**Supplementary Material:**

Five Safes Framework

The five dimensions within the framework can embody a range of values. ‘Safety’ should not be seen as a specific state, but rather as a measure. For example, ‘safe data’ does not mean that the data is non-disclosive; rather it is the dimension under which the safety of the data can be assessed. The model also does not specify how the dimensions should be measured. For example, it may be appropriate to assess the safety of the data using subjective scales (e.g. from very high to very low), or alternatively more formal statistical models could be applied. The outcome of which, however, is that the user has a better-informed idea of ‘more safe data’ and ‘less safe data’.

When considering any data access solution, it is important to consider all five dimensions, even if it is just to note that a specific dimension is not relevant. However, within the broader framework each element should be evaluated individually to identify its own risk characteristics and evidence of appropriate practice.

To enable it to be applied to evaluate the value of data governance across a range of CABI projects and programmes, a generic framework with example questions was created. Table S1 below describes the first iteration of the framework developed for this project. This reconfigured framework, which for the first time includes additional sub-domains, was redeveloped in order that the value of data governance and access can be evaluated across the whole data lifecycle, including planning, design, collection and dissemination.

The redeveloped framework also includes examples of the types of questions that could be asked. The precise nature of the questions asked should vary depending on the type of project/programme being assessed, the stage of the data governance and access lifecycle being assessed and the roles and responsibilities of the specific individual interviewed.

Table S1: Specific Case Study Questions

|  |  |  |
| --- | --- | --- |
| Domain | Sub domain | Example questions |
| Safe projects | 1. Project planning 2. Data management plan 3. Approvals process 4. Public Engagement | * What was the approval process? * How did it facilitate or delay the project? * What processes have been set up to make the next iteration more efficient? * Did you have a data management plan at beginning? * Were all stages identified? * How much did this change? * Who was responsible for the approval process? * Was the approval process developed from scratch? * How was the advice of specialists used? * Is explaining your role to the wider public part of your project's objectives? * How do you engage with the public? |
| Safe people | 1. Governance 2. Training 3. Access | * What model and structure of data governance and access is used? * How many people are employed in data governance functions and in what capacity? * What proportion of their time was spent on data governance issues? * What, if any, training is provided for the following groups:   + Data collectors   + IT staff and data processors/stewards   + Users * Is this training developed and delivered in house? * How do you differentiate access privileges by type of users? * What systems and procedures are in place to ensure data users operate in an appropriate manner? |
| Safe data | Data quality  Data need  Compliance | * How is the flow of data processed in the project? * How is data quality managed? * How is disclosure risk in the data managed? * How do you limit the level of detail available depending on the type of user and level of need? * What is the most detailed level of data made available to researchers? (e.g. geography) * What are the mechanisms through which breaches of data governance procedures are enforced? * What are the range of sanctions that can be applied in the case of a breach? |
| Safe settings |  | * How do you make data available from a safe setting?   + From where can data be accessed?   + What IT systems do you use to limit unauthorised access? * What are the challenges to users safely accessing the data from a safe setting? * How do you enforce standards and policies in relation to accessing data in a safe setting? |
| Safe outputs | Disclosure risk  Building capacity | * Are clear standards set and adhered to in relation to disclosure control? * What type of output checking for disclosure control is undertaken before release? * Are researchers trained in checking outputs for disclosure risk? * Is data created specifically for a project retained after the project is completed? * Are users allowed to archive their workspace (including code) once the project has finished? * What additional products and services have been developed as a result of using the data? |
| Five Safes: questions to be asked for each domain   * How much did the processes you commented on contribute to the costs and outcomes of the project? * What have you learned and what you would do differently next time? | | |

**Supplementary Material:**

Table S2: Assumptions made to generate quantitative estimates

|  |
| --- |
| 1. Internal direct value calculated as follows:  Value of activity/attendance = number of events\*number of people\*hours\*hourly wage rate |
|  |
| 2.Training/workshop events estimated to be six hours a day unless specifically asked in questionnaire |
|  |
| 3 When answers to questions should equal 100%, where this did not occur for an underestimation the residual would be allocated to ‘other’ while any overestimation would be taken off equally in absolute terms across the board. |
|  |
| 4. Weighted averages are conditioned on answer to confidence level questions; where no response has been given to confidence level questions they have been assigned a value of 0.5% (neither confident nor unconfident). |
|  |
| 5. Estimates of time were converted as follows: 1 month = 4 weeks = 20 days = 160 hours |
|  |
| 6. Number of publications was derived from using the highest figure for each category where information on the type of publication is given. However, where no information on the types of publications are provided the average figure is used |
|  |
| 7. Training in relation to activity 3.1 (Human Capacity and Development) is fully allocated to government officials as the training only relates to upskilling of staff of national IT centralised data repository. |
|  |
| 9. Value of time saved is calculated as follows:  Time = amount of time multiplied by average wage rate |
|  |
| 10. Average salary rates for the six sectors were provided by CABI’s Ethiopian staff. They were verified by checking against Salary Explorer.  <http://www.salaryexplorer.com/salary-survey.php?loc=69&loctype=1#:~:text=The%20average%20hourly%20wage%20(pay,ETB%20for%20every%20worked%20hour>.  Two amendments were made to the salary estimates provided by CABI. For scientists/academics/researchers CABI provide a range of estimate for annual income (1000-5000). In order not to overestimate the impact, the lowest estimate was used, as the highest estimate was attributed to international workers and was significantly greater than comparable estimates on Salary Explorer.  For international agency workers the estimate was 200-1000 per annum. As the spread was narrower and due to the fact that potentially a considerable proportion of the international agency staff would not be permanently be based in Ethiopia, the middle estimate (600) was used. |
|  |
| 11. The exchange rate as of 14th January 2021 and as listed on xe.com was used for conversion from Ethiopian Birr. The rates were as follows:  £1 = 44 ETB Br  €1 = 40 ETB Br  $1 = 35 ETB Br |