MODULE 6:

Data consolidation and categorisation

Standardised data consolidation procedures, well-defined categorisation rules, and strict documentation requirements are the basis for a transparent, reproducible and credible land-use finance mapping.

OBJECTIVE

Consolidate and classify quantitative data to construct a database against the land-use finance definition

KEY QUESTIONS

6.1 How to construct a consolidated database for analysis?

6.2 How to check quality of consolidated data?

6.3 How to categorise data to support objectives of mapping?

6.4 How to develop and meet strict documentation requirements?

TEMPLATE

Download **Template 10 - Database** (See Excel “Database Template”)

6.1 How to construct a consolidated database for analysis?

By now, you should have different sets of quantitative and qualitative data from different sources. These datasets need to be checked for their quality and, based on this assessment, included or excluded from further analysis. Datasets to be included must be cleaned and/or formatted so they can be consolidated in one spreadsheet or uploaded to a database that is set up in accordance with the mapping framework established in Module 4.

6.1.1 How to select the datasets and clean them

When assessing the quality of a dataset, one should:

- Identify the key fields (or columns) whose contents need to be assessed for their quality, because they will be used in the land-use finance mapping. The quality of fields outside of that list will not need to be assessed, and will not have an influence on whether a dataset will be selected for final use or not.
  - For a land-use finance mapping, the list of key fields could include location, year, source of finance, amount, currency, use of finance (activities, sectors) and qualitative markers (climate relevance).
  - Depending on the scope of the land-use finance mapping, key fields might also be intermediary of finance, financial instrument and recipient of finance.
  - Dataset fields determined to be unimportant should be excluded from the quality assessment. Such fields might include project lifetime, city, sub-regional classification, grant/loan number and so forth, However, information can be retained in case it becomes useful for the analysis later on.
- Check the data quality of the key fields in each dataset, and clean/correct if necessary. Quality indicators are: timeliness, completeness, consistency, accuracy, validity/integrity, and uniqueness (DAMA UK, 2013). See Table 9 for definitions, examples. and potential solutions to data quality issues.
- Compare different datasets reflecting the same key fields, and based on the result of the quality assessment, decide which (part of each) dataset to use for the later land-use finance mapping.
- Document for each (part of a) dataset for the land-use finance: quality assessment of key fields, and why the dataset can or cannot be used for the mapping exercise.

6.1.2 How to know which database format to use

A database will be customised to the scope of a land-use finance mapping. It depends on the scope, the schedule, available resources and planned frequency of a project whether a database or a spreadsheet is more feasible.

- **A database** should be designed and used if the focus is on managing large amounts of data efficiently, consistently and permanently. In a relational database, data would be split and stored in many different database tables, each of them reflecting a unique set of information. Avoiding multiple storage of the same kind of information helps minimise storage space and manipulate the data quickly. However, database queries are needed to ‘reconnect’ the information stored in these different tables, to perform mathematical operations and so forth. For a database approach, a team would need somebody who is familiar with setting up databases and queries, and who would be available throughout the project. In addition, team members would need to be trained to use the database. A database approach is recommended if a land-use finance mapping is planned to be undertaken frequently.

