http://dx.doi.org/10.14236/ewic/ODAK22.7

Open Data Literacy by Remote: Hiccups and Lessons

Alessia Antelmi
Dipartimento di Informatica
Università degli Studi di Salerno
aantelmi@unisa.it

Maria Angela Pellegrino
Dipartimento di Informatica
Università degli Studi di Salerno
mapellegrino@unisa.it

Open Data are published to ensure the creation of value and data exploitation, but limited technical skills are a critical barrier. Most users lack the skills required to assess data quality and its fitness to use, awareness of open data sources, and what they can do with the data. To advance the dialogue around methods to increase awareness of Open Data, improve users' skills to work with them, and deal with the requirement of letting future citizens develop data and information literacy according to 21st-century skills, this article proposes a series of workshops to let Italian high school learners familiarise themselves with effective communication based on Open Data. The article describes an ongoing activity, reporting preliminary results on engagement and learning. We discuss challenges in engaging learners remotely and the promising learning outcomes achieved by overcoming cultural and technical barriers to visualise Open Data.

Open Data, Data Literacy, Data Visualisation, Engagement, Learning, At a distance, High school

1. INTRODUCTION AND BACKGROUND

"*Open Data (OD) are data that can be freely used, shared and built-on by anyone, anywhere, for any purpose*" (Open Knowledge Foundation 2013). Data exploitation is based on the assumption that once data are discoverable, accessible, available in alternative formats, and with licensing schemes that allow free reuse, interested stakeholders will create value out of them (Chan 2013; Janssen et al. 2012).

However, limited technical skills are an important barrier as most users are unaware of available OD (Martin et al. 2015) and what they can do with the data (Safarov et al. 2017), and further lack data literacy skills. Even if user training is crucial in facilitating and spurring OD consumption, there is limited research on strategies to train users (Gascó-Hernández et al. 2018), scarce involvement of citizens (Safarov et al. 2017; Styrin et al. 2017), and only isolated efforts to consider skills and tasks that interested stakeholders desire to perform with data (Martin and Begany 2017; Susha et al. 2015).

To advance the dialogue around methods to increase awareness of OD, improve users' skills to work with OD, and deal with the requirement of letting future citizens develop data and information literacy according to 21st-century skills, this article proposes a series of remote workshops to let Italian high school learners familiarise themselves with effective communication based on OD. We report on an ongoing activity spanning from February to May 2022, and we discuss preliminary qualitative results concerning engagement and learning.

2. WORKSHOP DESIGN

2.1. Research Questions

The main research goals of this article relate to understanding participants' *engagement* and *learning* in the proposed workshops. Our aim translates into two main research questions (RQs):

RQ1 - To what extent participants are engaged in working with OD?

RQ2 - Which is the experienced easiness in learning and immediately exploiting the introduced concepts?

2.2. Participants and Setting

A total of 73 classical high school learners (i.e. from Italian Liceo Classico) joined the workshops (80% females). All of them were unfamiliar with concepts related to data literacy (e.g. data manipulation and chart creation) and tools used to perform data exploitation (e.g. Excel and Google Sheet). Participants' ages were heterogeneous, with a

Proceedings of the Symposium on Open Data and Knowledge for a Post-Pandemic Era ODAK22, UK

minimum of 14, a mean of 16, and a maximum of 18. Participants were divided into four groups based on their attended classes. The workshops were organised in four turns, one for each group. Meetings started in February 2022 and are scheduled until May 2022, entirely online and at a distance due to COVID-19 regulations. The workshops took place via the Webex synchronous videoconferencing tool. A researcher led each workshop, acting both as a moderator of the event and as an observer, annotating the notes collected during each lesson at the end of the workshop.

BIMED[1] curated the recruitment process, managing the agreement with an Italian high school specialised in classical studies and the recruitment with an occasional service contract of the researchers involved.

Researchers from the University of Salerno undertook post-workshop data processing anonymously to meet data protection requirements and constraints.

2.3. Protocol

The workshop spanned over five days, two hours per day split into a one-hour introductory phase and one-hour hands-on session. In the introductory phase, the moderator explained concepts, encouraged participants to reply to challenges formulated as questions and quick oral exercises, and replied to any request for clarification. During the hands-on session, participants worked in groups of three to five members, assisted by the moderator when needed, and co-created charts through Google Sheet and Google Data Studio. Finally, each group was invited to present the authored chart(s) by screen-sharing. Learning contents, defined by domain experts in OD and open knowledge, are inspired by the *Data, Information, Knowledge pyramid* (Frické 2019) that guides users to explore available (open) data, transform data into information with compelling visualisations, and let external actors take decisions by efficiently communicate data insights. Details on the learning content of each workshop ($W_{\#}$ with # progressive number) follow.

