Using Business Data in Customs Risk Management: Data Quality and Data Value Perspective

Wout Hofman, Jonathan Migeotte, Mathieu Labare, Boriana Rukanova, Yao-Hua Tan

Using Business Data in Customs Risk Management: Data Quality and Data Value Perspective

Wout Hofman[1], Jonathan Migeotte[2], Mathieu Labare[2], Boriana Rukanova[3(✉)], Yao-Hua Tan[3]

[1]TNO, The Netherlands
wout.hofman@tno.nl
[2]Belgian Customs, Belgium
[3]Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands
b.d.rukanova@tudelft.nl; y.tan@tudelft.nl

Abstract. With the rise of data analytics use in government, government organizations are starting to explore the possibilities of using business data to create further public value. This process, however, is far from straightforward: key questions that governments need to address relate to the quality of this external data and the value it brings. In the domain of global trade, customs administrations are responsible on the one hand to control trade for safety and security and duty collection and on the other hand they need to facilitate trade and not hinder economic activities. With the increased trade volumes, also due to growth in eCommerce, customs administrations have turned their attention to the use of data analytics to support their risk management processes. Beyond the internal customs data sources, customs is starting to explore the value of business data provided by business infrastructures and platforms. While these external data sources seem to hold valuable information for customs, data quality of the external data sources, as well as the value they bring to customs need to be well understood. Building on a case study conducted in the context of the PROFILE research project, this contribution reports the findings on data quality and data linking of ENS customs data with external data (BigDataMari) and other customs (import declaration) data and we discuss specific lessons learned and recommendations for practice. In addition, we also develop a *data quality and data value evaluation framework applied to customs* as high-level framework to help data users to evaluate potential value of external data sources. From a theoretical perspective this paper further extends earlier research on value of data analytics for government supervision, by zooming on data quality.

Keywords: Data Quality, Data Analytics, Value, Government Supervision, Customs, Risk Analysis.

1 Introduction

With the rise of data analytics use in government, government organizations are starting to explore the possibilities of using business data to create further public value [4]. This process, however, is far from straightforward: key questions that governments need to

address relate to the quality of this external data and the value it brings. In the domain of global trade, which is the domain of our investigation, with increased trade volumes due to for instance eCommerce, authorities like customs administrations rely more and more on IT-innovations to be able to perform their duties and deal with the large trade volumes. Customs administrations are responsible on the one hand to control the international trade for safety, security, and duty collection and on the other hand to facilitate fair trade and competition. Recently customs administrations around the world have turned their attention to the use of data analytics as part of their risk management framework.

Earlier research indicates that customs administrations face issues of low-quality data in customs declarations [10, 7, 18], which makes it hard to perform risk analysis and apply data analytics on such data [14]. Earlier e-government research has argued that government organizations can potentially create public value if they make use of business and other data beyond the data available in their government systems [4]. In the context of customs the goal is to improve its risk assessment processes and ensure public value such safety and security and revenue collection. More specifically, to improve the quality of data used in the customs risk assessment processes, customs starts exploring the value of business data provided by business infrastructures and platforms. But while external data sources hold a promise to enrich the customs data sets and provide better basis for analytics, the use of business data sources also raises issues and concerns such as: what is the quality of this data and is it of value [13,14]? Other relevant aspects are data protection issues and use of the data in real-time risk assessment. While earlier eGovernment research has recognized the issues of data quality of customs data and the potential of using business data, so far the studies have remained on a high level and limited research has examined the issues of how business data can help to address data quality of customs data and how this business data can add value to customs. Especially as nowadays there is a large number of digital infrastructures and platforms for sharing business data that is potentially useful for customs, there is a need for a systematic approach for evaluating these data sources, their data quality, and how they contribute to improving data quality of customs data and generate value. The empirical basis for our study is the research performed in the EU-funded PROFILE project where variety of external data sources were acquired and evaluated for their potential for addressing deficiencies in customs data. Based on the case findings, detailed case-specific lessons learned for customs related to data linking and data quality of customs and business data are defined. Based on the insights form the literature and the case we also developed a high-level *data quality and data value evaluation framework applied to customs*. The remaining part of this paper is structured as followed. In Section 2 we provide a literature review on data quality, big data, value of data, and customs risk management. In Section 3 we present our case study method. The results of our case analysis are presented in Section 4. In Section 5 we present our *Data quality and data value evaluation framework applied to customs.* We end the paper with discussion and conclusions.

