Mestrado em Engenharia Informática e Tecnologia Web

Data Model Definition and Implementation for BIRD and IReF Framework Enablement

DISSERTAÇÃO SUBMETIDA PARA A OBTENÇÃO DO GRAU DE MESTRE

Marco António Antunes Clara

Orientação: Prof. Doutor Henrique São Mamede
Coorientação: Prof. José Martins

Versão: fevereiro de 2026

Condições de utilização (licença *Creative Commons*)

Agradecimentos

Agradeço aos Professores Henrique Mamede e José Martins por todo o seu acompanhamento, pela orientação, pela motivação, pelo tempo que dedicaram e pela amizade que demonstraram ao longo do último ano, ajudando-me assim a desenvolver, completar e melhorar este meu projeto.

Gostaria de agradecer à minha esposa Sofia e aos meus filhos Alexandre e Pedro, pelo apoio ao longo dos últimos meses, durante os quais me ajudaram, construindo um ambiente positivo em meu redor que foi determinante para conseguir alcançar os meus objetivos académicos.

Um agradecimento também especial aos meus pais por, desde cedo, sempre me motivarem e incentivarem a dar o meu melhor e a alcançar todo o meu potencial.

Agradeço ainda aos meus colegas por todo o apoio e motivação que sempre fizeram questão de me fazer sentir e por me incentivarem ao longo desta jornada.

Um grande obrigado a todos aqueles que de alguma forma contribuíram para o meu sucesso neste projeto.

Declaração de Integridade

Declaro ter atuado com integridade na elaboração da presente dissertação/do presente relatório. Confirmo que em todo o trabalho conducente à sua elaboração não recorri à prática de plágio ou a qualquer outra forma de falsificação de resultados.
Mais declaro que tomei conhecimento integral do Regulamento Disciplinar da Universidade Aberta, publicado no Diário da República, 2.ª série, n.º 215, de 6 de novembro de 2013.

I hereby declare having conducted my thesis/report work with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of writing this thesis/report.

I further declare that I have fully acknowledged Disciplinary Regulations of the Universidade Aberta (regulation published in the official journal "Diário da República", 2nd series, No. 215, dated November 6, 2013).

Universidade Aberta, 1 de fevereiro de 2026

Nome completo / Full name: Marco António Antunes Clara

Assinado por: **Marco António Antunes Clara**
Num. de Identificação: 11707595
Data: 2026.04.30 11:03:21+01'00'

Assinatura / Signature: _____
(manuscrita ou digital / handwritten or digital)

CHAVE MÓVEL

Resumo

O setor bancário tem sido, nos últimos anos, alvo de uma vasta gama de regulamentos operacionais, com um foco significativo em processos, dados, aplicações e arquiteturas tecnológicas. Esta regulação e escrutínio visa estabelecer a garantia de resiliência destas instituições em todos os níveis, não apenas no setor financeiro. Neste contexto, o setor bancário será em breve obrigado a padronizar relatórios periódicos às autoridades supervisoras europeias (Banco de Portugal, Banco Central Europeu, Agência Bancária Europeia).

O BIRD (Banks' Integrated Reporting Dictionary), já criou um modelo conceptual para a respetiva especificação, que irá alimentar diretamente o IReF (Integrated Reporting Framework). Assim sendo, é urgente mapear os requisitos que permitam implementar um sistema informático com base nos conceitos subjacentes a estas referências.

No contexto do presente trabalho, pretende-se realizar uma análise de lacunas, bem como o desenvolvimento de um artefacto que permita dar resposta aos desafios inerentes à transição do atual modelo de reporte bancário para o novo modelo preconizado pelo BIRD e IReF.

As questões de investigação assumidas como essenciais para atingir os desafios acima descritos, são as seguintes: a) Quais são os conceitos e relações entre eles, que deveriam fazer parte de um modelo lógico de dados, que reúne toda a informação relevante de diferentes instituições bancárias, com vista a um relatório unificado num quadro centralizado para supervisores do setor bancário?; b) A partir dos conceitos e relações modeladas, é possível especificar requisitos que permitam a implementação de um modelo lógico que suporte o modelo conceptual definido no BIRD e a sua integração no IReF?; e c) Esta implementação pode ser realizada de acordo com a norma BCBS 239 (Comité de Basileia para Supervisão Bancária) ao abrigo dos "Princípios para agregação eficaz de dados de risco e reporte de risco"?

As principais conclusões incluem a cobertura e a integração do quadro, a abordagem arquitetónica, os requisitos de implementação, os desafios significativos, a governação de dados, o alinhamento regulatório, a direção futura, as limitações de investigação, a integração tecnológica e a garantia de qualidade. As contribuições-chave,

para além do artefacto desenvolvido e que poderá servir como um pontapé de saída para desenvolvimento aplicacional de novos processos, estão relacionadas com a definição de padrões arquitetónicos, de integração, garantia de qualidade, especificação de requisitos técnicos, metodologia de desenvolvimento, padrões de gestão de dados e padrões de desenho de APIs.

Palavras-chave: BIRD, IReF, banks integrated reporting dictionary, integrated reporting framework, BCBS 239

Abstract

The banking sector has, in recent years, been the target of a wide range of operational regulations, with a significant focus on processes, data, applications and technological architectures. This regulation and scrutiny aim to establish the resilience of these institutions at all levels, not just in the financial sector. In this context, the banking sector will soon be required to standardise periodic reports to the European supervisory authorities (Bank of Portugal, European Central Bank, European Banking Agency).

The BIRD (Banks' Integrated Reporting Dictionary) has already created a conceptual model for its specification, which will feed directly into the IReF (Integrated Reporting Framework). Therefore, it is urgent to map the requirements that allow the implementation of a computer system based on the concepts underlying these references.

In the context of this work, it is intended to carry out a gap analysis, as well as the development of an artifact that will allow to respond to the challenges inherent to the transition from the current banking reporting model to the new model recommended by the IBRD and IReF.

The research questions assumed as essential to achieve the challenges described above are the following: a) What are the concepts and relationships between them, which should be part of a logical data model, which brings together all the relevant information from different banking institutions, with a view to a unified report in a centralized framework for banking sector supervisors?; b) From the concepts and modelled relationships, is it possible to specify requirements that allow the implementation of a logical model that supports the conceptual model defined in the IBRD and its integration into the IReF?; and c) Can this implementation be carried out in accordance with BCBS 239 (Basel Committee on Banking Supervision) under the "Principles for Effective Risk Data Aggregation and Risk Reporting"?

Key findings include framework coverage and integration, architectural approach, implementation requirements, significant challenges, data governance, regulatory alignment, future direction, research limitations, technology integration, and quality assurance. The key contributions, in addition to the developed artifact that can serve as

a kick-off for the application development of new processes, are related to the definition of architectural standards, integration, quality assurance, specification of technical requirements, development methodology, data management standards and API design standards.

Keywords: BIRD, IReF, banks integrated reporting dictionary, integrated reporting framework, BCBS 239

Contents

Table Index

Figure Index

List of Abbreviations

Table 1 - List of Abbreviations

Abbreviation	Full Term	Description / Context
AnaCredit	Analytical Credit Datasets	AnaCredit is a dataset containing detailed information on individual bank loans in the euro area, harmonised across all Member States
BIRD	Banks' Integrated Reporting Dictionary	A harmonized data model developed by the ECB to support banks in regulatory reporting
BdP	Banco de Portugal	Portuguese central bank for regulation
COREP	Common Reporting	EBA framework for capital adequacy and risk reporting
CRD	Capital Requirements Directive	EU legislation governing capital adequacy of financial institutions
CRR	Capital Requirements Regulation	Key component of the European Union's prudential regulatory framework for financial institutions. It works alongside the Capital Requirements Directive (CRD) to implement the Basel III standards in the EU
DPM	Data Point Model	EBA's framework for defining regulatory reporting requirements using structured data points
EBA	European Banking Authority	EU authority responsible for banking regulation and supervision
ECB	European Central Bank	Central bank for the eurozone; initiator of BIRD and IReF
EIL	Extended Input Layer	Harmonized version of IL using BIRD definitions; enables consistent transformation

ESRB	European Systemic Risk Board	Institution monitoring systemic risks in the EU financial system
FINREP	Financial Reporting	EBA framework for financial reporting by institutions
IL	Input Layer	Raw data from banks' internal systems, unharmonized and institution-specific
IReF	Integrated Reporting Framework	ECB initiative to integrate statistical reporting requirements into a single framework
ITS	Implementing Technical Standards	Detailed rules issued by EBA to implement EU regulations
NACE	Nomenclature of Economic Activities	EU classification system for economic activities, used in counterparty reporting
NROL	Non-Reference Output Layer	Output layer using non-reference codes (e.g., EBA DPM) tailored to actual reporting formats
RIAD	Register of Institutions and Affiliates Data	ECB database of institutional reference data
ROL	Reference Output Layer	Standardized regulatory requirements using reference codes and definitions
SDMX	Statistical Data and Metadata eXchange	ISO standard for exchanging statistical data and metadata
SHS	Securities Holdings Statistics	Information on securities held by euro area resident sectors, broken down by instrument type, holder country and further classifications
XBRL	eXtensible Business Reporting Language	Standard for digital business reporting, used in regulatory submissions

Introduction

Over the past decade, the banking sector has experienced an unprecedented intensification of regulatory scrutiny, particularly regarding the integrity, governance, and interoperability of its data and technological infrastructures. Supervisory authorities across Europe have progressively reinforced requirements aimed at ensuring that financial institutions maintain high levels of operational resilience, not only from a financial standpoint, but also in their processes, data management practices, application landscapes, and underlying technology architectures. This ongoing regulatory evolution reflects a bigger goal: to strengthen stability, transparency, and comparability of information used for prudential supervision (Penczar et al., 2023).

Within this context, European regulators such as Banco de Portugal (BdP), the European Central Bank (ECB), and the European Banking Authority (EBA) are converging toward a more harmonised and standardised regulatory reporting model. Traditional reporting mechanisms, characterised by fragmented data definitions, heterogeneous formats and interpretations, are gradually being replaced by integrated, concept-driven frameworks. Two initiatives stand out in this transformation: the Banks' Integrated Reporting Dictionary (BIRD) and the Integrated Reporting Framework (IReF) (Flötgen et al., 2020).

BIRD provides a comprehensive and standardised data dictionary that defines the concepts, transformations, and structures required to meet the reporting obligations of financial institutions. Developed collaboratively by the ECB, national authorities, and participating banks, it aims to reduce reporting complexity by harmonising data semantics and promoting the use of reusable data models. IReF, on the other hand, proposes integrating multiple statistical and supervisory reports into a single, coherent reporting framework applicable across the euro area. In practice, BIRD offers the conceptual input data definitions and transformation rules that will feed the IReF reporting framework (Dorrer, 2024).

Given the strategic importance and implementation horizon of these initiatives, it becomes essential for financial institutions to understand and operationalise the

requirements associated with BIRD and IReF. This includes not only interpreting conceptual models but also designing technological solutions that translate regulatory specifications into functional, automated, and scalable systems. Mapping these requirements and exploring feasible architectural approaches is, therefore, both an urgent and necessary task.

It is within this framework that the present dissertation is situated. The goal is to develop a practical, implementable proof of concept that demonstrates how a computer system may be constructed using the conceptual foundations defined by BIRD to support the IReF reporting obligations. By aligning academic research with a real and imminent regulatory challenge, this project seeks to contribute meaningfully to the operational readiness of financial institutions and to the broader evolution of regulatory reporting in Europe.

Motivation

This study, which, from a functional point of view, concerns the regulation of the banking sector, addresses a cross-cutting theme across several areas of information systems, with a greater focus on topics such as **data management**, **analytics**, and **data governance**. Therefore, exploring these and other aspects throughout the study will be essential to achieve the proposed objectives.

It also aims to set a concrete objective: to discover a **path to build a document mapping the requirements** for regulatory reporting that the BIRD establishes, **allowing the creation of a data model** resulting from mapping these requirements and the concepts and relationships identified.

So far, to the best of our knowledge, there is no solution to this problem, and its relevance increases by the day, as it is a regulatory matter. Therefore, the definition of a due date is expected to be released soon. This type of unification in terms of reporting does not currently exist. When it does, it will be in the banks' best interest to be prepared to respond to this type of request.

One of the biggest challenges foreseen during this study is naturally related to the heterogeneity of data (and its models) across banking institutions. Over the last decades, banks have evolved and expanded their information systems ecosystem, with some shared

and familiar concepts, although defined differently across institutions. Thus, it will be essential to map these concepts to establish a common language that enables unified information exchange and sharing with regulatory authorities in the context of their supervisory tasks.

Document Structure

The present document is structured in 6 chapters:

Chapter 1 contains the study scope, namely by describing the **research questions** and **study objectives**.

Chapter 2 contains the **literature review,** including the theoretical and methodological aspects, limitations, search details, and, ultimately, the selected literature to be used in the following sections.

Chapter 3 contains the **methodology details**, namely specifying the action research methodology stages (planning, action, observation and reflection) and the attained results.

Chapter 4 contains the **implementation details**, namely, the specific information regarding a proof of concept as an artifact.

Chapter 5 aggregates the **results**, both for the state-of-the-art analysis and for the implementation results from the proof of concept.

Chapter 6 presents the **study's conclusions**, including key findings, implications, limitations, recommendations, and contributions.

Finally, after the **references** section, there is an **appendix** section where additional information is provided, including links to detailed search reports, an extraction matrix for the selected literature, a PRISMA 2020 diagram, and links to the source code for the proof of concept.

Chapter 1 - Objectives & Research Problem

1.1 Objectives

The central objective of this dissertation is to investigate how BIRD can support the implementation of the IReF and to demonstrate, through a practical proof of concept, how a logical data model may be designed to operationalise this integration. By combining conceptual analysis, literature review, and technical experimentation, the work aims to contribute academically and practically to the evolving landscape of regulatory reporting.

To achieve this, a literature review was conducted focused on the academic, institutional and technical foundations of integrated regulatory reporting. This effort involves identifying and synthesising research on conceptual data models, reporting frameworks, and the applicability of BCBS 239. The goal is to establish a scientific foundation for the architectural decisions and methodological approaches adopted throughout the study.

Following the literature review, the conceptual structure of the BIRD framework was examined. This analysis includes identifying the key entities, relationships and transformation rules that define BIRD's metadata model, as well as understanding how these support harmonised reporting across institutions. The study also explores data requirements and structural expectations of IReF, assessing how the ECB envisions consolidation of statistical and supervisory reporting into a unified framework.

Building on this understanding of BIRD and IReF, a mapping of BIRD's conceptual structures to the reporting requirements defined by IReF was conducted. This task includes identifying alignment, conducting gap analysis, and defining the logical correspondences necessary for an integrated reporting process. The goal is to establish how BIRD's semantic and transformation layers can support IReF's reporting.

Based on this conceptual mapping, a logical data model was designed to host BIRD concepts and enable the production of IReF-compliant outputs. The model is intended to be scalable, interoperable, and transparent, reflecting regulatory expectations for data

quality and automation. This architectural proposal serves as the theoretical blueprint for the practical implementation to be developed later.

To validate the proposed architecture, a proof-of-concept was developed to demonstrate the feasibility of integrating BIRD structures into an IReF reporting process. This prototype demonstrates the model's ingestion, transformation, and output capabilities, illustrating its practical relevance and feasibility.

Finally, there is a critical reflection on the findings obtained throughout the study, analysing the opportunities, challenges, and limitations of the proposed approach.

1.2 Research Problem

- What are the concepts and relationships between them, which should be part of a logical data model, which houses all relevant information from different banking institutions, with a view to unified reporting in a centralised framework for banking sector supervisors?
- From the concepts and modelled relationships, can requirements be specified to implement a logical model that supports the conceptual model specified in the BIRD and integrates it into the IReF?
- Can this implementation be carried out in accordance with the BCBS 239 standard (Basel Committee on Banking Supervision) under "Principles for effective risk data aggregation and risk reporting"?

The main issue defines the maximum objective, which in this case consists of enabling banks' centralised reporting process to supervisory entities by specifying requirements and implementing a logical model that supports BIRD and enables IReF.

The sub-questions define how this should be done, namely through the mapping between a logical model and a conceptual model (BIRD) and the feasibility of the reporting framework (IReF) through their integration, in compliance with the BCBS 239 standards.

Chapter 2 - Literature Review

The main objective of the current literature review is to collect all relevant information that will allow us to establish the current "state of the art" regarding this matter and start working based on that. This will enable the definition of requirements and the implementation of a logical model in SQL to support the BIRD concept model and the IReF framework.

Following this initial step, the main task will be to specify a highly complex model that allows mapping concepts in data models while strictly following the regulations. Exploring specific themes in information systems, particularly data management and analytics, will be essential to this project.

Given the relevance and criticality of privacy issues within this sector, any constraints on implementing the project may include access to data for the analysis and exploration phase of data models with banking information.

This section describes how the literature review was performed by showing two distinct approaches (Systematic Literature Review and Multivocal Literature Review). It is therefore structured in five subsections: **SLR methodology**, **MLR methodology**, **study limitations**, **literature search** and **selected literature**.

The explanation for the need to move from SLR to MLR is also described in more detail in this section. The outcomes and key interpretations of the adopted research methodology are described in the **results and discussion** section, and the **conclusion** section presents the key findings and contributions of the study.

2.1 Systematic Literature Review Methodology

The first methodological approach used to conduct the literature review involved eleven distinct steps (Figure 1).

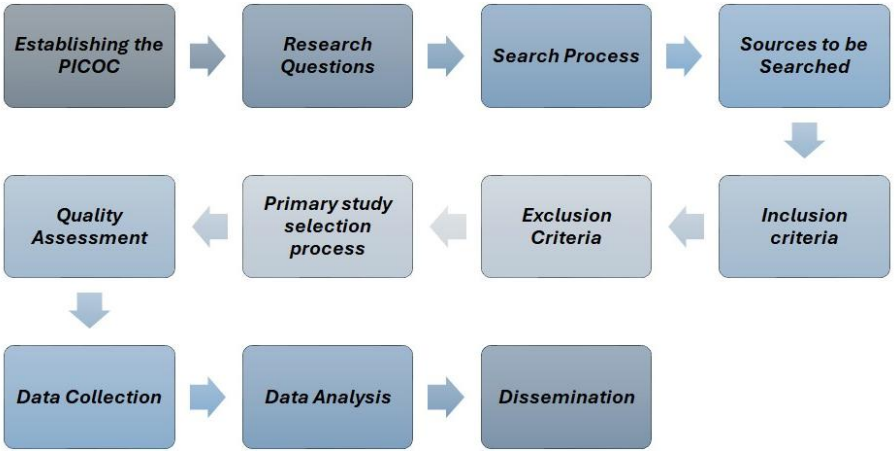


Figure 1 - SLR Protocol Steps

2.1.1 PICOC

As part of the research questions definition process, the first relevant step is to establish the PICOC (Population, Intervention, Comparison, Outcome and Context):

Table 2 - PICOC Definition

Criteria	Definition
Population	The population addressed in this study consists of European credit institutions and national central banks (Bank of Portugal, European Central Bank, European Banking Agency) that are subject to statistical and supervisory reporting obligations under the European System of Central Banks (ESCB). These entities represent the primary stakeholders in the reporting process, including data providers such as local banks (countrywide) and data receivers such as central banks and the European Central Bank (ECB). The relevance of this population lies in its direct exposure to regulatory requirements and the operational challenges associated with fragmented reporting frameworks across jurisdictions.
Intervention	The intervention under consideration targets the integration of the Banks' Integrated Reporting Dictionary (BIRD) with the Integrated Reporting Framework (IReF). This integration aims to streamline and standardize reporting requirements for credit institutions by leveraging BIRD as a metadata and transformation layer to implement IReF specifications. Through this approach, institutions can reduce complexity in reporting processes, automate data transformations and ensure compliance with harmonized standards, ultimately improving efficiency and data quality.

Comparison	The comparison is drawn against the current state of reporting, which is characterized by fragmented frameworks, multiple templates, and overlapping requirements. Traditionally, institutions rely on manual mapping and interpretation of reporting obligations, which increases operational burden and risk of inconsistencies. By contrasting this with an integrated approach using BIRD and IReF, this study highlights key differences such as automation, harmonization and the elimination of redundant data collection efforts.
Outcome	The expected outcome of integrating BIRD and IReF includes improved consistency and accuracy in regulatory reporting, reduced operational costs for banks, and enhanced data quality for supervisory authorities. Secondary benefits encompass greater timeliness and transparency in data exchange, contributing to a more efficient regulatory ecosystem. In the long term, this integration supports the development of a single European reporting framework aligned with digitalization and standardization objectives.
Context	This intervention occurs within a regulatory environment shaped by European banking supervision and statistical reporting mandates under the ECB. The technological context is defined by ongoing digital transformation initiatives in financial reporting, while the economic context reflects the need for cost-efficient compliance. Strategically, the integration aligns with objectives for data standardization, interoperability, and the reduction of administrative burden on financial institutions.

2.1.2 Research Questions

Following the PICOC definition, the next step of the protocol was to establish the research questions addressed in this study, namely:

- RQ1: Which conceptual data models or semantic frameworks have been proposed in the literature to support integrated or standardised regulatory reporting in the banking sector?
- RQ2: How does existing research address the use of metadata repositories or standardised data dictionaries (similar approaches to BIRD) to automate or harmonise regulatory data transformations?
- RQ3: Which methods or architectural approaches have been studied for integrating heterogeneous banking datasets into unified reporting frameworks comparable to IReF?

- RQ4: What evidence exists in the literature regarding the applicability of BCBS 239 principles to architectures for integrated regulatory reporting?
- RQ5: Which outcomes related to data quality, operational efficiency or reporting accuracy are associated with the adoption of standardised regulatory reporting models?

RQ1 supports the need to identify concepts and relationships for a logical model and targets literature focusing on data dictionaries, metadata models, harmonisation and integrated reporting. RQ2 targets understanding whether BIRD similar models are academically recognised and how they are being used for data transformation and harmonisation. RQ3 allows mapping literature to the integration challenge (between BIRD and IReF), focusing on integration, data pipelines, reporting consolidation, and cross jurisdictional frameworks. RQ4 supports the previous question (RQ3), however framed as an SLR question from a more neutral perspective. RQ5 matches the PICOC's outcome, synthesising literature that measures benefits of standardisation.

2.1.3 Search Process

The search process definition involves reviewing several documents and cross-referencing key concepts such as BIRD, IReF, and BCBS 239. This includes using specific sources and criteria further detailed in this document.

While determining sources to be searched, search engines such as **EBSCOhost, Google Scholar, Semantic Scholar, and Scopus** were used for this purpose, and the choice was based on criteria of coverage, academic relevance, and diversity of sources.

EBSCOhost was included because it aggregates multiple academic databases, providing access to peer-reviewed articles, reports, and specialized publications in areas such as economics, management, and technology. The following specific databases were selected and used for the initial search: *Academic Search Complete, Business Source Complete, Communication Source, eBook Collection (EBSCOhost), eBook ITCore (EBSCOhost), eBook University Press Collection (EBSCOhost), E-Journals, Library, Information Science & Technology Abstracts* and *OpenDissertations*.

Google Scholar was chosen for its broad coverage and ease of access to academic literature, including articles, theses, and technical documents, which helps identify relevant works that may not be indexed in more restrictive databases. **Semantic Scholar** adds value through its use of artificial intelligence to improve the relevance of search results, making it particularly useful for identifying recent studies and connections between related topics. Finally, **Scopus** was selected for its reputation as one of the largest multidisciplinary databases, offering rigorous indexing criteria and bibliometric tools that support quantitative analysis of the literature.

Other databases such as **Web of Science, IEEE Xplore, ACM Digital Library, and SSRN** were excluded due to the specific nature of the research topic. **Web of Science**, although robust, significantly overlaps with Scopus, making its inclusion redundant. **IEEE Xplore** and **ACM DL** are strongly oriented toward engineering and computer science, which are not the central focus of this study, as it concentrates on regulatory reporting and framework integration in the financial sector. As for **SSRN**, despite its relevance in economics and law, its coverage is mainly focused on working papers and preliminary studies, which do not guarantee the same level of consistency and peer review as the selected databases.

2.1.4 Inclusion & Exclusion Criteria

There is very little information regarding the concepts of BIRD and IReF (a few more regarding BCBS 239). The search was not restricted (e.g., only reviews or surveys) to reach the "grey literature." Therefore, all articles and e-books published between January 1st 2010 and January 1st 2025 will be included in this search.

A reduced number of exclusion criteria will be applied for the reasons mentioned in the previous section. Namely, only the following ones were considered:

- Duplicated papers
- Papers written in a non-English or non-Portuguese language
- Papers unrelated to the current subject
- Papers not subject to peer review

2.1.5 Selection Process

While performing the primary study selection process, the results will be listed as follows:

- Number of papers per year, per source
- Number of candidate papers per year, per source
- Number of selected papers per year, per source

As part of the documented process, a list of rejected candidate papers will be kept with the specific reasons for the rejection.

2.1.6 Quality Assessment

The proposed evaluation follows a documented protocol (Kitchenham, 2004), in which each SLR will be evaluated using the York University Centre for Reviews and Dissemination (CDR) Database of Abstracts of Reviews of Effects (DARE[1]) criteria. The criteria are based on four questions:

- Q1: Are the review's inclusion and exclusion criteria described and appropriate?
- Q2: Is the literature search likely to have covered all relevant studies?
- Q3: Did the reviewers assess the quality/validity of the included studies?
- Q4: Were the basic data/studies adequately described?

The questions are to be scored as follows:

Table 3 - Scoring Table

Question	Scoring
Q1	• Y (yes), the inclusion criteria are explicitly defined in the paper • P (partly), the inclusion criteria are implicit • N (no), the inclusion criteria are not defined and cannot be readily inferred

[1] http://www.york.ac.uk/inst/crd/crddatabase.htm#DARE

Q2	• Y, the authors have either searched four or more digital libraries and included additional search strategies or identified and referenced all journals addressing the topic of interest • P, the authors have searched 3 or 4 digital libraries with no extra search strategies, or searched a defined but restricted set of journals and conference proceedings • N, the authors have searched up to 2 digital libraries or a highly restricted set of journals
Q3	• Y, the authors have explicitly defined quality criteria and extracted them from each primary study • P, the research question involves quality issues the study addresses • N, no explicit quality assessment of individual papers has been attempted
Q4	• Y, information is presented about each paper • P, only summary information is presented about individual papers • N, the results of the individual studies are not specified

The scoring procedure is Y=1, P=0.5 and N/Unknown=0.

2.1.7 Data Collection

Data extracted from each paper would be the following:

- Source (i.e., the conference or journal)
- The year the paper was published
- Classification of paper (i.e., survey, review, or e-book)
- Main software engineering topic area (i.e., data modelling, business intelligence, or requirements engineering)
- Author(s) and affiliation (organisation and country)
- Research question/issue
- Summary of paper
- Quality score for the study

2.1.8 Data Analysis

At this point, it became apparent that the literature review process would need reassessment between the researcher and the coordination. The initial search results attained according to the process detailed in the previous sections only produced a list of 3 results (after excluding one duplicate) that were insufficient to proceed with the literature review via the **SLR** protocol. This was the drive to proceed to an **MLR** (Multivocal Literature Review) and extend the initial search to **GL** (grey literature) to enrich the list of publications to work with.

2.1.9 Dissemination

Although at this point, the approach is to move forward to an MLR process (with a distinct protocol or set of guidelines), the dissemination step of the SLR protocol still makes sense. Therefore, the study's results should interest the software engineering community and relevant stakeholders in the banking industry regarding this subject (i.e., CDOs and other people responsible for management information). Therefore, the plan was to report the results on a publicly available website, in line with the currently available information regarding BIRD (restrictions may be applied to the data model definition for information safeguard purposes).

2.2 Multivocal Literature Review Methodology

Although most existing literature about MLR and GL currently states no specific guidelines for performing Multivocal Literature Reviews, there are some recommendations for doing so. The current review was performed according to a particular MLR approach (Garousi et al., 2019), and the whole process (Figure 2) was documented in the following sections.