2.3.1. W_1: Basic visualisation techniques

W_1 introduced basic visualisation by simple text, tables, bar charts, line charts, and point charts. For each chart, the moderator first challenged participants to read and interpret it, to then clarify the chart objective, the data types required to obtain it, real and practical scenarios where the chart is used, how to read the chart correctly, and how to generate it with Google Sheet. All the examples in this workshop are based on The Avengers movie world to propose a topic close to participants' age and (potential) interests.

2.3.2. W_2: Advanced visualisation techniques

W_2 introduced geographical maps, organisation charts, treemaps, radar charts, and box plots. Due to the heterogeneity of the required data types, each chart was introduced by a different example as close as possible to the topics participants are familiar with, e.g., school subject hierarchies and the grading system.

2.3.3. W_3: Tips and tricks in data visualisation

W_3 aimed to raise awareness in participants and help them develop the critical spirit to pose the right questions while reading and interpreting charts as "a picture is worth a thousand words..." when you can read it properly. Even if charts seem to represent data objectively, authors can voluntarily or accidentally introduce errors in data visualisations and lead to wrong interpretation. Tricks can be classified in either data or chart manipulation. While tricks on data are related to the generation of charts starting from uncertain, incomplete, or de-contextualised data, tricks applied on charts concern axes manipulation, missing axes scales, extreme zoom, and percentages that do not sum to 100. In the hands-on session, participants were challenged to create a data table about any topic of interest and design two versions of the same chart. The goal was to exploit one of the discussed tricks to present different interpretations of the same phenomenon (e.g. a chart describing a rapidly growing trend vs a chart visualising the same data with a plateau).

2.3.4. W_4: Open Data repositories

W_4 clarified the key concepts of OD, distinguishing legal openness thanks to open licenses and technical openness related to published OD data formats. Real regional and national OD repositories were introduced, such as the portal of the Campania region[2], the ISTAT Open Data portal[3], and the government data platform[4]. Participants were guided in downloading CSV datasets, importing them into Google Sheet and dealing with the data format and technical challenges. In the hands-on session, participants were challenged to select an Open Dataset published by the Campania region and co-author effective and appropriate charts.

2.3.5. W_5: Open Data communication

W_5 focused on the advantages and techniques in proposing coherent, linked, and interactive data visualisations. It resulted in the introduction of data dashboards and Google Data Studio to easily author

[1] BIMED: https://www.bimed.net

[2] https://dati.regione.campania.it/opendata

[3] https://www.istat.it/it/dati-analisi-e-prodotti/banche-dati

[4] https://www.dati.gov.it

data reports without asking for technical skills in programming. Participants were challenged to start from a dataset available on Google Data Studio and author an interactive and shareable data report during the hands-on session.

2.4. Data Gathering for Engagement

The moderator collected observations concerning participants' engagement during and immediately after each workshop. Observers tracked data in diaries, both via notes and comments and in the form of codes from the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 (Ocumpaugh 2015). BROMP is a protocol for qualitative and quantitative field observations of behaviours and affective states in learning with technology, which indicate engagement in tasks. The categories used in this paper are listed in Table 2. We reported observations at the class or group level because participants' reluctance to switch on either their webcam or microphone made it impossible for observers in remote settings to track the actual engagement of each participant. Qualitative observations have been preferred to standardised and quantitative surveys due to the lack of continuity of the participants in the described workshops.

2.5. Data Gathering for Learning

During the hands-on session of each workshop, participants were invited to share the authored Google Sheet or Google Data Studio project, which was then used to assess their learning of OD literacy. The metrics used to evaluate the learning aspect, inspired by Krstevska (2021), consider the quality of the produced charts and the data they visualise. Two domain experts iteratively defined these metrics, and then they followed a two-step procedure to evaluate participants' projects. First, the experts independently reviewed each visualisation according to the metrics reported in Table 1; then they resolved inconsistencies through discussions. The final score associated with each chart is the sum of the score of each metric. Thus, the maximum score per chart is 8 (the higher, the better).

Table 2: *The BROMP categories used in our study.*

Positive	**Description**
Concentration	Proactive or on demand participation, outcome sharing.
Delight	Pleasurable expressions used to describe the workshops.
Negative	**Description**
Confusion	Difficulty in understanding tasks or using the right terminology.
Frustration	Feelings of distress in dealing with introduced concepts or tools.
Boredom	Lack of interest in joining the activity.

3. RESULTS AND DISCUSSION

The following sections report the results obtained according to the metrics previously defined and discuss them in relation to the relevant RQs.

