Detailed planning IT-report (Version 23 January 2018)

Id	Product	Deadline	Who*	Comments
8.3	**Report describing the IT-infrastructure used and the accompanying processes developed and skills needed to study or produce Big Data based official statistics.**	**31-01-2018**		
1.	Metadata management (ontology) It is important to have (high quality) metadata available for big data. This is essential for nearly all uses of Big Data. Ideally, an ontology is available in which the entities, the relations between entities and any domain rules are laid down.	10-01-2018	*Piet Daas (NL)* Jacek Maślankowski (PL) *Monica Scannapieco (IT)*	First draft is available
2.	Big Data processing life cycle Continuous improvement of Big Data processing requires capturing the entire process in a workflow, monitoring and improving it. This introduces the need to design and adapt the process and determine its dependence on external conditions.	10-01-2018	Piet Daas (NL) Jacek Maślankowski (PL) *Monica Scannapieco (IT)*	First draft is available
3.	Format of Big Data processing Processing large amounts of data in a reliable and efficient way introduces the need for a unified framework of languages and libraries.	20-12-2017	Piet Daas (NL) Jacek Maślankowski (PL)	First draft is available
4.	Datahub Sharing of multiple data sources is greatly facilitated when a single point of access, a so-called hub, is set up via which these sources are made available to others.	20-12-2017	Piet Daas (NL) Jacek Maślankowski (PL)	First draft is available
5.	Data source integration There is a need for an environment on which data sources, including Big Data, can be easily, accurately and rapidly integrated.	20-12-2017	Piet Daas (NL) *Jacek Maślankowski (PL)* Sónia Quaresma (PT)	First draft is available
6.	Choosing the right infrastructure A number of Big Data oriented infrastructures are available. Choosing the right one for the job at hand is key to assuring optimal use is made of the resources and time available.	20-12-2017	Piet Daas (NL) *Jacek Maślankowski (PL)* Sónia Quaresma (PT)	First draft is available
7.	List of secure and tested API's An application programming interface (API) is a set of subroutine definitions, protocols,	10-11-2017	Jacek Maślankowski (PL)	First draft is available

	and tools for building application software. It is important to know which API´s are available for Big Data and which of them are secure, tested and allowed to be used.			
8.	Shared libraries and documented standards Sharing code, libraries and documentation stimulates the exchange of knowledge and experience between partners. Setting up a GitHub repository (or something similar) would enable this.	24-11-2017	Jacek Maślankowski (PL)	First draft is available
9.	Data-lakes Combining Big Data with other, more traditional, data sources is beneficial for statistics production. Making all data available at a single location, a so-called data-lake, is a way to enable this.	20-12-2017	Piet Daas (NL) *Jacek Maślankowski (PL)*	First draft is available
10.	Training/skills/knowledge Facilitating the exchange of skills and knowledge, for instance by offering trainings or assistance, will greatly stimulate the use of Big Data by many National Statistical Offices.	8-12-2017	Piet Daas (NL) *Jacek Maślankowski (PL)*	First draft is available
11.	Speed of algorithms Making use of fast and stable implementations of algorithms will greatly stimulate the application of Big Data as this speeds up the processing of large amounts of data tremendously.	13-12-2017	Piet Daas (NL) *Jacek Maślankowski (PL)*	First draft is available

* alphabetical order and underlined main contributor, regular contributor and *italic "little" contributor*

Detailed planning quality report

Id	Product	Deadline	Who*	Comments
8.2	**Report describing the quality aspects identified in studies focussing on the use of Big Data for official statistics**	**02-03-2018**		
1.	Coverage Information on the population included in a big data source is vital for reliable statistics. Important for this issue are the lack of information on the units included, their duplication and their selectivity.	02-03-2018	*Manca Golmajer and Crt Grahonja(SI), Jacek Maslankowski (PL)* Tiziana Tuoto (IT)	First draft available
2.	Comparability over time To produce comparable statistics over time, it is essential that the source remains accessible, relevant and its content remain usable.	02-03-2018	Valentin Chavdarov (BG)	First draft is available
3.	Processing errors During the processing of Big Data, errors may be introduced that negatively affect the quality of the data. Examples of this are the way outliers and missing values are treated.	02-03-2018	Manca Golmajer and Crt Grahonja(SI), Magdalena Six (AT)	First draft is available
4.	Process chain control In a Big Data process it is very likely that multiple partners are involved. To assure a stable and timely delivery of data of high quality, the entire process needs to be controlled.	02-03-2018	Piet Daas (NL)	First draft is available
5.	Linkability It is to be expected that Big Data needs to be linked or combined with to other data sources. During this process, errors may occur which affect the quality of the output.	02-03-2018	Manca Golmajer and Crt Grahonja(SI), *Jacek Maslankowski (PL)* Tiziana Tuoto (IT)	Draft of each chapter, consolidated within chapter-group: 26th of January, Consolidated draft: 14th of February Final: 2nd of March
6.	Measurement error The values included in Big Data may not be all correctly measured; some may contain errors. This affects the outcomes produced, certainly when a systematic bias is introduced.	02-03-2018	Valentin Chavdarov (BG) Manca Golmajer and Crt Grahonja(SI),	First draft is available
7.	Model errors and Precision Big data based estimates are likely produced by models. The specifications of these models may be incorrect which	02-03-2018	*Manca Golmajer and Crt Grahonja(SI),* Magdalena Six (AT)	Draft of each chapter, consolidated within chapter-group: 26th of January, Consolidated draft:

	negatively affects the reliability of the estimates.		Tiziana Tuoto (IT)	14th of February Final: 2nd of March

