SOEN 6611 - SOFTWARE MEASUREMENT: THEORY AND PRACTICE
Project Report on Task 3
SUMMER 2022
Course Instructor: Dr. Olga Ormandjieva

Source: SEI Implementing Goal-Driven Measurement course material (adapted).

Team 11	
#Student ID	Name
40198687	Hasandeep Singh
40159259	Anushka Sharma
40218417	Jasleen Kaur
40205476	Kavleen Kour Sidhu

Revamping Goals of Step1: Quantify goals

Measurement Goal Label:	**Description**	**Corresponding business goal (write its label)**
MG1	Compare the change in volume of big data at different time intervals. The volume of preprocessed dataset must be greater than threshold value of 60% to generalize and predict future results.	**BG-01**(Volume)
MG2	Despite changes to the dataset, the goal is to process the dataset without many changes in the existing infrastructure and the system must be at least scaled up to 5 to 6 replicas under increased velocity.	**BG-02** (Velocity)
MG3	To have a dataset diversity between 10% to 12% with for unbiased dataset. Characterize the different types of big data gathered.	**BG-03** (Variety)
MG4	To determine the vincularity of the pipeline for linkage and connectivity. The system must ensure a traceability of at least 0.5 value.	**BG-04** (Vincularity)
MG5	Compare veracity and check if the data is correct and relevant to the final goal. Evaluate current ness and system must ensure data more than 10 to 15 years of age is eliminated.	**BG-05** (Veracity)
MG6	Evaluate the data source credibility which can be based on factors such as rating, scores.	**BG-06** (Validity)

Objective: develop and document Success Criteria and Indicators, derived measures and base measures

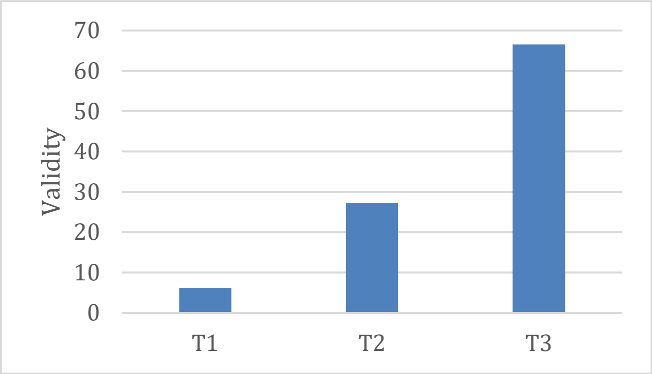
Step3-Part 1 (6 points): derive Success Criteria and Indicators (for Validity, Vincularity and Veracity)

The objective of Part 1 is to develop success criteria and success indicators.

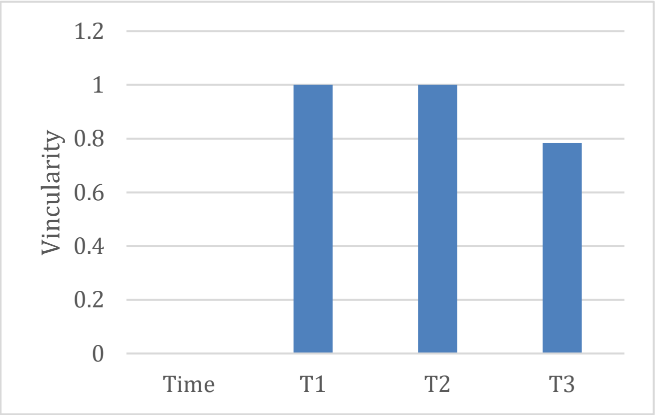
Success (answering the measurement question within the desired timeframe) that can only be achieved when certain conditions are in place. indicators that will allow you to answer the questions quantitatively and then communicate the results to others.

For each measurement question related to Validity, Vincularity and Veracity, develop success criteria that will allow you to answer the measurement questions quantitatively.

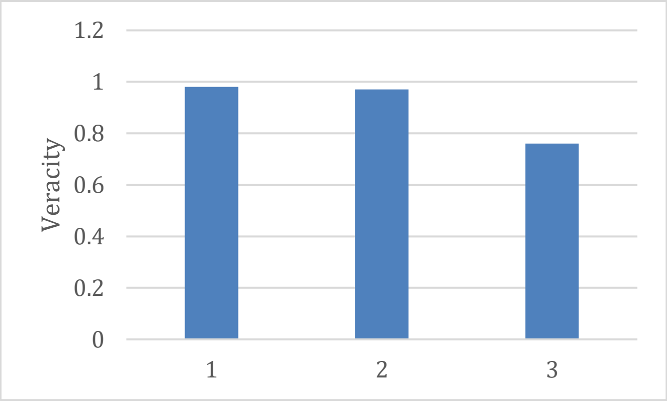
Mesurément Question Label / Operationalized Goal Label	MG1 - Validity Verifying the data source credibility of the data at regular time intervals against set rules.
Success Criteria Label and description	Success Criteria Label: SC01 Description: The success criteria of Validity compromises of increased compliance and credibility of the data to Topic (of Ukraine conflicts) according to ISO standards. According to the standards, data is aligned to avoid discrepancies among different stakeholders.
Indicator Label and description	Indicator Label: I1_Val Description: I1_Val measures the degree/ percentage of data compliance and credibility, which in turn gives us the Validity.
Indicator Analysis Model and Interpretation	Indicator Analysis Model: The compliance and credibility values derive the validity of the data. Both the values contribute equally to our final value. So, we are using the 50% summation of both values to give us the Validity. Interpretation: Ideally, data should be 100% valid at all times but in real-world data not all data points are valid towards the main goal/ topic. Validity will correspond directly to label SC01.

	Validity: > 90% - Highly meaningful data 80 - 90% - Slightly off/less significant data 60 - 80% - Somewhat relevant data < 60% - Unconnected data
Indicator Sketch	

Mesurément Question Label / Operationalized Goal Label	MG2 - Vincularity Data should be connected and traceable to reach data points in the entire landscape.
Success Criteria Label and description	Success Criteria: SC02 Description: The big data we have must ensure a 0.5 traceability value or above at all times. The percentage of recordings that are traceable grows as record length increases. Vincularity increases due to increased traceability.
Indicator Label and description	Indicator Label: I2_Mvin Description: I2_Mvin indicates the degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.