Quality indicators	Definition	Example	How to solve
Timeliness	Does the data represent reality for the required point in time?	There are records in the datasets describing a period before the chosen project period/year	See Module 5 on how to fill data gaps
Completeness	Are all datasets and data items recorded?	Not all data points covered (columns complete, while rows are incomplete)	See Module 5 on how to fill data gaps
		Not all relevant fields covered (columns incomplete, while rows are complete)	
Consistency (in structure)	Has the same definitions/ methodologies/ categories been used across all datasets over time?	Different currencies used throughout the dataset	Use consistent structures (formats, values, naming conventions), such as date formats, currency exchange rates, units (thousand vs million)
Consistency (in content)		Inconsistency between years, as ministries often change, merge, close or expand, with implications for budget coding/structure	Engage with data manager/ provider to see whether there is a translation table or a list that helps to convert budget codes/structure from one year to the other
		Datasets are based on different land-use finance definitions or use different categorisation	Engage with data manager/ provider to develop a conversion strategy
Accuracy	Does the data reflect the correct value?	Updates/corrections to disbursement data after reconciliation and audits	If schedule allows, update dataset with reconciled/ audited data, otherwise document carefully
Double counting/ uniqueness	Is there a single view of the data?	There may be ten ministries key to a land-use finance mapping, but the dataset captures 11, including 'Ministry of Environment' and 'Environment Ministry'	Use standards: unified spelling/abbreviations/ capitalisation in names
Validity	Does the data match the rules?	Invalid record content: Each govt. budget line should have a unique four-digits code/ identifier that describes: ministry (two digits) and programme (two digits). There are records in the dataset, whose identifier shows 'XXX' instead of a four-digit code/ identifier, hence the record cannot be related to the source (ministry) and the use (programme) of finance.	Engage with data manager/ provider to see what the reason for invalid records might be; correct if possible
		Invalid date format used: instead of using an English default date format MM/DD/ YY, a record or table shows the German format 'Mittwoch, 14. März 2018' and thus does not appear in a filtering result	Check whether all records use the same formats
Integrity	Is the information unchanged from its source, for example accidentally through programming errors?	Amount of finance recorded for a ministry's programme is not accurate because digits were transposed during recording	Compare recorded dataset with original data, for example by comparing the total number of records, using checksums, carrying out spot checks on the data and cross-checking sums in processed dataset with those in the original dataset, and so forth

Table 9: Ensure the quality of a record/dataset/database

- **A spreadsheet** might be sufficient if the focus of the land-use finance mapping is on analysing data (once) and datasets are reasonably sized. Spreadsheets are easy to create and to share, but difficult to handle. In most cases, data can be added/manipulated any time and by all team members without any built-in data quality checks. As a result, spreadsheets bare an increased risk of double-counting, manual errors and inconsistency. It is advisable to introduce mechanisms to avoid this. For example, applying the four-eyes principle or appointing a spreadsheet manager who is responsible for maintaining data quality, integrating new data and so forth. If they are too large, spreadsheets might also require long processing times, or even lead to the crash of the spreadsheet application. However, there are built-in functions for most mathematical and logical operations, statistical comparisons and visualisations in most spreadsheet applications. Most users will be familiar with the usage without further training.

Next, we will focus on spreadsheet design and use.

First, the structure of the spreadsheet must be determined (see Template 10 as an example):

- Key fields represent the minimum set of columns in the spreadsheet. For the quality assessment, the user already identified key fields (or columns) reflecting the information to be used in the land-use finance mapping. These key fields will become columns in the spreadsheet.
- Additional columns might be used to store data from the original datasets for documentation purposes. There might be no current plans to use this information for the final land-use finance mapping. This information could include, for example grant/loan number, programme description, region, whether it is blended finance and so forth. However, it might turn out at a later stage that this information from the original dataset as a table or chart could help to underpin the findings of the land-use finance mapping. It could therefore be valuable to upload this information to the consolidated dataset.
- Additional columns will be needed for:
  - Calculations, for example currency conversion
  - Flags or markers, for example green finance or grey finance
  - Notes, for example reference to data source, documentation about how records have been manipulated

As data is currently organised in different formats, including tables with various structures, all datasets have to be converted into a list format reflecting the structure of the spreadsheet.

Once all datasets reflect the standard structure, and blank columns and rows have been removed, they can be consolidated into one spreadsheet.

6.2 How to check the quality of consolidated data?

After the first quality assessment, the records in each of the selected original datasets now appear to be complete, consistent, accurate, unique, valid and whole. However, after the consolidation into one spreadsheet, there might be inconsistencies between datasets or errors resulting from the consolidation process. For example, a column format may have been changed erroneously. Hence, the consolidated spreadsheet has to be checked again for the described quality indicators (see Table 9: Ensure the quality of a record/dataset/database.) to avoid double-counting or underestimating funds.

