Innovative Data Analytics, Data Sources, and Architecture for European Customs Risk Management

D8.8 Policy, Research and Standardization Recommendations

Tan, Y.; Rukanova, B.D.; Alpsten, Anders; van Rijnsoever, Ben; Oosterman, Dion; Heijmann, Frank; Chen, Hao; Gisnås, Hallvar; Hintsa, Juha; Migeotte, Jonathan

Publication date
2021

Document Version
Final published version

Citation (APA)
Tan, Y., Rukanova, B. D., Alpsten, A., van Rijnsoever, B., Oosterman, D., Heijmann, F., Chen, H., Gisnås, H., Hintsa, J., Migeotte, J., Kacmajor, M., Kooij-Janic, M., Labare, M., Molenhuis, M., Johansson, R., Engoy, T., Gustavi, T., Männistö, T., Tsikolenko, V., ... Palaskas, Z. (2021). *Innovative Data Analytics, Data Sources, and Architecture for European Customs Risk Management: D8.8 Policy, Research and Standardization Recommendations*. PROFILE Consortium. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5f1716216&appId=PPGMS

PROFILE

Innovative Data Analytics, Data Sources, and Architecture for European Customs Risk Management

D8.8 Policy, Research and Standardization Recommendations

12/ 2021



Document Summary Information

Grant Agreement No	786748	**Acronym**	PROFILE
Full Title	Innovative Data Analytics, Data Sources, and Architecture for European Customs Risk Management		
Start Date	01/08/2018	**Duration**	36 months
Project URL	www.profile-project.eu		
Deliverable	D8.8 Policy, Research and Standardization Recommendations		
Work Package	WP8		
Contractual due date	M42	**Actual submission date**	
Nature	Report	**Dissemination Level**	Public
Lead Beneficiary	TUD		
Responsible Author	Yao-Hua Tan		
Contributions from	TUD, BCA, DCA, SCA, TNO, IBM, ILS, FOI, FFI, CBRA		
Remarks	A full and confidential version of this deliverable is published as an annex of the PROFILE D8.12 Exploitation Plan and Report (that confidential annex contains some customs sensitive information that could not be published in this public deliverable).		

Authors and contributors

Initials	Name	Organisation	Role
YHT	Yao-Hua Tan	TUD	Responsible Author
BR	Boriana Rukanova	TUD	Lead Author
AA	Anders Alpsten	SCA	Contributor
BvR	Ben van Rijnsoever	IBM	Contributor
DO	Dion Oosterman	TNO	Co-author
DvD	Dennis van Dijk	DCA	Contributor
FH	Frank Heijmann	DCA	Contributor
HC	Hao Chen	IBM	Contributor
HG	Hallvar Gisnås	FFI	Internal reviewer
JH	Juha Hintsa	CBRA	Contributor
JM	Jonathan Migeotte	BCA	Co-author
MK	Magdalena Kacmajor	IBM	Co-author
MKJ	Milena Kooij-Janic	TNO	Contributor
ML	Mathieu Labare	BCA	Co-author
MM	Marcel Molenhuis	DCA	Co-author
RJ	Ronnie Johansson	FOI	Co-author
TE	Thor Engoy	FFI	Co-author
TG	Tove Gustavi	FOI	Co-author
TM	Toni Männistö	CBRA	Co-author
VT	Vladlen Tsikolenko	CBRA	Co-author
WH	Wout Hofman	TNO	Co-author
WL	Wouter Langenkamp	TNO	Co-author
ZP	Zisis Palaskas	ILS	Co-author

Disclaimer

Copyright message

Executive Summary

The PROFILE project focussed on exploring and experimenting with data analytics innovations and solutions for customs risk management. Living Labs involving several customs administrations, data analytics providers and academia offered a real-life environment for developing and testing data analytics solutions and exploring the potential that external business data sources can offer.

This deliverable aims to reflect on the key lessons learned and recommendations that have been derived based on the work done in PROFILE and the results from the Living Labs. The main lessons learned and recommendations derived based on these lessons learned (presented in Sections 3-5, and consolidated in Section 6) are a result of a systematic process structured around activities related to the PROFILE policy, research and standardization recommendations (called RECS) activities. This process spanned over more than a year. A dedicated RECS committee was formed and met regularly and set-up structures to steer the process of partners to reflect on the lessons learned and recommendations and streamline recommendations into the flow of key deliverables. In addition to the RECS committee meetings, close interactions were maintained with the work packages and the different Living Labs to ensure convergence. Two dedicated PROFILE mini workshops (PMWs) with representatives from DG TAXUD were also used as a platform to present progress on technical results and recommendations and obtain feedback.

The lessons learned, as well as the recommendations based on these lessons learned, are structured logically into three broad themes, namely: Theme [1] Organizational aspects, discussed in Section 3; Theme [2] Linked data, semantic technologies and standards, discussed in Section 4, and Theme [3] Data analytics pilots conducted in the PROFILE Living Labs, discussed in Section 5. Subsequently, in Section 6, we clustered the recommendations into three broad areas, namely (a) Policy and organizational recommendations; (b) Standardization recommendations; (c) Further research recommendations. While for each of the three themes various detailed lessons learned and recommendations are formulated, several high-level observations can be derived. Looking at the three PROFILE themes, key high-level lessons learned and recommendation include:

Theme 1. Organizational aspects

Data analytics is not the holy grail - enforcement can never be done only by **data analytics**.

- You need **human expertise** to define risk thresholds.
- It is recommended to **conduct research in an environment closely connected to the operational environment** and to **collaborate** among customs administrations to arrive at sufficient amounts of data to develop reliable analytics. There is a need for clear legislative guidelines for data sharing, protocols for accessing and using data and data sharing environment to share data in a secure and simple way. Sufficient data preparation and understanding efforts should be taken into account for successful data analysis development. You need to prepare data first (e.g. select your risk focus (e.g. tax fraud versus narcotics) and variables) and only then consider more data and include more variables.
- On the management side, it is recommended to bring closer the **customs expertise and technical expertise** and encourage close collaboration between data scientists, customs officers with specific

domain knowledge and IT people. This, on a day-to-day, operational basis. You need people that can act as **mediators** (between customs and technical experts) internally and between the customs experts and external data analytics partners. And for the collaboration it is important to have the possibility to meet physically, identify and address both technical and legal issues related to data sharing early in the project, and find a common model for data sharing when it comes to exchanging or comparing declaration data from more than one country.

- It is also of key importance to do **careful scoping** and be explicit about the **assumptions**.

Theme 2. Linked data, semantic technology and standards

Using **external data sources,** next to the customs declaration data, may help to improve the development of data analytics.

- For **data linking** of customs data and external business data it is recommended to use a **semantic model**. The FEDeRATED model provides a useful ground for data linking. However, the use of such a model is still in its primary development stage. For applying the semantic model it is necessary to: (1) complement the node with a semantic adapter, (2) investigate the governance, (3) develop a roadmap for adoption of semantic architecture, and (4) address the issue of distributed data management.
- It is recommended to extend **EU CDM** and **UN/CEFACT** and make **links to ontologies** (e.g. links also to the FEDeRATED model). Having ontologies aligned with each other will enable data transformations.

Theme 3. Data analytics methods

- Based on experiments related toHS code prediction in the context of WP3 and WP5, various further research recommendations are provided. There recommendations are very technical in nature and relate to the specific context of where the experiments were conducted and they need to be considered having this technical, and case specific context into account. While we will not list all of them here, examples include: (a) consider, for **sentence embeddings**, using a model pre-trained on **customs data**; (b) **focus on chapter level** since some chapters can be more "descriptive" than others. (c) consider combining Random Forest method and Natural Language Processing (NLP) and explore opportunities to implement methods in the automated process; (c) consider running classifiers on foreign data.
- Based on experiments conducted related to the outliers and anomalies detection with declaration data by using Autoencoder (WP2/WP4), for further research is recommended to apply the autoencoder on (a combination of) different datasets, use semi-supervised learning (update and optimize the models using a feedback system), and multivariate analysis of the anomalies.
- Based on experiments related to automatic selection of customs declarations for inspection (WP4), further research can focus on reducing the number of input features, validation of the predictions generated for unlabelled data, investigation of the semantic value of entity embeddings used 'standalone'.

