HAL
open science

Report on Test and Final Development of the Cooperation Analytics

Jessica Pidoux, Dominique Boullier, Natalya Avanesova, Thibaut Soubrié

To cite this version:

Jessica Pidoux, Dominique Boullier, Natalya Avanesova, Thibaut Soubrié. Report on Test and Final Development of the Cooperation Analytics. EHESS. 2022. hal-04102124

HAL Id: hal-04102124

https://sciencespo.hal.science/hal-04102124v1

Submitted on 22 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

COESO

connecting research and society

COLLABORATIVE ENGAGEMENT ON SOCIETAL ISSUES

WP5 - Cooperation Quality Assessment

Report on Test and Final Development of the Cooperation Analytics

30.06.2022

This deliverable is licenced under
a Creative Commons Attribution 4.0 International Licence.

COESO has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020)
SwafS-27-2020 - Hands-on citizen science and frugal innovation, under Grant Agreement No.101006325
The content of this publication is the sole responsibility of the COESO consortium and can in no way be taken to reflect the views of the European Commission. The European Commission is not responsible for any use that may be made of the information it contains.

Deliverable 5.2

Report on Test and Final Development of the Cooperation Analytics

Grant Agreement number	:	101006325
Project acronym	:	COESO
Project title	:	Collaborative Engagement on Societal Issues
Funding Scheme	:	H2020-EU.5. - SCIENCE WITH AND FOR SOCIETY
Topic	:	SwafS-27-2020 - Hands-on citizen science and frugal innovation
Project's coordinator Organization	:	EHESS-OpenEdition
E-mail address	:	pierre.mounier@openedition.org, alessia.smaniotto@openedition.org
Website	:	https://coeso.hypotheses.org

WP and tasks contributing	:	WP5 - Cooperation quality assessment
WP leader	:	FNSP-CEE
Task leader	:	FNSP-CEE

Dissemination level	:	PU
Due date	:	30.06.2022
Delivery date	:	30.06.2022

Authors	:	Jessica Pidoux (FNSP-CEE), Dominique Boullier (FNESP-CEE), Natalya Avanesova (Preste), Thibaut Soubrié (Preste)
Reviewers	:	Sy Holsinger (OPERAS), Luca De Santis (Net7)

Contents

Executive Summary 4

Citizen Science Practices' Analysis 5

Interview Study 5

Questionnaire Study 7

Lessons Learned from the First Round of Pilots 10

Indicator Construction and Workflow Design **11**

Workflow design 11

Data Types and Formats 12

Data Structure Study 13

Preprocessing 13

Unstructured data 14

Structured data 15

Postprocessing 15

Extraction and Analysis Methods and Tools 16

Unstructured data 16

Structured data 20

Feasibility Test Results 21

Feasibility Analysis 21

List of 10 Indicators Retained 21

Recommendations for integrating Cooperation Analytics into VERA 24

Conclusion 26

Annex 27

Pilots' Interview Guide 27

Executive Summary

The Deliverable D5.2 WP5 - Report on Test and Final Development of the Cooperation Analytics is the second deliverable of Work Package 5 within the project COESO - Collaborative Engagement on Societal Issues. D5.2 focuses on the technical operationalization of the cooperation analytics developed within COESO deliverable D5.1, which presented the conceptual framework of cooperation indicators for citizen science projects in the social sciences and humanities.

Deliverable D5.2 relates to two main tasks within COESO's grant agreement: Task 5.1 "Development and implementation of the Cooperation Analytics", planned from March 2021 (M3) until March 2022 (M15), and Task 5.2 "Observation of Cooperation Analytics' usages during open call pilots", planned from March 2021 (M15) until January 2023 (M25). The current reporting period of this deliverable goes until June 2022 (M18).

The specific aims of D5.2 are the following:

1. To understand the diversity of citizen science projects for adapting the conceptual framework to the projects' practices
2. To review the cooperation indicators by confronting them with technical constraints so they are calculable within the Virtual Ecosystem for Research Activation (VERA), the platform under development within COESO.

The report first presents the analysis of the first round of five pilot projects and the lessons learned from those projects that should be considered for the second round of five pilots starting in July 2022, and more largely within COESO's development. Secondly, the report describes the tests that were conducted on the cooperation indicators and concludes with the feasibility test results. These tests are based on sample datasets collected to provide a first list of indicators implementable within VERA as cooperation analytics. D5.2 is organised into three sections.

Section 1 Citizen Science Practices' Analysis provides the data collected from the first round of five pilot projects to understand their practices, including the methods for collecting the data. Here, one can find the lessons learned from studying citizen science practices as they are understood from COESO's selected pilot projects. This section responds to the first deliverable aim (To understand the diversity of citizen science projects for adapting the conceptual framework to the projects' practices) and is useful to know the way cooperation analytics are suitable for pilot projects' practices in citizen science.

Section 2 Indicator Construction and Workflow Design is centred on the process of building the indicators in a computable way and the workflow design used for developing the cooperation analytics. This process includes first, a data structure study for processing the data so the data is useful for automated calculations. Second, the data extraction methods that can be used later within VERA. Third, the computing methods for data analysis used and the resulting tools developed. The tools are: tool for extracting, filtering, and cleaning GMAIL archived data or similar messaging platforms; scraping script for any multi-languages web source, tool for extraction indicators' data from GMAIL metadata, as well as from GMAIL message content; script for extracting indicators' data from Hypotheses.org blog and Log Book content; script for work with

Zoom metadata.

Section 3 Feasibility Test Results finally presents the feasibility study outputs of the cooperation analytics conceived. It provides a list of indicators actually computable for VERA and the second round of pilots' learning process. The indicators are classified into Cooperation features, with a taxonomy described in D5.1. This section concludes with improvements and recommendations for ensuring that the cooperation analytics are operational within VERA. This section responds to the second deliverable aim (To review the cooperation indicators by confronting them with technical constraints) and is useful as protocols, first sample codes, and documentation for technical purposes: the cooperation analytics' integration within VERA.

I. Citizen Science Practices' Analysis

In this section, we present the methods used for the study of five pilots' citizen science projects and their practices. The study was based on exploratory interviews and a questionnaire as explained in the following subsections.

Our study followed a data management plan. The ethical and legal aspects are in alignment with Sciences Po's and COESO project privacy requirements.

