Monica Scannapieco, Istat

Input Privacy: Towards a Logical Framework for Defining Official Statistics Scenarios

Monica Scannapieco[1], Fabrizio De Fausti[1], Massimo De Cubellis[1], Matjaz Jug[2],
Saeid Molladavoudi[3], Dennis Ramondt[2]

[1] Istat - Italian National Institute of Statistics
[2] Statistics Netherlands
[3] Statistics Canada

Abstract:
Input privacy takes into account privacy requirements in the data access stage of an Official Statistics production pipeline. Input privacy scenarios are quite complex : it is not easy to describe them by highlighting all the aspects that are necessary to both the technical solutions and a full specification of the privacy guarantees.
In this paper, we present the current effort towards the proposal of a logical framework for defining input privacy scenarios for Official Statistics, as a result of the UNECE project "Input Privacy-Preserving techniques".

Keywords:
Input privacy; data access; privacy-preserving techniques

1. Introduction

Modern statistical organizations are more and more investing on becoming part of a data ecosystem where they acquire and integrate data from multiple sources and provide richer statistical products. Such sources can be rather consolidated like administrative sources or more recent like Big Data sources.

In this scenario, the issue of privacy preservation is particularly relevant: the more sources are acquired and integrated, the higher are the privacy risks and the stronger are the protection mechanisms required to guard against those risks. From a legislative perspective there is a clear obligation to take privacy into account throughout the whole data treatment process, including at the access stage, through the "privacy by design" concept[1].

National Statistical Organizations (NSOs) are used to apply techniques for enforcing privacy by design on the *output side*, i.e., when publishing aggregated statistical data for dissemination purposes and when sharing microdata for research purposes. Statistical disclosure control (SDC) techniques as well as centers for remote microdata access are standard capabilities in statistical systems.

[1] See e.g. GDPR (General Data Protection Regulation), EU regulation 2016/679, where Article 25 is about "Data protection by design and by default".

However, NSOs have still to learn and invest in dealing with privacy protection on the *input side*, in a complementary but distinct way with respect to output privacy preservation [1].

Refer to [1] and [2] for a more detailed definition of input and output privacy concepts.

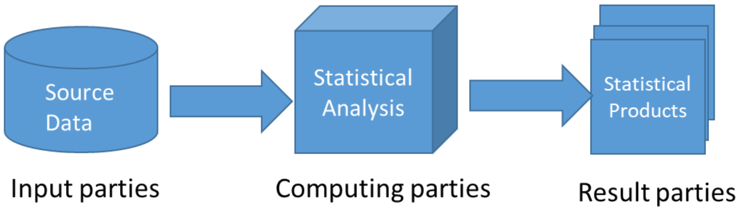


Figure 1 Abstract model for input/out privacy [2]

Looking at Figure 1, there are three roles, namely *Input parties*, *Computing parties*, *Result parties*: Input parties provide *Source data* that are processed by Computing parties through a *Statistical analysis* that is made available to Result parties as *Statistical products.* In this setting: ***"Input privacy** means that the Computing Party cannot access or derive any input value provided by Input Parties, nor access intermediate values or statistical results during processing of the data (unless the value has been specifically selected for disclosure)".*

In 2020, the UNECE High-level Group on Modernization of Official Statistics (HLG-MOS)[2] launched a project on Input privacy named "Input Privacy-Preserving techniques- IPP"[3], which is currently on-going and is scheduled until the end of 2021. The goal of this project is to investigate modern and innovative privacy-preserving techniques and methods that offer protection on the input side. The project currently involves Statistics Netherlands (acting also as project manager), Statistics Italy, Statistics Canada, ONS UK, INEGI Mexico, GSO Vietnam, Eurostat. The work of the project is coordinated with the UN Global Working Group Task Team on Privacy Preserving Techniques and activities of Eurostat planned in the same area.

Starting from this abstract model shown in Figure 1 and proposed in [2], in this paper we describe a first result of the IPP project, consisting of a logical framework aimed to define in a detailed way the scenarios of input privacy that are relevant for Official Statistics use.

In more detail, in Section 2, we introduce the framework, in Section 3 we instantiate it to the specific scenario of "Private Set Intersection with Analytics" and in Section 4 we draw some conclusions and next steps.

