: Center for Data Science and Public Policy, 2018b

Project Scoping

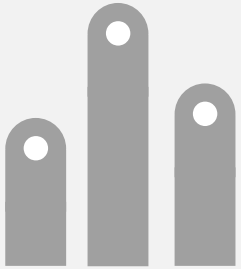


- Here you will find a worksheet aimed at non-profit organizations to plan actionable data science projects: http://www.datasciencepublicpolicy.org/wp-content/uploads/2021/09/ProjectScopingWorksheetBlank.pdf

- And here you can find the presentation „Data Science Project Scoping: A Guide for Social Good Organizations“: http://www.datasciencepublicpolicy.org/wp-content/uploads/2020/02/scopingslides.pdf

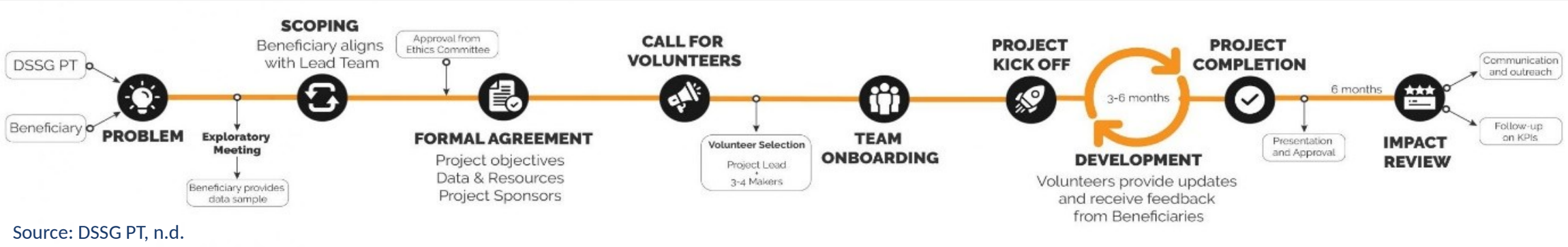
The typical Data for Good project lifecycle

Data maturity framework from DSSG Foundation



The typical Data for Good project lifecycle

A typical project lifecycle in a volunteer-driven organization is shown below:



Source: DSSG PT, n.d.

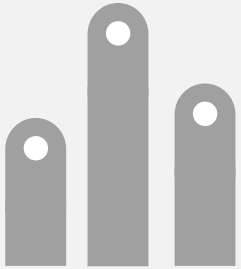
If you are an education institution that is interested in running DSSG-like summer fellowship, consider the DSSG summer fellowship curriculum here: https://github.com/dssg/hitchhikers-guide/tree/master/sources/curriculum.

The summer plan for the fellowship is available here: https://github.com/dssg/hitchhikers-guide/tree/master/sources/dssg-manual/summer-overview.

DSSG fellowship projects are typically challenging modeling projects. Other organizations may benefit more from the generic project lifecycle.

Best practices for running a project

A Guide created by the DSSG Organization CorrelAid.



Best practices for running a project

Best Practices & Resources

DSSG Best Practices: Explore a wealth of best practices on the DSSG summer fellowship GitHub: DSSG Curriculum

CorrelAid Best Practices:

- Establish a clear "definition of done" so all team members know when a task is truly complete
- Focus on explaining the "**why**" in your code documentation, not just the "**what**."
- Include "**how**" comments if the code is advanced or complex, especially for less experienced team members

Source: CorrelAid docs, n.d. d

Best practices for running a project

Data Visualization & Reporting

Best Practices:

- Beyond Bar and Box Plots: Essential reading for reporting projects. Concepts apply to Python as well: Beyond Bar and Box Plots
- R Specific: Comprehensive tips for beautiful plotting in R: ggplot2 Tutorial

R Projects – Best Practices:

- 1. General Best Practices:
  - **here:** Efficient file handling within your project
  - **renv:** Manage packages and maintain a consistent environment
  - **styler:** Customize and enforce code styling for consistency
  - **lintr**: Utilize linting for code quality