Interview Study

We conducted an exploratory study about the five Pilots projects for developing the conceptual framework of the cooperation analytics (see D5.1). The aim was to have an overview of Pilots' practices so the conception of the plurality of cooperation styles matches their reality to some extent. More specifically, we collected information about i) the specific goals and outputs of the project, ii) the network where the project is embedded in (including partners), (iii) leader's profile and role in the project, and (iv) the roles of the other team members.

The data for the analysis was collected from different sources: the Pilots' original plan as stated in the COESO project, the adaptations made in the course of the project development, and the actual practices of the main stakeholders in each Pilot that were interviewed by WP5 during September 2021, or by analysing the interviews conducted by Net7 from WP3. WP5 interviews were conducted online given the COVID-19's health protection measures at that time.

The methods consisted of reviewing existing documentation about the pilots, conducting exploratory interviews (*n*=4), and analysing the interviews conducted by WP3 (*n*=4). In total, there were 12 interviewees, of which were seven engaged stakeholders and five professional researchers. There was one group interview with five persons, and seven individual interviews (Table 1).

Table 1. Exploratory Interviews conducted with the first round of Pilots

Interview No.	Pilot No.	Participants	By
1	Pilot 1	1 professional researcher	WP5
2	Pilot 1	1 engaged stakeholder	WP5

3	Pilot 2	Team: 2 professional researchers and 3 engaged stakeholders	WP5
4	Pilot 3	1 engaged stakeholder	WP5
5	Pilot 4	1 professional researcher	Net7
6	Pilot 4	1 engaged stakeholder	Net7
7	Pilot 4	1 engaged stakeholder	Net7
8	Pilot 5	1 professional researcher	Net7

From this exploratory study, we draw five main observations about the pilots' state of progress that are relevant for the development of cooperation analytics.

Observation 1: the variety and the number of objectives set. Pilots 1 and 3 have a restricted number of objectives. On the contrary, Pilots 2, 4 and 5 have a large number of objectives to achieve that can vary from one member of the pilot to another one and in some cases they are not clearly defined for measurement. This is related to the type of project, for instance an art performance, which makes the cooperation process more complex to assess than the other Pilots.

Observation 2: the quantity and type of stakeholders involved in each project. While the members of Pilot 1 are in the same locality, Pilot 5 is formed at the international level by a large quantity and different types of stakeholders (e.g., university, archives, migrants' network): each stakeholder with different qualities to measure. While it is clear enough what can be assessed for the scientific part of the project, it is difficult to define what can be assessed for the engaged stakeholders (or citizens) of the project; in part because these engaged stakeholders do not necessarily have a spokesperson and are harder to reach for interviewing than the scientists. Research and "civic" objectives are not always clearly defined in a common way, they can be operationally distinctive like in Pilot 5 (The engaged stakeholders are transcribers but their implication can increase throughout time) or more challenging like in the case of Pilot 2 where a dancer and a philosopher, with opposite practices, are now working together and involved in the deployment of dance annotation features for a software called MemoRekall with an additional stakeholder that is more technical oriented.

Observation 3: the research process pipeline is not defined yet in some cases (i.e. Pilots 1, 2) and can be set as an open exploration with the public. Moreover, the actual phase of the project varies from one Pilot to another one, while some are advanced in producing results, others are at the very beginning of planning their research. This plurality is however an opportunity to design cooperation analytics that are capable of assessing projects throughout the whole chain of progress: from its initial work coordination to the dissemination of results.

Observation 4: the variety of working and communication tools. Concerning the working tools, some Pilots that include more data science and engineering work (Pilots 4 and 5) involve a high number of tools and databases used in parallel. Their operations are not centralised, they are multiplatform. Others have not yet found the best digital working tools adapted to their projects or are mainly interacting offline (Pilots 1 and 2). This observation constitutes another opportunity

for VERA to become the reference platform adopted for communicating and working together to some extent according to the operational tasks of each project.

Concerning the communication tools that also serve dissemination of results and collaboration practices, well-known platforms are recurrent across projects: social networks and Google docs. Slack is not favoured by any project as teams have tried it but without giving it much utility. The main channels remain mailing service (institutional, outlook or Gmail), videoconference tools like Zoom for those working mainly remotely (Pilot 1), and telephone calls for those which are mainly working face-to-face (Pilot 2). It is worth mentioning that some contradictory results can be found between interviews and members of the same pilot. The contradiction should be taken as a multiplicity of collaborations forms that are possible within the same pilot and the lack of formalism in daily-work practices.

Observation 5: Pilots present different cooperation challenges that have to be considered in order to be able to offer well established measures of cooperation. This way, the cooperation analytics are limited within a scope of what can, and cannot, be measured given the clarity of the activities of each project. It is important to note that Pilots 4 and 5 have a clearer view on what they expect from VERA, while other Pilots have no expectations and remain receptive to what can help them better cooperate.

Questionnaire Study

In February 2022, we created an online questionnaire for understanding better pilots' practices directly related to data processing. The goal was to know in detail which datasets from Pilots were possible to collect for building the cooperation analytics. The results of this questionnaire were key for initiating the indicator construction presented in section 2. The descriptive information about four[1] Pilot projects are presented in table 2.

Table 2. Four Pilots' Data Questionnaire and Results

	Pilot 1- Lisbon Tourism Observatory	**Pilot 2 Dancing Philosophy**	**PIlot 3- Social Evolution & Mafia**	**Pilot 4 - Investigative Reporting**
How many researchers are members of the project?	3	3	1	3
Among these researchers how many contribute (any kind) at least once a week?	2	3	1	3
How many engaged stakeholders (or citizens) are members of the project?	6	3	1	2
Among these engaged stakeholders (or citizens) how many contribute (any kind) at least once a week?	1	3	1	2

