Working Paper 27.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24-26 April 2017)

An information model for a metadata-driven editing and imputation system

Prepared by Rok Platinovsek, Statistics Finland

Abstract

Banff procedures can be seen as generic building blocks of the editing and imputation (E&I) process. Banff's status datasets partly contain E&I audit trail information but do not fully specify the E&I actions performed. In this paper we propose an information model for a metadata-driven E&I system that stores a full audit trail and allows for easy calculation of standardized indicators. Our metadata information model is influenced by Banff processor's use of linked information objects and takes into account the Generalized Statistical Data Editing models framework. The model's twelve metadata objects allow a Banff-based E&I process to be fully specified. The model is generic enough to also accommodate non-Banff methods with minimal or no alterations. We also discuss data organization and outline a simple data organization model that serves the needs of the E&I system.

I. Introduction

1. Efforts are underway at Statistics Finland to modernize the way editing and imputation (E&I) is conducted at our office. At present, many surveys implement editing and imputation with customized (SAS) programs containing hard-coded parameters with no versioning. This results in poor traceability of E&I activities, low comparability of E&I practices between surveys as well expensive maintenance of multiple programs aimed at achieving a similar purpose.

2. To improve the situation, we are considering developing a generic E&I system. The system should be fully metadata driven in order to dispense with hard-coded parameters. A full audit trail should be saved in order to allow E&I actions to be examined and reproduced even years after they were conducted. The system should also facilitate producing graphical checks and computing indicators like the imputation rate in a simple and standardized way.

3. The E&I system will utilize existing generic E&I solutions like Banff and Selekt. In order to meet the aforementioned demands, however, a metadata layer will need to be devised to manage and maintain the parameters used in E&I actions. This paper gives an overview of the E&I information model we are currently developing for the E&I system. The information model consists of two parts that will be discussed in turn: the metadata information model and the data organization model.

4. Because we intend to utilize Banff procedures and because our metadata information model is influenced by the Banff Processor, we will start by overviewing the Banff system and underlining those of its features that are important for our purposes.

II. Banff

5. Banff is a collection of SAS procedures used for editing and imputation. One of the strengths of the Banff system is that it is modular, meaning that the SAS procedures can be used independently of each other. The procedures can be combined somewhat freely to produce the desired E&I process, e.g., attempt to impute values with a preferred method and where that fails impute the remaining values with a more robust method (Statistics Canada 2014). The logical order in which Banff procedures are applied in the context of survey processing is given in Figure 1.

B. Banff's procedures as GSDEM constituents

6. According to the Generalized Statistical Data Editing models framework (UNECE 2015, see also Pannekoek and Zhang 2012) E&I activities fall into three broad categories in terms of their purpose. Examining the data and trying to identify potential problems are classified as *review* activities. *Selection* refers to selecting units or variables within units for further treatment. *Amendment*, in turn, refers to actually changing selected data values in a way that is considered appropriate to improve the data quality.

7. Applying this classification to the typical process flow of Banff procedures in Figure 1 reveals an unsurprising pattern: Banff procedures implementing review and selection are applied first and are flowed by amendment procedures. Banff's *Outlier Detection* module can be seen as implementing both review and selection functions as it, firstly, assesses a field's plausibility by calculating an outlier measure (review) and, secondly, marks up fields with measures exceeding the threshold for further treatment (selection). The *Error Localization* procedure, likewise, performs a dual function. It evaluates a record's validity according to a set of edit rules (review) and uses Fellegi and Holt's (1976) error localization approach to identify fields for treatment (selection of variables).

