The Hetor project: a joint effort to co-create Cultural Heritage Open Data in the Campania Region

Maria Anna Ambrosino[1], Vanja Annunziata[1], Maria Angela Pellegrino[1,*] and Vittorio Scarano[1]

[1]*Università degli Studi di Salerno, via Giovanni Paolo II, 132 84084 Fisciano (SA), Italy*

Abstract

Open Data are published to encourage their exploitation, but limited technical skills are a crucial barrier. Initiatives to let learners in particular and users in general exploit Open Data are rare in literature, and they mainly focus on the exploitation phase rather than the authoring one. To increase Open Data awareness and move users in the position of open data curators, the HETOR project regularly organise workshops to let participants create, publish, and exploit Open Data. This project started in 2016 and resulted in the co-creation of dozens of high-quality open datasets, publicly available on CKAN, involving hundreds of learners, public administration delegates, and volunteers in associations. This article describes the involved communities within the HETOR project and quantitatively and qualitatively details authored datasets covering any aspect of Cultural Heritage in the Campania Region.

Keywords

Open Data, Authoring, Local Communities, Repository, Cultural Heritage

1. Introduction

"*Open Data (OD) [...] can be freely used, shared and built-on by anyone, anywhere, for any purpose*" [1]. OD is a promising tool to raise curiosity about data sources, data availability, and the techniques underlying data access, extraction, and analysis [2], develop data literacy [3], enhance digital skills [4, 5], stimulate critical thinking, collect relevant information and produce reliable conclusions [6]. OD are published to let interested stakeholders exploit data and create value, but limited technical skills are a crucial barrier [7].

Initiatives to let learners and interested users exploit OD are rare in literature. The situation is even worse if we look for opportunities to move them into the position of OD publishers. To advance the dialogue around methods to increase OD awareness and improve users' skills to familiarise themselves with OD, the HETOR project regularly organizes workshops with different communities to let them create, publish, and exploit OD. This article reports the effort invested by HETOR in co-authoring OD with learners, associations, and Public Administrations (PAs).

Education can take place in a heterogeneous setting, traditionally classified as formal, informal, and non-formal learning [8]. *Formal learning* corresponds to an intentional and systematic

19th IRCDL (The Conference on Information and Research science Connecting to Digital and Library science), February 23–24, 2023, Bari, Italy

[*] Corresponding author.

mariaanna.ambrosino@gmail.com (M. A. Ambrosino); vanja.annunziata@gmail.com (V. Annunziata); mapellegrino@unisa.it (M. A. Pellegrino); vitsca@unisa.it (V. Scarano)

CEUR Workshop Proceedings (CEUR-WS.org)

education model, and it typically takes place at school. *Non-formal learning* is still intentional but takes place outside formal learning environments, typically occurring in community settings, such as associations or clubs. While the HETOR activities with learners take place as formal learning, the ones with PAs and associations are classified as non-formal learning.

The contribution of this manuscript is twofold: i) it reports the effort of the HETOR activities in preserving and digitizing Cultural Heritage (CH) of the Campania Region by co-creating OD involving communities of associations, PAs, and learners; ii) it details the HETOR datasets publicly available as CSV files on CKAN with an open license to let researchers, data lovers, or any interested user exploit available data to disseminate data, improve data quality by machine-learning based approach, or model tabular datasets via Semantic Web technologies.

The article is structured as follows. Section 2 overviews related work; Section 3 reports on the HETOR project, overviews the involved communities, and quantitatively and qualitatively details the authored open datasets; Section 4 discusses potentialities interpreted as success stories and limitations; then, the article concludes with final remarks and future directions.

2. Related work

More and more researchers and educators recognise the potentialities in using OD as an educational resource [9] targeting heterogeneous goals, such as focusing on deeper learners' skills in environmental education [10, 11] or improve data visualization and data literacy skills [12, 13, 14]. Learners usually experience OD in a formal setting, as including skills in educational curricula democratises the learning process [15]. However, reaching new audiences is an important benefit of OD [16, 17, 18, 19]. Gasco et al. [20] describe and compare interventions to increase awareness of OD, enhance users' skills and engage them in the use of OD by involving learners, PAs, non-governmental organizations, and citizens. Similarly, the HETOR project targets heterogeneous communities, i.e., schools, PAs, and associations.