* alphabetical order and <u>underlined main contributor,</u> regular contributor and *italic "little" contributor*

Detailed planning methodology report

Id	Product	Deadline	Who*	Comments
8.4	**Report describing the methodology (principles of finding and collecting Big data and assuring stable access, methodology of using Big Data as a single or major source of input, methodology of using Big Data as an additional data source in combination with others) of using Big data for official statistics and the most important questions for future studies**	**31-3-2018**		
1.	Assessing accuracy How accurate are Big Data based the estimates. Both bias and variance need to be considered, but bias is expected to be more important.	2-3-2018	Tiziana Tuoto (IT) Manca Golmajer (SI)	**Deadlines:** - **24th of Feb 2018** – Draft Report - **2nd of March 2018** - Final report - **26th of Jan 2018** - Structure/template
2.	What should our final product look like? Data driven statistics do not start with a predefined end product in mind. However, during this work it is important to start thinking about the product that can/will be delivered.	2-3-2018	Valentin Chavdarov (BG)	**Deadlines:** - **24th of Feb 2018** – Draft Report - **2nd of March 2018** - Final report - **26th of Jan 2018** - Structure/template
3.	Deal with spatial dimension Many Big data sources have a spatial component, such as a geolocation,. It is essential to make use of this kind of information. This means that attention has to be paid to the location of objects and aggregating spatial data.	2-3-2018	Piet Daas (NL)	**Deadlines:** - **24th of Feb 2018** – Draft Report - **2nd of March 2018** - Final report - **26th of Jan 2018** - Structure/template
4.	Changes in data sources Many Big Data sources are by-products of new technological developments. The content of these sources may therefore change rapidly. This will also be the case for data sources produced	2-3-2018	Valentin Chavdarov (BG) Piet Daas (NL)	First draft available

	by private companies. It is important to get a grip on these to enable the production of reliable time series.			
5.	<u>Machine learning in official statistics</u> Considering its rise in popularity, it is likely that in the future more and more (official) statistics will make use of machine learning based methods. It is vital to fully understand the implications of applying these kinds of methods in the production of official statistics. Estimation of variance, extracting features and knowing how to create good data sets for training purposes are examples of important considerations.	2-3-2018	Tiziana Tuoto (IT) Manca Golmajer (SI)	
6.	<u>Data linkage</u> To fully integrate Big Data in official statistics production it is essential that these data sources and/or their (intermediary) products can be combined with the data provided by other (more traditional) sources. Combining refers to the inclusion at either the unit or domain level here.	2-3-2018	Piet Daas (NL) Tiziana Tuoto (IT)	**<u>Deadlines:</u>** - **24th of Feb 2018** – Draft Report - **2nd of March 2018** - Final report - **26th of Jan 2018** - Structure/template
7.	<u>Secure multi-party computation</u> A lot of interesting data are produced by other organisations, such as private companies. Combining data from different organizations is challenging as many of them may not want the others to have complete access to their data; i.e. access at the individual record level. Being able to combine sources with a method that keeps the individual inputs of each partner private for the other parties involved is key to unlock the full potential of Big Data. National Statistical Offices are organizations that are ideally suited to function as a trusted party for this.	2-3-2018	Piet Daas (NL)	**<u>Deadlines:</u>** - **24th of Feb 2018** – Draft Report - **2nd of March 2018** - Final report - **26th of Jan 2018** - Structure/template
8.	<u>Inference</u> Methods that are able to reliably infer from Big Data and/or the combination of such data and other sources need to be developed. It is also essential to understand how this kind of inference is affected by the various sources of error that may occur.	2-3-2018	Piet Daas (NL) Tiziana Tuoto (IT)	**<u>Deadlines:</u>** - **24th of Feb 2018** – Draft Report - **2nd of March 2018** - Final report

				- **26th of Jan 2018** - Structure/template
9.	<u>Sampling</u> Not all data are available or can be used. The ability to deal with subsets of Big Data and draw valid conclusions from it is important in unleashing its full potential.	2-3-2018	Valentin Chavdarov (BG) Tiziana Tuoto (IT)	First draft is available
10.	<u>Data process architecture</u> Big data are often produced by one of the partners in the chain and subsequently transferred and/or used by others. For statistics production it is essential to have an overview of the whole chain and the individual steps performed by each partner as each step affects the other.	2-3-2018	Piet Daas (NL)	**<u>Deadlines:</u>** - **24th of Feb 2018** – Draft Report - **2nd of March 2018** - Final report - **26th of Jan 2018** - Structure/template
11.	<u>Unit identification problem</u> Big Data are produced by units of which (often) hardly any information is available in the source. This makes it challenging to identify the 'real world' units producing the data. For statistics it is essential to relate the units in Big Data with that of the (statistical) target population.	2-3-2018	Piet Daas (NL) Tiziana Tuoto (IT)	**<u>Deadlines:</u>** - **24th of Feb 2018** – Draft Report - **2nd of March 2018** - Final report - **26th of Jan 2018** - Structure/template