efsa European Food Safety Authority	**Standard Operating Procedure** **Data Collection and Validation**
Public	

Scope	Process for receiving scientific data collections through the Data Collection Framework and for including data into the EFSA Scientific Data Warehouse.

Special Requirements	This procedure is a controlled document maintained by Quality Management. It may not be deleted without comparable controls.

Responsibilities	• DATA Unit: Evidence Management Unit • Leading Unit: EFSA units requesting the DATA Unit to perform a data collection for a specific project/process

Abbreviations and definitions

BI	Business Intelligence
BR	Business Rule used for automatic data validation
Centralised DC	Data collection process for which data warehousing is implemented through all the following standardised steps: data transmission via the DCF (Data Collection Framework), ETL (Extract, Transform, Load) process, RT (Reference Terminology) checking, BR (Business Rule) validation, BI (Business Intelligence) analyses, Security configuration. Datasets managed as part of a centralised data collection are handled according to GDE2 (Guidance on Data Exchange, version 2) IT protocol.
DC	Data Collection
Data Analyst	Person responsible for: 1. Performing data collection, collation, validation and analysis activities, including verifying data quality, for scientific assessment; 2. Implementing tools for data collection (or integration) and analysis to support the work of units, Panels, Working Groups and the Scientific Committee; 3. Assisting data providers for the delivery of data to the organisation; 4. Supporting the development of a standardised framework for data management and analysis (process and procedures); 5. Supporting the development and maintenance of suitable technical environments for hosting the data management and analysis framework; 6. Defining, implementing and maintaining data management processes; 7. Managing contracts and procurements, including monitoring procedures.
Data Officer	Person responsible for: 1. Performing/coordinating data collection, collation, validation and analysis activities, including verifying data quality, for EFSA scientific assessment; 2. Defining methods and implementing or coordinating the implementation of tools for data collection (or integration) and data analysis for scientific assessment; 3. Coordinating or providing supporting activities for data providers and related networks; 4. Supporting the development of a standardised framework for data collection,

	management and analysis (process and procedures); 5. Supporting the development and maintenance of suitable technical environments, liaising with IT service providers, for hosting the data collection, management and analysis framework; 6. Defining, implementing and maintaining/supervising data management processes; 7. Defining and managing/supervising contracts and procurements, including monitoring procedures; 8. Supporting scientific data governance and the vision for data management in the organisation; 9. Supporting open data initiatives in the organisation.
Data Steward	Data Analyst/Officer responsible for coordinating the activities of a specific centralised DC and for ensuring that the collected data meet minimum quality criteria agreed with the LU.
DATA Unit	EFSA's Evidence Management Unit
Data-BA	Data Business Analyst: Data Analyst/Officer responsible for the analysis of data requirements and data modelling (i.e. data model definition)
DCF	Data Collection Framework
DCF Specialist	Data Analyst/Officer responsible for the configuration of the DCF
DMS	Document Management System
DP	Data Provider (e.g. Member State competent authority)
DWH Reporting Tool	Data reporting, analysis and visualisation tool accessing the data in the S-DWH
DWH Specialist	Data Analyst/Officer responsible for the creation and maintenance of reports for BI analysis generated from the S-DWH using the S-DWH Reporting Tool.
ETL	Extract, Transform, Load data warehousing process
ETL Specialist	Data Analyst/Officer responsible for the creation and maintenance of the ETL process from the DCF to the S-DWH
LU	Leading Unit: the EFSA Unit leading a specific project requiring a DC process.
RT	Reference Terminology (used by EFSA to harmonise data collections)
RT Specialist	Data Analyst/Officer responsible for the creation and maintenance of RTs
SAS	Statistical Analysis System
SAS-DI	SAS Data Integration
Self-Managed DC	DC process self-managed by the LU and not fully implementing GDE2 IT protocol. Self-managed DCs are generally managed using some but not all of the standardised steps followed for a centralised DC (i.e. mainly DCF configuration). A self-managed DC process might be followed, instead of a centralised one, due to project constraints (e.g., strict deadline for implementation, a piloting phase or DC not repeated over time).
SDCC	Scientific Data Collection Committee: EFSA body supporting the governance of scientific data collections. The composition of the SDCC is confirmed every year. The SDDC involves representatives of the different EFSA units requesting the DATA unit to perform data collection activities (i.e. all the LUs involved on specific projects/processes requiring the execution of a DC).
S-DWH	Scientific Data Warehouse

