SSHOC
social sciences & humanities open cloud

Response from SSHOC on COVID-19

April 2020

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-04-2018, Grant Agreement #823782.

RESPONSE FROM SSHOC ON COVID-19

Introduction

The EC has requested Research Infrastructures and RI Projects to respond how they can set up possible actions that can be oriented towards the objective ***to create a European data platform for COVID-19 related information exchange****.*

Summary

Our goal is to **remove barriers** that hinder high-quality, reproducible science leading to evidence-based interventions, such as

- Non-availability of relevant data – over countries, from various sources.
  Some data might be lost forever if they are not collected in due course.
- Need for a data panel to collect actions, attitudes and behaviours of citizens
  We need a coordinated, web-based platform to collect data of citizens. This must be done in multiple countries (languages, cultural differences).
- Limited accessibility of data.
  Some data require security or privacy-protection measures and can only be accessed by remote access techniques.
- Difficulties of finding the data (by humans and machines),
  There is a massive data lake on social behaviour. Depending on the type of the crisis we need to filter out relevant data quickly. This is hindered due to lack of standardised descriptions (metadata) and physical spread of the data over countries and locations.
- Extensive efforts needed to combine data (over countries, over types of data),
  Multilinguality and differences in data types hinder data comparison and data integration, and would require large and time consuming efforts by researchers.

The Data Portal should be designed to become a **Scientific Commons** and **Virtual Collection** on all the relevant (non-medical) social and political/policy data on the COVID-19 pandemic and its consequences.

It should cover data and research from the multiple disciplines that are relevant (e.g. demography, economics, linguistics/natural language processing, media and communication studies, migration studies, political science, psychology, sociology, urban studies, etc.).

It should incorporate the great variety of data formats (official statistics, surveys, registers, social media, qualitative, multi-media data, etc.) in which data relating to the COVID-19 emergency

Ron Dekker, CESSDA, Laura Morales, Sciences Po, OTHERS

SSHOC Standpoint

SSHOC can contribute the **Social Data** component of the European **COVID-19 Data Platform**.

Our goal matches the EMBL-EBI COVID-19 goal[1]:

> The goal is
> to collect and share rapidly available research data
> from different sources and of different types
> to enable synergies, cross-fertilisation and use of diverse data sets
> with different degrees of aggregation, validation and/or completeness
> so they can be accessed by the research community.

We will follow a two-track policy:

1. A **fast track** with existing elements.
2. A **knowledge development track**, making use of online surveys and AI techniques.

Fast track

The fast track consists of a catalogue of **key data sets** for social, economic, psychological analyses, supplemented with **contextual data** (economic, social, cultural, health, migration, …). It will also include **background infrastructures** such as multilingual thesauri, controlled vocabularies (and CV management systems), metadata profiles based on global standards, tagging of sensitivity (and required security) of the data.

Like the EMBL structure, there will be Data Hubs that contain the data. These hubs are organised by data producers and national service providers: research data from surveys, national registers and official statistics data from government agencies, social media and geo-related data from commercial providers, government policies and decisions at national archives, multimedia data from TV-stations and media platforms, etc.

SSHOC partners like CESSDA, CLARIN and DARIAH (and their national service providers and nodes) can coordinate and aggregate on the cataloguing, and make use of existing overview initiatives like the ERAC Standing Working Group on Open Science[2]. Data producing entities like ESS, SHARE, GGP, WageIndicator, EVS can adapt their questionnaires to collect new information on health, values, economics, family and relatives.

[1] EMBL-EBI COVID-19 Action Plan.

[2] https://docs.google.com/document/d/1wgkq_zr1wBYR2-r4YBQlX9W_ogi_ZcchUmXanMRZS5c/edit.

Specific challenges for the social data are their **multi-linguality** and large variety of data types. **Qualitative data** have high information value, but require specific techniques for annotations, analysis and extraction of information. We will use SSHOC expertise tools to deal with qualitative and multi-media data and to address multilinguality.