2 Related Research on Data Quality, Big Data, Value of Data, and Customs Risk Management

There are different definitions of data quality. Building on Vetrò et al. [20], we will focus on discussing two definitions of data quality First, looking at the inherent data quality characteristics rather than the technical characteristics, the ISO 25012 standard refers to the inherent data quality as "the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions"[1]. Second, researchers have highlighted the "fitness for use" aspect when defining data quality [21, 17, 1], namely fit for use by a data consumer. For example, Strong et al. [17] argue that data quality research needs to look beyond the intrinsic properties related to data quality, and to focus also on the wider context and include data users. Strong et al. [17] distinguish between data producer or the party that produces the data, data custodian or parties that provide computing power to store and manage data and data consumers, or people who use the data. Strong et al. [17] define high-quality data as data that is fit for use by data consumers and they use four dimension categories and related dimensions to discuss high-quality data. The dimension categories and related dimensions are as follows: (1) Intrinsic data quality (accuracy, objectivity, believability, and reputation; (2) Accessibility data quality (accessibility, access security); (3) Contextual data quality (Relevancy, value-added, timeliness, completeness, amount of data); (4) Representational data quality (interpretability, ease of understanding, concise representation, consistent representation) [17]. Based on that they define data quality problem as "any difficulty encountered along one or more quality dimensions that renders data completely or largely unfit for use" [17, p.104]. While this research dates back almost two decades ago, it is still used today [e.g. 9] and is very relevant for the context of our study. Especially linking data quality to use relates closely to the issue of value of data.

With the proliferation of platforms sharing large amount of business data, big data is becoming increasingly interesting for government organizations who may want to make use of this external data sources. Big data can be seen as "the information asset characterised by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value" [3, p.133]. Literature has identified various challenges related to big data. For example Sivarajah et al. [17] distinguish among (1) data challenges that are related to the data characteristics (i.e. volume, velocity, veracity, variability etc.); (2) process challenges (e.g. cleansing, data aggregation and integration etc.); (3) management challenges related to such topics as privacy, data ownership, security, data governance etc.

Cai et al. [2] specifically discuss the issue of data quality in the era of big data and identify several challenges, namely: (1) difficulty of data integration due to diversity of data sources and complex data structures; (2) the difficulty to judge data quality in a reasonable time due to the tremendous volumes; (3) fast changing time of data which adds requirements for processing time; and (4) lack of unified data quality standard.

[1] https://iso25000.com/index.php/en/iso-25000-standards/iso-25012

However, apart from these more technical aspects, research has argued that there is a need for understanding of value of big data and analytics performed on this data [5,6, 15].

Looking at the eGovernment research specifically, the topics of data quality and big data are also of interest for the eGovernment research, where in the recent years these topics have gained attention in the area of open data research [8, 23, 19, 20]. Recent studies have also examined the use of big data and analytics for customs risk assessment [13,14], where a value analysis framework was proposed. The value analysis framework [14] examines the value of big data and analytics from three views (interdependency view, process view and collective capability building view). The interdependency view examines value as an interdependency among the work practice level where the data and the related analytics are deployed in the customs risk assessment process, the organizational level where policies and priorities are set, as well as the supra-organizational level where interactions with external stakeholders are defined. Especially the supra-organizational view is very relevant for our study, as this is the interface with external data providers and it is at this interface where the engagement with the data providers will take place and the potential of the external data for customs will be examined. The process view from the value framework focusses on the processes of capability building and capability realization. The collective capability building view focussed on possible collaborative arrangements. While Rukanova et al. [14] explicitly include the external data providers and mention data quality issues, these are discussed on a very high-level and no in-depth understanding is provided on how customs can better understand what external data sources have to offer and how they add value in improving data quality of customs data for risk assessment purposes. This is the area which will further explore in this research.

3 Method

For this study we used interpretative case study approach [22], where in our study we were interested in exploring and gaining an in-depth understanding of the possibilities for customs to make use of external business data. The empirical context for our research was provided by the PROFILE project, more specifically the Belgian Living Lab, where research is focussed on examining the potential of external business data sources to improve the quality of customs declarations used for safety and security risk assessment (i.e. the Entry Summary Declarations (ENS)). It is important to notice that the intention is to improve data quality in the new customs regulation for incoming cargo, ICS2, but additional data can always be of value. Any changes will have to be implemented according to ICS2 from 2024 onwards, leading to different customs data sets. The external data source that is explored in the PROFILE project is BigDataMari (BDM). Customs declarations that are included in this paper are for outgoing cargo, the so-called Entry Summary Declaration (ENS), and import. We follow several steps in the case analysis, namely: (step 1) identification of data requirements for customs risk assessment where we took the focus on security and ENS declarations and (step 2) development of a domain model of data that is needed for customs risk analysis. We use

this domain model as a basis of assessing data quality of the individual data sets: customs data (in step 3[2]) and the BigDataMari (in step 4). Based on the results of the previous step, in step 5 we made an analysis of the added value of customs and business data sets by linking those. In our assessment of data quality, we applied the approach of Strong et al. [17]: high-quality data as data that is fit for use by data consumers according to four categories with different dimensions:

1. *Intrinsic data quality* (accuracy, objectivity, believability, and reputation). We assume there are no challenges at this level, since we deal with known data sources.
2. *Accessibility* (accessibility, access security). This is not part of our study, since we received the data from their sources to study data linking.
3. *Contextual data quality* (relevancy, value-added, timeliness, completeness, amount of data). This is especially of interest from the perspective of data linking. The dimension 'value-added' will not be assessed, since it is assumed for each data set to be of value for particular processes like customs risk assessment or container shipment.
4. *Representational data quality* (interpretability, ease of understanding, concise representation, consistent representation). This category is also of relevance: are we able to interpret different data sets and create links. Concise representation will not be assessed for the same reason that 'value-added' is not assessed. Data sets are based on customs – and business standards for declarations and container shipping by sea respectively.

Based on these categories and dimensions Strong et al. [17] define data quality problem as "any difficulty encountered along one or more quality dimensions that renders data completely or largely unfit for use" [17, p.104]. In our approach, we will focus on the last two categories: contextual and representational data quality. This refers to the aspects of Interpretation and Re-usability of data as explained by the FAIR principles[3] (FAIR – Findable, Accessible, Interpretable, and Re-usable). These principles can be expressed by the aforementioned data quality categories and dimensions.

4 Case Analysis

This section describes the case by first assessing customs data requirements, secondly map them to a domain model and thirdly relate the various data sets to this domain model. We will assess data quality by exploring links of BigDataMari data with Entry Summary Declarations (ENS) and ENS with import declarations.

[2] We first started with ENS customs data, then did the mapping of the BigDataMari data, based on gaps we examined BigDataMari, and subsequently added an additional customs data source, i.e. import declaration data.

[3] https://www.go-fair.org/fair-principles/

4.1 Customs Data Requirements

Customs has developed a taxonomy of risks. The main groups of fiscal, economic, security, safety, drugs trafficking, and environmental risks are identified. These can be further decomposed, e.g. fiscal risks are decomposed into VAT, excise tax, anti-dumping, countervailing and customs duties.

Risks are perceived along two dimensions, namely the supply – and the logistics chain dimension. The supply chain dimension is on the movement of products and their components from a supplier to for instance an (eCommerce) buyer. These products flows can be triggered in different ways. Buy-sell is one approach, stock replenishment based on foreign production another, and Vendor Managed Inventory (VMI) yet another one. Stock replenishment can be based on internal production or purchasing orders, depending on the structure of a company. Stock replenishment can be supported by an third party or internal department providing purchasing services to for instance retail stores. On supply chain level, origin and destination, product composition, (invoice) value and product classification for customs purposes needs to be known, i.e. the Harmonised Systems code.

Logistics chains are on the movement of these products from an origin to a destination. Products become goods by packing them to facilitate transport. A variety of packaging types can be applied like boxes on pallets, that are put into containers. Each movement and unpacking or repacking can give a potential risk, e.g. fiscal risks like anti-dumping and economic risks lie IPR protection, quotas, and licenses of trade agreements.

This paper focusses on data requirements for customs risks assessment based on logistics chains. The following data is required:

- Transport flow – it concerns the flow of transport means with their various operations. It is decomposed into:
    - Itinerary - a timed sequence of transshipment locations or hubs (place of call) passed by a transport means for loading/discharging goods/containers (synonym: voyage, conveyance, trip)
    - Route - the use of the infrastructure taken by a transport means between any two places of call of its itinerary
- Cargo flow – it concerns the flow of package products, decomposed into:
    - Goods flow - timed sequence of transport legs and/or container tracks
    - Container track - timed sequence of transport legs for a container
    - Transport leg - transport of goods or containers between two adjacent (in time) locations with one transport means (e.g. POL, POD)
- Logistics chain structure - combination of physical flows (cargo/transport) and parties involved

These concepts will determine the data that has to be available from logistics chains.

4.2 Capturing the Domain of Interest

The domain of interest, the logistics domain, is characterized by the following concepts:

- Physical objects like packages, pallets, containers, and transport means like vessels and trucks.
- Locations where these physical objects can be transhipped, originate from, are destined for, or are stored under customs regime.
- Events representing associations in time between physical objects, like a container loaded on a vessel, or between a physical object and a location, like a container located on a terminal. The time can be in the past, present, or future, where a future time is represented from the expectation of a customer or the estimation of a service provider.
- Parties involved based on commercial transactions between these parties. For instance, a forwarder acting as service provider for shippers with the possibility to combine logistics flows for optimization.

These concepts construct so-called logistics chains in logistics systems. The system boundaries of our domain of interest are goods flows from outside into the European Union with containers via sea. This requires various transport modes, groupage and stripping, loading and discharge. The following figure shows the physical flow from an origin, i.e. PLA – Place of Acceptance, to a destination, i.e. PLD – Place of Delivery.