- Based on experiments conducted related to visual analytics it is suggested to conduct further research to explore the applicability of visual analytics to support a targeting officer and visualize differences in data sets.

The recommendations presented in this deliverable will serve as a basis for the PROFILE sustainability blueprint deliverable (D8.11), as well as targeted recommendations to specific stakeholder communities such as DG TAXUD that are planned as part of the PROFILE final dissemination activities.

Table of Contents

Table of Tables

Table of Figures

Glossary

ABBR	Description
2D	2 Dimensional

ABIEs	Aggregate Business Information Entities
AI	Artificial Intelligence
API	Application Programming Interface
ASBIEs	Associations between such entities Associated Business Information Entitles
AWS	Amazon Web Services
B/L	Bill-of-Lading
B2B	Business-to-business
BCA	Belgian Customs Administration
BCA	Belgian Customs Administration
BDI	Basic Data sharing Infrastructure
BIEs	Business Information Entities
BigDataMari	Name of an external data provider
CBM-RDM	Cross Border Management Reference Data Model
CCL	Core Components Library
CEF	Connecting Europe Facilities
CNN	Convolutional neural network
DA	Data analytics
DCA	Dutch Customs Administration
DG HOME	Directorate General Migration and Home Affairs
DG MOVE	Directorate General Mobility and Transport
DG TAXUD	Directorate General Taxation and Customs Union
DTLF	Digital Transport and Logistics Forum
ENS	Entry Summary Declaration
ETA	Estimated Time of Arrival
EU	European Union
EU CDM	EU Customs Data Model
FCL	Full Container Load
FEDeRATED	FEDeRATED Network of Platforms for Data Sharing in the Freight Transport and Logistics, www.federatedplatforms.eu

FFI	Norwegian Defence Research Establishment
FOI	Swedish Defence Research Agency
G2G	Government-to-government
GAIA-X	A Federated and Secure Data Infrastructure (www.gaia-x.eu)
GDPR	General Data Protection Regulation
HS	Harmonized System
HTML	HyperText Markup Language
IAA	Identification, Authentication and Authorisation
IBM	International Business Machines
ICS2	Import Control System 2
ID	Identification
IDSA	International Data Space Association
ILS	Inlecom
IPEX	Intellectual Property and Exploitation group in PROFILE
IT	Information Technology
IT	Information Technology
IT	Italy
JSON-LD	Java Script Object Notation for Linked Data
LCL	Less than Full Container Load
LL	Living Lab
ML	Machine Learning
MS	Member State
NCA	Norwegian Customs Administration
NLP	Natural Language Processing
NN	Neural Network
NOR	Norway
OWL	Ontology Web Language
PMW	PROFILE Mini Workshops
RDF	Resource Description Framework

RDF	Resource Description Framework
RECS	Recommendations[1].
SC	Supply chain
SCA	Swedish Customs Administration
SCS	Supply Chain Security
SHACL	Shape Constraint Language
SSA	Safety and Security Analytics
SWE	Sweden
TARIC	EU Customs Tariff
TECH6	Group comprising the PROFILE technical partners
TNO	Netherlands Organisation for Applied Scientific Research
TSC	Technical Steering Committee
TUD	Delft University of Technology
UCC	The Union Customs Code
UK	United Kingdom
UN/CEFACT	United Nations Centre for Trade Facilitation and Electronic Business
URL	Universal Resource Location
USE	Universal Sentence Encoder
WG	Working Group
WP	Work Package
XML	Extensible Markup Language

[1] In PROFILE, the work of the RECS committee and working groups refers to the work related to the Policy, Standardization and Research Recommendations.

1 Introduction

1.1 General introduction

The PROFILE project focussed on experimenting with data analytics innovations and solutions for customs risk management. Living Labs involving several customs administrations, data analytics providers and academia offered a real-life environment for developing and testing data analytics solutions and exploring the potential that external business data sources can offer. This deliverable aims to reflect on the key lessons learned and recommendations that have been derived based on the work done in PROFILE and the results from the Living Labs.

The main lessons learned and recommendations presented in this deliverable are a result of a systematic process structured around the activities for deriving policy, research and standardization recommendations (the so called RECS activities). These activities included virtual meetings to present Living Labs results, elaboration and discussions of preliminary conclusions and recommendations with partners, discussion with deliverable leaders how they will capture and include lessons learned and recommendations in key Living Lab deliverables. Next to that, a dedicated process was set-up to collect the lessons learned from the different work packages, via structures forms through pre-defined tables, that were distributed to the Living Lab partners for collecting inputs. The lessons learned and recommendations were subsequently further summarized, resulting in the recommendations presented in Sections 3-5, which were subsequently consolidated in Section 6.

The RECS activities spanned over more than a year, where a dedicated RECS committee was formed and met regularly. In addition to the RECS committee meetings, close interactions were maintained with the work packages and dedicated workshops were organized in the last quarter of the PROFILE project to ensure convergence. Two dedicated PROFILE mini workshops with representatives from Directorate General Taxation and Customs Union (DG TAXUD) were also used as a platform to present progress on technical results and recommendations and obtain feedback. This deliverable reports on the final outcomes of this consultation process. The key lessons learned are structured along three themes, namely:

(1) Theme [1]: Organizational aspects- this theme largely covers recommendations at organizational/ policy level that are relevant for stakeholders like DG TAXUD and EU Member State Customs Administrations.
(2) Theme [2]: Data linking, semantic technologies, and standards- this theme covers recommendations related to standardization, but also research and policy and is of relevance to key stakeholders such as Directorate General Mobility and Transport (DG MOVE) (Digital Transport and Logistics Forum (DTLF) where the FEDeRATED[2] semantic model is developed), as well as other standardization organizations like United Nations Centre for Trade Facilitation and Electronic Business (UN/CEFACT) and EU Customs Data Model (EU CDM).
(3) Theme [3]: Data analytics methods- this theme covers recommendations mainly related to further research and further research directions derived based on the data analytics experiments performed in PROFILE. These further research recommendations can be used for formulating follow-up

[2] www.federatedplatforms.eu

research projects and may be relevant for stakeholders such as Directorate General Migration and Home Affairs (DG HOME) as input for future technical research on data analytics for customs risk management and the security research agenda.

Recommendations within each of the three themes can relate to different aspects, such as policy, standardization and further research. In Section 6 we cluster the recommendations derived from Themes [1], [2], and [3] into policy, standardization and research recommendations.

For readability purposes, in this document we will not elaborate the recommendation for specific stakeholder groups. Such detailed documents with recommendations per key stakeholders, based on this deliverable, and extended with other project results, will be prepared and shared with PROFILE key stakeholders as part of the dissemination strategy and related dissemination and engagement events. This deliverable will also serve as a basis for the PROFILE sustainability blueprint which will be covered in Deliverable D8.11.

1.2 Structure of the deliverable

This deliverable is structured as follows. In Section 2 we first discuss the methodology followed for deriving the recommendations and the RECS way of working. Subsequently we introduce the Evaluation Framework developed by Belgian Customs (Deliverable D4.2), which allows us to place the recommendations (structured along the Themes [1]. [2], and [3] as discussed earlier) in the broader customs risk management context. Subsequently we use the three themes to present the PROFILE recommendations in Chapters 3-5 as follows. Chapter 3 presents the recommendations related to Theme [1] on Organizational aspects; Chapter 4 focusses on recommendations related to Theme [2]- Linked data, semantic technologies and standards; Chapter 5 focusses on recommendations related to Theme [3] on data analytics pilots. In Chapter 6 we cluster the recommendations discussed in Sections 3-5 into organizational and policy recommendations, standardization recommendation, and further research recommendations. Finally we end this deliverable with Conclusions.