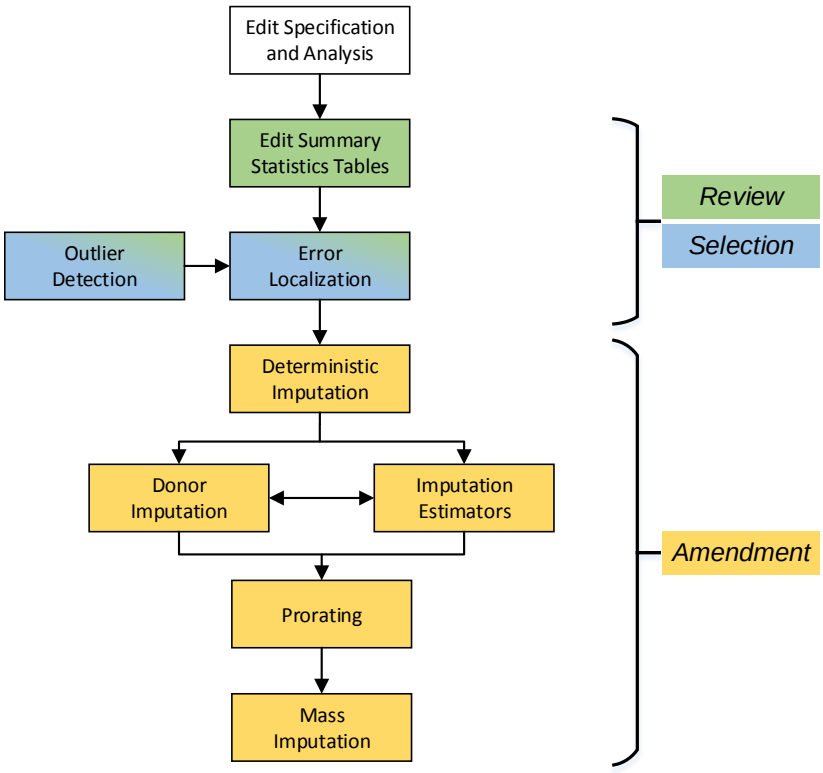

Figure 1: The logical order of the Banff procedures in the context of survey processing; source: Figure 1.1 (Statistics Canada 2014). Categorization with regard to purpose added by the author.

8. In the Banff system, selection is technically implemented by marking fields for further treatment in the so-called *status dataset*. This information is passed down to imputation procedures that actually change (amend) the fields' values. Each Banff's amendment procedure requires such a status dataset as input, but also writes a status dataset of its own wherein it marks the fields it imputed.

9. The use of status datasets is a central feature of the Banff system. On the one hand, it allows review and selection procedures to pass information to amendment procedures downstream in the E&I process. This enables the Banff system to be modular. On the other hand, the status dataset partially contains the audit trail information as each amendment procedure stamps treated fields with its own status identifier. We note that the audit trail contained in Banff's status datasets does not, however, sufficiently document the E&I actions to allow their results to be reproduced. The status dataset merely identifies the applied methods but does not specify the methods' parameters.

B. Banff Processor

10. The Banff Processor is an interface that enables the E&I process to be driven by metadata tables. The sequence in which Banff procedures (and possible user-defined SAS programs) are to be run is defined in the main driver table like the one shown in the top left corner of Figure 2. Each row of the main driver table pertains to an E&I method and can contain links pointing to secondary metadata tables further specifying the method's parameters.