Interventions to improve users' skills and knowledge proposed in the literature mainly focus on OD exploitation, to engage learners while letting them learn [21, 22, 23], improve their awareness of the environment and smart city development [24], master OD visualization [25, 26]. OD initiatives rarely move learners to the position of OD producers. Consequently, learners only sometimes experience OD production challenges, such as defining data schema, collecting information, dealing with licenses, and mastering OD authoring tools. Chen et al. [21] employ an instructional pervasive gaming model to deeper participants' CH knowledge. They exploit an OD Kit form that is used as the interface for implicitly gathering information from the mobile device. Similarly, HETOR's workshops move secondary school learners to the position of OD publishers, letting them experience the challenges inherent in the role of data curator. A key difference with related work is that learners *explicitly* author OD.

3. HETOR activities to co-create Open Data

The HETOR project[1] aims to collect and make available both the "Open Heritage" provided by the National Institutions, such as ISTAT, MIBACT, MIUR, and Campania Region, and the one

[1] The HETOR project: http://www.hetor.it

created by interested citizens concerning their local CH, improving the quality and quantity of OD at the local and national level. This article focuses on OD concerning the Campania Region. To reach these goals, the HETOR project co-creates OD in the tabular format working with schools, associations, and local PAs via a Social Platform for Open Data (SPOD)[2], reuses and exploits data via data visualizations, and disseminates data stories via social networks, such as Facebook, Instagram, and Telegram, and the Hetor website.

Communities. 3 communities actively contribute to the HETOR project, associations, schools, and PAs. By detailing agencies and number of users, HETOR collaborated with 39 users belonging to 14 associations, 67 users belonging to 3 PAs, and 596 learners belonging to 9 schools. All the associations, but one, are in small municipalities, all belonging to the province of Salerno. The effort from *Nocera Inferiore* is remarkable, with the participation of 11 associations joining the HETOR project. The school community is the largest in terms of involved users, with Avellino holding a record of 215 users. School agencies cover all the provinces of the Campania Region but Benevento, mainly collaborating with municipalities. Moreover, schools are heterogeneous in terms of involved school type, involving both High Schools and technical institutes. The PA community is the smallest group, represented by mayors, cultural advisors, school professors, and politicians. They cover all the Campania region provinces. While some municipalities join two communities, such as Montoro and Avellino, it is remarkable the participation of Nocera Inferiore in all the communities. While activities with the schools take place as formal learning, collaborations with associations and PAs represent non-formal learning. While PAs and associations freely join the HETOR project to digitise, document, and preserve local CH, schools join it to let learners develop data literacy skills.

The HETOR datasets. This section overviews datasets authored within the HETOR project by learners, local PAs, and associations, quantifies the effort invested in preserving and digitizing CH of the Campania Region, and reports the quality of the authored datasets. All the datasets[3] are publicly available on CKAN with the Creative Commons License, in the CSV format, and in the Italian language. Datasets are manually authored and refined via SPOD. Table 1 reports the English dataset name, the community that authored the dataset, quantitative details in terms of the number of rows, columns, and cells, and qualitative details in terms of completeness and accuracy. When we report that a dataset is authored by a given community, such as the school, we mean that learners created the dataset supervised by the HETOR group. Datasets are classified according to the CH definition in Tangible CH, further split into movable and immovable, Intangible CH, Natural CH, Food & Wine, and other that includes geographical information and details about companies and associations. The completeness metric reports the percentage of non-empty values. The accuracy metric is computed by verifying how many textual geographical fields (such as municipalities) are correctly reconciled with Wikidata towns or municipalities. The accuracy metric also considers how many ZIP codes (if any) in the datasets match the ones retrieved by Wikidata. The qualitative information is computed by Open Refine, exploiting the facet and the reconciliation mechanisms.