3.1. Engagement Results

The first aspect worth noting is how the number of participants reduced over time. Nevertheless, we have to clarify that not all classes attended their last lessons when submitting the article. Further, the discrepancy in the number of shared and expected projects sheds light on the technical challenges that need to be faced when dealing with data manipulation tools and the cultural barrier to exploiting OD training workshops in the best possible way. Most of the participants actively engaged in the workshops, being *concentrated* and primarily interacting during the hands-on sessions

Table 1: *Metrics for assessing the learning outcomes.*

Metrics	**Description**	**Score**
Chart-related		
Right chart choice	Chart chosen according to the available data and not based on aesthetics.	{0,0.5,1}
Chart correctness	Avoid broken Y scale, comparison between not comparable series.	{0,0.5,1}
Use of convention	Follow standard practices of visualisation.	{0,0.5,1}
Amount of displayed data	Avoid displaying too much data.	{0,0.5,1}
Reader support	Use of qualifying numbers, legends, title and axis labels.	{0,0.5,1}
Effective use of decoration	Avoid distracting and misleading decorations (such as 3D modeling).	{0,0.5,1}
Data-related		
Original data	Use of data different from the ones used during the introductory phase	{0,0.5,1}
Real data	Use of real OD or publicly available data.	{0,0.5,1}

Table 3: *Engagement results.*

	Participants	Expected projects	Shared projects	Shared charts	Avg. number of charts per project	1st BROMP code	2nd BROMP code
$\mathbf{W_1}$	73	19	15	18	1.2	Concentration	Boredom
$\mathbf{W_2}$	70	20	18	25	1.4	Concentration	Boredom/Confusion
$\mathbf{W_3}$	55	17	9	18	2.0	Concentration	Confusion
$\mathbf{W_4}$	57	16	6	13	2.2	Concentration	Boredom/Confusion
$\mathbf{W_5}$	24	6	5	17	3.4	Concentration	Confusion

(RQ1). In some cases, active participants showed *confusion*, mainly caused by the theoretical aspects of advanced charts in W_2, challenges in immediately applying tricks to manipulate data and visualisations in W_3, the complexity of real OD in W_4, and technical aspects concerning the Google Data Studio platform in W_5. The reluctance of attendees to switch on their cameras and their attitude to only partially contribute in an active way to the workshops is interpreted as negative engagement, classified as *boredom* by observers. Table 3 summarises these results.

3.2. Learning Results

Learning results are graphically represented by the box plots in Figure 1. It is worth recalling that the maximum score is 8. Further, in some tasks, the minimum score is bounded by the task itself. Specifically, all the participants must use OD in W_4; hence, the minimum score is 1. In W_5, all participants must use the same real and original dataset. Hence, the minimum score is 2.

In general, the results are rather satisfying, as the minimum score in all workshops is at least 5. This outcome demonstrates that participants who actively joined the workshops learnt how to exploit (open) data by visualisations. In all workshops but W_3, the maximum score of 8 underlines that active participation in the proposed workshops led to satisfying learning achievements (RQ2).

The youngest participants tend to have the best learning outcomes in the first workshops, while the last workshops were challenging for them. In fact, they reported they required more time to internalise the introduced concepts. In particular,

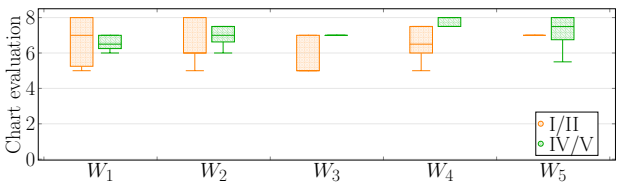

Figure 1: *Attendees' learning results grouped by class. I/II corresponds to the youngest participants, IV/V to the oldest ones.*

most of the youngest participants misunderstood the assignment in W_3 and experienced technical challenges posed by Google Data Studio in W_5. On the contrary, it seems that older participants learnt less during the first workshops, probably because they were not stimulated by the discussed concepts (maybe too easy for them), while they were particularly productive in exploiting real open datasets in W_4 and authoring data dashboards in W_5.

An important aspect worth commenting on is that moderators noticed how participants demonstrated interest in using real OD also when the task did not explicitly require it. In fact, attendees stressed both the real data source used when reporting the authored charts and the time spent looking for stimulating and real scenarios. Covered topics span from sports to health, from economy to gender gaps and discrimination. Moreover, moderators were also surprised by observing that participants proactively explored charts and tool features not overviewed during the introductory phase, demonstrating independence and interest in the workshop topics.