Another challenge is the **quality** of the data. As there are multiple sources, we need quality assurance to filter out fake news and poor quality data. There will be a multi-layer approach: use the CoreTrustSeal to certify data service providers, make use of the FAIRification tool developed and deployed by EOSC-Nordic and Go FAIR to check metadata quality of studies (data) and repositories, DDI-converter tool developed by CESSDA-GESIS to standardise social metadata, checks on the provenance of the data.

The social data COVID-19 platform will focus on the **behaviour of people**. E.g. responses and attitudes to medical advises and government rules. How people react and behave is crucial in the spread of the virus. It will also have a broader scope, like psychological effects, impact on well-being, long-term effects of being restricted in social interactions. The data include **policy measures and political decisions**, as well as contextual data (geographical data, mobility of people, opinions).

The fast track will give an overview of data that are available (Findability) and the Accessibility of the data.

Although we focus on Europe, the goal is to realise a global network. Especially for policy analyses we would need data from China, Japan, Australia, New Zealand, US, etc.

Knowledge development track

The knowledge development track will focus on the knowledge cycle [3]. It will focus on **Interoperability** of data, including non-hierarchical data, via semantic techniques like Knowledge Graphs[4].

To optimise **Reusability**, there should be concerted actions to group data together and to enrich these data. This grouping will be dynamic, as it depends on the topic or research interest, and in cooperation with research communities.

[3] Cf. Draft ERAC Opinion on the Future of the ERA, WK 13883/2019 INIT
[4] Blumbauer and Nagy (2020), The Knowledge Graph Cookbook, Open Access.

Bringing relevant data together will increase **efficiency** (data preparation comprises up to 40-60% of research time), improve comparability and provide a level-playing field as all researchers have same opportunities for using the data.

To facilitate cooperation and to provide seamless access even to sensitive data, we will set up **secured environments** for bringing data and researchers together. In principle, data will stay where they are, using secured connections (e.g. via light paths) or remote access techniques to analyse the data.

Besides existing data, there will be a **web-based Europanel** that builds on existing European SSH Surveys. Using the internet, we can immediately investigate **opinions and attitudes** of the European people during a crisis. In combination with contextual data, we can build a **monitoring system** on human behaviour, occurrence and spread of a virus, etc.

Connecting to the COVID-19 Platform

From the start the goal is to connect this social data infrastructure to the COVID-19 Platform. This implies that catalogues, metadata, protocols need to be aligned with the life sciences and other parts of this platforms.

The Cluster projects (ENVRI-FAIR, ESCAPE, LIFE-Watch, PaNOSC, SSHOC) have started cooperation on several aspects, like secured access, market place, governance[5].

The COVID-19 platform could serve as a facility for analysis and policy evaluations. E.g. what would have been optimal decision-making given the demography, structure etc. of a country or region. Other topics can be on studying the effects of cultural differences. New techniques for text and data mining could be used.

The platform can also serve cross-national networks of academics (political scientists, computer scientists, language processing specialists, media studies scholars, etc.), policy experts and citizens. We can introduce citizens' science projects as there will be a need for capacity to work on the vast amount of annotations and translations of documents, interviews, etc.

[5] Cf. the Clusters' joint position paper: https://zenodo.org/record/3675081#.X322Ey2w21s

Annex Social Sciences and Humanities Open Cloud - SSHOC

The European SSHOC project

- covers multiple disciplines;
- brings in expertise on multilingual issues;
- deals with a great variety of data formats;
  (surveys, registers, social media, qualitative multi-media data)

Social Sciences and Humanities study human behaviour which is one of the key parameters in the spread of the virus, and are also used to deal with, and understanding different cultures.

- We can provide the substantive, methodological and data management expertise on the connections between the **COVID-19 emergency and other major global challenges (cross-cutting dimension)**, notably migration and climate change, so that data collection and analysis can examine the interconnections with larger global processes.
- We already work within the EOSC and FAIR parameters, so we are able to build such a Data Portal within an **efficient and sustainable framework for EOSC** to ensure that relevant, but ad-hoc, initiatives last over time and establish the basis for future situations of need.
- Our infrastructures have **ample experience in generating data, collecting and curating data, and make the data available for reuse** – even if it concerns sensitive data that need to remain in a protected and secured environment. New developments tackle social media data, or combinations of health, geographical and social data. And SSH has experience in dealing with qualitative data (texts, audio, tv) and assign annotations and references to these data.