Common challenges after consolidating different datasets into one spreadsheet include the variation in date, time or number formats, currencies, naming conventions and so forth. The user needs to check the consolidated records under each column, and in most cases manually adjust so that all entries are consistently named to allow analysis in pivot tables and so forth.

6.3 How to categorise data against the national land-use finance definition

Once the user has a clean consolidated dataset, records need to be filtered for their relevance, and categorised applying the land-use finance definition and typology developed earlier in this project (see Module 3) according to the following steps:

1. Classification of budget lines, activities and programmes into climate-aligned/ misaligned/conditionally-aligned
2. Applying a weighting strategy
3. Categorisation into dimensions in line with mapping framework (defined in Module 4)

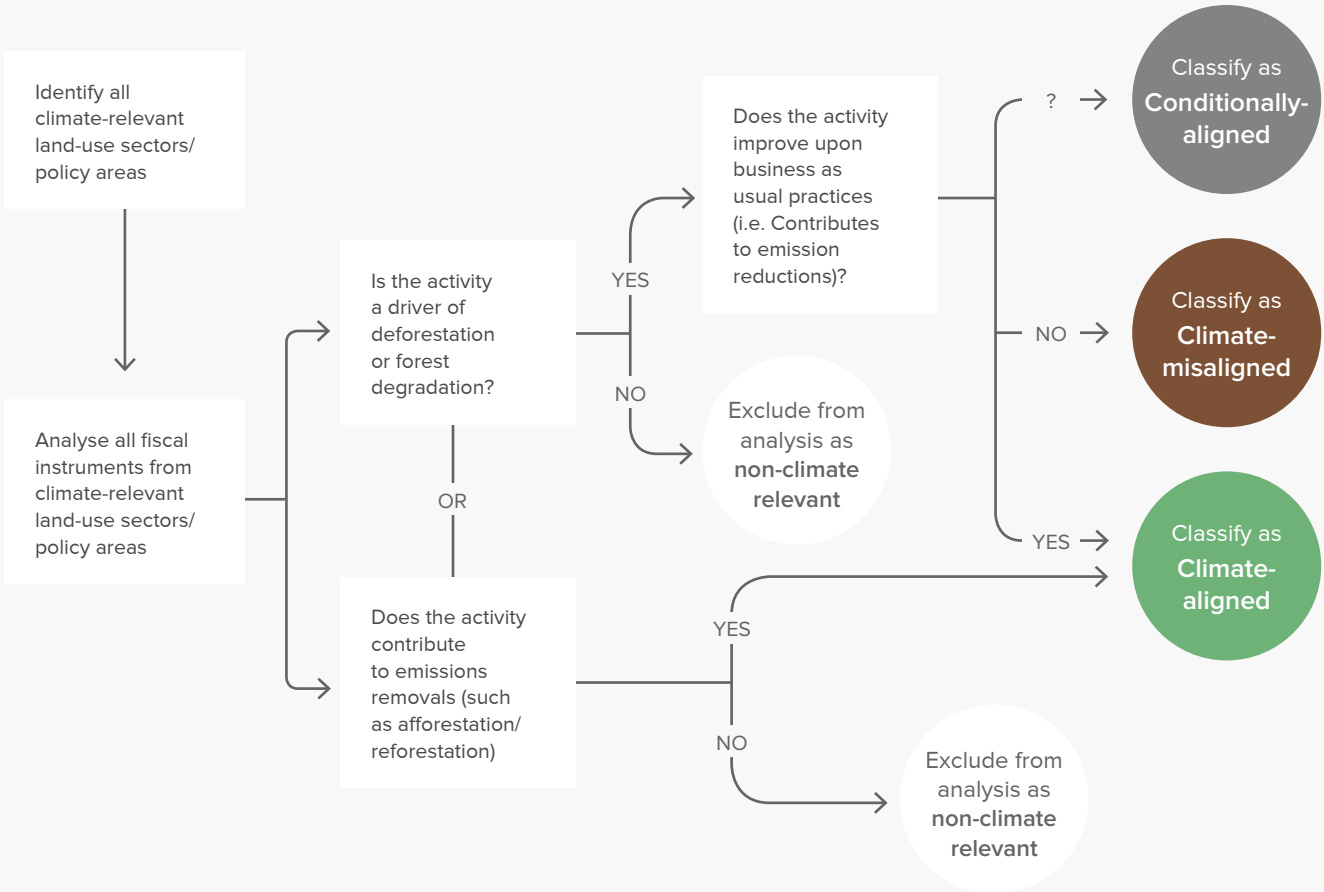
6.3.1 Classifying budget lines, activities and programmes