Best practices for running a project

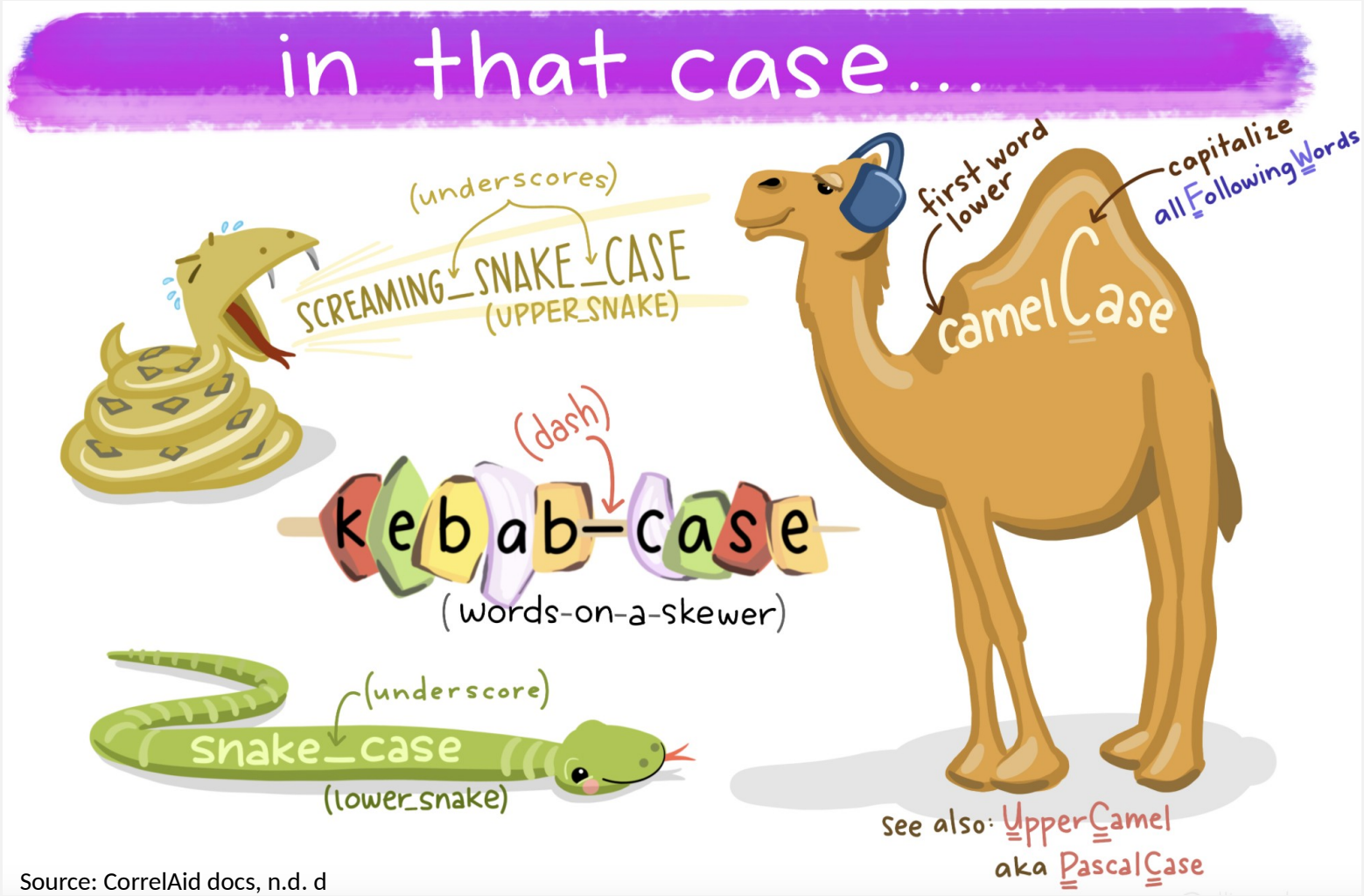
Data Visualization & Reporting

2. Organize Your Files and Code:

- Folder and File Organization:
  - Explore best practices for structuring R projects
- File Naming and Good Code Practices:
  - Project-Oriented Workflow
  - File Naming Guide
  - Project Structure Guide
- Naming Objects:
  - Use snake_case in R (e.g., my_obj) and camelCase in Shiny
  - Ensure consistency by agreeing on a naming convention within your team