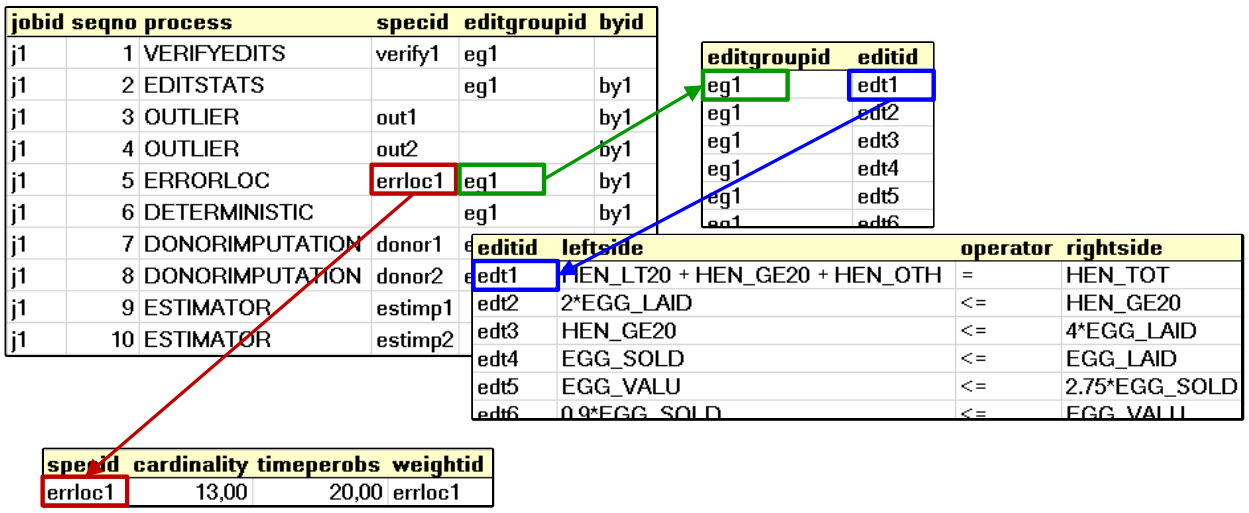

Figure 2: Banff Processor's linked metadata tables.

11. The advantage of having a metadata driven system is that the maintenance of the E&I process boils down to managing tables, rather than code (Statistics Canada 2012). In our opinion, this approach also emphasizes the E&I process as a whole rather than focusing on its individual components.

12. As the example in Figure 2 illustrates, certain parameters – most importantly the set of edits – are defined as separate metadata objects and can be referred to from several methods. One benefit of this design is that edits do not need to be defined separately for each method. This approach also facilitates testing "what if" scenarios: the set of edits can be modified to test how the change affects the entire E&I process instead of having to make the same change for each module separately (Statistics Canada 2012).

13. The Banff Processor's features we have outlined have influenced the metadata information model that we describe in the following section. We begin by giving a short overview of Statistics Finland's metadata infrastructure.

III. Metadata

14. Efforts to incorporate metadata into the statistical production process at Statistics Finland span back to the early 2000s (see Rouhuvirta 2001, 2004). XML-form metadata records have been made available in a central repository and serve, firstly, as the cornerstone of an XML-based publishing system, allowing many tasks involved in producing statistical publications to be automated (see Lehtinen 2006). In recent years, efforts have also been made to integrate the metadata into earlier phases of statistical production (see Platinovsek and Piirainen 2016).

15. In principle, each dataset (but also each table or publication) used at Statistics Finland should have a corresponding metadata record in the repository. Employees of the statistical office may freely access any metadata record in the repository: while data may be confidential, their corresponding metadata records are public. The metadata records are structured according to the CoSSI information model (Common Structure of Statistical Information) whereby the content of each metadata record is divided into three modules (Lehtinen 2006).

(a) The *statistical* metadata module describes a particular dataset's variables: their operational definitions, measurement units, classifications etc.
(b) The *document* metadata contain information about the dataset, table, or publication as a whole (creator, subject, keywords etc.).
(c) The *processing* metadata module contains instructions that determine the functioning of certain software.

B. The metadata information model

16. While the processing metadata module has been in the CoSSI model since the beginning, it has not seen extensive use. In order to implement a Banff-Processor-like metadata driven imputation application, the processing metadata module will be expanded to include a description of the E&I process. A preliminary version of the E&I information model is currently under discussion at Statistics Finland. In formulating the information model, the GSDEMSs framework was taken into account with regard to terminology and features like the classification of functions according to purpose and the separation of function and method.