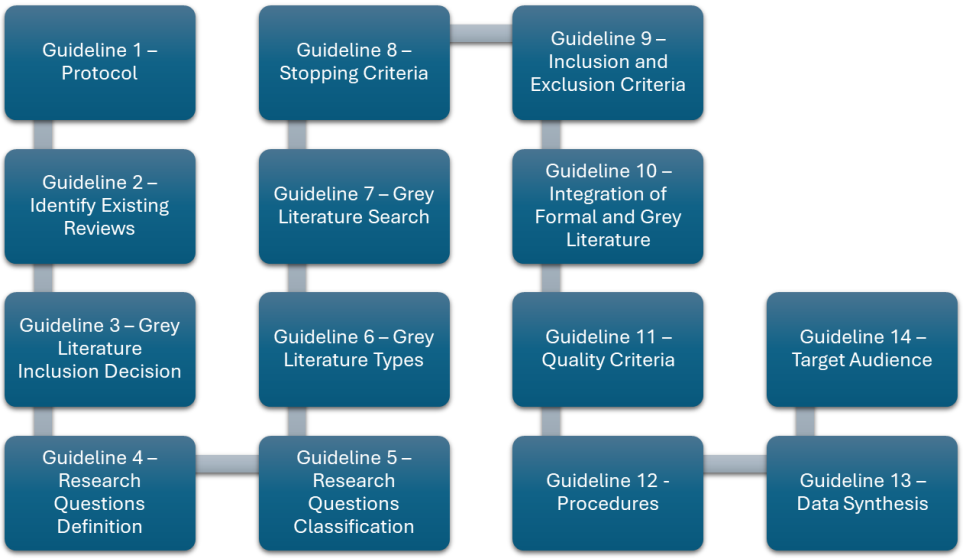


Figure 2 - MLR Guidelines

1. **Protocol:** The typical process for conducting an MLR can be used to define a protocol to determine how the review will be undertaken. One alternative way of doing so is to use the base structure of the SLR protocol and apply the MLR guidelines to it. Since the process started as an SLR and then moved towards MLR, the last option has been chosen.
2. **Identify Existing Reviews:** A key phase of the MLR process is identifying existing reviews. This will be followed by planning and executing the following steps to provide relevant outcomes for researchers.
3. **Grey Literature Inclusion Decision:** To determine the need for including grey literature (which will result in performing an MLR rather than a standard SLR), one needs to clearly define the criteria and relevant questions, resulting in a systematic process.
4. **Research Questions Definition:** The research questions should consider the research goal and the target audience (e.g., other researchers). They should be related to the review goal, address that same goal, match the needs of the defined target audience, and be as objective and measurable as possible.
5. **Research Questions Classification:** There is a need to classify the research questions under different types, although some studies may not contribute to some specific question types to be provided with an answer.

6. **Grey Literature Types:** Still, classification should also be applied to grey literature types and data sources early in the review process.
7. **Grey Literature Search:** Regarding the search process, general search engines (web), specific and specialised databases, and other websites should also be identified. Direct contact with authors can also be undertaken and documented whenever it happens.
8. **Stopping Criteria:** Some stopping criteria for the grey literature search process need to be established. Some examples of stopping criteria for searches are the theoretical saturation of a subject (whenever search results stop providing new concepts), including a determined number of search engine hits (e.g., top 8 results), or evidence exhaustion.
9. **Inclusion and Exclusion Criteria:** The inclusion and exclusion criteria for the grey literature search should also be combined with quality assessment criteria. This means the search can be narrowed or broadened according to the attained results at a specific time.
10. **Integration of Formal and Grey Literature:** The selection process of sources for both grey and formal literature should be coordinated to aggregate all results properly.
11. **Quality Criteria:** The quality criteria can be applied to several aspects of the literature, such as the author (whether he is an authority on the subject), methodology, objectivity, publication date, impact, etc. The application or adaptation of such criteria can be used to determine the quality of grey literature.
12. **Procedures:** While retrieving information, all procedures undertaken systematically should be appropriately documented for traceability purposes, including sources for extracted data. One should not forget the purpose of addressing the research questions; therefore, extracting and recording as much data as possible is very relevant for the following steps of the process (including the synthesis phase).
13. **Data Synthesis:** Data synthesis requires method selection. Some grey literature sources are more suitable for qualitative synthesis. In contrast, others can be the target of a quantitative analysis (e.g., StackOverflow databases can be the target of a quantitative analysis).

14. **Target Audience:** Usually, the target audience of an MLR paper can be researchers or practitioners (or both). The writing style in which the MLR paper is written should be adequate for that specific target audience. Whenever targeting researchers, the writing should be as transparent as possible, covering the research methodology, online repositories, and research findings relevant for future work. If practitioners are the target audience, a more plain and objective writing with less detail about the research methodology should be a better option.

2.3 Study Limitations

In an early stage, following the proposed guidelines for a systematic literature review, there was a struggle to search for relevant literature. For example, only four results were found using relevant search criteria and terms over the initially determined search database (EBSCOhost). Due to this lack of literature in the form of academic reviews or surveys, it was necessary to proceed with a different approach, focused on grey literature. This represented additional time to perform the literature search and process the retrieved information.

2.4 Literature Search

A specific approach was considered to establish the **protocol**, according to the MLR process and as per the guidelines followed, a specific approach was considered (Garousi et al., 2019). In this study, the same three phases described in the SLR process (Kitchenham, 2004) were considered, namely:

1. **Planning the review:** Conducting the review as an MLR was challenging because it required performing it within the initially proposed timeframe for the literature review, which was planned to be finished by the end of January 2025. The original plan needed some modifications, and the target date for finishing this step needed to be delayed by two months to the end of March.
2. **Conducting the review:** To conduct the review while trying to avoid further delays to the original plan's schedule, several tools were selected in order to try and accelerate

some of the tasks; some examples of the used tools in the process are Harzing's Publish or Perish (for assisting with the multiple database search) (Figure 3), Zotero (for managing the retrieved existing literature) and Elicit (to process the selected information).

3. **Reporting the review:** The fully detailed process and results of the conducted review are provided as part of this document, in the methodology section, so that one can understand the several steps that were taken, the achieved results, and (if necessary) replicate or repeat the entire process.

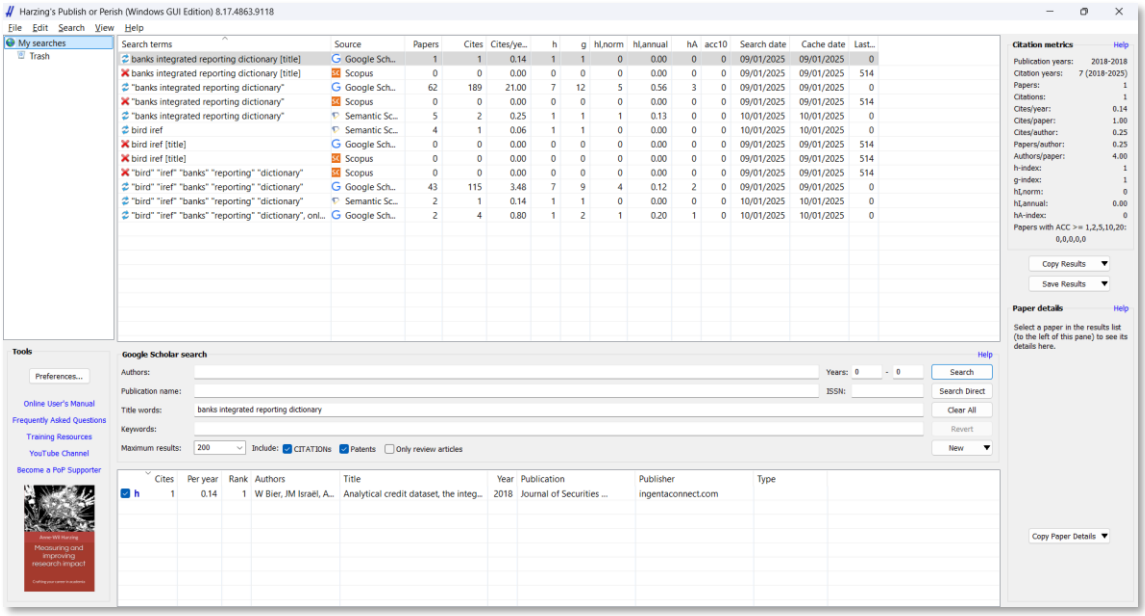


Figure 3 - List of Searches Performed via Publish or Perish

The **grey literature inclusion decision** while undertaking a Multivocal Literature Review was taken considering the following questions (and answers):

1. Is the subject "complex" and not solvable by considering only the formal literature? **Yes**
2. Does the formal literature lack volume, evidence quality, or outcome measurement consensus? **Yes**
3. Is the contextual information important to the subject under study? **Yes**
4. Is it the goal to validate or corroborate scientific outcomes with practical experiences? **No**

5. Is it the goal to challenge assumptions or falsify results from practice using academic research or vice versa? **No**
6. Would synthesising insights and evidence from the industrial and academic communities benefit one or both communities? **Yes**
7. Is there a large volume of practitioner sources indicating high practitioner interest in a topic? **No**

A positive answer to one or more of the above questions suggests the inclusion of grey literature. Hence, the decision to incorporate it within the literature review was final based on this preliminary analysis.

The **research questions** addressed by this study have already been mentioned in the previous sections.

Some different classification types (Table 4) are to be considered when establishing a **research question classification**. Following a provided example, they can be categorised in the following manner:

Table 4 - Research Questions Classification

Research Question	Category	Subcategory
Which conceptual data models or semantic frameworks have been proposed in the literature to support integrated or standardised regulatory reporting in the banking sector?	Exploratory	Classification
How does existing research address the use of metadata repositories or standardised data dictionaries (similar approaches to BIRD) to automate or harmonise regulatory data transformations?	Relationship	Relationship
Which methods or architectural approaches have been studied for integrating heterogeneous banking datasets into unified reporting frameworks comparable to IReF?	Design	Design
What evidence exists in the literature regarding the applicability of BCBS 239 principles to architectures for integrated regulatory reporting?	Exploratory	Classification
Which outcomes related to data quality, operational efficiency or reporting accuracy are associated with the adoption of standardised regulatory reporting models?	Relationship	Relationship

Regarding **grey literature types** (Figure 4), producers, and data sources, and considering the subject's scope, relevance, and sensitivity, it was considered **first-tier** only (high outlet control/high credibility, such as books, magazines, government reports, and white papers).

For reference purposes, **second-tier** grey literature refers to moderate outlet control and credibility, such as annual reports, news articles, presentations, videos, Q&A sites and wiki articles, and third-tier grey literature refers to low outlet control and credibility, such as blogs, emails and tweets (Garousi et al., 2019).

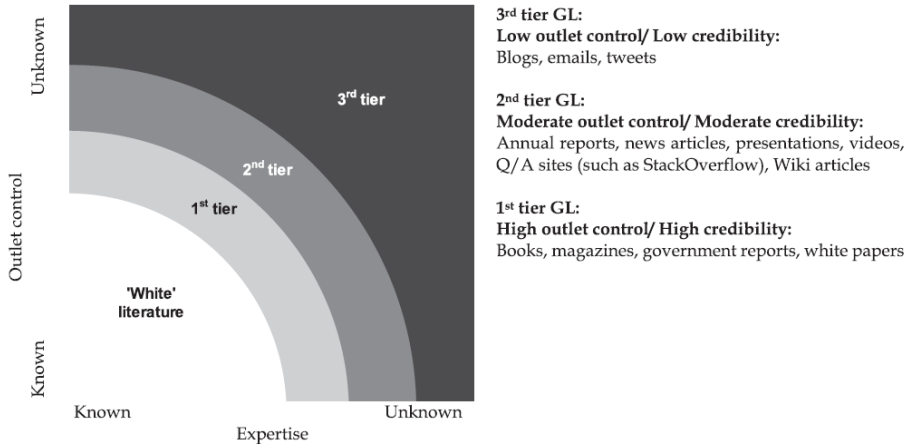


Figure 4 - Shades of Grey Literature (Garousi et al., 2019)

When considering ways to search for grey literature, search engines such as **EBSCOhost, Google Scholar, Semantic Scholar, and Scopus** were used for this purpose, and the choice was based on criteria of coverage, academic relevance, and diversity of sources.

EBSCOhost was included because it aggregates multiple academic databases, providing access to peer-reviewed articles, reports, and specialized publications in areas such as economics, management, and technology. The following specific databases were selected and used for the search:

- Academic Search Complete
- Business Source Complete
- Communication Source
- eBook Collection (EBSCOhost)

- eBook ITCore (EBSCOhost)
- eBook University Press Collection (EBSCOhost)
- E-Journals
- Library, Information Science & Technology Abstracts
- OpenDissertations

Google Scholar was chosen for its broad coverage and ease of access to academic literature, including articles, theses, and technical documents, which helps identify relevant works that may not be indexed in more restrictive databases. **Semantic Scholar** adds value through its use of artificial intelligence to improve the relevance of search results, making it particularly useful for identifying recent studies and connections between related topics. Finally, **Scopus** was selected for its reputation as one of the largest multidisciplinary databases, offering rigorous indexing criteria and bibliometric tools that support quantitative analysis of the literature.

Other databases such as **Web of Science, IEEE Xplore, ACM Digital Library, and SSRN** were excluded due to the specific nature of the research topic. **Web of Science**, although robust, significantly overlaps with Scopus, making its inclusion redundant. **IEEE Xplore** and **ACM DL** are strongly oriented toward engineering and computer science, which are not the central focus of this study, as it concentrates on regulatory reporting and framework integration in the financial sector. As for **SSRN**, despite its relevance in economics and law, its coverage is mainly focused on working papers and preliminary studies, which do not guarantee the same level of consistency and peer review as the selected databases.

EBSCOhost database usage allowed the very few results for publications retrieved while doing searches during the SLR approach to be kept. Although it was decided to document Scopus database usage in this document, it was irrelevant to the process since the performed searches did not provide any results.

For the fully detailed literature search report and used query strings (generated with Publish or Perish tool), please refer to Appendix 1 – Literature Search Reports.

For the complete list of both formal and grey literature publications included in this review, please refer to Appendix 2 – Aggregated Formal and Grey Literature List with Classification.

For 14 of the attained results (Table 5), it was a challenge to retrieve the related full documents since they were not publicly available for download on any of the platforms used during the search.

Table 5 - Literature Not Available

Year	Author	Title
2017	Scapeccia, Andrea	The Evolution in Banks' Reporting
2017	Broersen, P.; Koppen, R.	Preparing for AnaCredit: A timely start is crucial
2017	Dorval, M.	AnaCredit: Why it matters to look at other regulations in context
2018	Bier, Werner; Israël, Jean-Marc; Colangelo, Antonio; Bonci, Riccardo	Analytical credit dataset, the integrated reporting framework, and the banks' integrated reporting dictionary: Do we overshoot? Or do we undershoot?
2018	Lemmens, I.; Laar, B. van de; Saton, J.; Bulles, J.	How to fulfil regulatory requirements consistently: a semantic-based approach
2019	Hauet, Olivia	The AnaCredit regulation is in place — a reflection on data-driven reporting: What has been achieved by the banks and regulators, and what remains to be done?
2019	Shah, B.	The road to making regulation more efficient: A case study in the application of best practices and data standards in regulatory reporting
2023	Amzallag, A.	54,000 PRIIPs KIDs-how to read them (all)
2023	Hauet, O.	Reviving securitisation in the EU: A critical analysis of the reporting requirements
2024	Araujo, D.; Bruno, G.; Marcucci, J.; ...	Data science in the economy and finance: A central bank perspective
2024	Breymann, H. E.; Hauf, P.; Künzle, C.	Venturing into New Ways of Regulatory Reporting and Systemic Risk Analysis
2024	Dorrer, D.	Empirical Analysis
2024	Dorrer, D.	Analysis of the Expert Interviews
2024	Dorrer, D.	Introduction of IReF: The Development of an Integrated System for Achieving Harmonised Supervisory Reporting in the European Union

Several authors were directly contacted for 12 of the previously mentioned publications listed under ResearchGate. However, no response has been received in a timely manner (Figure 5).

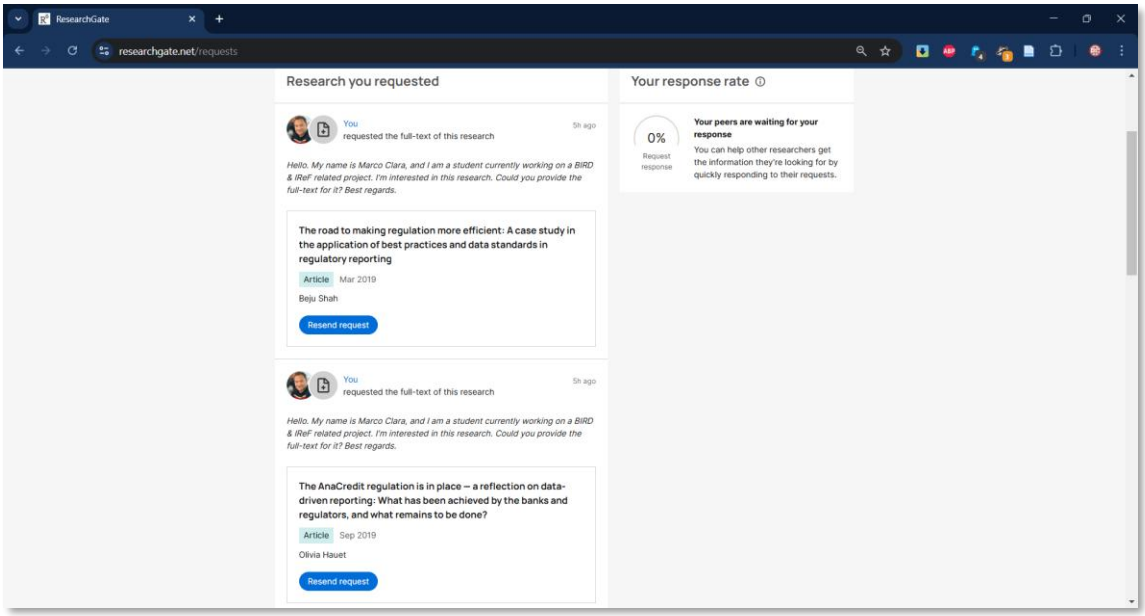


Figure 5 - List of Full-Text Research Requests via ResearchGate

The **stopping criteria** previously applied to get the final list of selected publications was a direct consequence of the decision to include the considered top search engines (Google Scholar, Semantic Scholar, and Scopus) and evidence exhaustion, since all the results have been collected and (at this point) not discarded.

While establishing **inclusion and exclusion criteria** for the search, it is important to consider that currently, there is very little information regarding the concepts of BIRD and IReF (a few more regarding BCBS 239). So, the search was not very restrictive (e.g., only reviews or surveys) to reach the "grey literature." Therefore, all articles and e-books published between 2010 and 2024 will be included in this search.

For the very same reason, there is a reduced amount of exclusion, namely, only:

- **Duplicated** (for obvious reasons): A total of 39 publications were dropped
- **Language** (non-English or non-Portuguese written publications were discarded, including German, French, Italian, and Russian – reasonably commonly applied

criteria with no significant impact considering the range of available publications for the current subject): a total of 13 publications were dropped

- **Unrelated** (publications from unrelated subjects were discarded – e.g., "a framework to estimate the risk of noise exposure from vessels for endangered cetacean species" is of null importance for the review…): a total of 22 publications were dropped

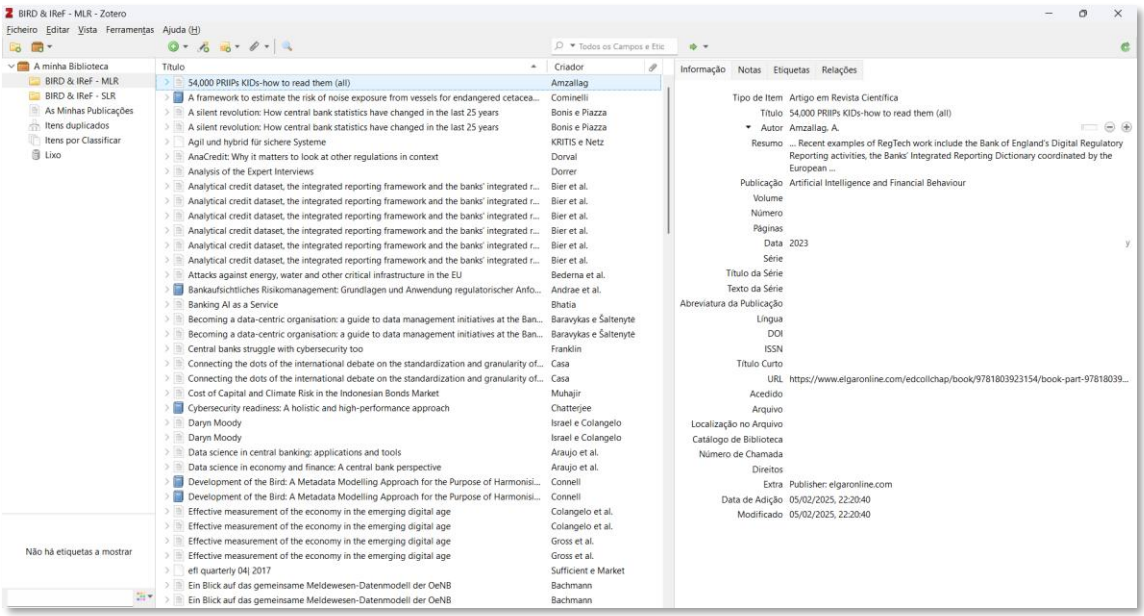


Figure 6 – View of the Initial Grey Literature Collection in Zotero

After applying the exclusion criteria, the initial list with 119 publications (Figure 6) was significantly reduced to 45 relevant publications (Figure 7).

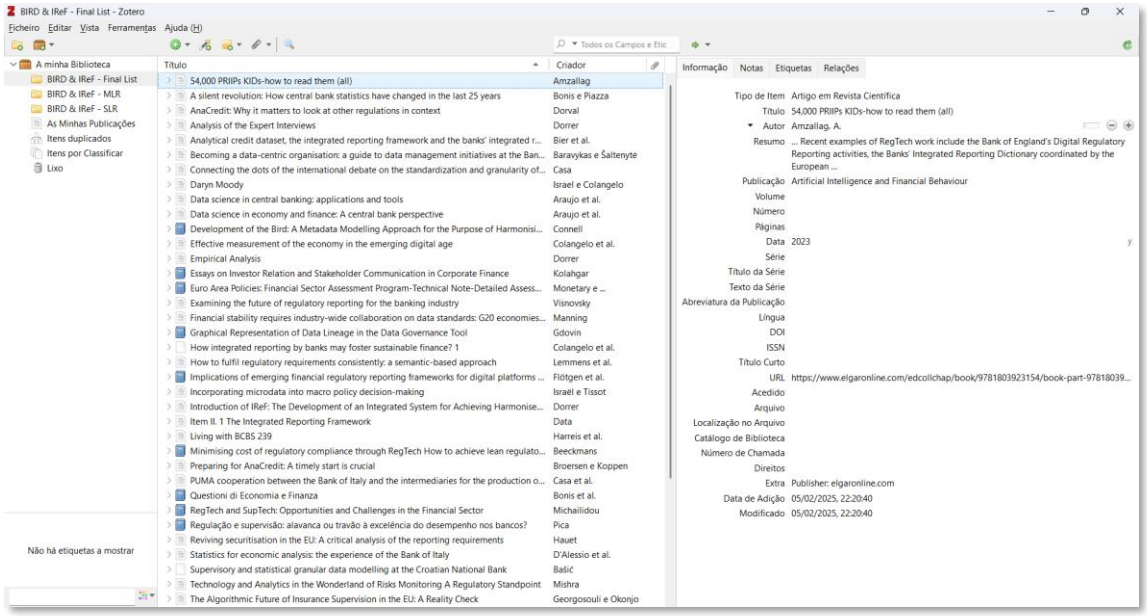


Figure 7 – View of the Aggregated Collection of Formal and Grey Literature in Zotero

Also, as previously mentioned, some documents were not publicly available for download. Therefore, for 11 publications, direct contact was established with the authors via ResearchGate, asking for full-text research. No source for the other 4 publications was available for download or information allowing direct contact with the authors.

The adoption of the MLR to incorporate the grey literature has mainly resulted from the lack of formal literature found while performing an SLR (Figure 8). The defined database was EBSCOhost, and the search only allowed for four results (from which 2 were duplicates). Since then, the search has expanded to other databases, namely Google Scholar, Semantic Scholar, and Scopus, to complement (significantly) the number of results and the amount of information to analyse. Therefore, both types of results are combined to ensure the **integration of formal and grey literature**.

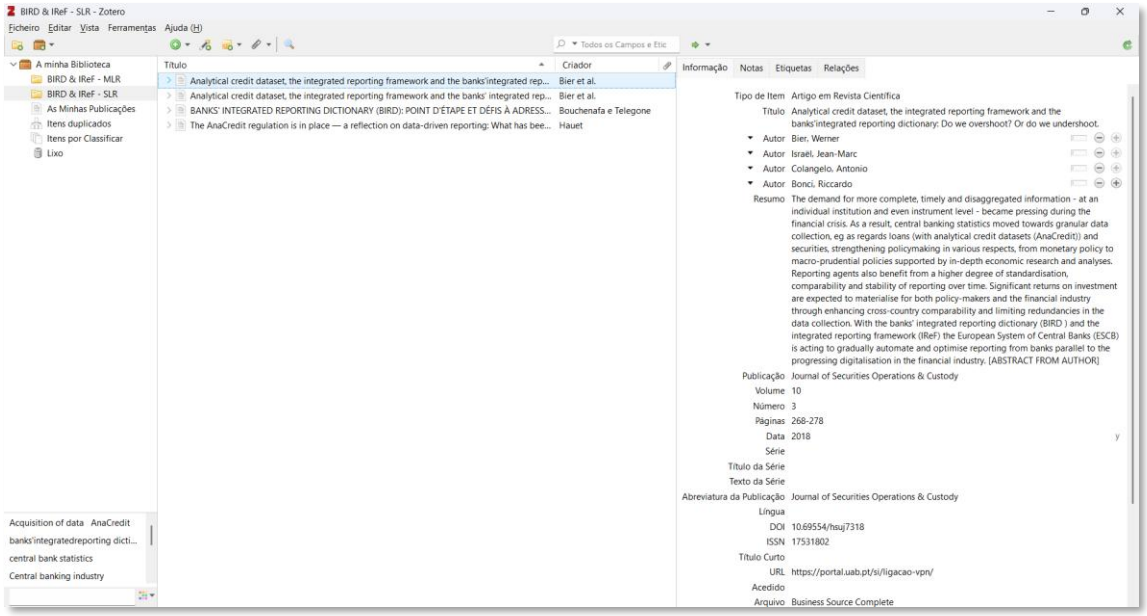


Figure 8 - View of the Formal Literature Collection in Zotero

In terms of **quality criteria** assessment, it was considered the application of the **AACODS (Authority, Accuracy, Coverage, Objectivity, Date, Significance)**. This quality assessment framework in the review of grey literature is essential for ensuring methodological rigor and credibility. Grey literature often lacks the standardized peer-review processes associated with scholarly publications, which can introduce variability in reliability and validity. Systematically evaluating sources against AACODS criteria allows to critically appraise the authority of authors and institutions, verify the accuracy and comprehensiveness of content, assess objectivity and potential bias, confirm temporal relevance, and determine the significance of contributions to the field. This structured approach (Table 6) mitigates the inherent risks of incorporating non-peer-reviewed materials, enhancing the transparency, consistency, and trustworthiness of the review process.

Table 6 - AACODS Quality Assessment of the Literature Review

Field	Rating	Assessment
Authority	**Moderate**	The review includes sources from highly authoritative institutions including the European Central Bank (ECB), Bank of Italy, Croatian National Bank, European Banking Authority and 3 others. However,

		the research notes that several studies are conceptual analyses and framework development papers rather than peer-reviewed empirical research, which may limit their academic authority.
Accuracy	**Limited Assessment Possible**	The report demonstrates internal consistency in presenting findings but notes a significant limitation: none of the reviewed studies provided concrete evidence of successful implementation in banking environments. This suggests that while the frameworks are theoretically sound, their practical accuracy remains unvalidated.
Coverage	**Moderate to Good**	The review covers a specific set of 11 studies, out of the selected literature (detailed in the following section), spanning from 2016 to 2024, addressing multiple frameworks including BIRD (6 studies), IReF (4 studies), and various other initiatives. The coverage includes diverse geographical perspectives (European focus with specific country examples) and multiple types of analysis from framework development to policy review.
Objectivity	**Good**	The research maintains objectivity by systematically categorizing studies by type, framework coverage, and research focus. It acknowledges limitations transparently, noting that the studies did not provide empirical evidence of successful implementation and that real-world testing of these frameworks through empirical studies would be necessary to validate their effectiveness.
Date	**Good**	The literature spans 8 years (2016-2024) with recent publications included, ensuring currency of the topic. The review includes the most recent developments in regulatory frameworks like IReF and BIRD implementations.
Significance	**High**	The review addresses a highly significant topic in banking regulation and supervision. The research question directly relates to critical regulatory compliance issues, particularly BCBS 239 standards, and addresses practical challenges in banking sector reporting that affect global financial stability.

Strengths: the sources demonstrate high institutional authority, ensuring credibility and trustworthiness. They offer comprehensive coverage of relevant frameworks, providing a solid foundation for analysis. Additionally, there is transparent acknowledgment of limitations, which enhances reliability and integrity. The topics addressed are current and

relevant, and the approach taken is systematic and analytical, contributing to a well-structured evaluation.

Limitations: despite these strengths, the reviewed studies lack empirical validation, which limits the robustness of findings. The focus remains predominantly theoretical and conceptual, reducing practical applicability. Geographic diversity is limited, with most sources concentrated in Europe, and several references do not indicate peer-review status, raising questions about academic rigor.

Recommendation: the literature review provides a solid foundation for understanding theoretical frameworks for unified banking reporting but would benefit from inclusion of empirical studies and broader geographic coverage to enhance its practical applicability and generalizability. Therefore, as an overall assessment of this research, it could be rated as **moderate to good**.