Indicator Analysis Model and Interpretation	Indicator Analysis: The model used in this helps us define vincularity of data in terms which are relevant to the topic of our data. Traceability derive the vincularity for the data. Traceability is ensuring all data is traceable across the entire landscape. Interpretation: The data increase over time is measured, compared, and the findings are displayed in the graph. Traceability is directly related to vincularity, so any increase/decrease in the former will infer a similar change in the latter. While the traceability should never drop below 0.5 as that will directly indicate the disconnect in data has become such that it has become irrelevant. The data similarity or data rule must always be 1. If it falls below then the data set becomes decoupled.
Indicator Sketch	

Mesurément Question Label / Operationalized Goal Label	MG3 - Veracity Compare the veracity of big data at different time intervals. To check if the data is correct and relevant to the final goal.
Success Criteria Label and description	Success Criteria: SC03 Description:

	- Accuracy- The dataset needs to be 90% accurate over the given time frames. The redundant values in all columns should correspond to zero (null values). - Completeness- The subject data should have values for all attributes and records i.e it should be at least 60% absolute (unique values). - Correctness- With regards to the indicator plot, all acceptable values should be in between lower and upper range.
Indicator Label and description	Indicator Label: I3_Mver Description: I2_Mver indicates the accuracy of evaluating the quality results. It needs to be range compliant with truth and rules need to be satisfied.
Indicator Analysis Model and Interpretation	Indicator Analysis Model: The accuracy, completeness and correctness values derive the veracity of the data,So, we are using approximately 33.33% summation of all values to give us the final serving of veracity. Interpretation: In ideal scenario, data should be 100% versatile but the in reality difference between the consecutive timeframes should remain nearly constant. For example, if we have three time frames => T3-T2= T2-T1 ~ 0.
Indicator Sketch	

Step3-Part 2 (4 points): The objective of Part 2 is to define all measures required to derive your V's indicators (for **Validity, Vincularity and Veracity**) and decide on the achievement of the corresponding operationalized goals.

3.2.1 Identification of the V's measures (for Validity, Vincularity and Veracity), tracing them to the corresponding indicators, their availability and source

For each of the **V's** indicators (for **Validity, Vincularity and Veracity**), identify all required measures (derived and base). The table below will be used to complete each of these measures in sections 3.2 and 3.3. It is also recommended that you review and complete this table after all measures have been defined.
This table therefore gives a good summary of all the measurements to be collected and analyzed.

Indicator Level	**Indicators**	**Formula**
I1	Mval	$Mval\ (MDS) = Credability\ (MDS) * W_{Cred} + Compliance(MDS) * W_{Compli}$
		$Compliance\ (MDS) = \frac{\sum_{\forall\ DS \in MDS} Nrec_{comp}(DS)}{Nds(MDS)}$
		$Credability\ (MDS) = \frac{Nds_{cr}(MDS)}{Nds\ (MDS)}$
I2	Mvin	$Mvin\ (MDS) = \frac{\sum_{\forall\ DS \in MDS} Traceability\ (DS)}{Nds\ (MDS)}$
		$Traceability\ (DS) = \frac{Rec_{Trace}(DS)}{Ldst\ (DS)}$
I3	Mver	$Mver\ (MDS) = Accuracy\ (MDS) * W_{Acc} + Completness(MDS) * W_{Comp} + Currentness\ (MDS) * W_{Curr} + Availability * W_{Avail}$
		$Accuracy\ (MDS) = \frac{H_{acc}}{H\_\max}$
		$Hacc(MDS) = \mathrm{Log}\ 2\ (Lbd) - \left(\left(\frac{1}{Lbd}\right) * \sum_{j = [1..k]} P_j\ \mathrm{Log}_2 (P_j)\right)$
		$H_{max}\ (MDS) = \log_2(Lbd)$
		$Com_m\ (MDS) = \frac{[rec\_no\_null\ (MDS)]}{Lbd(MDS)}$
		$Currentness\ (MDS) = \frac{[rec\_acc\_age\ (MDS)]}{Lbd(MDS)}$
		$Availability\ (MDS) = \frac{[n\_succ\_req\ (MDS)]}{n\_req(MDS)}$

Measures					Indicator(s) label		
#	Identification (name of the measure)	Type	Availability	Source	<l1>	<l2>	<I3>
1	Credibility	Derived	B	Dataset	X		
2	Compliance	Derived	B	Dataset	X		
3	**Mval (Validity)**	Derived	B	Dataset	X		
4	Traceability	Derived	B	Dataset		X	
5	**Mvin (vincularity)**	Derived	B	Dataset		X	
6	Accuracy	Derived	B	Dataset			X
7	Hmax-Max entropy	Derived	C	Dataset			X
8	Hacc-Entropy of Multiple Datasets	Derived	C	Dataset			X
9	Currentness	Derived	B	Dataset			X
11	Availability	Derived	B	Dataset			X
12	**Mver(Veracity)**	Derived	B	Dataset			X

13	Completeness	Derived	B	Dataset			X
14	Nds- Number of datasets	Base	A	Dataset	X	X	
10	Nds_cr - Number of credible Datasets	Base	C	Dataset	X		
15	Nrec_comp- Number of compliant records in a Dataset	Base	C	Dataset	X		
16	Rec_Trace-Provides the total number of records that are traceable in MDS	Base	C	Dataset		X	
17	Ldst	Base	C	Dataset		X	
18	Rec_no_null (MDS) -Frequency of records (in MDS) with no null values	Base	C	Dataset			X
19	Lbd-Total Number of records in MDS	Base	A	Dataset			X
20	Rec_acc_age - Provides the total number of records with ages that fall within the acceptable range	Base	C	Dataset			X
21	Pj - Provides the total number of duplicate items and their specific count in each dataset	Base	C	Dataset			X
22	N_succ_req - Number of successful requests	Base	C	Dataset			X

23	N_req - Number of Requests	Base	A	Dataset			*X*

***Type**: "Derived" or "Base".*

Availability:

"A": Already available and collected;
"B": Can be derived from other data fairly directly;
"C": Possibly obtained with minor effort;
"D": Not available at the moment;
"E": Very difficult, if not impossible to obtain at the moment.

