The EWC database data quality guide

Stan De Spiegelaere

February 2017
european trade union institute

etui.

Stan De Spiegelaere is a researcher at the European Trade Union Institute (ETUI) in Brussels. Contact: sdespiegelaere@etui.org

Brussels, 2017

The ETUI is financially supported by the European Union. The European Union is not responsible for any use made of the information contained in this publication.

Contents

Introduction

The European Works Council database (EWCdb) comprises information on European works councils (EWCs), the founding agreements of EWCs and the companies they are established in. The European Trade Union Institute (ETUI) first established a database on companies with EWCs in 1995 in cooperation with external partners and sectoral European Trade Union Federations. Since then, the EWCdb has evolved, been merged with several other databases and extended to cover other forms of transnational information and consultation institutions, such as works councils in *societas europaea* (SEs) or *societas cooperative europaea* (SCEs). Currently, the EWCdb is the only reference database on EWCs.

The EWCdb is managed by the ETUI and has a threefold aim: (1) monitoring EWCs and SE works councils, (2) analysing the evolution of EWCs and (3) providing information to practitioners, policymakers and researchers on EWCs.

In this publication, we present the structure, aims and content of the EWCdb to help users and researchers understand the background to, and content of, the available information. We also describe the efforts made by the ETUI to monitor, assess, correct and prevent errors from occurring in the database.

1. The EWC database: a brief history

— In collaboration with over 20 research institutes and the European Trade Union Federations, the ETUI initially created a database that identifies multinational companies that fall within the scope of the EWC Directive. This database was put together for the first time in 1995.

— In 1999, the ETUI created another database to contain English texts of agreements setting up EWCs.

— In 2004 the collection of agreements and the data on multinational companies were integrated into a single database and supplemented with an analysis of the content of the agreements, provided by SDA-Infopoint. The latest published update was issued in 2006 on CD-ROM.

— In 2005 a selection of the data was put online, which enabled EWC members to inform us about changes in the composition of their EWC or to send us copies of renegotiated agreements.

— In 2007 the database was made accessible online via www.ewcdb.eu and complemented with new features.

— In 2008 a decision was taken to integrate content analysis of EWC agreements, hitherto performed by the Social Development Agency, into the ETUI database of EWCs.

— In 2008 the SE works council agreements were added to the database.

— In 2010 a content analysis of EWC and SE works council agreements was added to the EWC database. This content analysis is based on a scheme of over 100 questions on which all EWC agreements are assessed.

— In 2015 a new website was launched to improve the usability of the EWCdb for practitioners, researchers, policymakers and the general public.

2. Structure of the database

The EWCdb is a **relational database**, divided into three tables of information:

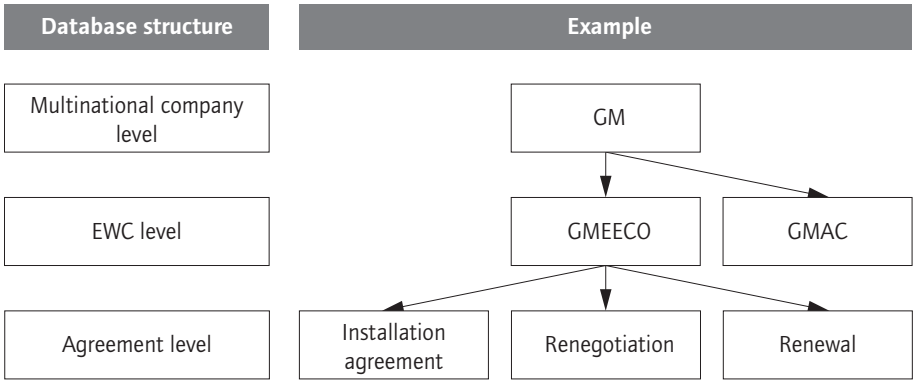
1. First, the database comprises **EWC agreements**. These are signed agreements between the management representatives and the employee representatives of a certain company for the establishment (or subsequent renegotiations and subsidiary agreements on functioning, resources, etc.) of transnational information and consultation bodies or procedures (at the moment, European works councils, SE works councils and SCE works councils). These agreements (or amendments or other related documents) are registered in the database, including some demographic data (e.g. date of signature, agreement type, signatories, etc.) and an analysis of the agreement content using a coding guide.