Due to the nature of the current project, the **target audience** is mainly practitioners. This software engineering project has a sense of urgency regarding implementing a solution for the established objective. There is a greater need to establish a practical approach to reach that solution, rather than doing much research to understand how it can be done.

2.5 Selected Literature

The following table (Table 7) shows the summarised list of selected publications, following the application of exclusion criteria and removal of literature that is currently not publicly available (nor was it possible to attain through direct contact with the authors).

Table 7 - Selected Literature

Year	Author	Title
2016	Pica, C. A.	Regulação e supervisão: alavanca ou travão à excelência do desempenho nos bancos?
2016	Visnovsky, F.	Examining the future of regulatory reporting for the banking industry

Year	Author	Title
2017	Gdovin, F.	Graphical Representation of Data Lineage in the Data Governance Tool
2017	Bašić, Ines	Supervisory and statistical granular data modelling at the Croatian National Bank
2017	Harreis, H.; Tavakoli, A.; Ho, T.; Machado, J.; ...	Living with BCBS 239
2017	Israel, J. M.; Damia, V.; Bonci, R.; Watfe, G.	The Analytical Credit Dataset. A magnifying glass for analysing credit in the euro area
2017	Manning, M.	Financial stability requires industry-wide collaboration on data standards: G20 economies promote data sharing through collaboration
2018	Data, M.	Item II. 1 The Integrated Reporting Framework
2018	Israel, M.; Colangelo, A.	Daryn Moody
2018	Monetary, International Monetary Fund; ...	Euro Area Policies: Financial Sector Assessment Program-Technical Note-Detailed Assessment of Observance of Basel Core Principles for Effective Banking …
2019	Mishra, D. R. N.	Technology and Analytics in the Wonderland of Risks Monitoring A Regulatory Standpoint
2020	Connell, M. P.	Development of the Bird: A Metadata Modelling Approach for the Purpose of Harmonising Supervisory Reporting at the European Central Bank-Directorate of General …
2020	Flötgen, R. J.; Gomm, S.; Böhm, M.; Krcmar, H.	Implications of emerging financial regulatory reporting frameworks for digital platforms boundary resources
2020	Kolahgar, S.	Essays on Investor Relations and Stakeholder Communication in Corporate Finance
2020	Michailidou, F.	RegTech and SupTech: Opportunities and Challenges in the Financial Sector
2020	Bonis, R. De; Piazza, G. M.	A silent revolution: How central bank statistics have changed in the last 25 years
2020	Drvar, M.; Turner, J.; Piechocki, M.; ...	The future of data collection and data management: Agile RegOps for digitalising the regulatory value chain
2021	Boggavarapu, S.	The Effect of Third-Party Service Providers on Information Security Breaches at Financial Institutions
2021	Baravykas, R.; Šaltenytė, U.	Becoming a data-centric organisation: a guide to data management initiatives at the Bank of Lithuania

Year	Author	Title
2021	Colangelo, A.; Gross, F.; Schuster, F.	Effective measurement of the economy in the emerging digital age
2021	Israël, J. M.; Tissot, B.	Incorporating microdata into macro policy decision-making
2022	Iwanicz-Drozdowska, M.; Malinowska-Misiąg, E.; ...	The Role of Crises in Shaping Financial Systems: From the Global Financial Crisis to COVID-19
2022	Colangelo, A.; Israël, Jean-Marc; Ahsbahs, Catherine; Continanza, D.; Damia, Violetta; Devillers, Corinne; Gross, Francis; Hiebert, P.; Saponara, Angelo; Schuster, Florian	How integrated reporting by banks may foster sustainable finance?
2022	Casa, M.; Palmieri, L. Graziani; Mellone, L.; ...	The integrated approach adopted by Bank of Italy in the collection and production of credit and financial data
2022	Casa, M.; Sabatini, R.; Carnevali, M.; ...	PUMA cooperation between the Bank of Italy and the intermediaries for the production of statistical, supervisory and resolution reporting
2022	D'Alessio, G.; Bonis, R. De; Infante, L.; ...	Statistics for economic analysis: the experience of the Bank of Italy
2022	Georgosouli, A.; Okonjo, J.	The Algorithmic Future of Insurance Supervision in the EU: A Reality Check
2023	Araujo, D.; Bruno, G.; Marcucci, J.; Schmidt, R.; Tissot, B.	Data science in central banking: applications and tools
2023	Casa, M.	Connecting the dots of the international debate on the standardisation and granularity of regulatory data
2024	Poloni, P.	The evolution of the supervisory reporting framework for the EU banking sector

Chapter 3 - Methodology

Action Research is a participatory, iterative, and problem-solving research methodology widely used in education, social sciences, management, IT, and organizational transformation studies. It is especially suitable for those who want to investigate and improve real-world processes while generating academically rigorous insights.

Figure 9 - Action Research Methodology

For this dissertation, the chosen model consists in four specific stages that will be further detailed in the following sections, namely:

1. Planning
2. Action
3. Observation
4. Reflection

3.1 Planning

Initially and considering the iterative nature of the selected research methodology, a timeline was proposed to have multiple sprints with a month duration each (see Figure 10). These sprints would follow an initial task with greater focus on the first task of the literature review (right after problem identification). At this point, there was a need to address the research questions from a different perspective.

Figure 10 - Plan Baseline

Data synthesis was initially performed focusing on each research question individually, but after some iterations, the following reflection was required:

- **Single Research Question**: This approach works well if the literature review focuses on a specific aspect or hypothesis. It is straightforward and keeps the review tightly organised around that single question. This can make it easier to dive deep and provide detailed analysis.
- **Aggregated Research Questions**: If the research encompasses broader themes or multiple aspects of a topic, one might consider aggregating several related research questions. This allows for a more comprehensive review, covering different angles and perspectives. It is important that the questions are interconnected and that the review is structured in a way that flows logically.

So, there was a decision to work on aggregated multiple research questions, namely:

1. *What are the concepts and relationships between them, which should be part of a logical data model, which houses all relevant information from different banking*

institutions, with a view to unified reporting in a centralised framework for banking sector supervisors?

2. *From the concepts and modelled relationships, is it possible to specify requirements that allow the implementation of a logical model that supports the conceptual model specified in the BIRD and its integration into the IReF?*
3. *Can this implementation be carried out by the BCBS 239 standard (Basel Committee on Banking Supervision), which is titled "Principles for effective risk data aggregation and risk reporting"?*

Considering the aggregated research question option and the above three individual research questions, the following research was elaborated using Elicit. This indicates that logical data models for unified banking reporting require specific components and technical infrastructure aligned with BCBS 239 but lack empirical validation of successful implementation.

This report analysed **11 studies** examining concepts and relationships needed for a logical data model to support unified banking sector reporting. The reviewed papers, primarily theoretical analyses and framework proposals focused on initiatives like the Banks' Integrated Reporting Dictionary (BIRD) and the Integrated Reporting Framework (IReF).

The reviewed studies identified specific components for logical data modelling, including granular data collection systems, standardised data dictionaries, cross-domain reporting integration, data governance frameworks, and standardised data exchange protocols. Most proposed frameworks incorporated four layers: **data storage**, **semantic interpretation**, **data processing** and **reporting functions**.

The reviewed papers described technical requirements regarding BIRD implementation and IReF integration, including robust IT infrastructure, comprehensive data governance structures, and quality assurance mechanisms. While these align with some BCBS 239 principles for risk data aggregation, the studies did not provide empirical evidence of successful implementation.

The papers highlighted several implementation challenges: complexity in harmonising diverse reporting requirements, substantial technology investment needs, and difficulties in standardising data across jurisdictions.

The authors noted that real-world testing of these frameworks through empirical studies would be necessary to validate their effectiveness. **None of the reviewed studies provided concrete evidence of successful implementation in banking environments**.

3.2 Action

3.2.1 Paper Search

To answer the research question, "What are the concepts and relationships between them, which should be part of a logical data model, which houses all relevant information from different banking institutions, with a view to unified reporting in a centralised framework for banking sector supervisors? From the concepts and modelled relationships, is it possible to specify requirements that allow the implementation of a logical model that supports the conceptual model specified in the BIRD and its integration into the IReF? Can this implementation be carried out per the BCBS 239 standard (Basel Committee on Banking Supervision) under "Principles for effective risk data aggregation and risk reporting"? We looked over only the papers uploaded.

Twenty-nine papers were uploaded to Elicit for screening.

3.2.2 Screening

We screened papers that met these criteria:

- **Banking Data Models and Regulatory Focus**: Does the study address data models, frameworks, or architecture specifically in the context of banking supervisory or regulatory reporting?
- **Standardised Reporting Frameworks: Does the research examine standardised** reporting frameworks (such as BIRD, IReF, or BCBS 239) or their implementation in banking supervision?

- **Multi-Institution Scope**: Does the study address data integration or standardisation across multiple banking institutions (rather than single-institution reporting)?
- **Banking Sector Specificity**: Does the study focus on the banking sector (rather than other financial institutions or general data modelling)?
- **Regulatory/Supervisory Focus**: Does the study address regulatory/supervisory reporting aspects (rather than solely operational or retail banking processes)?
- **Framework Level Analysis**: Does the research address conceptual frameworks or models (rather than focusing solely on specific technology implementations)?
- **Implementation Context**: If the study discusses implementation, does it include an analysis of the broader regulatory reporting framework or requirements (rather than purely technical details)?

We considered all screening questions and critically judged whether to screen in each paper.

3.3 Observation

3.3.1 Data Extraction

We asked a LLM (large language model) to extract each data column from each paper, giving the model the instructions below for each column.

Research Approach and Methodology:

Describe the primary methodological approach used in the study:

- Identify the type of study (e.g., conceptual analysis, framework development, policy review)
- Specify the primary research methods used (e.g., document analysis, ecosystem modelling, institutional analysis)
- Note any specific analytical techniques or frameworks employed (e.g., e3-value method, institutional perspective analysis)

Instructions:

- Look in the methods, introduction, or methodology sections
- If multiple approaches are used, list them in order of prominence
- Be precise about the specific analytical approach
- If the approach is not explicitly stated, infer from the study's methodology.

Conceptual Framework and Key Concepts:

Extract the primary conceptual frameworks or models discussed in the study:

- List the main conceptual models or frameworks proposed
- Identify key concepts related to banking supervisory reporting
- Note any specific references to existing frameworks (e.g., BIRD, IReF, BCBS 239)

Instructions:

- Search throughout the complete text, particularly the introduction and discussion sections.
- If multiple frameworks are mentioned, list them in order of importance
- Provide brief descriptions of how these frameworks are conceptualised
- If frameworks are partially described, note this explicitly

Regulatory Reporting Requirements and Challenges:

Identify and describe:

- Specific regulatory reporting requirements discussed
- Challenges in current supervisory reporting frameworks
- Proposed solutions or improvements to reporting mechanisms

Instructions:

- Focus on explicit discussions of reporting requirements
- Extract specific challenges mentioned by the authors
- Note any proposed improvements or recommendations
- Look in the results, discussion, and conclusion sections

- If multiple challenges are identified, list them in order of significance

Data Governance and Sharing Mechanisms:

Extract information about:

- Data governance approaches proposed
- Mechanisms for data sharing between institutions
- Principles for data collection and management

Instructions:

- Search for explicit discussions of data governance
- Note any specific governance models or principles
- Identify any proposed centralisation or standardisation approaches
- Look in the methodology, results, and discussion sections
- If multiple approaches are mentioned, list them comprehensively

Compliance with Regulatory Standards:

Identify:

- Specific regulatory standards discussed (e.g., BCBS 239)
- Compliance requirements or challenges
- Proposed strategies for meeting regulatory standards

Instructions:

- Focus on explicit discussions of regulatory compliance
- Extract specific standard requirements mentioned
- Note any implementation challenges
- Look at the discussion and conclusion sections
- Be precise about the specific standards referenced

3.3.2 Characteristics of Included Studies

The following table summarises the included studies (Table 8) that contributed to the presented findings in this paper. The research focus, framework coverage, study type, and key findings were considered for each of the listed publications.

Table 8 - Characteristics of Included Studies

Study	Research Focus	Framework Coverage	Study Type	Key Findings
Regulation (EU) 2016/867 of the ECB, 2016	AnaCredit requirements and implementation	AnaCredit, Banks' Integrated Reporting Dictionary (BIRD), European Reporting Framework (ERF)	Framework development	AnaCredit was established as a shared multipurpose database for granular credit data collection.
Bašić, 2017	Granular data system for regulatory reporting	Croatian National Bank (CNB) Banks' Integrated Reporting Dictionary	Framework development	CNB BIRD presented as a standardised framework for data extraction and transformation
Casa et al., PUMA Cooperation Between the Bank of Italy	Procedura Unificata di Matrice Aziendale (PUMA) initiative for data collection and management	PUMA	Conceptual analysis and policy review	PUMA highlighted as an integrated approach to data collection and management
Casa et al., The Integrated Approach by Bank of Italy	Integrated approach to data management	Bank of Italy's integrated approach	Framework development	Proposed integrated approach to data governance with a

Study	Research Focus	Framework Coverage	Study Type	Key Findings
				single statistical dictionary
Casa, 2024	Regulatory data standardisation	Integrated Reporting Framework (IReF), BIRD	Conceptual analysis and policy review	Emphasised the need for a standard data dictionary and an integrated reporting system
Colangelo et al., How Integrated Reporting by Banks	Granular data in central banking	IReF, BIRD	Conceptual analysis and framework development	Discussed IReF and BIRD as key initiatives for standardising and integrating reporting requirements
Colangelo et al., 2021	Addressing the Data Problem in Banking	IReF, BIRD	Conceptual analysis and framework development	Proposed radical standardisation across countries with central banks as catalysts
Drvar et al., The Future of Data Collection	RegOps approach to regulatory reporting	RegOps	Conceptual analysis with framework development	Introduced RegOps as an agile approach to regulatory reporting
Connell, 2020	BIRD project and Single Data Dictionary (SDD)	BIRD, SDD	Framework development	Focused on the BIRD project and SDD for harmonising data models across banks
Jago Flötgen et al., Digital Platforms Boundary Resources	Digital platforms for financial reporting	BIRD, IReF	Framework development with ecosystem modelling	Proposed a standardised reporting framework as a public good

Study	Research Focus	Framework Coverage	Study Type	Key Findings
Poloni, 2024	EU supervisory reporting framework	European Banking Authority (EBA) reporting framework, Basel Committee on Banking Supervision 239 (BCBS 239)	Policy review and institutional analysis	Discussed the development of a standard data dictionary and joint governance under the Joint Board of Reporting Committees (JBRC)

3.3.3 Framework Coverage

- BIRD: Most frequently covered (6 out of 11 studies)
- IReF: Second most common (4 studies)
- Other frameworks: Each covered in 1 study (AnaCredit, ERF, CNB BIRD, PUMA, Bank of Italy's integrated approach, RegOps, SDD, EBA reporting framework, BCBS 239)
- Multiple frameworks: Covered in some studies

3.3.4 Study Types

- Framework development: Most common (4 studies)
- Conceptual analysis with policy review or framework development: 4 studies (2 each)
- Other types: One study each for conceptual analysis with framework development, framework development with ecosystem modelling, and policy review and institutional analysis

3.3.5 Research Focus

- Diverse range: Included regulatory reporting, data standardisation, integrated approaches to data management, and digital platforms for financial reporting
- No single dominant focus across all studies

3.4 Reflection

3.4.1 Data Analysis

In this section two distinct types of analysis are provided, namely a thematic analysis and a bibliometric analysis, as described in the following subsections. Additionally, as part of the study an extraction matrix is also provided for further detail (please refer to Appendix 3 – Extraction Matrix).

3.4.1.1 Thematic Analysis

For a thematical analysis of the data model components and their relationships (Table 9), a list of concept categories and related elements was systematized, including integration points with each framework and mapping with specific regulatory requirements.

Table 9 - Data Model Components and Relationships

Concept Category	Related Elements	Framework Integration Points	Regulatory Requirements
Granular Data	Credit data, counterparty data, and instrument-level data	AnaCredit, BIRD, IReF	BCBS 239 principles on data aggregation
Data Dictionary	Metadata, data definitions, transformation rules	BIRD, SDD, CNB BIRD	Standardisation of reporting requirements
Integrated Reporting	Statistical, prudential, and resolution reporting	IReF, PUMA	Harmonisation of reporting across domains
Data Governance	Data quality, data lineage, data ownership	BCBS 239, RegOps	Effective risk data aggregation and reporting

Standardised Interfaces	Application Programming Interfaces (APIs), data exchange protocols	Digital platforms, RegOps	Facilitating data sharing and interoperability
Semantic Integration	Common definitions, taxonomies	Single statistical dictionary, BIRD	Ensuring consistency across reporting frameworks
Data Aggregation	Consolidation rules, aggregation methods	AnaCredit, IReF	Meeting supervisory and statistical reporting needs
Data Validation	Quality checks, consistency rules	PUMA, CNB BIRD	Ensuring data accuracy and reliability
Regulatory Compliance	Reporting requirements, disclosure standards	BCBS 239, EBA framework	Meeting supervisory expectations and legal obligations
Data Sharing	Confidentiality rules, access controls	European data space concept	Balancing transparency with data protection
Metadata Managemen t	Data lineage, versioning, and change management	SDD, BIRD	Supporting data governance and traceability

3.4.1.2 Bibliometric Analysis

Based on our analysis of the included studies, we identified 23 distinct types of related elements across the concept categories. Data lineage was mentioned twice, while all others were mentioned once each. We found 12 different framework integration points. The most frequently mentioned were **BIRD** (4 mentions), **IReF** (3 mentions), **AnaCredit**, **SDD**, **CNB BIRD**, **PUMA**, **BCBS 239** and **RegOps** (2 mentions each) and the remaining four frameworks were mentioned once each. We found three types of reporting mentioned: **statistical**, **prudential**, and **resolution reporting**. The concept categories covered a wide range of data management aspects, including **data types** (e.g., credit data, counterparty data), **data governance** (e.g., data quality, data ownership), **technical aspects** (e.g., APIs, data exchange protocols), **regulatory aspects** (e.g., reporting requirements, disclosure standards) and **metadata management** (e.g., versioning, change management). Framework integration points included both specific initiatives (e.g., AnaCredit, BIRD) and broader concepts (e.g., Digital platforms, European data space concept).

According to the research we have performed the following bibliometric analysis:

Temporal Distribution: publication years range from 2016 to 2024, having the following distribution: 2016 (1 publication), 2017 (1 publication), 2019-2020 (1 publication), 2021 (1 publication), 2024 (2 publications) and other papers' years are not explicitly stated in the report.

Research Methodology Distribution: framework development (4 studies), conceptual analysis with policy review (2 studies), conceptual analysis with framework development (2 studies), framework development with ecosystem modelling (1 study), policy review and institutional analysis (1 study). Figure 11 shows the describe framework coverage distribution.

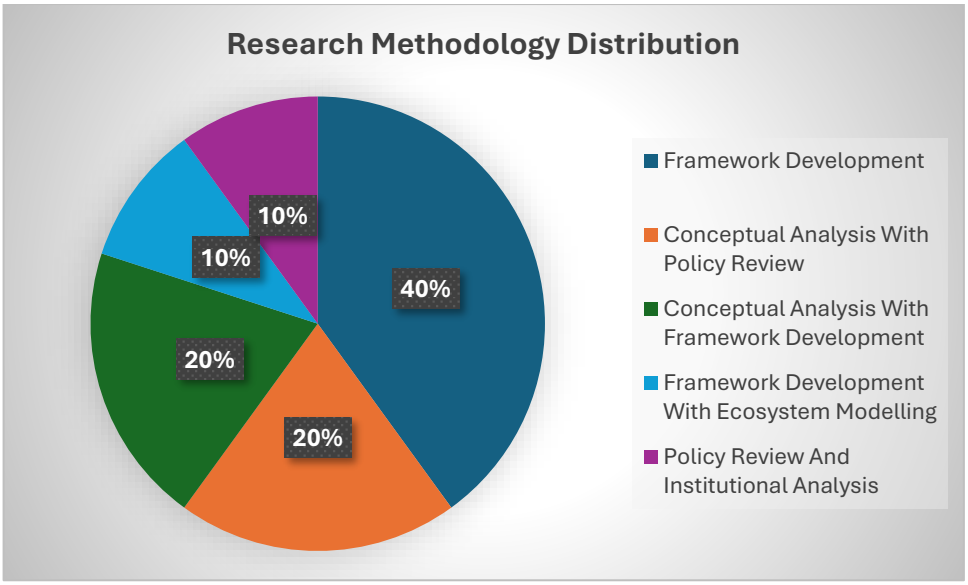


Figure 11 - Research Methodology Distribution

Framework Coverage Analysis: BIRD (6 studies, 54.5% of total), IReF (4 studies, 36.4% of total), single framework studies (AnaCredit, ERF, CNB BIRD, PUMA, Bank of Italy's integrated approach, RegOps, SDD, EBA reporting framework, BCBS 239). Figure 12 shows the describe framework coverage distribution.

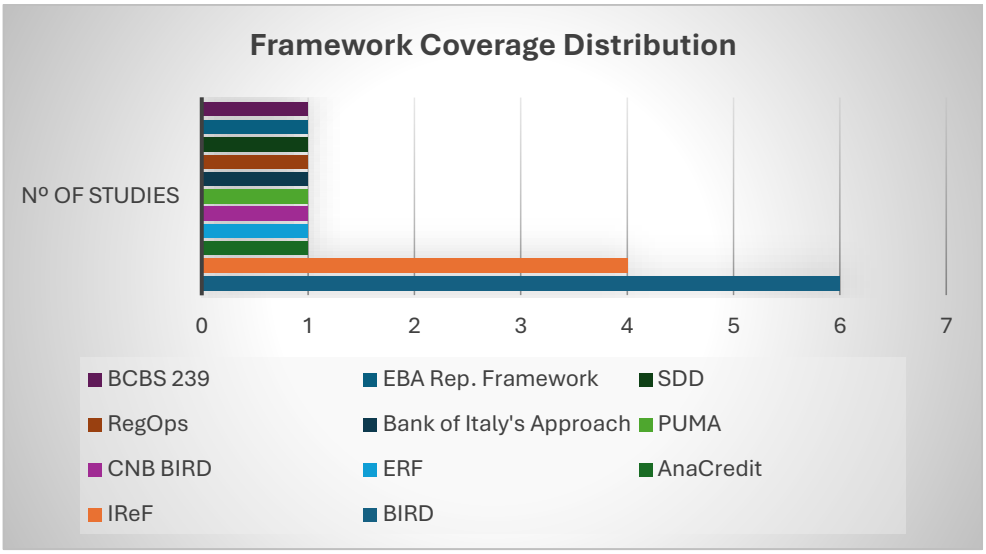


Figure 12 - Framework Coverage Distribution

Study Type Distribution (primary categories): framework development (36.4%, 4 studies), conceptual analysis combinations (36.4%, 4 studies), other specialised approaches (27.2%, 3 studies)

Institutional Representation: Connell, 2020; Bank of Italy Casa et al., "PUMA Cooperation Between the Bank of Italy" & 1 others; Croatian National Bank Bašić, 2017; other regulatory and academic institutions.

Research Focus Areas (major themes identified): regulatory reporting requirements, data standardisation, integrated approaches to data management, digital platforms for financial reporting, framework development and implementation, data governance and quality assurance.

Document Types: research reports, institutional papers, policy documents, framework proposals, conceptual analyses.

This bibliometric analysis reveals a diverse range of studies focusing strongly on framework development and conceptual analysis, particularly centred around BIRD and IReF implementations. The temporal distribution shows ongoing research activity from 2016 to 2024, with recent publications indicating the continued relevance of the topic.

3.4.1.3 Synthesis Table

A synthesis table is relevant to better understand the research evidence and gaps regarding each of the research questions. Below (Table 10) there is a representation for RQ1 to RQ3 containing those evidence and gaps.

Table 10 - Synthesis Table (RQ1-RQ3)

Research Questions	Evidence	Gaps
RQ1: Concepts and Relationships for Logical Data Model	• Strong evidence for core components: Granular data systems, standardized data dictionaries, cross-domain integration, data governance frameworks • Architectural consensus: Four-layer model (data, semantic, processing, reporting) proposed across multiple studies • 11 concept categories identified: Including granular data, data dictionary, integrated reporting, data governance, standardized interfaces	• No empirical validation: "None of the reviewed studies provided concrete evidence of successful implementation in banking environments" • Limited relationship modelling: Conceptual relationships identified but not formally modelled • Missing quantitative specifications: No detailed technical specifications for logical model implementation
RQ2: BIRD Implementation and IReF Integration Requirements	• Technical requirements specified: Robust IT infrastructure, comprehensive data governance, quality assurance mechanisms • Framework coverage: BIRD covered in 6/11 studies, IReF in 4/11 studies • Integration approach defined: "Integration with IReF seen as key step towards comprehensive architecture"	• No implementation roadmap: Studies lack concrete implementation timelines or phases • Missing cost-benefit analysis: No evidence of resource requirements or ROI assessments • Limited interoperability testing: No evidence of successful BIRD-IReF integration pilots
RQ3: BCBS 239 Compliance Alignment	• Partial alignment confirmed: Several studies show frameworks align with BCBS 239 principles for data aggregation, quality, timeliness • Governance alignment: Proposed frameworks include "robust data governance structures" consistent with BCBS 239 • Quality assurance mechanisms: Data validation, lineage tracking, and quality metrics align with BCBS 239 requirements	• No formal compliance assessment: "While these align with some BCBS 239 principles...the studies did not provide empirical evidence" • Missing compliance metrics: No quantitative measures of BCBS 239 adherence • Limited regulatory validation: No evidence of supervisor approval or compliance certification

Regarding the research **overall evidence strength assessment**, there is a **strong theoretical foundation** (considered 11 studies and consistent frameworks), **moderate regulatory compliance** (conceptual alignment although with no formal validation) and **weak practical implementation** (absence of empirical evidence).

There are **critical research gaps** identified, namely the need of **empirical validation studies** for all three RQ's, missing **quantitative implementation metrics**, the need for **real-world pilot testing** and the need for **formal regulatory compliance assessment**.

In summary, in terms of strength of evidence metrics, current evidence base appears too preliminary for robust quality-weighted synthesis, as the research concludes that while the conceptual foundation is well-established, real-world testing of these frameworks through empirical studies would be necessary to validate their effectiveness.

Basic Vote Counting (unweighted):

- Total studies: 11
- BIRD support: 6/11 studies (55%)
- IReF support: 4/11 studies (36%)
- Other frameworks: 1/11 studies each (9% each)

Study Type Distribution:

- Framework development: 4 studies
- Conceptual analysis with policy review: 2 studies
- Conceptual analysis with framework development: 2 studies
- Other types: 3 studies (1 each of different types)

To establish **quality weighting metrics** for this research, it would be necessary:

1. **Develop quality criteria** (such as methodological rigor, empirical evidence vs. theoretical, peer review status and implementation evidence)
2. **Assign quality weights** (e.g., 0.5 for theoretical, 1.0 for empirical with implementation data)

3. **Calculate weighted support** (using a formula such as Σ(quality weight × direction of evidence) / Σ(quality weights))

However, the research points out that the studies were primarily theoretical analyses and framework proposals and did not provide empirical evidence of successful implementation. This presents challenges and **limitations for quality weighting** because:

1. **No formal quality assessment**: the research doesn't include risk of bias evaluation, study quality scores, or methodological rigor assessments
2. **Limited empirical evidence**: most studies are theoretical rather than empirical
3. **No effect sizes**: the studies don't provide quantitative outcomes that could be weighted

3.4.2 Framework Integration Architecture

Key Components

- Integrated Reporting Framework (IReF):
    - Overarching framework
    - Integrates existing reporting lines
    - Aims for a single, standardised system across the euro area
    - Consolidates statistical reporting requirements
    - Reduces redundancy and ensures consistency

- Banks' Integrated Reporting Dictionary (BIRD):
    - Complements IReF
    - Provides a harmonised data model and standardised transformation rules
    - Helps banks organise internal data and meet reporting requirements
    - Integration with IReF seen as key step towards comprehensive architecture

3.4.3 Common Goal

Create a unified and efficient reporting system for the banking sector.

Several studies propose a layered architecture approach:

1. **Data Layer**: This foundational layer involves collecting and storing granular data, as exemplified by initiatives like AnaCredit. The emphasis is on capturing detailed, loan-by-loan or transaction-level data to support various reporting and analytical needs.
2. **Semantic Layer**: This layer includes the standard data dictionary and metadata repository, which provide standardised definitions and transformation rules. The Single Data Dictionary (SDD) and various BIRD implementations (e.g., CNB BIRD) are key components of this layer.
3. **Processing Layer**: This layer includes the standardised processing logic and aggregation rules. The PUMA initiative is an example of a system that provides generalised software components for data processing.
4. **Reporting Layer**: This top layer generates various regulatory reports and disclosures, leveraging the standardised data and processes from the lower layers.

The RegOps approach introduces a more agile and technology-driven architecture. It proposes using APIs and big data-enabled platforms to facilitate data exchange and analysis. This approach emphasises the need for real-time or near-real-time data processing and reporting capabilities.