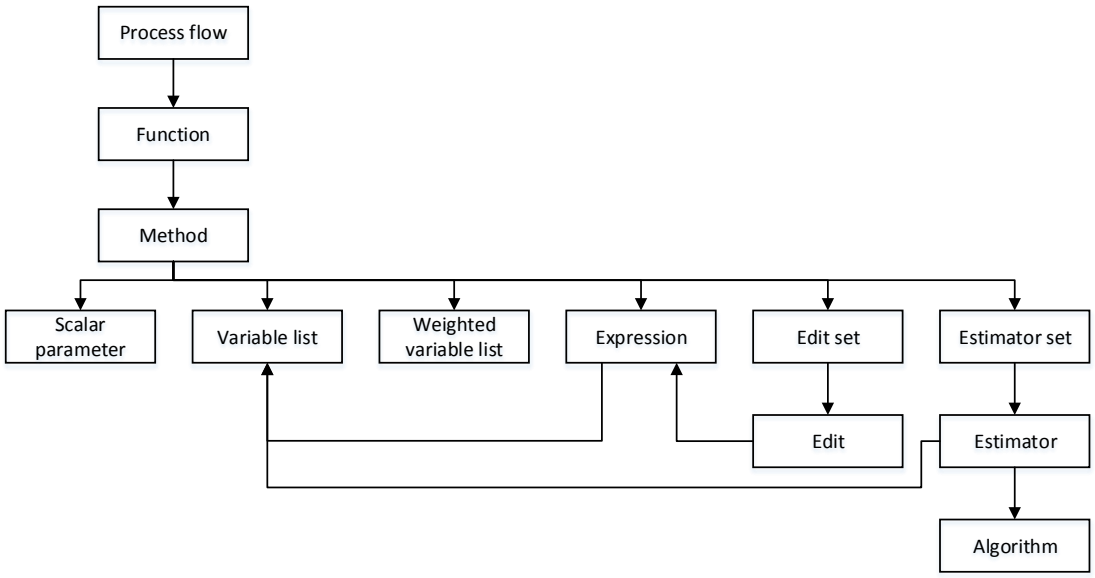

Figure 3: The E&I metadata information model.

17. Figure 3 depicts twelve metadata objects and their interrelations that allow the E&I process (implemented by Banff procedures) to be completely specified. The Document Type Definition (DTD) used to specify the model can be obtained from the author upon request. We hereby briefly describe the model's main features.

(a) The *process flow* is the topmost object in the hierarchy. There can be several process flows in a metadata record, e.g., one used for production and another used for development and testing.
(b) A process flow contains *functions*. A function describes what E&I activity is performed, but leaves to the method to specify how the activity in question is implemented (UNECE 2015). A function has a *purpose* attribute that classifies it as review, selection or amendment.
(c) A function is carried out by one or several *methods*. The method object is the central metadata object in the information model. A method has an *implementation* attribute identifying the concrete procedure with which it is implemented (e.g. Banff donor imputation). The method's set of input parameters, naturally, depends on the implementation. The input parameters may include a number of scalar parameters and other parameter objects that we describe in turn.
(d) The *scalar parameter* is the simplest parameter object and is essentially a name-value pair. Banff's donor imputation module, e.g., requires the user to specify the minimum number of donors per by-group that is required to perform imputation. This parameter can be stored in metadata as a name-value pair, e.g., *name="mindonors" value="10"*.
(e) Certain methods require a list of variables as an input parameter (e.g. by-variables, the donor module's must-match variables). The *variable list* is an ordered vector of variable names. A single variable serving as a parameter is implemented as a variable list of length one. Each variable in the list can be tagged with a unique identifier linking the variable to its documentation in the descriptive metadata (i.e. the CoSSI model's statmeta module). Consistently tagging variables with their IDs allows the metadata system to be queried with queries like "what roles does variable X appear in?"
(f) A *weighted variable list* is a vector of variable-weight pairs (used, e.g., in Banff's error localization procedure).
(g) An *expression* is a SAS-expression (e.g. "*strat=1*") that can be used, e.g., to select a subset of the dataset to which an E&I action is applied. The expression also contains a variable list of all variables appearing in the expression.
(h) An *edit set* serves as a named container for edit rules.
(i) An *edit* specifies the edit rule through an (in)equation and defines its certain other characteristics (described below). The (in)equation is expressed as a SAS expression that, in turn, contains a variable list of all variables appearing in the edit. By using this structure, we tag each variable appearing in the edit with its ID, allowing for queries to the metadata of the type mentioned above.
(j) An *estimator set* works as a named container for estimators.
(k) An *estimator* object specifies the imputation algorithm, defines its parameters, and casts variables into roles using two variable lists (variable to be imputed, auxiliary variables).
(l) An *algorithm* object defines a custom imputation algorithm (one not found on Banff's list of pre-defined algorithms).

18. The metadata information model is intended to be generic enough to be easily extendable. Specifying scalar parameters as name-value pairs means that the information model itself does not need to be changed when new methods are added provided that the method to be added can be parametrized via the parameter objects already in the model.