2. The second table of information deals with the **EWC body**. The EWC bodies are officially installed councils for transnational information and consultation under the EU EWC regulation. These bodies involve a number of agreements: an agreement officially installing the body and, if applicable, possible renegotiations or amendments to that original installation agreement, plus any subsidiary agreements (e.g. transnational company agreements) or other relevant documents (news items, court judgments, etc.). A single agreement can also be linked to two different bodies if, in a single company, several EWCs are established.

3. The third table focuses on the **company**. This includes, where available, information on the company size in terms of workforce, its ownership structure and its activities in various EU countries. Again, a single company can be linked to several EWC bodies (usually for specific divisions) if there is more than one EWC active in its subsidiaries.

The table containing most of the (unique) information in the database is that of the EWC agreements. The agreements provide the basic information for the 'EWC body' table (creation, renegotiation date etc.), and this, in turn, enables us to search for information on the company and introduce it into the database.

Figure 1 **EWCdb tables of information**



Depending on the purpose of the study, the database can subsequently be analysed on different levels:

— **Company:** The data on the EWC bodies and agreements can be aggregated at the company level to address, for example, issues of coverage (how many of the companies covered by the Directive have an EWC?) or effectiveness (how does the presence of an EWC affect company performance?).

— **Body:** The data on EWC agreements can be aggregated at the body level, and the company information disaggregated, in order to establish an overview of the state of play in a given year. This can help to answer such questions as: how many EWCs currently have a negotiation competence or a select committee?

— **Agreement:** All the body and company data can be disaggregated on the agreement level. Using this population, we can assess the evolution of EWC agreements over time, answering such questions as: are competences on corporate social responsibility more common in recent agreements? Did the Recast Directive have an effect on the content of founding agreements?

3. Data quality

A common primary concern when thinking about data quality is the accuracy of the data, i.e. whether or not the values in the database reflect the true values in the population. Accuracy is indeed a central issue, but data quality is a broader concern and includes other aspects such as the completeness, consistency and timeliness of the data (Batini and Scannapieco 2006; Dasu and Johnson 2003). These four dimensions are traditionally referred to as the central dimensions of data quality:

1. **Accuracy** refers to the proximity of the value in the database to the true value. Incorrect coding of the data is the main cause of a lack of accuracy.

2. **Completeness** is the degree to which the population is completely covered in the database and whether or not important subgroups are excluded. Here, we consider missing values both on the unit level (observations not included) and on the item level (missing information on a certain variable in an observation).

3. **Timeliness** refers to the updated character of the data and whether or not the data reflect the current situation or whether there are fields that require an update.

4. **Consistency** refers to the internal logical consistency in data points. If one variable indicates someone is male, s/he cannot be categorised as female in another variable.

These four dimensions of data quality are all of importance for the overall data quality. Nevertheless, some trade-offs between the dimensions can occur. Depending on the object of the study, timeliness might be a more important factor than completeness (or vice versa).

For the EWCdb, efforts regarding data quality optimisation are performed on all four dimensions. We here distinguish between efforts focused on preventing errors, assessing errors and fixing identified errors.

4. The steps of data collection, input and analysis

In line with the 'total survey error' approaches to surveying quality (Weisberg 2009), we map here the different steps of data collection, input and analysis, and assess the various types of error that can occur (see Figure 2).

The EWCdb is a database that aims to include and cover all **established** EWCs and SE works councils, as well as all their related agreements and companies. The first step of the data process is therefore the agreement **identification**. Using various sources (expert networks, trade unions, submissions via ewcdb.eu), the EWCdb obtains information on EWC agreements. After the identification, access to the agreement needs to be ensured using different channels. If the agreements and bodies are not all included in the database, a *coverage error* can occur.

Once accessed, all agreements are **processed** internally. If necessary, a translation of the agreement is obtained and all language versions are put into a common lay-out. In these phases, *processing error* is a risk. Pages can go missing, translations can be substandard, etc.