A European data space is also proposed as an overarching architectural element, providing a framework for secure data sharing and reuse among various stakeholders. This would require robust data governance mechanisms and standardised interfaces to ensure interoperability and data protection.

Several studies emphasise the importance of a centralised approach to data management and reporting, with central banks playing a key role in defining standards and coordinating integration efforts. However, this centralisation is balanced with the need for flexibility to accommodate differences in national reporting requirements and banking systems.

Integrating statistical and supervisory reporting frameworks is a recurring theme, with proposals for aligning data models and processes across these domains. This integration aims to reduce the reporting burden on banks while providing a more comprehensive view of the financial system for regulators.

While the studies present various architectural approaches, they converge on the need for a highly standardised, integrated, and flexible framework that can accommodate evolving regulatory requirements and technological advancements. The challenge lies in implementing such an architecture across diverse banking systems and regulatory landscapes.

Regulatory Compliance and Implementation

The implementation of integrated reporting frameworks and their alignment with regulatory standards, particularly BCBS 239, is a key focus across the studies. While not all papers explicitly discuss BCBS 239, many proposed frameworks and approaches align with its principles for effective risk data aggregation and reporting.

Poloni's research most directly addresses compliance with BCBS 239 principles (Poloni, 2024), emphasising the need for consistent data collection across banks to enable meaningful comparisons. The paper suggests that applying BCBS 239 principles should be communicated to avoid inconsistent behaviour among stakeholders.

Several key aspects of regulatory compliance and implementation emerge from the studies:

1. **Data Granularity**: The shift towards granular data collection, as exemplified by AnaCredit, aligns with BCBS 239's emphasis on detailed risk reporting. This granularity enables more accurate risk assessment and reporting.
2. **Data Quality and Consistency**: Many studies highlight the importance of data quality assurance mechanisms, a key principle of BCBS 239. For instance, the CNB Banks' Integrated Reporting Dictionary includes validation rules to ensure data quality and compliance with reporting standards.
3. **Timeliness of Reporting**: The RegOps approach proposed by Drvar et al. emphasises real-time or near-real-time data processing, which aligns with BCBS 239's requirement for timely risk reporting.
4. **Data Aggregation Capabilities**: The integrated reporting frameworks discussed in several studies aim to improve banks' ability to aggregate risk data, a key requirement of BCBS 239. The IReF, for instance, is designed to consolidate existing reporting lines, potentially enhancing data aggregation capabilities.

5. **Comprehensive Risk Data**: The emphasis on integrating various reporting domains (statistical, prudential, resolution) in frameworks like IReF and PUMA aligns with BCBS 239's principle of comprehensiveness in risk reporting.
6. **Data Governance**: Several studies discuss the importance of robust data governance structures, a fundamental principle of BCBS 239. For instance, the integrated approach proposed by the Bank of Italy emphasises company-wide governance of data management processes.
7. **Adaptability**: The proposed frameworks, such as RegOps and digital platform approaches, emphasise adaptability to changing regulatory requirements, which aligns with BCBS 239's principle of adaptability in risk data aggregation and reporting practices.

Implementation challenges highlighted across the studies include:

- Complexity of integrating diverse reporting requirements and data models
- Need for significant investment in IT infrastructure and data management systems
- Challenges in standardising data across different jurisdictions and banking systems
- Balancing the granularity of data collection with the reporting burden on banks
- Ensuring data privacy and security while facilitating necessary data sharing

Proposed strategies for implementation and ensuring compliance include:

- Phased implementation of integrated reporting frameworks
- Collaborative development of common data dictionaries and reporting standards
- Leveraging technology for automated data validation and reporting
- Establishing clear governance structures for data management and reporting processes
- Ongoing dialogue between regulators and banks to refine reporting requirements and processes

While the studies present various approaches to implementing integrated reporting frameworks, they consistently emphasise the need for standardisation, automation, and robust data governance. These elements align well with the principles of BCBS 239, suggesting that the proposed frameworks could support compliance with these regulatory standards.

3.4.4 Implementation Considerations

3.4.4.1 Technical Requirements

1. **Data Integration Infrastructure**
   - Robust infrastructure for integrating data from various sources and systems
   - Ability to handle structured and unstructured data
   - Integration of legacy systems with new data management platforms
2. **Granular Data Management**
   - Capability to collect, store, and process highly granular data
   - Significant storage capacity and efficient data retrieval mechanisms
3. **Real-time or Near-real-time Processing**
   - Emphasised in the RegOps approach
   - Support for timely reporting and decision-making
4. **Standardised Interfaces and APIs**
   - Facilitate data exchange between different systems and institutions
   - Support secure and efficient data transmission
5. **Data Transformation and Aggregation Capabilities**
   - Apply complex transformation rules and aggregation methods
   - Convert granular data into required reporting formats
6. **Metadata Management**
   - Maintain data lineage
   - Support data governance
   - Ensure consistency in data definitions across systems
7. **Data Quality and Validation Tools**
   - Automated data quality checks

- Validation mechanisms for accuracy and reliability

8. **Scalability and Flexibility**
    - Accommodate changes in reporting needs and data volumes
    - Adapt to evolving regulatory requirements
9. **Security and Access Control**
    - Robust security measures
    - Granular access control mechanisms
10. **Analytical Capabilities**
    - Support advanced analytics
    - Derive insights for regulatory compliance and business decision-making

3.4.4.2 Governance Structure

The studies emphasise the importance of robust governance structures to support the implementation and ongoing management of unified reporting frameworks. Key aspects of governance highlighted across the studies include:

1. **Centralised Oversight:** Several studies propose a centralised governance model, with central banks or regulatory authorities playing a key role in defining standards and coordinating integration efforts.
2. **Collaborative Development**: This paper emphasises the importance of collaboration between regulatory authorities, banks, and other stakeholders in developing and refining reporting standards.
3. **Clear Data Ownership**: Establishing clear data ownership and responsibilities across domains (e.g., statistical, prudential, resolution) is crucial for effective data management.
4. **Data Quality Management**: Governance structures should include transparent processes and responsibilities for ensuring data quality, including validation, error handling, and continuous improvement.
5. **Change Management**: Given the evolving nature of regulatory requirements, governance structures should include robust change management processes to handle updates to reporting standards and data models.

6. **Cross-functional Coordination**: Governance models should facilitate coordination across different functional areas within banks (e.g., risk management, finance, IT) to ensure a holistic approach to data management and reporting.
7. **Regulatory Alignment**: Governance structures should ensure ongoing alignment with regulatory requirements, including mechanisms for interpreting and implementing new regulations.
8. **Data Ethics and Privacy**: Governance frameworks should address ethical considerations in data use and ensure compliance with data protection regulations.
9. **Transparency and Accountability**: Clear mechanisms for ensuring transparency in data management processes and accountability for data quality and reporting accuracy should be established.
10. **Continuous Improvement**: Governance structures should support continuous evaluation and improvement of data management and reporting processes.

These governance considerations highlight the need for a comprehensive and well-structured approach to managing the complex ecosystem of data, processes, and stakeholders involved in banking sector reporting.

3.4.4.3 Data Quality Assurance

As highlighted in several studies, ensuring data quality is critical to implementing unified reporting frameworks. Key considerations for data quality assurance include:

1. **Standardised Data Definitions:** Using common data dictionaries and standardised metadata, as proposed in initiatives like BIRD and SDD, is crucial for ensuring data interpretation and reporting consistency.
2. **Automated Validation Rules**: The emphasis is on implementing automated data validation checks at various stages of data collection and processing. This includes technical and business rule validations (e.g., format checks).

3. **Data Lineage and Traceability**: Maintaining clear data lineage from source systems to final reports is crucial for auditing and ensuring data quality. This is particularly emphasised in the context of BCBS 239 compliance.
4. **Reconciliation Processes**: Regular data reconciliation across different reporting domains (e.g., statistical, prudential) is proposed to ensure consistency and identify discrepancies.
5. **Quality Metrics and Reporting**: To monitor and improve data quality over time, it is suggested that clear data quality metrics be established and that regular reporting on these metrics be conducted.
6. **Feedback Loops**: Implement feedback mechanisms to capture and address data quality issues identified by end-users (e.g., regulators, analysts) is proposed.
7. **Data Governance Role**: Several studies emphasise the importance of clear roles and responsibilities for data quality within the overall data governance framework.
8. **Continuous Monitoring**: Implementing continuous monitoring processes to identify and address data quality issues near-real-time, particularly emphasised in the RegOps approach.
9. **Training and Awareness**: Ensuring all data management and reporting stakeholders know data quality requirements and best practices.
10. **External Audits**: Regular external audits of data quality processes and outcomes are suggested to provide independent assurance.

These data quality assurance measures are essential for ensuring the reliability and usefulness of the data collected and reported through unified reporting frameworks. They also align closely with the principles of BCBS 239, particularly in terms of ensuring the accuracy and integrity of risk data aggregation and reporting.

Chapter 4 - Implementation

A proof of concept for BIRD & IReF Integration was implemented. The resulting artifact (please refer to Appendix 5 – Proof of Concept Source Code for access to full project and available scripts) is a small example for the automation of regulatory reporting via harmonised data models. This PoC demonstrates how the proposed framework can be easily and successfully utilized by any bank to leverage the BIRD data model and automatically prepare and submit IReF-compliant reports, reducing manual effort, improving data quality, and ensuring regulatory alignment.

As previously stated, BIRD (Banks' Integrated Reporting Dictionary) provides a standardised data model for banks to map internal data to regulatory requirements, and IReF (Integrated Reporting Framework) is the ECB's initiative to unify statistical, supervisory, and resolution reporting into a single framework.

This proof of concept will simulate the process by which a bank can use BIRD to automatically generate IReF reports from its internal systems.

4.1 Proposed Architecture for the PoC

First, a PoC architecture needs to be established to proceed. For this purpose, the following architectural layers were considered (Figure 13):

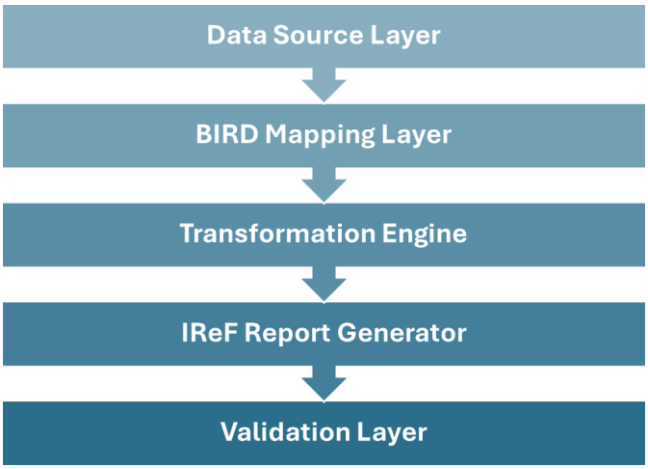


Figure 13 - PoC Architecture

Data Source Layer: Simulated bank data (e.g. loans, deposits, counterparties) in a relational database or flat files.

BIRD Mapping Layer: Use BIRD templates to map internal data to ECB-required attributes.

Transformation Engine: ETL process (e.g. Python or Talend) to convert raw data into BIRD-compliant format.

IReF Report Generator: Generate mock IReF reports (e.g. AnaCredit, SHS, etc.) using ECB's draft specifications.

Validation Layer: Include basic validation rules to check data completeness and consistency.

4.2 Research & Planning

Following the previously presented research methodology, this step aims to develop a proof of concept that demonstrates how a bank's internal data can be transformed and mapped using the BIRD data model to generate reports aligned with one of the modules under the IReF framework. For simulation purposes AnaCredit was the selected module. Other modules could be considered (e.g. SHS - Securities Holdings Statistics), however, due to the status of the available frameworks, it makes sense to use AnaCredit due to current development maturity and the nature and relevance of the information within (risk and credit data).

AnaCredit's conceptual data model includes **13 entities** that represent the core data structures used to capture granular credit and credit risk information for regulatory reporting, namely:

- **Reporting Institution**: identifies the credit institution responsible for submitting AnaCredit data to the national central bank
- **Counterparty**: represents the legal entity (e.g. corporation or other borrower) that is party to the credit agreement
- **Contract**: defines the legal agreement between the reporting institution and the counterparty, encompassing one or more instruments
- **Instrument**: refers to the financial product (e.g. loan, credit line) issued under a contract and subject to reporting

- **Loan**: a subtype of instrument that involves disbursement of funds with repayment obligations
- **Guarantee/Collateral**: assets pledged by the counterparty to secure the instrument and mitigate credit risk
- **Guarantor**: a third party that provides a guarantee for the counterparty's obligations under the contract
- **Credit Protection**: mechanisms (e.g. guarantees, insurance) that reduce the credit risk associated with the instrument
- **Interest Rate**: specifies the terms and conditions of interest applied to the instrument
- **Payment Schedule**: details the timing and structure of repayments associated with the instrument
- **Default**: indicates whether the counterparty has failed to meet contractual obligations
- **Risk**: captures the credit risk assessment and exposure metrics related to the instrument or counterparty
- **Accounting**: provides financial reporting data such as valuation, impairment, and provisioning.

These entities are interrelated and form the foundation of the AnaCredit data model (AnaCredit Reporting Manual - Part II - Datasets and Data Attributes, 2017). They are used to describe the relationships between banks, borrowers, financial instruments, and risk mitigation mechanisms.

The BIRD portal was used to get an initial overview of the specific AnaCredit framework, which contains detailed information on individual bank loans in the euro area, harmonised across all Member States. The data model for AnaCredit's collection showed 9 cubes (Figure 14):

- **Accounting**: development of the instrument in accordance with the relevant accounting standard
- **Counterparty Default:** identify the default status of the counterparty in accordance with Article 178 of the CRR

- **Counterparty Instrument**: counterparties that take on certain roles, such as the creditor to the instrument, the servicer to the instrument or the debtor to the instrument, related to an instrument reported in the instrument dataset
- **Counterparty Risk**: assessment of the counterparty's credit risk in accordance with the CRR
- **Financial**: instrument's financial development. The data reported are the actual data reflecting the situation of the instrument at the reporting reference date
- **Instrument**: instrument according with the definitions in the AnaCredit Regulation
- **Instrument Protection**: relationship between an instrument and a protection item that is used to secure this instrument
- **Joint liabilities**: instruments with a plurality of debtors. The joint liabilities dataset is reported only for instruments which have a multiple debtor
- **Protection Received**: characteristics of any protection (both funded and unfunded) that serves to secure the repayment of instruments reported in the instrument dataset.

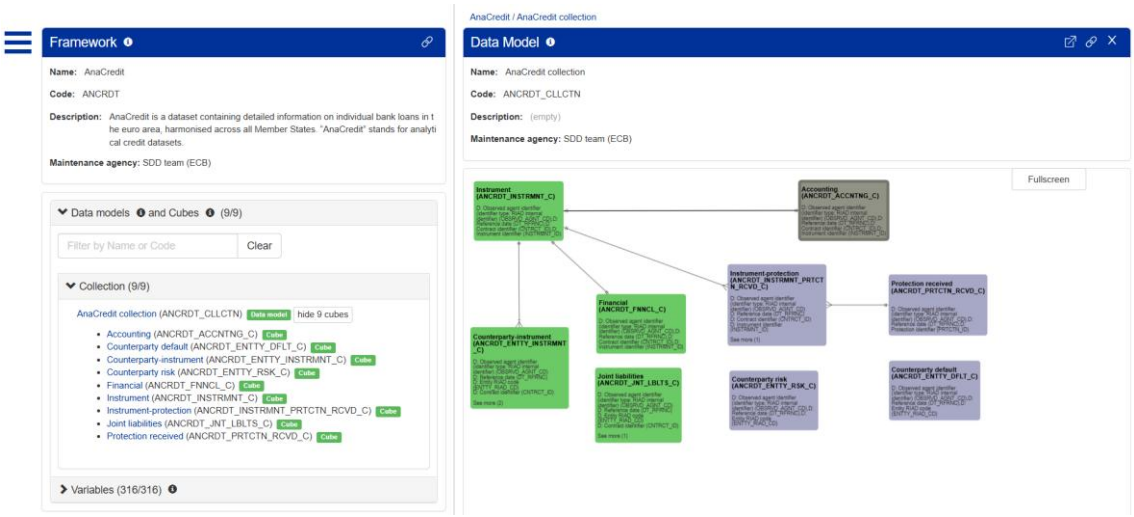


Figure 14 - BIRD's AnaCredit Framework

Also, a total of 316 variables can be found for this framework, related to AnaCredit collection and the previously identified cubes. This shows the complexity of this model and therefore the current PoC will naturally focus on a subset of the

framework. Once established AnaCredit as the target framework within BIRD for the PoC, the actions that fall within the **scope** of the PoC shall be the following (Figure 15):

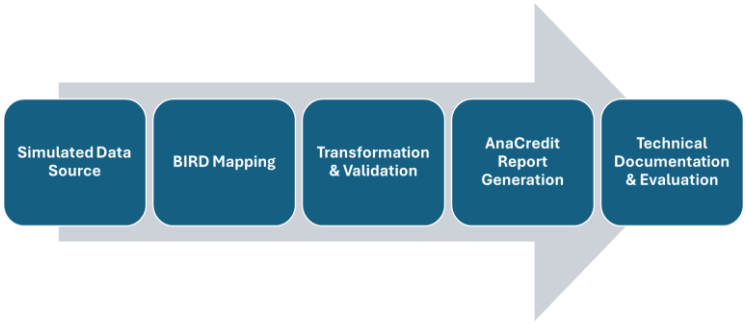


Figure 15 - Scope of the PoC

1. **Simulated Data Source**
   - Create a database with sample credit contracts, counterparties, guarantees, and other relevant entities that reflect real banking data
2. **BIRD Mapping**
   - Apply the BIRD dictionary to map internal data to AnaCredit's required attributes
3. **Transformation & Validation**
   - Build ETL scripts (using Python and SQL) to convert raw data into BIRD-compliant format
   - Implement basic validation rules (e.g. mandatory fields, format checks, cross-entity consistency)
4. **AnaCredit Report Generation**
   - Produce a mock output file that simulates an AnaCredit report in accordance with IReF technical specifications
5. **Technical Documentation & Evaluation**
   - Document the process, technical decisions, and limitations
   - Evaluate the scalability and real-world applicability of the solution

The **success criteria** of this proof of concept will be measured according to the conditions described in Table 11.

Table 11 - Success Criteria

Criteria	Description
Data Coverage	All mandatory AnaCredit attributes for specified entities are correctly mapped and populated
Technical Compliance	The output file follows the structure and format defined by IReF for AnaCredit
Automation	The transformation and report generation processes are automated via scripts

Data Quality Validation	Validation rules are implemented to detect errors or inconsistencies before report generation
Complete Documentation	The project includes clear documentation of architecture, mappings, rules, and limitations
Academic Presentation	The final presentation clearly explains technical choices and demonstrates understanding of BIRD and IReF concepts

4.2 Data Modelling

The goal of this step is to build a **simulated AnaCredit data source**, creating a realistic, structured dataset that mimics the internal data of a bank, aligned with the BIRD Input Layer (IL) and AnaCredit reporting requirements under IReF.

The first part of this step is to **select an adequate database engine** that supports the data model creation process. For this, three of the most well-known database engines were assessed according to their features (Table 12).

Table 12 - Database Engines

Engine	Features
PostgreSQL	Best for relational modelling, supports complex joins, constraints, and foreign keys. Ideal for simulating AnaCredit's entity relationships
SQLite	Lightweight, file-based, portable. Good for prototyping and academic use
MySQL	Easy to set up, widely supported, good for integration with web apps or dashboards

The choice came down to **PostgreSQL** due to its ability to simulate complex relationships and validations, as well as compatibility with tools like Python.

For the **initial definition of the data model**, some core entities were selected for simulation purposes, based on BIRD IL and AnaCredit ERM. According to the BIRD Metadata Manual[2], the AnaCredit module includes key entities described in Table 13.

[2] https://www.ecb.europa.eu/stats/ecb_statistics/co-operation_and_standards/reporting/html/BIRD_metadata_manual_latestversion.pdf

Table 13 - Data Model Entities

Entity	Description
Counterparty	Borrowers (individuals or legal entities), including identifiers like LEI, sector, residency
Instrument	Credit instruments (loans, credit lines), with attributes like type, currency, maturity
Contract	Links between counterparties and instruments, including contract dates and conditions
Guarantee	Collateral or guarantees associated with instruments
InterestRateTerms	Interest rate structure, fixed/variable, spread, reference rate
AccountingData	Outstanding amounts, arrears, provisions, write-offs

The selected entities are considered representative of the conceptual data model for AnaCredit (**6 out of 13 entities**, representing an approximate coverage of **47%**) and, therefore, adequate for the PoC.

Each of the tables to be created will include **primary keys, foreign keys** to link entities, **mandatory fields** based on AnaCredit specs, and **validation constraints** (e.g. NOT NULL, UNIQUE, CHECK).

For the **creation of sample datasets** (e.g. loan contracts, customer data), some data generation techniques were considered. Initially, manual entries seemed like an option, while using Excel or CSV files to manually define a significant set of rows for each table and then importing them to PostgreSQL. Ultimately, the solution was to utilize a tool for generating realistic synthetic data, specifically Python and the Faker library.

For **validation** purposes, foreign key constraints were used to enforce relationships, and CHECK constraints were applied to fields such as currency codes and maturity dates. Also, for the same purpose, it was considered to include sample anomalies (e.g. missing fields, invalid dates) to test validation logic in a later stage.

Regarding the generated data, it was also necessary to consider exporting it using specific **export formats**. Therefore, once populated, the data is to be exported in comma-separated values files (CSV for transformation and mapping), SQL dumps (for portability), or JSON (for integrating with APIs or web apps).

So, now, the relevant technical tasks were designing the PostgreSQL script containing the database schema (refer to *db_tables.sql*) and writing the Python scripts

(refer to: *01-anacredit_model.py* for database and table creation and *02-anacredit_to_sql.py* for database table population). The proposed data model is represented in Figure 16.

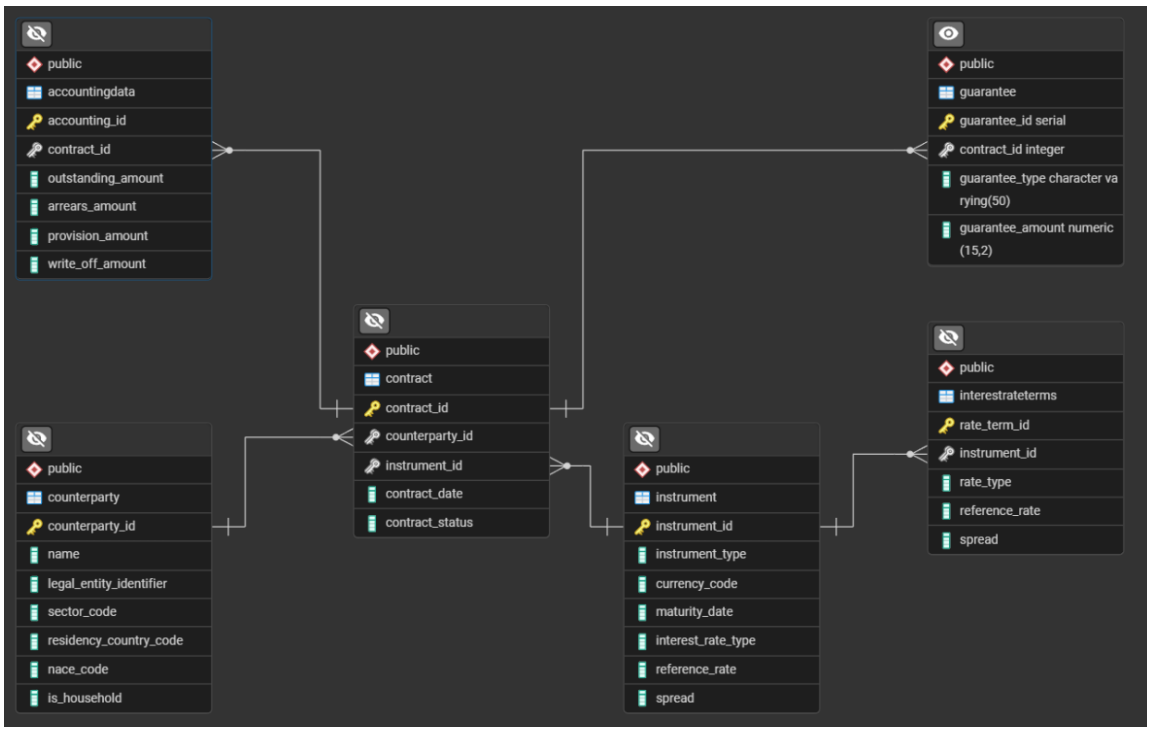


Figure 16 - Entity Relationship Diagram for PoC

4.3 Mapping & Transformation

It is now important to proceed by mapping AnaCredit data to BIRD attributes, which involves aligning the internal database fields (e.g. contract_id, maturity_date, counterparty_id) with the Input Layer (IL) defined in the BIRD model. This ensures that data can be transformed into IReF-compliant reports using standardized transformation rules. The mapping process relies on the following steps:

4.3.1 Understanding the BIRD Input Layer (IL) Structure

The BIRD IL is composed of **entities** such as:

- Counterparty
- Instrument
- Contract

- Guarantee
- Interest Rate Terms
- Accounting Data

Each entity has a set of **attributes** defined in BIRD metadata repository (SDD), such as:

- Counterparty.LEI
- Instrument.MaturityDate
- Contract.ContractDate
- Guarantee.GuaranteeAmount
- AccountingData.OutstandingAmount

These correspond directly to the fields in the PostgreSQL schema.

4.3.2 Using the Mapping Package from the SDD

The ECB's Single Data Dictionary (SDD) includes a **Mapping Package** that defines:

- **Variable_Mapping_Item**: links internal variables to BIRD attributes
- **Member_Mapping_Item**: maps code lists (e.g. currency codes, sector codes)

Example: field **instrument_type** is mapped to **Instrument.Type** (BIRD IL)

4.3.3 Applying Transformation Rules

The transformation process follows three phases (Figure 17).

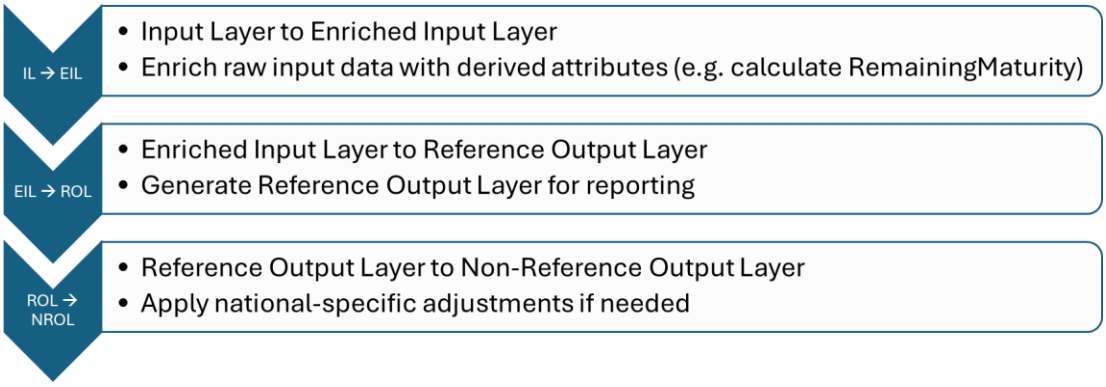


Figure 17 - Transformation Phase for PoC

Transformation rules can be defined in Python or SQL, e.g.:

```
-- Calculate Remaining Maturity

SELECT instrumentid,
maturitydate - CURRENTDATE AS remainingmaturity
FROM Instrument;
```

4.3.4 Validating Against BIRD Metadata

Using the metadata definitions (referring to BIRD Metadata Manual[3] for attribute definitions and validations rules) to:

- Check data types (e.g. DATE, NUMERIC)
- Validate mandatory fields
- Apply code lists (e.g. ISO country codes, NACE codes)

4.3.5 Document the Mapping

For documenting all significant attribute mapping the creation of a mapping table (Table 14) is required. This will allow to establish correspondence between internal fields and BIRD attributes, as well as determine necessary transformation rules to be applied.

Table 14 - Mapping Table Example

Internal Field	BIRD Attribute	Transformation Rule	Notes
contract_date	Contract.ContractDate	Direct	Mandatory
sector_code	Counterparty.SectorCode	Code list mapping	Use ECB sector codes
spread	Instrument.Spread	Direct	Optional

[3] https://www.ecb.europa.eu/stats/ecb_statistics/co-operation_and_standards/reporting/html/BIRD_metadata_manual_latestversion.pdf

It is also necessary to document the process of mapping the current schema to BIRD as per the following example (Table 15). This will allow to identify tables and specific fields within the schema and map them to the corresponding IL attribute.

