IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

New strategy of data sharing and data access in statistics: the view from Banco de Portugal[1]

Ana R Gonçalves, Mário Lourenço, Daniel V Sousa, and
Thomas Verheij,
Banco de Portugal

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

New strategy of data sharing and data access in statistics

The view from Banco de Portugal

Ana R. Gonçalves[1], Daniel V. Sousa[2], Mário Lourenço[3], Thomas Verheij[4]

Abstract

Information management assumes a central role in the strategy of every organization. The increasing number of data sources, submitted according to a variety of formats and concepts, treated by different systems and subject to different quality control procedures, determined the need for Banco de Portugal to reflect on how it intends to structure its statistical production systems. Based on pillars such as the use of centralised data repositories, the integration of data acquisition and statistical compilation processes, efficient reference data and data catalogue management, this paper focuses on the harmonisation of statistical concepts as the cornerstone of this strategy, moving towards the goal of having a unified statistical production system based on a broad dictionary of concepts capable of addressing the needs of all statistical outputs.

[1] Banco de Portugal, Statistics Department – anragoncalves@bportugal.pt.

[2] Banco de Portugal, Statistics Department – dvsousa@bportugal.pt.

[3] Banco de Portugal, Statistics Department – mfllourenco@bportugal.pt.

[4] Banco de Portugal, Statistics Department – tjverheij@bportugal.pt.

Contents

1. Introduction

Information is the major asset of any modern organization. Managing information must be, therefore, at the very heart of every organization's business strategy. This is particularly true for National Central Banks, given the increasing need for data to evaluate and provide insights on ever more complex and dynamic economic phenomena.

In Banco de Portugal, the need to deal with increasingly larger datasets, transitioning from "stable" aggregated reports to unstructured microdata, from "controllable" data sources to a stage where virtually any economic agent can be a provider of relevant data, combined with the widespread dissemination of self-service data processing tools, provided the need to evaluate what could be done in order to have more effective information management policies and procedures. In the context of Banco de Portugal Statistics Department, the increasing number of data sources, submitted according to a variety of formats and concepts, treated by different systems and subject to different quality control procedures, determined the need to reflect on how to structure our statistical production systems. Section 2 provides more details on our departing point and why we started to envision the need to change it.

Based on pillars such as the use of centralised data repositories, the integration of data acquisition and statistical compilation processes, as well as efficient reference data and data catalogue management, a vision for the future of our statistical information systems was defined. Section 3 refers to this vision and to the steps we have already taken towards its implementation.

One of the cornerstones of our strategy is the definition of a common set of concepts aiming towards the standardisation of our statistical production systems. For this to happen, these need to progressively adopt the same broad dictionary of common concepts. This dictionary needs to be capable of addressing the needs of all statistical outputs (from more "traditional" ones, such as Monetary and Financial Statistics or Balance of Payments Statistics, to outputs concerning Non-Financial Corporations' Balance Sheet Statistics, for instance). Section 4 provides an in-depth description of the work done towards harmonising statistical concepts and implementing this dictionary of statistical terms.

Section 5 closes and presents the steps we foresee to implement Banco de Portugal's new strategy of data sharing and data access in statistics.

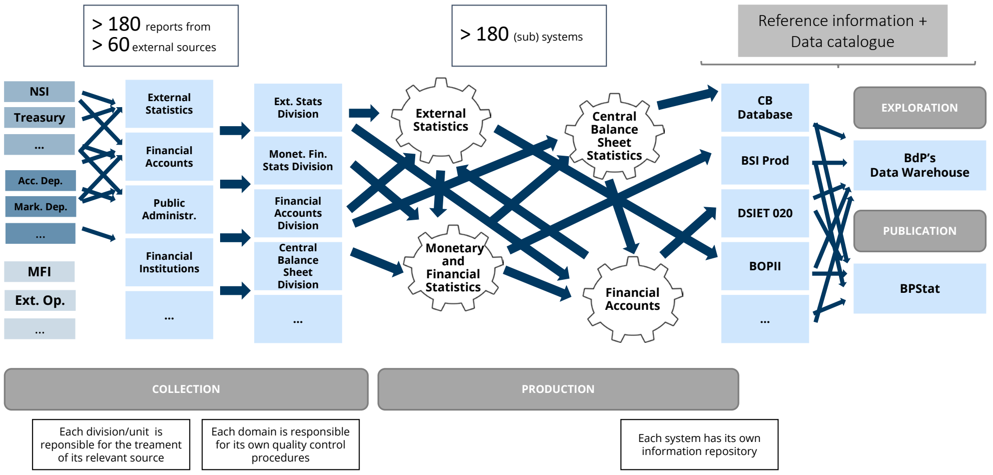
2. Where it all started: untangling the web

The information landscape has shifted dramatically in recent years for all organizations. The need to address additional information requirements or the availability of processes that allow the gathering of unstructured data, produced or sourced by various economic agents, have determined a proliferation of systems specifically designed to deal with each dataset. The global financial crisis also provided the motivation to obtain more detailed data on a wide variety of subjects. The growing use of big data and administrative data sources together with the need to produce statistics and explore datasets from different perspectives and with greater detail, contributed to the growing complexity of statistical production systems.

Banco de Portugal's statistical production systems are no different. In 2020, we listed more than 180 reports of data obtained from over 60 different external sources. These reports were collected, analysed, controlled, and used in production procedures carried out by over 180 statistical systems or subsystems (Figure 1).

Mock-up of Banco de Portugal's statistical production system in 2020

Figure 1



In this scenario, each division or unit assumes the responsibility to deal with the data it considers relevant for its own purposes. Data is, therefore, analysed several times, from different perspectives, according to the goals of each unit.

The organizational setup of Banco de Portugal's Statistics Department also contributed to the proliferation of datasets and data processing procedures. The fact that the same department is responsible for a variety of statistical production systems, with different requirements and a wide number of reporting agents, together with the growing availability of relevant administrative and public data (capable of providing further insights into the economic phenomena object of each divisions fields of expertise) led to the proliferation of datasets and data formats which were handled according to their specific sets of concepts and codes, hence undermining the possibility to implement a common conceptual framework that would capture this complexity.

Quality control procedures were designed according to the needs of each production system and its specific conceptual framework. This ultimately meant that several units conducted similar quality control checks over the same datasets. These checks, however, could not be standardised in a way that enabled them to be useful for different production systems.

Data was then used within each production system – with several connections established directly between them – (i) to provide relevant data for comparability checks (to assure the consistency of different statistical outputs with similar characteristics) and (ii) to provide input for other statistical production systems.

Statistical outputs were stored in each system's specific database. Statistical series published in our statistics portal (BPstat), as well as shared within Banco de Portugal to be used by other departments (within our corporate data warehouse, BPW) were derived from each system's database. Each statistical system was, in the end, responsible for their own dissemination processes. Most of the production processes were, in conclusion, specifically implemented by each system for its own purposes, albeit some information management principles were already in place (for instance, data shared with internal users was recoded according to a reference information system and a data catalogue was increasingly put into place with metadata that helped users have easier access to our outputs).

The interdependences of our systems while conducting IT maintenance or overall revisions of methodology to address new needs, to fulfil new reporting requirements by international organizations or to define new data inputs became more and more challenging through the years. The fact that each system had its own specific aspects, built according to its own needs, based on its own concepts, resulted in increasing difficulties whenever the teams responsible for each system changed. The potential dissemination of know-how was reduced by these constraints, leading to an overall consciousness that day-to-day activities could be more efficient.

This determined the need to think about how this could change and led to the implementation of a set of pilot experiences towards a new setup, in line with the new vision for Banco de Portugal's statistical production systems.

3. The future vision for Banco de Portugal's statistical information systems

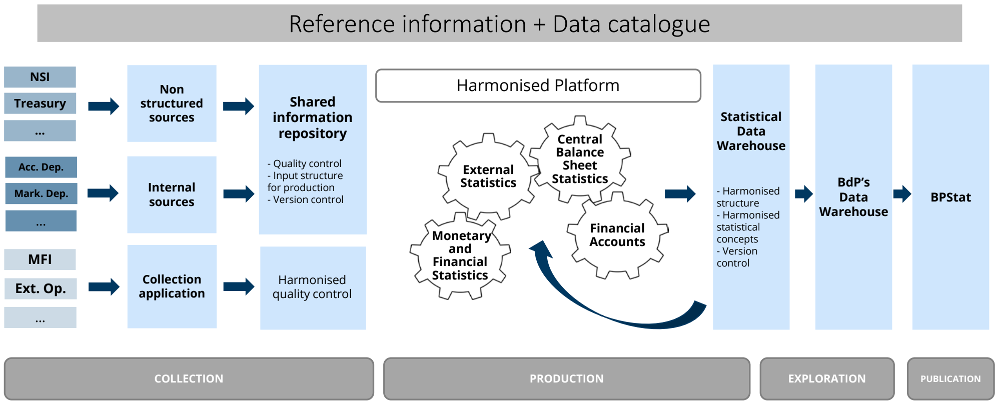
In 2017, Banco de Portugal launched an Integrated Data Management (IDM) Programme as part of its Strategic Plan for the 2017-2020 period, aiming at better using the available data by means of rationalisation of the processes associated with its collection and processing, promoting its effective sharing within the organization (MORENO, 2021).

The importance of Data Management was also consecrated in Banco de Portugal's Strategic Plan for the 2021-2025 period, with the need to consolidate the IDM Programme and the evolution of our internal data governance models and data architecture being explicitly referred as a central goal of the institution, turning Banco de Portugal into an organization focused towards data.

Within the Statistics Department, data management was one of the main concerns when a more streamlined conceptual framework for the future of our systems was designed (Figure 2). This vision highlights the role of our reference information and our data catalogue as cornerstones of our strategy, considering the importance of using concepts, dimensions and terms which are commonly understood by all systems, giving our teams full knowledge of all the data available in-house (where it is stored, how can it be accessed, which relevant terms does it include, who is the expert on a given topic in case of any doubts, etc.).

Future vision for Banco de Portugal's statistical production systems

Figure 2



This vision tries to simplify the statistical production process, basing itself in the traditional stages of any statistical production system, while trying to harmonise procedures and isolating the impacts of any changes that occur over time.