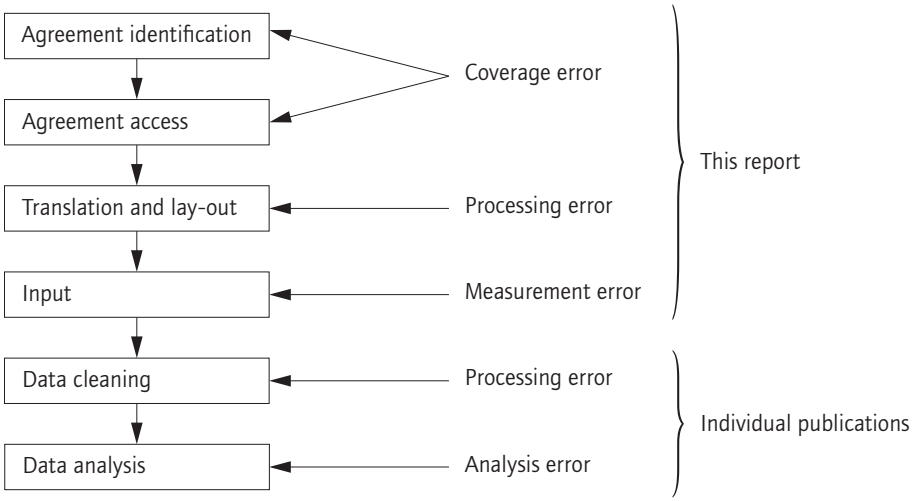
In a third step, the contents of the agreements are **analysed and coded** in the database. This step requires a great deal of manual effort and can allow for

measurement error to occur as the codes in the dataset can divert from the true value.

After the data input, the data are subjected to a process of **data cleaning and data transformation** to make the data ready for analysis. In this phase, again, a form of *processing error* can affect the data quality.

In the last step of **analysis**, it is important to choose the correct analysis techniques, otherwise a further type of error, *analysis error*, is caused (interpretation and coding of agreement).

Figure 2 **Error in the EWCdb**



In this report, the focus will be on the first three types of error (coverage, processing and measurement), as they are intrinsically linked to the database. The other types of error are more relevant for the analysis stage when the gathered and inputted data are used for further analysis. Choices of data cleaning, transformation and analysis are to be discussed in further individual publications.

5. Coverage error

Coverage error refers to the mismatch between the observations in the database and the population. In this case, some members of the population (agreements) are not registered in the database. For a database that aims to encompass the whole population, any exclusion of an EWC or SE works council constitutes an important error. Traditionally, distinction is made between different types of item non-response (missing data).

The first type of missing data refers to data, which are **missing completely at random** (MCAR). MCAR occurs when the reason for the data being missing

has no relation at all to the variable of interest, or to any of the variables related to the variable of interest. In our case this would mean that missing agreements are not different in terms of content to the registered agreements and are not generally present in, for example, a certain sector. This type of missing data is not likely to cause any bias in the results and can be ignored.

Missing data that is **missing at random** (MAR) refers to missing data that is not related to the variable of interest, but is related to another possible explicatory variable. In our case this would mean that missing agreements are not different in terms of content, but there are more missing agreements in certain sectors than in others. For this type of missing data, controlling for the relevant variables normally suffices (Weisberg 2005: 141).

A third situation occurs when data are **missing not at random** (MNAR). In this situation, the missing data are related to the variable of interest, the agreement content. MNAR could occur when, for example, the best or worst EWC agreements are significantly less present in the database. This type of non-response cannot be ignored.

A normal way of handling unit non-response would be the inclusion of a **weighting factor**. This gives a higher importance (weight) to agreements in categories that are underrepresented and less to overrepresented categories. In order to design a weight, the population distribution has to be known. In our case, this is not evident. There is no exhaustive list of all the EWCs in Europe, nor is there any reliable survey-based material that could give us an insight into the distribution over sectors. Moreover, the EWCdb aims to be the reference database of all EWCs in Europe.

For this reason, we do not apply or develop any weighting procedures. We do, however, introduce some procedures to minimise the coverage error and regularly estimate the coverage error in the EWCdb.