***Source**: Place or tool where data is collected. In the case of base measures, this is obvious; in the case of derived measures, it depends on where the base data is stored after collection.*

***Indicator (s):** Mark an "X" when this measurement is required for each of your indicators.*

3.2.2 Validity, Vincularity and Veracity: Derived measures definitions and operationalization

Using the template for defining a derived measure (see the file < Derived-Base-Mesures-templates.docx>), complete the fields required for each of the derived measures identified in 3.2.1

1)

<table>
<tr><td colspan="6"><b>Derived measure or indicator:</b></td></tr>
<tr><td>#</td><td>Derived measure or indicator
Big Data Validity : <b>Mval</b></td><td colspan="4">Formula
$Mval\ (MDS) = Credability\ (MDS) * W_{Cred} + Compliance(MDS) * W_{Compli}$
Compliance – degree to which data has attributes that adhere to standards
Credibility - degree to which data has attributes that are regarded as true and believable by users
W_{Cred} : Weight of Credibility (Set to 1/2 by default)
W_{Compli} : Weight of Compliance (Set to 1/2 by default)
Sum of all weights is equal to 1</td></tr>
<tr><td colspan="3">Link with the measurement goal (which goal)
Measurement Goal 1</td><td>Responsible (who analyzes)
Big data managers</td><td>Stakeholder (who uses)
User, Strategy manager</td><td>Frequency (when)
<b>Monthly</b></td></tr>
<tr><td colspan="3">Data source (where the measurement data will be extracted from)
<b>https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds</b></td><td>Storage of the result (where data will be stored after the extraction)
Google drive or local machine (to store metadata)</td><td colspan="2">Data interpretation rules
The compliance and credibility values derive the validity of the data. The</td></tr>
</table>

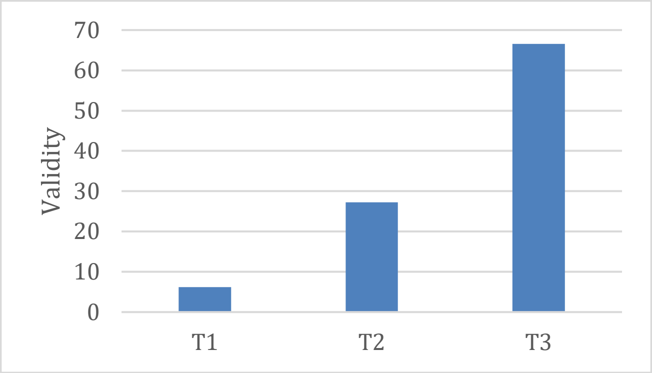
validity of the data above 90% is considered meaningful.

Analysis procedure

The compliance and credibility values derive the validity of the data. Both the values contribute equally to our final value. So, we are using the 50% summation of both values to give us the Validity.

$$Mval\ (MDS) = Credability\ (MDS) * W_{Cred} + Compliance(MDS) * W_{Compli}$$

Presentation of the results (sketch illustrating what it looks like):



Potential decision making depending on the results

The calculated Mval must be above 90% for highly meaningful data and between 80-90% for slightly significant data.

2)

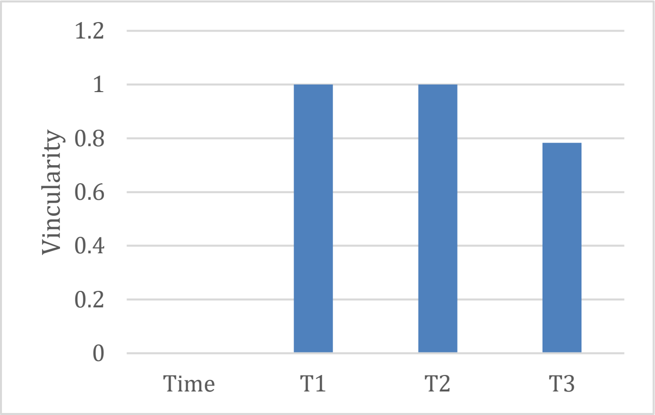
Derived measure or indicator:		
#	Derived measure or indicator **Mvin**	Formula $Mvin\,(MDS) = \frac{\sum_{\forall\, DS \in MDS} Traceability\,(DS)}{Nds\,(MDS)}$ Traceability measures provide the degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use. Nds – Number of Datasets

Link with the measurement goal (which goal) Measurement Goal 2	Responsible (who analyzes) Big data managers	Stakeholder (who uses) User, Strategy manager	Frequency (when) **Monthly**

Data source (where the measurement data will be extracted from) **https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds**	Storage of the result (where data will be stored after the extraction) Google drive or local machine (to store metadata)	Data interpretation rules **The Vincularity is measured using the ratio of traceable data within all datasets, which ranges from 0-100, where 100 indicates that all of them are traceable across MDS.**

Analysis procedure	

<table>
<tr>
<td></td>
<td rowspan="2">Presentation of the results (sketch illustrating what it looks like):



</td>
</tr>
<tr>
<td>Potential decision making depending on the results

The sets of data recorded are traceable and reliable.</td>
</tr>
</table>

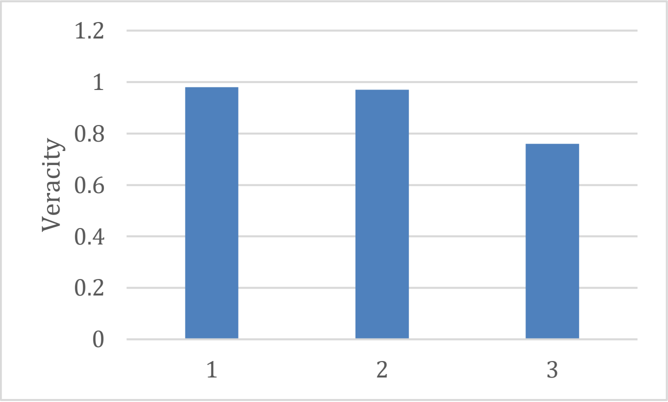
3)

Derived measure or indicator:		
#	Derived measure or indicator **Mver**	Formula $Mver\ (MDS) = Accuracy\ (MDS) * W_{Acc} + Completness(MDS) * W_{Comp} + Currentness\ (MDS) * W_{Curr} + Availability * W_{Avail}$ Accuracy - Degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use. Completeness - Degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use Currentness- Degree to which data has attributes that are of the right age in a specific context of use. Completeness - Degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use. W_{Ndde} : Weight of Ndde (Set to 1/4 by default) W_{Lbd} : Weight of Lbd (Set to 1/4 by default) W_{Nds} : Weight of Nds (Set to 1/4 by default) Sum of all weights is equal to 1

Link with the measurement goal (which goal) **Measurement Goal 3**	Responsible (who analyzes) Big data managers	Stakeholder (who uses) Strategy manager	Frequency (when) **Monthly**

Data source (where the measurement data will be extracted from)	Storage of the result (where data will be	Data interpretation rules

https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds	stored after the extraction) Google drive or local machine (to store metadata)	Since all of Veracity's sub values fall between 0 and 1, and since all weights are taken into account to be 0.25, it is clear that the ideal value for Veracity is 1.0. The higher values appear to be more successful than those with lower veracity rates.