The following guiding questions can be used to classify individual financial flows according to the definition developed in Module 3:

- Is the activity a known driver of deforestation or forest degradation, or does the activity contribute to atmospheric CO_2 removals?
- Does the activity improve upon business-as-usual practice, for example by contributing to emission reductions?
- Is the activity aimed at improving the enabling environment, for example towards implementation of the National REDD+ Strategy?

The process of classifying financial flows may identify additional sectors or activities that are considered climate-relevant, that were not considered during the initial definition setting stage. These additional activities can be incorporated into the typology to improve the overall definition of climate-relevant land-use finance.

Figure 11: Process flow to classify climate-relevant land-use flows as climate-aligned, conditionally-aligned, or climate-misaligned



However, a classification of financial flows is only possible if the data is detailed enough, and if qualitative information about the nature of programmes and activities can be accessed. Access to project data information, reports, as well as bilateral interviews with implementers are usually needed to be able to categorise activities according to their expected impact (see Module 5).

In Vietnam, for example, there is very limited qualitative information on projects available. Information focuses on the general objectives of projects and programmes. In various cases, these are purposely kept broad to accommodate changes depending on local priorities. These circumstances and the forward-looking nature of the study made it challenging to classify budget lines according to the typology (EFI and CIEM, 2018).

6.3.2 Applying a weighting

During the classification process, weighting may also be applied to individual financial flows. It can be done for several reasons:

- Financial flows often have multiple sub-components, not all of which may be related to climate change. In these cases, and where data is available, analysis can distinguish between sub-activities and only include a corresponding amount of the overall financial flow in the final analysis.
- More detailed weighting strategies can also include apportioning expenditures based on the degree of relevance and impact of climate change mitigation and adaptation objectives.

The determination of if, and how, to apply such weightings is a matter of national priority and outcomes of consultations.

There is no one-size-fits-all approach to this analysis, nor a ‘right’ or ‘wrong’ way of classifying flows. This is a national or jurisdiction-level decision, based on needs and availability of information.

Data classification is usually resource-intensive and the responsibility of the project team. Considerable effort is required to ensure consistency of approach in how definitions are applied and in classification decisions. However, it can be of great added value to involve relevant sectoral experts in the process if possible, both to add accuracy to the analysis, as well as to build awareness of partners on potential misalignment of spending with climate objectives.

Once financial flows are categorised according to the definition developed during this module, and further analysis is conducted based on these results (see Module 6), a final consultation can be conducted to validate the results of the classification. This can include a discussion on how financial flows were classified, as well as the resulting qualitative and quantitative analysis that categorises flows according to the definitions. If multiple categories of finance have been classified, for example climate-aligned and climate-misaligned, it is often helpful at this stage to validate if the results of the classification match expectations at the national level.

In Côte d’Ivoire, for example, weighting based on individual subcomponents was applied, but was not in Papua New Guinea. In Vietnam, the analysis took a simple approach to apportioning individual budget lines. Many investments have multiple sub-components, not all of which related to the National REDD+ Action Plan. In these cases, and where data is available, the study distinguished between sub-activities and only included a corresponding amount of the overall financial flow in the final analysis (Source: EFI and CIEM, 2018).