2. Defining Input Privacy Scenarios in Official Statistics: A logical framework

A relevant work when setting up an IPP project is the detailed description of the use cases. Indeed, most of the solutions that can be proposed are highly dependent on those, requiring

[2] https://statswiki.unece.org/pages/viewpage.action?pageId=187891840

[3] https://statswiki.unece.org/download/attachments/293536151/HLG-MOS%202021%20project%20proposal_Input%20Privacy%20Preservation.pdf?version=1&modificationDate=1607525791997&api=v2

a level of specification that can be by far more detailed than the one used for typical Official Statistics use cases.

To address this issue, it can be useful to have a logical framework, i.e. a template, that *guides* the specification of such use cases, and the following proposal goes exactly in this direction. The framework that we propose for defining input privacy scenarios in OS, in the following IPP-Template for OS includes:

- General features, which describe the general setting of the scenario.
- Solution-specific features, which detail the requirements and the solution, and that starting from Figure 1, are organized along three lines, namely Input parties, Computing Parties and Result Parties

The specific components of the template are detailed in Figure 2.

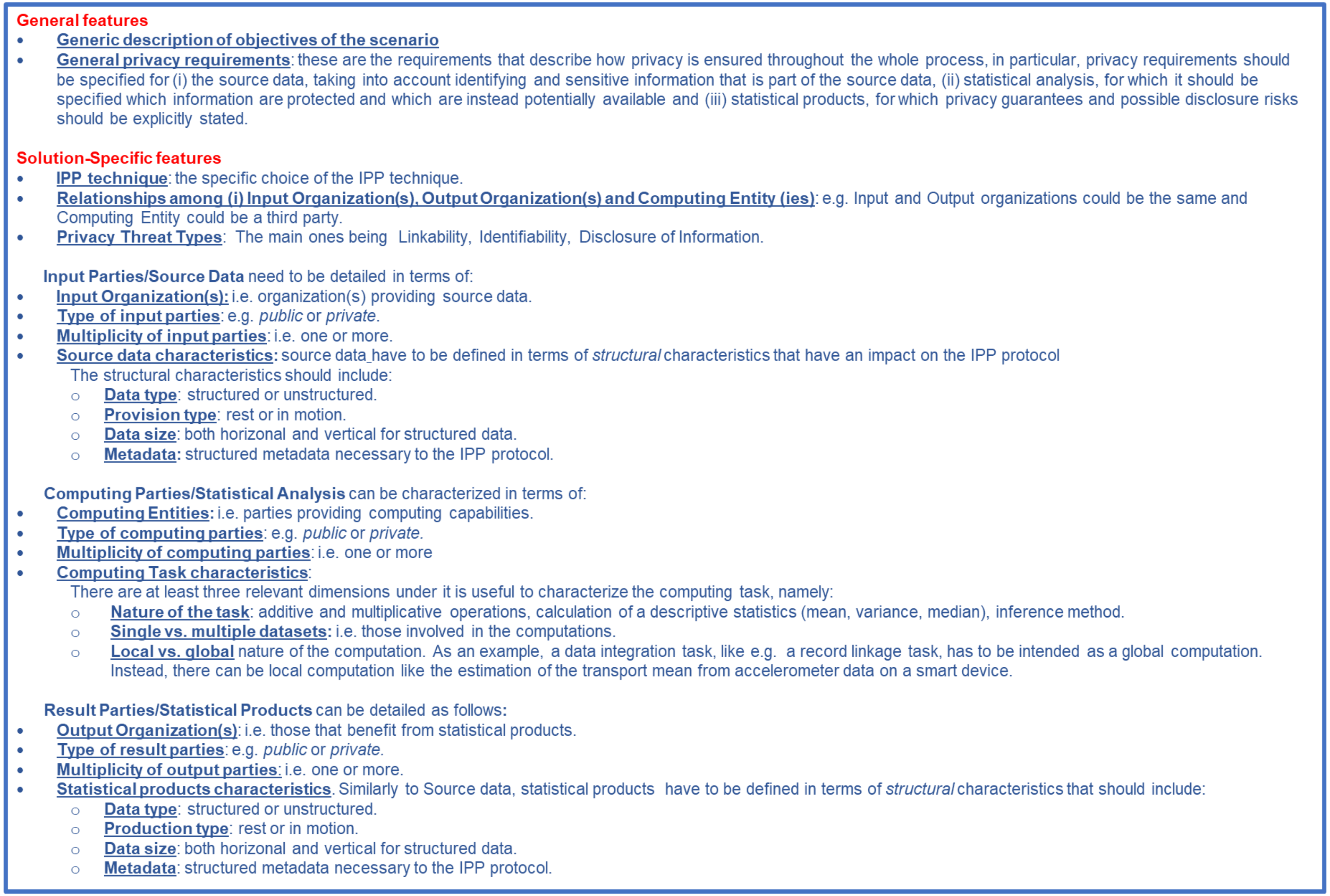
3. Instantiating the IPP Template for OS: Private Set Intersection with Analytics