Source: CorrelAid docs, n.d. d

Best practices for running a project



Source: CorrelAid docs, n.d. d

CorrelAid

Best practices for running a project

Data Visualization & Reporting

3. Reports - RMarkdown

- Use RMarkdown: Ensure reproducibility and version control in reports
- Alternative Formats: Microsoft Word may be acceptable for documentation intended for non-profit partners

Output Formats:

- **PDF:** Ideal for third-party reports, such as grant documentation
- **HTML:** Best for internal use with interactive elements
- **bookdown:** Suitable for longer, book-like reports

Resources:

- RMarkdown Tutorial: Beginner Guide
- RMarkdown Driven Development: Emily Riederer | Posit Video

Best practices for running a project

Data Visualization & Reporting

4. Interactive Dashboards

Flexdashboard (Without Shiny):

- Ideal for simple, semi-interactive dashboards
- Suitable for static or infrequently updated data with low interactivity demands

Shiny:

- Best for fully interactive dashboards
- Requires hosting after development unless local use is sufficient
- Consider maintainability—future support may be needed if issues arise after project completion.

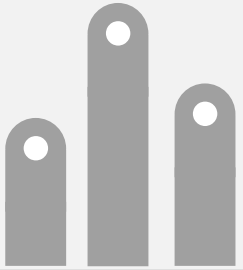
Flexboard with Shiny:

- Combines RMarkdown with Shiny for small interactive dashboards.

Source: CorrelAid docs, n.d. d

Ethical and legal concerns

What is a good way to deal with these concerns when interacting with Data for Good organizations or volunteers?



Ethical and legal concerns

Ethical Use of Data Science & Legal Considerations

Critical Importance:

- Address ethical and legal concerns, **especially** with personal data or bias-prone data
- If these cannot be managed, **pause** the project

Ethics Committees:

- Mature organizations like DSSG Portugal and CorrelAid have Ethics Committees and GDPR advisors to guide teams

Resources:

- DSSG Ethics Presentation: Download here
- Materials on Ethics, Bias, and Fairness: Access here

Consultation:

- In case of doubt, consult a Data Protection Officer or legal experts

Ethical and legal concerns

Data privacy and security

CorrelAid additionally provides a document for volunteers about the Data Security containing the following information:

- Types of Data
- Personal Data

> ‘ ‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person‘

Source: CorrelAid docs, n.d. e

Ethical and legal concerns

Anonymization & Pseudonymization

Only when personal data is fully anonymized does it fall outside the scope of the General Data Protection Regulation (GDPR). Anonymization involves altering the data so that individuals cannot be identified, making re-identification impossible. The standards for true anonymization are stringent, requiring an irreversible process to ensure that the data remains non-identifiable.

Pseudonymized data is data where identification of a person is possible **only** with the use of additional, separately kept information. Such data remains within the scope of GDPR, as there is still a potential for re-identification. Often, data referred to as "anonymized" may actually be pseudonymized according to GDPR standards, meaning it is still subject to regulation

Source: CorrelAid docs, n.d. e

Ethical and legal concerns

Non-Personal Data & Data Security

Non-Personal Data:

- E.g. data on organizational processes and institutions
- **Open Data:** Information available for unrestricted use, reuse, and distribution, with minimal conditions like attribution
- Non-personal data is not subject to the General Data Protection Regulation (GDPR)

Data Protection and Security Declaration:

- Depending on the project, all team members and the non-profit organization may need to sign a data security declaration
- CorrelAid's template: Data Security Declaration
- Customize the declaration based on specific project requirements
- The project coordinator should collect signed declarations from all team members before data sharing begins.

Source: CorrelAid docs, n.d. e

Ethical and legal concerns

Data Encryption

Responsibility: The project lead and team must ensure secure data storage on local machines.

Methods:

- **Encrypt Home Folder:** Secure data by encrypting the user's home directory
- **VeraCrypt:**
  - Use VeraCrypt to create encrypted containers that require a password for decryption
  - Decrypted containers appear as new drives, with all stored data automatically encrypted
  - Programs can access encrypted data in memory
  - Note: The container size cannot be changed after creation, so the project manager must estimate the required storage volume accurately.