19. The edit rule contains attributes *error message* and *error level* that are specified for the Validation and Transformation Language's *define datapoint ruleset* object (VTL, 2016). The value of the error level attribute denotes the severity associated to the edit rule. The error message element, in turn, contains a textual description of the edit involved that can be used in reports, tables etc. Including the edit rule related information required by the VTL in the information model enables VTL-compliant documentation to be generated on the basis of the metadata record should that be required at some point in the future.

20. Our final note in this section concerns the versioning of E&I parameters. Once a set of E&I parameter values has been used in production, it should be retained indefinitely to ensure the outcomes of the E&I actions can be reproduced and checked. In the document type definition, we thus included a modification of the v-document versioning mechanism (Wang and Zaniolo 2008). The versioning is implemented by introducing two additional arguments *vstart* and *vend* to versioned XML-elements. After the timestamp given in vend, the element in question is either changed or removed. The value of vend can

be left empty to denote the ever-increasing time-stamp of the current version (see the example in the Appendix). This versioning mechanism allows temporal queries of the type "what E&I parameter set was valid at time point X" to be used to extract information from the metadata.

B. Software implications

21. The metadata information model is, of course, intended to serve the software utilizing it. We are presently considering the main console of the E&I system to be implemented as a SAS Enterprise Guide custom task (see Hemedinger 2013). Like SAS EG's native tasks, custom tasks, too, are code-generating graphical user interfaces (see Platinovsek and Piirainen 2016 for a description of metadata-related custom tasks already in use at Statistics Finland). We envision the E&I custom task to be used in the following way.

(a) The user connects to the metadata repository and reads therefrom (or creates) the appropriate metadata record.
(b) The E&I process flow with its subcomponents can be created or maintained via the custom task's graphical user interface. If changes are made, the metadata are updated accordingly.
(c) Upon user confirmation, the sequence of E&I actions defined in the process flow is processed. The custom task generates and executes the appropriate SAS code containing, among other things, calls to Banff procedures.

22. One Banff Processor's feature that we would like to reproduce in the E&I application is parameter re-use. The user should be able to define and maintain *libraries* of information objects like methods, edits, estimators, custom algorithms etc. An edit rule can therefore appear in several edit rule sets, e.g. in donor imputation the same rule can appear both in the "regular" rule set and in the post-imputation rule set (see Statistics Canada 2014). As mentioned, defining independent metadata objects and referring to them dispenses with unnecessary multiple definitions of the same object as well as facilitates easy testing of "what if" scenarios. Changing a particular edit rule, e.g., affects all edit sets in which it appears and this, in turn, affects all methods relying on the edit sets in question.

23. One important feature of the proposed metadata information model is also that it imposes a driver table setup for E&I activities. This means that a main driver table like the one shown in Figure 4 would be prominent in the application's graphical user interface. The user would be able to access the methods' parameters by clicking on the rows of the table.

Function	Method name	Purpose	Implementation
Verify the set of edits		**prep**	
	Verify the set of edits		BANFF veryfyedits
Edit summary tables		**review**	
	Edit summary tables		BANFF editstats
Identify outliers		**selection**	
	Identify outliers; historic method		BANFF outlier
	Identify outliers; current method		BANFF outlier
Identify inconsistent observations and select fields for imputation		**selection**	
	Identify inconsistent observations and select fields for imputation		BANFF errorloc
Impute missing values and fields identified by error localization		**amendment**	
	Deterministic imputation		BANFF deterministic
	Donor imputation within areas		BANFF donorimputation
	Donor imputation (unrestricted)		BANFF donorimputation
	Estimator imputation (negative values not accepted)		BANFF estimatorimputation
	Estimator imputation for QR_PROF (negative values accepted)		BANFF estimatorimputation

Figure 4: Mock-up of the main driver table in the graphical user interface.

IV. Data organization

24. The metadata information model outlined in the previous section allows the complete set of parameters determining the functioning of E&I activities to be described in the metadata. Ensuring the traceability and reproducibility of E&I activities thereby boils down to including the following information in the audit trail.