When focusing on IPP scenarios that involve multiple organizations, a possible characterization is provided by [3], where four scenarios are envisioned to support the "generic" information sharing need, namely:

- Private set intersection (PSI): Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an exact join to A and B without revealing any unnecessary information about their individual databases. That is, ideally, the only information learned by P1 about B is A∩B and vice versa
- Private set intersection with enrichment (PSI-E): Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an exact or approximate join to A and B without revealing any unnecessary information about their individual databases. After that, they wish to enrich joined records with variables by both parties. At the end of the process P1 will learn additional P2 variables on A∩B and vice versa
- Private set intersection with analytics (PSI-A); Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an analytics function to the intersection of A and B in a private way. At the end of the process, the only information learned by the parties (beyond the keys of the records belonging to the intersection) is the result of the analytics function.
- Private Set Union with Analytics (PSU-A), which is a common private data mining scenario: Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an analytics function to the union of A and B without revealing any unnecessary information about their individual databases. At the end of the process, the only information learned by the parties) is the result of the analytics function.

Among these scenarios, we focus on Private set intersection with analytics PSI-A to instantiate the template. The result is shown in Figure 3.

The template synthetically describes an IPP scenario between Istat and Bank of Italy involving PSI-A, by highlighting input privacy requirements and main features of the implemented solution.



General features

- **Generic description of objectives of the scenario**
- **General privacy requirements**: these are the requirements that describe how privacy is ensured throughout the whole process, in particular, privacy requirements should be specified for (i) the source data, taking into account identifying and sensitive information that is part of the source data, (ii) statistical analysis, for which it should be specified which information are protected and which are instead potentially available and (iii) statistical products, for which privacy guarantees and possible disclosure risks should be explicitly stated.

Solution-Specific features

- **IPP technique**: the specific choice of the IPP technique.
- **Relationships among (i) Input Organization(s), Output Organization(s) and Computing Entity (ies)**: e.g. Input and Output organizations could be the same and Computing Entity could be a third party.
- **Privacy Threat Types**: The main ones being Linkability, Identifiability, Disclosure of Information.

Input Parties/Source Data need to be detailed in terms of:

- **Input Organization(s):** i.e. organization(s) providing source data.
- **Type of input parties**: e.g. *public* or *private*.
- **Multiplicity of input parties**: i.e. one or more.
- **Source data characteristics:** source data_have to be defined in terms of *structural* characteristics that have an impact on the IPP protocol

  The structural characteristics should include:
  - **Data type**: structured or unstructured.
  - **Provision type**: rest or in motion.
  - **Data size**: both horizonal and vertical for structured data.
  - **Metadata:** structured metadata necessary to the IPP protocol.

Computing Parties/Statistical Analysis can be characterized in terms of:

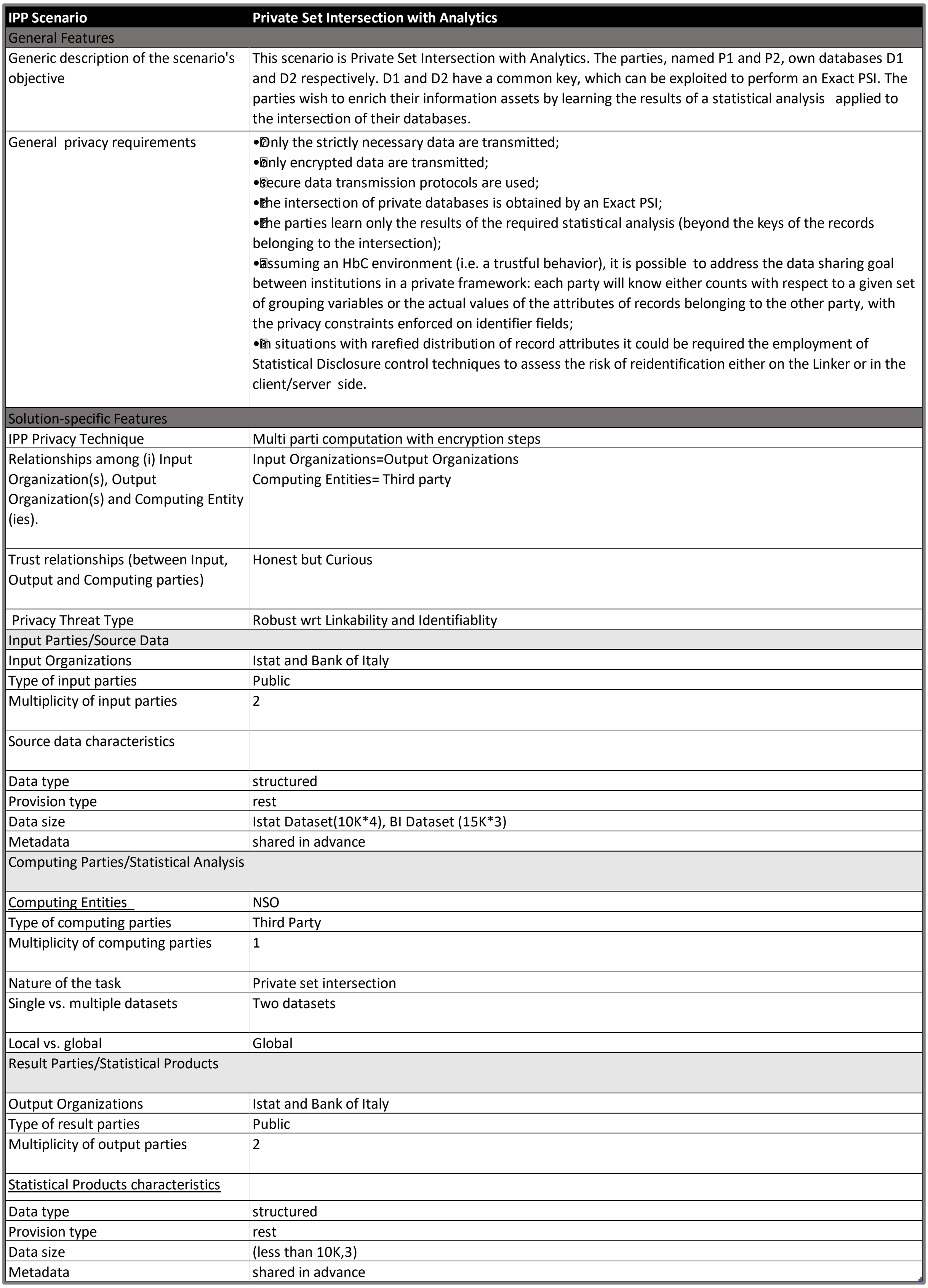
- **Computing Entities:** i.e. parties providing computing capabilities.
- **Type of computing parties**: e.g. *public* or *private*.
- **Multiplicity of computing parties**: i.e. one or more
- **Computing Task characteristics**:

  There are at least three relevant dimensions under it is useful to characterize the computing task, namely:
  - **Nature of the task**: additive and multiplicative operations, calculation of a descriptive statistics (mean, variance, median), inference method.
  - **Single vs. multiple datasets:** i.e. those involved in the computations.
  - **Local vs. global** nature of the computation. As an example, a data integration task, like e.g. a record linkage task, has to be intended as a global computation. Instead, there can be local computation like the estimation of the transport mean from accelerometer data on a smart device.

Result Parties/Statistical Products can be detailed as follows**:**

- **Output Organization(s)**: i.e. those that benefit from statistical products.
- **Type of result parties**: e.g. *public* or *private*.
- **Multiplicity of output parties:** i.e. one or more.
- **Statistical products characteristics**. Similarly to Source data, statistical products have to be defined in terms of *structural* characteristics that should include:
  - **Data type**: structured or unstructured.
  - **Production type**: rest or in motion.
  - **Data size**: both horizonal and vertical for structured data.
  - **Metadata**: structured metadata necessary to the IPP protocol.

