United Nations
Global Working Group on Big Data for Official Statistics
Task Team on Cross-Cutting Issues

Deliverable 2: Revision and Further Development of the Classification of Big Data

Version 12 October 2015, for discussion at the 2015 Global Conference on Big Data for Official Statistics, Abu Dhabi, 20-22 October, 2015

1. Introduction

According to its Terms of Reference (TOR, [1]), the broad objectives of the Task Team on Cross-cutting issues, classifications, frameworks and taxonomy (TTCC) are “to examine cross-cutting issues related to the integration of Big Data into official statistics' production, such as classification, taxonomy, data methodologies and quality frameworks for collecting, analysing and disseminating statistics derived from Big Data. In this context, the TTCC should develop and share knowledge of methodologies, data analytics and visualisation tools as well as quality assurance frameworks for the use of Big Data in official statistics and to refine the classification of Big Data.”

The TOR defines four deliverables. The second one concerns the classification of Big Data. In 2013 UNECE already developed a classification of Big Data [2]. The deliverable is meant to look at possible extensions of that classification, such as adding dimensions related to characteristics of Big Data, for instance velocity, variety (structure), owners or collectors, data subject, or statistical methods. Such dimensions could be used to create subsets of Big Data according to the needs of users. The needs might be related to access, processing analysis or publication of statistics from Big Data sources. The resulting classification would be more flexible than the existing one, according to the TOR.

The need for this deliverable was underlined by the outcomes of the 2015 UNSD Global Survey on Big Data for Official Statistics. The survey included the question: “On which topics do you see an urgent need for statistical guidance for your office or national statistical system?”; one of the topics listed was “classification of Big Data”. The question was answered by 89 respondents. Of these, 73% indicated that guidance on the classification of Big Data had a “high” (37%) or “medium” (36%) urgency.

This document describes the conclusions reached by the TTCC. First, the approach taken is explained (chapter 2), followed by an identification of possible uses of the classification (chapter 3). For any classification, its subject and scope has to be clearly defined. For the Big Data classification, this means defining the notion of Big Data source. This is done in chapter 4, which is followed by the identification of possible classification criteria (chapter 5), derived from the intended uses of the classification. Conclusions on the need for extending the UNECE classification and further work to be done are given in the last chapter. For reference, the Annex to this document contains the UNECE classification.

It should be noted that the TTCC has *not* been asked to give a definition of the concept of Big Data. A definition for statistical purposes is already available from UNECE [3]:

> *Big Data are data sources that can be – generally – described as: “high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making”.*

2. The approach to the classification of Big Data

<u>the nature of classifications</u>

Classifications usually have a <u>subject</u>, a <u>scope</u> (or universe), and one or more levels of (sub)<u>classes</u> describing possible <u>characteristics</u> of the subjects, based on explicit or implicit <u>classification criteria</u>. Classifications are designed on the basis of their <u>intended uses</u>.

For instance, SICs (standard industrial classifications) are meant to be used for classifying businesses (the subject) of, as a minimum, the corporate sector of SNA (the scope). SICs consist of a number of levels of classes and subclasses, describing the economic activities (the characteristic) of businesses. The layers of the SIC are nested, each class covering the same range of economic activities as its combined subclasses. The economic activities are defined in terms of, for instance, the types of outputs, inputs and processes of the businesses (the classification criteria). Applied to a population of businesses, the set of businesses classified to the same class is called an industry. The demand for information on industries (the intended use) determines the design of the SIC.