(a) Marking the field that was reviewed, selected or amended. As mentioned, Banff status datasets already contain this information.
(b) Providing a reference to the E&I method that was used. The reference points to the method description in the metadata where the method's parameters are specified.
(c) Storing the timestamp when the E&I action was performed. This is required in order to be able to retrieve from the metadata the parameter set for the production cycle in question since methods' parameter values may differ from one production cycle to another.

25. As mentioned, we wish to move beyond stove-pipe oriented production and develop generic E&I applications that can be used for a variety of surveys. The challenge is that any application operating on data must assume the data to have a certain structure. Developing generic E&I applications therefore requires that a certain degree of standardization be introduced with regard to the way data are organized.

26. We will describe a naïve data organization model that works well for illustration purposes, but is rather inefficient: the data organized with respect to this model contain numerous missing and repeated values. We will discuss how the naïve data organization model facilitates E&I activities. A more efficient data organization model will need to be specified that should fulfill the same needs as the naïve model presented here.

27. Figure 5 gives a minimal example of an edited dataset organized according to the naïve model. The variable name prefix *edt_* is reserved for E&I related variables in the example data. The *edt_status* variable either has the value "original" or holds the information on the E&I method applied in the row in question. The value of the *edt_mref* variable points to the description of the E&I method in the metadata, while the *edt_time* records the time when the E&I action in question was performed. Each variable that is subject to E&I actions has a corresponding indicator variable added (prefixed *edt_n_* in Figure 5). The indicator variable holds the information on whether the field in question was identified as invalid/implausible, selected for further treatment, or amended.

	edt_status	edt_mref	edt_time	id	year	class	var1	var2	edt_n_var1
1	original		2016-03-01 09:31	001	2015	2	45	150	.
2	original		2017-02-15 15:01	001	2016	2	51	156	.
3	original		2016-03-01 09:31	002	2015	9	12	99	.
4	original		2017-02-15 15:01	002	2016	9	60	110	.
5	selection/banff_errorloc	ref5	2017-02-16 10:03	002	2016	9	.	.	1
6	amendment/banff_estimator	ref7	2017-02-15 10:23	002	2016	9	13	110	1

Figure 5: A minimal example of a cumulative data with several production cycles and version history.

28. The example data contain two units (id=001 and id=002) observed at two time points (year=2015 and year=2016). The fact that the second unit was edited in the second production cycle can be discerned from the fact that there are several rows pertaining to this unit and time point (rows 4 through 6). Row 4 gives the original (unedited) values for the record. As E&I actions were performed, rows were added to the same dataset describing how fields were selected for treatment and amended. Row 5 gives the information that variable *var1* was marked up for further treatment (*edt_n_var*=1) by the error localization method (as denoted by the value of *edt_status*). Row 6, in turn, contains the record's amended values. The amended field is marked as having been treated by the estimator imputation method (*edt_n_var*=1).

29. The outlined data organization model is inefficient for the following reasons. The indicator variables (prefixed *edt_n_*) mostly contain missing values. The values of survey variables are missing for review and selection rows (i.e. *var1* and *var2* in row 5 in the above example). For amendment rows, all untreated fields' values are repeated with regard to the original row.

30. Despite its inefficiency the data organization model does satisfy the following demands.
   (a) Data pertaining to different production cycles can be extracted in a standardized way. The extracted data have a standard structure. This is necessary since certain E&I methods require as input a current and historical dataset with the same structure. A current and historical dataset pair is used, e.g., in selective editing score calculation, outlier detection, and LOCF-type imputation.
   (b) A particular editing version can be extracted in a standardized way (by conditioning on the value of *edt_status*). When performing LOCF-type imputation, we typically want to use untreated, rather than imputed, historical data so that values that were imputed in the previous production cycle are not carried forward into the current one.
   (c) Indicators like the imputation rate can be calculated in a standardized way. Calculating the record-level imputation rate using the naïve data organization model boils down to counting the number of imputed rows and comparing it to the number of original rows.
   (d) It is possible to review an observation's the audit trail to verify that the E&I actions were performed as intended. Using the values of *edt_mref* and *edt_time* allows us to extract from the metadata the parameter set that was valid at the time the method in question was run.