To do so, common data repositories (or warehouses) accessible by different production systems were defined, enabling the elimination of direct links between different production platforms. A primary repository stores data obtained from external sources, as well as from other departments within Banco de Portugal. Each dataset is stored in this repository only once, eliminating duplicated reports. Information can then be used by every statistical production system as they see fit, bearing in mind there is only one unified vision of the data reported and shared.

Statistical production processes will be increasingly defined within a harmonised production platform which will consist of simpler, easier to handle processes, enabling systems to use the same procedure within several production routines. For instance, if there is the need to implement a procedure to estimate a certain statistical phenomenon, the same procedure should be used in more than one system, hence ensuring the same set of tools and methodology is used to perform similar tasks.

This is also true for quality control procedures, given that most quality control checks have generally similar purposes (the identification of outliers or possibly incorrect observations which then need to be confirmed or corrected, for instance).

Another data repository stores data compiled by the different production systems, using standardised data structures defined according to harmonised statistical concepts (a set of principles explained in more detail in the next section). This statistical data warehouse will enable all statistical outputs to be stored in the same infrastructure. Systems that need the output of another system will only need to connect themselves to this data warehouse to gather the most recent information made available regarding any data domain.

Statistical outputs will cover all data made available to other departments within Banco de Portugal through our corporate data warehouse, as well as all statistical series published in our statistics portal, BPstat. These are, usually, subsets of the information produced by our statistical production systems. Storing each unrestricted statistical output in a single warehouse will allow all outputs for dissemination, reporting, etc, to be generated from the datasets stored in this warehouse (basing such reports in the "single truth" each system shares as its output).

This vision will enable us to simplify our systems, harmonise our processes and make them more efficient, giving way to leaner statistical production procedures, consolidated within a single framework, easier to maintain and evolve, ensuring greater homogeneity. This will also allow our processes to be more transparent (eliminating black boxes or situations where team members do not fully know the expected result of a certain action or change in an everyday procedure). Auditability is maintained to keep track of how processes change and how data is handled. It is our goal to make systems more driven towards information sharing, which will help reduce adaption costs for team members.

In 2022, the first Banco de Portugal's Statistical Department's centralised data repository was implemented. It corresponds to a platform where non-structured reported data has been increasingly integrated and shared among internal users.

Working with Banco de Portugal's IT department, the possibility to have more standard, flexible and agile processes which can be applied to different information domains has also been evaluated in the past year, namely through the conduction of pilot experiments or case studies focusing on the development of new statistical production systems with these concerns in mind.

An in-house pilot platform where standardised quality control checks are conducted (applied to different production systems and stages of the statistical production) is also being developed, as well as the repository where statistical production outputs will be stored.

For this to be possible, nonetheless, it became evident that a harmonised set of statistical concepts was needed as a corner stone of our strategy. Therefore, a task force specifically mandated towards this goal was created. The following Section describes the work developed by this task force and the issues addressed when defining Banco de Portugal's Statistics Department's harmonised statistical concepts.

4. Working towards the definition of a common set of concepts

4.1. Preliminary work and guiding principles

In March 2022, a preliminary study[5] regarding the conceptual framework of our systems, developed by members of different business areas within Banco de Portugal's Statistics Department, was presented. This study showed that, to ensure the standardisation and consolidation of our production systems, the harmonisation of "the language" in which datasets are coded is crucial, ultimately leading to:

- Greater efficiency in data processing;
- Synergies in information sharing processes;
- Easier maintenance and implementation of further developments in our information systems;
- Smaller cost of entry for team members when they change their role within the department or when they need to work with different datasets (other than the usual ones).

Between November 2022 and March 2023, a second team[6] developed a set of harmonised concepts to be adopted in Banco de Portugal's Statistics Department. This team's work started by summarising the work developed by the previous group, detailing the harmonisation of concepts and mappings that were to be implemented in a third stage.

Some proposals regarding the guiding principles and methodology for the construction of the codes and for the strategy to achieve harmonisation were considered, namely: i) the need for the codes to have clear designations, (ii) the importance of avoiding the combination of information of several dimensions into a single dimension (e.g. the sector dimension should only have information regarding the sector, excluding concepts that can be obtained by cross-referencing other dimensions), (iii) the importance of sharing the most granular data available as the output of each statistical system, (iv) the preferable use of already structured encodings, prioritizing *Statistical Data and Metadata eXchange* (SDMX) codes, and (v) the need to validate with all business areas the result of the harmonisation procedures. With these principles in mind, to find a method of harmonisation that would facilitate the development of the project, different dictionaries already available within several data catalogues and metadata repositories of Eurosystem working groups were thoroughly analysed, among which *Single Data Dictionary* (SDD), SDMX, *Data Point Modelling* (DPM), *Informal Coordination Group on Integrated Reporting by the Banking sector* (ICG IRB), *Banks' Integrated Reporting Dictionary* (BIRD) and the *Integrated Reporting Framework* (IReF).

[5] This study was conducted by Alexandra Miguel, Ana Colaço, José Alexandre Neves, Lídia Brás and Rita Poiares.

[6] The second team was composed by Diogo Barbosa, Ana R. Gonçalves, Ana Francisco, Elena Bucea and Daniel V. Sousa.

Regarding the methodology, three basic rules were considered: (i) the construction of the codes would, desirably, guarantee the same logic for all common dimensions; (ii) searches and aggregations should be simple, as "automatic" as possible, intuitive and, whenever possible, consider hierarchies; and (iii) the new harmonised codes should begin with letters, following a clear logic in the way they are designed.

4.2. Defining what to harmonise

The definition of a single set of concepts applicable to every statistical output could be done considering either aggregated datasets (usually, statistical time series) or granular datasets (individual data on specific economic agents or operations).

Evaluating the available data from a bottom-up approach (considering granular data as the starting point) would expectedly increase the comparability between datasets, providing greater flexibility regarding data aggregation and dissemination. It would imply, however, that more variables needed to be analysed, introducing more complexity in the process, which necessarily implied a more in-depth analysis of specific aspects of each set of the current concepts used within each statistical production system. Possible difficulties arising from different aggregation procedures would also not be directly addressed at this stage, as they would need to be evaluated individually, when aggregated outputs are analysed.

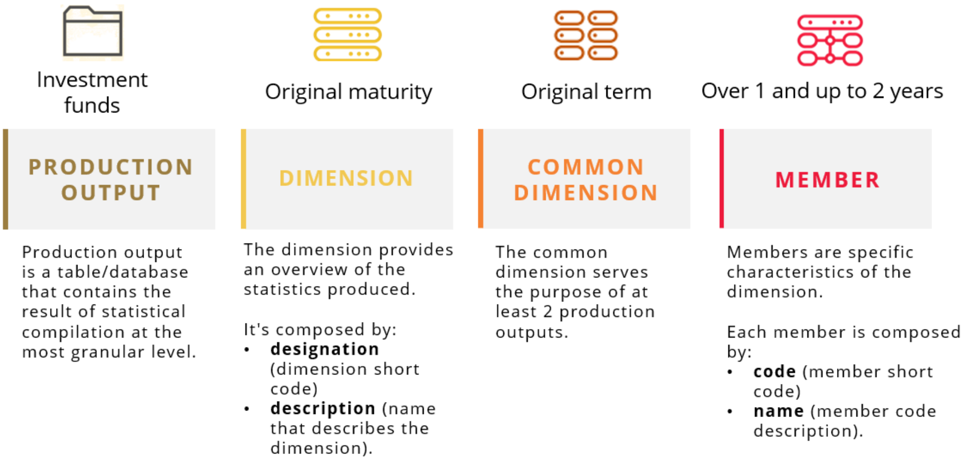
A top-down approach (starting from aggregated datasets, composed by data disseminated through reports to different international organizations or through Banco de Portugal's own publications), on the other hand, would lead to faster results and to an easier implementation, although it would be more difficult to evaluate all relevant variables and statistical concepts basing the analysis solely on aggregate data (which might not include all variables relevant for all statistical outputs). Additionally, hierarchies associated with different statistical outputs can differ, making it difficult to link different datasets. Finally, this approach would also provide additional difficulties when common dimensions were to be recoded according to the new harmonised set of concepts.

This ultimately led to the decision of harmonizing according to a bottom-up approach, starting from the most granular information available at each statistical production system, evaluating concepts autonomously but working closely with data experts to have detailed discussions whenever specific issues needed to be addressed in the definition of harmonised concepts.

The harmonisation procedure was based on a glossary of terms where statistical production outputs were assumed to be the cornerstone of the process, as exemplified in Figure 3.

Example of the application of the glossary of terms

Figure 3



Investment funds

PRODUCTION OUTPUT

Production output is a table/database that contains the result of statistical compilation at the most granular level.

Original maturity

DIMENSION

The dimension provides an overview of the statistics produced.

It's composed by:

- **designation** (dimension short code)
- **description** (name that describes the dimension).

Original term

COMMON DIMENSION

The common dimension serves the purpose of at least 2 production outputs.

Over 1 and up to 2 years

MEMBER

Members are specific characteristics of the dimension.

Each member is composed by:

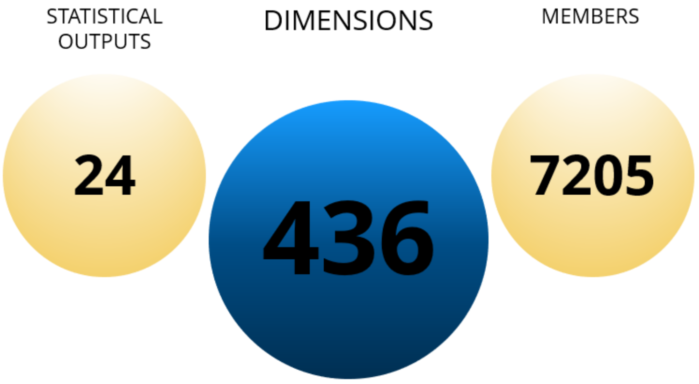
- **code** (member short code)
- **name** (member code description).

Statistical production outputs were defined as a set of tables or databases containing the results of the statistical compilation process (at the most granular level possible). These tables contain dimensions that stand for the perspective (attribute or variable) of such statistics. These dimensions are characterised using a designation (short dimension code) and a description (name describing the dimension) and contain different members, i.e., characteristics shown in the dimension, with each member being composed by a code (short member code) and a name (description of the member code). The harmonisation procedure might imply that current members could be ultimately split into members of different dimensions (designated as "complex members"). Dimensions that are part of at least two production outputs are defined as common dimensions. Dimensions that are part of only one production output are defined as specific dimensions.