An extensive literature exists on classifications. For instance, best practices have been described by Statistics New Zealand [4].

the approach of the UNECE classification of Big Data

The desired classification of Big Data follows a similar pattern. We may adopt the definition of Big Data given above, which treats Big Data as data sources. These are the subjects of the classification. Examples from the UNECE classifications are social networks such as Twitter, business data such as credit card information, or sensor data such as traffic sensors. The scope is all sources of Big Data that fulfil the definition of Big Data given above. The UNECE classification has four levels of classes, e.g. Internet of Things (one of three classes at the highest level), data from sensors (one of two classes at the second level), fixed sensors (one of two classes at the third level), and home automation (one of three classes at the lowest level). The classification criteria are not always explicit, but the highest level of the UNECE classification, for example, is apparently based on the type of generator of the data. The intended uses of the UNECE classification may not have been clear from the outset, but the classification has been used for several purposes (see chapter 3 below).

questions for the TTCC

When designing a Big Data classification, or extending the one from UNECE, several issues have to be solved. First of all, the intended uses have to be identified, otherwise the design cannot be started. This is done in chapter 3. In line with the TOR of the TTCC, we will restrict ourselves to uses of Big Data for official statistics. Then the subject and scope have to be chosen. This is linked to the notion of “Big Data”. Definitions of the concept of Big Data tend to be highly disputable. For practical reasons we will use the definition given in the introduction of this document. Nevertheless, even if it is clear what “Big Data” is in a general sense, its delineation as the subject of the classification may not be clear. For instance, if we follow the definition of Big Data and look at it as a Big Data *source,* it should be noted that one single organisation may provide the platform for entirely different types of Big Data, e.g. Google providing e-mail, search and other services. Chapter 4 deals with such issues.

A related question is whether any instance of Big Data sources must belong to one and only one class of the classification of Big Data. Again, there is a parallel with SICs. The subject of an SIC, the business, needs a definition. Defining the notion of the business has a long history, and for the ESS (the European System of Statistics) a number of different business concepts have been defined, for different statistical uses, which are regulated by EU law. The definition of the business has implications for the SIC, and vice versa. For instance, designing an SIC for enterprises would result in different classes than designing one for local units. Enterprises may have secondary activities (resulting in heterogeneity of industries), but so-called units of homogeneous production by definition do not have secondary activities (resulting in homogeneous industries). For the Big Data classification one question to answer is whether homogeneity must be a requirement.

It may be pointed out that the question of defining the subject of the classification and the question of choosing classification criteria are not independent from each other. Certain criteria, such as “country” may be used as a defining criterion of Big Data sources, making them homogeneous in respect of country by definition, or may be used as a classification criterion, resulting in different (sub)classes for different countries, although these (sub)classes may not be homogeneous.

The classification criteria have to be based on the intended uses of the classification. We will look at the actual uses of the UNECE classification, other existing listings of Big Data sources and foreseen uses, such as in the context of SDG indicators. These result in potential classification criteria, from which the actual classification criteria will have to be selected. The potential criteria are discussed in chapter 5, but it is too early to make the selection of the actual criteria, as will be reasoned in the concluding chapter.

3. Uses of Big Data classifications

The UNECE classification has been used already for several purposes. At the same time a number of ad hoc classifications or lists of Big Data sources exist for various purposes. In addition, new potential uses of a Big Data classification can be distinguished.

uses of the UNECE classification

The UNECE classification was drafted at a time when NSIs were seeking guidance for using Big Data for official statistics. At that – still very recent – time there was no comprehensive overview of potential Big Data sources, it was not clear in what ways Big Data could be used for official statistics, and what would be the most important issues to be solved. There was no precise definition of Big Data, in the sense that it would give, for any specific data source, an answer to the question whether or not it was considered Big Data. In this situation the Big Data classification provided a very useful list of Big Data sources that could be used as an *extensive* definition of Big Data: it showed what was actually considered to be Big Data sources.

This also proved to be very useful for managerial and policy purposes. The classification was referred to in the 2014 UNECE Big Data project [5]. For instance, it was used as a list from which Big Data sources could be selected in the context of the so-called Sandbox of that project [6]. The 2014 Global Survey on Big Data for Official Statistics of UNSD in collaboration with UNECE also referred to the UNECE classification. Some NSIs have used the classification as well for managerial and policy purposes.

