EUROPEAN DATAWAREHOUSE

September 2016

Special Report

# Quality of loan level data - lessons learned for the AnaCredit project


## Authors

**Markus Schaber**
Managing Director
Markus.Schaber@eurodw.eu

**Dr. Christian Thun**
Managing Director
Christian.Thun@eurodw.eu

European DataWarehouse GmbH
Walther-von-Cronberg-Platz 2
60594 Frankfurt am Main
www.eurodw.eu


The centralised collection of granular credit data and credit risk information at the individual loan level is not a very recent innovation in Europe. At national level this information has already been gathered for decades by several national central banks (e.g. in France since 1946, in Italy since 1964).[1] What is new, though, is that with the Analytical Credit Dataset (AnaCredit) project, the European Central Bank (ECB) will create a pan-European collection of loan level data for exposures above €25,000 in a standardised format from 2018 onward that will enable comparability across European loan portfolios.

One of the main challenges that this exercise – like any other data collection effort – will face is to achieve and maintain good and consistent data quality delivered by the financial institutions (also called reporting agents). Given the relatively high complexity of AnaCredit, the data quality dimension might become particularly relevant in a pan-European context given the limited standardisation at this stage.

In order to assess how these challenges might emerge going forward, it is worthwhile to compare AnaCredit with other pan-European data initiatives. A good example is the European DataWarehouse (ED) which was founded in 2012 in the context of the ECB's ABS loan level initiative as the first central data warehouse in Europe for the collection, validation and dissemination of standardised and asset class specific loan level data for Asset-Backed Securities (ABS) transactions. ED's main objectives are to improve transparency in order to enable better risk assessment of underlying portfolios and to remove information asymmetries between investors and rating agencies. Ultimately, enhanced transparency should help to restore confidence in the European ABS market. As part of its mandate, ED checks the collected loan level data for completeness and correctness. Given the high level of resources, systems and processes deployed for data quality management, ED has gathered practical knowledge on data quality issues which might be of relevance for the AnaCredit project.

[1] See Jentzsch, N.; Do We Need a European Directive for Credit Reporting?; p. 50.

## Differences and similarities between AnaCredit and the ECB's ABS loan level initiative

Unlike for AnaCredit, there are no formal regulatory reporting obligations for the ABS loan level initiative. Submitting the data to ED and the subsequent publication to the ABS market is voluntary, albeit the intention to make an individual ABS transaction eligible as collateral for the Eurosystem repo framework requires the upload and publication of the loan level data. Consequently, there is a strong economic incentive for loan level reporting. Other key differences are the complexity, volume of preparatory work and longer phase-in periods for AnaCredit whereas for the ABS loan level initiative phase-in periods were comparatively short.

The main similarity between AnaCredit and the ECB's ABS loan level initiative is the collection of very granular data (i.e. loan-by-loan or instrument-by-instrument) by financial institutions across Europe but with a focus on the Eurozone. For both initiatives, uniform and standardised definitions across Europe for the reported data items (so called 'attributes') are key elements. As part of the ABS loan level initiative, standardised data templates for different asset classes and corresponding data manuals (so called 'taxonomies') were jointly developed by representatives of the ABS industry, the ECB and national central banks with the ECB taking a leading role. The taxonomies contain explanations for each attribute including some examples. National characteristics of particular data fields have been taken into account where differences across countries were too significant to have a fully uniform definition. Each data template comprises mandatory as well as optional fields. Optional fields are not required to achieve eligibility of the ABS transaction in the Eurosystem collateral framework but should, nevertheless, help in the overall transparency effort. In addition, further clarifications and examples are discussed on a specific ECB website in the form of "frequently asked questions (FAQ)". Moreover, for specific questions raised by a reporting agent, the ECB maintains a helpline for individual clarifications.

Given the complexity of the AnaCredit project, a number of related projects have been set up in order to prepare financial institutions for the reporting requirements. The 'Bank's Integrated Reporting Dictionary (BIRD)' project is particularly relevant as it examines the integration of data from various source systems into a clearly defined output layer. The BIRD project aims to develop a compendium with very detailed definitions for the input positions required for AnaCredit including a description of the necessary transformation that will translate input data into the required uniform data field definition where necessary. Achieving a high level of data quality and data consistency are the main objectives of BIRD. However, since adhering to the BIRD requirements is voluntary, it remains to be seen how many financial institutions will implement the suggestions in time for the launch date given the potentially significant investments needed at reporting agent level.