[1] We were not able to reach Pilot 5

In which languages do you communicate with others and publish your work?	Portuguese	English/French/ Italian	French/Italian	English/Italian
Do you use translation platforms? Please name them all, you can select multiple choices	no	deepl/google translate/linguee	no	google translate/ context reverso
Which platforms do engaged stakeholders (or citizens) and researchers use to chat and talk? Please name them all, you can select multiple choices	phone calls/whatsapp	phone calls/whatsapp	phone calls/whatsapp	phone calls/whatsapp
Which platforms do engaged stakeholders (or citizens) and researchers use to email? Please name them all, you can select multiple choices	outlook manager app/gmail/yahoo	gmail/yahoo/institutional mail service/mac mail manager/mail from rennes2	gmail	gmail/institution mail service
Which platforms do engaged stakeholders (or citizens) and researchers use for meetings? Please name them all, you can select multiple choices	zoom/offline meetings	offline meetings/ framatalk/zoom	zoom/offline meetings	offline meetings/ Teams
If you record your online and offline meetings, please tell us which tools do you use	dictaphone/photos/ zoom	photos/video camera/zoom	dictaphone	No recording
Which platforms do engaged stakeholders (or citizens) and researchers use for scheduling meetings and work in general? Please name them all, you can select multiple choices	none	google calendar/doodle	google calendar	google calendar/ outlook calendar
Do you take meeting minutes in a formal way ?	no	no	no	no
Do you take meeting minutes and informal notes in any of the options presented below?	google docs/handwritten notes in paper/evernote	google docs/handwritten notes in paper/evernote	google docs/handwritten notes in paper	no
Do engaged stakeholders (or citizens) and researchers use task management tools? If yes, please specify	no	github	no	github
How do engaged stakeholders (or citizens) and researchers share documents and data?	email/google drive	email/google drive/own cloud / institutional cloud/shared docs	email/google drive/ institutional cloud	email/own cloud

		(HUMAN-NUM)		
Where are the results and activities of your citizen science project published?	institutional repositories/hypotheses.org blog/zenodo	twitter/youtube/facebook/HAL/Hypotheses.org blog/personal or institutional website	twitter/facebook/ Hypotheses.org blog	research gate/academia.edu/Hypotheses.org blog/personal or institutional website
What type of image formats have you produced so far in your citizen science project?	JPEG	JPEG/RAW	JPEG/RAW	PNG/JPEG
What type of video formats have you produced so far in your citizen science project?	mp4	mp4/mov	AAC	none
What type of audio formats have you produced so far in your citizen science project?	wav	none	wav	none
What type of text formats have you produced so far in your citizen science project?	word/powerpoint	word/powerpoint/ excel/LibreOffice Writer - Open Office	word/pdf/excel/LibreOffice Writer - Open Office	word/pdf/excel /txt
What type of data formats have you produced so far in your citizen science project?	none	xml	none	csv
Where do engaged stakeholders (or citizens) and researchers store the data and files produced?	google drive/emails/ local storage phone/local storage computer/external hard drive	shared docs HUMAN-NUM/google cloud/google drive/emails/ local storage phone/local storage computer/external hard drive/other	google drive/emails/local storage phone/local storage computer/external hard drive	external hard drive/own cloud
Please tell us if engaged stakeholders (or citizens) and researchers use other data storage platform, name all the platforms below	COESO google drive	no	Basecamp, wetransfer (share), USB key	no
How much data and files engaged stakeholders (or citizens) and researchers have stored? Please give an exact estimation	1-2GB	20-30GB	20GB	1GB
Do engaged stakeholders (or citizens) and researchers handle personal and sensitive data (i.e., name, sexual	yes	yes	yes	yes

orientation, location or any information that identifies a person) ?				
If you selected other platforms for publishing results and activities please list them below	CRIA website, CRIA facebook	-	-	-
Please provide citizen and researcher username accounts, orcid, URLs, or key authors to find you on the platforms where you publish your results	CRIA website, https://cria.org.pt/pt, CRIA facebook, https://www.facebook.com/cria.centroemredeinvestigacaoantropologia/	ttps://orcid.org/, 0000-0002-8126-8249, https://www.cadmiumcompagnie.com/, https://www.facebook.com/, TWITTER: @LupettiLupino, https://dansophie.hypotheses.org/, http://memorekall.com/home.php	twitter, @dorcadie, facebook, Mathilde Dorcadie, twitter, @fabrizzoli, facebook, fabrice rizzoli, Hypotheses.org, https://usbc.hypotheses.org/	IRPI, https://irpimedia.irpi.eu/
If you, or your institution, have your own platform developed internally for working with engaged stakeholders (or citizens) and researchers please provide the URL or name	Instituto Universitário de Lisboa, https://repositorio.iscte-iul.pt/	MemoRekall web, no API, files xml available	-	-

Lessons Learned from the First Round of Pilots

What we learned from the existing pilots is that they make extensive use of multiple platforms and tools for communication, storage, and reporting. Consequently, this situation generates a challenge for collecting their data in order to compute the cooperation analytics. However, we found strategies to recover some data from platforms like Gmail by using Google Takeout based on every person's data access rights. But privacy concerns were also encountered. The main problem for developing the cooperation analytics is the low level of formalisation of the pilots dynamics, and of the research as well: small teams, informal meetings, personal relationships, fuzzy methodologies, limited accountability.

When doing citizen science, we are aware that the process of knowledge production becomes complex. Engaged stakeholders (or citizens) and professional scholars have to cooperate to deliver new insights that traditional scientific practices would have missed or disqualified. They have to conciliate different standpoints on a specific issue, as well as their multiple protocols to validate knowledge. That is why all stakeholders of a pilot project must take into account the quality of the cooperation process. Our definition of quality does not refer to criteria defined in a self-evaluation survey at the end of the project. This will not help stakeholders to gain a better

understanding of their research process. The COESO project advocates precise and useful feedback for pilots within the VERA platform as to learn about their on-going cooperation practices; that is to observe citizen science cooperation "in the making".

For developing the cooperation analytics, we require collecting reliable and sufficient amounts of data from the second round of pilots. Therefore, we will ask them to adopt some platforms and protocols that will help us to provide them useful insights about their project. Then comes the issue of acceptability (not limited to Cooperation analytics). We should be able to give incentives like offering templates, comparisons, resources developed only on this platform including the cooperation analytics benefits for the projects. Moreover, we should provide recommendations about the tools and protocols to be used when messaging, storing data, and taking minutes for every meeting, including documentation and tutorials that should be available immediately after the selection of the new pilots.

These recommendations will be discussed on a one-to-one basis with each pilot to understand their constraints, and if necessary to set up a guidance process or to find an arrangement with the existing uses. The data processing for computing the cooperation analytics follows our Data Management Plan, including an information sheet.