2 Methodology

2.1 RECS way of working

In PROFILE, in the summer of 2020, dedicated activities were initiated to arrive at policy, standardization and research recommendations (the so-called RECS activities). The RECS activities relate to the Work Package 1 (WP1) activity of innovation management and WP8 activity on recommendations. As a WP1 activity, the RECS focus was to identify what are innovations/ lessons learned from the different WPs (in close collaboration with the so-called IPEX group on Intellectual Property and Exploitation that was initiated in PROFILE). As a WP8 activity, the RECS focus was on the consolidation of the lessons learned and drafting the recommendations.

The work of RECS was organized around the RECS Committee and the RECS Working Group (WGs). The RECS committee composed of the members of the Technical Steering Committee (TSC) and the members of the so-called TECH6 group, which included the PROFILE technical partners. The purpose of the RECS Committee meetings were to discuss the progress, identify the relevant topics for the RECS, and set-up high-level planning of the process of making the PROFILE lessons learned explicit and streamlining these towards recommendations. Towards the end of the PROFILE project, the RECS Committee Meetings focused on discussing the findings on the lessons learned. Key partners were invited to present their findings and lessons learned and the presentations were structures along the three themes that were introduced earlier, namely Theme [1] Organizational aspects; Theme [2] Linked data, semantic technologies and standards; Theme [3] Data analytics pilots. This process aimed to encourage parties to be more alert for identifying their findings and lessons learned and making them explicit. The RECS committee held meetings (on-line) on regular basis (approximately 3 times per year, once in 4 months). Furthermore RECS committee members were consulted when needed via e-mail or dedicated meetings/ workshops to provide comments on the RECS input (prepared by the members of the RECS Working Groups (WGs)).

Next to the general RECS Committee meetings, the work of the RECS related to the preparations and consolidation of inputs was done by the RECS Working Groups (WGs). The Primary role of the RECS WGs was to consolidate inputs from partners and prepare follow-up meetings. The work in the Working Groups was coordinated by working group coordinators. Working group coordinators have been identified for the Working Groups on: (a) Policy (CBRA, TUD); (b) Standardization (TNO, ILS), Research (CBRA, TUD). The Working Group coordinators kept close contact to ensure synergies and alignment.

The RECS, through the RECS WG coordinators, kept in close contact and aligned its activities with several other initiatives that were set in PROFILE, namely the Technical Steering Committee (TSC), the TECH6 (group comprising of the PROFILE technical partners) for discussing progress on the technical developments, and the dedicated IPEX group that was set-up in PROFILE and focused on tasks related to innovation and exploitation.

In the final year of the PROFILE project, a series of PROFILE Mini Workshops (PMW) were organized. Two of these workshops were related to RECS and also representatives of DG TAXUD were invited. The first PMW focused on presenting initial technical results from the project, the second PMW was focused on presenting the preliminary lessons learned and recommendations and obtaining feedback.

During the RECS Committee meetings, initially a long list of stakeholders was identified (see Annex 1). Subsequently a decision was made that RECS will focus on a short-list of key stakeholders. The key stakeholders include: DG TAXUD (the PROFILE customs partners have strong links to DG TAXUD); DG MOVE (in particular, the Digital Transport and Logistics Forum (DTLF)) (TNO maintains close links with this stakeholder group); UN/CEFACT (ILS is closely linked to this stakeholder group). We kept in touch with these key stakeholders via the key partners in PROFILE that had links to them.

The table below provides an overview of the RECS Committee meetings, as well as the two Profile Mini Workshops that took place. We also listed key topics discussed.

Table 1: Overview of RECS meetings

Meeting No	Date	Key issues discussed
1St RECS	6 July, 2020	The meeting focussed on the setting-up the RECS organization and way of working, initial identification of a list of stakeholders (see Annex 1), as well as initial identification of a short list of potential themes for recommendations and lessons learned. These served as the basis for the follow-up activities.
2nd RECS	3 November, 2020	This meeting discussed the alignment of RECS with other groups set-up in PROFILE, including IPEX with focus on Intellectual Property and Exploitation and TECH6 which comprised the technical partners and discussed progress on the technical experiments. Furthermore this RECS meeting was focus on updates on relevant developments such as the Customs Action Plan and implications for RECS, as well as a discussion for identifying the key stakeholders for the RECS.
3rd RECS	16 February, 2021	During the 3rd RECS meeting, key stakeholders for the RECS were selected from the broader stakeholder list. Content-specific presentation were given by WP3 on lessons learned on using data analytics in the context of eCommerce; Next to that RECS focussed on alignment of RECS and TECH6, inviting TNO to present lessons learned on data linking and the use of the FEDeRATED semantic model; alignment RECS and the IPEX activities (CBRA), and RECS and dissemination activities. Presentations focussed also on monitoring relevant developments related to research and policy. Initial discussion on the draft recommendations from the Belgian Living Lab also took place.
PMW with representatives from DG TAXUD	30 March, 2021	The focus on this PMW was to present initial findings from data analytics pilots developed in the Living Labs and initial lessons learned. Presentations were given by the technical experts of the Dutch, Belgian, and Sweden- Norway Living Labs.
4th RECS	11 May, 2021	The 4th RECS meeting focussed on revisiting the organizational recommendations from the Belgian Living Lab presented during the 3rd RECS meeting and discussing the initial organizational recommendation of the Dutch Living Lab. Furthermore, building on the collaboration RECS- TECH6, the technical partners from IBM, TNO and FOI were invited to present the technical lessons learned related to the data analytics pilots from the Living Labs (WP3, WP4, WP5).

5th RECS	15 June, 2021	The 5th RECS meeting focussed on standardization lessons learned. The convergence semantic model/ DTLF and EU CDM/ UN/CEFACT were discussed, as well as the lessons learned from using graph-based methods.
6th RECS	21 September, 2021	The 6th RECS meeting was the final RECS Committee meeting to discuss the lessons learned and recommendations. Specific attention was paid on the organizational lessons learned and recommendations and the lessons learned and recommendations related to semantic model/ DTLF and EU CDM/ UN/CEFACT. As the technical work in the Living Labs was still being finalized, additional activities were discussed to finalize the work on the technical lessons learned and recommendations related to the data analytics pilots.
PMW with representatives from DG TAXUD	3 November, 2021	Presentation to representatives from DG TAXUD of provisional lessons learned and recommendations related to organizational recommendations, the data analytics pilots lessons learned, as well as the lessons learned related to the semantic model and standards.

In addition to the formal meetings, the PROFILE partners collaborated very intensively, especially in the last half of year of the project, to consolidate the lessons learned. The consolidated lessons learned that served as a basis for the recommendations, were identified in close collaboration with the project partners through a series of iterations as part of the regular RECS activities. Structured forms were distributed to project partners to collect lessons learned and recommendations and these raw inputs from partners served as a basis for the consolidated lessons learned and recommendations. These consolidated lessons learned and recommendations were an intermediary step from the detailed knowledge and lessons learned (as accumulated in the different work packages and described in PROFILE deliverables) towards the more abstract and high-levels recommendations presented in the main body of this deliverable (sections 3-6). As such the lessons learned and the recommendations captured in this document present the best of our efforts to provide a consolidated view, while abstracting from the specific technical details.

It is important to highlight that the PROFILE work package deliverables present also a key source of information as they contain a very detailed documentation of the work done. Partners were encouraged to document as far as possible the lessons learned as part of their own deliverables.

2.2 Themes used for structuring the recommendations

As discussed in the Introduction, the PROFILE recommendations as presented in Section 3-5 of this deliverable are structures along three themes as follows:

- Theme [1]: Organizational aspects
- Theme [2]: Data linking, semantic technologies and standards
- Theme [3]: Data analytics pilots

Recommendations related to each of these themes further capture a number of different topics. The table below provides an overview of the topics covered per theme. These topics are used also in the recommendations tables in Sections 3-5 to structure further the recommendations related to each of the Themes.

As illustrated in Table 3, in Theme [1] we discuss the topics of: (1) Machine learning in the customs context; (2) Data quality, and data preparation, and the effort for data preparation; (3) Building-up customs business expertise; (4) Long-term view on innovation; (5) Upscaling to the operational environment; (6) Collaborative data analytics development for customs administrations.