Thus, 24 statistical production outputs (produced by 6 different business areas within Banco de Portugal's Statistics Department) were explored, which translated into a total of 436 dimensions to be evaluated (Figure 4).

Collection of statistical outputs, dimensions and members to be harmonised

Figure 4

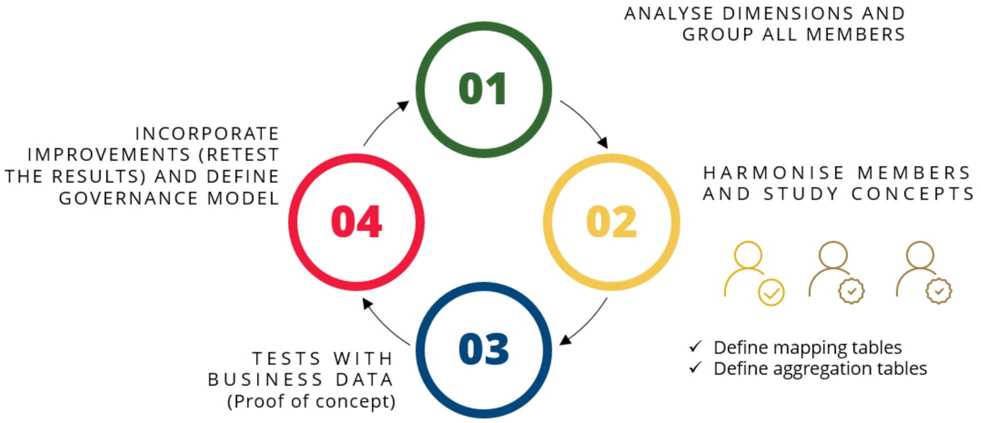


4.3. Defining how to harmonise – the harmonisation cycle

Having defined the level of detail on which the project would be based, with the objective of ensuring an unbiased and efficient process, a harmonisation cycle was defined considering several interactions (Figure 5).

Harmonisation cycle

Figure 5


ANALYSE DIMENSIONS AND GROUP ALL MEMBERS
01
HARMONISE MEMBERS AND STUDY CONCEPTS
02
✓ Define mapping tables
✓ Define aggregation tables
TESTS WITH BUSINESS DATA (Proof of concept)
03
INCORPORATE IMPROVEMENTS (RETEST THE RESULTS) AND DEFINE GOVERNANCE MODEL
04

The first phase of the process consisted of an in-depth analysis of the dimensions and the members of each statistical output to identify opportunities of harmonisation (considering both variables, or dimensions, and observations, or members).

In the second stage, the concepts' nature was thoroughly studied, trying to find similarities which would allow the unequivocal definition of a single set of concepts, broad enough to enclose currently analogous classifications. The outcome of this analysis was always validated by more than one team member, hence ensuring that it would lead to a uniform and intuitive new set of concepts. At this stage, the team also defined mapping tables (linking old and new concepts), as well as aggregation tables that facilitate comparison between different levels of data granularity.

In a third moment, proofs of concept were conducted to check the validity and appropriateness of the proposed solution, testing the harmonisation procedure applied to the evaluated statistical outputs.

This was relevant to ensure the consistency of the outputs produced based on the harmonised set of concepts (keeping the new dictionary of dimensions and terms permanently updated and organised), while also providing further validation that the new set of harmonised concepts was in fact generating the intended results, namely facilitating the comparability of different statistical outputs.

Finally, in the last stage, all improvements that derived from the previous stages were implemented, redefining the harmonised set of concepts, as well as detailing a governance model for its future application and maintenance.

4.4. Results

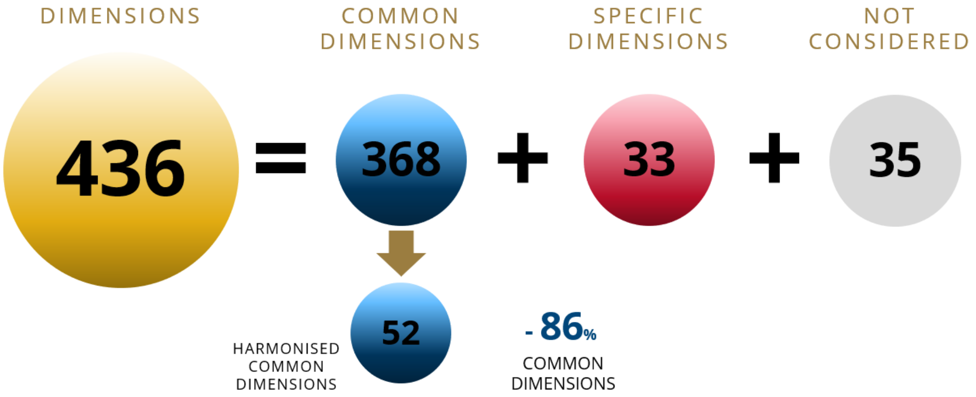
The set of 436 dimensions collected from the 24 statistical production outputs analysed was subject to the harmonisation procedure and recoded into new harmonised dimensions according to the referred guidelines and methodology.

Within the 436 dimensions analysed, 368 were considered common to at least two statistical outputs, 33 were classified as specific dimensions and 35 were deemed redundant (and, hence, excluded from the harmonisation procedure, considering it would be possible to obtain such information from other dimensions).

The set of 368 common dimensions was reduced to 52 dimensions after the harmonisation of its names and descriptions. This meant that, at the start, the 436 dimensions initially listed were reduced to a set of 85 dimensions to be further analysed (52 common dimensions and 33 specific dimensions), a reduction of 79% in the set of dimensions available (86%, if only common dimensions are considered) (Figure 6).

Harmonisation process results regarding the set of dimensions

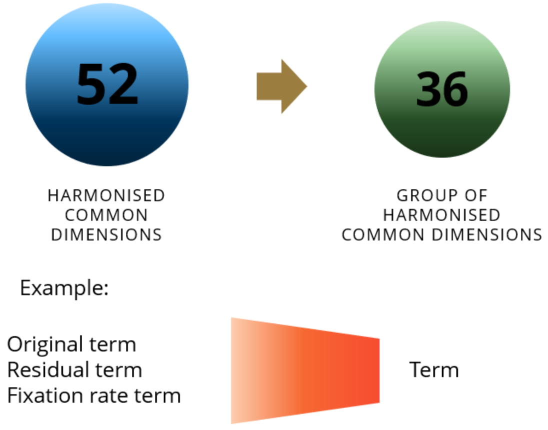
Figure 6



The 52 common dimensions were summarised into 36 groups of dimensions, considering the same set of members were shared between more than one dimension. This was the case, for example, for the dimension "Term", since it included the same members as dimensions "Original term", "Residual term" and "Refixation term".

Group of harmonised dimensions

Figure 7

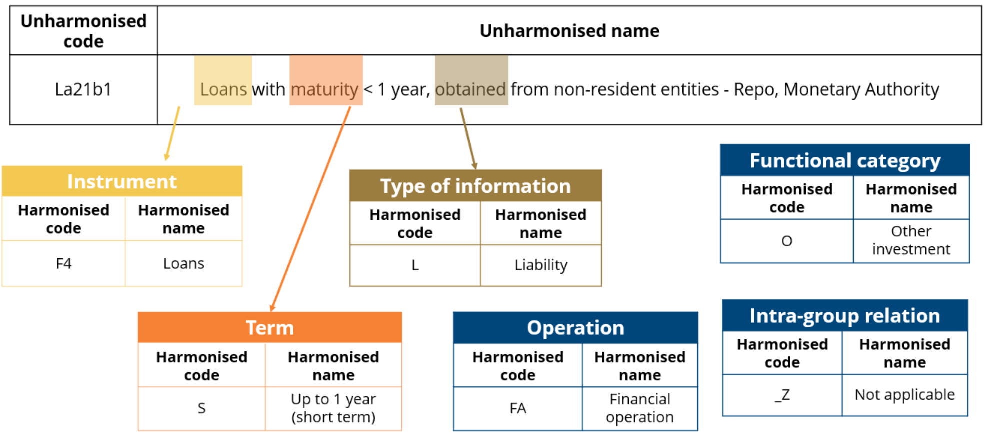

52
HARMONISED COMMON DIMENSIONS
36
GROUP OF HARMONISED COMMON DIMENSIONS
Example:
Original term
Residual term
Fixation rate term
Term

By this stage, the focus shifted into the harmonisation of members, i.e., recoding the list of attributes associated with each dimension. The harmonisation of the members of common dimensions was prioritised. From the 36 groups of common dimensions, 8 were evaluated as “dynamic“, which led to the option to have them harmonised through the definition of good practices regarding its coding and description (as was the case of the “Period” dimension, for instance). The remaining 28 groups of common dimensions were analysed one-by-one, having all members potentially included in the respective lists manually evaluated, combining business codes and descriptions to fully understand the underlying concepts. Of the remaining 33 specific dimensions, 14 were also analysed at this stage, considering that, although they were deemed specific of a particular statistical output or business area, they shared members with other dimensions (as was the case of dimensions regarding specific uses of territory codes, for instance). The remaining dimensions were to be evaluated whenever needed (i.e., when the need to harmonise the specific outputs that use such dimensions arises).

The full list of members of the various outputs analysed led to the identification of 7205 members to be harmonised, 1610 of which were considered complex (which meant they would be segregated into a combination of several dimensions). Figure 8 exemplifies the analysis of one of these cases.

Harmonisation example for complex members

Figure 8



Unharmonised code	Unharmonised name
La21b1	Loans with maturity < 1 year, obtained from non-resident entities - Repo, Monetary Authority

Instrument

Harmonised code	Harmonised name
F4	Loans

Type of information

Harmonised code	Harmonised name
L	Liability

Functional category

Harmonised code	Harmonised name
O	Other investment

Term

Harmonised code	Harmonised name
S	Up to 1 year (short term)

Operation

Harmonised code	Harmonised name
FA	Financial operation

Intra-group relation

Harmonised code	Harmonised name
_Z	Not applicable

This analysis added complexity to the harmonisation process, generating new members as part of the attributes of common dimensions.