Analysis procedure	Presentation of the results (sketch illustrating what it looks like): 
Potential decision making depending on the results	

4)

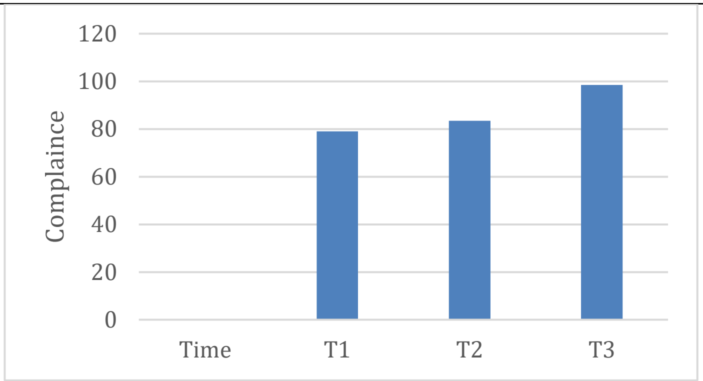
Derived measure or indicator:

#	Derived measure or indicator Compliance	Formula $Compliance\ (MDS) = \frac{\sum_{\forall\, DS \in MDS} Nrec_{comp}(DS)}{Nds(MDS)}$ • Nrec_comp: Number of compliant records in a Dataset • Nds_cr: Number of credible Datasets

Link with the measurement goal (which goal) **Measurement Goal 1**	Responsible (who analyzes) Strategy manager	Stakeholder (who uses) Data Engineers, Users	Frequency (when) **Monthly**

Data source (where the measurement data will be extracted from) **https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds**	Storage of the result (where data will be stored after the extraction) Google drive or local machine (to store metadata)	Data interpretation rules Compliance: Nrec_comp: Compliant records Nds: Total Number of datasets The value of compliance must be closer to 1 which means a highly compliant dataset.

Analysis procedure	Presentation of the results (sketch illustrating what it looks like):

The compliance and credibility values derive the validity of the data. Both the values contribute equally to our final value. So, we are using the 50% summation of both values to give us the Validity. The closer the compliance value to 1, the more valid the dataset is.	

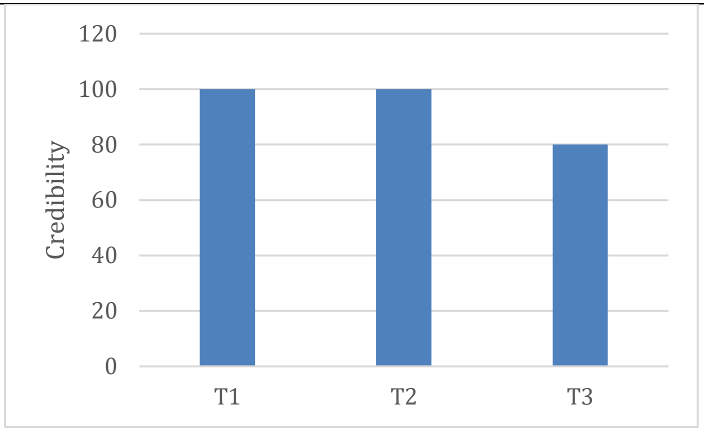
5)

Derived measure or indicator:		
#	Derived measure or indicator Credibility	Formula $Credability\ (MDS) = \frac{Nds_{cr}(MDS)}{Nds\ (MDS)}$ • Nds_cr: Number of credible Datasets • Nds : Number of datasets

Link with the measurement goal (which goal) **Measurement Goal 1**	Responsible (who analyzes) Strategy Managers	Stakeholder (who uses) Data Engineers,Users	Frequency (when) **Weekly**

Data source (where the measurement data will be extracted from) **https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds**	Storage of the result (where data will be stored after the extraction) Google drive or local machine (to store metadata)	Data interpretation rules Credibility: Nds_cr : Credible Datasets Nds: Total number of datasets The value of credibility must be closer to 1 which means a highly credible dataset.

Analysis procedure The compliance and credibility values derive the validity of the data. Both the values contribute equally to our final value. So, we are using	Presentation of the results (sketch illustrating what it looks like):

the 50% summation of both values to give us the Validity. The closer the credibility value to 1, the more valid the dataset is.	

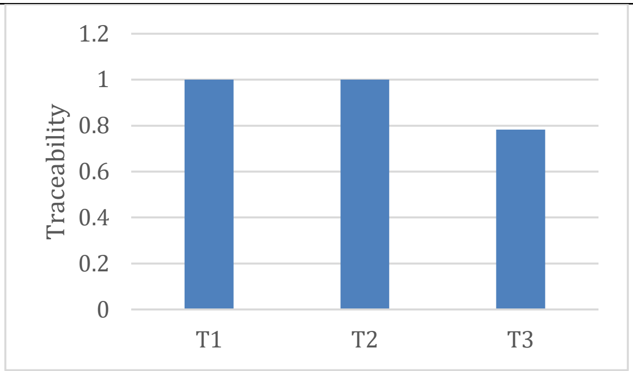
6)

Derived measure or indicator:		
#	Derived measure or indicator Traceability	Formula $Traceability\ (DS) = \frac{Rec_{Trace}(DS)}{Ldst\ (DS)}$ Ldst (Length of the Record): Total number of occurrences of data elements in dataset (DS) Rec_Trace: Provides the total number of records that are traceable in MDS

Link with the measurement goal (which goal) **Measurement Goal 2**	Responsible (who analyzes) Big data managers	Stakeholder (who uses) Strategy manager	Frequency (when) **Monthly**

Data source (where the measurement data will be extracted from) **https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds**	Storage of the result (where data will be stored after the extraction) Google drive or local machine (to store metadata)	Data interpretation rules The big data we have must always ensure a 0.5 traceability value or above.