Compared to the ABS loan level initiative, the AnaCredit project will require far more substantial efforts in terms of time and resources given the significantly larger scope. Current projects such as BIRD, as well as the development of a comprehensive AnaCredit Manual, will help to formulate a detailed set of guidelines and definitions and will hence clarify many questions in

advance. Nevertheless the following examples of the ABS loan level initiative experience will still be relevant for the AnaCredit project[2]:

- Different national accounting standards
- Different interpretations at national level of specific regulatory or accounting standards
- Regulatory requirements have not yet fully converged at European level and can require multiple and diverse data capture at national level for financial institutions
- Unavailability of data due to outdated IT systems and non-collection of certain data fields in the past, including different availability of and access to data within financial institutions due to mergers and acquisitions
- Diverging loan servicing practices, specifically in relation to the management of defaults and distressed loans which, in turn, prevent standardised reporting
- Data errors given incorrect data capture or handling

Given the various challenges European DataWarehouse has faced since becoming operational, a comprehensive data quality management process had to be installed and was further extended in recent years. The key data quality issues found across many portfolios will be highlighted in the next sections.

## Data quality management as part of the ABS Loan Level Initiative

While there are many different reasons for data quality issues, in general they can be allocated to one of the following three categories:

1. Data quality problems due to insufficient clarity of definitions of data fields
2. Data quality problems due to erroneous data entries
3. Data quality problems due to inconsistencies of the data field content

[2] See also for comparison the 'EBF comments on the ECB draft regulation on AnaCredit' with respect to potential AnaCredit challenges.

## 1. Data quality problems due to insufficient clarity of definitions of data fields

As outlined, the ECB provides and publishes standardised data templates for each asset class and dedicated manuals (taxonomies) for each template on its website.[3]

The taxonomies are meant to help facilitate issuers in the process of mapping bank-internal data to the required data fields in the data templates and to avoid potential misunderstandings. However, despite the fact that the taxonomies contain comprehensive explanations at field level, there can still be open questions for certain situations which are not clearly addressed in the taxonomy and which require further guidance. Since a taxonomy – as detailed as it might be – cannot ever cover all possible eventualities, the only way to solve this problem is to prepare a continuously growing collection of case-by-case questions and answers.

Given the large number of questions it receives, the ECB maintains a dedicated webpage providing a list of 'frequently asked questions (FAQs)' concerning the data templates and other reporting issues.[4]

Two examples from the taxonomy for Residential Mortgage Backed Securities (RMBS) shall illustrate the above mentioned problems.

| | |
|---|---|
| Field name and number | Foreign National (AR16) |
| Status of the field | Static |
| Explanation in the taxonomy | Indicating whether the borrower is a national of the country in which the property and mortgage loan resides. If no data available use the following input ND (No Data) |
| Required input | Yes / No |

In this first example the field name is not consistent with the explanation in the taxonomy. The field name creates the impression that information is gathered for foreign borrowers while the explanation requires a confirmation for domestic clients. The ECB clarified this point on its website as follows:

| | |
|---|---|
| Field name and number | Foreign National (AR16) |
| Frequently asked question | How should this field be completed? |
| Explanation ECB | The field name is only intended to be informative and the data provider should follow the instructions in the field definition and criteria.<br>If the borrower is a national of the country in which both the property and mortgage loan reside (e.g. an Italian borrower with an Italian law mortgage over a property in Italy), then the appropriate response would be "Y". |

[3] See European Central Bank, Data templates, https://www.ecb.europa.eu/mopo/assets/loanlevel/transmission/html/index.en.html.
[4] See European Central Bank, Frequently Asked Questions, https://www.ecb.europa.eu/mopo/assets/loanlevel/faq/html/index.en.html.