The UNECE classification was essentially a taxonomy reflecting easy-to-observe characteristics of data sources, such as the purpose of sensors or the platform of social media. It did not take into account, in its structure, what would be involved when using these data sources. As a consequence, the classification did not figure prominently in the deliverables of the 2014 UNECE Big Data project, which described the findings on methodology, privacy and partnerships [7].

uses of other Big Data lists

The UNECE classification was also used as a reference when the GWG started its work, but when the Global Survey was launched again in 2015 by UNSD, no reference was made to the UNECE classification. There was one question specifically on the use of Big Data sources, with a list of 13 broad Big Data source types (plus a category "other"), but this list was different from the UNECE classification, of which some categories were split, some were combined, and some had a different wording. There were also new categories, such as ships identification data.

Other lists of Big Data sources that are not clearly linked to the UNECE classification are used elsewhere. Many Big Data overview papers contain lists of Big Data sources. A recent example is a paper by Kitchin [8], which contains a table linking Big Data sources to data types and statistical domains, but there are many more cases of ad hoc classifications of Big Data. Companies that offer services related to Big Data use their own classifications, such as IBM [9].

A special case worth mentioning is the use of Big Data categories in the context of the ESS Task Force on Big Data. When preparing for a future ESSnet on Big Data, pilots were going to be selected. In order to obtain relevant information, a matrix was drafted with 18 types of Big Data sources on one axis, and a number of aspects considered important on the other [10, 11]. The aspects comprised considerations such as sustainability of the source, geographic dimension, privacy and costs. Task Force members were asked to indicate the importance of aspects for each Big Data source for their selection in the pilots. Since both axes were the result of consultations with Task Force members, the Big Data sources listed may be pertinent in their own right. The aspects considered could be relevant when defining potential classification criteria.

foreseen uses

One possibly important future use of the classification is as a reference in the discussion on the possible use of Big Data for compiling SDG indicators. These indicators still have to be formally adopted, and only a small minority of countries have started looking at the usability of Big Data for deriving indicators to measure progress on the SDGs, as was shown in the 2015 UNSD Global Survey on Big Data for Official Statistics. Therefore, it may be too early to know how it will be used, but it is clear that the usability of Big Data for such indicators will be a relevant factor, possibly with a further decomposition such as the SDG goals, targets or indicators that could be measured using each Big Data source. A parallel Task Team is exploring this (the Task Team on Using Big Data for the SDGs).

The 2015 Global Survey itself may also be used for identifying possible requirements for the classification of Big Data, since the questionnaire not only contained a list of Big Data sources itself, but data was also collected on their possible use for various statistical domains, among other things. In addition, the results of that survey will be used for an on-line repository of Big Data projects, and the classification of Big Data sources may be used for structuring this repository by looking at the Big Data sources referred to in the projects.

The possible uses described above are leading when defining the subject of the envisaged Big Data classification and its classification criteria.

4. The subject and scope of the classification

scope

From the definition of UNECE [3], reproduced in the introduction, it follows that the Big Data classification is about *data sources* of a certain type. Whether a data source is of that type (i.e., is one of the set of data sources that can be – generally – described as: "high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making") is not relevant to the question whether or not it is a *data source*, but it does determine whether or not the data source is within the scope of the classification.

Thus the scope is implicit in the definition of Big Data, which we assumed already chosen. In practice this may not resolve all scope issues encountered. However, as is often done when designing classifications, flexibility is ensured by listing all classes that are reasonably well established and adding a category "other". In principle, this can be done at every level of the classification. Given the uses described in chapter 3, further specifying the requirements for a data source to be considered Big Data does not deserve a high priority. This issue will not be further discussed in this document.

subject

More important is illuminating the concept of data source (in the case of Big Data). A couple of questions come to mind. First, if certain Big Data is not used (yet) for official statistics, do we still consider it a data source? The answer to this question is clearly affirmative, since the intended uses of the classification refer, in many cases, to situations where a data source is not actually used, at least not yet, but may be considered.