Table 15 - Mapping Schema Example

PostgreSQL Table	Field	BIRD IL Attribute
Counterparty	legal_entity_identifier	Counterparty.LEI
Instrument	maturity_date	Instrument.MaturityDate
Contract	contract_date	Contract.ContractDate
Guarantee	guarantee_amount	Protection.Amount
AccountingData	outstanding_amount	Accounting.AmountOutstanding

Based on the BIRD metadata modelling approach and the AnaCredit data model, a **full mapping table** that aligns each field in your PostgreSQL schema with the corresponding **BIRD Input Layer (IL) attribute** can be created. This mapping – perceivable from Table 16 to Table 21 – is essential for transforming synthetic data into a format suitable for IReF reporting.

Table 16 - Counterparty Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
Counterparty	counterparty_id	Counterparty.Identifier	Internal ID; mapped to BIRD's unique identifier
Counterparty	name	Counterparty.Name	Direct mapping
Counterparty	legal_entity_identifier	Counterparty.LEI	Use LEI if available; fallback to national ID
Counterparty	sector_code	Counterparty.SectorCode	Map to ECB sector codes (e.g. NFC, HH)
Counterparty	residency_country_code	Counterparty.ResidencyCountry	ISO 3166-1 alpha-2 format
Counterparty	nace_code	Counterparty.NACECode	Use NACE Rev.2 classification
Counterparty	is_household	Counterparty.Type	Boolean → 'Household' or 'Legal Entity'

Table 17 - Instrument Data Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
Instrument	instrument_id	Instrument.Identifier	Internal ID
Instrument	instrument_type	Instrument.Type	Map to BIRD instrument types (e.g., Loan, CreditLine)
Instrument	currency_code	Instrument.Currency	ISO 4217 format
Instrument	maturity_date	Instrument.MaturityDate	Direct mapping
Instrument	interest_rate_type	Instrument.InterestRateType	'Fixed' or 'Variable'
Instrument	reference_rate	Instrument.ReferenceRate	e.g. EURIBOR, LIBOR
Instrument	spread	Instrument.Spread	Numeric value

Table 18 - Contract Data Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
Contract	contract_id	Contract.Identifier	Internal ID
Contract	counterparty_id	Contract.CounterpartyIdentifier	Foreign Key to Counterparty
Contract	instrument_id	Contract.InstrumentIdentifier	Foreign Key to Instrument
Contract	contract_date	Contract.ContractDate	Direct mapping
Contract	contract_status	Contract.Status	Map to BIRD status codes (e.g., Active, Terminated)

Table 19 - Guarantee Data Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
Guarantee	guarantee_id	Protection.Identifier	Internal ID
Guarantee	contract_id	Protection.ContractIdentifier	Foreign Key to Contract
Guarantee	guarantee_type	Protection.Type	Map to BIRD protection types
Guarantee	guarantee_amount	Protection.Amount	Direct mapping

Table 20 - Interest Rate Terms Data Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
InterestRateTerms	rate_term_id	InterestRateTerms.Identifier	Internal ID
InterestRateTerms	instrument_id	InterestRateTerms.InstrumentIdentifier	Foreign Key to Instrument
InterestRateTerms	rate_type	InterestRateTerms.Type	'Fixed' or 'Variable'
InterestRateTerms	reference_rate	InterestRateTerms.ReferenceRate	Same as Instrument
InterestRateTerms	spread	InterestRateTerms.Spread	Numeric value

Table 21 - Accounting Data Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
AccountingData	accounting_id	Accounting.Identifier	Internal ID
AccountingData	contract_id	Accounting.ContractIdentifier	Foreign Key to Contract
AccountingData	outstanding_amount	Accounting.AmountOutstanding	Direct mapping
AccountingData	arrears_amount	Accounting.AmountInArrears	Direct mapping
AccountingData	provision_amount	Accounting.ProvisionAmount	Direct mapping
AccountingData	write_off_amount	Accounting.WriteOffAmount	Direct mapping

EIL Layer

Considering the previous mapping, the following step is to prepare a **complete SQL transformation script** that creates **EIL views** for all six entities in the AnaCredit simulation schema, followed by a **Python script** to export each EIL view to a **BIRD compliant CSV file**. So, at this point there is a database named ***poc_anacredit_sim***, already populated with simulated data for the IL tables (Counterparty, Instrument, etc.), that we will use to simulate the **IL → EIL transformation** phase of the BIRD model.

So, now, the relevant technical tasks are designing the EIL views SQL script (refer to: *eil_views.sql*), creating the EIL views (refer to: *03-create_eil_views.py*) and exporting the EIL views data to CSV (refer to: *04-export_eil_views_csv.py*).

As an output of the export process six CSV files will be generated. Each of these CSV files will contain a header with the exported attributes and extracted data from the corresponding views, as per the following example (Figure 18). These files simulate the **BIRD compliant EIL output**, ready for transformation into the **Reference Output Layer (ROL)** or for mock submission for this proof of concept:

- EIL_Counterparty.csv
- EIL_Instrument.csv
- EIL_Contract.csv
- EIL_Guarantee.csv
- EIL_InterestRateTerms.csv
- EIL_AccountingData.csv

Figure 18 - Sample CSV File Generated from EIL View

ROL Layer

The **Reference Output Layer (ROL)** is the structured output derived from the EIL, formatted according to the reporting templates and technical specifications defined by the ECB. It represents the **final reporting ready dataset** before submission to regulators under IReF. ROL creation involves:

- **Filtering** EIL data based on reporting criteria
- **Aggregating** or **reshaping** data into ECB-defined templates
- **Applying rendering rules** (e.g. formatting, naming, code lists)

Each AnaCredit report (e.g. Credit Instrument, Counterparty, Protection) has a **Rendering Table** defined in BIRD. These specify required fields, data types, code lists and

format rules. For exporting purposes, one can reuse the previous Python script, replacing view names:

```
views = [
    "ROL_Counterparty",
    "ROL_Instrument",
    "ROL_Contract",
    "ROL_Protection",
    "ROL_InterestRateTerms",
    "ROL_Accounting"
]
```

So, now, the relevant technical tasks are designing the ROL views SQL script (refer to: *rol_views.sql*), creating ROL views from EIL data (refer to: *05-create_rol_views.py*) and exporting ROL views data to CSV files (refer to: *06-export_rol_views_csv.py*).

As an output of the export process six CSV files will be generated. Each of these CSV files will contain a header with the exported attributes and extracted data from the corresponding views, as per the following example (Figure 19). These files simulate the **BIRD compliant ROL output**, ready for mock submission for this proof of concept:

- ROL_Counterparty.csv
- ROL_Instrument.csv
- ROL_Contract.csv
- ROL_Guarantee.csv
- ROL_InterestRateTerms.csv
- ROL_AccountingData.csv

metadata > ROL_Counterparty.csv > data

```
1  counterpartyidentifier,lei,sectorcode,residencycountry,nacecode
2  1,wyIwNSxCUJybNILwNPdN,NFC,GA,Fm.Vq
3  2,rdMnVlKRgXzQAbYhFAvd,GOV,KH,AQ.IL
4  3,laUnbkDZPKcxdZqdsTct,NFC,KR,HB.Xt
5  4,RrlHYQLUvbwzkDuAYULL,GOV,SL,Ft.Wl
6  5,VWVFUxqsKpxogRFvVepl,GOV,MY,dC.BU
7  6,ZaMTiZlYvJFayIhzuEFj,NFC,GQ,lF.Kv
8  9,naPzwCRNljXBVeHneHZq,NFC,TO,HY.lM
9  10,gSeyXsnTtuKdqAvTBlmh,GOV,OM,iG.fZ
10 11,fQMMTBtbKORQXPvtyTuz,GOV,GD,xh.vl
```

Figure 19 - Sample CSV File Generated from ROL View

4.4 Report Generation

Moving towards report generation phase, it is important to validate ROL CSV files against ECB Rules. For doing so, it is necessary to use BIRD transformation and validation rules. The ECB's BIRD Metadata Manual includes both transformation and validation rules in a unified format. These rules define:

- **Mandatory fields** per reporting template
- **Accepted code lists** (e.g. ISO country codes, NACE, currency)
- **Logical constraints** (e.g. maturity date must be after contract date)
- **Cross-entity consistency** (e.g. instrument currency must match accounting currency)

The latest validation rules can be downloaded from the BIRD Portal[4] under the "Transformation Rules" section. Validation checklist for CSV files is described in Table 22.

Table 22 - Validation Rules

Validation Rule	How to Check
Field presence	Ensure all mandatory columns are present in each CSV
Data types	Dates in YYYY-MM-DD, numeric fields with "." as decimal separator
Code lists	Validate fields like SectorCode, Currency, NACECode against ECB lists
Referential integrity	Ensure foreign keys (e.g. ContractIdentifier) exist in related files
Logical rules	e.g. RemainingMaturity > 0, AmountOutstanding ≥ AmountInArrears
Uniqueness	Primary keys (e.g. ContractIdentifier) must be unique per file

Initially, manual validation can be considered for smaller datasets, using tools such as Excel or LibreOffice to filter for missing values, apply conditional formatting, and use formulas to check logic. However, in this case, the option was to use automated validation using Python and Pandas to script validation checks.

Some national central banks (e.g. Banco de España, De Nederlandsche Bank) offer **validation engines** or **Excel templates** for AnaCredit submissions. These can be used to simulate real-world validation.

[4] https://bird.ecb.europa.eu/

In this case, a **complete Python validation script** for all six **ROL CSV files** was generated from AnaCredit simulation. It is designed to be modular and extensible, allowing for easy adaptation **to future BIRD/IReF modules**.

This script checks for:

- Mandatory fields
- Data types
- Code list compliance
- Logical consistency
- Uniqueness of identifiers

So, now, the relevant technical task is to perform validation of ROL CSV files (refer to: *07a-validation_rol_csv_files.py*).

Following first script execution, a problem was identified. Validation returned error "*AssertionError: Outstanding < Arrears*.", which means that in at least one row of ROL_Accounting.csv, the value in the AmountOutstanding column is **less than** the value in AmountInArrears (overdue), which violates a basic **AnaCredit validation rule**: "**Outstanding amount must always be greater than or equal to the amount in arrears".** To fix this error, the following steps were taken:

4.4.1 Identify and Fix Problematic Rows

Following the assertion error traceback, data was corrected manually, and a new CSV file was generated. Besides the manual fix for the assertion error data resolution, the data generation script was also updated to fix this issue permanently:

```
# Generate Accounting Data

for contract_id in contract_ids:
    outstanding = round(random.uniform(1000, 500000), 2)

    # Ensure arrears ≤ outstanding
    arrears = round(random.uniform(0, outstanding), 2)

    provision = round(random.uniform(0, 10000), 2)
    write_off = round(random.uniform(0, 2000), 2)

    cursor.execute("""
        INSERT INTO AccountingData (contract_id, outstanding_amount,
        arrears_amount, provision_amount, write_off_amount)
        VALUES (%s, %s, %s, %s, %s)
    """, (contract_id, outstanding, arrears, provision, write_off))
```

4.4.2 Re-run the Validation

Once the data was corrected, the validation script was re-run. The error now disappeared, and the [ROL_Accounting] validation completed successfully:

```
[ROL_Counterparty] Column 'sectorcode' passed code list validation.
[ROL_Counterparty] Invalid values in column 'residencycountry'
[ROL_Counterparty] Validation complete.
[ROL_Instrument] Column 'instrumenttype' passed code list validation.
[ROL_Instrument] Column 'currency' passed code list validation.
[ROL_Instrument] Validation complete.
[ROL_Contract] Column 'contractstatus' passed code list validation.
[ROL_Contract] Validation complete.
[ROL_Protection] Column 'protectiontype' passed code list validation.
[ROL_Protection] Validation complete.
[ROL_InterestRateTerms] Column 'ratetype' passed code list validation.
[ROL_InterestRateTerms] Column 'referencerate' passed code list
validation.
[ROL_InterestRateTerms] Validation complete.
[ROL_Accounting] Validation complete.
```

Additionally, to the previously performed validation, it is also considered relevant to integrate **ECB code lists** into the **Python validation logic** for the AnaCredit ROL CSV files. ECB provides **standardized code lists** for fields such as:

- **SectorCode** (e.g. NFC, HH, GOV)
- **Currency** (ISO 4217: EUR, USD, GBP, etc.)
- **NACECode** (economic activity classification)
- **ContractStatus**, **InstrumentType**, **ProtectionType**, etc.

To integrate ECB code lists into the existing validation script, first, it is required to define ECB Code Lists (based on BIRD metadata exports), and second, it is required to update the validation functions. This Integration ensures **regulatory compliance with AnaCredit and IReF standards**, improves **data quality,** and **reduces the risk of rejection** during submission, and **aligns this proof of concept with real-world ECB validation logic**.

So, now, the relevant technical task was to perform validation of ROL CSV files with ECB validation logic (refer to: *07b-validation_rol_csv_files_with_ecb.py*).

At this point, while reviewing and testing the validation logic, it was considered relevant for future usage of this script to automate some tasks. As an example, it was decided to automate the download of ECB code lists from the BIRD portal (Figure 20), therefore converting this validation logic into a reusable module for future work.

So, now, the relevant technical task was to perform validation of ROL CSV files with ECB validation logic while including download of up-to-date NACE code list from ECB portal (refer to: *07c-validation_rol_csv_files_with_ecb_dl.py*).

Figure 20 - Automated List of Values Update for Validation Purposes

Once validation is performed, the next step is to generate mock IReF reports using ECB templates. For this, the most suitable format for mocking a submission is xBRL-CSV, part of the Open Information Model (OIM) suite. It is designed for granular, high-volume datasets such as AnaCredit and is easier to generate from CSVs than traditional XML/XBRL.

For **converting CSVs to xBRL-CSV**, it is first necessary to prepare the CSV files. To do so, it needs to be ensured that the ROL CSVs are cleaned up and validated, that they use consistent column headers matching BIRD attributes, and that they are stored in a data folder. Then, it is necessary to create **metadata files** (two files will be necessary for this):

- Taxonomy and structure definition (refer to: *xbrl.json*)
- Mapping of CSV columns to XBRL concepts (refer to: *table*.json* – one per entity)

Sample XBRL file for Counterparty:

```
{
  "table": {
    "columns": [
      { "name": "counterpartyidentifier", "concept": "ana:CounterpartyIdentifier" },
      { "name": "lei", "concept": "ana:LegalEntityIdentifier" },
      { "name": "sectorcode", "concept": "ana:SectorCode" },
      { "name": "residencycountry", "concept": "ana:ResidencyCountryCode" },
      { "name": "nacecode", "concept": "ana:NACECode" }
    ],
    "file": "ROL_Counterparty.csv"
  }
}
```

Once the files are created, some tools (OIM Validator, Arelle or XML Spy) can be used to validate xBRL-CSV packages, converting them to standard XBRL or XML if needed and simulating submission to ECB or national authorities.

For preparing the submission package, a specific folder structure is required. In the following example, this package simulates what a bank might need to submit to a regulator under the IReF framework using BIRD-compliant data:

```
submission_package
├── metadata
│   ├── submission_manifest.xml
│   ├── version_info.txt
├── data
│   ├── ROL_Counterparty.csv
│   ├── ROL_Instrument.csv
│   ├── ROL_Contract.csv
│   ├── ROL_Protection.csv
│   ├── ROL_InterestRateTerms.csv
│   ├── ROL_Accounting.csv
├── validation
│   ├── validation_report.txt
│   ├── validation_script.py
├── documentation
│   ├── mapping_table.xlsx
│   ├── transformation_rules.pdf
│   ├── thesis_methodology_section.md
```

For each of the mentioned folders, description of each file is as follows:

metadata/

- **submission_manifest.xml**: simulates the header of a regulatory submission (e.g. institution ID, reporting period, schema version).
- **version_info.txt**: notes the BIRD version and transformation rule set used.

data/

- Contains all six **ROL CSV files** generated from PostgreSQL views.

validation/

- **validation_script.py**: full Python script built with ECB code list integration.
- **validation_report.txt**: output summary of validation results (e.g. "All files passed", or list of errors).

documentation/

- **mapping_table.xlsx**: IL → EIL → ROL mapping table.
- **transformation_rules.pdf**: summary of transformation logic applied.
- **thesis_methodology_section.md**: markdown or word document describing validation and submission process.

4.5 Adaptation to Other Datasets

Following the initial approach for this proof of concept, which utilized a self-generated dataset with simulated data, it was deemed essential to run the process against an existing and validated sample dataset. For this purpose, various sources were consulted to look for publicly available datasets, including Kaggle and Google's Dataset Search. However, this type of information, due to its sensitivity and data privacy concerns, is not easily accessible. The second approach was to look for this type of information within the regulators. One option was to look up information within CRC at Banco de Portugal; however, while reviewing available information from the ECB, a sample dataset for AnaCredit was found on the ECB portal[5]. Within the Data Reporting

[5] https://www.ecb.europa.eu/stats/ecb_statistics/anacredit/html/index.en.html

and Standards section, alongside AnaCredit's manuals, some examples[6] of complete reports were found to be available (Figure 21). Although this information was last updated in 2019, it remains relevant for advancing this PoC to the next level.

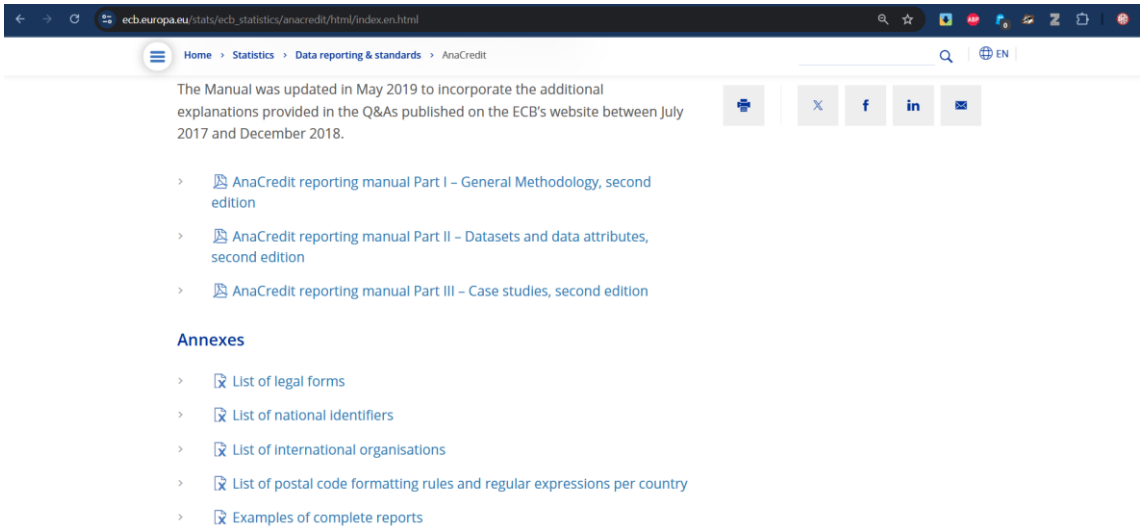


Figure 21 - ECB's AnaCredit Manuals & Examples

Following the link to the examples of provided complete reports, an Excel document provides access to **19 distinct cases, each** with a subset of **10 entities**, from AnaCredit's conceptual model.

4.5.1 Dataset Preparation

Due to the nature of data aggregation for multiple cases within a single Excel file, the option was to split this information to test and manage each case individually, thereby allowing for testing the PoC with several distinct scenarios. Therefore, the first technical step was to create a script for automating the transformation process from the source report to multiple Excel files (one per case) with structured information per entity (one sheet per entity). The primary objective of this process is to prepare these files for export into a new database and then proceed with the transformation tasks, as outlined in the self-generated dataset example.

6 https://www.ecb.europa.eu/stats/money/aggregates/anacredit/shared/pdf/AnaCredit_Manual_Part_III_Examples_of_complete_reports.xlsx

A new Python script was created: *02a-anacredit_to_excel.py*. This script was used to perform the following transformations:

1. Ignore the cover/summary sheet from the original spreadsheet report (*AnaCredit_Manual_Part_III_Examples_of_complete_reports.xlsx*)
2. Convert each of the individual sheets prefixed with "CASE" (Figure 22) to a new standalone spreadsheet (one per case, e.g. *CASE 17.xlsx*)
3. Each of the standalone case spreadsheets contains 10 sheets, one per entity from the original AnaCredit report (e.g. Instrument, Financial, Accounting, etc.)
4. All individual case spreadsheets (Figure 23) are stored within an output folder for the next steps

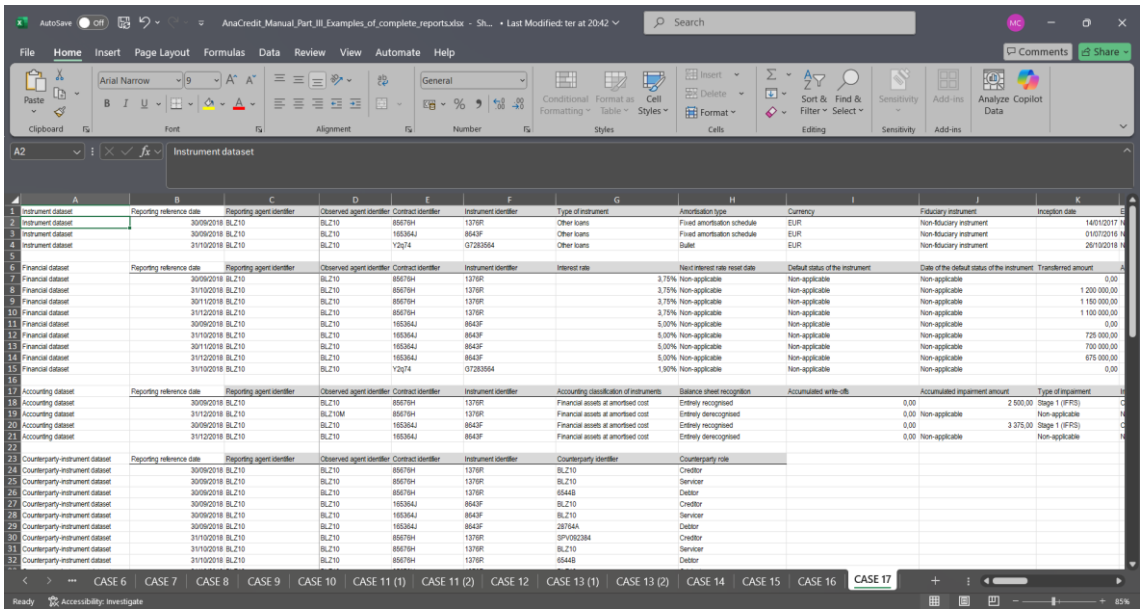


Figure 22 - Sample Case from Original AnaCredit Report

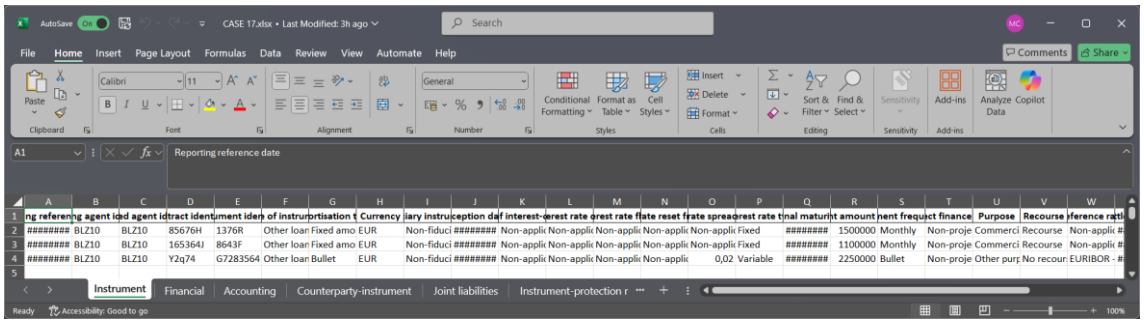


Figure 23 - Sample Individual Case Generated Spreadsheet

Some adjustments were made with Python script (Figure 24), such as naming entities from “Instrument dataset” to “Instrument”, that will later be useful for mapping purposes.

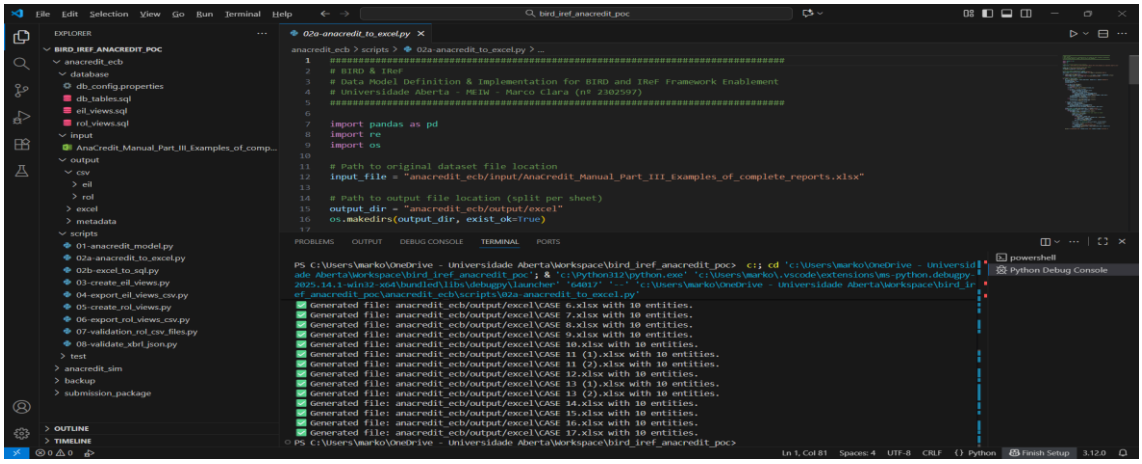


Figure 24 - AnaCredit Sample Dataset Split

Once the individual files for each of the 19 cases are generated, the next step is to prepare a SQL script for creating the database model. This database will hold 10 tables corresponding to each of the entities within the sample dataset, namely: **Instrument, Financial, Accounting, Counterparty reference, Counterparty-instrument, Instrument-protection received, Protection received, Counterparty default, Counterparty risk,** and **Joint liabilities**.

Both entity names and attribute descriptions were normalized to remove spaces, special characters, etc., for the SQL script creation.

A new SQL script (Figure 25) was created: *db_tables.sql* (and Python script: *01-anacredit_model.py*).

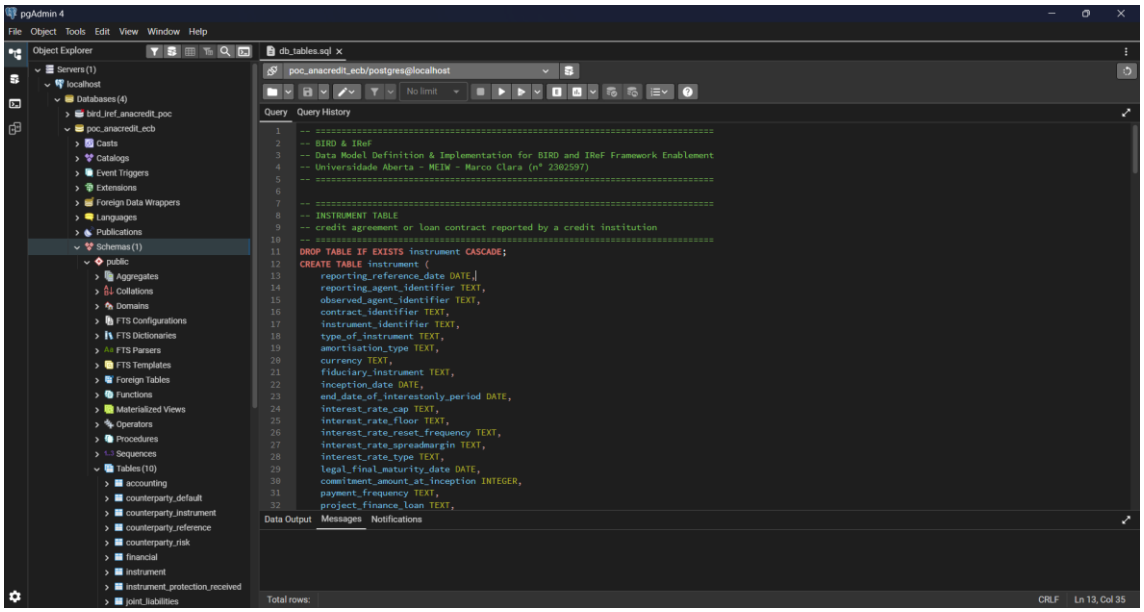


Figure 25 - AnaCredit Sample Database

The generated database and table structure will be used for testing each of the previously generated cases individually. Therefore, the next step would naturally be preparing the loading process for those tables. A new script was created that would read from the output folder where all case spreadsheets were generated, or from a specific folder where a single case would be made available.

From this point onwards, the relevant part of the script involves mapping the entities within each spreadsheet to the tables and their corresponding attributes. Some transformation rules were included in this script, namely by replacing some literals for null values (e.g.: “not applicable”).