After the harmonisation procedure, 4104 distinct harmonised members were determined, reflecting a 43% reduction from the starting number of members considered in this exercise.

4.5. Project outputs

The thorough analysis of the dimensions and members resulted into several deliverables relevant for the organization and documentation of the process as a whole, for its effective implementation.

A dimension matrix that centralises the exhaustive survey of the dimensions collected was built, presenting the correspondence between the old dimensions of the statistical production outputs evaluated and the new harmonised dimensions, as well as a classification of whether the dimensions were considered common or specific.

A set of reference tables was also defined, including i) a dictionary of the harmonised terms (dimensions and members); ii) a glossary of terms which includes further detail on the harmonised members (to harmonise several concepts, a trade-off between the detail to be presented in the designation of the member and its specific nature had to be considered, leading to the need to have more detailed definitions documented in this glossary).

One of the main outputs of this project was the definition of mapping tables, which enable the correspondence between information currently available in the statistical production outputs and the new harmonised codes. This is a key component of this exercise, allowing systems to progressively transition from the current situation (where specific concepts are in place). Aggregation tables were also defined, enabling the comparability of the outputs produced according to different levels of granularity.

Although these deliverables are quite relevant, they would be pointless (as the whole harmonization exercise, probably) if nothing was envisioned to enable the permanent update and future implementation of the harmonised concepts and its principles.

This led to the definition of a governance that would enable the harmonisation process to outlast the initial analysis carried in the last couple of years, making it a permanent part of the statistical production in Banco de Portugal's Statistics Department's. Whenever changes in production systems occur or constraints in the current harmonised sets of dimensions and members are identified, these rules enable the permanent update of the harmonisation effort. Generic principles were agreed upon, among which the need to:

1. Follow, whenever possible, the same logic of harmonisation – at dimension (common and specific) and member levels;

2. Continue defining intuitive codes, possibly with an implicit hierarchy that will allow simple and/or automatic searches and aggregations;

3. Have member codes beginning with letters;

4. Keep having the ability to distinguish different concepts.

Specific rules were also defined for each analysed dimension. In general, priority was given to international code lists, such as SDMX codes and SDD dictionaries, adapted whenever needed to fulfil the principles agreed upon. Among others, rules were made for the way new member descriptions should be defined, as well as on how business areas should act when the need to adjust or add new elements to the set of harmonised concepts arises.

This governance clearly states that the harmonisation work conducted would be permanently subject to revisions, redefinitions, and improvements. A timeline for further developments regarding the effective implementation of these principles was also defined, as well as a set of use cases to be tested in the months following the second stage of the conceptual harmonisation process. Additionally, it was determined that a unit within the Data Management Division would be responsible for ensuring the continuity of the harmonization process and the correct application of its governance.

5. Closing remarks – Where do we go from here?

The need for a harmonised set of statistical concepts to be used by all statistical domains was, from the beginning, considered to be the cornerstone of Banco de Portugal's Statistics Department's new strategy for data sharing and data access. Efficient and homogeneous statistical production systems that can rely on centralised data repositories and quality control procedures are only possible if the information management programme is imposed in such fashion that virtually all systems and its outputs speak the same language and are defined according to the same standards.

A thorough analysis of the current business concepts adopted within our systems was conducted and a team was mandated to define a new set of dictionaries which would be the basis of the integration of our statistical outputs into a single repository, enabling its centralised dissemination.

Harmonisation could either be implemented within every single statistical production system from the start, or it could be progressively implemented whenever systems need to be changed. Given the constraints to a "big bang" implementation, a stepwise approach was considered more appropriate, defining the need, at this stage, to increasingly add statistical harmonised outputs to our final data repository. This is being done with no impact in the current production systems which remain virtually unchanged, except for the fact that they need to be able to produce their statistical outputs coded according to the rules defined by this task force. This is being done using a set of tools that allow systems to translate their current concepts into new ones. This was considered determinant to fully grasp the needs of conceptual harmonisation bearing in mind all datasets. The final stage of our production processes was also considered to be the most suitable stage to initiate the harmonisation procedures, given the need to define the data structures that are being materialised in our integrated statistical outputs data warehouse. Once harmonised codes are consolidated, they will progressively be implemented in the remaining stages of our statistical production processes, as the systems are being remodelled and further developments are implemented.

Nonetheless, this does not mean all work is done. Conceptual harmonisation is an essential step towards implementing the vision we have for our statistical production systems in the future, but it must keep being updated and revised to address new needs and bypass unidentified constraints. It is, however, a vital step to fully achieve a situation where data can be shared among different data specialists and users within Banco de Portugal, facilitating data usage and its comparability.

6. References

MORENO, M. Carmo (2021), Data Governance: an orchestra of people, processes, and technology, in IFC Bulletin No 54 (2021), Issues in Data Governance.