Second, how do we determine whether, in a specific situation, we are dealing with one or with several data sources? Given the uses described in chapter 3, it would not be helpful to consider, for instance, Google, as a single data source: it would not make sense *not* to distinguish between Google Earth, Google Trends, etc. Apparently the company or organisation to which the data is associated is not automatically the same as the data source. But if it is not the organisation, what then delineates the data source?

It stands to reason to take the type of data into account when delineating the data source, but it is not clear how this can best be done. There is a possibility that the delineation of the subject of the classification becomes dependent on the classification criteria, and vice versa. In order to avoid confusion, it may be best to delineate the subjects of the classification in three steps:

1. Start with the set of organisations that are associated with Big Data. (This would include, e.g., Google.)
2. Identify relevant classification criteria on the bases of envisaged uses of the classification.
3. Make a very limited selection of these classification criteria for which the subject of the classification must be homogeneous and, where applicable, split the organisations of point 1 accordingly. (This selection would presumably preclude, e.g., having e-mail and searching services comprised in the same classification subject).

A related question concerning the definition of the subject of the classification concerns the inclusion or exclusion of data that is associated to a data source. Consider the example of mobile phone data. The questionnaire of the 2015 UNSD Global Survey listed a number of types of data sources. For mobile phone data it made a distinction between CDR data (only outgoing), CDR data (all calls and messages), CDR data (including internet traffic), Location Area updates and Abis data. If recorded by the same company, would these be considered one data source or five data sources? Similarly, what data would be included in a social media data source, would subscription data be considered part of the data source? And what about data sources providing only aggregate data, such as Google Trends? Would Google Trends be a data source in its own right, or must the search database behind Google Trends be considered the subject – or a different subject – of the classification?

Given the possible and likely uses described in chapter 3, it seems best to take an inclusive approach. For most purposes, it would not make sense to specify the actual use of Big Data from a potential source in advance. The classification is meant to facilitate the exploration and analysis of potential uses, so options need to be kept open. This means that all five types of mobile phone data would be part of a single data source (i.e., this data source would be the subject of the classification), a social media source would include all data kept, and the search database of Google would be a data source, whether or not it is Google Trends that is actually used. This does not preclude analysing the optimal selection of data, or aggregate data, from the data source, or making distinctions between different options for purposes of analysis and exploration.

Similarly, it would not be of use to delineate the subject of the classification according to language (e.g., for social media) or country (e.g., for mobile phone providers). If one company provides the platform for social media messages in several languages, splitting it up accordingly would not increase the value of the Big Data classification. When looking at the possible use of such sources, the selection of language or region would of course be a relevant consideration, but it does not have to be reflected in the classification. However, in terms of classification criteria, it may be useful to distinguish between “local” and “global” Big Data sources. We come back to this in the next chapter.

5. Classification criteria

Once the Big Data sources have been delineated, the classification criteria are used to split the set of Big Data sources into groups. This can be done at several levels and in several ways, depending on the importance attached to the criteria. In this chapter possible classification criteria are identified, based on the uses described in chapter 3 and the references mentioned there. Selecting classification criteria and giving them weights is not done in this document, this remains to be done in the future. The reason for this is that a discussion must first be held with the intended users of the classification in order to know their preferences, and it is also too early to judge the importance of certain classification criteria, given uncertainties such as the fact that the SDG indicators still have to be chosen.

The possible classification criteria are listed below. Many of the criteria listed are not independent from each other and some may be highly correlated. The list is not necessarily complete. When discussing this with the intended users, more potential classification criteria may be added.

- Group Big Data sources according to **characteristics of the data** itself.