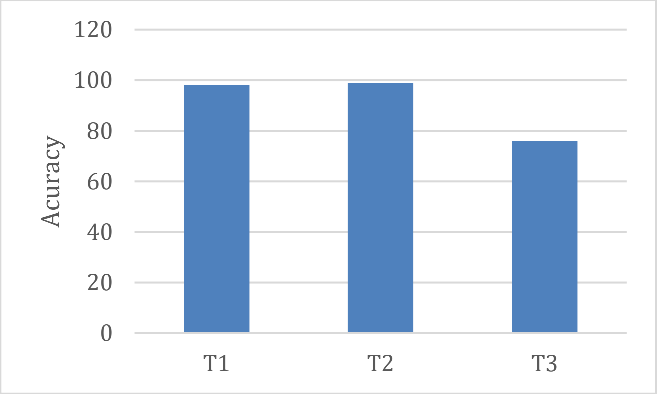
Analysis procedure	Presentation of the results (sketch illustrating what it looks like):

The data increase over time is measured, compared, and the findings are displayed in the graph Traceability derive the vincularity for the data. Traceability is ensuring all data is traceable across the entire landscape	

7)

Derived measure or indicator:		
#	Derived measure or indicator **Accuracy**	Formula $$Accuracy\ (MDS) = \frac{H_{acc}}{H\_\max}$$ Hacc: Entropy of Multiple Datasets. How much duplication exists? Hmax: Max entropy. Theoretically the least amount of duplication

Link with the measurement goal (which goal)	Responsible (who analyzes)	Stakeholder (who uses)	Frequency (when)
Measurement Goal 2	Strategy Managers	Developers, Data Engineers	**Monthly**

<table>
<tr>
<td>Data source (where the measurement data will be extracted from)
<b>https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds</b></td>
<td colspan="2">Storage of the result (where data will be stored after the extraction)
Google drive or local machine (to store metadata)</td>
<td>Data interpretation rules
Estimating the quality parameter value and overall data accuracy in determining which characteristics are applicable to improve the veracity</td>
</tr>
<tr>
<td colspan="2">Analysis procedure
The dataset needs to be 90% accurate over the given time frames. The redundant values in all columns should correspond to zero</td>
<td colspan="2">Presentation of the results (sketch illustrating what it looks like):
 </td>
</tr>
</table>

8)

<table>
<tr><th colspan="7">Derived measure or indicator:</th></tr>
<tr><td>#</td><td>Derived measure or indicator
Accuracy</td><td colspan="5">Formula
$Accuracy\ (MDS) = \frac{H_{acc}}{H\_\max}$</td></tr>
<tr><td colspan="3">Link with the measurement goal (which goal)
Measurement Goal 2</td><td colspan="2">Responsible (who analyzes)
Strategy Managers</td><td>Stakeholder (who uses)
Developers, Data Engineers</td><td>Frequency (when)
Monthly</td></tr>
<tr><td colspan="3">Data source (where the measurement data will be extracted from)
https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds</td><td colspan="2">Storage of the result (where data will be stored after the extraction)
Google drive or local machine (to store metadata)</td><td colspan="2">Data interpretation rules
Estimating the quality parameter value and overall data accuracy in determining which characteristics are applicable to improve the veracity</td></tr>
<tr><td colspan="4">Analysis procedure</td><td colspan="3">Presentation of the results (sketch illustrating what it looks like):</td></tr>
</table>

The dataset needs to be 90% accurate over the given time frames. The redundant values in all columns should correspond to zero	

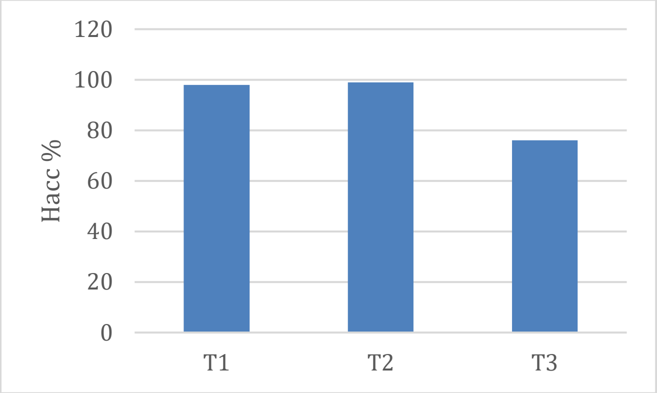
9)

Derived measure or indicator:		
#	Derived measure or indicator Hacc-Entropy of Multiple Datasets	Formula $H_{acc}\,(MDS) = \log_2(Lbd) - \frac{1}{Lbd * \sum_{j=[1\ldots k]} p_j \log_2(p_j)}$ Pj: Provides the total number of duplicate items and their specific count in each dataset

Link with the measurement goal (which goal) **Measurement Goal 2**	Responsible (who analyzes) Strategy Managers	Stakeholder (who uses) Developers, Data Engineers, Users	Frequency (when) **Weekly**

Data source (where the measurement data will be extracted from) **https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds**	Storage of the result (where data will be stored after the extraction) Google drive or local machine (to store metadata)	Data interpretation rules Hacc is used to calculate the degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.

Analysis procedure	Presentation of the results (sketch illustrating what it looks like):

$Accuracy\ (MDS) = \frac{H_{acc}}{H\_\max}$ **Hacc is used to calculate how much duplication exists in the dataset.**	

10)

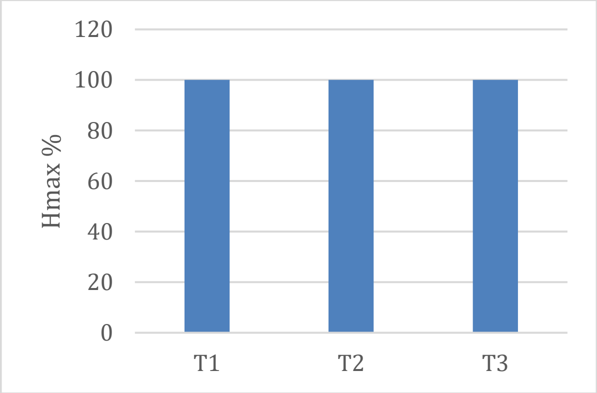
Derived measure or indicator:			
#	Derived measure or indicator Hmax-Max entropy	Formula $H_{max}\ (MDS) = \log_2(Lbd)$	
Link with the measurement goal (which goal) **Measurement Goal 2**	Responsible (who analyzes) Strategy Managers	Stakeholder (who uses) Developers, Data Engineers, Users	Frequency (when) **Weekly**
Data source (where the measurement data will be extracted from)	Storage of the result (where data will be	Data interpretation rules	