V. Conclusion and future work

31. We have outlined in this paper an information model for a metadata driven E&I system. The information model consists of two main parts: the metadata information model and the data organization model. Banff Processor's idea of linked metadata objects has influenced our metadata information model and was developed further by taking the GSDEMSs framework into account. Fully specifying the E&I process in the metadata means that only a few pieces of information need to be saved in the data to ensure full traceability and reproducibility of E&I activities. The most important of these is a reference to the metadata object that defines the E&I method's parameters.

32. The metadata information model is currently under discussion at Statistics Finland and will be developed further. As for the data organization model, various organization principles will need to be considered and tested. Once both parts of the information model have sufficiently matured, we can begin developing applications utilizing it.

References

Fellegi, I.P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. Journal of the American Statistical Association (71), 17-35.

Hemedinger, C. (2013). Custom Tasks for SAS Enterprise Guide Using Microsoft .NET, SAS Institute.

Lehtinen, H. (2006). Dissemination of Statistical Data and Metadata: Process Based on Common Structure of Statistical Information (CoSSI), paper presented at the UNECE/Eurostat/OECD work session on statistical metadata, Geneva 3-5 April 2006,
Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2006/wp.20.e.pdf (Accessed 14 February 2017).

Pannekoek, J. and Zhang, L. (2012). On the general flow of editing. Oslo: UNECE Conference of European statisticians.

Platinovsek, R. and Piirainen, A. (2016). Metadata-enriched statistical production at Statistics Finland, paper presented at the Nordic Statistical Meeting, Stockholm 22-24 August 2016.
Available at: http://www.scb.se/Upload/NSM2016/theme4/B_2_Platinovsek_Piiranen.pdf
(Accessed 14 February 2018).

Rouhuvirta, H. (2001). On the structuring of statistical information, paper presented at the First MetaNet Conference, Voorburg 2-4 April 2001,
Available at: http://www.stat.fi/org/tut/dthemes/papers/structuring_statistical_information_2001.pdf
(Accessed 14 February 2017).

Rouhuvirta, H. (2004). An alternative approach to metadata – CoSSI and modelling of metadata, paper presented at the CODACMOS project meeting, Bratislava 7. October 2004,
Available at:
http://www.stat.fi/org/tut/dthemes/papers/alternative_approach_to_metadata_codacmos_2004.pdf
(Accessed 14 February 2017).

Statistics Canada. (2012). Banff Processor User Guide. Ottawa: Statistics Canada.

Statistics Canada. (2014). Functional description of the Banff system for edit and imputation, Version 2.06. Ottawa: Statistics Canada.

UNECE. (2015). Generic Statistical Data Editing Models. UNECE: Conference of European statisticians.

VTL, 2016. Validation and Transformation Language version 1.1, Part 2 - Reference Manual, SDMX Technical Working Group: VTL Task Force.

Wang, F. & Zaniolo, C., 2008. Temporal queries and version management in XML-based document archives. Data & Knowledge Engineering, Issue 65, p. 304–324.

Appendix

```
<edit editid="edt1" vstart="2017-01-11" vend=""
        direction="pass" implementation="banff" errorlevel="2" >
    <errormessagegrp>
        <errormessage vstart="2017-01-11" vend="2017-15-02" xml:lang="en">Total number
              of hens unequal to the sum of its parts</errormessage>
        <errormessage vstart="2017-15-02" vend=" " xml:lang="en">Total number
              of hens not equal to the sum of its parts</errormessage>
    </errormessagegrp>
    <sas_expression vstart="2017-01-11" vend=""
            expression="HEN_LT20+HEN_GE20+HEN_OTH=HEN_TOT">
        <varlist vstart="2017-01-11" vend="" type="expression">
            <listvariable vstart="2017-01-11" vend="" fieldName="HEN_LT20"/>
            <listvariable vstart="2017-01-11" vend="" fieldName="HEN_GE20"/>
            <listvariable vstart="2017-01-11" vend="" fieldName="HEN_OTH"/>
            <listvariable vstart="2017-01-11" vend="" fieldName="HEN_TOT"/>
        </varlist>
    </sas_expression>
</edit>
```

Figure 6: An example of an edit rule in XML form. In the example, the value of the errormesage element was changed on 15. February 2017.