In Theme [2] the discussion relates to: (1) declaration data and external data sets; (2) semantic model; (3) data exploration and the use of semantic technologies (e.g. graph-based methods), and discussions about relations to EU CDM (4) and UN/CEFACT (5).

Theme [3] on data analytics pilots discusses: (1) Harmonized System (HS) code predictor; (2) Random Forest commodity code (HS code) classifier, Natural Language Processing (NLP) and 2 Dimensional (2D) visualisation for outlier detection; (3) Outliers and anomalies detection with declaration data by using Autoencoder; (4) Automatic selection of customs declarations for inspection, and (5) Visual analytics, (6) Other further research related to eCommerce platforms, and data pipelines.

Table 2: Themes and topics for structuring the recommendations

Themes	**Theme [1]** **Organizational aspects**	**Theme [2]** **Linked data, semantic technology, and standards**	**Theme [3]** **Data analytics pilots**
Topics	1. Machine learning in the customs context 2. Data quality, data preparation and the effort for data preparation 3. Building-up customs business expertise 4. Long-term view on innovation 5. Upscaling to the operational environment 6. Collaborative DA development for customs administrations	1. Declaration data and external data sets 2. Semantic model 3. Data exploration and the use of semantic technologies (e.g. graph-based methods) 4. Relation to EU CDM 5. Relation to UN/CEFACT	1. HS code predictor 2. Random Forest commodity code (HS code) classifier, NLP and 2D visualisation for outlier detection 3. Outliers and anomalies detection with declaration data by using Autoencoder 4. Automatic selection of customs declarations for inspection 5. Visual analytics 6. Other further research related to eCommerce platforms, and data pipelines

2.3 Using the Evaluation Framework[3] for placing the lessons learned and the recommendations in the broader customs risk management context

While we derived a long list of recommendations related to the Themes and Topics as presented in Table 2, it is important to position and relate these back to the broader context of customs risk management. To do that we will make use of the Evaluation Framework developed by Belgian Customs in the context of the Belgian Living Lab.

This Evaluation Framework of Belgian Customs is visually represented in the form of a triangle as can be seen in Figure 1.

The framework consists of a number of elements as follows. In the center of the framework is the customs risk management strategy as the key reference point, with three key dimensions gravitating around it:

- the data quality dimension,
- the technical expertise dimension,
- the customs business expertise.

The triangle highlights the high level of interdependency between the different dimensions as well as the crucial points where they interact with each other. The critical points at the interaction of two dimensions are represented as the three corners of the triangle as follows:

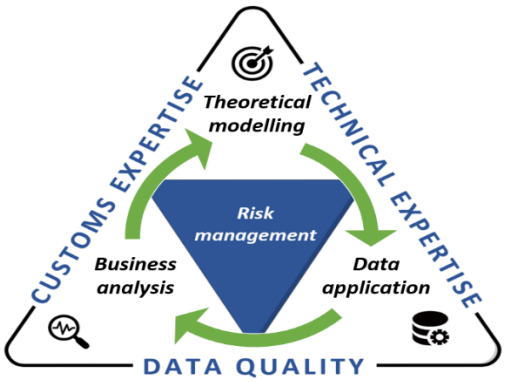

Figure 1: Data Analytics Evaluation Framework for Customs Risk Management

- Business analysis (on the interaction between data quality and customs expertise)
- Data management (on the interaction between data quality and technical expertise)
- Theoretical risk models (on the interaction between customs expertise and technical expertise).

For developing data analytics that will reach specific objectives of the customs risk management strategy, it is important that data quality, customs expertise and technical expertise are closely aligned to contribute to the specific customs risk management objective. This is an iterative process, where through iterative loops these dimensions get better aligned to lead to better data analytics that are fit for the objectives.

[3] The text in this section is adopted from D4.2. General overview on upgraded risk indicators and profiles (Belgian Living Lab)

The dimensions are not limited to a single organization but can also be used for reflecting on issues crossing organizational boundaries. For example the data quality dimension can be used to reason about the data quality of the customs data, as well as data quality of external data sources and data linking. The technical expertise dimension can also be used to reason about both internally available technical expertise and external data analytics expertise provided by external technical partners.

In Figure 2 we plotted the three Themes and related Topics that we used to structure the recommendations around the Evaluation framework of Belgian Customs. We positioned Theme [1], which addresses topics related to the technical aspects in the data analytics pilots, close to the *technical expertise* dimension of the Evaluation Framework. Theme [2] has a lot to do with issues related to data quality and data linking, we therefore positioned it close to the *data quality dimension* of the framework. Theme [1] covers different organizational aspects that stem more from the specifics of the customs environment, we therefore placed it close to the *customs expertise* dimension of the framework.

As illustrated in the Evaluation framework, these customs expertise, technical experts and the data quality dimensions need to come together in order for data analytics to support customs risk management, which is at the heart of the Evaluation Framework. Therefore we linked Themes [1], [2], and [3] to the core of the Evaluation Framework. This is because the lessons learned and recommendations, even if they have a strong link to one of the dimensions of the framework , they inevitably touch upon and reflect on aspects related to the other dimensions as well.

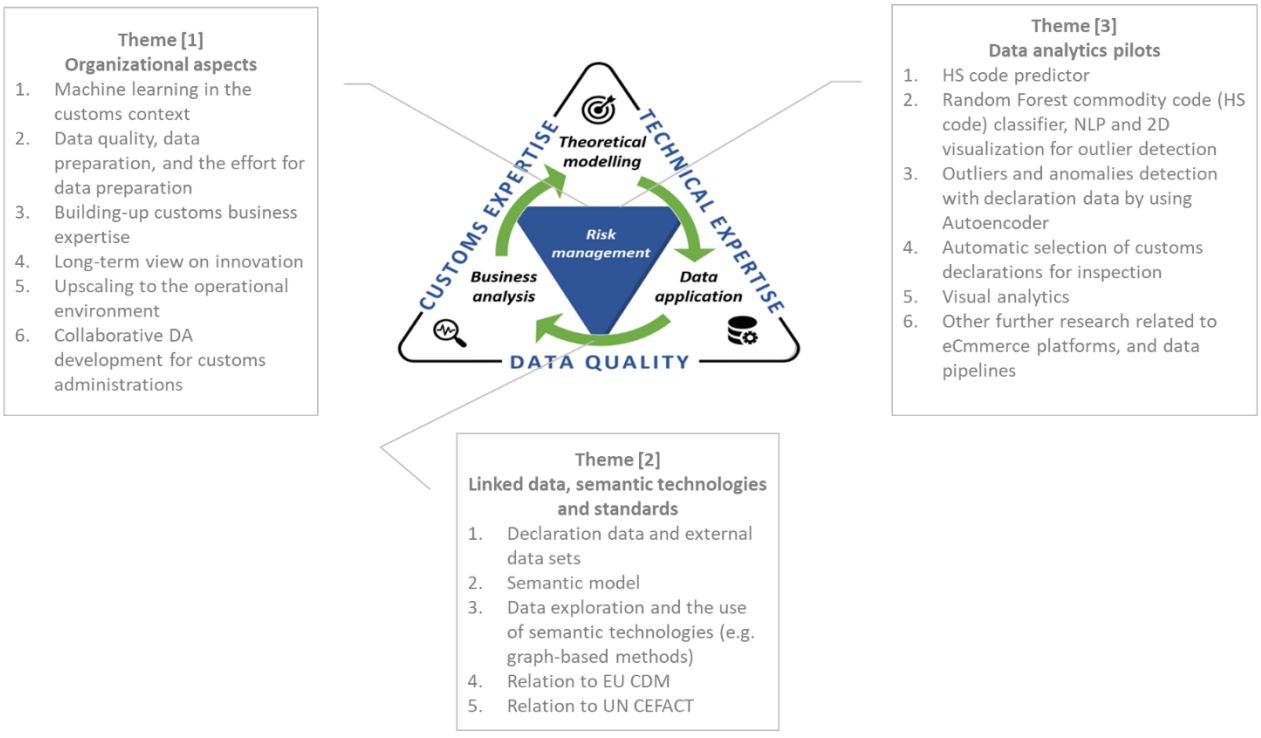

Figure 2: Positioning the themes for the recommendations along the Evaluation Framework

In the following chapters 3-5 we present the PROFILE recommendations, structured along Themes [1], [2], and [3] as listed in Table 2. In chapter 6 we further structure the recommendations into: (1) organizational and policy recommendations, (2) standardization recommendations; and (3) further research recommendations. Overall, the recommendations related to Theme [1] represent mostly organizational and

policy recommendations, the recommendations from Theme [3] represent largely further research recommendations about further technical research on data analytics based on the Living Lab experiments. The recommendations around Theme [2] are more spread along the three categories and cover policy, standards and further research recommendations.