5.1 Prevent: ETUI coverage policy

The identification of EWC agreements is currently done using multiple sources:

1. **EWC coordinators**: the EWC coordinators of the European Trade Union Federations, who are in constant contact with many EWCs, frequently notify the EWCdb database managers of new or changed agreements.

2. **ETUI trainers:** through the training of the ETUI, constant updates on EWC agreements (including renegotiated agreements) are submitted to the database.

3. **Press:** specialised labour-oriented press is used to identify new or renegotiated EWC agreements, such as Planet Labor, IR notes and Ewc-news.com.

4. **EWC agreements:** in the renegotiation of EWCs or of installation agreements after mergers or acquisitions, frequent reference is made to previous agreements. When missing, these agreements are added to the database.

5. **ETUI networks**: using the personal and professional networks of the ETUI researchers, additional agreements are collected and registered.

6. **Crowdsourcing:** one of the identified potential solutions for the coverage error is *data sharing* and open data availability. Through the sharing and consultation process, errors are identified and can be treated in the original database (Dasu and Johnson 2003: 109). As the EWCdb is an open database frequently used by practitioners to gain access to a certain agreement, the users contact the database managers regularly with updated information or agreements. Moreover, in the new EWCdb website, registered users are explicitly invited to provide updates of their EWC to the database managers.

Once an agreement is identified, it is included in the database with some very basic demographic data. In a next step, the access to the agreements is ensured. When the identification and access phases do not run simultaneously, queries are sent to the sectoral European Trade Union Federations (ETUF), coordinators and other personal contacts to enable the agreement access.

5.2 Assess: non-coverage error estimation

As the EWCdb is not a sample-based database, the ambition is to cover, exhaustively, the whole EWC field, a potential coverage error can by caused by (1) identified, yet missing agreements (mainly early agreements from the 1980s and 1990s) and (2) non-identification of agreements.

5.2.1 Missing full texts of agreements

One source of non-coverage error (and possibly bias) is the agreements that are registered and known, but of which the full text is not available. These agreements, therefore, cannot be analysed. If these agreements are significantly different from the agreements that are available, this non-coverage can introduce a bias in the aggregate results of the content analysis.

Box 1 **Missing agreements**

To assess the degree of this potential bias, we perform several logistic regressions to identify how the missing agreements differ from the overall population in terms of demographic variables like company size, sector, EWC type or country of the company HQ. Such an analysis guides future efforts to get more full-text access to agreements in order to minimise this source of non-coverage error.

5.2.2 Non-identified agreements

A second source of potential non-coverage error (and bias) is the agreements that are not registered in the database, yet exist. EWCs can be installed or renegotiated by agreements not recorded within the EWCdb. Again, if missing agreements differ significantly from the agreements that we have access to, this non-coverage can be a source of bias in the aggregate analysis.

As we do not have any knowledge of those agreements, it is impossible to make a comparison. According to the '**late response' argument**, we can nevertheless assume that the agreements that are registered very late, after their signature, might share the same characteristics as agreements that we have not yet registered. The technique of using late respondents (in our case late registration) as a proxy for assessing the non-coverage error is often used in survey analysis (e.g. Billiet and Waege 2003: 313; Helasoja *et al.* 2002; Korkeila *et al.* 2001).

Box 2 **Non-identified agreements**

In a second non-coverage error assessment, we therefore look at (1) the type of agreements that we registered very late and (2) how they differ in terms of content from agreements that we obtained knowledge of and access to soon after their conclusion. This analysis helps us in assessing the importance of this non-coverage error and which populations to focus our efforts on.

6. Processing error

Dasu and Johnson (2003: 110) talk about *data mutilation* and *data loss* as two potential forms of processing error. With data mutilation, the translation and lay-outing steps might change some of the data due to omissions or inaccurate translation of the agreements. In this case, observations (agreements) can be **censored**, in the sense that their observation is incomplete or incorrect. Data loss, meanwhile, refers to the simple omission of some observations from the database.

For the EWCdb, such forms of processing error are avoided in several ways. When news about a new/renegotiated agreement is received, a new record is introduced in the database immediately. When the agreement is accessed, it goes through different stages (archiving, translation, layout) before it is analysed and made available on the website. At the analysis phase, a check is regularly performed on the status of different agreements to ascertain whether they have been correctly archived and made available.