A new Python (Figure 26) script was created: *02b-excel_to_sql.py.*

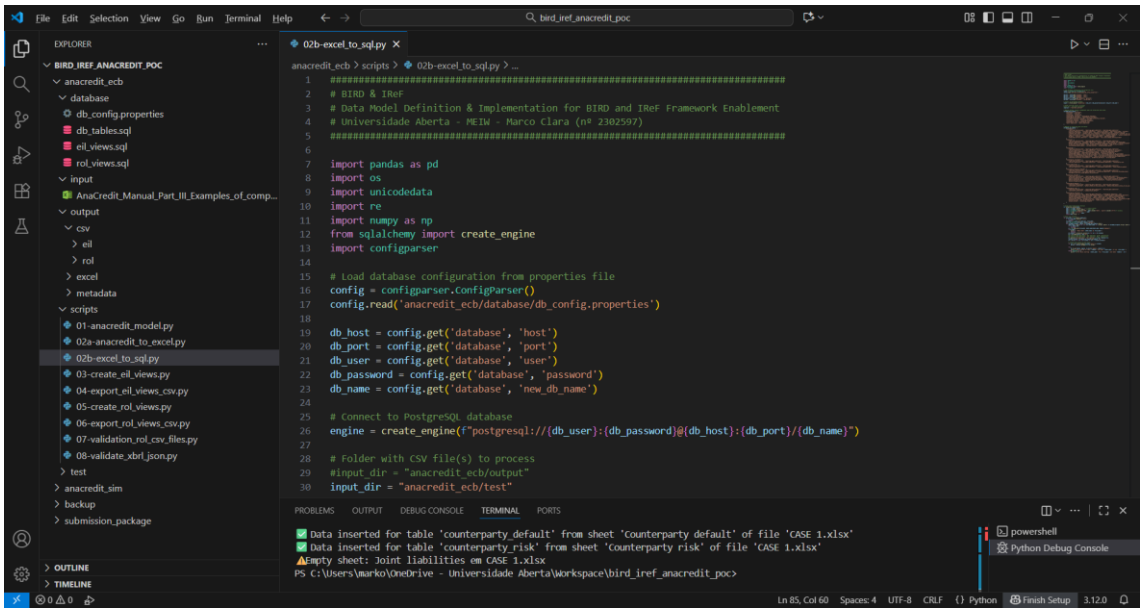


Figure 26 - Database Loading with Sample Case

Not all available cases contain information for every entity, and therefore, it is OK to have database loading script execution results with warnings for specific entities (e.g., Joint Liabilities entity does not have data in several of the sample cases).

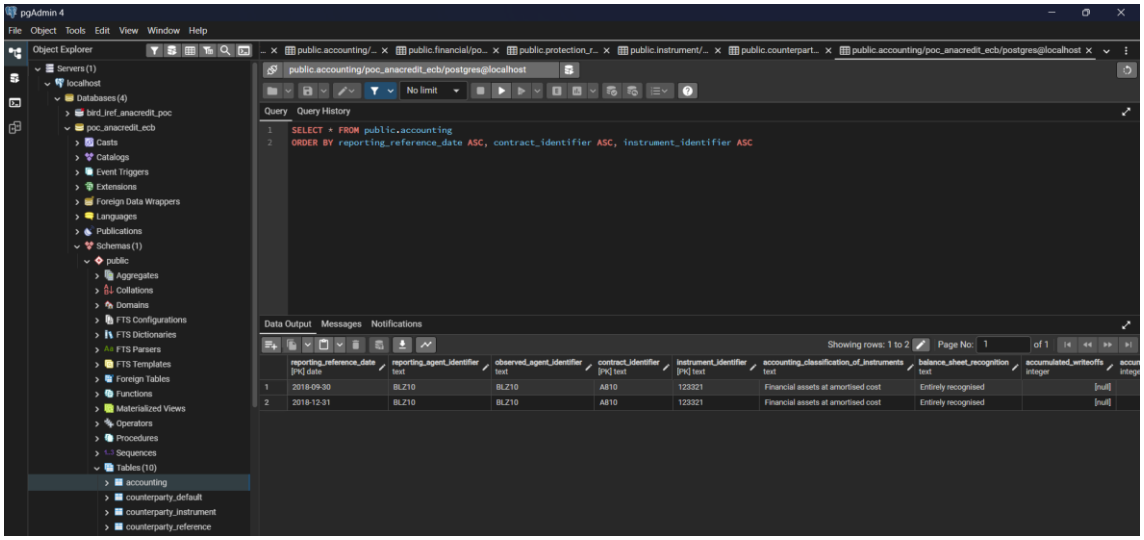


Figure 27 - Sample Data for Accounting Table

At this point, we have the database created and loaded with sample data (Figure 27) from the AnaCredit example report, which can now be used for the remaining transformation process defined within the scope of this PoC.

After the transformation implementation, the approach will be to use each case individually for experimentation purposes, trying out the process with multiple distinct scenarios.

4.5.2 Data Analysis and Preparation

After loading the database with data from the test cases from ECB's AnaCredit example reports, some analysis is required to proceed to the previously established transformation steps, starting with **IL to EIL** using the same target data model structure that was used with the simulated dataset. One issue was promptly identified during analysis, regarding data for the **counterparty** entity. While analysing IL data, two types of counterparty classifications needed to be considered: sector-based (e.g. **NFC**, **FC**, **HH**) and type-based (e.g. **household**, **legal entity**).

Sector Classification (Institutional Sector Codes)

In AnaCredit (and more broadly in ECB statistical frameworks), counterparties are classified by **institutional sector**. These codes are based on the **ESA 2010** (European System of Accounts) sector classification. The most common codes are those presented in Table 23.

Table 23 - Sector Classification Codes

Code	Sector Name	Description
NFC	Non-Financial Corporations	Enterprises producing goods and non-financial services.
FC	Financial Corporations	Banks, insurance companies, investment funds, etc.
HH	Households	Individuals or groups of individuals consuming goods/services or producing for own use.
GG	General Government	Central, regional, and local government institutions.
NPISH	Non-Profit Institutions Serving Households	Charities, foundations, religious institutions.
RoW	Rest of the World	Non-resident entities.

Type Classification (Counterparty Type)

This classification is used in the **Counterparty Reference Data** section of AnaCredit reporting. It refers to the **legal nature** of the counterparty, and is used in AnaCredit to distinguish between, as we can perceive in Table 24.

Table 24 - Type Classification Codes

Code	Type Name	Description
LE	Legal Entity	Any organization with legal personality (e.g., companies, banks, associations).
HH	Household	Natural persons acting in a personal capacity.
IE	Individual Entrepreneur	Natural persons acting in a professional/business capacity.
IG	International Organisation	Entities like the UN, IMF, etc.
GG	Government Entity	Public sector bodies.

The key differences between these two classifications are described below, in Table 25.

Table 25 - Differences Between Classifications

Aspect	Sector Classification	Type Classification
Purpose	Economic role in the economy	Legal form of the counterparty
Examples	NFC, FC, HH, GG, NPISH	Legal Entity, Household, Individual Entrepreneur
Used for	Sectoral analysis, macroeconomic statistics	Legal identification, risk segmentation

According to the official ECB document that contains the full list of **Institutional Sector Codes** used for **counterparty classification** in AnaCredit (the list can be found in **Annex IV** of the mentioned document, based on the **ESA 2010** framework and used in AnaCredit to identify the economic sector of counterparties), the main codes are the ones presented in Table 26.

Table 26 - Institutional Sector Codes

Code	Sector Name	Description
S11	Non-financial corporations	Corporations producing goods and non-financial services

S12	Financial corporations	Includes banks, insurance companies, investment funds, etc.
S121	Central bank	National central banks and ECB
S122	Deposit-taking corporations except the central bank	Commercial banks, savings banks, etc.
S123	Money market funds (MMFs)	Investment funds that invest in short-term debt instruments
S124	Non-MMF investment funds	Other investment funds
S125	Other financial intermediaries	Leasing companies, factoring companies, etc.
S126	Financial auxiliaries	Brokers, insurance agents, etc.
S127	Captive financial institutions and money lenders	Holding companies, special purpose entities, etc.
S128	Insurance corporations	Life and non-life insurance companies
S129	Pension funds	Institutions providing retirement benefits
S13	General government	Central, regional, and local governments
S1311	Central government	Ministries, national agencies
S1313	Local government	Municipalities, local authorities
S14	Households	Individuals or groups of individuals
S15	Non-profit institutions serving households (NPISH)	Charities, foundations, religious institutions, etc.
S2	Rest of the world	Non-resident entities

Mapping: Institutional Sector Codes vs. Counterparty Type Codes

To clarify the relationship between **sector** and **type** classifications, Table 27 presented a **mapping** between institutional sector codes and counterparty type codes, based on ECB documentation (Annex IV and V of the AnaCredit Manual – Part II).

Table 27 - Mapping Between Sector and Type Codes

Institutional Sector Code	Sector Name	Typical Counterparty Type(s)
S11	Non-financial corporations (NFC)	**LE** (Legal Entity)
S12	Financial corporations (FC)	**LE** (Legal Entity)
└— S121	Central bank	**GG** (Government Entity)
└— S122	Deposit-taking corporations	**LE** (Legal Entity)
└— S123–S129	Other financial subsectors	**LE** (Legal Entity)
S13	General government (GG)	**GG** (Government Entity)

└— S1311–S1314	Central, state, local, and social security funds	**GG** (Government Entity)
S14	Households (HH)	**HH** (Household), **IE** (Individual Entity)
S15	Non-profit institutions serving households (NPISH)	**LE** (Legal Entity) or **IG**
S2	Rest of the world (RoW)	**LE**, **IG**, **HH**, depending on case

Legal Entities (LE) are used across most institutional sectors, particularly for corporations and non-profits. **Households (HH)** and **Individual Entrepreneurs (IE)** are typically mapped to **S14**, depending on whether the person is acting in a personal or professional capacity. **Government Entities (GG)** are mapped to **S13** and its subcategories. **International Organisations (IG)** may appear under **S15** or **S2**, depending on residency and legal form.

These data mapping rules are of extreme relevance to prepare the EIL while ensuring compliance according to relevant data standards. These rules are therefore required to be included within the transformation process described in the following section.

4.5.3 IL to EIL Layer

In this step, the **Enriched Input Layer (EIL)** views will be created according to the preliminary data analysis and aligned with the IL data model for the selected AnaCredit dataset. Once again, it will be necessary to create a **comprehensive mapping table** that aligns all database fields with the corresponding **BIRD Input Layer (IL) attributes**, enabling the transformation of synthetic data into a format suitable for IReF reporting (Table 28 to Table 33).

Table 28 - AnaCredit's Counterparty Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
Counterparty	counterparty_identifier	Counterparty.Identifier	Internal ID; mapped to BIRD's unique identifier
Counterparty	name	Counterparty.Name	Direct mapping
Counterparty	lei	Counterparty.LEI	Use LEI if available; fallback to national ID
Counterparty	institutional_sector	Counterparty.SectorCode	Map to ECB sector codes (e.g. NFC, HH)
Counterparty	address_country	Counterparty.ResidencyCountry	ISO 3166-1 alpha-2 format
Counterparty	*NULL*	Counterparty.NACECode	Note: not available in the source dataset
Counterparty	institutional_sector	Counterparty.Type	Mapped from ECB Sector Codes

Table 29 - AnaCredit's Instrument Data Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
Instrument	instrument_identifier	Instrument.Identifier	Internal ID
Instrument	type_of_instrument	Instrument.Type	Maps instrument types (e.g., Loan, CreditLine)
Instrument	currency	Instrument.Currency	ISO 4217 format
Instrument	legal_final_maturity_date	Instrument.MaturityDate	Direct mapping
Instrument	interest_rate_type	Instrument.InterestRateType	'Fixed' or 'Variable'
Instrument	reference_rate	Instrument.ReferenceRate	e.g. EURIBOR, LIBOR
Instrument	interest_rate_spreadmargin	Instrument.Spread	Numeric value

Table 30 - AnaCredit's Contract Data Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
Counterparty Instrument	contract_identifier	Contract.Identifier	Internal ID
Counterparty Instrument	counterparty_identifier	Contract.CounterpartyIdentifier	Internal ID
Counterparty Instrument	instrument_identifier	Contract.InstrumentIdentifier	Internal ID
Financial / Accounting	date_of_the_default_status_of_the_instrument / date_of_the_status_of_forbearance_and_renegotiation	Contract.ContractDate	Direct mapping (from financial data or accounting data)
Financial / Accounting	default_status_of_the_instrument / status_of_forbearance_and_renegotiation/ performing_status_of_the_instrument	Contract.Status	Direct mapping (from financial data or accounting data) Map to BIRD status codes (e.g., Active, Terminated)

Table 31 - AnaCredit's Guarantee Data Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
Protection Received	protection_identifier	Protection.Identifier	Internal ID
Instrument Protection Received	contract_identifier	Protection.ContractIdentifier	Internal ID
Protection Received	type_of_protection	Protection.Type	Map to BIRD protection types
Instrument Protection Received	protection_allocated_value	Protection.Amount	Direct mapping

Table 32 - AnaCredit's Interest Rate Terms Data Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
Instrument	reporting_reference_date, reporting_agent_identifier, instrument_identifier	InterestRateTerms.RateTermIdentifier	Internal ID
Instrument	instrument_identifier	InterestRateTerms.InstrumentIdentifier	Internal ID
Instrument	interest_rate_type	InterestRateTerms.Type	'Fixed' or 'Variable'
Instrument	reference_rate	InterestRateTerms.ReferenceRate	Same as Instrument
Instrument	interest_rate_spreadmargin	InterestRateTerms.Spread	Numeric value

Table 33 - AnaCredit's Accounting Data Mapping Table

PostgreSQL Table	Field	Mapped BIRD IL Attribute	Transformation Rule / Notes
Accounting	reporting_reference_date, reporting_agent_identifier, instrument_identifier	Accounting.AccountingIdentifier	Internal ID
Accounting	contract_identifier	Accounting.ContractIdentifier	Internal ID
Financial	outstanding_nominal_amount	Accounting.AmountOutstanding	Direct mapping
Financial	arrears_for_the_instrument	Accounting.AmountInArrears	Direct mapping
Accounting	provisions_associated_with_offbalancesheet_exposures	Accounting.ProvisionAmount	Direct mapping
Accounting	accumulated_writeoffs	Accounting.WriteOffAmount	Direct mapping

So, now, the relevant technical tasks are:

- Designing the EIL views SQL script (refer to: *eil_views.sql*): this is where all identified and necessary rules were applied to properly map data from AnaCredit's new and original dataset to the previously defined EIL
- Creating the EIL views (refer to: *03-create_eil_views.py*)
- Exporting the EIL views data to CSV (refer to: *04-export_eil_views_csv.py*): this was the same script used in the simulated data scenario, with no modifications required

Once again, as an output of the export process six CSV files (Figure 28) will be generated, simulating the **BIRD compliant EIL output**, ready for transformation into the **Reference Output Layer (ROL)** or for mock submission for this proof of concept: *EIL_Counterparty.csv, EIL_Instrument.csv, EIL_Contract.csv, EIL_Guarantee.csv, EIL_InterestRateTerms.csv* and *EIL_AccountingData.csv*

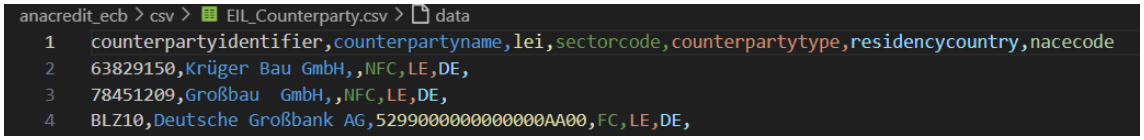


Figure 28 - Sample CSV File Generated from EIL View

One of the goals at this point was to achieve the same type of output for EIL that was generated with the previously simulated dataset, so that the following steps can be taken by reusing previously prepared scripts for transformation and validation purposes. As this goal is achieved, the remaining process can be considered standard and fit for any output that can come out of the EIL.

4.5.4 EIL to ROL Layer

In this step, the **Reference Output Layer (ROL)** views will be created according to the structured output derived from the EIL, formatted according to the reporting templates and technical specifications defined by the ECB, representing the **final reporting ready dataset** before submission to regulators under IReF.

So, now, the relevant technical tasks are creating ROL views from EIL data (refer to: *rol_views.sql* and *05-create_rol_views.py*) and exporting ROL views data to CSV files (refer to: *06-export_rol_views_csv*).

Both tasks were performed running the same scripts that were used in the simulated data scenario, with no modifications required.

As an output of the export process six CSV files (Figure 29) will be generated, simulating the **BIRD compliant ROL output** for mock submission for this proof of concept

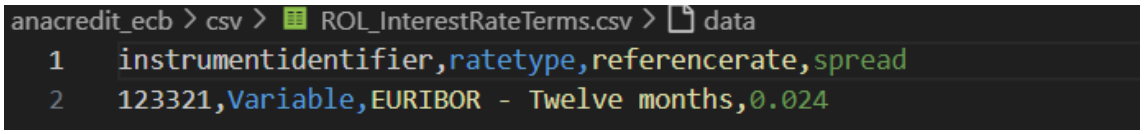


Figure 29 - Sample CSV File Generated from ROL View

4.5.5 Validation

Following the ROL transformation, it is now important to proceed with the validation steps. While directly running the validation script (*07-validation_rol_csv_files.py*) for the first case from the dataset, it was clear that something was missing from the initial transformation to EIL that later affected ROL data. Although most of entities passed validation criteria:

```
[ROL_Counterparty] Column 'sectorcode' passed code list validation.
[ROL_Counterparty] Column 'residencycountry' passed code list validation.
[ROL_Counterparty] Validation complete.
[ROL_Instrument] Column 'instrumenttype' passed code list validation.
[ROL_Instrument] Column 'currency' passed code list validation.
[ROL_Instrument] Validation complete.
[ROL_Instrument] Column 'currency' passed code list validation.
[ROL_Instrument] Validation complete.
[ROL_Instrument] Validation complete.
[ROL_Contract] Column 'contractstatus' passed code list validation.
[ROL_Contract] Validation complete.
(…)
```

One revealed an issue:

```
AssertionError: Duplicate ProtectionIdentifier
```

The cause of this issue was promptly identified, and it related to the existence of an attribute for reporting date within the original dataset from AnaCredit, which means

that for this entity, the same record can have multiple entries with different reporting dates. Once analysed, the problem was traced back from ROL to EIL, so it was necessary to go back to the EIL transformation step and perform the required adjustments to correct this. In this case, it was a minor modification to the corresponding view:

```
-- EIL Guarantee View
CREATE OR REPLACE VIEW EIL_Guarantee AS
SELECT DISTINCT
  pr.protection_identifier AS ProtectionIdentifier,
  ipr.contract_identifier AS ContractIdentifier,
  pr.type_of_protection AS ProtectionType,
  ipr.protection_allocated_value AS ProtectionAmount
FROM instrument_protection_received ipr
JOIN protection_received pr
  ON ipr.protection_identifier = pr.protection_identifier
WHERE ipr.reporting_reference_date = pr.reporting_reference_date;
```

Once the previous issue was resolved, it was also considered relevant to enhance the validation script built for the simulated dataset experiment. In this case, further automation was implemented regarding the list of validations, specifically by retrieving additional lists of codes from official sources and incorporating them into the validation functions (Table 34).

Table 34 - Code List Sources

Code List	Source
NACE Codes	Already implemented via EU Publications SDMX XML
Currency Codes	ECB Reference Exchange Rates XML (ECB FX Rates)
Residency Countries	ISO 3166 country codes via https://datahub.io/core/country-list or ECB SDMX
Sector Codes, Instrument Types, Protection Types, Contract Status, Rate Types, Reference Rates	Still hardcoded as no direct SDMX source found yet but can be modularized for future extension via BIRD/IReF metadata

The key enhancements with this update are retrieving ISO currency codes from ECB FX rates XML and retrieving ISO 3166 country codes from Datahub CSV. This is done with modularized code list fetching for future extension, replacing hardcoded lists where (and whenever) possible.

After validation of ROL CSV files, it is now possible to move forward to xBRL validation using Arelle. A set of **xBRL JSON** files is prepared and used for validation

purposes, aiming to prepare the mocked submission package. The structure of xBRL files includes the main file (refer to: *xbrl.json*) that references a JSON target file for each table, namely:

```
"tables": [
   "table_counterparty.json",
   "table_instrument.json",
   "table_contract.json",
   "table_protection.json",
   "table_interestrateterms.json",
   "table_accounting.json"
 ]
```

Each of the referenced tables contains a specification regarding validation that takes place for each column. E.g. counterparty table (refer to: *table_counterparty.json*):

```
"columns": [
   { "name": "counterpartyidentifier", "concept": "ana:CounterpartyIdentifier" },
   { "name": "lei", "concept": "ana:LegalEntityIdentifier" },
   { "name": "sectorcode", "concept": "ana:SectorCode" },
   { "name": "residencycountry", "concept": "ana:ResidencyCountryCode" },
   { "name": "nacecode", "concept": "ana:NACECode" }
 ]
```

Once the xBRL JSON files are prepared, a new and last script is prepared for running this set of validations (refer to: *08-validate_xbrl_json.py*). At this point and upon successful validation for all entities, it is now possible to move forward and prepare the package for submission.

Chapter 5 - Results & Discussion

5.1 Research Report Results

Based on the current research, several key interpretations emerge, namely:

Framework Integration Trends: the analysis reveals a strong movement toward integrated reporting frameworks, with BIRD and IReF emerging as dominant approaches (Casa, 2024). This suggests a clear industry direction toward standardisation and harmonisation of reporting requirements across the banking sector.

Layered Architecture Approach: the studies consistently support a four-layer architecture for data management (Figure 30), consisting of a Data Layer (granular data collection), a Semantic Layer (standardised definitions), a Processing Layer (transformation rules), and a Reporting Layer (output generation).

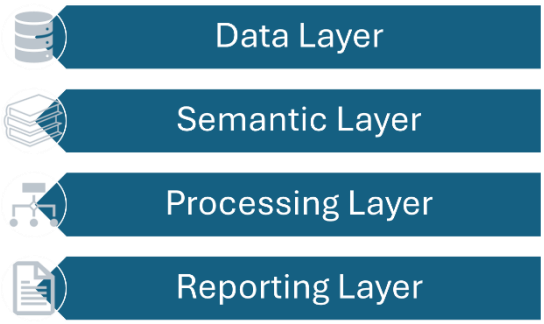


Figure 30 - Four-Layer Architecture

Implementation Challenges: the research identifies significant implementation hurdles, including complex integration of diverse reporting requirements, substantial IT infrastructure investments and cross-jurisdictional standardisation challenges (Drvar et al., 2021).

Regulatory Alignment: the findings suggest that the proposed frameworks align with BCBS 239 principles, particularly regarding data aggregation and reporting (Poloni, 2024). However, the research notes a lack of empirical evidence for successfully implementing these frameworks in real banking environments.

Evolution of Reporting Approaches: there is a clear trend toward more agile, technology-driven approaches to regulatory reporting, as evidenced by the emergence of RegOps and digital platform solutions (Drvar et al., 2021).

Data Quality and Governance: the research emphasises the critical importance of robust data governance structures and quality assurance mechanisms (Casa et al., "PUMA Cooperation Between the Bank of Italy"). This suggests that successful implementation depends not just on technical solutions but also on strong organisational frameworks.

Future Implications: the findings point to a future where centralised, standardised reporting frameworks become the norm, though significant work remains to achieve this vision (Colangelo et al., 2021). The lack of empirical evidence of implementation suggests that practical validation of these frameworks remains a crucial next step.

This interpretation indicates that while there is a strong theoretical foundation and agreement on the direction of regulatory reporting evolution, significant practical challenges remain in implementing these frameworks effectively.

5.2 State-of-the-Art Gap Analysis

Throughout the current study, several key gaps and distinguishing features emerged, namely:

Lack of Empirical Validation: None of the reviewed studies provides empirical evidence of successfully implementing integrated frameworks; the studies are primarily theoretical analyses and framework proposals, lacking real-world testing.

Integration Challenges: While studies propose various frameworks, none demonstrate successful simultaneous implementation of BIRD, IReF, and BCBS 239 compliance; the complexity of harmonising diverse reporting requirements across jurisdictions remains a significant challenge.

Technical Implementation Gaps: Studies focus on conceptual frameworks but lack detailed technical specifications for implementation; there is limited guidance on the practical aspects of transitioning from legacy systems to integrated frameworks.

Data Quality Assurance: While data quality requirements are discussed theoretically, there is a lack of empirical validation of proposed quality assurance mechanisms; the effectiveness of proposed validation frameworks remains untested in real-world scenarios.

Governance Structure: Studies propose various governance models but lack evidence of their effectiveness in practice; the practical challenges of implementing centralised governance across different jurisdictions are not fully addressed.

Cost-Benefit Analysis: There is limited analysis of the actual costs versus benefits of implementing these integrated frameworks; the economic impact of implementation on different-sized institutions is not thoroughly examined.

Standardisation Challenges: While standardisation is emphasised as crucial, there is no empirical evidence of successful standardisation across multiple jurisdictions; the practical challenges of achieving semantic integration across different reporting domains remain unaddressed.

This gap analysis highlights the primarily theoretical nature of existing research and the need for empirical studies to validate the proposed frameworks' effectiveness in real-world banking environments.

5.3 Proof of Concept

This section provides a critical evaluation of the proof of concept developed to integrate selected entities from the AnaCredit conceptual model into a mock reporting package aligned with the Integrated Reporting Framework (IReF). The assessment focuses on three key dimensions: **data quality**, **automation benefits**, and alignment with **regulatory compliance**. These dimensions are essential to validate the feasibility, robustness, and readiness of the proposed solution.

Prior to the three dimensions assessment, it is also important to mention that the initially specified success criteria were successfully covered, namely the ones described in Table 35.

Table 35 - Success Criteria

Criteria	Description	Result
Data Coverage	All mandatory AnaCredit attributes for specified entities are correctly mapped and populated	100%
Technical Compliance	The output file follows the structure and format defined by IReF for AnaCredit	Yes
Automation	The transformation and report generation processes are automated via scripts	Yes
Data Quality Validation	Validation rules are implemented to detect errors or inconsistencies before report generation	Yes
Complete Documentation	The project includes clear documentation of architecture, mappings, rules, and limitations	Yes
Academic Presentation	The final presentation clearly explains technical choices and demonstrates understanding of BIRD and IReF concepts	Yes

5.3.1 Data Quality Assessment

High-quality data is fundamental to effective regulatory reporting. The BIRD framework provides a structured metadata model and transformation logic that ensures:

- **Semantic consistency** across entities and attributes
- **Technical validation** through derivation and transformation rules
- **Traceability** from source systems to regulatory outputs

In this proof of concept, entities such as Counterparty, Instrument, Contract, Guarantee, Interest Rate Terms, and Accounting Data were mapped to BIRD input layer cubes. This mapping enabled:

- Elimination of ambiguity in attribute definitions (a full set of mapping tables was created to ensure 100% alignment between attributes)
- Application of validation rules to ensure data integrity (a validation script was created, improved, and successfully tested, to automatically check for any data integrity issues)
- Simulation of reporting scenarios with complete and coherent datasets (simulated data was generated, tested, and corrected to ensure coherence)

- Testing of real-life scenarios with existing datasets (real-world data from existing reports was used to run the end-to-end process and further validate its reusability)

Regarding the real-world dataset, it not only provided a more accurate way to test the process but also enabled the creation of multiple distinct scenarios for experimentation (Table 36).

Table 36 - AnaCredit Entities per Case

Entities per Case	Instrument	Financial	Accounting	Counterparty Reference	Counterparty Instrument	Instrument Protection	Protection Received	Counterparty Default	Counterparty Risk	Joint Liabilities
CASE 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
CASE 2	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No
CASE 3	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
CASE 4	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
CASE 5	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No
CASE 6	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No
CASE 7	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No
CASE 8	Yes	Yes	No	Yes	Yes	No	No	No	Yes	No
CASE 9	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	No
CASE 10	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
CASE 11 (1)	Yes	Yes	No	Yes	Yes	No	No	Yes	Yes	No
CASE 11 (2)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
CASE 12	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
CASE 13 (1)	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No
CASE 13 (2)	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No
CASE 14	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
CASE 15	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
CASE 16	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CASE 17	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No

By leveraging BIRD's input layer structure, the reporting package met the expected standards of completeness, consistency, and reliability, which are critical for submission to national central banks under AnaCredit and IReF.

5.3.2 Automation Benefits

Automation is a central objective of both BIRD and IReF, aiming to reduce manual effort and increase efficiency in regulatory reporting. The proof of concept demonstrated automation through:

- **Automated data extraction** from internal systems (mocked): data extraction between the different layers was performed using SQL and Python scripts for automation, relying on database data extraction and transformation
- **Transformation logic** to generate output reports based on BIRD rules: several views were created to assist the different layer transformation logic, applied using Python scripts for automation purposes (EIL, ROL), followed by data export to standard data file formats such as CSV, JSON, and xBRL
- **Validation mechanisms** to ensure readiness for submission: validation scripts were created using Python to check for basic data integrity over the exported data (CSV files) and to perform validation against ECB codes (NACE, currency, and country codes)

These components showed that:

- Automation significantly reduces reporting preparation time
- It minimizes human error and improves data consistency
- It supports scalability across multiple reporting frameworks

The modular structure of BIRD, with reusable input cubes and transformation rules, proved effective in supporting the generation of IReF-ready reports with minimal manual intervention.