Source: CorrelAid docs, n.d. e

Ethical and legal concerns

Encrypted home directory

Purpose: Encrypt project data to prevent unauthorized access if the device is lost

Automatic Encryption: Data is decrypted/encrypted upon user login/logout

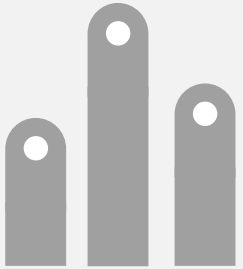
Operating System Options:

- **Windows:**
  - Home directory encryption is only available in Windows Pro and Enterprise editions
  - Solution: Windows users typically need to use VeraCrypt for encryption
- Mac:
  - Use FileVault (available from Mac OS X 10.4) to encrypt the home directory or entire hard disk
  - Caution: Losing both user and recovery passwords can result in irretrievable data loss
- Linux:
  - Most recent Linux distributions offer encryption during new user account setup via ecryptfs-utils
  - Alternative: If not enabled initially, use VeraCrypt as a fallback option

Source: CorrelAid docs, n.d. e

Typical output

DSSG Portugal & CorrelAid

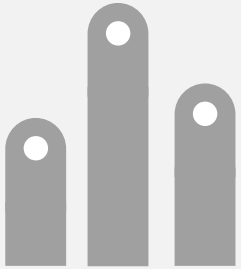


Typical output

What is a typical output of a Data for Good project and how is a project handed over?

- **Prototype Development:** Projects (up to 6 months) typically produce a prototype of a working model that requires further development and testing
- **Delivery Method:** Prototypes are usually shared via a GitHub repository
- **Handover Workshop:**
  - Includes detailed explanations for partners on code usage and maintenance
  - Provides guidance on the path for testing and deploying the model

Typical Organizational structure



Typical Organizational structure

DSSG Portugal is a non-profit organization, that connects volunteer data scientists with associations that benefit from the use of data

Source: DSSG PT, n.d.

Typical Organizational structure

DSSG Portugal: Beneficiaries and Impact

Beneficiaries: Public, governmental, private, or non-profit institutions working on socially impactful projects

Requirements for Beneficiaries:

- Problems to Solve: Clear issues that data can address
- Data Availability: Existing data or the potential to collect it
- Collaboration: Commitment to actively participate, crucial for project success

DSSG Portugal's Contributions:

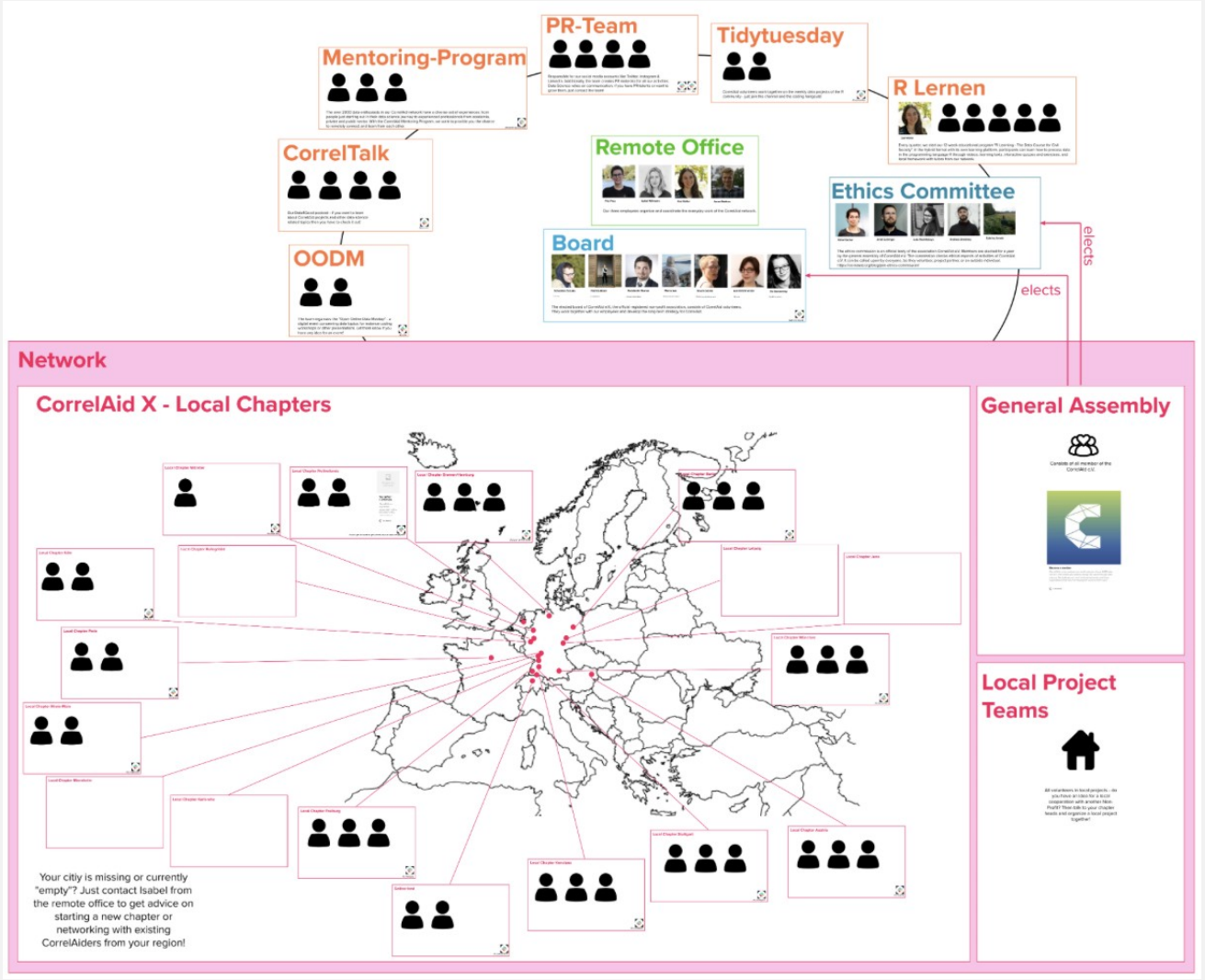
- Better resource utilization through data
- Data-driven analysis enhances accuracy
- Data insights can measure and enhance program impact