Box 3 **Inconsistencies in demographics**

Some inconsistencies can nevertheless occur that affect the accuracy of the database. These issues are checked on a regular basis and procedures are developed to prevent such inconsistencies occurring. Examples include missing (incorrect) names, missing information on dates, missing information on dissolution, ambiguous/incorrect data on sector due to incorrect company data, etc.

7. Measurement error

Measurement error is the error that occurs in the database during the measurement process. In this process, the true value of a certain variable is translated to a value in the database. If the true value differs from the value in the database, a measurement error has occurred.

Distinction is generally made between two types of measurement errors. When the measurement error is completely random, the errors tend to balance out in the end. This is what we call **variance**. An extra (erroneous) variance is included in the database. This will make the estimates less certain but is not likely to affect the validity of the estimated parameters. Applied to the EWCdb, some errors in coding (due to wrong classification/interpretation for example) label an EWC as being a 'joint' or an 'employee-only' body. If these mistakes are random, about the same amount of agreements will be (erroneously) labelled 'joint' as 'employee only'. Furthermore, these errors will not significantly relate to other variables. Such mistakes are not likely to affect the estimates regarding the proportion of joint or employee-only agreements. Nor are they likely to change the possible relation between joint bodies and certain other aspects, although they might inflate the standard errors of such estimates.

When the measurement errors are systematic, however, **bias** occurs. Using the same example, joint bodies are systematically more often labelled as employee-only bodies than vice versa. Consequently, the errors will not balance out and the estimated proportional distribution will be erroneous.

While the first type of measurement error is problematic in terms of obtaining clear results, the second is problematic because it leads to wrong results and conclusions. To minimise both types of measurement error, the EWCdb uses several preventive techniques. The coder effect is then also analysed in order to assess the overall measurement error in the data.

7.1 Prevent: the EWCdb coding process

Once the agreement is accessible, an ETUI researcher reads the agreement and completes the analysis framework. This framework is a list of topics in which the researcher should select the appropriate codes according to the agreement

content. While for the most part the codes are relatively straightforward (e.g. applicable law, duration of the agreement) or binary values (e.g. items on the EWC agenda), others might require differing degrees of interpretation when the text is ambiguous or the available analytical categories do not match the reality. In this coding process, different errors can affect the data quality:

1. Simple errors can be made in which the researcher selects the wrong code in the coding frame.
2. Researchers can forget to assign a certain code because they missed a paragraph in the analysed agreement.
3. Interpretation errors can be made if the text is ambiguously worded.
4. The codes can be misinterpreted by the researcher and be incorrectly assigned.
5. The interpretation of certain provisions may have evolved with the number of agreements analysed.

To minimise these errors, several preventive measures and tools are developed:

1. **Limited amount of coders:** So far, the analysis of EWC agreements has been undertaken primarily by four different researchers. These researchers have always liaised with each other in the coding phase and exchanged questions and notes on how to interpret and code several passages. Moreover, the researcher performing the coding is registered in the database.

2. **Coding guide:** To improve the communication and identical interpretation of the coding scheme, a guide is created in which the interpretation is further developed and some examples are given.

3. **Free format coding:** If some agreement stipulations do not fit the coding scheme, or some exceptional provisions are included in the agreement, the researchers can note these in a free format part of the analysis framework called 'selected provisions'.

4. **Optimised data interface:** In the data-input interface, several input controls are built in to avoid coding errors. As such, the data type of all fields are predefined and several dependencies are established (options that can only be selected if another field is completed) (Hellerstein 2008).

7.2 Assess and correct: outliers

Outliers are 'observations which deviate so much from the other observations as to arouse suspicions that they were generated by a different mechanism' (Aggarwal 2013). They are, in other words, observations in which it is very likely (but not definite) that some kind of measurement error occurred in the data input.

Outliers can be detected by focusing on the scores of individual variables and detecting those observations that are significantly different from the overall norm. While such an approach is easy for continuous variables, this is not the case for categorical variables, which have a predetermined range of possible values. Therefore, we look at multivariate outliers, observations with such a dissimilar combination of different values that error is likely.