But still, there will be some reluctance to change routines, which is quite normal and should be handled with empathy. Through one-to-one discussions and coaching, we could obtain a limited number of very largely diffused apps and platforms to be selected by the pilots and set up a range of choices for the activities. For those apps and platforms, we could design a number of APIs as we started to do with Gmail so that we still obtain some traceability.

The last point would be how to implement these specific developments within the VERA platform or on another module that could be used for the data processing and just connected to the platform for the data visualisation. We will discuss these recommendations with respect to their technical implementation in VERA after the results of our feasibility study in section 3.

II. Indicator Construction and Workflow Design

In this section, we present the construction of cooperation indicators conceived, referring to D5.1, as well as the workflow design we followed for building them.

Workflow design

For the indicator's construction, we followed a workflow design consisting of three mains steps:

1. **Data collection** -> 2. **Extraction and Analysis** -> 3. **Feasibility Tests**

First, we collected different data types in the raw format as produced by the pilots on different platforms. Second, we extracted the data with different tools in order to process it and build a data structure for its analysis. Finally, the feasibility tests were conducted as after the data processing it was possible to define consistent data types to either reject or successfully build the cooperation indicators.

The workflow design was key for adapting and challenging the cooperation definitions we initially

established in D5.1 against technical constraints that are proper to computational calculations.

Data Types and Formats

The initial analysis of data produced by Pilot teams showed that we are dealing with rather heterogeneous types, defining which terms we use such as structured and unstructured data. According to the principles of natural language processing, any textual data that is produced in the result of communication and cooperation between people is considered an exclusively unstructured data type, since it lacks consistent formatting and requires a specific approach to its analysis and processing. Based on that we define as unstructured category data derived from messaging content, project notes, meeting notes, and blog content. In the structured category, we include information that can be easily extracted directly from metadata of the used platforms (calendars, mails, task management, and meeting tools), as this information is well-organised and structured under the tags and columns names.

Data sources analysis also showed that we need to introduce one more additional technical category, which would be at the intersection of the two main ones, namely semi-structured data. This category includes information that is more or less structured. By this we mean that the document has a predefined form, such as final reports or project proposals, but the language itself is natural so it cannot be extracted and stored in table format using some tags or name of column for example like we do with purely structured data. Nevertheless, having the understanding of document structure and expected data (targeted to be extracted) we are able to build some matching linguistic rules and patterns with the aim to process this kind of data in a more standardised way. On a high-level, we consider this category as a sub-category of unstructured.

As a result, we have collected all tools/platforms used by Pilot teams during their cooperation and classified them into one of predefined data categories, based on the nature of data presented:

Table 3. Tools and platforms used by the first round of Pilots

Tools and platforms used by Pilots	**Processed Data**	**Data Category**
Profile Pages	yes	Structured
Google calendar	yes	Structured (metadata)
Whatsapp	no	Unstructured
Gmail	yes	Structured (metadata)
		Unstructured
Yahoo	no	Structured (metadata)
		Unstructured
Outlook	no	Structured (metadata)
		Unstructured
Zoom	yes	Structured (metadata)
		Unstructured

Teams	no	Structured (metadata)
Hypotheses.org blog	yes	Unstructured
photos (JPEG)/PDF (used for meetings minutes)	no	Semi-structured
dictaphone (used for meetings minutes)	no	Semi-structured
video camera (used for meetings minutes)	no	Semi-structured
google docs (used for meetings minutes in informal way)	no	Semi-structured
google docs (used for free notes)	no	Semi-structured
evernote (used for meetings minutes in informal way)	no	Unstructured
evernote (used for free notes)	no	Unstructured
powerpoint (used for free notes)	no	Unstructured
xml/csv	no	Semi-structured
github	no	Structured (metadata)
Doodle	no	Structured (metadata)
shared docs (HUMAN-NUM)	no	Unstructured
facebook	no	Semi-structured
CRIA website	no	Semi-structured
website	no	Semi-structured
twitter	no	Semi-structured
ORCID	no	Semi-structured
Workfield Log Book	yes	Unstructured

Obtained categorization helped us to group tools/platforms that have similar structural content and thus we were able to bulk them at the preprocessing stage. This is explained by the need to develop different and quite specific preprocessing pipelines to prepare structured/unstructured data for further work.

Data Structure Study

Preprocessing

Text preprocessing is traditionally an important step for natural language processing (NLP) tasks. It transforms textual data into a more digestible form so that automatic approaches can be implemented. Different data types require the application of different specific preprocessing steps. Thus, we have elaborated two preprocessing schemes for structured and unstructured data respectively.

Unstructured data

To illustrate the difference between the two preprocessing schemes, let's consider a Gmail platform used by one of Pilot teams as a messaging tool. As mentioned above in the classification table, Gmail as a data source has 2 data types useful for building indicators: structured data in the form of metadata and unstructured data in the form of message content. To be able to work with unstructured message content, first we need to clean it properly from noisy data and apply following steps:

- remove user's personal data
- remove tabulation symbols (>>>) and other noisy symbols (@/*)
- remove emoji (:)) and exclamation (Eh eh, he, hahahahah)
- remove different URLs (https://)
- convert other HTML entities to recognisable characters
- remove duplicate text, e.g., the original message in a mail reply
- remove extra spaces
- remove nonsense sentences (containing less than 2 words, not NOUN, VERB, ADJ)
- prepare a list of custom stop-words (persons names, expression of politeness, e.g., "Thank you for your interest and availability", etc. and greetings)
- remove punctuation (except period "." that is needed for sentence splitting)
- sentence splitting

Taking into consideration the fact that collaboration teams may include participants from different countries, they can use different languages for communication, so there is a need to normalise the content and to translate it to one single language, preferably English. This is explained by the fact that for some indicators ("Type of meeting" etc) we need to process all the data at the same time, so we need to have it in the English language, since the existing NLP tools have the highest accuracy rating for this particular language. It is important to emphasise that this last translation-related step is not mandatory for building all indicators (for "Language diversity degree" indicator we need to keep original languages), compared to other methods described above, such as removing duplicates, extra spaces, etc., which are mandatory steps for further work on indicators from unstructured data. Moreover, for example, to build an indicator like "the degree of language diversity", it is necessary to preserve the original languages of communication between project participants, as well as to extract information about the level of formality of communication, which is one of the components for adjusting the indicator "social balance" (see all indicators in D5.1).