As discussed in the Introduction, the recommendations as captured in Sections 3-6 of this deliverable will feed further into deliverable D8.11 on the PROFILE sustainability blueprint. They will be also used as a basis for deriving targeted recommendations for specific stakeholder groups, which will form the basis of tailored dissemination materials for these stakeholder groups that will be used in further PROFILE dissemination and engagement events.

3 Theme [1] Organizational aspects

Table 3 below captures the recommendations related to Theme [1]. The table is structures as follows. The first column provides a number related to the Theme. In this section we discuss only recommendations related to theme [1], this is reflected in column one. As discussed earlier in Section 2.2 (see also Table 2) each Theme is further structured into topics. Columns two of Table 3 reflect the topic number and topic description related to Theme [1] as listed in Table 2 of Section 2.2. Column three of Table 3 is called Findings. To keep the tables concise, here we provide a very short description of findings and context for the specific recommendations. Column four of Table 3 contains the specific recommendations per topic. It is important to mention that per topics often more than one recommendations are provided. These recommendations are further numbered as well for ease of referencing. The recommendations covered in Table 3 are predominantly organizational and policy recommendations. The recommendations text as presented in column four of Table 3, as well as the related structured numbering is used as input for the consolidated table with recommendations as presented in Section 6.

Table 3: Recommendations related to Theme [1]: Organization and policy

Theme	Topic	Findings	Recommendations
[1]	1. Machine learning in the customs context	There is difference between the physical world and the digital world. Even with all the data in the digital world there is still uncertainty how well this data reflects the physical world and what one may find in the physical world. This uncertainty is a given.	[1] 1.1. Data Analytics is not a not the holy grail- acknowledging uncertainty and the human factor. Data Analytics does not work as a standalone - enforcement can never be done only by data analytics. You have also the physical world and you need to have an overall strong processes and business knowledge in order to leverage the potential of data analytics.
		The outcomes of the analytics is not binary. Non-compliance risks needs to be seen as a scale and even a low risk is a risk.	[1] 1.2.You need human expertise to define thresholds.
		Because of the uncertainty, it is hard to develop Data Analytics for Customs Risk Management.	[1] 1.3. Conduct research in environment closely connected with the operational environment.
			[1] 1.4. Focus on flexible and dynamic data analytics development approaches; Due to the changing nature of the customs landscape, it is probably more efficient to work on either on developing simple models and algorithms that are easily retrained, or focusing on parameters that are stable over time.

			[1] 1.5. Collaborate among customs administrations to arrive at amount of data sufficient to develop reliable analytics.
		Information available in customs declarations is limited for development of data analytics for customs risk analysis	[1] 1.6. Using external data sources next to the customs declaration data may help to improve the development of analytics and risk management. In terms of Policy recommendations, one direction is to drive the way towards the closer linking and co-operation between Agencies (i.e. customs) and Freight Forwarders, carriers, eCommerce sellers and eMarkets, having a more transparent access to the transaction data, providing a right set of initiatives and the motivation for this to happen.
		Rules and regulations and the definitions related to customs procedures, prohibitions and restrictions are so different and so complex that is very difficult to apply data analytics as the sample sets for each case are too small and too different. Every time a new regulation is added, it also distorts the trends a bit more. Adding new data sources does not solve the problem.	[1] 1.7. Further simplification of the system of duties, prohibitions and restrictions can make it easier to develop data analytics.
[1]	2. Data quality, data preparation, and the effort for data preparation	More data does not necessarily mean better data analytics. Data needs to be prepared and understood in order to bring added value.	[1] 2.1.: You need first to prepare data (e.g. select your risk focus (e.g. fiscal fraud versus narcotics) and variables) and only afterwards take more data into account.
		It appeared that most efforts were needed for data preparation and understanding. Only after this is done well one can start with data analytics.	[1] 2.2. Sufficient efforts for data preparation and understanding should be taken into account for successful data analytics development.
		Collaborating with external data analytics developers and other customs for customs risk management turned to be very difficult in the project due to legislative constrains.	[1] 2.3. There is a need for clear legislation guidelines concerning the application of General Data Protection Regulation (GDPR) and Artificial Intelligence (AI) Act in the customs risk management environment in such cases.
			[1] 2.4. There is a need for data access and use protocols within the government services to easily share data with external data analytics development partners.

			[1] 2.5. There is a need for a trusted environment with trusted protocols for sharing data in a secure and simple manner, internally across customs departments, across customs administrations, and between customs and external data analytics providers.
		Customs data has issues with data quality	[1] 2.6. For improving data quality explore the possibilities for customs providing services to businesses. •To improve the declaration data quality (e.g. using HS code predictor to suggest inaccuracies in filled-in HS codes), and •in the long run possibilities for prefilled declarations (e.g. using also business data).
[1]	3. Building-up customs business expertise	A lot of time and efforts can be spent if the assumptions are not made explicit and the scope and objectives are not sufficiently aligned.	[1] 3.1. When you start with a data analytics innovation do careful balancing between scope and objectives, and be explicit about the piloting assumptions.
		Data analysts typically have limited customs risk domain expertise.	[1] 3.2. The best way to build up customs business expertise efficiently is that data analysts collaborate closely with customs risk domain experts.
			[1] 3.3. You need people that can act as mediators/translators/liaisons internally between the technical and business people, as well as externally between customs and the external data analytics providers.
[1]	4. Long-term view on innovation	Upscaling of Data analytics innovations will require a series of projects to allow for the continuity from initial Research and Development (R&D) towards implementation and upscaling.	[1] 4.1. It would be beneficial if funding programs allocate funds for multi-phase innovation tracks, in order to push TRL further towards implementation.
[1].	5. Upscaling to the operational environment	Upscaling Data Analytics innovations to the operational environment will require organizational changes, IT changes and adoption from the customs experts involved in customs risk management.	[1] 5.1. It is recommended to consider three broad areas when thinking about upscaling of data analytics innovation for customs risk management: • How to organize for data analytics innovation: Consider transforming the existing organization(with separate IT, risk management, data analytics and detection technology departments) towards a new data-driven organization where these departments are better aligned and work together.

			• How to bring data analytics innovation to the real-time environment: Consider creating data sharing infrastructures that enable using different data sources to generate new insights needed for customs risk management real-time. • How to bring data analytics innovation to the operational customs officers: Consider issues like explainability, training and human resource development.
[1]	6. Collaborative Data Analytics Development between Customs Administrations	Some challenges observed in the collaboration between customs administrations included limited possibilities for physical meetings, technical and legal issues.	[1] 6.1. It is valuable to meet physically and work together to a greater extent. [1] 6.2. Identify and address both technical and legal issues related to data sharing early in the project, so that during the project itself you can focus more directly on innovations and technical solutions. [1] 6.3. Find a common model for data sharing (e.g. fields, anonymization, restrictions) when it comes to exchanging or comparing declaration data from more than one country.

4 Theme [2] Linked data, semantic technologies and standards Recommendations

Table 4 below provides a summary of the recommendations related to Theme [2] Data linking, semantic technologies and standards. The structure of the table follows a similar format as Table 3. The recommendations covered in Table 4 relate to organizational, standardization and research aspects. The recommendations text as presented in column four of Table 4, as well as the related structured numbering is used as input for the consolidated table with recommendations as presented in Section 6.