6.3.3 How to verify categorised numbers?

In most cases, consultants and/or technical staff will be responsible to help with the management and processing of data. However, their ability to categorise projects and activities might be limited. Hence, data gathering and categorising is best conducted in close collaboration with stakeholders, for example government officials at key ministries or land-use experts from research institutes/academia. This will increase the likelihood that all relevant data is gathered, the appropriate typology is assigned to records of the dataset/expenditure items, and data analysis or interpretation issues are easily solved.

As data categorisation creates the base for a land-use finance mapping, it is advisable to verify the results of this step with advisers/principle staff of those organisations financing/implementing the land-use activities categorised. A verification at this point can help to gather further input, keep stakeholders at a higher hierarchy level informed and engaged, and ease a later acceptance and value of final results. Potential formats could be (see also Module 2 on Stakeholder Engagement):

- Presenting categorisation results to the organisation/department in question, explain implications and answer questions.
- Sending an extract of the relevant actor data that was filled in the database.

6.4 How to document data robustly

When consolidating different datasets, it will be important to document (either in a separate spreadsheet, in an extra column/row of the same spreadsheet, or similar):

- Where data in the spreadsheet comes from (document for each column in the spreadsheet, which column(s) in which original dataset it corresponds to).
- Calculation rules and assumptions used to manipulate or process original data, for example flag set to ‘non-climate,’ if the value of another cell equals a certain threshold/number/code.
- Conversion tables, for example currency exchange rate table used for currency conversion.

Clearly documenting data sources and processing procedures ensures transparency and replicability in future years.

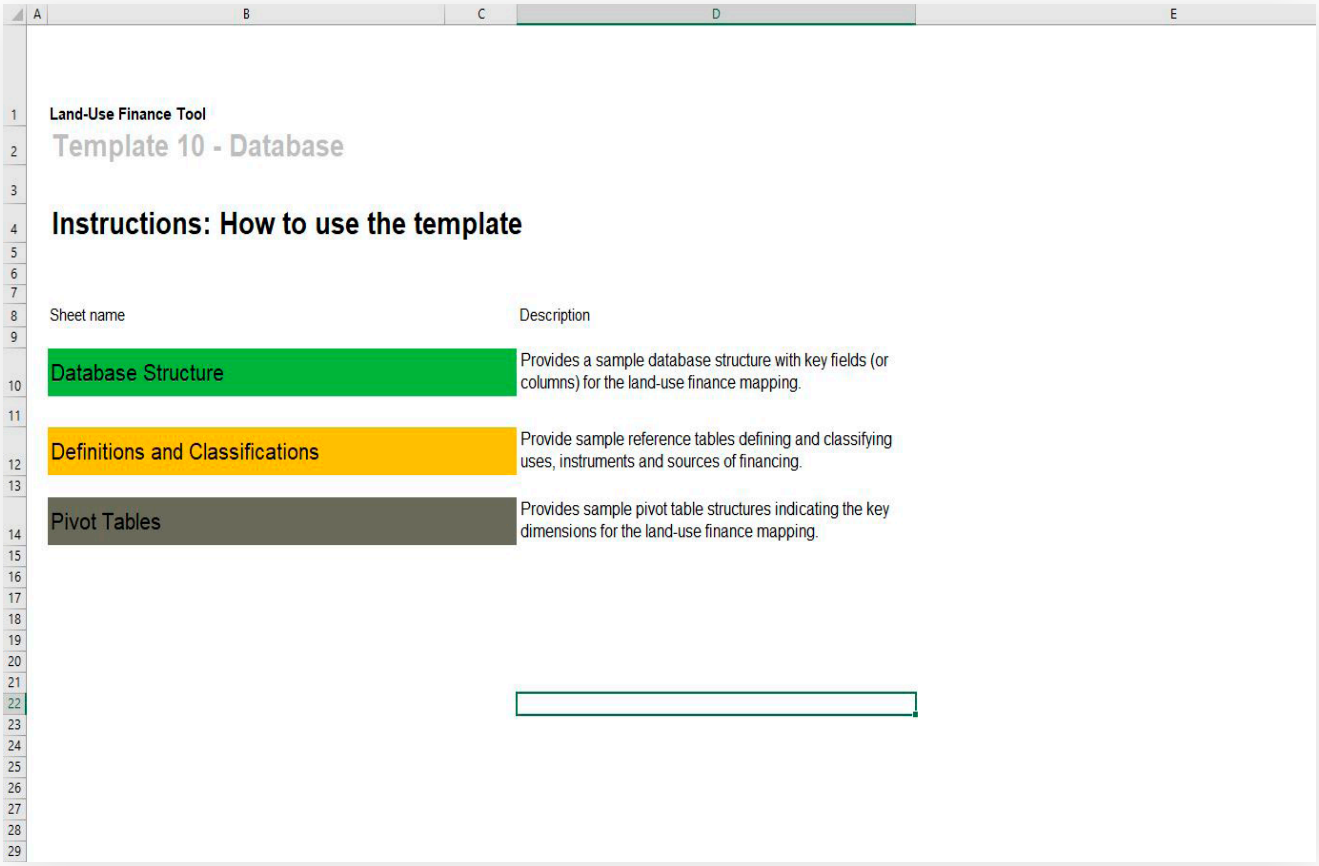
When processing many sources of data and/or using different versions of the same datasets, it can be helpful to catalogue each dataset according to their characteristics (see Table 10). This could be done in a specific documentation file for all data sources or as an attachment to each data source. For example, a workbook consisting of spreadsheet A with the data, and spreadsheet B with data source characteristics.