Structured data

Processing of metadata requires less steps in comparison with the message content. Among the

most important techniques that should be applied are following:

- remove duplicates
- remove URLs (https://)
- convert other HTML entities to recognisable characters
- remove unnecessary spaces
- remove leading and trailing spaces
- remove different noisy abbreviations (fw:|fwd:|Re:Fw:|Re:|Fwd:|Fw:|RE: FW:|FW:|RE)
- convert to lowercase and lemma for extracting info needed for some of indicators

While the above described techniques are basic preprocessing schemes, we applied additional parameters when necessary based on five factors. First, additional modifications are affected by the specifics of the tools themselves: not many changes in the pipeline are required to process metadata from Zoom or for a calendar, whereas there is a need to add a date conversion step for Gmail as the format differs. Second, the step of removing forwarded messages was not necessary for tools other than messaging platforms like emails (e.g. Slack, WhatsApp). Third, a punctuation check step is added to process content from the Field-Log Book, since the notes are less organised than in messaging platforms and the usual standardised approach to breaking text into sentences does not cope. Fourth, the Log Book (Carnet) required the elaboration of a very specific preprocessing plan: not only the use of non usual sentence splitting based on a new line, but also grammar correction and spelling mistakes checking. Finally, another factor that affects the change in the basic preprocessing scheme lies in the indicators themselves, which are subsequently built from the data we expect to extract when VERA is released. As an example, there may be the addition of the stage of translating source text into a single language.

Postprocessing

Postprocessing steps cannot be formulated in such a universal way as we did for preprocessing schemes, as it totally depends on the specifics of indicators that we aim to build. As an example: to obtain the Attendance rate indicator that belongs to the organisation of citizen/research participation Cooperation feature, we need to calculate it as the real participant's percentage value out of the total number of invited persons. The number of invited participants can be retrieved from the meeting invitation emails (Gmail metadata), number of the real participants - from the scheduler tool (Google Calendar) in form of number values, so the percentage calculation is applied here as a postprocessing step. In comparison, to build indicator “Type of meeting” belonging to the same Cooperation feature, it is necessary to take the list of extracted meetings (Gmail or Google calendar) and classify each of them according to the following conditions: if the information in the scheduler tool for a meeting contains keywords such as weekly, daily, monthly, etc., we categorise such a meeting as regular. If there are no such indicators of regularity, we define them as scheduled ones. For the scheduled meetings, based on the calculated difference in minutes between the time of sending a meeting invitation and the time of meeting itself, we label it as spontaneous if the time difference is less than 720 min (12 hrs), otherwise it is left as scheduled. To summarise the above examples, the postprocessing required for each of the indicators is quite different.

Extraction and Analysis Methods and Tools

Extraction of the necessary data types is further complicated by the fact that each data source has its own technical features, and also by the fact that it is not possible to build a universal pipeline for extracting all structured or all unstructured data since everything depends on what kind of information we are interested in and from what is exactly the source. For example, there are data tools that cannot be accessed easily, that's why before implementing particular NLP approaches we need to deal with the issue of getting all available information from a data source, filtering it, or converting to the specific format in order to be able to process it and extract data type/s needed for the construction of the indicators.

Unstructured data

Let's consider the difference in data extraction comparing some data sources for unstructured information:

1. In the context of Gmail, the mail backup can be exported only in the MBOX format and there is no direct way to import it in any other easily processed format, like CSV for example. Taking this into account, we have written a script for converting the mbox archive to CSV format with synchronous extraction of only useful information from there, such as the Date/Subject/From/To and the body of the email itself. In addition, considering the need to clean and anonymize data, an additional functionality was added to the mentioned script that allows data filtering by parameters such as Date/Subject/From/To, and it also helps to create a specific user email dataset for further NLP processing. The script is provided below:

```
import mailbox
import csv

def getcharsets(msg):
    charsets = set({})
    for c in msg.get_charsets():
        if c is not None:
            charsets.update([c])
    return charsets

def handleerror(errmsg, emailmsg,cs):
    print()
    print(errmsg)
    print("This error occurred while decoding with ",cs," charset.")
    print("These charsets were found in the one
email.",getcharsets(emailmsg))
    print("This is the subject:",emailmsg['subject'])
    print("This is the sender:",emailmsg['From'])
```

```
def getbodyfromemail(msg):
    body = None
    #Walk through the parts of the email to find the text body.
    if msg.is_multipart():
        for part in msg.walk():
 # If part is multipart, walk through the subparts.
            if part.is_multipart():

                for subpart in part.walk():
                    if subpart.get_content_type() == 'text/plain':
                        # Get the subpart payload (i.e the message body)
                        body = subpart.get_payload(decode=True)
                        #charset = subpart.get_charset()

            # Part isn't multipart so get the email body
            elif part.get_content_type() == 'text/plain':
                body = part.get_payload(decode=True)
                #charset = part.get_charset()

    # If this isn't a multi-part message then get the payload (i.e the message
body)
    elif msg.get_content_type() == 'text/plain':
        body = msg.get_payload(decode=True)

   # No checking done to match the charset with the correct part.
    for charset in getcharsets(msg):
        try:
            body = body.decode(charset)
        except UnicodeDecodeError:
            return ''
        except AttributeError:
            return ''
    return body

def filter_data_include(message, filters):
    for filter in filters:
        for (k, f) in filter.items():
            if k not in message or f not in message[k]:
                return False
    return True

def filter_data_exclude(message, filters):
    for filter in filters:
        for (k, f) in filter.items():
            if k in message and f in message[k]:
```

```
                return False
    return True

def write_csv(filters_include=None, filters_exclude=None):
    writer = csv.writer(open("cleanedmails.csv", 'w', newline='',
encoding='utf-8'))
    for message in mailbox.mbox("inpit your mbox file here"):
        if filters_include and not filter_data_include(message, filters_include):
            continue
        if filters_exclude and not filter_data_exclude(message, filters_exclude):
            continue

        writer.writerow([
                message['Date'],
                message['Subject'],
                message['From'],
                message['To'],
                getbodyfromemail(message)] )

###Fields for patterns are below###

filter_include = [{"From": "Jessica"}]
filter_exclude = []

write_csv(filter_include, filter_exclude)
```