5.3.3 Compliance Alignment

Regulatory compliance is ensured by adhering to the standards set by the European Central Bank. The proof of concept aligns with:

- The **AnaCredit conceptual model**, as defined in the *AnaCredit Manual Part II*
- The **BIRD-IReF integration roadmap**, outlined in the *BIRD Workplan 2025*
- The **harmonization principles** of IReF for statistical reporting

The reporting package developed in this project reflects the ECB's vision for a unified and streamlined reporting framework. It anticipates future regulatory requirements and demonstrates interoperability between AnaCredit and IReF, ensuring long-term compliance and adaptability (Figure 31).

BIRD WORKPLAN 2025		2025											
		Q1			Q2			Q3			Q4		
ACTIVITY	TASK	1	2	3	4	5	6	7	8	9	10	11	12
IReF	Support to IReF project:												
	- Modelling IReF Logical Data Model [*1]												
	- Prototyping/Testing IReF Logical Data Model												
	IReF Transformation rules [*2]												
FINREP & Asset Encumbrance	FINREP [*3] and Asset Encumbrance Subgroup:												
	- Enhancement of FINREP based on initial feedback received												
	- Finalization of Asset Encumbrance Transformation Rules												
AnaCredit	Finalize the integration of AnaCredit into BIRD [*4]												
SSM Loan Tape	Start integrating SSM Loan Tape into BIRD												
COREP	Start integrating COREP (Credit Risk) into BIRD												
Regular update of included Frameworks	Investigation phase [*5]												
Work Stream on Data Modelling	Modelling of IReF requirements												
	Modelling of Asset Encumbrance requirements												
	Modelling of AnaCredit requirements												
	Modelling of SSM Loan Tape requirements												
	Modelling of COREP requirements												
Work Stream on Prototyping	Executing BIRD community supporting activities												
Communication and Promotion	Regular work												
Transformation Rules format survey	Assessment of survey results												
	Development of a proposed way forward												

[*1] The IReF Team develops and finalises the IReF Logical Data Model. The BIRD IReF Subgroup offers consultation and input to ensure that the data model aligns with practical implementation needs and user perspectives.
[*2] The start of this activity in October 2025 depends on the availability of an IReF draft regulation.
[*3] FINREP Scope: 77 out of 122 Templates.
[*4] Creation of AnaCredit Transformation Rules and complementary BIRD deliverables in accordance with current formats.
[*5] Outline a strategy to update BIRD Frameworks after their initial incorporation and report findings to the Steering Group for review and decision-making.

Figure 31 - BIRD Work Plan 2025

5.4 Standards & Implementation

The evolution of regulatory reporting frameworks in Europe reflects a paradigm shift from template-based approaches toward metadata-driven, harmonized models. This transformation is driven by the need for granular data, interoperability, and automation in supervisory and statistical reporting. Three key standards underpin this ecosystem:

- **Data Point Model (DPM)**, developed by the European Banking Authority (EBA) for supervisory reporting
- **eXtensible Business Reporting Language (XBRL)**, an XML-based encoding standard for financial data exchange
- **Statistical Data and Metadata eXchange (SDMX)**, an ISO standard for statistical data sharing among international organizations

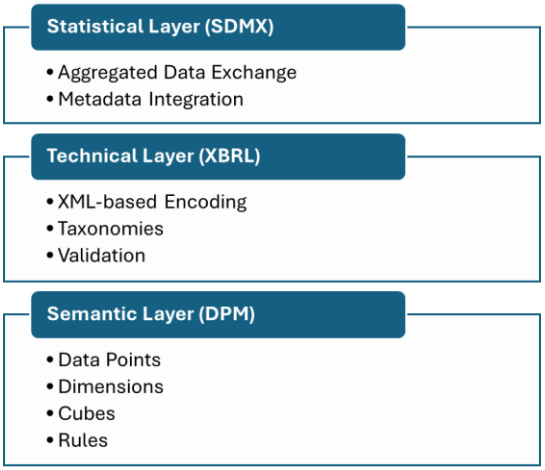


Figure 32 - Layered Architecture of Standards

These standards (Figure 32) form the foundation for initiatives such as the Banks' Integrated Reporting Dictionary (BIRD) and the Integrated Reporting Framework (IReF), which aim to harmonize input data models and reporting processes across the European System of Central Banks (ESCB) and supervisory authorities.

5.4.1 DPM - Data Point Model

The DPM is a semantic framework that defines regulatory reporting requirements at the most granular level. It organizes data into data points, each representing a unique reporting concept, and structures them through dimensions, domains, and cubes for multidimensional analysis.

Key components include:

- **Core Package:** defines fundamental concepts and relationships
- **Mapping Package**: specifies how input data maps to regulatory outputs
- **Validation Rules**: ensure data integrity and compliance

DPM supports the generation of XBRL taxonomies, enabling machine-readable encoding of supervisory templates such as FINREP and COREP. Its design promotes semantic consistency, reducing ambiguity in interpretation and facilitating automation in data processing.

5.4.2 XBRL - Technical Encoding for Financial Reporting

XBRL (eXtensible Business Reporting Language) is an international open standard based on XML, designed for the electronic communication of business and financial data. It uses taxonomies to define concepts, relationships, and validation rules, ensuring uniformity in data exchange.

In the European context, XBRL is the mandatory transmission format for EBA's Implementing Technical Standards (ITS) on supervisory reporting. Its main advantages include:

- **Automation**: Facilitates validation and reduces manual intervention
- **Standardization**: Harmonizes reporting formats across jurisdictions
- **Transparency**: Enhances comparability of financial statements

While XBRL provides the technical layer for encoding and transmitting data, it relies on semantic models like DPM to define the meaning and structure of the data points.

5.4.3 SDMX - Statistical Data Exchange

SDMX (Statistical Data and Metadata eXchange) is an ISO standard developed for the exchange of statistical data and metadata among international organizations and national authorities. Unlike XBRL, which focuses on financial and prudential reporting, SDMX addresses:

- **Aggregated Data**: Supports macro-level statistical reporting
- **Metadata Integration**: Provides rich descriptive structures for datasets
- **Interoperability**: Enables cross-domain data sharing among statistical agencies

SDMX is widely used by institutions such as the ECB and Eurostat for macroeconomic and financial statistics, complementing granular frameworks like DPM for micro-level supervisory data.

5.4.4 Integration with BIRD and IReF

The Banks' Integrated Reporting Dictionary (BIRD) and the Integrated Reporting Framework (IReF) represent strategic initiatives to harmonize data collection and reporting across supervisory and statistical domains. Their integration with DPM, XBRL, and SDMX enables:

- **End-to-End Digitalization**: From input data extraction to output encoding
- **Semantic and Technical Alignment**: BIRD defines the input layer and transformation rules, while DPM structures the output layer and XBRL encodes it for transmission
- **Statistical Interoperability**: SDMX ensures that aggregated data flows seamlessly into macroeconomic frameworks

The ecosystem model developed by the ESCB and EBA illustrates this synergy:

- **BIRD** provides banks with a harmonized data dictionary and transformation logic
- **DPM** translates these requirements into granular data points for supervisory reporting
- **XBRL** operationalizes the transmission of these data points in a machine-readable format
- **SDMX** integrates statistical outputs for broader policy analysis

This layered approach supports the principle of "define once, report once", reducing reporting burdens and enhancing data quality across regulatory value chain (Figure 33).

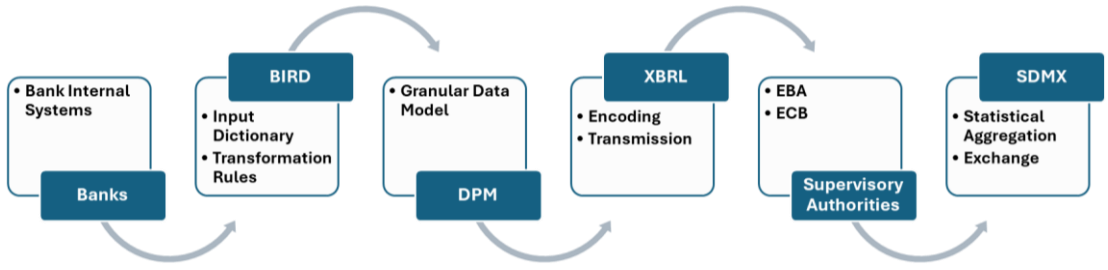


Figure 33 - Data Flow from Banks to Authorities

DPM, XBRL, and SDMX are complementary standards that collectively enable a harmonized, metadata-driven reporting ecosystem. Their integration with BIRD and IReF represents a significant step toward platform-based regulatory reporting, fostering interoperability, automation, and transparency in the European financial system. Future developments, such as API-driven architectures and RegOps paradigms, will further enhance this integration, paving the way for real-time supervision and advanced analytics.

5.5 Quality Controls

Data quality is a cornerstone of effective regulatory reporting. Supervisory authorities and central banks rely on accurate, consistent, and timely data to ensure financial stability and support policy decisions. In the context of metadata-driven frameworks such as BIRD and IReF, quality controls are embedded throughout the reporting lifecycle from data extraction at banks to final dissemination for supervisory and statistical purposes.

5.5.1 Data Quality

Quality in regulatory reporting is multidimensional, encompassing:

- **Accuracy**: Data must faithfully represent observed phenomena
- **Consistency**: Logical coherence across datasets and over time
- **Comparability**: Harmonization across institutions and jurisdictions
- **Timeliness**: Availability of data within prescribed deadlines
- **Integrity**: Protection against errors, manipulation, and system failures
- **Cost-effectiveness**: Balancing the depth of checks with operational burden

These principles align with international standards such as the IMF Data Quality Assessment Framework and the Code of European Statistics, ensuring trust in official statistics and supervisory data.

5.5.2 Quality Control Mechanisms

Quality controls are applied at multiple stages of the reporting process, leveraging metadata-driven IT systems and automated validation rules. The Bank of Italy's integrated approach offers a reference model widely adopted in Europe.

Automated Checks - upon data submission, a layered set of controls is activated:

- **Format Checks**: Verify compliance with expected technical standards (e.g., SDMX or XBRL). Files failing these checks are rejected immediately
- **Formal Checks**: Validate coding against predefined dictionaries and domains (e.g., economic activity codes)
- **Deterministic Checks**: Ensure logical relationships between variables (e.g., balance sheet totals match within tolerance)
- **Statistical-Probabilistic Checks**: Assess plausibility of trends using tolerance thresholds and advanced algorithms

Error Severity and Actions - errors are classified by severity:

- **Very Serious**: Block data release until corrected
- **Non-Serious**: Data may be provisionally accepted with remarks sent to reporters for confirmation or correction

5.5.3 Governance and Quality Assurance

Specialized units within central banks manage quality controls, ensuring independence from data users while maintaining close cooperation for algorithm design. Quality improvement is iterative, requiring continuous interaction with reporting agents and periodic refinement of validation rules. The growing complexity and volume of granular data demand advanced techniques:

- **Machine Learning**: Enables dynamic adjustment of validation rules based on historical patterns, reducing false positives and negatives.
- **Big Data Analytics**: Supports anomaly detection and predictive quality indicators.

These innovations enhance efficiency and accuracy, reducing manual intervention and improving responsiveness to evolving reporting requirements.

5.5.5 Integration with BIRD and IReF

Quality controls are integral to the **BIRD-IReF ecosystem**:

- **BIRD**: Provides harmonized input data definitions and transformation logic, reducing inconsistencies at source.
- **DPM**: Structures output data points for supervisory reporting, enabling precise validation.
- **XBRL and SDMX**: Ensure technical compliance and interoperability, embedding validation rules in taxonomies and data structures. Together, these frameworks support **end-to-end quality assurance**, reinforcing the principle of **"define once, report once"** across supervisory and statistical domains.

5.6 Data Lineage

In modern regulatory reporting ecosystems, data lineage plays a critical role in ensuring transparency, traceability, and compliance. It refers to the ability to track the origin, transformations, and movement of data across systems and processes. With initiatives like BIRD and IReF promoting metadata-driven reporting, lineage frameworks such as OpenLineage and Apache Atlas provide the technological foundation for governance and auditability.

Data lineage addresses fundamental questions about the lifecycle of data within regulatory reporting processes. It seeks to clarify **where the data originates**, typically from source systems within financial institutions. It also examines **how the data is transformed** through processes such as ETL (Extract, Transform, Load) operations and aggregation logic, which prepare the information for reporting purposes. Finally, data lineage tracks **where the data ultimately resides**, whether in regulatory templates or statistical outputs used by supervisory authorities.

In the context of regulatory reporting, data lineage plays a vital role in ensuring **compliance**, **auditability**, and **quality assurance**. Compliance is achieved by demonstrating that data handling and transformations adhere to supervisory requirements and established standards. Auditability is supported by providing clear evidence of data flows and transformations, enabling both internal and external audits

to verify accuracy and integrity. Additionally, lineage contributes to quality assurance by allowing institutions to perform root-cause analysis when errors occur, thereby improving reliability and reducing risks associated with incorrect or incomplete reporting.

5.6.1 OpenLineage

OpenLineage is an open standard for metadata and lineage collection in data pipelines. It integrates with modern orchestration tools to capture:

- **Job-level metadata**: execution context, parameters
- **Dataset-level lineage**: input-output relationships
- **Event-based tracking**: real-time lineage updates

Its relevance for regulatory reporting lies in:

- **Granular traceability**: mapping transformations from raw banking data to supervisory templates
- **Integration with BIRD**: capturing lineage for transformation rules defined in BIRD
- **Support for RegOps**: enabling automated governance in agile regulatory operations

5.6.2 Apache Atlas

Apache Atlas is a metadata management and governance platform that provides:

- **Business and technical metadata cataloguing**
- **Lineage visualization**: End-to-end flow of data assets
- **Policy enforcement**: Security and compliance controls

For IReF and BIRD, Atlas can:

- Maintain a central metadata repository for input dictionaries and transformation logic
- Visualize lineage across semantic (DPM) and technical (XBRL/SDMX) layers
- Enable impact analysis when regulatory requirements change

5.6.3 Integration with BIRD and IReF

The integration of lineage frameworks with BIRD and IReF provides a robust foundation for **metadata-driven governance**. This approach ensures that lineage is aligned with semantic models such as the Data Point Model (DPM) and technical standards like XBRL and SDMX. By embedding governance principles into the metadata layer, institutions can maintain consistency and compliance across all stages of the reporting process.

Another critical aspect of this integration is **end-to-end traceability**. Data lineage enables organizations to track the entire journey of data, starting from bank source systems and moving through the BIRD input layer, the DPM semantic layer, and finally into XBRL encoding for transmission. Once processed, data can be aggregated and exchanged using SDMX for statistical reporting. This comprehensive traceability ensures that every transformation and movement of data is documented and auditable.

Finally, lineage frameworks enhance **regulatory transparency** by providing supervisors with clear evidence of data transformations and compliance. This visibility allows regulators to verify that reporting processes adhere to established standards and that the data submitted is accurate, complete, and reliable. By integrating lineage with BIRD and IReF, institutions not only meet compliance requirements but also strengthen trust and accountability in the regulatory reporting ecosystem. This approach strengthens the principle of “define once, report once”, ensuring that every transformation is documented and auditable within a unified metadata ecosystem.

OpenLineage and Apache Atlas are essential components of a modern regulatory reporting architecture. By embedding lineage into the BIRD & IReF ecosystem, institutions can achieve full traceability, improved data quality, and regulatory compliance, while enabling advanced analytics and automation for future supervisory frameworks.

The proof of concept successfully demonstrated the integration of AnaCredit entities into an IReF-compatible reporting package. It validated the practical application of BIRD standards and highlighted the benefits of structured metadata, automation, and regulatory alignment. This evaluation confirms the viability of the approach and its potential for broader implementation in real-world regulatory environments.

Chapter 6 - Conclusion

6.1 Key Findings

At an early stage of this study, the literature review revealed several key findings across multiple domains:

Framework Coverage and Integration: BIRD emerged as the most widely studied framework, covered in 6 of 11 studies, followed by IReF, covered in 4 studies (Casa, 2024). There is a strong emphasis on framework integration and standardisation across the banking sector (Connell, 2020) and a movement toward unified multipurpose databases for granular credit data collection (ECB 2016/867).

Architectural Approach: The research consistently identified a four-layer architecture that includes a Data Layer (granular data collection), a Semantic Layer (standardised definitions), a Processing Layer (transformation rules) and a Reporting Layer (output generation) (Casa et al., "The Integrated Approach by Bank of Italy").

Implementation Requirements: Key technical requirements identified include robust data integration infrastructure, real-time processing capabilities, standardised interfaces and APIs and comprehensive data quality assurance mechanisms (Drvar et al., 2021).

Major Challenges: several significant challenges were identified, namely complexity in harmonising diverse reporting requirements, substantial technology investment needs, difficulties with cross-jurisdictional standardisation, and data quality maintenance (Poloni, 2024).

Data Governance: the studies emphasised the importance of centralised oversight mechanisms, clear data ownership structures, robust quality assurance processes and comprehensive governance frameworks (Casa et al., "PUMA Cooperation Between the Bank of Italy").

Regulatory Alignment: proposed frameworks generally align with BCBS 239 principles, focus on effective risk data aggregation and reporting and show the need for standardised reporting approaches (Poloni, 2024).

Future Direction: the findings indicate movement toward more integrated reporting systems, increased standardisation across jurisdictions, greater emphasis on

real-time data processing and enhanced focus on data quality and governance (Drvar et al., 2021).

Research Limitations: important limitations were noted, including a lack of empirical evidence for successful implementation, reliance on primarily theoretical analyses and framework proposals, and limited real-world testing of proposed solutions (Colangelo et al., 2021).

Technology Integration: there is a growing emphasis on RegOps and digital platform solutions, a need for sophisticated API and integration capabilities, and a focus on automated validation and quality control (Jago Flötgen et al., 2020).

Quality Assurance: the review highlighted requirements for automated validation processes, continuous monitoring systems and comprehensive data quality frameworks (Bašić, 2017).

A significant finding is that while theoretical frameworks and approaches are well-developed, there is limited empirical evidence of successful implementation in real banking environments (Colangelo et al., 2021). This suggests a need for practical validation of these frameworks through real-world implementation studies.

The review also indicates that successful implementation requires a balanced approach combining technical solutions with robust governance structures and quality assurance mechanisms (Casa, 2024). This highlights the importance of considering technical and organisational aspects in implementing unified reporting frameworks.

6.2 Implications for Practice

Based on the research findings, there are several key implications (Figure 34) for software engineering practice in the banking sector:

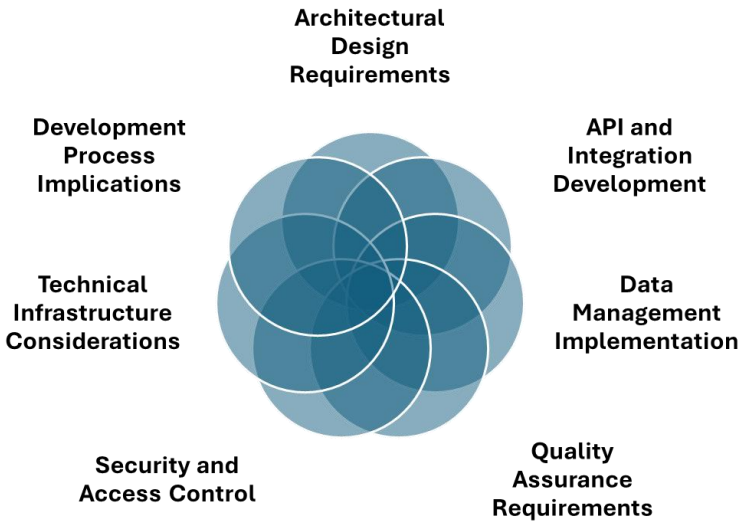


Figure 34 - Implications for Practice

Architectural design requirements include layered architecture, incorporating data, semantic, processing, and reporting layers (Casa et al., 2022). Additionally, there is a requirement for flexible, scalable architectures that can accommodate evolving regulatory requirements (Drvar et al., 2021). Moreover, the integration of real-time or near-real-time processing capabilities is essential through RegOps approaches (Drvar et al., 2021).

API and integration development highlights the need for standardised interfaces and APIs to facilitate data exchange between systems (Jago Flötgen et al., 2020). Additionally, there is a requirement for robust integration capabilities to handle both legacy systems and new data management platforms (Connell, 2020).

Data management implementation involves several key requirements: systems must support granular data collection and processing capabilities (ECB 2016/867). There is a need for sophisticated metadata management systems to maintain data lineage and ensure consistency (Casa, 2024). Additionally, efficient data transformation and aggregation mechanisms are required (Bašić, 2017).

Quality assurance requirements include the implementation of automated validation rules and quality checks (Poloni, 2024), continuous monitoring and automated error-detection systems, and the development of comprehensive data quality metrics and reporting capabilities.

In the realm of **security and access control**, implementing granular access controls is critical to ensuring that only authorised users can access sensitive information. Additionally, the development of secure data transmission protocols is essential for protecting data integrity during transfer. Furthermore, integrating robust security measures throughout the system helps safeguard against potential threats and vulnerabilities.

When **considering technical infrastructure**, several key factors must be addressed. Firstly, there is a need for scalable storage solutions to handle large volumes of granular data. Secondly, high-performance processing capabilities are required to ensure efficient data handling. Lastly, the implementation of reliable backup and recovery systems is crucial for maintaining data integrity and accessibility.

The **development process has several implications** that need to be addressed. Firstly, there is a need for agile development approaches to accommodate changing regulatory requirements (Drvar et al., 2021). Additionally, the importance of continuous integration and deployment practices cannot be overstated. Finally, robust testing frameworks are required to ensure regulatory compliance.

These implications suggest that software engineers in the banking sector must focus on **developing highly flexible, scalable, and integrated systems while maintaining strict compliance with regulatory requirements and data quality standards**. The findings emphasise the importance of standardisation and interoperability in system design and implementation (Colangelo et al., 2021). However, it is important to note that the research does not provide empirical evidence of successful implementations, suggesting that practical validation of these approaches is still needed in real-world banking environments.

6.3 Implications for Future Research

Based on the analysis of the performed research, several key areas emerge as priorities for future work:

Empirical validation studies emphasise the need for real-world implementation studies of the BIRD and IReF frameworks (Colangelo et al., 2021). They focus on evaluating the actual effectiveness of the proposed data models in banking

environments and assessing the practical challenges associated with framework implementation.

The **implementation challenges** include investigating specific technical barriers to integrating diverse reporting requirements (Drvar et al., 2021), analysing the cost-benefit implications for different-sized banking institutions, and examining change management approaches for framework adoption (Poloni, 2024).

Cross-jurisdictional implementation involves several key components. First, there is a need for research on harmonising reporting requirements across different regulatory jurisdictions. Additionally, it is important to investigate country-specific adaptation needs for standardised frameworks, as outlined by Casa (2024). Lastly, an analysis of international data-sharing mechanisms and the associated challenges is essential.

Technical architecture validation involves testing proposed layered architecture approaches in real banking environments, evaluating the effectiveness of RegOps implementation (Drvar et al., 2021), and assessing scalability and performance in large-scale deployments.

Data quality and governance involve several critical components. First, an investigation into effective data quality assurance mechanisms within unified reporting frameworks is conducted. Additionally, it includes an analysis of the effectiveness of governance structures, as highlighted by Casa et al. in "The Integrated Approach by Bank of Italy." Furthermore, the research emphasises automated validation approaches and their reliability in ensuring data integrity.

Regulatory compliance involves several key areas of focus. First, there are studies on the alignment between implemented frameworks and BCBS 239 requirements (Poloni, 2024). Additionally, an investigation into compliance verification methods is necessary. Finally, it is important to analyse the accuracy and timeliness of regulatory reporting.

Integration challenges include research on legacy system integration approaches as outlined in the ECB Internship Report for 2019-2020 (Connell, 2020), an investigation of technical interoperability solutions, and an analysis of data transformation effectiveness.

Performance and efficiency involve studies of the processing efficiency of integrated reporting systems, analyses of resource requirements across different implementation approaches, and investigations of real-time reporting capabilities.

Security and privacy are crucial considerations in today's digital landscape. This involves conducting research on data protection mechanisms within integrated frameworks, investigating the effectiveness of access control measures, and analysing the security of cross-border data sharing practices.

The **impact** of unified reporting frameworks on different **stakeholders** involves several key areas of focus. First, studies examine how these frameworks affect various groups. Additionally, an analysis of the training and adaptation requirements is needed for stakeholders to effectively engage with the frameworks. Finally, the investigation includes assessing user acceptance and the challenges associated with adopting these reporting methods.

Figure 35 - Implications for Future Research

These research areas (Figure 35) are critical, given that the current literature is primarily theoretical (Colangelo et al., 2021) and provides limited empirical evidence of successful implementation. Future research should provide practical validation of the proposed frameworks and address the identified implementation challenges through real-world case studies and empirical analysis.

The research report indicates that while theoretical frameworks are well-developed, a significant gap exists in understanding their practical application and

effectiveness (Casa, 2024). This suggests that priority should be given to empirical studies that can validate these frameworks and provide concrete evidence of their effectiveness in actual banking environments.

6.4 Limitations & Recommendations

Regarding the performed study, there are a few limitations of theoretical nature (some studies are primarily theoretical analyses and framework proposals, lacking empirical evidence of successful implementation in real banking environments), limited empirical validation (none of the reviewed studies provided concrete evidence of successful implementation in actual banking environments) and scope constraints (some studies focused mainly on framework development and conceptual analysis, with limited attention to practical implementation challenges (Colangelo et al., 2021 & 1 others)).

6.4.1 Validation Bias Risk

The current research exhibits several significant validation bias risks that compromise the reliability and generalizability of findings across all three research questions.

Primary Validation Bias Concerns:

1. Absence of Empirical Validation: the research explicitly states that the reviewed studies were primarily theoretical analyses and framework proposals and did not provide empirical evidence of successful implementation. This creates a fundamental validation bias where proposed frameworks lack real-world testing, theoretical benefits cannot be verified against actual outcomes and implementation challenges remain unidentified and unaddressed.

2. Confirmation Bias in Framework Selection: the research shows heavy concentration on specific frameworks, with “BIRD: Most frequently covered (6 out of 11 studies)” and “IReF: Second most common (4 studies)”. This suggests potential confirmation bias where researchers may be selectively focusing on preferred solutions,

alternative approaches receive insufficient consideration and framework limitations may be underexplored.

3. Publication and Selection Bias: the research points out that none of the reviewed studies provided concrete evidence of successful implementation in banking environments. This pattern suggests possible publication bias favouring theoretical over empirical studies, selection bias toward studies supporting integrated reporting frameworks and absence of studies reporting implementation failures or challenges.

Methodological Validation Risks:

4. Lack of Comparative Analysis: the research lacks systematic comparison between proposed frameworks and existing systems, creating validation bias through insufficient baseline performance metrics, missing cost-benefit empirical comparisons and unvalidated assumptions about improvement potential.

5. Stakeholder Validation Gaps The report indicates limited evidence of stakeholder validation, particularly no evidence of banking industry validation of proposed requirements, missing regulatory authority endorsement of compliance claims and absence of end-user testing or feedback.

BCBS 239 Compliance Validation Bias:

6. Assumed Compliance Without Testing: while the report suggests frameworks align with some BCBS 239 principles, it acknowledges the studies did not provide empirical evidence. This creates validation bias where theoretical alignment is assumed to equal practical compliance, implementation gaps in BCBS 239 adherence remain unidentified and compliance claims lack independent verification.

6.4.2 Technical Recommendations

Once stated the previously identified limitations in multiple dimensions, there are however some recommendations that can be provided, namely:

- **Technical Implementation**: develop robust IT infrastructure for data integration (Casa, 2024), implement standardised interfaces and APIs for data

exchange and establish real-time or near-real-time processing capabilities (Drvar et al., "The Future of Data Collection")

- **Governance and Standardisation**: create centralised oversight mechanisms (Poloni, 2024), establish clear data ownership and responsibilities (Casa et al., "The Integrated Approach by Bank of Italy") and implement comprehensive data quality assurance measures (Bašić, 2017)
- **Future Research and Development**: conduct empirical studies to validate framework effectiveness, test implementations in real banking environments and evaluate the practical application of proposed frameworks across different jurisdictions (Colangelo et al., 2021)
- **Regulatory Alignment**: ensure ongoing alignment with evolving regulatory requirements (Poloni, 2024), develop mechanisms for interpreting and implementing new regulations (Casa, 2024) and harmonise reporting requirements across different jurisdictions (Colangelo et al., "How Integrated Reporting by Banks")
- **Stakeholder Collaboration**: foster collaboration between regulatory authorities and banks (Casa et al., "PUMA Cooperation Between the Bank of Italy"), establish continuous dialogue for refining reporting requirements (Poloni, 2024) and develop shared approaches to data standardisation (Jago Flötgen et al., "Digital Platforms Boundary Resources")

6.5 Contributions of the Study

The study makes several significant contributions to software engineering practice and research, in several distinct areas (Figure 36).