<table>
<tr>
<td colspan="2"><b>https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds</b></td>
<td colspan="2">stored after the extraction)

Google drive or local machine (to store metadata)</td>
<td>Hmax is used to calculate the degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.</td>
</tr>
<tr>
<td colspan="3">Analysis procedure

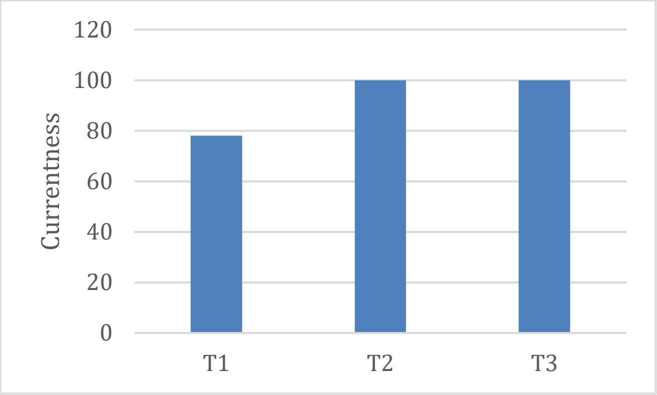
$$Accuracy\ (MDS) = \frac{H_{acc}}{H\_\max}$$

<b>Hmax is used to calculate the max entropy of the dataset.</b></td>
<td colspan="2">Presentation of the results (sketch illustrating what it looks like):


</td>
</tr>
</table>

11)

Derived measure or indicator:			
#	Derived measure or indicator Currentness	Formula $Currentness\ (MDS) = \frac{[rec\_acc\_age\ (MDS)]}{Lbd(MDS)}$	
Link with the measurement goal (which goal) **Measurement Goal 3**	Responsible (who analyzes) Strategy Managers	Stakeholder (who uses) Developers, Data Engineers	Frequency (when) **Weekly**
Data source (where the measurement data will be extracted from) **https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds**	Storage of the result (where data will be stored after the extraction) Google drive or local machine (to store metadata)	Data interpretation rules Currently refers to the degree to which data has attributes that are of the right age in a specific context of use. The higher the value of the currently the more relevant the dataset if for the frame.	

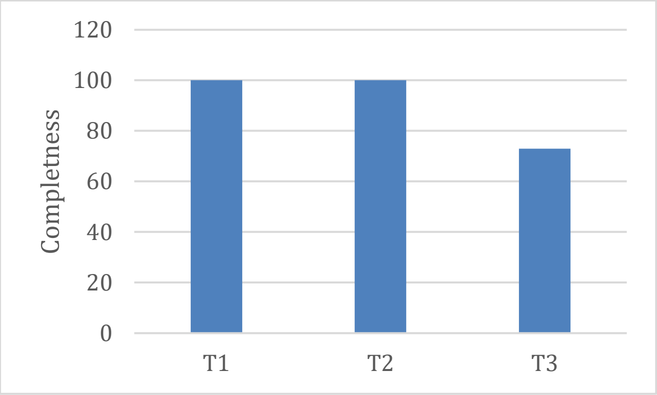
Analysis procedure Calculate the rec_acc_age as the **the total number of records with ages that fall within the acceptable range based on the upper and lower quartiles of the Box and Whisker and use the ratio of rec_acc_age to lbd for calculating the currentness. The higher the currentless, the more the veracity of the system.**	Presentation of the results (sketch illustrating what it looks like): 

12)

Derived measure or indicator:		
#	Derived measure or indicator Completeness	Formula $Com_m\,(MDS) = \frac{[rec\_no\_null\,(MDS)]}{Lbd(MDS)}$ • **Rec_no_null (MDS)** : Frequency of records (in MDS) with no null values • **Lbd (MDS):** Total Number of records in MDS

Link with the measurement goal (which goal) **Measurement Goal 3**	Responsible (who analyzes) Strategy Managers, Big Data Managers	Stakeholder (who uses) Developers, Data Engineers, Users	Frequency (when) **Weekly**

Data source (where the measurement data will be extracted from) **https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds**	Storage of the result (where data will be stored after the extraction) Google drive or local machine (to store metadata)	Data interpretation rules Completeness refers to the degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use. The higher the value of the completeness the more it's veracity is.

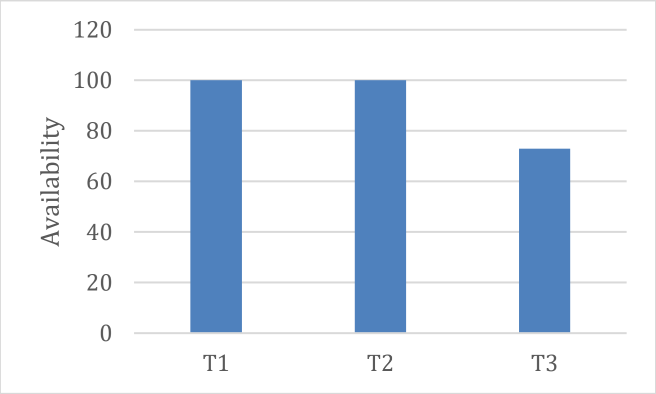
Analysis procedure	Presentation of the results (sketch illustrating what it looks like):
The completeness is calculated as the ratio of frequency of records that have no null value to the total number of records. The higher the value of completeness is, the more the the veracity as they are directly proportional.	

13)

Derived measure or indicator:		
#	Derived measure or indicator Availability	Formula $Availability\,(MDS) = \frac{[n\_succ\_req\,(MDS)]}{n\_req(MDS)}$ • N_succ_req (MDS) : Number of successful requests (from an API, server, datastore, origins of data, etc) • N_req (MDS): Number of requests

Link with the measurement goal (which goal) **Measurement Goal 3**	Responsible (who analyzes) Developers	Stakeholder (who uses) User	Frequency (when) **Daily**

Data source (where the measurement data will be extracted from) **https://www.kaggle.com/code/akuppps/ukrainianconflict-subreddit-eda-and-wordclouds**	Storage of the result (where data will be stored after the extraction) Google drive or local machine (to store metadata)	Data interpretation rules Availability refers to the degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of use. The system must be available at all times.