4. CHALLENGES AND SUGGESTIONS

OD initiatives encourage the publication of OD to let interested stakeholders create value out of them. There is limited evidence for such a transformation due to several factors influencing and minimising the use of OD (mainly technical and cultural barriers). According to the survey authored by Saddiqa et al. (Saddiqa et al. 2019), interventions to improve users' skills and knowledge are rare in the literature. Towards this direction, this article reports on an ongoing activity concerning a series of at-a-distance workshops to let Italian high school learners familiarise themselves with effective communication based on OD. We are currently working on a systematic literature review to compare such initiatives. Considering the preliminary qualitative results concerning engagement and learning reported in this article, we briefly comment on the experienced challenges and suggestions for future editions of similar projects.

Remote workshops, as the initiative described in this article, seem to limit the potential benefit of OD training sessions due to participants' reluctance to switch on cameras. This attitude can be partially justified by shyness in being recorded during the event, the tendency to participate in informal activities not in a stable setting (from school to home), and the perception of being evaluated due to the presence of a reference professor within their school. Participants might be further encouraged to join OD training sessions actively after revising some of these inhibitors' factors.

During the workshops, *technical challenges* were not rare. They were mainly due to the technological immaturity of most of the participants, lack of experience in data management and exploitation tools, and difficulties posed by the mobile version of video-conferencing tools that are not always user-friendly. Considering the impact that age had on the learning outcomes, it is crucial to *tailor workshops' content and tools* to participants' age and technical skills. A revision of the proposed protocol might consider focusing longer on W_3 and avoid using Google Data Studio with the youngest participants.

Observers noticed that participants freely explored real and significant social issues during the hands-on sessions. This outcome suggests that workshop organisers should not be afraid to introduce charts and OD value creation using examples borrowed from political and health contexts as also youngest participants are ready to deal with them.

Participants retrieved and exploited real OD also when it was not explicitly required by the task demonstrating that workshops similar to the ones introduced in this article are positively perceived by future citizens and have the potential to spur OD use for possible stakeholders. Still, the decreasing *number of participants* indicates the presence of a cultural barrier to exploiting OD and stresses the need to further encourage OD stakeholders to join training activities.

Acknowledgments: The authors thank BIMED, represented by its president Andrea Iovino, for letting them conduct the research at the basis of this article.

REFERENCES

Chan, C. M. (2013), From open data to open innovation strategies: Creating e-services using open government data, *in* '2013 46th Hawaii International Conference on System Sciences', IEEE, pp. 1890–1899.

Frické, M. (2019), 'The knowledge pyramid: the DIKW hierarchy', *Knowledge Organization* **46**(1), 33–46.

Gascó-Hernández, M., Martin, E. G., Reggi, L., Pyo, S. and Luna-Reyes, L. F. (2018), 'Promoting the use of open government data: Cases of training and engagement', *Government Information Quarterly* **35**(2), 233–242.

Janssen, M., Charalabidis, Y. and Zuiderwijk, A. (2012), 'Benefits, adoption barriers and myths of open data and open government', *Information systems management* **29**(4), 258–268.

Krstevska, S. (2021), 'Interpreting data visualizations: The basics'. [Online, last access April 2022]. **URL:** *https://guides.zsr.wfu.edu/interpretdataviz*

Martin, E. G. and Begany, G. M. (2017), 'Opening government health data to the public: benefits, challenges, and lessons learned from early innovators', *Journal of the American Medical Informatics Association* **24**(2), 345–351.

Martin, E. G., Helbig, N. and Birkhead, G. S. (2015), 'Opening health data: what do researchers want? early experiences with new york's open health data platform', *Journal of Public Health Management and Practice* **21**(5), E1–E7.

Ocumpaugh, J. (2015), 'Baker rodrigo ocumpaugh monitoring protocol (bromp) 2.0 technical and training manual', *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences* **60**.

Open Knowledge Foundation (2013), 'Defining open data'. [Online, Last access April 2022]. **URL:** *https://blog.okfn.org/2013/10/03/defining-open-data/*

Saddiqa, M., Kirikova, M., Magnussen, R., Larsen, B. and Pedersen, J. M. (2019), 'Enterprise architecture oriented requirements engineering for the design of a school friendly open data web interface', *Complex Systems Informatics and Modeling Quarterly* (21), 1–20.

Safarov, I., Meijer, A. and Grimmelikhuijsen, S. (2017), 'Utilization of open government data: A systematic literature review of types, conditions, effects and users', *Information Polity* **22**(1), 1–24.

Styrin, E., Luna-Reyes, L. F. and Harrison, T. M. (2017), 'Open data ecosystems: an international comparison', *Transforming Government: People, Process and Policy* .

Susha, I., Grönlund, Å. and Janssen, M. (2015), 'Driving factors of service innovation using open government data: An exploratory study of entrepreneurs in two countries', *Information polity* **20**(1), 19–34.