Figure 2: IPP Template for OS



IPP Scenario	Private Set Intersection with Analytics
General Features	
Generic description of the scenario's objective	This scenario is Private Set Intersection with Analytics. The parties, named P1 and P2, own databases D1 and D2 respectively. D1 and D2 have a common key, which can be exploited to perform an Exact PSI. The parties wish to enrich their information assets by learning the results of a statistical analysis applied to the intersection of their databases.
General privacy requirements	•Only the strictly necessary data are transmitted; •only encrypted data are transmitted; •secure data transmission protocols are used; •the intersection of private databases is obtained by an Exact PSI; •the parties learn only the results of the required statistical analysis (beyond the keys of the records belonging to the intersection); •assuming an HbC environment (i.e. a trustful behavior), it is possible to address the data sharing goal between institutions in a private framework: each party will know either counts with respect to a given set of grouping variables or the actual values of the attributes of records belonging to the other party, with the privacy constraints enforced on identifier fields; •in situations with rarefied distribution of record attributes it could be required the employment of Statistical Disclosure control techniques to assess the risk of reidentification either on the Linker or in the client/server side.
Solution-specific Features	
IPP Privacy Technique	Multi parti computation with encryption steps
Relationships among (i) Input Organization(s), Output Organization(s) and Computing Entity (ies).	Input Organizations=Output Organizations Computing Entities= Third party
Trust relationships (between Input, Output and Computing parties)	Honest but Curious
Privacy Threat Type	Robust wrt Linkability and Identifiablity
Input Parties/Source Data	
Input Organizations	Istat and Bank of Italy
Type of input parties	Public
Multiplicity of input parties	2
Source data characteristics	
Data type	structured
Provision type	rest
Data size	Istat Dataset(10K*4), BI Dataset (15K*3)
Metadata	shared in advance
Computing Parties/Statistical Analysis	
Computing Entities	NSO
Type of computing parties	Third Party
Multiplicity of computing parties	1
Nature of the task	Private set intersection
Single vs. multiple datasets	Two datasets
Local vs. global	Global
Result Parties/Statistical Products	
Output Organizations	Istat and Bank of Italy
Type of result parties	Public
Multiplicity of output parties	2
Statistical Products characteristics	
Data type	structured
Provision type	rest
Data size	(less than 10K,3)
Metadata	shared in advance

Figure 3 Instantiation of the IPP Template for OS to Private Set Intersection with Analytics Scenario

4. Discussion and Conclusion

In this paper we illustrate the current state of a proposal of a generic template for describing input privacy scenarios for Official Statistics, as one of the result of the UNECE IPP project, which is a project under the umbrella of the UNECE HLG-MOS. We showed an application of it to a private data integration scenario, namely private set intersection with analytics, also introducing a further classification aimed to shed some light on the complexity of privacy preserving data sharing scenarios.

The UNECE IPP project used the template also to describe further scenarios, including private machine learning and privacy preserving smart surveys. These contributions are the first step towards an ongoing effort of the project, namely producing clear definition and technical elements to support the introduction of IPP techniques in Official Statistics.

Acknowledgement:

We would like to thank Fabio Ricciato (Eurostat) and the full IPP project team that participated to the discussion. We also would like to thank the UNECE team, Taeke Gjaltema and Yan Qin, for their support to the project.

References:

1. F. Ricciato, A. Bujnowska, A. Wirthmann, M. Hahn, E. Barredo-Capelot, A reflection on privacy and data confidentiality in Official Statistics, ISI 2019.
2. UN Handbook on Privacy-Preserving Computation Techniques, http://publications.officialstatistics.org/handbooks/privacy-preserving-techniques-handbook/UN%20Handbook%20for%20Privacy-Preserving%20Techniques.pdf
3. P.D. Falorsi, B. Liseo, M. Scannapieco: Dealing with Privacy Issues in Data Integration Systems, extended version of proceedings 'Law via the Internet 2018', Knowledge of the Law in the Big Data Era, Series Frontiers in Artificial Intelligence and Applications, ISBN 978-1-61499-984-3, IOS Press, 2019