2. In the Hypotheses.org blog, the scraping script was deployed to obtain articles from the web pages where pilots published. Considering the fact that articles in blogs are written in different languages, we chose the Newspaper3k python package that supports seamless language extraction and detection and is also suitable for processing multiple URLs at once. Important notice: the mentioned above scraping script can be used to scrap any other multi-language web source. The scraping script is presented below:

```
!pip install newspaper3k
import newspaper
import nltk
nltk.download('punkt')
from newspaper import Article
#Define article's URL from hypoteses.blog
url = "https://civtur.hypotheses.org/151"
#For different language newspaper refer above table
coeso_article = Article(url, language="pt") # en for English
```

```
#Download the article
coeso_article.download()
#Parse the article
coeso_article.parse()
#Perform nlp
coeso_article.nlp()
#Extract title
print("Article's Title:")
print(coeso_article.title)
print("n")
#Extract text
print("Article's Text:")
print(coeso_article.text)
print("n")
```

After we have gained access to all information from the required source, it is important to decide which methods will be the most efficient way to process them in order to subsequently obtain the data from which the indicator will be built in the future. When choosing a method, the decisive role is played by which category the data belongs to: is it structured metadata or unstructured text written in natural language.

Thereby, data from Gmail metadata was extracted using SpaCy open source library based on custom patterns and taxonomies. See below an extraction code using SpaCy and a taxonomy for meeting keywords variations and a similar example with the usage of time zone taxonomy.

```
#find mails about meetings
import spacy
from spacy.matcher import Matcher
nlp = spacy.load("en_core_web_sm")
matcher = PhraseMatcher(nlp.vocab, attr='LOWER')
variants = ["meeting", "reunion", "call",
           "zoom"]
patterns = [nlp.make_doc(text) for text in variants if text != "Reminder"]
matcher.add("Matching", None, *patterns)

def get_phrases(all):
 prob = []
 for s in all:
   doc = nlp(s)
   matches = matcher(doc)
   if matches:
     prob.append(s)
 return prob
matched_mails = get_phrases(unique_mails)
meetings = remove_duplicates(matched_mails)
print(meetings)
```

```
print(len(meetings))        ###number of meetings
```

```
#find data about meeting duration
import spacy
from spacy.matcher import PhraseMatcher
nlp = spacy.load('en_core_web_sm')
matcher = Matcher(nlp.vocab)
time = ["CEST", "EST", "EET", "GET", "CET"]
# the list containing the phrases to be matched
pattern = [[{"TEXT": {"REGEX": "([0-9][0-9]:[0-9][0-9])"}},{"IS_PUNCT": True},
{"TEXT": {"REGEX": "([0-9][0-9]:[0-9][0-9])"}}], [{"TEXT": {"REGEX":
"([0-9][0-9]:[0-9][0-9])"}},{"IS_ALPHA": True}, {"TEXT": {"REGEX":
"([0-9][0-9]:[0-9][0-9])"}}]]


matcher.add("Matching", None, *pattern)
doc = nlp(str(meetings))
matches = matcher(doc)
for match_id, start, end in matches:
   span = doc[start:end]
   #print(span.text)
   #print(match_id, string_id, start, end, span.text)
results =  [doc[start:end].text for match_id, start, end in matches]
print(results)
```

In contrast, data from Gmail message content was processed with the help of IBM Natural Language Understanding API.

As for the content of the blog, it is a semi-structured form of data that has to be processed using combined approaches: pattern-based with the help of SpaCy matcher library, fully automatic with IBM Natural Language Understanding API, and a statistical one using YAKE library. Also, during the analysis, we have discovered that only pattern-based approaches can be applied for such unstructured and free text as it's presented in the Log Book.

Structured data

There are of course data sources that don't require a specific approach to be applied to get access to its information, for example, Zoom or Google Calendar metadata that can be directly pulled from the platform in the needed CSV format. Having info in this format, metadata processing can be easily processed by table queries, as it's shown in the script below made for Zoom metadata processing:

```
# reading .csv dataset
```

```
import pandas as pd
data = pd.read_csv("/content/zoom_data.csv")
# previewing the first 5 lines of the loaded data
data.head()
# getting the values of Subject column
values = data['Sujet'].dropna().astype(str).tolist()
print(values)
# getting the values of Duration column
values = data['Durée (minutes)'].dropna().astype(int).tolist()
print(values)
# getting the values of Participants column
values = data['Participants'].dropna().astype(int).tolist()
print(values)
```

It is also worth emphasising that extraction from various tools and sources was manual for the Pilot team's data, such as downloading the Gmail archive using Google Takeout, as well as pulling metadata from Google Calendar and Zoom. In the future, the process can be automated. The data automatic extraction in the context of the VERA platform remains to be defined with the COESO consortium (see our recommendations in section 3): it is possible to create data exchanges between tools via an API, another form of integration, or extraction pipeline.

III. Feasibility Test Results

In this final section, we present the results of the feasibility tests conducted for building the indicators. We list the indicators retained with their corresponding definition as expressed in D5.1 and their technical description for their implementation within VERA. We finally present our improvements and recommendations for the operational development of cooperation analytics in the remaining working period.

Feasibility Analysis

List of 10 Indicators Retained

The construction of indicators is a two-stage process. This is explained by the fact that each indicator can consist of either one or several types of data, therefore, to obtain a final indicator, it is necessary to extract required data types from different sources and then build an indicator from the collection of the extracted data types. Based on that, the first 16 data types were successfully processed and obtained from data sources such as Gmail metadata, Zoom metadata, Google Calendar, Gmail message content, Worfield Log Book content, and Hypotheses.org content. From these 16 data types, we built the following 10 indicators in Table 4 that belong to different Cooperation Features as defined in D5.1, these are:

Table 4. Indicators for measuring cooperation features with their technical description