The second example refers to the data field „Default or Foreclosure" (AR 177).

| | |
|---|---|
| Field name and number | Default or Foreclosure (AR 177) |
| Status of the field | Dynamic |
| Explanation in the taxonomy | Total default amount before the application of sale proceeds and recoveries. If no data available use the following input ND |
| Required input | 12-digit numerical value incl. two decimals |

This field contains the outstanding amount of a defaulted loan before it is offset with the recoveries from collateral foreclosure. The problem with this field definition is the requirement for the amount to be dynamic. The outstanding amount before offsetting with sale proceeds and recoveries is, however, meant to be a static value at a certain point in time. The ECB clarified this point on its website as follows:

| | |
|---|---|
| Field name and number | Default or Foreclosure (AR 177) |
| Frequently asked question | Is this field dynamic or static? |
| Explanation ECB | Although this field is labelled as dynamic, once it has been populated (at the point of default), the value should remain unchanged thereafter. |

In total, the ECB provided detailed answers for more than 350 individual data template-related questions on its website. [5]

With regard to the much larger number of reporting institutions, as well as AnaCredit's much wider coverage of the lending business, similar issues cannot be fully avoided despite significant definition clarifications upfront. An example of this is the data attribute "Probability of Default (PD)" which is collected as part of the Counterparty Risk Data.[6] While the input format is clearly defined (numeric value between 0 and 1) and there is reference to a regulatory definition, there is no specific date when the PD was last assessed or the related period. While arguably it should be always the current one as of the last data submission and hence changes could be tracked based on the succession of submissions. However, the actual review frequency and date can only be implied if changes occur. Therefore, it might be more difficult to derive meaningful transition matrices.

[5] See European Central Bank, Frequently Asked Questions, https://www.ecb.europa.eu/mopo/assets/loanlevel/faq/html/index.en.html.

[6] See European Central Bank, Regulation (EU) 2016/867 of 18 May 2016 on the collection of granular credit and credit risk data (ECB/2016/13), Annex IV.

## 2. Data quality problems due to erroneous data entries

European ABS issuers submit the loan level data via the populated data templates to the European DataWarehouse. The actual upload into the database follows the xml syntax and, during the upload process, a series of data formatting checks (so called syntax / schema checks) are performed. Should the data not meet these formal criteria the data upload will be rejected. Examples of these formal errors are:

- Entries as text instead of numeric values
- Incorrect date formats (e.g. 12-2015 instead of 2015-12)
- Values outside predefined ranges, e.g. in list fields (e.g. data entries can range from "1" to "6" but the actual entry is "7" or the data entry should be "Y" or "N" but the entry is "X")
- Data entry "ND"(No Data) even though this option is not possible according to the taxonomy
- Ignoring the required syntax (e.g. currency codes according to ISO 4217, regional names according to NUTS or industry classification according to NACE Rev. 2)

Irrespective of these formal checks the data has to be further checked for data errors. While many of these errors can be easily identified at the individual loan level, the high number of loans in any given transaction (there are more than 50m loans in the database of the European DataWarehouse as of July 2016) requires a multi-step data quality analysis. Typical examples for these data errors are:

- Missing decimal point: data entry 10000000 instead of 100000.00
- Use of proxies or dummies, e.g. borrower's primary income "0.00" for 80% of all loans
- Data entry "ND5" (no data – data not relevant for this loan) even though information is undoubtedly relevant
- Duplication of loans with the same loan identifier

For the AnaCredit data collection process, the ECB, as well as the national central banks, provide data templates with up to 95 attributes per loan. Since the information that will be collected by AnaCredit for many data fields will be identical to that collected by the ABS loan level initiative (e.g. regional names according to NUTS or industry classification according to NACE Rev. 2, annual turnover or interest margin), it is likely that AnaCredit will face similar data quality problems due to erroneous data entries.

## 3. Data quality problems due to inconsistencies of the data fields content

A comprehensive and focused data quality analysis does not only check for compliance with criteria at data field level but also questions the content-related consistency of the data. In the case of loan level data such assessment comprises the reconciliation of individual fields in a data template. In addition, it should be checked if the data entry has been in principle correct (i.e. the data correctly reflects the underlying fact) but the taxonomy requires a different definition. In such cases, a transformation would then be necessary.

The check for content-related consistency takes into account that a single loan is described through numerous data fields / attributes. In order to assess the reliability of the individual information, data must be organised in a logical context and the content must then be verified. For example in the case of RMBS once the amortisation of a loan is made through one single payment at the end of the loan term (field AR72 Payment Type, data entry: "6" Bullet), there should be a corresponding data entry in the field for the repayment method (field AR69 data entry should be "1" for Interest only). Another example in the case of SMEs is the allocation of collateral. A loan labelled as senior secured (field AS26 Seniority, data entry: "senior secured") must show a corresponding collateral type (field CS6) and a corresponding collateral value (field CS4).