Many papers about Big Data start with discussing the definition of Big Data and characteristics. It appears that a number of characteristics, including the “three V’s”, occur quite often, but no single characteristic seems to be an unconditional requirement, not even the high volume of Big Data. This implies that Big Data sources can be grouped according to the presence or absence of such characteristics, or the degree to which they apply. Big Data characteristics are relevant to NSIs for several reasons, such as the usability of the data, methodological and IT implications and quality considerations. The following characteristics may be relevant:

  - high volume
  - high velocity
  - high variety and number of variables
  - high veracity
  - selectivity
  - (lack of) structure
  - high population dynamics
  - event based, continuous

- Make a distinction between **“local” and “global”** Big Data sources.

Websites or Twitter, for instance, are essentially not restricted to national territories, whereas, for example, systems of road sensors are operated by national authorities. This is not a sharp distinction. Mobile phone operators are a point in case. Vodafone is a worldwide company, but it works within national jurisdictions. For data, NSIs would have to contact mobile phone operators at the national level. However, CDRs are internationally standardised.

- Group Big Data sources according to **regulatory framework**.

Different regulatory frameworks may apply to different Big Data sources, for instance a financial regulatory framework for holders of financial transactions data or a framework regulating frequency bands for mobile phone operators. Similarly, different frameworks concerning privacy protection may apply.

- Distinguish between Big Data sources for which the data is a **by-product** or not.

Some Big Data sources, for instance administrative databases with road sensor data or satellite data, concern Big Data collected purposefully for the value of the data itself. However, Big Data may also be a by-product of the processes carried out by organisations to achieve their goals, for instance location data collected by mobile phone providers. This may affect the quality of the metadata and data accessibility.

- Group Big Data sources according to the **purpose and subject of the data** recorded.

For instance, there are sensors for various purposes, platforms for social networks with different purposes, etc. Websites may cover all imaginable subjects. The data may refer to persons, personal devices, cars, ships, transactions, etc., and may or may not constitute a data network.

- Distinguish between **original and derived** Big Data sources.

There are quite a few organisations that collect Big Data from other Big Data sources and make them available to others. This is mainly done for commercial reasons, and sometimes some value is added to the data. An example is the company Coosto [12], which collects pubic social media messages, adds value by adding a sentiment value to the messages, and sells subscriptions to its database. There are also a number of companies engaged in web scraping.

- Group Big Data sources according to their **relationship with the data**.

There are several kinds of relationships between the organisations behind the Big Data sources and the data. Some own the data, others don't. Some provide an exchange platform for others, some merely provide storage capacity, some process the data, etc.

- For the Big Data sources, make a distinction between **public and private organisations**.

If the Big Data source is kept by a government or other public organisation, access for NSIs may be easier than may be the case with private organisations that keep data that is not already publicly accessible. Big Data sources from government may even be influenced by NSIs. Further subdivisions of public and private organisations are conceivable. The distinction between the categories may be based on SNA.

- Make a distinction between Big Data sources that are sourced by **humans and machines**.

In particular social networks are about data provided by humans, whereas the Internet of Things is machine-based. In addition to human and machine sourced Big Data, the UNECE classification also mentions process-mediated data from business systems that record for instance transaction data. Thus, in addition to humans and machines, organisations could be a third source category. It may be noted that in the case of human and organisation sourced data, there is always interaction with machines. The distinction may be difficult to apply to some Big Data sources such as private body sensors used in sports.

- Group Big Data sources according to the degree of **stability of the source**.

Big Data sources come and go. For example, some of the earlier providers of internet search services and some earlier social platforms do not exist anymore. Big Data sources adapt to changing circumstances and evolve in nature. This is all very relevant to their usability for official statistics. Unfortunately, the degree of stability is not self-evident and the future of many Big Data sources cannot reliably be predicted.

- Group Big Data sources according to their **accessibility**.