NEW STRATEGY OF DATA SHARING AND DATA ACCESS IN STATISTICS

THE VIEW FROM BANCO DE PORTUGAL

OCTOBER 2023

BANCO DE PORTUGAL
EUROSISTEMA

PRESENTATION OVERVIEW

01 | BANCO DE PORTUGAL'S STATISTICS DEPARTMENT

02 | WHERE IT ALL STARTED

03 | FUTURE VISION OF OUR STATISTICAL INFORMATION SYSTEMS

04 | DEFINING A COMMON SET OF CONCEPTS

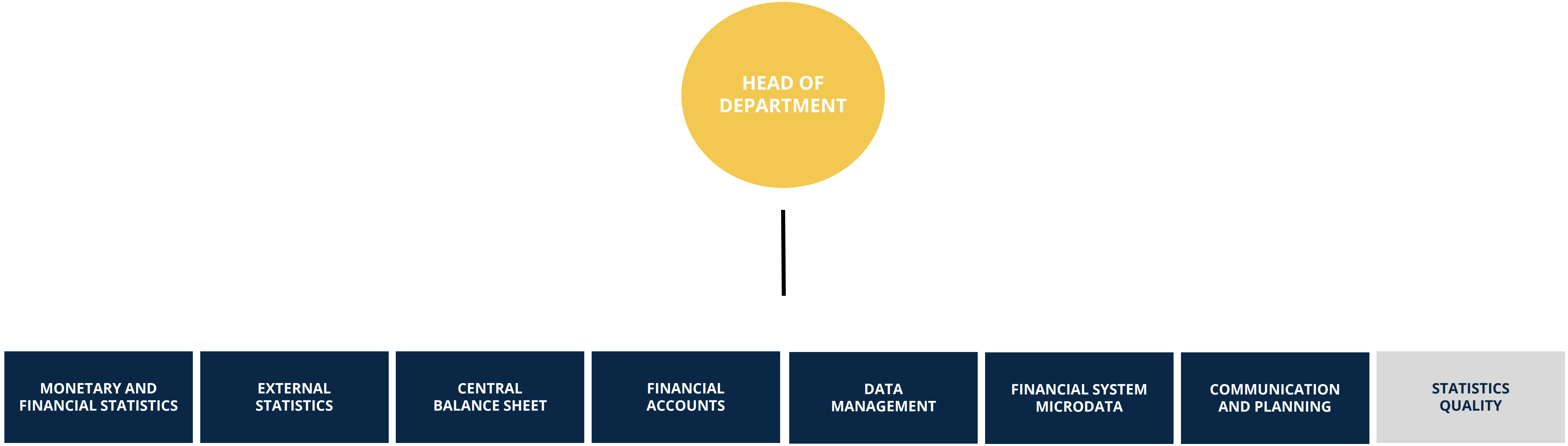
05 | CLOSING REMARKS

BANCO DE PORTUGAL'S STATISTICS DEPARTMENT

01

BANCO DE PORTUGAL'S STATISTICS DEPARTMENT

ORGANIZATIONAL STRUCTURE



HEAD OF DEPARTMENT

MONETARY AND FINANCIAL STATISTICS

EXTERNAL STATISTICS

CENTRAL BALANCE SHEET

FINANCIAL ACCOUNTS

DATA MANAGEMENT

FINANCIAL SYSTEM MICRODATA

COMMUNICATION AND PLANNING

STATISTICS QUALITY

BANCO DE PORTUGAL'S STATISTICS DEPARTMENT

OUR STATISTICS

CENTRAL BALANCE SHEET STUDIES

INDEBTEDNESS OF THE NON-FINANCIAL SECTOR

EXCHANGE RATES

INCOME AND SAVING

BUDGET OUTTURN

INTEREST RATES

GENERAL GOVERNMENT FINANCIAL ACCOUNTS

SECURITIES

ECONOMIC AND FINANCIAL INDICATORS

INTERNATIONAL RESERVES

PENSION FUNDS

MFI BALANCE SHEET

GENERAL GOVERNMENT FINANCING

BALANCE OF PAYMENTS

EXCHANGE RATES INDICES

NATIONAL FINANCIAL ACCOUNTS

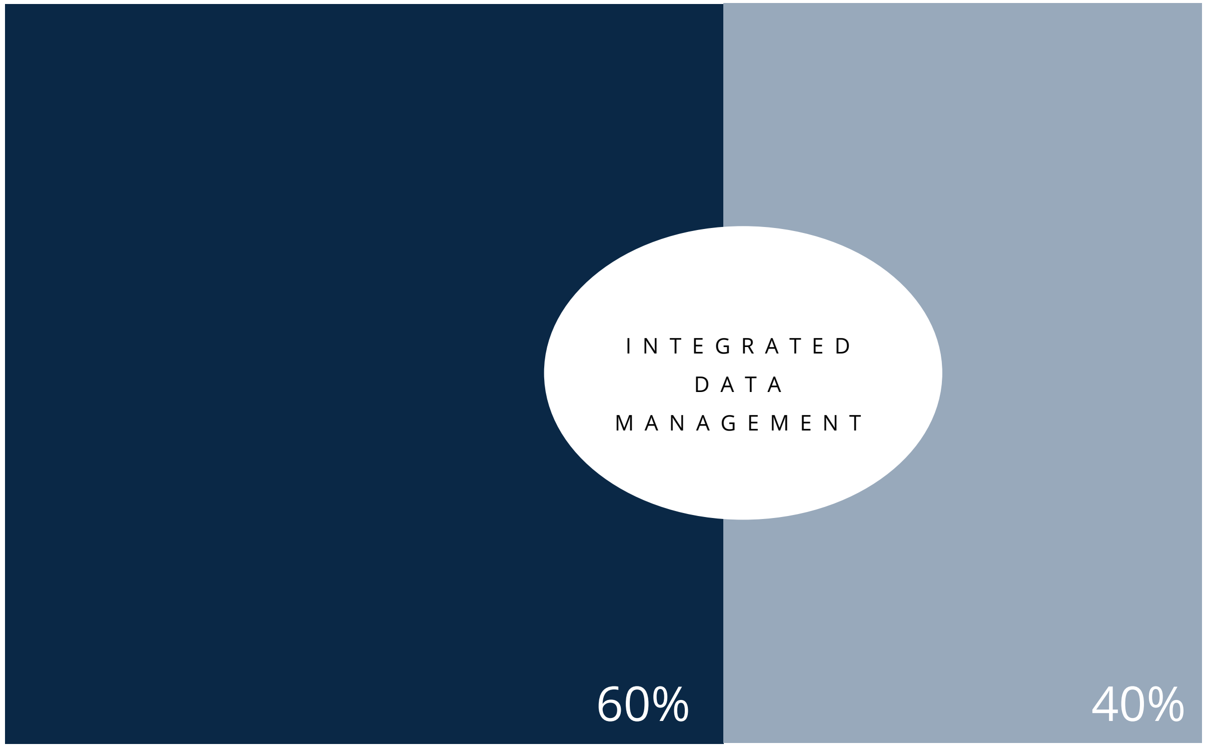
INTERNATIONAL INVESTIMENT POSITION

GENERAL GOVERNMENT DEBT

BANCO DE PORTUGAL'S STATISTICS DEPARTMENT

WHAT DO WE DO?

PRODUCE AND DISSEMINATE STATISTICS



MANAGE OTHER DATABASES

[CENTRAL CREDIT REGISTER,

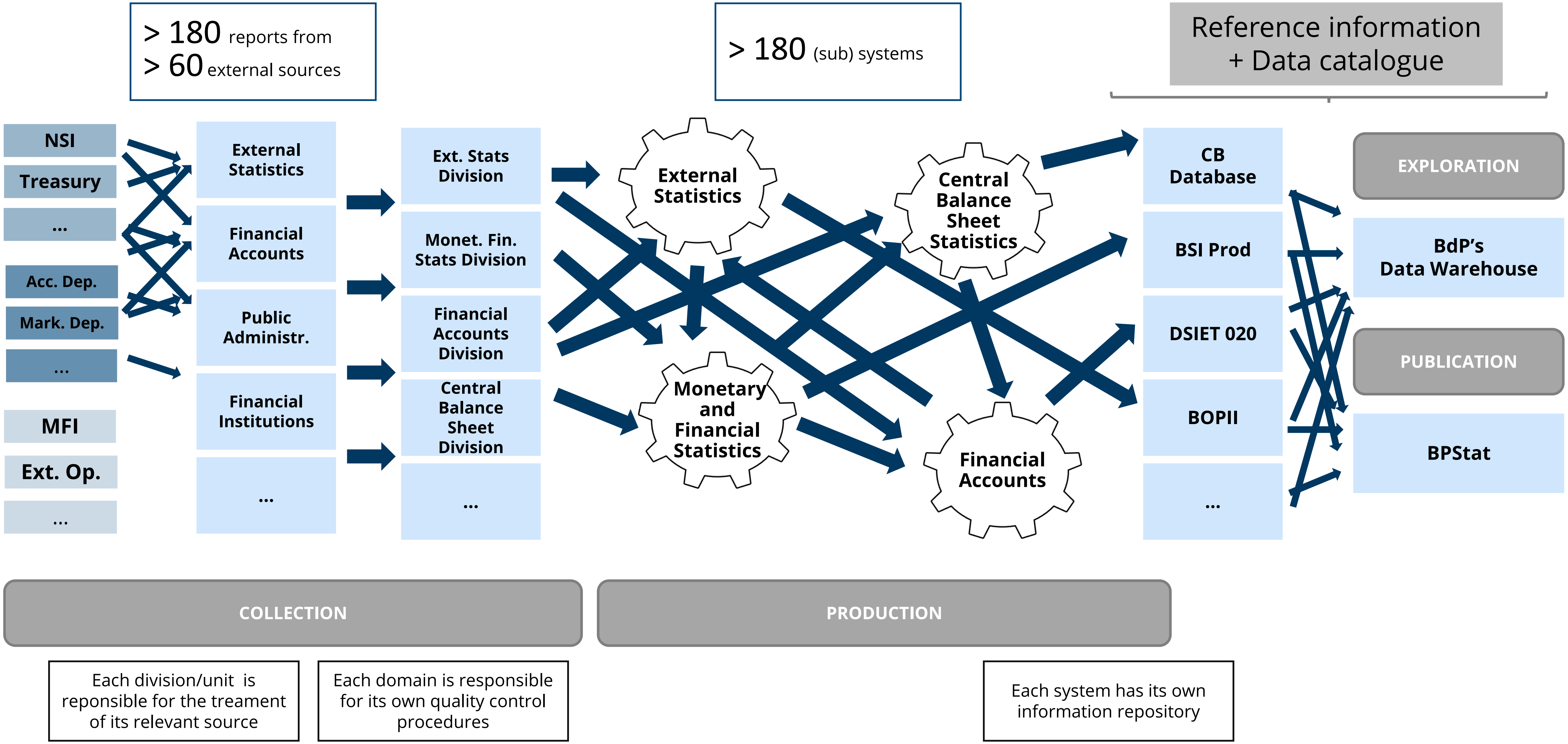
BANKING ACCOUNTS DATABASE,

IN-HOUSE CREDIT ASSESSMENT SYSTEM]

WHERE IT ALL STARTED

02

WHERE IT ALL STARTED



> 180 reports from
> 60 external sources

> 180 (sub) systems

Reference information
+ Data catalogue

NSI

Treasury

...

Acc. Dep.

Mark. Dep.

...

MFI

Ext. Op.

...

External Statistics

Financial Accounts

Public Administr.

Financial Institutions

...

Ext. Stats Division

Monet. Fin. Stats Division

Financial Accounts Division

Central Balance Sheet Division

...

External Statistics

Central Balance Sheet Statistics

Monetary and Financial Statistics

Financial Accounts

CB Database

BSI Prod

DSIET 020

BOPII

...

EXPLORATION

BdP's Data Warehouse

PUBLICATION

BPStat

COLLECTION

PRODUCTION

Each division/unit is reponsible for the treament of its relevant source

Each domain is responsible for its own quality control procedures

Each system has its own information repository

WHERE IT ALL STARTED

Main thoughts

Each domain was responsible for the treatment of its own relevant set of data.

Adapted to the purpose of each division.

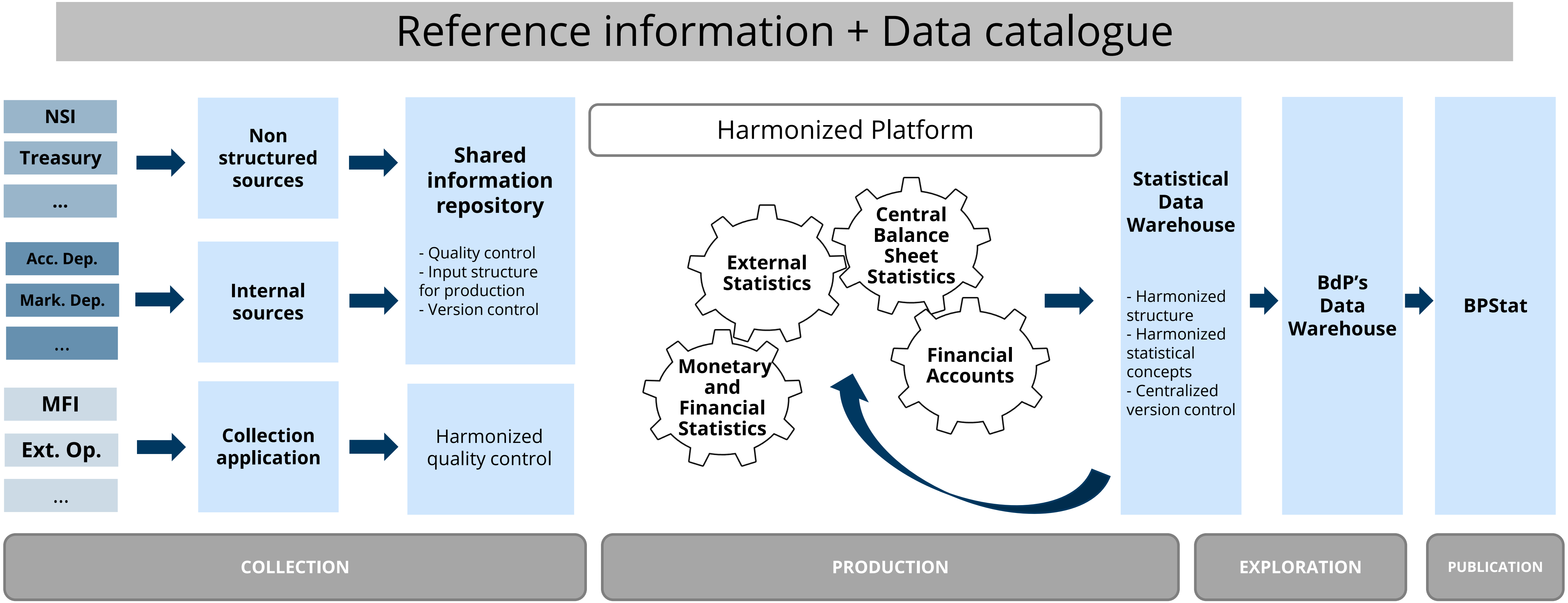
No harmonized concepts, processes or systems.

FUTURE VISION OF OUR STATISTICAL INFORMATION SYSTEMS

03

FUTURE VISION OF OUR STATISTICAL INFORMATION SYSTEMS

UNTANGLING THE WEB



Reference information + Data catalogue

NSI

Treasury

...

Acc. Dep.

Mark. Dep.

...

MFI

Ext. Op.

...