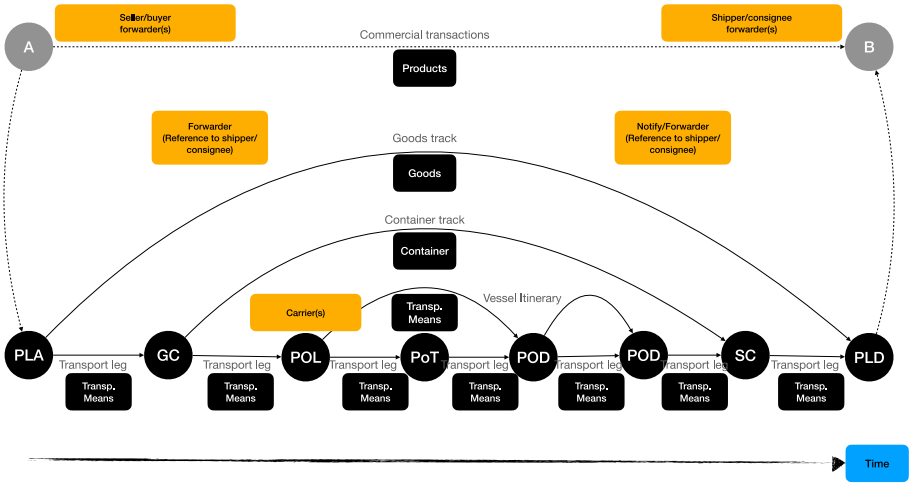

Fig. 1. Logistic chain.

Where the goods track is between PLA and PLD, container tracks are from a Groupage Centre (GC) to a Stripping Centre (SC) via a Port of Loading (POL; outgoing) and a Port of Discharge (POD; incoming). In-between a POL-POD, containers might be transhipped in a Port of Transhipment (POT). The figure shows two Ports of Discharge, where the first POD is the first port of call of a vessel in the EU and the second is where a container is discharged (these can be identical). The different transport legs visualize the movement of goods or containers from one location to another. At all locations, the relation between physical objects in time can be given. The goods – and container track

information can be based on any reporting of a transport means of its load/unload operation of particular goods or containers. For instance, a container track is based on reporting by a carrier from a GC to POL, another carrier from POL to POD and a final one from POD to SC. This might be the same carrier, i.e. the shipping line, coordinating transport from GC-POL and POD-SC, in which case it is called carrier haulage.

The figure shows commercial transactions, e.g. purchasing orders and replenishment orders, trigger the processes. It also shows that location A can be a third party purchasing organization acting on behalf of buyers at location B, where location A differs from PLA and B differs from PLD.

4.3 Mapping and Analysing the Quality of the Different Data Sets

For maritime transport, the ENS, Entry Summary Declaration, has to be submitted 24 hours prior to loading by a carrier in the port of loading, although there are exceptions. It provides data of containers that will be loaded and transported via sea to a European port. The Uniform Customs Code – Incoming Cargo System identifies two important European ports in this context, namely the first port of call of a vessel in an EU Member State (MS), the so-called Customs Office of Entry (COFE) and the actual port of unloading with the Customs Office of Unloading (COU). Figure a1 in Annex 1 shows the mapping of the relevant ENS concepts to the domain model. Table 1 presents the data quality of the ENS in relation to the data requirements expressed in the domain model.

The second data source is called 'BigDataMari' (BDM). The three data sets mentioned are linked, e.g. a shipping instruction refers to a booking and a container status message to a shipping instruction.

The functionality of these three data sets can be described as follows:

- Booking – an indication provide by a customer like a forwarder or a shipper to a carrier for the requirement of transport of a number of containers between two ports. Transport requirements are expressed by the number of Twenty feet Equivalent Units (TEU) to be transported from a POL to a POD with estimated dates/times.
- Shipping instruction – this data set contains details of containers and refers to a vessel for loading.
- Container Status Message – this data set has the actual status of container movements like loading in a POL onto a vessel and discharging in a POD, potentially from another vessel.

All relevant data elements in these three data sets can be mapped to the domain model, as shown in Figure a2 in Annex 1. The figure shows that a container can also be pickup at a GC, called place of receipt, and dropped off at an SC, called place of delivery. This relates to variants in logistics operations. The place of receipt can also be the PLA and the place of delivery can be identical to the PLD, in which case only container data is known and no data on goods is available. The latter case is known as carrier haulage with a Full Container Load (FCL).

The third data set is that of import, linked to incoming cargo movements. Import declarations are the basis for paying import duties. Undervaluation by an importer is one of the risks that needs to be assessed. Undervaluation can be a basis for unfair

competition. Incoming cargo movements can refer to a container that might be present in an ENS and/or import declaration.