Analysis procedure The availability of the system is defined as the successful requests to the total request. The availability of the system must be close to 1. The closer the system availability to 1, the higher the veracity.	Presentation of the results (sketch illustrating what it looks like): 

3.2.3 Validity, Vincularity and Veracity: Base measures definitions and operationalization

Using the template for defining a base measure (see the file < Derived-Base-Mesures-templates.docx>, complete the fields requested for each of the base measures identified in 3.2.1.

<table>
<tr><td colspan="6">Base measure: <b>Lbd</b></td></tr>
<tr><td>#</td><td colspan="2">Measure (what: entity, attribute)
<b>Lbd-Total Number of records in Big Data</b>
<b>Entity: Dataset</b>
<b>Attribute: Size</b></td><td>Scale type
<b>Absolute Scale</b></td><td colspan="2">Applicability
<b>It gives the actual length of the dataset and can be used to evaluate veracity in the dataset.</b></td></tr>
<tr><td>Who measures?
<b>Developer/ Data Scientist</b></td><td>Source of measurement
https://www.kaggle.com/akuppps/ukrainianconflict-top-comments</td><td colspan="2">Where to store the result
<b>Local Storage or any distributed File System</b></td><td>Tool
<b>Google Colab</b></td><td>Time (when to measure)
<b>During each time frame, the length of the new dataset is calculated.</b></td></tr>
<tr><td colspan="3">Collection procedure (how to collect the data)
<b>By using the Python code in Google collab online platform to find the distinct elements in the dataset.</b></td><td colspan="3">Notes or comments:
<b>This measure is to calculate the variety</b></td></tr>
</table>

<table>
<tr><td colspan="6">Base measure: : **Nds**</td></tr>
<tr><td>#</td><td colspan="2">Measure (what: entity, attribute)
Nds : No of Dataset in Big Data
Entity: Data set
Attribute: number of datasets</td><td>Scale type
Absolute scale</td><td colspan="2">Applicability
It counts the number of datasets which are present to analyze the variations coming in each measure during different time period</td></tr>
<tr><td colspan="2">Who measures?
Developer/ Data Scientist</td><td>Source of measurement
https://www.kaggle.com/akuppps/ukrainianconflict-top-comments</td><td>Where to store the result
Local Storage or any distributed File System</td><td>Tool
Google Colab</td><td>Time (when to measure)
During each time frame, the count of the new dataset is calculated.</td></tr>
<tr><td colspan="3">Collection procedure (how to collect the data)
By using the Python code in Google collab online platform to divide the entire dataset and allocate it to different time frames.</td><td colspan="3">Notes or comments:
This measure is to calculate the validity.</td></tr>
</table>

Base measure: **Nds_cr - Number of credible Datasets**				
#	Measure (what: entity, attribute) **Nds_cr - Number of credible Datasets** **Entity: Data set** **Attribute: size**	Scale type **Absolute scale**	Applicability **Calculates the total credible data values in the dataset giving us credibility which further is used to calculate Validity.**	
Who measures? **Developer/ Data Scientist**	Source of measurement **https://www.kaggle.com/akuppps/ukrainianconflict-top-comments**	Where to store the result **Local Storage or any distributed File System**	Tool **Google Colab**	Time (when to measure) **During each time frame, the count of the credible data is calculated.**
Collection procedure (how to collect the data) **By using the Python code in Google collab online platform to divide the entire dataset and allocate it to different time frames.**	Notes or comments: **Used to calculate Credibility that further is used to calculate Validity.**			

Base measure: **Nrec_comp- Number of compliant records in a Dataset**

#	Measure (what: entity, attribute) **Nrec_comp - Number of compliant records in a Dataset** **Entity: Data set** **Attribute: size**	Scale type **Absolute scale**	Applicability **Calculates the total compliant data values in the dataset giving us compliance which further is used to calculate Validity.**

Who measures? **Developer/ Data Scientist**	Source of measurement **https://www.kaggle.com/akuppps/ukrainianconflict-top-comments**	Where to store the result **Local Storage or any distributed File System**	Tool **Google Colab**	Time (when to measure) **During each time frame, the count of the compliant data is calculated.**

Collection procedure (how to collect the data) **By using the Python code in Google collab online platform to divide the entire dataset and allocate it to different time frames.**	Notes or comments: **Used to calculate Credibility that further is used to calculate Validity.**

Base measure: **Rec_Trace-Provides the total number of records that are traceable in MDS**

#	Measure (what: entity, attribute) **Rec_Trace-Provides the total number of records that are traceable in MDS** **Entity: Data set** **Attribute: size**	Scale type **Absolute scale**	Applicability **Checks what all records can be traced, calculating traceability. Used to calculate vincularity.**

Who measures? **Developer/ Data Scientist**	Source of measurement [**https://www.kaggle.com/akuppps/ukrainianconflict-top-comments**](https://www.kaggle.com/akuppps/ukrainianconflict-top-comments)	Where to store the result **Local Storage or any distributed File System**	Tool **Google Colab**	Time (when to measure) **During each time frame, the count of the traceable data is calculated.**

Collection procedure (how to collect the data) **By using the Python code in Google collab online platform to divide the entire dataset and allocate it to different time frames.**	Notes or comments: **Used to calculate vincularity.Traceability is measured both from “Inherent” and “System dependent” point of view.**

Base measure: **Ldst - Total number of occurrences of data elements in dataset (DS)**

#	Measure (what: entity, attribute) **Ldst - Total number of occurrences of data elements in dataset (DS)** **Entity: Data set** **Attribute: size**	Scale type **Absolute scale**	Applicability It evaluates the frequency of data elements in the data set.