Non structured sources

Internal sources

Collection application

Shared information repository

- Quality control
- Input structure for production
- Version control

Harmonized quality control

Harmonized Platform

External Statistics

Central Balance Sheet Statistics

Monetary and Financial Statistics

Financial Accounts

Statistical Data Warehouse

- Harmonized structure
- Harmonized statistical concepts
- Centralized version control

BdP's Data Warehouse

BPStat

COLLECTION

PRODUCTION

EXPLORATION

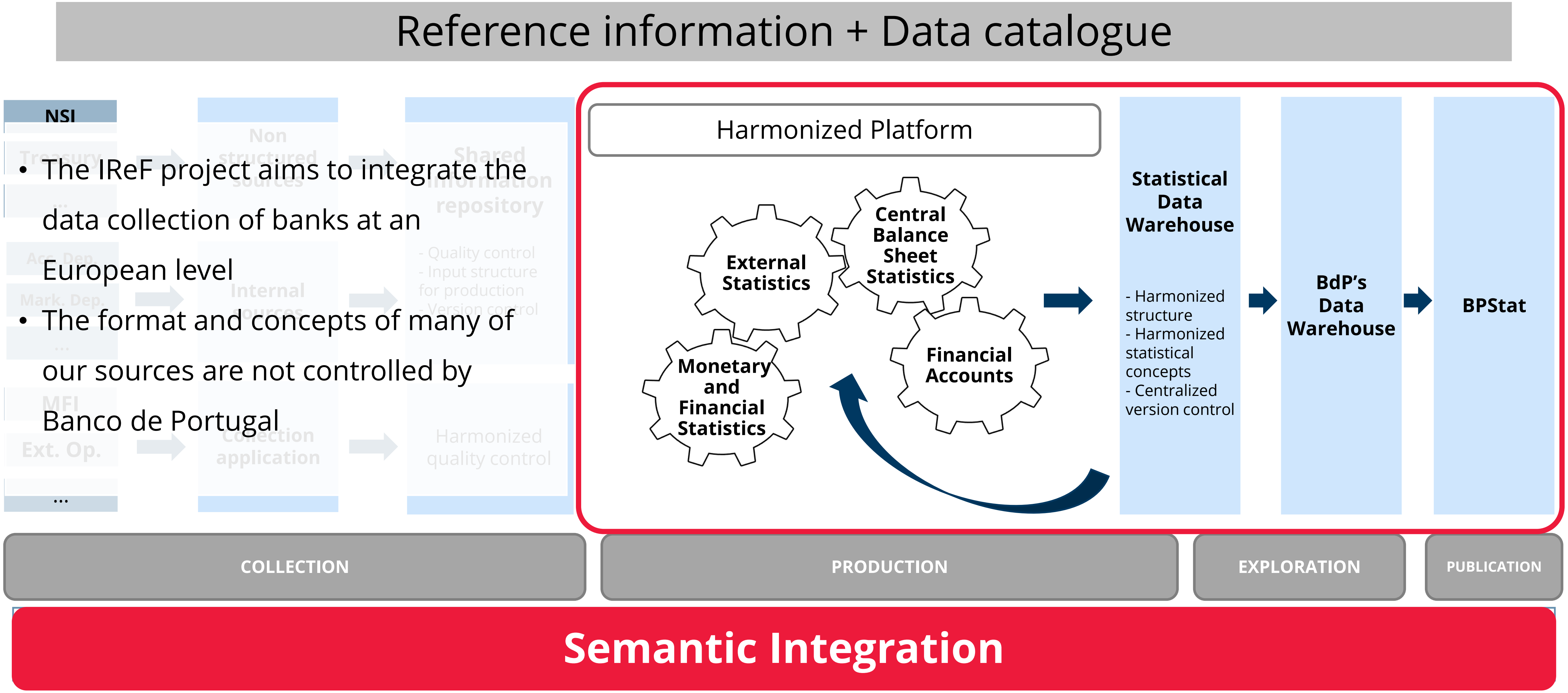
PUBLICATION

Simplify, Harmonize and Enable.

FUTURE VISION OF OUR STATISTICAL INFORMATION SYSTEMS

UNTANGLING THE WEB

- The IReF project aims to integrate the data collection of banks at an European level
- The format and concepts of many of our sources are not controlled by Banco de Portugal



FUTURE VISION OF OUR STATISTICAL INFORMATION SYSTEMS

UNTANGLING THE WEB

Main thoughts

Simplify and harmonize our domains.

Enhance our teams' knowledge on where and how to intervene.

Homogeneous and consolidated domains, easier to maintain and evolve.

Promote transparency, auditability and efficiency in our processes.

Information sharing, with less adaption costs.

Semantic integration as a cornerstone of our strategy.

DEFINING A COMMON SET OF CONCEPTS

04

DEFINING A COMMON SET OF CONCEPTS

WHY DOES IT MATTER?

ALL DOMAINS “TALK” THE SAME LANGUAGE.

HARMONIZED CONCEPTS FACILITATE COMPARISON.

PRODUCTION “COSTS” ARE REDUCED.

DIFFERENT INTERPRETATIONS ARE AVOIDED.

FULL KNOWLEDGE OF ALL THE DATA AVAILABLE IN-HOUSE.

DEFINING A COMMON SET OF CONCEPTS

UNHARMONIZED DATA – THE ISSUE

WHEN WE SEARCH FOR THE DIMENSION **INSTRUMENT**, WHAT DO WE FIND IN THE DATABASES?

TRA_IDBP_TITLE TYPEINSTRUMENTID EMI_IDBP_TITLE

DESIGLINE VARIABLE COD_SEC INSTRUMENT

INSTRUMENT

ITEM INST_FINANCIAL CCE

CI INSTRUMENT TYPE

FINANCIALINSTRUMENT

DEFINING A COMMON SET OF CONCEPTS

UNHARMONIZED DATA – THE ISSUE

MULTIPLE AND DIFFERENT SOURCES, REGULATIONS AND REPORTS.

CONCEPTS WITH HETEROGENEOUS DESCRIPTIONS.

DIMENSIONS (VARIABLES) WITH MULTIPLE CONCEPTS.

↓

DIFFICULTIES WHEN CROSSING DATA BETWEEN DOMAINS.

DEFINING A COMMON SET OF CONCEPTS

UNHARMONIZED DATA – HOW CAN WE SOLVE IT?

A TASK FORCE TO HARMONIZE CONCEPTS.

DEFINING A COMMON SET OF CONCEPTS

HARMONIZATION STRATEGY

A | DEFINE A GLOSSARY

B | COLLECT STATISTICAL OUTPUTS

C | HARMONIZATION CYCLE

D | RESULTS

DEFINING A COMMON SET OF CONCEPTS

A) DEFINE A GLOSSARY

Investment funds (example)

Original maturity (example)

Over 1 and up to 2 years (example)

PRODUCTION OUTPUT

Production output is a table/database that contains the result of statistical compilation at the most granular level.

DIMENSION

The dimension provides an overview of the statistics produced.

It's composed by:

- **designation** (dimension short code)
- **description** (name that describes the dimension).

MEMBER

Members are specific characteristics of the dimension.

Each member is composed by:

- **code** (member short code)
- **name** (member code description).