Table 4: Recommendations related to Theme 2: Linked data, semantic technologies and standards

Theme	Topic	Findings	Recommendations
[2]	1. Declaration data and external data sets	Customs declaration data alone is not always sufficient for customs risk management.	[2] 1. It is recommended that customs declaration data is enriched with external data for improving customs risk management and that there is a closer collaboration between customs and external data providers.
[2]	2. Semantic model	It is hard to link customs data and external data	[2] 2.It is recommended that the use of a semantic model (e.g. the Digital Transport and Logistics Forum (DTLF) semantic model developed by the FEDeRATED project) can enable better linking of customs data and external data. Still the semantic model is under development and in a pilot phase. For applying the semantic model it is necessary to: •create data pipelines based on a 'node' concept sharing linked (event) data; existing platforms will act as 'node' in line with the Data Governance Act, •complement the node with a semantic adapter for integration with existing solutions , •install governance in a EU context, and •develop a roadmap for adoption of semantic architecture by customs administrations.
[2]	3. Data exploration and the use of	Data exploration and data preparation of external data sets is challenging. Graph-based methods and	[2] 3. It is recommended to explore further the possibilities of semantic technologies that include graph based solutions for visual analytics, data preparation and data linking.

	semantic technologies	semantic technologies show promising results for data exploration and data preparation.	
[2]	4. Relation to EU CDM	EU CDM is limited when it comes to linking customs data to external supply chain and logistics data contained in external business data sources	[2]-4 It is recommended to extend EU Customs Data Model (EU CDM) so that it can integrate/ adopt the semantics of supply chain logistics data eventually aligning with DTLF, having in mind a short, mid, and long-term strategy. This can be done: • In the short run by integrating the concept 'time' related to logistics activities and applying these structures only in binary relations between a customer and logistics service provider. • In the mid-run restructure the EU CDM into a structured set of ontologies and align these with the DTLF semantic model (ontology). • Long term recommendation is to adopt one semantic model and semantic architecture, with the objective to phase out existing data sharing mechanisms and create a more flexible and extendible environment.
[2]	4. Relation to UN/ CEFACT	While UN/CEFACT standards are widely used by businesses, it is hard to use data structured along these standards for data linking and data analytics.	[2]-5 It is recommended that the UN/CEFACT standard makes steps towards ontologies, namely: • Medium term recommendation is to restructure the UN/CEFACT buy, ship, pay model into a structured set of ontologies and align these with the DTLF semantic models (ontologies). Combined with the similar recommendation for EU CDM, having these ontologies aligned with each other will enable data transformations. • Long term recommendation is to adopt one semantic model and semantic architecture, with the objective to phase out existing data sharing mechanisms and create a more flexible and extendible environment .

5 Theme [3] Data analytics pilots recommendations

Table 5 below provides a summary of the recommendations related to Theme [3] Data analytics pilots. The structure of the table follows a similar format as Tables 3.and 4 discussed earlier with the difference that in Table 5 we do not discuss findings, as it is difficult to describe the set-up of the pilots in a concise way and with sufficient detail. Most of the recommendations in Table 5 are future research recommendations that were derived based on the different experiments performed in the PROFILE Living Labs. The recommendations presented in Table 5 are very technical in nature and stem from very specific experiments. They should therefore be interpreted in this narrow context. Due to the highly technical nature of these recommendations, they are predominantly targeted to experts involved in developing data analytics for customs risk management. Nevertheless, they may have also a wider relevance for managers and policy makers that set directions and projects for data analytics at national and EU level.

Table 5: Recommendations related to Theme [3]: Data analytics pilots

Theme	Topic	RECOMMENDATIONS
[3]	1. HS code predictor	[3] 1.1. For sentence embeddings, consider using a model pre-trained on customs data that takes into account specific abbreviations and text truncations used in the goods description. In order to avoid training big models like USE from scratch (due to the process complexity, significant computational power and extremely large datasets), a solution might be to find a way to use transfer learning on models e.g. by studying the model training process and adding Customs data as an extra layer to the model .
		[3] 1.2. Consider using language detectors to separate goods descriptions by language (e.g. NLP models like spaCy[4], Language Identification models like CLD3[5] etc.).
		[3] 1.3. Consider combination of methods using also numerical and categorical data in customs declaration e.g. price, weight, value, loading location (place + country), consignee, consignor etc.
		[3] 1.4. Modify the approach to classification, e.g. it could be based on sentence embedding plus additional binary classifiers and dictionaries associated a with every HS code group to detect the commodity nomenclature.

[4] https://spacy.io/usage/models

[5] Language Identification using the 'fastText' package

		[3] 1.5. Focus on HS code chapter level since some chapters can be more explanatory details than others.
		[3] 1.6. Study the classification methods (principles) and databases used by commodity certification/classification authority
		[3] 1.7. Goods description, a free text field, would benefit from more structuring and further standardization.
		[3] 1.8. Consider HS code predictors that have been developed by commercial parties, taking into account that they may be biased to have (optimal) results for businesses.
		[3] 1.9. Safeguards on material composition, quality and assurance are essential for customs and also in the context of circular economy. Further research can examine how the HS code prediction models developed in PROFILE (WP3) can be further extended to also better differentiate similar goods that have different material composition and fall under different HS codes and nomenclatures.
[3]	2. Random Forest commodity code (HS code) classifier, NLP and 2D visualization for outlier detection (WP5)	[3] 2.1 Random Forest and NLP based methods were used to predict commodity codes. The methods have potential for further development. To increase accuracy one should consider combining the methods with each other.
		[3] 2.2 The NLP and Random Forest classification methods could theoretically be incorporated directly in the automatic process so that declarations can be analysed in real time. This has not been tested in this project though and can be a subject for further research.
		[3] 2.3 The classifiers were set up identically, but had higher performance for the Swedish data in all cases. This cross-border ”meta” result suggests that since the characteristic patterns of the Norwegian data are harder to learn, the quality of the Norwegian data could possibly be enhanced, perhaps through improving instructions to declarants.
[3]	3. Outliers and anomalies detection with declaration data by using Autoencoder (WP2/WP4)	[3] 3.1. Apply the autoencoder on (a combination of) different datasets. Improvement of the output of the autoencoder requires interaction with targeting officers. There are various ways to include them, but the most obvious way is to provide targeting officers with the output of the autoencoder in their work processes and let them provide feedback. This approach is taken in the targeting dashboard that has been developed.
		[3] 3.2. Apply semi-supervised learning (update and optimize the models using a feedback system)
		[3] 3.3. Apply multivariate analysis of the anomalies.
[3]	4. Automatic selection of customs declarations for inspections (WP4)	[3] 4.1. Reducing the number of input features. There is a trade-off between retaining features that are valuable source of information and removing features that contribute most strongly to dataset shift. Future work may leverage explainability for getting better insight into the role of different features in generating predictions and for identifying the optimal set of features.

		[3] 4.2. Validation of the predictions generated for unlabelled data. Entity embeddings – meaningful representations of categorical features – are potentially powerful tool for extracting features that are truly representative of full dataset (not affected by the bias). The results of the experiments with unsupervised learning will require validation in the field, i.e., real feedback from the inspectors on the quality of the outputs of unsupervised learning.
		[3] 4.3. Investigating the semantic value of entity embeddings used 'standalone'. Entity embeddings (learned representations of categorical features) form a semantic vector space, in which the similarity among entities is represented by the proximity of the vectors. The semantic representations of entities that can facilitate insights into mutual relationships between these entities. Future work should explore vector space formed by distributed representations of most important features (e.g., operators or commodity codes) to validate that they indeed reflect real-life dependencies among entities.
[3]	Visual analytics	[3] 5.1 Explore further the applicability of visual analytics to support a targeting officer. A first effort is made in the targeting dashboard (WP2/WP4) towards visual analytics. This could visualize anomalies in trade flows that could be of interest to targeting officers.
		[3] 5.2 Visualize differences in data sets. The objective is to compare different data sets on particular features and visualize these differences. Examples of differences are on values of 'weight', 'time/duration' and 'route/itinerary/locations'. The actual relevant differences have to be defined by targeting officers.
[3]	6. Other research related to eCommerce platforms, and data pipelines	[3].6. 1 Follow-up research can conduct a comparative study by examining different eCommerce platforms and providing an overview of which data can be of value for customs risk assessment purposes. Such an overview can lead towards a knowledge base on the eCommerce platform data sources.
		[3] 6.2. Further research can examine the possibilities for setting up a data pipeline/trusted trade lane, in which customs can immediately approach the supply chain data in the real-time logistics process. Advantages for customs based on this set-up can then be further examined.