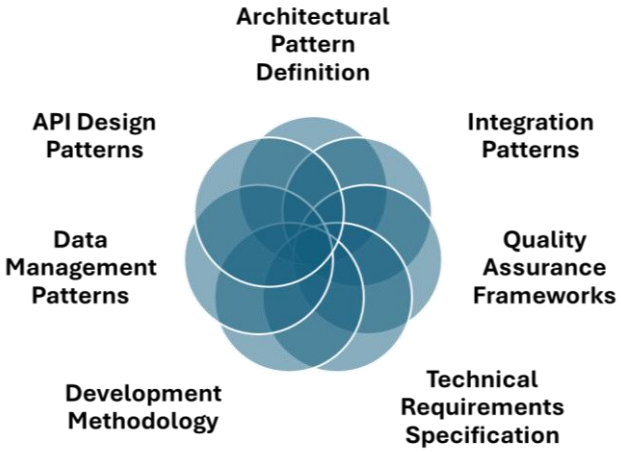


Figure 36 - Contributions of The Study

The **architectural pattern** is defined by establishing a validated four-layer architectural pattern specifically for financial reporting systems, as highlighted by Casa et al. in "The Integrated Approach by Bank of Italy." It provides a detailed framework for implementing data, semantic, processing, and reporting layers, as noted in the ECB Internship Report for 2019-2020 (Connell, 2020). Additionally, it introduces RegOps as an agile architectural approach for regulatory reporting systems, discussed by Drvar et al. in 2021.

Integration patterns define standardised approaches to system integration in financial environments (Jago Flötgen et al., 2020), establish frameworks for integrating legacy systems with modern data platforms (Casa, 2024), and provide models for implementing cross-system data exchange protocols (Connell, 2020).

The **quality assurance frameworks** introduce comprehensive validation frameworks for financial data systems (Poloni, 2024), establish patterns for automated quality control implementation, and define approaches for continuous monitoring and validation (Bašić, 2017).

The **technical requirements specification** includes several key components: real-time data processing systems, data transformation engines, integration interfaces, and security implementations (Drvar et al., 2021).

The **development methodology** introduces RegOps as an agile approach specific to regulatory software development (Drvar et al., 2021). It provides frameworks for

implementing continuous integration in regulatory contexts and establishes patterns for iterative development of compliance systems.

Data management patterns define patterns for implementing granular data management systems, as noted in the Regulation (EU) 2016/867 of the ECB, 2016. They establish frameworks for metadata management (Casa, 2024) and also provide models for data lineage implementation.

API design patterns establish standardised interface patterns for financial data exchange (Jago Flötgen et al., 2020). They define frameworks for API implementation in regulatory contexts and provide models for secure data transmission.

The primary value of software engineering lies in establishing specialised patterns and frameworks for financial regulatory systems, though practical validation through implementation case studies remains needed (Colangelo et al., 2021). The study provides a foundation for future empirical research in this domain.

One important question to keep in mind is: from the banks' perspective, what are the main advantages (Table 37) of adopting a unified reporting framework, as described in the present study (using BIRD and IReF)?

Based on current research, there are several **key advantages for banks** in adopting unified reporting frameworks such as BIRD and IReF.

Table 37 - Unified Reporting Key Advantages for Banks

Key Benefit	Description
Reduced Reporting Burden	Integration of multiple reporting requirements into a single standardised framework reduces redundancy (Colangelo et al., "How Integrated Reporting by Banks") and eliminates the need to report the same information multiple times in different formats (Casa et al., "PUMA Cooperation Between the Bank of Italy")
Improved Data Organisation and Efficiency	BIRD helps banks organise their internal data systems more effectively (Colangelo et al., 2021) and provides standardised transformation rules that simplify data processing and reporting (Connell, 2020), while a standard data dictionary reduces inconsistencies and interpretation errors (Poloni, 2024)
Cost Reduction	Standardisation and automation reduce operational costs associated with reporting (Drvar et al., 2021); this integrated approach minimises

	the need for multiple reporting systems and processes (Casa et al., "The Integrated Approach by Bank of Italy")
Enhanced Data Quality	Harmonised data models and standardised definitions improve data consistency (Bašić, 2017), while automated validation rules and quality checks reduce errors (Casa et al., "PUMA Cooperation Between the Bank of Italy")
Better Adaptability to New Requirements	A unified framework makes it easier to implement new reporting requirements (Casa, 2024) while representing a standardised approach that facilitates faster response to regulatory changes (Colangelo et al., 2021)
Cross-border Operations Support	A harmonised reporting framework particularly benefits banks operating across multiple jurisdictions (Colangelo et al., 2021), and it also reduces complexity in managing different national reporting requirements (Jago Flötgen et al., 2020)

This study indicates these advantages are theoretical proposals, as none of the studies provided empirical evidence of successful implementation in real banking environments.

References

AnaCredit Reporting Manual - Part II - Datasets and Data Attributes. (2017), from https://www.ecb.europa.eu/pub/pdf/other/AnaCredit_Manual_Part_II_Datasets_and_data_attributes.en.pdf

Antonio Colangelo, Catherine Ahsbahs, Davide Continanza, Violetta Damia, Corinne Devillers, Francis Gross, Paul Hiebert, Angelo Saponara, Florian Schuster, and Jean-Marc Israël. "How Integrated Reporting by Banks May Foster Sustainable Finance?", 2021

Antonio Colangelo, Francis Gross, and Florian Schuster. "Effective Measurement of the Economy in the Emerging Digital Age Effective Measurement of the Economy in the Emerging Digital Age", 2021.

Bank, E. C. (2024). Making banks' data reporting more efficient. https://www.ecb.europa.eu/stats/ecb_statistics/reporting/html/index.en.html

Banks' Integrated Reporting Dictionary. (no date). April 10th, 2024, from https://bird.ecb.europa.eu

Bundesbank - Guidelines on Credit Data Statistics (AnaCredit). July 1st, 2025, from https://www.bundesbank.de/resource/blob/803824/bc6d190b29937d944b5517067fd9deaf/472B63F073F071307366337C94F8C870/anacredit-guidelines-data.pdf

Capital Requirements Regulation (CRR). (no date). April 10th, 2024, from https://eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/12674

Connell, M. P. (2020). Development of the Bird: A Metadata Modelling Approach for the Purpose of Harmonising Supervisory Reporting at the European Central Bank-Directorate of General …. search.proquest.com. https://search.proquest.com/openview/62440693e8689ce517f20bc3e2517e61/1?pq-origsite=gscholar\&cbl=2026366\&diss=y

Dorrer, D. (2024). Empirical Analysis. In D. Dorrer (Ed.), Introduction of IReF: The Development of an Integrated System for Achieving Harmonised Supervisory Reporting in the European Union (pp. 45–85). Springer Fachmedien. https://doi.org/10.1007/978-3-658-46447-9_4

ECB - AnaCredit Q&A and Manual. (no date). September 1st, 2025, from https://www.ecb.europa.eu/stats/ecb_statistics/anacredit/questions/

Flötgen, R., Gomm, S., Böhm, M., & Krcmar, H. (2020). Implications of Emerging Financial Regulatory Reporting Frameworks for Digital Platforms Boundary Resources. AMCIS 2020 Proceedings. https://aisel.aisnet.org/amcis2020/accounting_info_systems/accounting_info_systems/2

Garousi, V., Felderer, M., & Mäntylä, M. V. (2019). Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. Information and Software Technology, 106, 101–121. https://doi.org/10.1016/j.infsof.2018.09.006

Ines Bašić. "Supervisory and Statistical Granular Data Modelling at the Croatian National Bank", 2017.

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. Keele, UK, Keele Univ., 33.

Martina Drvar, Johannes Turner, Maciej Piechocki, Regtech, Eric Stiegeler, Bearingpoint Regtech, and Daniel Münch. "The Future of Data Collection and Data Management: Agile RegOps for Digitalising the Regulatory Value Chain", 2021

Massimo Casa. "Connecting The Dots of The International Debate on The Standardization and Granularity of Regulatory Data". *Social Science Research Network*, 2024.

Massimo Casa, Laura Graziani Palmieri, Laura Mellone, and Francesca Monacelli. "The Integrated Approach Adopted by Bank of Italy in the Collection and Production of Credit and Financial Data", 2022

Massimo Casa, Marco Carnevali, Silvia Giacinti, and Roberto Sabatini. "Puma Cooperation Between the Bank of Italy and The Intermediaries for The Production of Statistical, Supervisory and Resolution Reporting", 2022

Paolo Poloni. “The Evolution of the Supervisory Reporting Framework for the EU Banking Sector”. *Social Science Research Network*, 2024.

Penczar, M., Nowak, S., & Liszewska, M. (2023). Playing by the rules: Do the post-crisis regulations influence banks’ funding model in the EU? International Journal of Finance & Economics, 28(4), 4572–4591. https://doi.org/10.1002/ijfe.2665

REGULATION (EU) 2016/867 OF THE EUROPEAN CENTRAL BANK - of 18 May 2016 - On the collection of granular credit and credit risk data (ECB/ 2016/ 13). (n.d.).

Rob Jago Flötgen, Sebastian Gomm, Markus Böhm, and Helmut Krcmar. “Digital Platforms Boundary Resources for Financial Reporting Frameworks”, 2020

Appendixes

Appendix 1 – Literature Search Reports

Search Reports on Google Scholar:

Query String: banks integrated reporting dictionary [title]

Query String: banks integrated reporting dictionary

Query String: "bird" "iref" "banks" "reporting" [only review]

Query String: "bird" "iref" "banks" "reporting" "dictionary"

Search Reports on Semantic Scholar:

Query String: banks integrated reporting dictionary

Query String: "bird" "iref" "banks" "reporting" "dictionary"

Query String: bird iref

Appendix 2 – Aggregated Formal and Grey Literature List with Classification

Source	Year	Classification	Author	Title	Comments
	2016	Book	Pica, C. A.	Regulação e supervisão: alavanca ou travão à excelência do desempenho nos bancos?	
Journal of Securities Operations &Custody	2016	Journal Article	Visnovsky, F.	Examining the future of regulatory reporting for the banking industry	
	2017	Book	Gdovin, F.	Graphical Representation of Data Lineage in the Data Governance Tool	
	2017	Document	Bašić, Ines	Supervisory and statistical granular data modelling at the Croatian National Bank	
	~~2017~~	~~Document~~	~~Scapeccia, Andrea~~	~~The Evolution in Banks' Reporting~~	~~No download available~~
~~Journal of Securities Operations & ...~~	~~2017~~	~~Journal Article~~	~~Broersen, P.; Koppen, R.~~	~~Preparing for AnaCredit: A timely start is crucial~~	~~Pending author(s) feedback~~
~~Journal of Securities Operations &Custody~~	~~2017~~	~~Journal Article~~	~~Dorval, M.~~	~~AnaCredit: Why it matters to look at other regulations in context~~	~~Pending author(s) feedback~~
McKinsey & …	2017	Journal Article	Harreis, H.; Tavakoli, A.; Ho, T.; Machado, J.; ...	Living with BCBS 239	
ECB Occasional Paper	2017	Journal Article	Israel, J. M.; Damia, V.; Bonci, R.; Watfe, G.	The Analytical Credit Dataset. A magnifying glass for analysing credit in the euro area	

Bank of England Quarterly Bulletin	2017	Journal Article	Manning, M.	Financial stability requires industry-wide collaboration on data standards: G20 economies promote data sharing through collaboration	
pdfs.semanticscholar.org	2018	Journal Article	Data, M.	Item II. 1 The Integrated Reporting Framework	
Journal of Securities Operations &Custody	2018	Journal Article	Israel, M.; Colangelo, A.	Daryn Moody	
	2018	Book	Monetary, International Monetary Fund; ...	Euro Area Policies: Financial Sector Assessment Program-Technical Note-Detailed Assessment of Observance of Basel Core Principles for Effective Banking …	
~~Journal of Securities Operations & Custody~~	~~2018~~	~~Journal Article~~	~~Bier, Werner; Israël, Jean-Marc; Colangelo, Antonio; Bonci, Riccardo~~	~~Analytical credit dataset, the integrated reporting framework and the banks' integrated reporting dictionary: Do we overshoot? Or do we undershoot?~~	~~Pending author(s) feedback~~
~~On the Move to Meaningful …~~	~~2018~~	~~Journal Article~~	~~Lemmens, I.; Laar, B. van de; Saton, J.; Bulles, J.~~	~~How to fulfil regulatory requirements consistently: a semantic-based approach~~	~~Pending author(s) feedback~~
~~Journal of Financial Compliance~~	~~2019~~	~~Journal Article~~	~~Hauet, Olivia~~	~~The AnaCredit regulation is in place — a reflection on data-driven reporting: What has been achieved by the banks and regulators, and what remains to be done?~~	~~Pending author(s) feedback~~

Available at SSRN 3407188	2019	Journal Article	Mishra, D. R. N.	Technology and Analytics in the Wonderland of Risks Monitoring A Regulatory Standpoint	
~~Journal of Securities Operations &Custody~~	~~2019~~	~~Journal Article~~	~~Shah, B.~~	~~The road to making regulation more efficient: A case study in the application of best practices and data standards in regulatory reporting~~	~~No download available~~
	2020	Book	Connell, M. P.	Development of the Bird: A Metadata Modelling Approach for the Purpose of Harmonising Supervisory Reporting at the European Central Bank-Directorate of General …	
	2020	Book	Flötgen, R. J.; Gomm, S.; Böhm, M.; Krcmar, H.	Implications of emerging financial regulatory reporting frameworks for digital platforms boundary resources	
	2020	Book	Kolahgar, S.	Essays on Investor Relations and Stakeholder Communication in Corporate Finance	
	2020	Book	Michailidou, F.	RegTech and SupTech: Opportunities and Challenges in the Financial Sector	
Bank of Italy Occasional Paper	2020	Journal Article	Bonis, R. De; Piazza, G. M.	A silent revolution: How central bank statistics have changed in the last 25 years	

BearingPoint …	2020	Journal Article	Drvar, M.; Turner, J.; Piechocki, M.; ...	The future of data collection and data management: Agile RegOps for digitalising the regulatory value chain	
	2021	Book	Boggavarapu, S.	The Effect of Third-Party Service Providers on Information Security Breaches at Financial Institutions	
Becoming a data-centric organisation: A guide to …	2021	Journal Article	Baravykas, R.; Šaltenytė, U.	Becoming a data-centric organisation: a guide to data management initiatives at the Bank of Lithuania	
Proceedings 63rd ISI World Statistics …	2021	Journal Article	Colangelo, A.; Gross, F.; Schuster, F.	Effective measurement of the economy in the emerging digital age	
Journal of Digital Banking	2021	Journal Article	Israël, J. M.; Tissot, B.	Incorporating microdata into macro policy decision-making	
	2022	Book	Iwanicz-Drozdowska, M.; Malinowska-Misiąg, E.; ...	The Role of Crises in Shaping Financial Systems: From the Global Financial Crisis to COVID-19	
	2022	Document	Colangelo, A.; Israël, Jean-Marc; Ahsbahs, Catherine; Continanza, D.; Damia, Violetta; Devillers, Corinne; Gross, Francis; Hiebert, P.; Saponara, Angelo; Schuster, Florian	How integrated reporting by banks may foster sustainable finance? 1	

Bank of Italy …	2022	Journal Article	Casa, M.; Palmieri, L. Graziani; Mellone, L.; ...	The integrated approach adopted by Bank of Italy in the collection and production of credit and financial data	
Bank of Italy Occasional …	2022	Journal Article	Casa, M.; Sabatini, R.; Carnevali, M.; ...	PUMA cooperation between the Bank of Italy and the intermediaries for the production of statistical, supervisory and resolution reporting	
Bank of Italy …	2022	Journal Article	D'Alessio, G.; Bonis, R. De; Infante, L.; ...	Statistics for economic analysis: the experience of the Bank of Italy	
The Governance of Insurance …	2022	Journal Article	Georgosouli, A.; Okonjo, J.	The Algorithmic Future of Insurance Supervision in the EU: A Reality Check	
~~Artificial Intelligence and Financial Behaviour~~	~~2023~~	~~Journal Article~~	~~Amzallag, A.~~	~~54,000 PRIIPs KIDs how to read them (all)~~	~~Pending author(s) feedback~~
IFC Bulletin	2023	Journal Article	Araujo, D.; Bruno, G.; Marcucci, J.; Schmidt, R.; Tissot, B.	Data science in central banking: applications and tools	
Bank of Italy Occasional Paper	2023	Journal Article	Casa, M.	Connecting the dots of the international debate on the standardisation and granularity of regulatory data	
~~Journal of Financial Compliance~~	~~2023~~	~~Journal Article~~	~~Hauet, O.~~	~~Reviving securitisation in the EU: A critical analysis of the reporting requirements~~	
~~Journal of Digital …~~	~~2024~~	~~Journal Article~~	~~Araujo, D.; Bruno, G.; Marcucci, J.; …~~	~~Data science in economy and finance: A central bank perspective~~	

~~BANKING RESILIENCE: New …~~	~~2024~~	~~Journal Article~~	~~Breymann, H. E.; Hauf, P.; Künzle, C.~~	~~Venturing into New Ways of Regulatory Reporting and Systemic Risk Analysis~~	~~Pending author(s) feedback~~
~~Introduction of IReF: The Development of an Integrated …~~	~~2024~~	~~Journal Article~~	~~Dorrer, D.~~	~~Empirical Analysis~~	~~Pending author(s) feedback~~
~~… for Achieving Harmonised Supervisory Reporting in …~~	~~2024~~	~~Journal Article~~	~~Dorrer, D.~~	~~Analysis of the Expert Interviews~~	~~Pending author(s) feedback~~
ECB Occasional Paper	2024	Journal Article	Poloni, P.	The evolution of the supervisory reporting framework for the EU banking sector	
~~books.google.com~~	~~2024~~	~~Journal Article~~	~~Dorrer, D.~~	~~Introduction of IReF: The Development of an Integrated System for Achieving Harmonised Supervisory Reporting in the European Union~~	~~Pending author(s) feedback~~

Appendix 3 – Extraction Matrix

A comprehensive matrix for the studies analysed during research, can be used to provide a better accuracy assessment for the study.

Study	Source	Objective	Method	Scope	Frameworks	Evidence	Metrics	Limitations
Regulation (EU) 2016/867 of the ECB, 2016	European Central Bank	Establish shared multipurpose database for granular credit data collection	Framework development; document analysis, stakeholder consultation, empirical data gathering through workshops	European banking sector credit data	AnaCredit, BIRD, European Reporting Framework (ERF)	Stakeholder workshops and fact-finding exercises	Not specified	Not explicitly mentioned
Casa, 2024	Bank of Italy	Emphasize need for common data dictionary and integrated reporting system	Conceptual analysis and policy review; document analysis, institutional analysis	EU banking regulatory reporting	IReF, BIRD, CRR III, Digital Regulatory Reporting (DRR), PUMA	Document analysis of existing frameworks	Not specified	Not explicitly mentioned
Colangelo et al., "How Integrated Reporting by Banks"	European Central Bank	Discuss IReF and BIRD as key initiatives for standardizing and integrating reporting requirements	Conceptual analysis and framework development; document analysis, policy review	Euro area banking statistical reporting	IReF, BIRD	Collaborative development by ESCB, EBA, and financial authorities	Not specified	Not explicitly mentioned
Colangelo et al., 2021	European Central Bank	Propose radical standardization across countries with central banks as catalysts	Conceptual analysis and framework development; theoretical exploration, stakeholder dialogue	Global banking data standardization	IReF, BIRD	Scientific observation combined with design thinking	Not specified	Not explicitly mentioned

Study	Source	Objective	Method	Scope	Frameworks	Evidence	Metrics	Limitations
Connell, 2020	European Central Bank	Focus on BIRD project and Single Data Dictionary for harmonizing data models across banks	Framework development; relational database modelling, document analysis	European banking supervisory reporting	BIRD, Single Data Dictionary (SDD)	Technical analysis of database structures	Not specified	Not explicitly mentioned
Jago Flötgen et al., "Digital Platforms Boundary Resources"	Academic research	Propose standardized reporting framework as a public good	Framework development with ecosystem modelling; e3-value method, qualitative content analysis, expert interviews	Digital platforms for financial reporting	BIRD, IReF, Data Point Model (DPM)	Expert interviews, ecosystem modelling	e3-value method metrics	Not explicitly mentioned
Bašić, 2017	Croatian National Bank	Present CNB BIRD as standardized framework for data extraction and transformation	Framework development; system design, document analysis, metadata rule development	Croatian banking sector granular data	CNB Banks' Integrated Reporting Dictionary	System implementation in Croatia	Not specified	Not explicitly mentioned
Casa et al., "PUMA Cooperation Between the Bank of Italy"	Bank of Italy	Highlight PUMA as integrated approach to data collection and management	Conceptual analysis and policy review; document analysis, institutional analysis	Italian banking data collection	PUMA, BIRD, IReF, BCBS 239	Metadata-driven process implementation	Not specified	Not explicitly mentioned
Casa et al., "The Integrated	Bank of Italy	Propose integrated approach to data	Framework development; conceptual analysis, matrix model, data-driven model	Italian banking credit and financial statistics	Bank of Italy's integrated	Matrix model and data-driven model implementation	Not specified	Not explicitly mentioned

Study	Source	Objective	Method	Scope	Frameworks	Evidence	Metrics	Limitations
Approach by Bank of Italy"		governance with single statistical dictionary			approach, BIRD, IReF			
Drvar et al., "The Future of Data Collection"	Austrian Financial Market Authority	Introduce RegOps as agile approach to regulatory reporting	Conceptual analysis with framework development; case studies, qualitative analysis	Regulatory reporting digitalization	RegOps, CNB BIRD, Austrian Reporting Services (AuRep)	Case studies of existing practices	Big data platform metrics	Not explicitly mentioned
Poloni, 2024	European Banking Authority	Discuss development of common data dictionary and joint governance under JBRC	Policy review and institutional analysis; document analysis	EU supervisory reporting framework	EBA reporting framework, BCBS 239, BIRD, IReF	Document analysis of existing reports and strategies	Not specified	Not explicitly mentioned

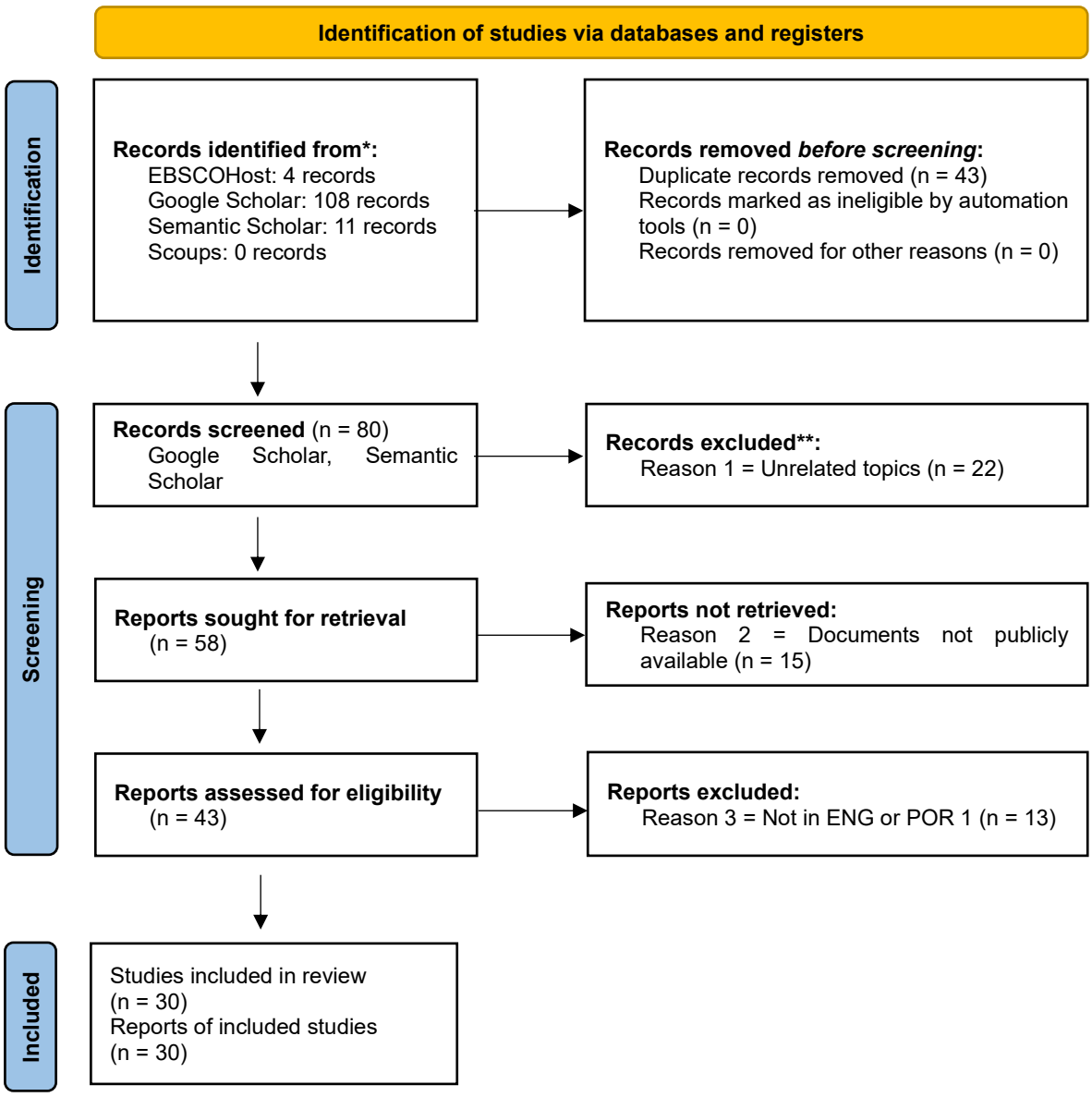
Key Observations from the Matrix:

1. **Source Distribution**: most studies originate from central banks and regulatory authorities (ECB, Bank of Italy, CNB, EBA), with one academic study.
2. **Common Objectives**: all studies focus on standardization, integration, and harmonization of banking reporting requirements.
3. **Methodological Patterns**: framework development: most common (4 studies); conceptual analysis with policy review or framework development (4 studies).
4. **Framework Convergence**: BIRD: most frequently covered (6 out of 11 studies) and IReF: second most common (4 studies).
5. **Evidence Limitations**: none of the reviewed studies provided concrete evidence of successful implementation in banking environments.
6. **Metrics Gap**: lack of quantitative metrics across studies, noting the need for real-world testing of these frameworks through empirical studies.

This matrix reveals that while the studies provide comprehensive theoretical frameworks, they lack empirical validation and quantitative metrics for measuring implementation success.

Appendix 4 – PRISMA 2020 Flow Diagram

The following PRISMA 2020 flow diagram provides a transparent overview of how many records were retrieved from different sources, how duplicates were removed, and how many studies were excluded at each stage of the current study's literature review.



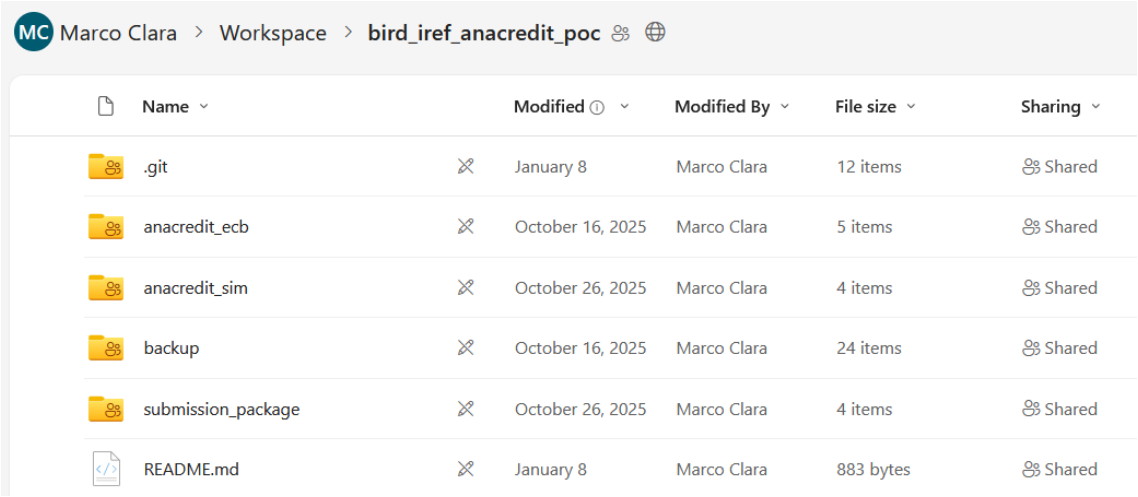
*Consider, if feasible to do so, reporting the number of records identified from each database or register searched (rather than the total number across all databases/registers).
**If automation tools were used, indicate how many records were excluded by a human and how many were excluded by automation tools.

Source: Page MJ, et al. BMJ 2021;372:n71. doi: 10.1136/bmj.n71.

Appendix 5 – Proof of Concept Source Code

BIRD & IReF – Anacredit Proof of Concept Source Code:

Link: [OneDrive]



Alternative Link: [GitHub]