DEFINING A COMMON SET OF CONCEPTS

B) COLLECT STATISTICAL OUTPUTS

DIMENSIONS	MEMBERS	HUMAN RECOURCES	BUSINESS AREAS	STATISTICAL OUTPUTS
436	7205	5	6	24

DEFINING A COMMON SET OF CONCEPTS

B) COLLECT STATISTICAL OUTPUTS

MEMBERS

7205

HUMAN RECOURCES

5

BUSINESS AREAS

6

STATISTICAL OUTPUTS

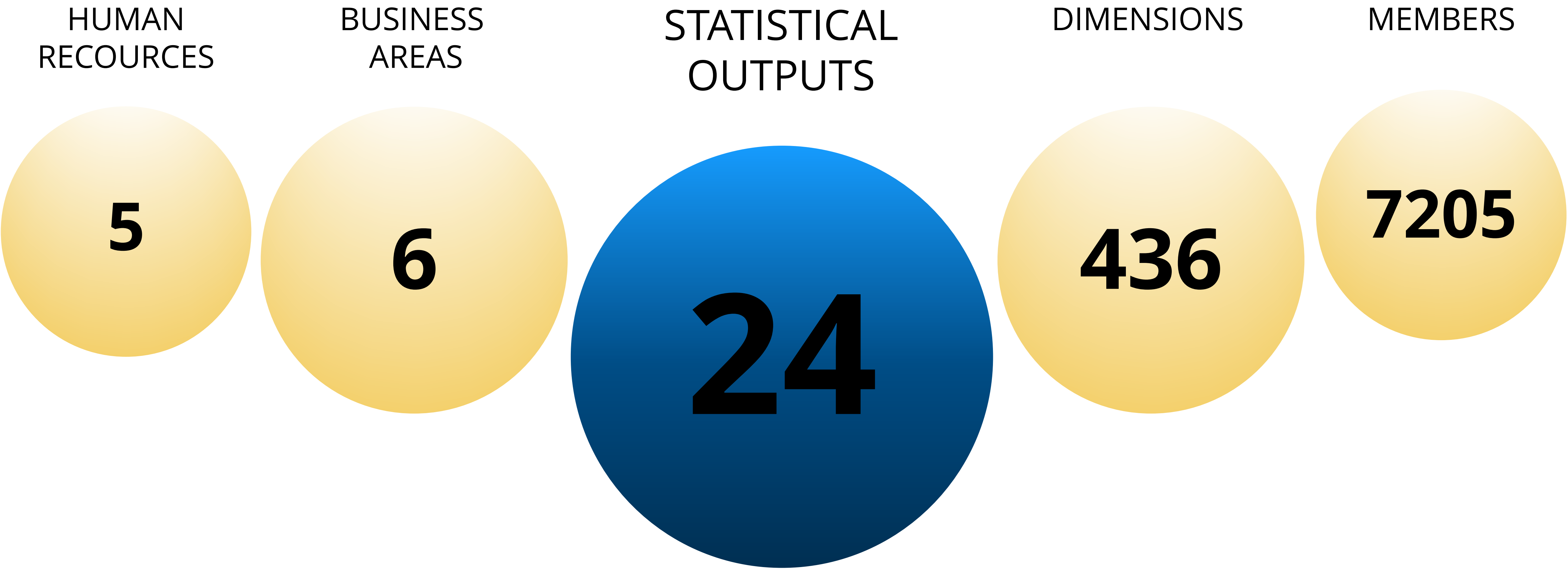
24

DIMENSIONS

436

DEFINING A COMMON SET OF CONCEPTS

B) COLLECT STATISTICAL OUTPUTS



HUMAN RECOURCES

5

BUSINESS AREAS

6

STATISTICAL OUTPUTS

24

DIMENSIONS

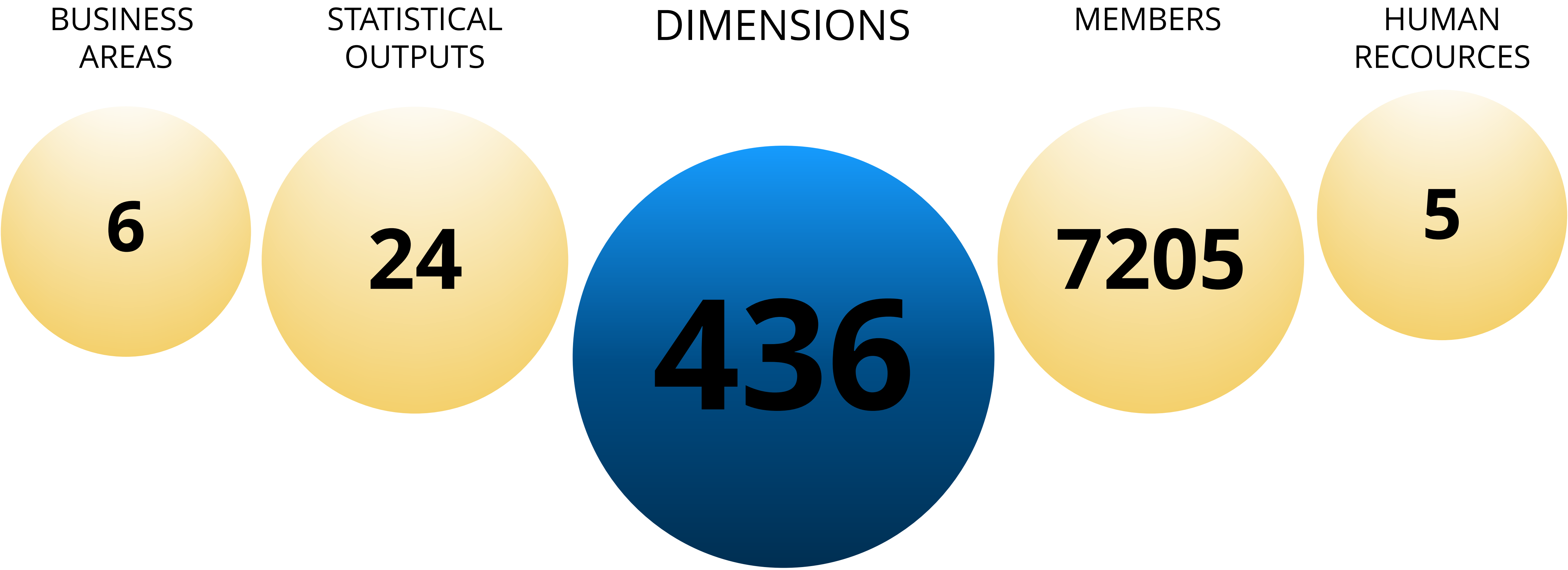
436

MEMBERS

7205

DEFINING A COMMON SET OF CONCEPTS

B) COLLECT STATISTICAL OUTPUTS



BUSINESS AREAS

6

STATISTICAL OUTPUTS

24

DIMENSIONS

436

MEMBERS

7205

HUMAN RECOURCES

5

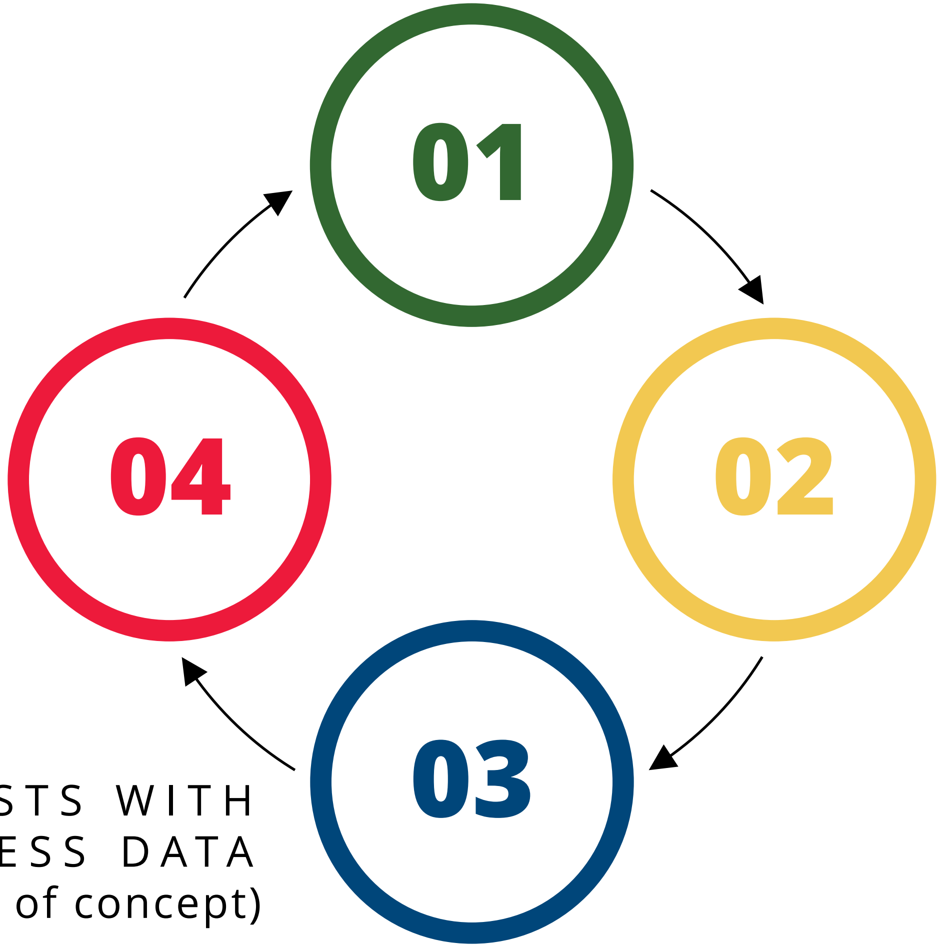
DEFINING A COMMON SET OF CONCEPTS

B) COLLECT STATISTICAL OUTPUTS

STATISTICAL OUTPUTS	DIMENSIONS	MEMBERS	HUMAN RECOURCES	BUSINESS AREAS
24	436	7205	5	6

DEFINING A COMMON SET OF CONCEPTS

C) HARMONIZATION CYCLE



ANALYSE DIMENSIONS AND
GROUP ALL MEMBERS

HARMONIZE MEMBERS
AND STUDY CONCEPTS

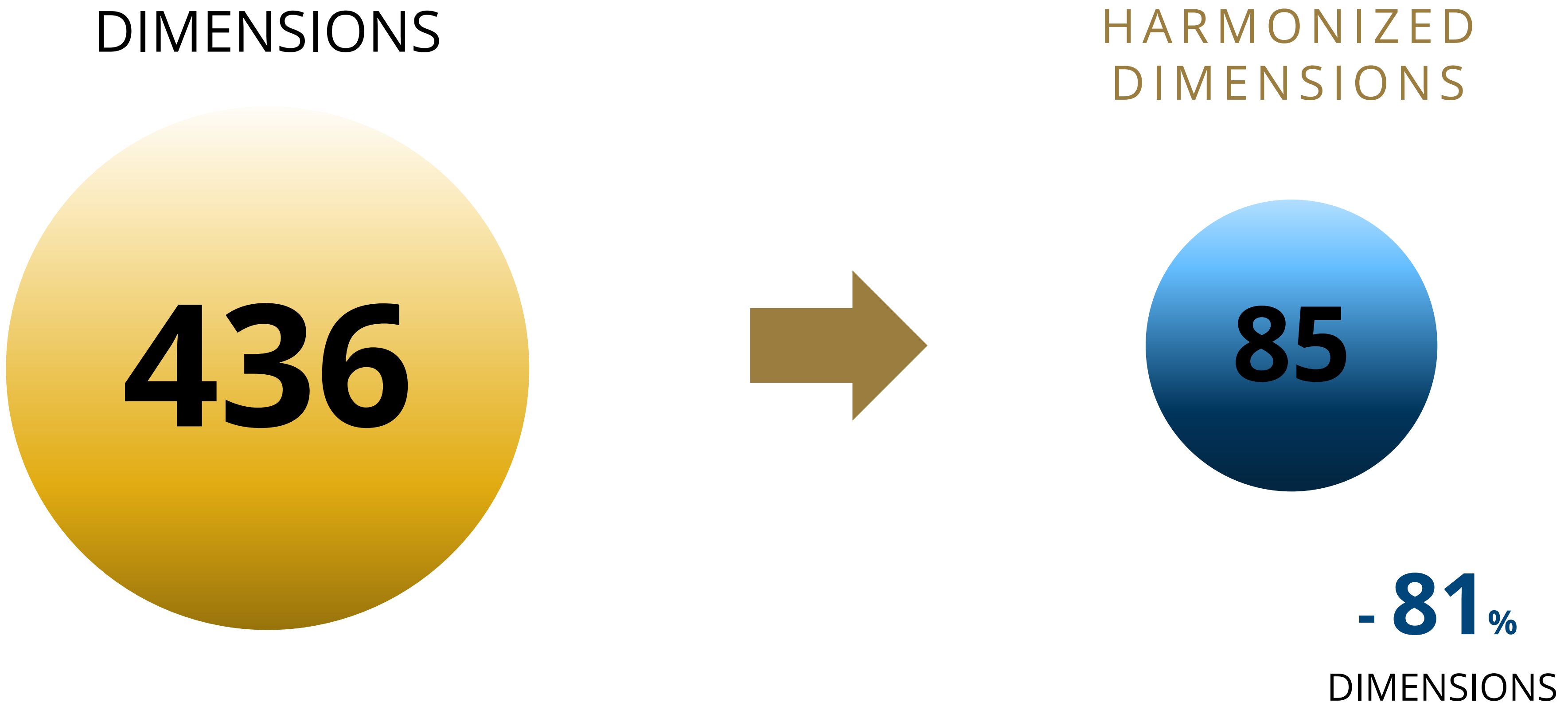
- ✓ Define mapping tables
- ✓ Define aggregation tables