6 Clustering the recommendations into policy, standards and research recommendations

Working with Themes allowed us to logically group lessons learned and derive recommendations. For this deliverable we preserved this logic when drafting the recommendations as presented in Sections 3-5. However, for further steps based on these recommendations it is useful to provide an indication which of these recommendations relate to:

(1) Policy and organizational aspects that may be relevant to policy makers at DG TAXUD or at national level for Member State Administrations,
(2) Standardization aspects which may be relevant to standardization bodies and initiatives like UN/ CEFACT and EU CDM
(3) Further research aspects which can relevant for stakeholders that define further research directions for customs innovation research (e.g. DG HOME for security research).

Table 6 below clusters the recommendations in the categories (1) organizational and policy recommendations; (2) standardization recommendations, and (3) further research recommendations that stem from the technical data analytics pilots. Table 6 is intended to serve two purposes.

First of all, for a general audience, it can provide a quick overview of which recommendations are related to policy, standardization or research or a mix there-of. This division is only indicative, as in discussions with experts form the Customs administrations it became clear that many of the recommendations were multi-faceted.

Second, Table 6 can serve also as a support tool and input for the PROFILE management for drafting the PROFILE sustainability blueprint deliverable (D8.11) and targeted PROFILE recommendations to key PROFILE stakeholders such as DG TAXUD, EU Customs Administrations, standardization bodies or bodies responsible for drafting future security research such as DG HOME. For this purpose, tailor-made dedicated communication materials, based on these recommendations and other PROFILE results will be developed at the end of PROFILE for presenting the PROFILE results to the key stakeholder groups.

Table 6: Clustering the recommendations into policy, standardization and research recommendations

	Recommendations
Organizational and Policy	**Theme [1]: Organizational** [1] 1. Machine learning in the customs context [1] 1.1. Data Analytics is not a not the holy grail- acknowledging uncertainty and the human factor. Data Analytics does not work as a standalone - enforcement can never be done only by data analytics. You have also the physical world and you need to have an overall strong processes and business knowledge in order to leverage the potential of data analytics. [1] 1.2.You need human expertise to define thresholds. [1] 1.3. Conduct research in environment closely connected with the operational environment. [1] 1.4. Focus on flexible and dynamic data analytics development approaches; Due to the changing nature of the customs landscape, it is probably more efficient to work on either on developing simple models and algorithms that are easily retrained, or focusing on parameters that are stable over time. [1] 1.5. Collaborate among customs administrations to arrive at amount of data sufficient to develop reliable analytics. [1] 1.6. Using external data sources next to the customs declaration data may help to improve the development of analytics and risk management. In terms of Policy recommendations, one direction is to drive the way towards the closer linking and co-operation between Agencies (i.e. customs) and Freight Forwarders, carriers, eCommerce sellers and eMarkets, having a more transparent access to the transaction data, providing a right set of initiatives and the motivation for this to happen. [1] 1.7. Further simplification of the system of duties, prohibitions and restrictions can make it easier to develop data analytics. [1] 2. Data quality, Data preparation and the Effort for Data Preparation [1] 2.1.: You need first to prepare data (e.g. select your risk focus (e.g. fiscal fraud versus narcotics) and variables) and only afterwards take more data into account. [1] 2.2. Sufficient efforts for data preparation and understanding should be taken into account for successful data analytics development. [1] 2.3. There is a need for clear legislation guidelines concerning the application of GDPR and AI Act in the customs risk management environment in such cases. [1] 2.4. There is a need for data access and use protocols within the government services to easily share data with external data analytics development partners. [1] 2.5. There is a need for a trusted environment with trusted protocols for sharing data in a secure and simple manner, internally across customs departments, across customs administrations, and between customs and external data analytics providers. [1] 2.6. For improving data quality explore the possibilities for customs providing services to businesses. • To improve the declaration data quality (e.g. using HS code predictor to suggest inaccuracies in filled-in HS codes), and • in the long run possibilities for prefilled declarations (e.g. using also business data). [1] 3.. Building-up customs business expertise

	[1] 3.1. When you start with a data analytics innovation do careful balancing between scope and objectives, and be explicit about the piloting assumptions. [1] 3.2. The best way to build up customs business expertise efficiently is that data analysts collaborate closely with customs risk domain experts. [1] 3.3. You need people that can act as mediators/translators/liaisons internally between the technical and business people, as well as externally between customs and the external data analytics providers. [1] 4. Long-term view on innovation [1] 4.1. It would be beneficial if funding programs allocate funds for multi-phase innovation tracks, in order to push TRL further towards implementation. [1] 5. Upscaling to the operational environment [1] 5.1. It is recommended to consider three broad areas when thinking about upscaling of data analytics innovation for customs risk management: • How to organize for data analytics innovation: Consider transforming the existing organization(with separate IT, risk management, data analytics and detection technology departments) towards a new data-driven organization where these departments are better aligned and work together. • How to bring data analytics innovation to the real-time environment: Consider creating data sharing infrastructures that enable using different data sources to generate new insights needed for customs risk management real-time. • How to bring data analytics innovation to the operational customs officers: Consider issues like explainability, training and human resource development.. [1] 6. Collaborative Data Analytics Development between Customs Administrations [1] 6.1. It is valuable to meet physically and work together to a greater extent. [1] 6.2. Identify and address both technical and legal issues related to data sharing early in the project, so that during the project itself you can focus more directly on innovations and technical solutions. [1] 6.3. Find a common model for data sharing (e.g. fields, anonymization, restrictions) when it comes to exchanging or comparing declaration data from more than one country. **Theme [2]. Linked data, semantic technologies, and standards** [2] 1. It is recommended that customs declaration data is enriched with external data for improving customs risk management and that there is a closer collaboration between customs and external data providers. [2] 2.It is recommended that the use of a semantic model (e.g. the Digital Transport and Logistics Forum (DTLF) semantic model developed by the FEDeRATED project) can enable better linking of customs data and external data. Still the semantic model is under development and in a pilot phase. For applying the semantic model it is necessary to: • create data pipelines based on a 'node' concept sharing linked (event) data; existing platforms will act as 'node' in line with the Data Governance Act, • complement the node with a semantic adapter for integration with existing solutions , • install governance in a EU context, and • develop a roadmap for adoption of semantic architecture by customs administrations.
	Recommendations

Standardization	**Theme[2] Linked data, semantic technologies, and standards** [2]-4 It is recommended to extend EU Customs Data Model (EU CDM) so that it can integrate/ adopt the semantics of supply chain logistics data eventually aligning with DTLF, having in mind a short, mid, and long-term strategy. This can be done: ▪ ·In the short run by integrating the concept 'time' related to logistics activities and applying these structures only in binary relations between a customer and logistics service provider. ▪ ·In the mid-run restructure the EU CDM into a structured set of ontologies and align these with the DTLF semantic model (ontology). ▪ ·Long term recommendation is to adopt one semantic model and semantic architecture, with the objective to phase out existing data sharing mechanisms and create a more flexible and extendible environment. [2]-5 It is recommended that the UN/CEFACT standard makes steps towards ontologies, namely: ▪ Medium term recommendation is to restructure the UN/CEFACT buy, ship, pay model into a structured set of ontologies and align these with the DTLF semantic models (ontologies). Combined with the similar recommendation for EU CDM, having these ontologies aligned with each other will enable data transformations. ▪ Long term recommendation is to adopt one semantic model and semantic architecture, with the objective to phase out existing data sharing mechanisms and create a more flexible and extendible environment . **Theme [3]. Data analytics pilots** [3]-1. HS code predictor [3] 1.7. Goods description, a free text field, would benefit from more structuring and further standardization.
	Recommendations
Further research	**Theme[2] Linked data, semantic technologies, and standards** [2]3. It is recommended to explore further the possibilities of semantic technologies that include graph based solutions for visual analytics, data preparation and data linking. **Theme [3]. Data analytics pilots** [3] 1. HS code predictor [3] 1.1. For sentence embeddings, consider using a model pre-trained on customs data that takes into account specific abbreviations and text truncations used in the goods description. In order to avoid training big models like USE from scratch (due to the process complexity, significant computational power and extremely large datasets), a solution might be to find a way to use transfer learning on models e.g. by studying the model training process and adding Customs data as an extra layer to the model . [3] 1.2. Consider using language detectors to separate goods descriptions by language (e.g. NLP models like spaCy , Language Identification models like CLD3 etc.). [3] 1.3. Consider combination of methods using also numerical and categorical data in customs declaration e.g. price, weight, value, loading location (place + country), consignee, consignor etc.