The data quality checks described above have to entirely rely on the submitted loan level data as additional data outside the data templates is typically not available. While many data quality issues can be eliminated through these checks, the data analysis process of the European DataWarehouse nevertheless takes one more step whereby the individual reporting agents are also scrutinised.

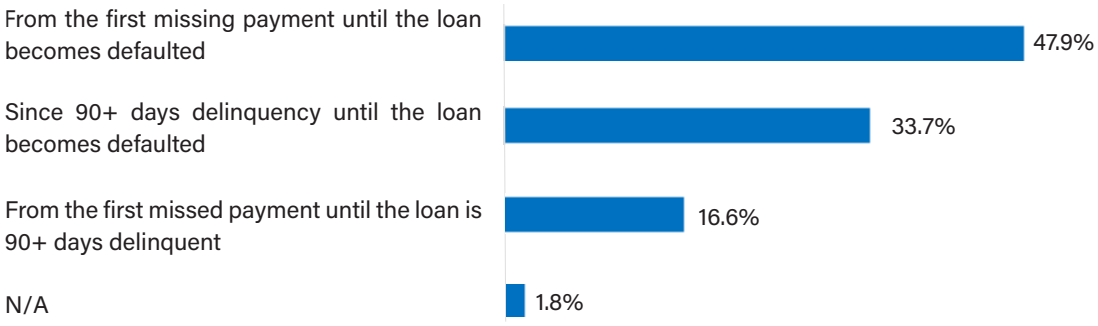
A survey among Spanish issuers of RMBS in the autumn of 2015 showed that there were significant deviations between the reported values from issuer to issuer due to the lack of standardised definitions within individual ABS transactions. This issue arises due to diverging processes and practices within banks as well as the use of proxies when required information was unavailable.

The following examples taken from the survey[7] illustrate these challenges that cannot be solved as part of a pure data-orientated quality analysis:

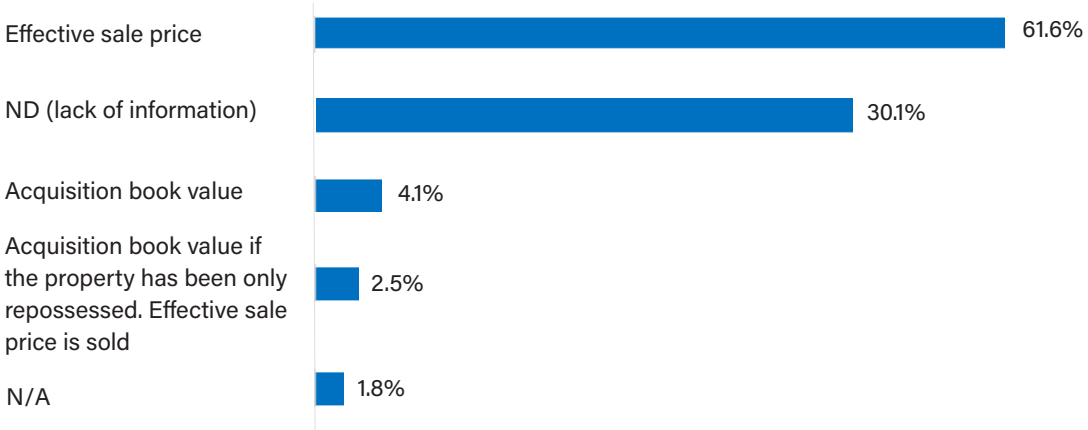
A. Due to the lack of standardised definitions within the European securitisation market issuers typically refer to the transaction documentation during the reporting process. During the survey this became clear for the data reported for loans in arrears (field AR166 Account Status, data entry: "2" in arrears). The following graph shows that about half (47.9%) of the issuers report loans to be in arrears from the first day a payment is missed.

[7] European DataWarehouse, Special Report – European DataWarehouse Commentary on Spanish RMBS Loan Level Data, January 2016, p. 4-6.