Some Big Data sources, such as public website information or Twitter messages, can be accessed relatively easily, other sources such as internet search databases (micro level) or individual mobile phone subscription records are currently virtually beyond the reach of NSIs. Accessibility depends on a number of factors, such as privacy, data ownership, culture, costs, technical accessibility, etc., and may change over time.

- Make a distinction between Big Data sources with **real-time or accumulated data**.

This criterion may be important to NSIs, because it has implications for the way data can be processed, the methodology to be used and the IT requirements. It may also influence the timeliness and other aspects of statistics. This criterion may also be made to include the periodicity of observations.

- Group Big Data sources according to the **statistical methodology** needed for using the data.

There is a high demand for methodological guidance, as was shown in the 2015 UNSD Big Data Survey. However, although grouping Big Data sources according to the methodological dimension would be most helpful, it is questionable whether this is feasible. A Big Data source may be used for several different purposes (e.g., for making nowcasts, for replacing traditional surveys, as auxiliary data in combination with other data sources, etc.), each requiring its own methodology.

- Group Big Data sources according to the **statistical domains** for which they can be used.

Among other things, the 2014 and 2015 Big Data Surveys generated information about the statistical domains for which Big Data are already used or might be used in the future. The problems with grouping Big Data sources according to end-use are similar to those of the statistical methodology criterion. A Big Data source may be used for different statistical domains (e.g. mobile phone data or social media data), and domains may change, may not yet be known or may be added in time. Moreover, statistics belonging to a certain domain may be based on various types of Big Data sources.

- The grouping of Big Data sources may take their usability for **SDG indicators** into account.

This criterion is similar to the statistical domain criterion, and it has similar associated problems. The two criteria may be combined. At the moment not much is known about the usability of Big Data sources for SDG indicators, as was shown in the 2015 Survey of Big Data Initiatives for SDGs, carried out by the Task Team on Using Big Data for SDGs, although additional insights may be drawn from that survey. That task team is also compiling a table linking the Big Data projects in a consolidated project inventory to each of the 169 Big Data targets. These products could be the basis for the compilation of a matrix of types of Big Data sources versus SDG goals or targets in 2016, based on the knowledge available then.

Although the classification criteria to be actually used are not selected in this document, some considerations may be mentioned. A grouping of Big Data sources which reflects similarities in the way they can be used, for instance in respect of applicable statistical methods or privacy treatment, may be convenient to users, but one could also decide not to use such classification criteria at all, because this can be dealt with separately from the classification. For example, one could design and fill a matrix in which the Big Data sources represent one dimension, and statistical methods another. This could also be done in respect of other dimensions, such as the usability of Big Data sources for SDG indicators, as mentioned above. However, in all such cases complexity may be reduced by already incorporating use-related criteria in the grouping of Big Data sources.

The more the classification of Big Data is used as a checklist and a list to be crossed with other dimensions considered of interest, the less classification criteria are needed. As seen in chapter 3, the UNECE classification of Big Data has so far mainly been used as a checklist, but the needs are growing, as shown by the 2015 UNSD Survey on Big Data.

6. Conclusions

summary

The TTCC proposes that a new classification of Big Data sources be constructed, building on and replacing the 2013 UNECE Classification of Big Data for Official Statistics. The new classification would use the UNECE definition of Big Data, in which Big Data in the context of official statistics is defined as Big Data *sources*. The definition determines the scope of the classification.

The construction of the classification is to be based on the intended uses. Several uses of the UNECE classification were mentioned, and other needs were also identified. Future uses may include informing the discussion on the use of Big Data for compiling SDG indicators. The 2015 UNSD Survey on Big Data also provided valuable information on intended uses.

The Big Data sources to which the classification is meant to refer need delineation. This can be done by starting with the set of Big Data organisations and if needed – i.e. if they harbour completely different types of Big Data – distinguish more than one Big Data source for them.

Fifteen potential classification criteria have been identified. A selection of these may be used for the construction of the classification.

work to be done

The actual construction of the classification of Big Data sources remains to be done. This has to be based on a user assessment of the proposed approach to the classification and the potential classification criteria identified.