The following table lists the data quality of each of these three data sets on the relevant aspects (section 2).

Table 1. Analysis of quality of individual data sets.

Data quality	**ENS**	**BDM**	**Import**
Contextual			
Relevancy	The data set provides a part of a vessel itinerary or container track of containers shipped to the EU or transhipped in an EU country to a non-EU country.	The data set provides a (part of) a container track. The data set contains sea containers shipped to a particular country.	This data set refers to commercial transactions. It provides details of products imported in the EU like their value.
Timeliness	Time is not related to the physical activity, but to sharing the data set (creation – and issue date). In ICS2, the actual date of departure and estimated time of arrival will be included.	The data set contains an expected container track and the actual one, given by Container Status Messages.	Submission of an import declaration is completely independent of a container track. A declaration can be submitted before container arrival or many days later.
Completeness	The part of the data provided is not complete. It does not cover vessel itinerary (which can change) or container track.	Data might be present in different data fields, e.g. a customs code (HS-code) might be in free text goods description. The container track might change due to changes in a vessel itinerary and/or transhipment. The data set does not cover all container flows to a particular country.	The import declaration does not fully relate to commercial transactions, e.g. multiple invoices per declaration or multiple declarations per invoice. With respect to the link to incoming containers, completeness is specified by the write-off process. This is an error prone (fuzzy) process, not always leading to complete links.
Representational			
Interpretability	A customs goods item is generated from a free text goods description, which does not make it reliable.	Fits with the logistics perspective of the domain.	The customs goods item will have an HS-code optimizing duty payment of an importer.
Ease of understanding	Goods item is to be interpreted from a customs perspective – based on the Harmonised Systems Code. This differs from logistics.	Fits with the logistics perspective of the domain.	Same as for ENS. The write-off (error prone and fuzzy) relates an import declaration to container track and the HS-code should relate to a (commercial) product.
Consistent representation	Different use of free text fields like parties involved.	Different use of data fields by different users, especially the free text fields, for instance goods description and parties involved.	Different use of free text fields like parties involved. An importer can differ from a consignee or buyer.

4.4 Linking Data Sets

Linking of data is based on (1) identifying data field similarities of and (2) finding data of those similar fields in two or more data sets.

Data field similarity relates to ontology alignment. An overview of ontology alignment approaches, algorithms, and indicators can be found in Mohammadi [11]. Data

fields similarity is by mapping the data fields of different data sets to our domain model with data requirements and using data quality assessment. Figure a3 in Annex 1 shows potential similarities of BDM and ENS and Figure a4 in Annex 1 the ENS to the import via the write-off process. The links on containers of BDM to import has not yet been assessed.

Table 2. Issues of linking.

Data quality		**ENS – BDM**	**ENS – Import**
Contextual			
	Relevancy	BDM – booking and shipping instructions are the basis for ENS declarations. These should map on container and potentially HS-code	Combination of ENS and import should provide more data on logistics – (ENS) and commercial flows (import). It might enable customs to configure data analytics for customs risk assessment using inspection results of both data sets. However, risk categories for both types of declarations might not completely overlap.
	Timeliness	There is a fuzzy relation on date/time between ENS and BDM, since ENS does not contain a logistics time	Matching on 'time' cannot be done.
	Completeness	BDM could complete the ENS data set with actual status date providing details of transhipment and itinerary deviations. However, BDM does not cover all incoming containers.	The write-off of import data would complete the data requirements. However, fuzziness and error prone of write-off makes it difficult to construct this completeness.
Representational			
	Interpretability	Linking is only feasible on containers.	Linking only feasible via the write-off process
	Ease of understanding	Customs goods item can only be used if it is copied from BDM.	The customs goods item (HS-codes) of ENS and Import will differ. Thus, they cannot be applied for linking, although they have the same definition.
	Consistent representation	Different use of free text fields does not allow to use these for linking.	Different use of free text fields does not allow to use these for linking.

All data sets are of 2018. The ENS and import declaration data sets are of one EU Member State. The BDM data set contains container shipments to two EU Member States, including the one for which we have the ENS and import declaration data sets. Figures of the actual mapping cannot be given in this paper, since these are confidential. Our findings are as follows: the union of ENS and import consists of 20% of containers in ENS and 6% of the containers in import declarations, and the union of ENS and BDM consists of 20% of the containers in ENS and 11% of the containers in BDM.

The relative low percentage of linking data sets relates to issues on data quality and similarities. Other reasons are due to the fact that a customs administration links a declaration to a previous one, using a write-off process. A further issue on the low percentage of linking relates to different procedures on a data level. For instance, import can relate to products coming into the EU via sea, air, road, or rail, where ENS only refers to sea. Import can also be transit from another EU Member State, in which case there is not a link to ENS. Of the number of actual incoming sea container, a number will also be put into transit to other Member States or outside the EU countries, e.g. to Norway.