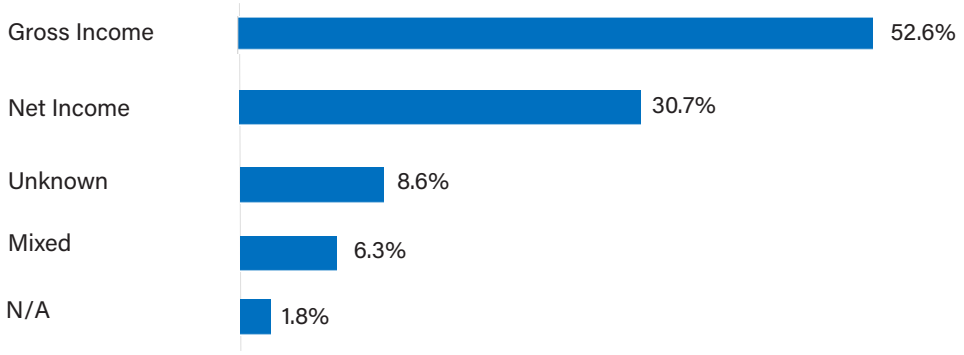
About a third (33.7%) of the issuers, however, only report loans to be in arrears once the payment is at least 91 days overdue and an additional 16.6% of the issuers report these loans as being in arrears anytime between the first and the 91st day. Accordingly, the two latter groups of issuers draw a very different picture about the number and amounts of loans in arrears.




B. Different processes, IT systems and practices within banks are other reasons for deviations that can only be clarified by a discussion with the issuer. These deviations can be caused by information challenges in the liquidation of defaulted loans, for example. The survey among Spanish RMBS issuers showed that about a third of the issuers did not possess the necessary data about the realised recovery values of residential properties that served as collateral for the loans (see graph below).



C. The lack of exact information often forces issuers to work with proxies. The reason might be that data requirements do not comply with the data that has been collected during the loan origination process. In the case of the Spanish RMBS issuers, lack of information regarding the borrower's primary income has been a recurring problem. The ECB taxonomy requires that gross annual income should be reported in field AR 26 (primary income). This value was only reported by about half (52.6%) of the Spanish issuers while more than a third (37.0%) used the net income or a mix of different income types as proxies (see following graph).




This example shows that even with clear definitions, significant data quality problems can occur in practice. As a side note, this example is also an important topic for the transformation process in the BIRD project. While a transformation from monthly income to annual income would be trivial (unless variable remuneration would be included in the annual but not the monthly income), a conversion from net annual income to gross annual income would be difficult to achieve because individual borrower information (e.g. individual tax rate) that is not usually collected during the loan origination process would be required

## Conclusion

The examples outlined above are based on the experience the European DataWarehouse has been gathering during its daily work with loan level data since 2013. As explained, observed data quality problems have three main causes: insufficient clarity of definitions of data fields, erroneous data entries and content-related deviations in the collected data.

Looking at the data quality dimension of the AnaCredit project, these experiences lead to two fundamental issues which must be addressed in the setup of a Europe-wide data collection effort:

1. The description of the required attributes should not leave room for interpretation by the reporting institutions. Given the vast range and individual features of financial instruments in corporate lending, as well as national lending rules and practices, this task will require substantial resources. The AnaCredit manual and the BIRD project will provide extensive definitions and case studies, but it is almost impossible to cover all eventualities ex ante.
2. A comprehensive data quality process is needed that comprises not only a rules-based data analysis but also individual analysis by experts as well as further input by reporting agents. In practice, there will be numerous data quality feedback loops regarding unresolved data quality issues between the reporting institutions and the recipients in the Eurosystem. A high level of data quality is unlikely to be achieved on day one and ongoing communication between reporting agents and information recipients is needed to achieve a high and consistent level of data quality.

## References

*Jentzsch, Nicola*; Do We Need a European Directive for Credit Reporting? in: CESinfo DICE Report, No. 2, 2007, S. 48 – 54.

*European Banking Federation* (EBF); EBF comments on the ECB draft regulation on AnaCredit, 27.1.2016, http://www.ebf-fbe.eu/wp-content/uploads/2016/01/EBF_019220-EBF-Positioning-on-AnaCredit-tracked-changes.pdf.

*European Central Bank,* Regulation (EU) 2016/867 of 18 May 2016 on the collection of granular credit and credit risk data (ECB/2016/13).

*European Central Bank*, Data templates, https://www.ecb.europa.eu/mopo/assets/loanlevel/transmission/html/index.en.html.

*European Central Bank*, Frequently Asked Questions, https://www.ecb.europa.eu/mopo/assets/loanlevel/faq/html/index.en.html.

*European DataWarehouse*, Special Report – European DataWarehouse Commentary on Spanish RMBS Loan Level Data, January 2016.

EUROPEAN DATAWAREHOUSE

**IMPORTANT DISCLOSURES:**