Table 10: Dataset characteristics

Characteristics	Example
Definition of the dataset, for example time series, sectors, and sub-sector detail, coverage	Audited govt. budget covering: • Ministry of Forestry • Ministry of Agriculture • Year 2016
Definition of the format (spreadsheet) and structure (what different tables are needed and their structure) of the dataset	Spreadsheet (soft copy) providing information on: • Ministry • Level of activity (programme, activity, activity component, etc.) • Activity name • Source of funds • Recipient/executing agency • Expenditure for staff • Expenditure for goods • Expenditure for assets
Description of any assumptions made regarding coverage, the sectors included, representative year, technology/activity level	Coverage: • Expenditure data for Ministries complete. See climate finance definition for selection of Ministries • Data on 'source of funds' can give some indication on intl. development partners' expenditure • Data on 'recipient' can give some indication on local govt. and national fund activities • Year: Latest available year. Budget auditing takes 18 months on average, so only numbers on the fiscal year 2016 are currently audited (June 2018) • Quality assessment: Some records for 'level of activity' are corrupt/invalid. Need to double-check with ministry in response
Identification of the routines and timescales for data collection activities (for example, how often is the dataset updated and what elements are updated)	Elements are updated yearly after the budget audit, in line with the planning and budgeting cycle. Audited numbers will be available by July of the following year at the latest Latest dataset (13 June 2018) compared to former result of database query (May 18, 2018) – no difference, but extra columns (source of funds). Hence, the routines already developed for processing can be applied to this dataset too Note: Budget codes were restructured in 2016, so processing routines and results cannot be compared between 2016 and the preceding years
Contact name and organisation	Mr. XYZ/Director of Budgetary Revenues and Expenditures/Ministry of Finance, building/floor/room/tel./Email
Date of availability	Database query from 13 June 2018

Template 10 - Database

See Excel template accompanying this document.



Land-Use Finance Tool

Template 10 - Database

Instructions: How to use the template

Sheet name	Description
Database Structure	Provides a sample database structure with key fields (or columns) for the land-use finance mapping.
Definitions and Classifications	Provide sample reference tables defining and classifying uses, instruments and sources of financing.
Pivot Tables	Provides sample pivot table structures indicating the key dimensions for the land-use finance mapping.