TESTS WITH
BUSINESS DATA
(Proof of concept)

INCORPORATE
IMPROVEMENTS (RETEST
THE RESULTS) AND DEFINE
GOVERNANCE MODEL

DEFINING A COMMON SET OF CONCEPTS

D) RESULTS



DIMENSIONS

436

HARMONIZED DIMENSIONS

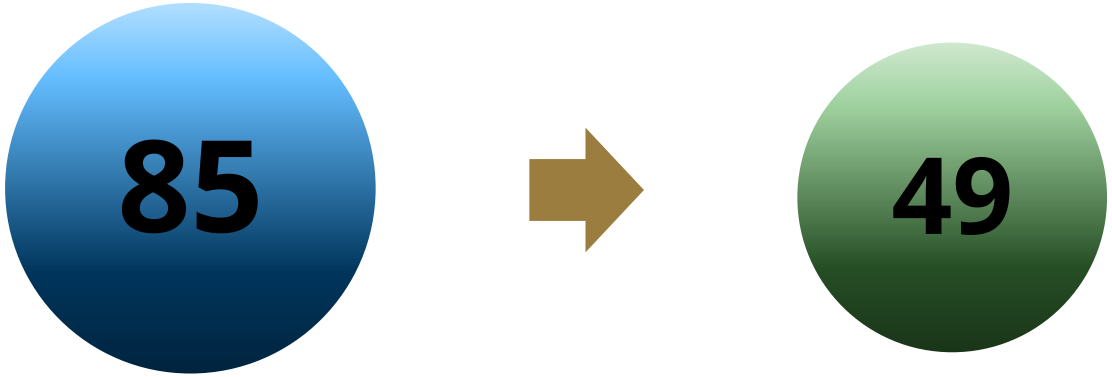
85

- 81%

DIMENSIONS

DEFINING A COMMON SET OF CONCEPTS

D) RESULTS



HARMONIZED DIMENSIONS

CODELISTS

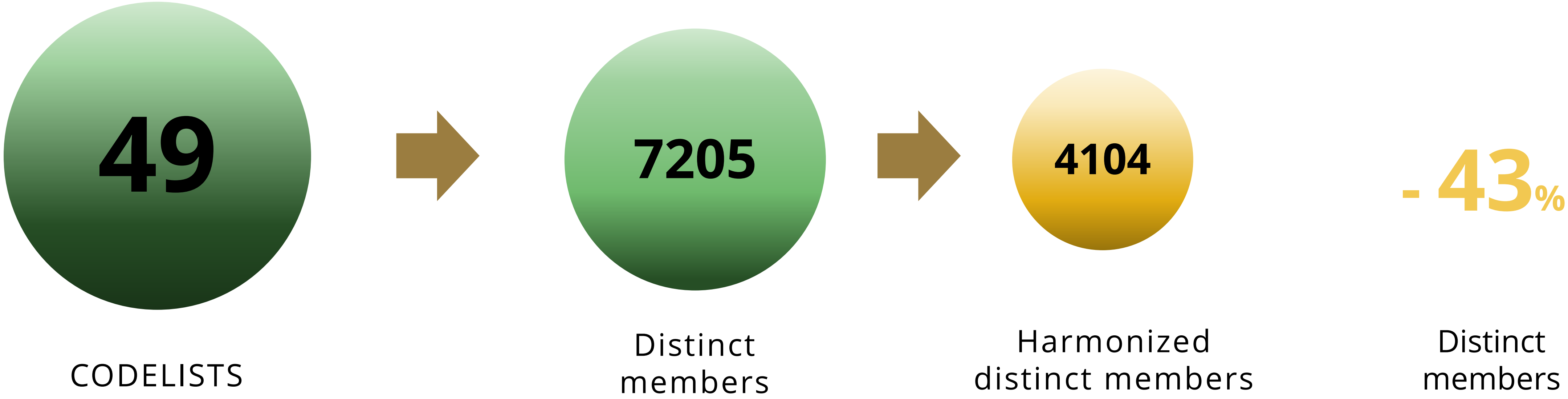
Example:

Original maturity
Residual maturity
Fixation rate maturity

Maturity

DEFINING A COMMON SET OF CONCEPTS

D) RESULTS



49

CODELISTS

7205

Distinct members

4104

Harmonized distinct members

- 43%

Distinct members

DEFINING A COMMON SET OF CONCEPTS

GOVERNANCE PRINCIPLES

USE ALPHANUMERIC CODES, STARTED WITH A LETTER.

FOLLOW THE STANDARD SDMX CODELISTS DEFINED IN THE SINGLE DATA DICTIONARY (SDD).

FOLLOW OTHER CODELISTS IN SDD, IF THERE IS NO MATCH FOR THE CONCEPT IN SDMX.

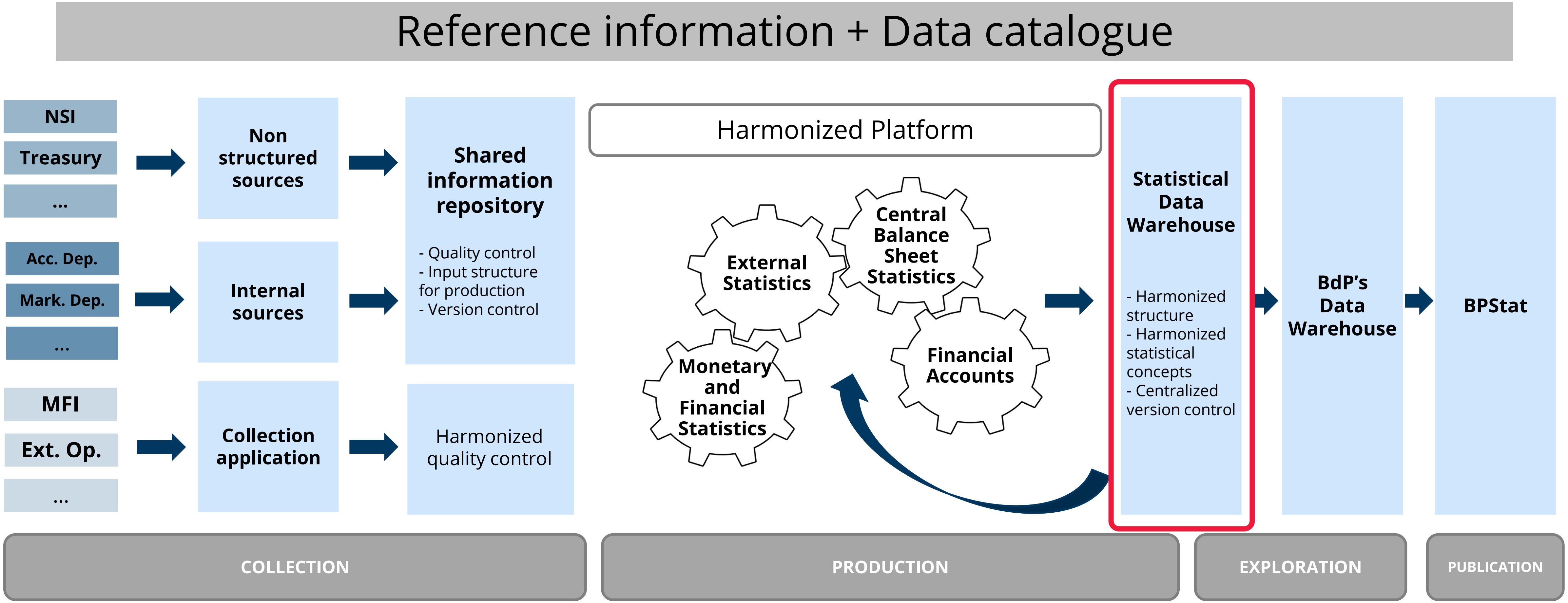
USE INTELLIGIBLE DESIGNATIONS AND CODES.

DESCRIBE THE UNDERLYING RULES FOR THE CODES OF EACH DIMENSION.

LIMIT CODE SIZE.

DEFINING A COMMON SET OF CONCEPTS

GOVERNANCE PRINCIPLES: WHERE TO START?



Reference information + Data catalogue

NSI
Treasury
…

Acc. Dep.
Mark. Dep.
…

MFI
Ext. Op.
…

Non structured sources

Internal sources

Collection application

Shared information repository

- Quality control
- Input structure for production
- Version control

Harmonized quality control

Harmonized Platform

External Statistics

Central Balance Sheet Statistics

Monetary and Financial Statistics

Financial Accounts

Statistical Data Warehouse

- Harmonized structure
- Harmonized statistical concepts
- Centralized version control

BdP's Data Warehouse

BPStat

COLLECTION

PRODUCTION

EXPLORATION

PUBLICATION

Simplify, Harmonize and Enable.

DEFINING A COMMON SET OF CONCEPTS

GOVERNANCE PRINCIPLES: WHERE TO START?

STATISTICAL DATA WAREHOUSE.

REFORMULATION OF STATISTICAL PRODUCTION PROCESSES.

PRODUCTION OF NEW STATISTICS.

NEW DIMENSION AND MEMBERS.

DEFINING A COMMON SET OF CONCEPTS

IMPLEMENTATION

WITHIN PRODUCTION PROCESSES

4 processes already adopted the new set of concepts

WHEN STORING STATISTICAL OUTPUTS IN THE DATAWAREHOUSE

3 statistical outputs translated into the new set of concepts

WORK IN PROGRESS

Planned integration of all statistical outputs translated into the new set of concepts by the end of 2024

CLOSING REMARKS

05

CLOSING REMARKS

SEMANTIC INTEGRATION IS A CORNERSTONE OF OUR STRATEGY.

CENTRALIZED STATISTICAL PRODUCTION PLATFORM.

CROSS-REFERENCE OF DIFFERENT DOMAINS (COHERENCE AND OVERALL QUALITY).

HARMONIZATION IS ALWAYS AN ON-GOING PROCESS.

IMPLEMENTATION IS STILL A CHALLENGE: A STEPWISE APPROACH WAS CONSIDERED MORE FEASIBLE.

“COMMON LANGUAGE”.

QUESTIONS