Source: DSSG PT, n.d.

Typical Organizational structure

CorrelAid is formed as an association.

In practice, the official association bodies only play a subordinate role. Most of the structures in which they work are informal, constantly changing and ad hoc.



Source: CorrelAid docs, n.d. f

Typical Organizational structure

CorrelAid Structure

The Network:

- Comprised of analysts who subscribe to the mailing list and participate in projects for civil society organizations
- Network members receive project announcements and information on how to engage with CorrelAid beyond individual projects
- Annual Data for Good Summit: The entire network is invited for workshops, networking, and collaboration

CorrelAid Teams:

- Internal work is organized into teams to streamline decision-making and structure activities
- Teams Include: Projects, Education, Fundraising, Infrastructure
- Teams are the first point of contact for those interested in organizational involvement and are re-evaluated annually during a strategy meeting

Source: CorrelAid docs, n.d. f

Typical Organizational structure

CorrelAid Structure

The Core Team:

- Includes all members actively involved in CorrelAid's organization
- A dynamic, flexible team with around 20 members currently
- Communication: Bi-weekly calls, Slack workspace, and an annual strategy meeting

The Project Teams:

- Tasked with project implementation, separate from the core team
- After defining a data project with an organization, a call for applications is published within the network
- Emphasis on assembling diverse teams with the necessary skills
- Projects begin with training on workflow and data security, followed by a kick-off workshop
- CorrelAid currently manages 5-10 projects per year, with plans to increase this number to expand their impact

Source: CorrelAid docs, n.d. f

Typical Organizational structure

CorrelAid Structure & Routine Standards

The Board:

- Elected annually at the Data for Good Conference during the general meeting
- Consists of seven members, with at least four being women
- Board members are dedicated to guiding the direction and development of CorrelAid, initiating and supporting future projects

Routine Standards:

- CorrelAid has established routine standards, stored as protected documents accessible within their drive
- Board members are responsible for regularly updating these standards
- The goal is to transparently document experiences and processes, easing the integration of new volunteers and enhancing the quality of CorrelAid's work

Source: CorrelAid docs, n.d. f

Typical Organizational structure

CorrelAid Communication

Slack Workspace:

- CorrelAid's primary communication tool, offering better file sharing and features than traditional group chats
- All online communication occurs within Slack, with channels for organizational tasks, technical support, and sharing interesting links and projects
- Most tasks are conducted in public Slack channels

Calls:

- **Weekly Calls: Held every Monday at 8 pm**
  - General Call (1st Monday): Provides current information and introductions for new members
  - Local Chapter Call (3rd Monday): Discusses current topics, projects, and challenges
- These calls are essential for decision-making and coordination, especially since CorrelAid team members are spread across Europe
- Participation is voluntary but crucial for staying informed

Source: CorrelAid docs, n.d. f

Typical Organizational structure

CorrelAid Meetings and strategic planning

Data-for-Good Summit:

- Annual Gathering: Weekend event for networking, learning, and planning future projects
- Schedule:
  - Friday Evening: Public event
  - Saturday: Data Science workshops
  - Sunday: CorrelAid development workshops
- Typically held in the fall, open to all participants

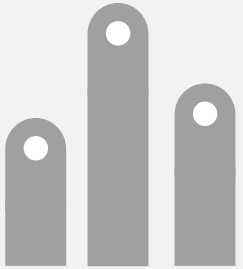
The Retreat:

- Core Team Meeting: Annual in-person meeting to reflect on the past year and plan upcoming internal projects
- Value: Facilitates valuable personal exchange beyond regular online communication

Source: CorrelAid docs, n.d. f

Necessity of a legal entity

CorrelAid – Finding a legal form



Necessity of a legal entity

CorrelAid Meetings and strategic planning

Many volunteer groups operate without a legal entity. However, a legal entity is needed to receive donations or enter legal agreements.

CorrelAid Example:

- As chapters grow, it may become necessary to establish a legal structure, especially for those outside of Germany
- CorrelAid has a license agreement to support this process
- Dutch Chapter: Chose to become its own foundation instead of operating under the German association due to legal and financial reasons
  - FridayChallenges: Cross-border charitable work is complicated without a specific legal structure
  - Benefits of a Foundation: Easier to enter agreements, transfer funds, and secure funding from local institutions, and necessary for Dutch government recognition as a charitable organization
  - Sunday: CorrelAid development workshops

Source: CorrelAid docs, n.d. f

Sources I

Center for Data Science and Public Policy, University of Chicago (2018a): Data Maturity Framework, http://www.datasciencepublicpolicy.org/our-work/tools-guides/datamaturity/

Center for Data Science and Public Policy, University of Chicago (2018b): Data Science Project Scoping Guide, http://www.datasciencepublicpolicy.org/our-work/tools-guides/data-science-project-scoping-guide/

CorrelAid docs (n.d. a): When do we do a project? When not? https://docs.correlaid.org/project-manual/project-decision-guide

CorrelAid docs (n.d. b): Team selection, https://docs.correlaid.org/project-manual/project-coordinators/team-selection

CorrelAid docs (n.d. c): Scoping & Call for Applications, https://docs.correlaid.org/project-manual/project-coordinators/scoping

CorrelAid docs (n.d. d): Best practices, https://docs.correlaid.org/project-manual/project-team/best-practices

Sources II

CorrelAid docs (n.d. e): Data privacy & security, https://docs.correlaid.org/project-manual/data-security-and-privacy#declaration-on-data-security

CorrelAid docs (n.d. f): How We work - Hitchhiker's Guide to CorrelAid, https://docs.correlaid.org/wiki/hitchhikers-guide#our-organization

CorrelAid docs (n.d. g): Finding a legal form, https://docs.correlaid.org/correlaidx-manual/finding-a-legal-form

Data Orchard (2022): Data maturity framework for the not-for-profit sector, https://www.dataorchard.org.uk/resources/data-maturity-framework

Data Science for Social Good Foundation (2021): What Makes a Good Data Science for Social Good Project? https://www.dssgfellowship.org/2015/11/04/what-makes-a-good-dssg-project/

DSSG PT (n.d.): How we work, https://www.dssg.pt/en/how-we-work/

Open Educational Resources

ATTRIBUTION 4.0 INTERNATIONAL - Deed

- **You are free to:**
- Share - copy and redistribute the material in any medium or format.
- Adapt - remix, transform, and build upon the material for any purpose, even commercially.
- **Under the following terms:**
- Attribution - You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. If you wish to use this work in a way not covered by the license, please contact:

Harz University of Applied Science
Friedrichstraße 57 – 59
38855 Wernigerode
E-mail: info@hs-harz.de