Once the criteria have been assessed and prioritised, they have to be worked out further, by providing more elaborate descriptions and listing the value range for each of them. They can then be applied to construct the levels of the classification. However, using a large number of dimensions would not be practical, the number of criteria to be used has to be limited.

The classification should be flexible, and should be able to evolve over time. Initially, this would probably mean relatively short periods between revisions. Flexibility may also be obtained by constructing a system for classifying Big Data sources on demand rather than a fixed classification. In that case, methods and rules would be needed, and possibly a larger number of criteria could be accommodated.

The construction of the classification – or classification system – will take some time, but in the meantime there will be more clarity on the SDG indicators, which will be decided on in February 2016. The aim is to have a draft classification ready by the end of 2016.

References

[1] Terms of Reference of the Task Team on Cross-cutting issues, classifications, frameworks and taxonomy, Revision 8, 3 October 2015.

[2] http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

[3] UNECE (2013). What does "Big Data" mean for official statistics? Conference of European Statisticians, 10 March 2013.

[4] Statistics New Zealand, (2010). Best Practice for Classifications, Statistics New Zealand, Christchurch.

[5] http://www1.unece.org/stat/platform/display/msis/Final+project+proposal%3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production

[6] http://www1.unece.org/stat/platform/display/bigdata/Sandbox

[7] http://www1.unece.org/stat/platform/display/bigdata/2014+Project

[8] Rob Kitchin (2015). The opportunities, challenges and risks of big data for official statistics. Statistical Journal of the IAOS 31 (2015), pp 471-481.

[9] IBM (2014). Big Data Roadmap Assessment for Statistics Netherlands. Leen Molendijk and Puja Nanda, IBM Global Business Services, 19 May 2014.

[10] https://trello-attachments.s3.amazonaws.com/54de47ebc943c0d418dcd166/5534ce99303d810783e9bb1b/96100f4e273a567c83c862b1bf94e6c1/Selection_criteria_v2_UN.xlsx

[11] https://trello-attachments.s3.amazonaws.com/54de47ebc943c0d418dcd166/5534ce99303d810783e9bb1b/c787d28003cb75020c1ce6fc7052db58/Explanatory_remarks_to_the_decision_matrix.docx

[12] http://www.coosto.com/

Annex: The 2013 UNECE classification of Big Data

1. Social Networks (human-sourced information): this information is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video. Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. Data are loosely structured and often ungoverned.

1100. Social Networks: Facebook, Twitter, Tumblr etc.
1200. Blogs and comments
1300. Personal documents
1400. Pictures: Instagram, Flickr, Picasa etc.
1500. Videos: Youtube etc.
1600. Internet searches
1700. Mobile data content: text messages
1800. User-generated maps
1900. E-Mail

2. Traditional Business systems (process-mediated data): these processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected is highly structured and includes transactions,reference tables and relationships, as well as the metadata that sets its context. Traditional business data is the vast majority of what IT managed and processed, in both operational and BI systems. Usually structured and stored in relational database systems. (Some sources belonging to this class may fall into the category of "Administrative data").

21. Data produced by Public Agencies
    2110. Medical records
22. Data produced by businesses
    2210. Commercial transactions
    2220. Banking/stock records
    2230. E-commerce
    2240. Credit cards

3. Internet of Things (machine-generated data): derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches.

- 31. Data from sensors
  - 311. Fixed sensors
    - 3111. Home automation
    - 3112. Weather/pollution sensors
    - 3113. Traffic sensors/webcam
    - 3114. Scientific sensors
    - 3115. Security/surveillance videos/images
  - 312. Mobile sensors (tracking)
    - 3121. Mobile phone location
    - 3122. Cars
    - 3123. Satellite images
- 32. Data from computer systems
  - 3210. Logs
  - 3220. Web logs