Procedure

	Previous SOPs in the process: SOP_001_S Receiving a request
Step 1 **LU, DATA unit**	**1. Envisioning phase of data collection: chartering the data collection activity within project/process**
	1.1 A data collection activity originates from specific data needs of an EFSA unit that is formalised in a draft project/process charter. The EFSA LU drafting the charter submits a request for a data collection to the DATA unit. 1.2 A Data-BA undertakes a high level analysis of the data collection request and, together with the LU, clarifies the data collection, management and the data analysis requirements. Depending on the project context, the data collection and management can be classified and operated in two ways: • Centralised DC (steps 2, 3, 4, 5 and 6 apply) • Self-managed DC (steps 4 and 7 apply) 1.3 The Data-BA agrees with the LU the support required and finalises, in the project charter, the tasks and estimates related to DATA unit contribution. 1.4 The Data-BA maintains a log/registry of the data collection and management requests in the DMS 1.5 For each request, a high level analysis document is filled in, stored in DMS and linked to the relevant record in the registry of requests. A standard template for the high level analysis document is available in the DMS
Step 2 **DATA Unit**	**2. Design of a centralised data collection**
	Step 2 does not apply to Self-managed DCs. 2.1 The Data-BA, in cooperation with the Data Steward, RT, ETL and DWH specialists writes a document of refined requirements for the data collection (extending the high level analysis performed in step 1) and agrees it with the LU. The Data Steward identifies the organisations and related users involved in the DC, to be registered in the DCF as DPs. The document containing the DC refined requirements shall be stored in DMS and linked to the record in the registry of requests. 2.2 The Data-BA, with support of the relevant specialists must also perform an impact assessment on resources shared with other DCs (i.e. data models, data architecture, RTs, BRs, ETL, BI analysis). When issues are identified, the relevant data stakeholders, members of the SDCC, are informed. 2.3 The Data Steward prepares test data and a test plan to be

	executed by the RT, ETL and DWH specialists in step 3.
Step 3 **DATA Unit**	**3. Implementation of a Centralised Data Collection**
	Step 3 does not apply to Self-managed DCs. 3.1 Based on the requirements defined at step 2, the DCF, ETL and DWH Specialists proceed with the implementation of the DC and document the executed tests performed within the test plan. The DC test plan shall be stored in DMS and linked to the relevant record in the registry of requests. 3.2 The Data Steward and the LU may optionally execute some additional tests and shall confirm that the DC set-up is completed. 3.3 The DC is opened and the Data Steward informs DPs that they can start reporting data.
Step 4 **DPs**	**4. Data provision**
	4.1 Data are uploaded in the DCF by DPs (directly or with support from EFSA). 4.2 Data are processed and automatically validated by the DCF (against RTs and BRs). 4.3 If, after the automatic validation, the dataset is in status "Rejected" or "Rejected editable", the DP must correct the dataset and re-upload it. This step is repeated until the status of the dataset is "Valid" or "Valid with warnings". 4.4 If the status of the dataset is "Valid" or "Valid with warnings" the DP confirms the dataset submission to EFSA by changing the status of the dataset to "Submitted". "Submitted" data cannot be modified by the DP in the DCF.
Step 5 **DATA Unit, DPs, LU**	**5. Data processing through ETL, S-DWH storage, manual validation by data stewards and data validation feedback**
	Step 5 does not apply to Self-managed DCs. 5.1 Once the datasets are flagged as "Submitted", they are automatically processed by an ETL process (e.g. standardisation of measure units, enrichment with residues maximum limits, categorisation of reported data into legislative classes, integration with hierarchies of analysis to correctly aggregate raw data) and loaded into the S-DWH. 5.2 The result of the ETL process can be viewed by the DP

	through a validation report available in the S-DWH reporting tool. The validation report allows DPs to verify the plausibility of the reported data. 5.3 The Data Steward in collaboration with the LU may optionally further evaluate the results of the validation report and perform additional analysis to validate the plausibility of the data. In case the Data Steward identifies the need for corrections to the data, she/he can reject the datasets and inform the DP by email. Correspondence with DPs must be saved in the functional mailbox assigned to the data collection. In this case the DP restarts the validation process from step 4.3. 5.4 In the validation report the DP can also see the list of the related 'Submitted' datasets and, for each of them, select the "confirm' or 'reject' operation. The DP can then trigger the selected operation by clicking on the 'submit' button at the end of the list. Confirmed datasets are consolidated by the ETL process into the S-DWH, while rejected datasets can be corrected and resubmitted by the DP (in this case the process restarts at step 4.3). 5.5 If the DP does not confirm the dataset within a predefined and configurable timeframe (e.g. 20 working days), the dataset is assumed to be confirmed by the DP. The DP is automatically notified by the DCF about the change of dataset status to "Accepted DWH".
Step 6 **Data Unit**	**6. Final deadline and closure of data collection**
	Step 6 does not apply to Self-managed DCs. 6.1 After the data collection final deadline, the Data Steward checks that there are no datasets pending in the DCF in a not-final state (e.g. "Valid", "Valid with warnings", "Submitted"). If needed, the Data Steward sends a notification to the DPs to complete the required actions. 6.2 "Accepted DWH" datasets are retained and archived in the S-DWH. 6.3 In case the DP needs to apply corrections on "Accepted DWH" datasets, these changes can only be applied as amendments according to the GDE2 IT protocol. Amendments are uploaded as new datasets and the process restarts from step 4.2.
Step 7 **LU**	**7. Accept and retain datasets for a Self-managed DC**

	Step 7 does not apply to Centralised DCs. 7.1 The LU may further evaluate the quality of the "Submitted" datasets stored in the DCF and, in case issues are identified, datasets can be rejected. In this case the DP restarts the validation process from step 4.3. 7.2 The LU moves datasets into the "Accepted DCF" status. Datasets are retained and archived in the DCF.
	Following SOPs in the process: SOP_040_S Process for the analysis of data stored in the S-DWH for the assessment of dietary exposure