5 Discussion

5.1 Discussion on the Case Findings and Case-Specific Lessons Learned

In this contribution, we have explored data quality categories and dimensions for assessing the potential value of linking different customs data sets and linking a business data set to a customs data set. The value is expressed in terms of data requirements for customs risk assessment.

There is value in linking the business data set BDM to the ENS data set. It provides more information on vessel itinerary and container track required by customs authorities. In case the import data set can be properly linked to commercial transactions and incoming cargo, there could also be value in linking the import data set to ENS. This value could be realized by training data analytics based on inspection results, but only if risk assessment and targeting of import and ENS is identical, i.e. on identical risks in a risk taxonomy.

Due to data quality issues of the different data sets and differences in customs procedures like ENS and import, it is difficult to link different data sets as shown by our analysis on data level. Only structured data fields have been used to identify links. This is to do with dimensions 'time' and 'ease of understanding', where the latter refers to differences in HS-code for different customs procedures. The extension of 'time' with actual departure date and estimated arrival data in ICS2 is expected to improve linking.

Concluding, the value of linking data is based on data requirements. We have formulated these data requirements in terms of our domain and identified similarities in different data sets expressed in the domain. By increasing the similarities, data completeness and thus data quality will improve. However, it requires to address other dimensions like 'time', 'consistency', and 'ease of understanding'. These can be improved by for instance encapsulating logistics events like present in BDM in customs declarations, at least for outgoing (ENS) and incoming declarations. These logistics events can be generated by each transport leg, part of an itinerary. A second improvement would be to create a better link between import and cargo flows. It requires stuffing of containers, packing lists with reference to products and their (invoice) value, and an explicit relation between a customs goods item and (commercial) products in import declarations.

In fact, all types of links need to be created, resulting in complex data sets. A proposed approach is to create a semantic model reflecting all potential associations as a basis for analysing data from different perspectives. Such a model is under development in the CEF (Connecting European Facilities) funded FEDeRATED Action (www.federatedplatforms.eu). Experiments for linking data sets to this model are performed by the H2020 PROFILE project.

5.2 Evaluation Framework for Data Value and Data Quality Applied to Customs

Beyond the case-specific findings, based on the insights from the theory and from the case we derived a general framework (Figure 2) that can be used by customs for reasoning about the data quality and data value of external data that can be used for customs risk analysis.

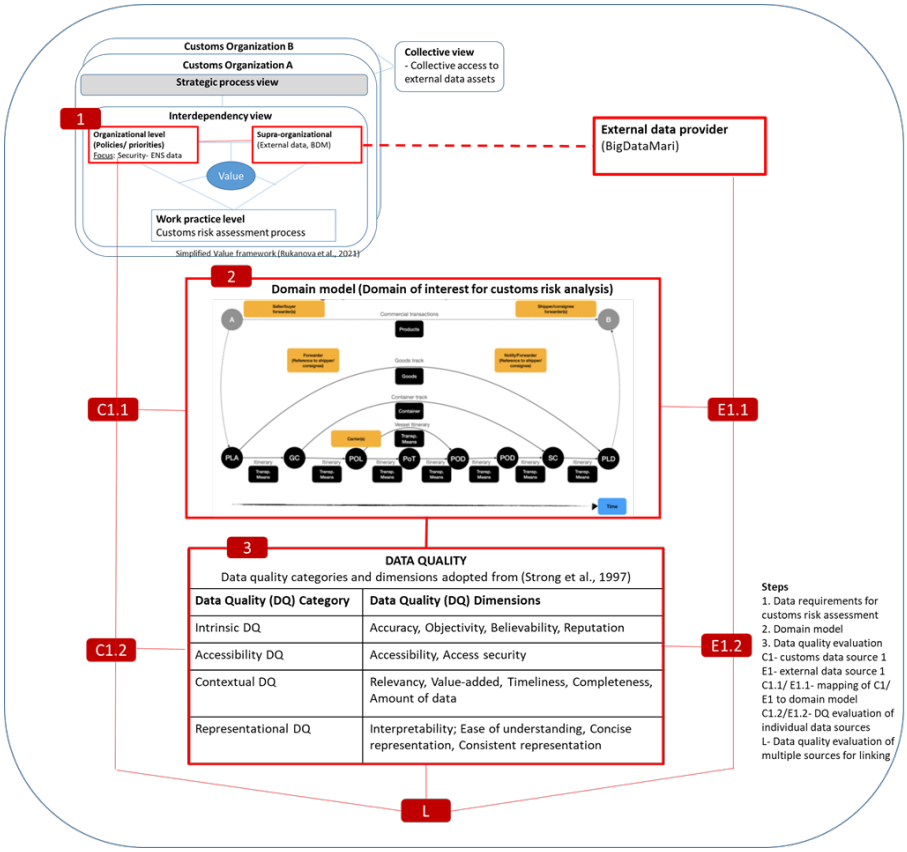

Fig. 2. Data quality and- value evaluation framework applied to customs.