Cooperation Features	Indicators	Technical Description
Organisation of citizen/research participation. It defines the type and configuration of actors' participation for cooperating in the project.	1. Type of meeting	Each of the meetings' subjects separately for a particular period of time for a project with labelling to the specific type: regular/scheduled/spontaneous. If in the scheduler tool data for a meeting there are such indicator words as weekly, daily, monthly etc., categorise such a meeting as regular. If there are no such indicators of regularity - define them as scheduled ones. For the scheduled meetings, based on the calculated difference in minutes between the time of sending meeting invitation and the time of meeting itself, label a meeting as spontaneous if the time difference is less than 720 min (12 hrs), otherwise - leave it as scheduled.
	2. Attendance rate	Calculated as the real participants percentage value of the total number of invited ones. Number of invited participants retrieved from the meeting invitation emails, number of the real participants - from the scheduler tool. By default those who did not reject the invitation - real participants.
	3. Scale of meeting	Calculated as a number of the invited participants for each of the meetings separately.
	4. Medium of meeting	List of media used for meeting
	5. Duration of meeting	For each of the meeting retrieved from the scheduler tool data as a difference between the start and end time of the meeting. Calculated in minutes.
	6. Frequency of meeting	For each of the 3 categories of meetings calculate the percentage value from the total number of meetings.
Idiom management. It enables us to detect the idiomatic tension and flexibility of actors	7. Language diversity degree	Calculated as a percentage value of messages in which some language was detected of the

considered when communicating with others in different media.		total number of messages in the source content for a period of time. If the percentage value is less than 2%, we consider it as non-accurate language detection and do not include it in the result representation.
Knowledge diversity processing. It enables us to detect the idiomatic tension and flexibility of actors considered within knowledge production processes when cooperating for producing a result (e.g. writing a report, an article).	8. Knowledge convergence degree	Calculated as a fraction of a unit for each of the disciplines detected in the source content, where 1 is a perfect match. The values of match less than 0.2 were not included.
Degree of participation is an adaptation of the initial feature conceived "Distribution of roles in scientific/citizen participation" in D5.1. Degree of participation refers to the extent the role or status of actors in the project configures the direction of contributions distributed among the parties involved.	9. Degree of asymmetry	The score is calculated based on the data from its components. Symmetry is defined when the value is equal to 0
Governance principles. It is mainly based on Boltanski and Thévenot's (2006) reference "The orders of worth" for describing the principles that guide the cooperation practices. Other authors, like Etienne Wenger, have highlighted the relevance of principles in organising communities of practice.	10. Social world balance	Balance is provided based on the values about the language formality + orders of worth. As for formality: If the difference between the previous and the last (starting from month 3) values degree of language formality is less than 0%, it means the decrease of formality degree in the used languages. (e.g. 80% - 93% = - 13 %). If the difference is more than 0%, it means the increase of formality degree in the languages used (e.g. 95% - 93% = 2 %). Zero value means no difference in formality level detected. As for orders of worth: Consider setting threshold of 5%: if the change is less than 5% - do not provide the result of change for the category.

We successfully implemented the above list of indicators by using the data samples from the pilots and we obtained confirmation of the feasibility of their computability. However, there are data types necessary to build some indicators that could not be fully checked due to several reasons (now they are on hold or in progress status). The main reason is the lack of well-organised and formalised data. As already mentioned before, we are dealing with both structured and unstructured data. At the same time, the amount of unstructured data

significantly prevails, which complicates the process of their processing. For some cases, we have tried to artificially translate messy data into a semi-structured form and apply patterned approaches that work at the syntactic and lexical level. This means that we had to write conditional rules, based on language patterns of speech construction, in order to extract, for example, such data as the mention of the tools used in free, unstructured text, or the extraction of specific keywords or phrases. But this approach also has its drawbacks, since without reference to the semantic context we cannot guarantee the accuracy of the extraction, which can provoke the appearance of false positive results. Another problem, which in principle correlates with what is written above, is the use of different languages in communication, which leads to the need to artificially translate all information to one language. Therefore, on the one hand, we simplify the data processing in order to build indicators, while on the other hand, we may lose some context, since we cannot exclude the possibility of data distortion during automatic translation.

Summing up all of the above, it is possible to simplify the processing and, accordingly, extraction of required data by adding formalism to communication between participants, as well as by stimulating the more-intensive use of such tools as a task scheduler, calendar, and others which store information in form of metadata.

Recommendations for integrating Cooperation Analytics into VERA

The recommendation of WP5 to the COESO consortium is to create explicit workflows to record the pilots' ongoing collaborations. Protocols have to be adopted by the second round of pilots so their activity can be used for the cooperation analytics, which provide new learnings for the pilots. We present these recommendations below according to different cooperative dynamics mediated by the technologies identified.

Meetings

- Online meetings should be recorded.
- Minutes for each meeting should be written.
- The minutes' meetings should have a formal structure using the COESO template provided.
- Use calendar invitations to organise meetings and add in C.C COESO-WP5 dedicated address provided.
- The word "Meeting" and key information should be identified in the invitation, for instance in the email/calendar subject line or in the email text, including topic, date and time.

Mails

- Mail subjects should include the word "COESO" when writing about pilots' activities. This keyword will allow us to quickly filter COESO-related emails and avoid considering other personal content in the data analysis.
- Write COESO-related mails from your personal address, adding in C.C a COESO-WP5 dedicated address created for analytic purposes.
- New tools for processing data coming from external platforms like an institutional mail service or Gmail can be integrated within VERA. We have these tools at our disposal as

presented in section 2. This requires alignment with VERA's general development scope and more broadly with COESO's values: although commercial and non open-source platforms are privileged by pilots like Gmail, Basecamp, Zoom, etc., VERA should encourage and limit the communication with open-source platforms for the project coherence with the general infrastructure.

Discussions

The messaging platform Mattermost should be a discussion space for pilots. If pilots are using other platforms, a minimum request is to test Mattermost.

Documents

The use of ShareDocs should be mandatory. If files are stored in clouds and local storage devices, a copy of those files should be kept in ShareDocs.

APIs

The use of Application Programme Interfaces (APIs) would allow VERA to communicate directly with the multiplicity of platforms used by pilots: Gmail, Zoom, and others. This implementation requires consequent technical work and further privacy protection measures. One API could for instance ask user permission to access gmail data, limited to COESO-related mails, of a user when registering to VERA. APIs could also help in analysing pilots' documents, so far mainly stored in Google drive or institutional clouds.