	[3] 1.4. Modify the approach to classification, e.g. it could be based on sentence embedding plus additional binary classifiers and dictionaries associated a with every HS code group to detect the commodity nomenclature. [3] 1.5. Focus on HS code chapter level since some chapters can be more explanatory details than others. [3] 1.6. Study the classification methods (principles) and databases used by commodity certification/classification authority [3] 1.8. Consider HS code predictors that have been developed by commercial parties, taking into account that they may be biased to have (optimal) results for businesses. [3] 1.9. Safeguards on material composition, quality and assurance are essential for customs and also in the context of circular economy. Further research can examine how the HS code prediction models developed in PROFILE (WP3) can be further extended to also better differentiate similar goods that have different material composition and fall under different HS codes and nomenclatures. [3] 2. Random Forest commodity code (HS code) classifier, NLP and 2D visualisation for outlier detection [3] 2.1 Random Forest and NLP based methods were used to predict commodity codes. The methods have potential for further development. To increase accuracy one should consider combining the methods with each other. [3] 2.2 The NLP and Random Forest classification methods could theoretically be incorporated directly in the automatic process so that declarations can be analysed in real time. This has not been tested in this project though and can be a subject for further research. [3] 2.3 The classifiers were set up identically, but had higher performance for the Swedish data in all cases. This cross-border ”meta” result suggests that since the characteristic patterns of the Norwegian data are harder to learn, the quality of the Norwegian data could possibly be enhanced, perhaps through improving instructions to declarants. [3] 3. Outliers and anomalies detection with declaration data by using Autoencoder [3] 3.1. Apply the autoencoder on (a combination of) different datasets. Improvement of the output of the autoencoder requires interaction with targeting officers. There are various ways to include them, but the most obvious way is to provide targeting officers with the output of the autoencoder in their work processes and let them provide feedback. This approach is taken in the targeting dashboard that has been developed. [3] 3.2. Apply semi-supervised learning (update and optimize the models using a feedback system) [3] 3.3. Apply multivariate analysis of the anomalies. [3] 4 Automatic selection of customs declarations for inspections [3] 4.1. Reducing the number of input features. There is a trade-off between retaining features that are valuable source of information and removing features that contribute most strongly to dataset shift. Future work may leverage explainability for getting better insight into the role of different features in generating predictions and for identifying the optimal set of features. [3] 4.2. Validation of the predictions generated for unlabelled data. Entity embeddings – meaningful representations of categorical features – are potentially powerful tool for extracting features that are truly representative of full dataset (not affected by the bias). The results of the experiments with unsupervised learning will require validation in the field, i.e., real feedback from the inspectors on the quality of the outputs of unsupervised learning. [3] 4.3. Investigating the semantic value of entity embeddings used ‘standalone’. Entity embeddings (learned representations of categorical features) form a semantic vector space, in which the similarity among entities is represented by the proximity of the vectors. The semantic representations of entities

	that can facilitate insights into mutual relationships between these entities. Future work should explore vector space formed by distributed representations of most important features (e.g., operators or commodity codes) to validate that they indeed reflect real-life dependencies among entities. [3] 5. Visual analytics [3] 5.1 Explore further the applicability of visual analytics to support a targeting officer. A first effort is made in the targeting dashboard (WP2/WP4) towards visual analytics. This could visualize anomalies in trade flows that could be of interest to targeting officers. [3] 5.2 Visualize differences in data sets. The objective is to compare different data sets on particular features and visualize these differences. Examples of differences are on values of 'weight', 'time/duration' and 'route/itinerary/locations'. The actual relevant differences have to be defined by targeting officers. [3] 6 Other research related to eCommerce platforms, data pipelines [3].6. 1 Follow-up research can conduct a comparative study by examining different eCommerce platforms and providing an overview of which data can be of value for customs risk assessment purposes. Such an overview can lead towards a knowledge base on the eCommerce platform data sources.. [3] 6.2. Further research can examine the possibilities for setting up a data pipeline/trusted trade lane, in which customs can immediately approach the supply chain data in the real-time logistics process. Advantages for customs based on this set-up can then be further examined.

7 Conclusions

This deliverable reports on the lessons learned and recommendations from the PROFILE Project. It will serve as an input for the deliverable on the PROFILE sustainability blueprint D8.11. It will also serve as a basis for deriving targeted recommendations towards PROFILE key stakeholder groups which will be disseminated in dissemination activities and events of the PROFILE project.

8 References

[1] De Neve, J. E., Imbert, C., Spinnewijn, J., Tsankova, T., & Luts, M. (2021). How to improve tax compliance? Evidence from population-wide experiments in Belgium. Journal of Political Economy, 129(5), 1425-1463.

[2] Oosterman, D. T., Langenkamp, W. H., & Bergen, E. L. V. (2021, September). Customs Risk Assessment Based on Unsupervised Anomaly Detection Using Autoencoders. In Proceedings of SAI Intelligent Systems Conference (pp. 668-681). Springer, Cham.

[3] Deliverable D4.2 General overview on upgraded risk indicators and profiles (Belgian Living Lab). PROFILE Project (confidential deliverable).

9 Annex 1. List of stakeholders

This Annex contains the long list of Stakeholders that was identified at the start of the RECS activities. The results of the PROFILE project may be of potential interest to these stakeholders.

Network overview

International - customs

- WCO
  - WCO Data Model;
  - WCO Working Group on Data Analytics

EU

- DG TAXUD
  - Import Control System (ICS2), Safety and Security Analytics (SSA), Joint Analytics Capabilities (JAC)
  - EU Customs Data Model (CDM),
  - DMPG Expert group on DA ;
- DG HOME
- DG MOVE
  - eFreight via the EC expert consultation group Digital Transport and Logistics Forum (DTLF)
  - various regulations like:
    - eFTI (electronic Freight Transport Information) –
    - eMSW (electronic Maritime Single Window) Regulations
- DG RESEARCH
  - REA and the new Horizon Europe programme
- Agencies
  - JRC
- DG CONNECT

  - The European Blockchain Services Infrastructure (EBSI);
  - DLT4EU (open calls) and various other relevant developments like
  - European Data Strategy;
  - European Data Spaces;
  - GAIA-X
- Innovation and Networks Executive Agency (INEA)
- Connecting Europe Facility (CEF).
- PEN-CP

Businesses

- Traders/ Supply chain partners
- Associations
  - DCSA (e.g. eB/L, event data sharing),
  - IMO
  - IATA (e.g. ONE Record),
  - IRU
    - eCMR and associated platforms,
  - ESC
  - UPU
- Trade platforms
  - Tradelens,
  - BigDataMari,
  - Port Community Systems (PCSs)
- Payment service providers
- Data providers, either open –, paid, or private data. Access to private data requires a Regulation. These data providers might apply data analytics functionality.
- Technology providers
- Data analytics providers
- Data sharing communities

- International Data Space Association (IDSA) and its different hubs in Europe.

Standardization organizations (Business)

- UN/CEFACT (buy-ship-pay model)
- GS1 (various data sharing standards and identifications)
- CEN CENELEC
- ISO
  - Blockchain interoperability