Theoretically our framework combines the data value perspective (building on the value framework [14]), and the data quality categories and dimensions [17], and we explicitly added the domain model of the domain of interest for customs risk analysis derived from the case. In addition, our framework also captures the process steps that customs can follow when evaluating data quality and potential value of external data, as well as link to engagement with external data providers.

Starting form the value model and the organizational level, step 1 in our framework relates to data requirements for customs risk assessment. Step 2 in our framework focusses on the development of the domain model. In our framework we included the

domain model of the domain of interest for customs risk analysis that we derived based on the case. We consider this model to be quite generic for different risk analysis purposes however in practice it may be limited for some situations and may need to be extended. Step 3 focusses on data quality evaluation using the domain model. Subsequently, our framework indicates that specific customs data sources (e.g. Customs data source 1 -C1 or external data source 1-E1) can be mapped to the domain model (step C1.1 and E1.1 respectively). Subsequently a data quality evaluation of each of the individual customs and externals data sources can be performed individually by using the categories of Strong et al. [17]. (see C1.2 and E1.2 in Figure 2). Finally, data quality evaluation of multiple sources for linking can be also performed (marked with L in Figure 2).

The detailed illustration of how the data quality assessment of the individual data sources and the linking is done was already discussed in the case analysis and these detailed illustrations can be used to guide such analysis on other data sources as well.

6 Conclusions

In this paper we provided a detailed case study on linking customs and business data and assessing their data quality and – value in the context of data requirements. Based on our analysis, we provide detailed lessons learned and recommendations for practice. In addition, we developed an evaluation framework for data quality and – value assessment of linked data sets based on data requirements and a domain model. This framework can be used as a support tool for customs experts (nationally and internationally) involved in data analytics for customs. The framework allows to go in-depth in order to provide detailed insights into what kind of data can be found in specific business data sources and how it adds/ complements existing customs data. At the same time the framework allows for a level of abstraction from very technical details, which makes it also useful for people involved in data analytics projects at a managerial and policy level. The framework might, for instance, be useful to assess proposed changes of the ICS2 regulation. From a theoretical perspective this paper further extends earlier frameworks on value of data analytics for government supervision, by zooming in more specifically on data quality.

Our framework also has a number of limitations which open also possibilities for further research. First of all, our analysis and the resulting framework focus on specific aspects of the value framework of Rukanova et al. [15], mostly on the inter-dependency view and the organizational level. Further research can also examine what the effects of using the external data are at a work practice level, where data analytics based on combined customs and external data takes place. Furthermore, further research can also explore the effects taking the process view, as well as the collective view of the value model [15] into account.

Acknowledgement

This research was partially funded by the PROFILE Project (nr. 786748), which is funded by the European Union's Horizon 2020 research and innovation program. Ideas and opinions expressed by the authors do not necessarily represent those of all partners.

Annex 1. Using the Domain Model for Mapping of Data from ENS, BDM and Import Declarations

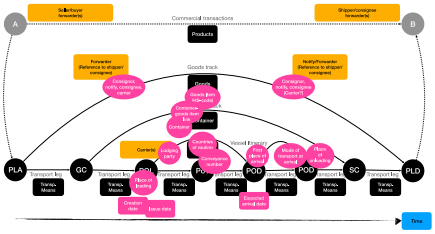

Fig. a1. Data in an ENS.

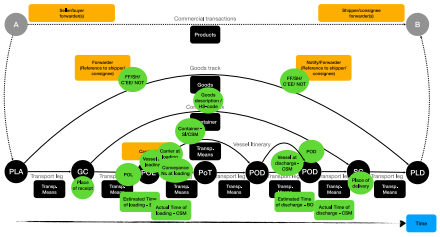

Fig. a2. Mapping BDM to the domain and linking data to ENS.

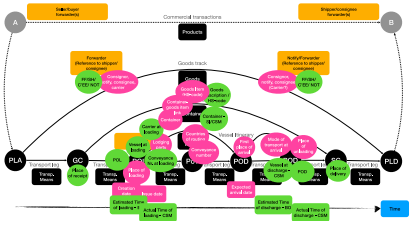

Fig. a3. Linking BDM to ENS.

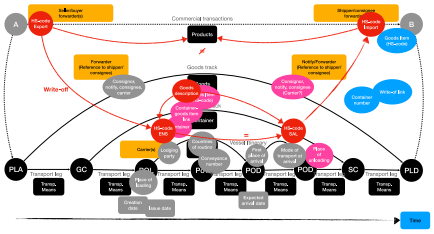

Fig. a4. ENS and import declarations.