Who measures? **Developer/ Data Scientist**	Source of measurement **https://www.kaggle.com/akuppps/ukrainianconflict-top-comments**	Where to store the result **Local Storage or any distributed File System**	Tool **Google Colab**	Time (when to measure) **During each time frame, the count of the data elements is calculated.**

Collection procedure (how to collect the data) **By using the Python code in Google collab online platform to divide the entire dataset and allocate it to different time frames.**	Notes or comments: **Used to calculate vincularity. Traceability is measured both from “Inherent” and “System dependent” point of view.**

Base measure: Rec_no_null (MDS) -Frequency of records (MDS) with no null values				
#	Measure (what: entity, attribute) **Entity: Dataset** **Attribute: no of unique non-null values**	Scale type **Absolute Scale**	Applicability **It counts the no of field entries in dataset that are unique and non-redundant. All duplicate and null values are removed.**	
Who measures? **Developer/Data Scientist**	Source of measurement **"UkrainianConflict" Reddit Top Comments/Posts** [**https://www.kaggle.com/akuppps/ukrainianconflict-top-comments**](https://www.kaggle.com/akuppps/ukrainianconflict-top-comments)	Where to store the result **Local Storage or any File Distributed System**	Tool **Google Colab**	Time (when to measure) **The dataset divided into constituent time frames undergoes processing and with each iteration frequency of records with no null values is calculated**
Collection procedure (how to collect the data) **The pre-processing of data resulted in dividing the entire dataset and allocate to different time frames**	Notes or comments: **This measure is to calculate completeness, which further acts as an indicator to base measure veracity**			

Base measure: Rec_acc_age - Provides the total number of records with ages that fall within the acceptable range

#	Measure (what: entity, attribute) **Entity: Dataset** **Attribute: Coverage level (Fenton’s List)**	Scale type **Absolute Scale**	Applicability **It counts the no of records in dataset that are within the acceptable range (lower quartile-upper quartile) of BoxPlot**

Who measures? **Developer/Data Scientist**	Source of measurement **"UkrainianConflict" Reddit Top Comments/Posts** **https://www.kaggle.com/akuppps/ukrainianconflict-top-comments**	Where to store the result **Local Storage or any File Distributed System**	Tool **Google Colab**	Time (when to measure) **Should be calculated after specific context of use has been identified and fixed. The upper and lower limits should not be dynamic and rather be static**

Collection procedure (how to collect the data) **By using coefficient of determination at box plot, quartiles can be calculated**	Notes or comments: **This measure is to calculate correctness, which further acts as an indicator to base measure veracity**

<table>
<tr><td colspan="6"><b>Base measure: Pj - Provides the total number of duplicate items and their specific count in each dataset</b></td></tr>
<tr><td>#</td><td colspan="2">Measure (what: entity, attribute)
<b>Entity: Dataset</b>
<b>Attribute: no of unique elements</b></td><td>Scale type
<b>Absolute Scale</b></td><td colspan="2">Applicability
<b>It counts the no of records in dataset that are redundant and returns their count from each dataset</b></td></tr>
<tr><td colspan="2">Who measures?
<b>Developer/Data Scientist</b></td><td>Source of measurement
<b>"UkrainianConflict" Reddit Top Comments/Posts</b>
<b>https://www.kaggle.com/akuppps/ukrainianconflict-top-comments</b></td><td>Where to store the result
<b>Local Storage or any File Distributed System</b></td><td>Tool
<b>Google Colab</b></td><td>Time (when to measure)
<b>Should be calculated after specific context of use has been identified and fixed. The upper and lower limits should not be dynamic and rather be static</b></td></tr>
<tr><td colspan="3">Collection procedure (how to collect the data)
<b>By using coefficient of determination at box plot, quartiles can be calculated</b></td><td colspan="3">Notes or comments:
<b>This measure is to calculate correctness, which further acts as an indicator to base measure veracity</b></td></tr>
</table>

Base measure: N_succ_req - Number of successful requests				
#	Measure (what: entity, attribute) **Entity: Dataset** **Attribute: control – flow structuredness**	Scale type **Absolute Scale**	Applicability **It counts the no of successfull request given through an API server**	
Who measures? **Developer/Data Scientist**	Source of measurement **"Ukrainian Conflict" Reddit Top Comments/Posts** **https://www.kaggle.com/akuppps/ukrainianconflict-top-comments**	Where to store the result **Local Storage or any File Distributed System**	Tool **Google Colab**	Time (when to measure) **It is measured when the data attributes are retrieved and authorised enough with respect to specific context of use**
Collection procedure (how to collect the data) **Counting the number of requests from API, server or datastore**		Notes or comments: **This measure is to calculate availability which acts as an base measure indicator for Veracity.** $Availability\ (MDS) = \frac{[n\_succ\_req\ (MDS)]}{n\_req(MDS)}$		

<table>
<tr><td colspan="7"><b>Base measure</b>: <b>N_req - Number of Requests</b></td></tr>
<tr><td colspan="2">Measure (what: entity, attribute)
<b>Entity: Dataset</b>
<b>Attribute: control – flow structuredness</b></td><td colspan="2">Scale type
<b>Absolute Scale</b></td><td colspan="3">Applicability
<b>It counts the total no. of requests made and without categorising them as successful or unsuccessful. The wholesome count of requests is returned at end.</b></td></tr>
<tr><td>Who measures?
<b>Developer/Data Scientist</b></td><td colspan="2">Source of measurement
<b>"Ukrainian Conflict" Reddit Top Comments/Posts</b>
<b>https://www.kaggle.com/akuppps/ukrainianconflict-top-comments</b></td><td colspan="2">Where to store the result
<b>Local Storage or any File Distributed System</b></td><td>Tool
<b>Google Colab</b></td><td>Time (when to measure)
<b>The dataset divided into constituent time frames undergoes processing and with each iteration frequency of records with no null values is calculated</b></td></tr>
<tr><td colspan="3">Collection procedure (how to collect the data)
<b>The dataset retrieved should be accessible to authorized users using the Google colab , an online platform to separate dataset into different time frames.</b></td><td colspan="4">Notes or comments:
<b>This measure is to calculate a</b>vailability, <b>which further acts as an indicator to base measure veracity.</b>
$Availability\ (MDS) = \frac{[n\_succ\_req\ (MDS)]}{n\_req(MDS)}$</td></tr>
</table>

Bibliography:

[1] Ormandjieva, Olga et al. “Measuring the 3V's of Big Data: A Rigorous Approach.” IWSM-Mensura (2020). [

2] Lecture 11 Notes for performing Step 3 of Project.

[3] Dave Bharadvaj, “Measurement Framework for Assessing Quality of Big Data (Mega) in Big Data Pipeline