Box 4 **Multivariate categorical outliers**

We use **latent class analysis** to identify these multivariate categorical outliers. Based on a set of categorical variables, different types of EWCs can be distinguished from each other. In such an analysis, it is possible to identify ***influential cases*** based on their Cooks' D value. This value gives insight into the influence of a specific observation in the model. Observations with unique and very different patterns have higher Cooks' D values and are thus worth attention.

7.3 Assess: the 'coder effect'

Most of the data in the EWCdb are gathered at the agreement level. These agreements are nested in EWC bodies, as one single EWC can have several agreements (installation agreement, amendment, renegotiation, etc.). Several coders using a unique coding frame (see above) undertake the data input. Agreements are thus not only nested in EWCs, but also in coders.

For most of the fields, we do not expect a large coder bias, as the information is generally 'objective' in the sense that it reflects the coding of easy-to-determine characteristics (e.g. amount of EWC members, signature date, presence of a select committee, etc.). For some other questions, the coder is required to make a larger interpretation of the agreement. A significant difference in interpretation of the different coders can thus occur. Based on the coding frame, a selection of questions is used to assess the coder effect in the EWCdb.

To see whether the coder has a significant effect on the agreement analysis, one would normally perform a multilevel analysis with the coder as a second-order random variable. As only four researchers are currently involved in the coding, we opted for a multilevel analysis with 'company' as a random variable, but added the coders in the fixed effect part.

Next to the overall coder effect, a second coder-related source of error can be the **coder drift**. Coder drift can occur because, over time, coders tend to code some agreements differently. As such, the interpretation of a variable can be subject to a subtle evolution, which introduces error and bias in the results. In order to assess such a possible coder drift, the date of the last update is included in the analysis, both on an overall level (to analyse the coder-independent coder

drift) and in interaction with the coder (in order to identify individual coder drift) (Orwin and Vevea 2009).

Box 5 **Coder effect**

To inspect the coder effect, different (multilevel logistic) regression analysis are performed including the coder as a fixed variable, the last updated date, and the interaction between both in combination with a series of control variables.

Conclusion

In this report, we discussed the data quality approach of the ETUI towards the European Works Council database (EWCdb). This database is a complex relational database including information from several sources gathered using a multiplicity of means.

One of the central aims of the database is to monitor the evolutions in the EWC population. For this, the EWCdb is currently the only reference database of EWCs and aims to be as exhaustive as possible. A second central aim of the database (the third being to make information accessible) is to analyse the content of EWC agreements and their evolutions over time. For this, the content analysis should be as complete and correct as possible.

This report described the ETUI policies to prevent errors from occurring, identify the errors, assess their importance and correct them. The preventive measures constitute a continuous task, while the assessment of the potential errors will be the subject of regular updates on the website.

References

Aggarwal C.C. (2013) Outlier analysis, New York, Springer.

Batini C. and Scannapieco M. (2006) Data quality: concepts, methodologies and techniques, Berlin, Springer.

Billiet J. and Waege H. (2003) Een samenleving onderzocht onderzocht: methoden van sociaal-wetenschappelijk onderzoek, Antwerpen, De Boeck.

Dasu T. and Johnson T. (2003) Exploratory data mining and data cleaning, New York, Wiley-Interscience.

Helasoja V., Prättälä R., Dregval L., Pudule I. and Kasmel A. (2002) Late response and item nonresponse in the Finbalt Health Monitor Survey, The European Journal of Public Health, 12 (2), 117–123. doi: 10.1093/eurpub/12.2.117

Hellerstein J. (2008) Quantitative data cleaning for large databases. http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf

Korkeila K. *et al*. (2001) Non-response and related factors in a nation-wide health survey, European Journal of Epidemiology, 17 (11), 991–999. doi: 10.1023/A:1020016922473

Orwin R.G. and Vevea J.L. (2009) Evaluating coding decisions, in Cooper H., Hedges L.V. and Valentine J.C. (eds.) The handbook of research synthesis and meta-analysis, New York, Russell Sage Foundation, 177–203.

Weisberg H.F. (2005) The total survey error approach: a guide to the new science of survey research, Chicago, University of Chicago Press.