References

1. Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* 41, 3, Article 16 (July 2009), 52 pages. DOI=10.1145/1541880.1541883 http://doi.acm.org/10.1145/1541880.1541883
2. Cai, L. and Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002
3. De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 65(3), 122-135.

4. Gil-Garcia, J. R. (2012). Towards a smart State? Inter-agency collaboration, information integration, and beyond. *Information Polity, 17*(3, 4), 269–280.
5. Grover, V., Chiang, R. H., Liang, T. P., & Zhang, D. (2018). Creating strategic business value from big data analytics: A research framework. *Journal of Management Information Systems, 35*(2), 388-423.
6. Günther, W.A., Mehrizi, M.H.R., Huysman, M., Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems, 26* (3), 191-209, https://doi.org/10.1016/j.jsis.2017.07.003.
7. Heijmann, F., Tan, Y.H., Rukanova, B., Veenstra, A. (2020). The changing role of Customs: Customs aligning with supply chain and information management. *World Customs Journal*, 14 (2).
8. Higgins, A., Klein, S.: Introduction to the Living Lab Approach (2011). In: Tan Y.H., Björn-Andersen N., Klein S., Rukanova B. (eds) *Accelerating Global Supply Chains with IT-Innovation*, pp.37-54. Springer, Berlin, Heidelberg.
9. Janssen, MFWHA., & van den Hoven, MJ. (2015). Big and Open Linked Data (BOLD) in government: A Challenge to Transparency and Privacy? *Government Information Quarterly,* 32(4), 363-369.
10. Juddoo, S., George, C., Duquenoy, P., & Windridge, D. (2018). Data Governance in the Health Industry: Investigating Data Quality Dimensions within a Big Data Context. *Applied System Innovation*, 1(4), 43. MDPI AG. Retrieved from http://dx.doi.org/10.3390/asi1040043
11. Klievink, B., Van Stijn, E., Hesketh, D., Aldewereld, H., Overbeek, S., Heijmann, F., & Tan, Y.-H. (2012). Enhancing visibility in international supply chains: the data pipeline concept. *International Journal of Electronic Government Research* , 8(4), 14–33.
12. Mohammadi, M. (2020). *Ontology alignment: Simulated annealing-based system, statistical evaluation, and application to logistics interoperability* . https://doi.org/10.4233/uuid:7d8ac519-f3f7-425f-82ce-1df481bc1c34
13. Pipino, L.; Yang, L.; Wang, R. (2002). Data Quality Assessment. *Commun. ACM*, 45, 211–218.
14. Rukanova, B., Henningsson, S., Henriksen, H. Z., Tan, Y.-H. (2018). Digital Trade Infrastructures: A Framework for Analysis. *Complex Systems Informatics and Modeling Quarterly* (14). DOI: 10.7250/csimq.2018-14.01.
15. Rukanova, B., Tan, Y.H., Slegt, M., Molenhuis, M., van Rijnsoever, B., Migeotte, J., Labare, M.L.M., Plecko, K., Caglayan, B., Shorten, G., van der Meij, O., Post, S. (2021). Identifying the value of data analytics in the context of government supervision: Insights from the customs domain, *Government Information Quarterly*, https://doi.org/10.1016/j.giq.2020.101496.
16. Seddon, P.B., Constantinidis, D., Tamm, T., Dod, H. (2017). How does business analytics contribute to business value? *Information Systems Journal*, 27(3), pp. 237-269.
17. Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 17*, 263-286.
18. Strong, D. M., Lee, Y., Wang, R. (1997) Data quality in context, *Communications of the ACM*, 40 (5), 103-110.
19. Tan, Y.-H., Bjørn-Andersen, N., Klein, S., & Rukanova, B. (2011). *Accelerating global supply chains with IT-innovation: ITAIDE tools and methods*. Springer Science & Business Media.
20. Umbrich, J., Neumaier, S., & Polleres, A. (2015). Towards assessing the quality evolution of Open Data portals. ODQ2015: *Open data quality: From theory to practice workshop* (Munich, Germany).

21. Vetrò, A., Canova, L., Torchiano, M., Minotas, C.O., Iemma, R., Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data, *Government Information Quarterly*, 33 (2), 325-337,
22. Wang, R.; Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst*., 12, 5–33.
23. Yin, R.. K. (1984). *Case study research: Design and methods*. Beverly Hills, CA: Sage.
24. Zuiderwijk, A., & Janssen, M. (2015). Participation and data quality in open data use: Open data infrastructures evaluated. *Proceedings of the 15th European Conference on eGovernment 2015*: ECEG 2015.