New developments tackle social media data, combinations of cross-domain health, geographical and social data. And SSHOC partners use text recognition, semantic techniques and AI to filter data and to track fake news. In the project, there are pilots with the use of Knowledge Graphs and outreach to research communities.

Annex Overview of existing initiatives

A non-exhaustive overview of current COVID initiatives.

- EMBL/EBI: a COVID Research Data platform to find a vaccine integrated into EOSC to facilitate data sharing
- OpenAIRE: Dashboard/Scientific Gateway where all metadata about the research products related to COVID19 and SARS-CoV-2 (clinical data, epidemiological, scientific articles, software, methodologies, processes, etc.) will be automatically aggregated from relevant sources world-wides for research discovery
- FAIRsharing.org are curating a collection of databases and standards on COVID-19 but as far as I can see only in the bio-medical sciences: https://fairsharing.org/collection/COVID19Resources
- RDA: workgroup on Data sharing Guidelines (https://www.rd-alliance.org/groups/rda-covid19)
- Zenodo dedicated space: https://zenodo.org/communities/covid-19/
- Data Together COVID-19 Appeal and Actions: positioning against a centralized data warehouse and calling for all and any emerging platforms to adhere to the FAIR principles and for them to include rich, machine-actionable metadata https://www.rd-alliance.org/sites/default/files/attachment/Data%20Together%20COVID-19%20Statement%20FINAL.pdf
- Michigan's ICPSR New Report (http://hdl.handle.net/2027.42/154682) reviewing best practices for using data resources from ICPSR, its projects, and its collaborating partners for measuring the impact of epidemics.
- WAPOR: is compiling a list of surveys on COVID-19 https://wapor.us11.list-manage.com/track/click?u=80f5ebf95750d28938fb8c1c4&id=8efaa95113&e=1fb46e030b
- OECD: a Google doc with a list of Open Government initiatives on COVID-18 https://docs.google.com/document/d/1BdSnXzCZ1Z7ovOrPue3O0osRUpiqTKlu8pwG9U4DwWw/edit
- GitHub: https://github.com/datasets/covid-19
- SparcEurope has curated a list with a number of resources and calls for action around COVID-19: https://sparceurope.org/coronaopensciencereadsandusecases/
- The Melinda & Bill Gates Foundation with Exaptive have created a data-centric platform https://covid-19.cognitive.city/cognitive/welcome

Annex Functionalities

The core element of the Platform would be the **metadata and data catalogue**, that should be structured through a clear topic-focused index. It can build on the user-interface of the current COVID-19 Platform and possibly add functionality that is relevant for social data, cf. official statistical websites that are structured in the Statistics by theme.

The **indexing** structure needs to be dynamic. It must allow the population of new themes not yet covered and proposed by data producers). It could be based on the FAIRsharing approach and lay out: https://fairsharing.org/collection/COVID19Resources).

Each dataset (the smallest unit at which each data type is available) should be described with full metadata (relying on DDI, preferably version 2 or 3), including a DOI or other permanent identifier linking to the actual dataset (also a FAIR and an EOSC requirement).

Additionally, the Portal should provide the following functionalities:

- A powerful advanced search tool that allows to search without using the thematic Index.
- An area that allows to register the interest in generating new data in collaboration with others for a specific topic/theme. This will require the ability to create user profiles where users can register and provide and update the roles they want to take on.
- A set of tools that allow to self-archive data (cf. DataverseEU which is part of the SSHOC project), and add them to the overall index of the catalogue after passing some (expeditious but sufficient) quality controls.
- A set of tools that allow for (quick) data visualization of data already in the catalogue.

And we need background functionalities:

- Metadata harvesters for metadata based on international standards
- Controlled vocabularies for indexing and search functions
- Possibility of thematic tagging
- Possibility of access and re-use level tagging (cf. Harvard tagging)

SSHOC
social sciences & humanities open cloud

www.sshopencloud.eu

@SSHOpenCloud

in/company/sshoc

info@sshopencloud.eu