The required protocol would be creating a common ShareDocs (TGIR Huma-Num) space where all pilots' documents can be transferred for analytic purposes. The upload would require a communication between VERA and the platform used by a pilot like Google Drive for accessing the documents and transferring them into ShareDocs.

VERA's Profile Page

We reviewed WP3's profile page proposal and after discussing some improvements with Net7, we agree on its structure that is in accordance with the indicators we have designed. The classifications that are under development for Triple's project and VERA platform are relevant for our own purposes within WP5.

During the feasibility analysis, it was determined that the profile is a key data source that can and should be structured to obtain unambiguous data directly from the individual user or project leader. It means that we need to offer a list of options to select to keep structured the user's answers for textual and numerical values. Therefore, we present some propositions in table 5 on how to formulate questions, what to offer as a drop-down menu list, and possible sources for creating necessary taxonomies. However, for VERA release additional criteria have to be considered to define non-academic profiles. This is still under discussion given the fluidity of engaged stakeholders' definition and role in the scientific process.

Table 5. Declarative data and input values for VERA's profile page

Declarative Variables	Input Values
Select the disciplines that better define your profile	For interoperability purposes, the disciplines taxonomy is the same as used in TRIPLE: Archaeology and Prehistory/Architecture and Space Management/Art and Art History/Biological Anthropology/Classical Studies/Communication Sciences/Cultural Heritage and Museology/Demography/Economies and Finances/Education/Environmental Studies/Gender Studies/Geography/History/History, Philosophy and Sociology of Sciences/Law/Linguistics/Literature/Management/Methods and Statistics/Musicology and Performing Arts/Philosophy/Political Science/Psychology/Religions/Social Anthropology and Ethnology/Sociology
Select the languages you speak, indicate your proficiency level and your main working language	The list of official languages by country can be extracted from WikiData- open knowledge base: https://query.wikidata.org/#SELECT%20DISTINCT%20%3FcountryLabel%20%3FlanguageLabel%0A%7B%0A%20%20%3Fcountry%20wdt%3AP31%20wd%3AQ6256%20%3B%0A%20%20%20%20%20%20%20%20%20%20%20%20wdt%3AP37%20%3Flanguage%20.%0A%20%20SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22en%22%20%7D%0A%7D%0AORDER%20BY%20%3FcountryLabel%0A The Common European Framework for grading an individual's language proficiency is A - Basic User (A1/A2) B - Independent User (B1/B2) C - Proficient User (C1/C2). Source: https://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages Another useful list can be extracted from Linkedin: elementary proficiency, limited working proficiency, professional working proficiency, full professional proficiency, native or bilingual proficiency
Select the country where you work	The list of countries can be extracted from the International Standard for country and region codes https://www.iso.org/obp/ui/#search
Please indicate your gender (only collected for gender balance measurement)	male/female/non-binary/prefer not to respond

VERA's Organisational and Team Page

In addition, we determined that info blocks in profiles should be divided not only into user and project, but also into team ones. This way collective indicators could be built as a result of the aggregation of data received from individual users involved in the same project.

For example, the profile page could request a user to list all participants involved in the project, and indicate their roles. Other questions can be asked to the project leader in the block dedicated to the Team info such as recruitment mode (e.g., word-of-mouth, personal network, open call, open-administrative competition), contract duration or weekly effort required (this information will help measuring for instance the degree of "contractual formalism" within the project).

Moreover, the organisational/project profile page could request information about sources of funding and their types, the funded amount for every funding source, the types of data storage used for the project, if an access control policy is defined for the project's database, and other data-management related questions.

Conclusion

This report presented the Deliverable D5.2 of COESO's WP5 entitled "Report on Test and Final Development of the Cooperation Analytics". Its main goal was to show the state of completion of the cooperation indicators' construction for extreme citizen science in the social sciences and humanities, according to five pilot projects. These indicators will serve the cooperation analytics' integration into the VERA platform.

Sections 1 and 2 present the results of the pilots' practices study and the workflow design we followed to review the cooperation indicators. In section 2, the methods and tools for extracting and analysing pilots' data are detailed according to the different data types and structures studied.

The main two outputs can be consulted in section 3 where first, 10 cooperation indicators are constructed following the feasibility tests' results. These indicators are ready to be technically implemented. Secondly, in the same section, we provided recommendations to be considered by the COESO consortium to carry out the correct functioning of cooperation analytics in the future.

Annex

Pilots' Interview Guide

Interview Guide by WP5 for COESO PILOT PROJECTS

Septembre 2021

1. **The specific goals and outputs of the project**
   1.1. Could you tell me what the main goal of your project is?
   1.2. What are the outputs expected on your project?
   1.3. How do you plan to measure the quality of your project?
   1.4. Have you already defined quality indicators in your project?
      1.4.1.If yes, which ones?
   1.5. What is the current phase of the project?

2. **The network where the project is embedded in (including partners)**
   2.1. Who are the main partners of your project? Please name the institutions and persons that are working with you. Please indicate the Degree of collaboration and frequency with partners?

Institution	Direct contact person	Degree of Collaboration	Frequency	Degree of dependence

2.2. What is the profile of engaged stakeholders (or citizens) you are working with?
2.3. What are the personal skills you are looking for from engaged stakeholders (or citizens)?
2.4. From which culture are the engaged stakeholders (or citizens)? (diversity)
2.5. How do you embrace cultural diversity in your project?
2.6. What do you think are the problems and risks you can face in this project? In respect with your partners
2.7. What are the characteristics of engaged stakeholders (or citizens)?
2.8. Who are the key –external- actors for the success of this project?

3. Leader's profile and role in the project

3.1. What is your professional background?
3.2. What is your position at this project?
3.3. What are your main responsibilities in this project?
3.4. Can you describe a typical day at work within this project?
3.5. What are the specific tasks you are responsible for?

4. Roles of the other team members

4.1. What type of technical resources do you use to work with others?
4.2. What is missing in your opinion?
4.3. What type of features or functionalities are missing in your opinion?
4.4. What means of communication do you use?
4.5. Which communication means you feel the most at ease with?
4.6. What type of documents do you use to work with others?
4.7. Where do you store them?
4.8. How do you share them with others?
4.9. With which team members do you work closely?

Please provide their names, position, and the tasks you are involved in with them.

Name	Position	Task	Which part of the project this concerns