[2] SPOD: http://spod.databenc.it
[3] Hetor datasets: http://www.hetor.it/dataset

Table 1: Overview of Open datasets co-created within the HETOR project.

Dataset details		Quantitative info			Qualitative info	
Name	**Author**	**Rows**	**Cols**	**Cells**	**CMP.**	**ACC.**
Tangible Cultural Heritage - Immovable Cultural Heritage						
Castels and coast towers	Hetor	523	31	16213	45%	96%
Rock cults	Hetor	88	25	2200	50%	82%
Theatres and odeons	Hetor	32	27	864	77%	87%
Noble palaces in Fisciano	Assoc.	22	16	352	89%	100%
Churches and art in Calitri	School	64	19	1216	86%	97%
Cilento resources	Hetor	145	5	725	98%	87%
Abandoned factories	School	69	23	1587	72%	99%
CH of San Nicola la Strada	School	22	13	286	83%	100%
Calitri buldings	School	27	18	486	75%	100%
Novera Inferiore Itineraries	School	49	14	686	98%	98%
Agrometeorological network	School	33	12	396	92%	70%
Collina del Parco risk map	Assoc.	8	13	104	70%	-
Caserta contemporary itineraries	School	31	16	496	98%	94%
Caserta modern itineraries	School	31	14	434	100%	100%
Caserta medieval itineraries	School	42	16	672	99%	93%
Capua & Aversa Churches	School	133	20	2660	57%	94%
Agriculture assistance centres	School	161	10	1610	99%	89%
Clinical records of Psychiatric Hospital in Nocera Inferiore	Assoc.	200	10	2000	90%	86%
Bio companies	School	203	16	3248	100 %	89%
Solidarity Purchasing Groups	School	29	12	377	99%	85%
Didactic farms	School	267	14	3738	99%	-
Gate crests of Nocera Inferiore	Assoc.	9	7	63	70%	-
Nocera Inferiore votive shrines	Assoc.	49	12	588	74%	-
Photografic Safari @Paestum	School & Assoc.	115	9	1035	98%	100%
Touring club	Assoc.	29	19	551	72%	97%
Avellino POI	School	1439	13	18707	83%	99%
Caserta POI	School	1314	13	17082	81%	100%
Nocerino - Sarnese POI	School	285	13	3705	82%	99%
Artistic High Schools	Assoc.	36	37	1332	92%	94%
Museums of Cilento and the Gulf of Policastro	School	69	17	1173	68%	93%
Hidden treasures	Assoc.	83	50	4150	77%	96%
Tangible Cultural Heritage - Movable Cultural Heritage						
Trademarks	School	32	43	1376	98%	58%
Peasant civilization	School	196	14	2744	87%	-
Open Museum	School	213	18	3834	90%	-
Irpino Museum: Epigraphs	School	11	53	583	93%	91%

Irpino Museum	School	21	30	630	100%	100%
Ancient arts and jobs	School	95	13	1235	80%	-
San Nicola La Strada churches decor elements	School	35	12	420	97%	-
Forino company trademarks	Assoc.	109	23	2507	100%	-
Chronicle of Nuceria Alfaterna and its territory: the Agro Nocerino museum	Assoc.	327	27	8829	69%	95%
Handicrafts	Hetor	26	12	312	100%	54%
Monumental complex of the former Bourbon prison of Avellino	School	95	18	1710	98%	85%
Art at UNISA	School	7	9	63	94%	-
Mathematics Museum	School	53	18	954	96%	-
Intangible Cultural Heritage						
Central Political Records Office	School	509	22	11198	88%	81%
Provincial political records of Caserta during the Kingdom of Italy	School	464	23	10672	88%	79%
“La torre” press	School	1287	10	12870	100%	-
Uses and customs of Upper Irpinia	School	65	11	715	79%	89%
Ancient arts and crafts of the Beniamino Tartaglia Museum of Aquilonia - Crafts Section	School	130	13	1690	80%	-
Traditional games	Assoc. & PA	29	15	435	60%	-
The Nocerina industry from the unification of Italy to the economic miracle	Assoc.	385	12	4620	30%	-
Proverbs and ancient words	Assoc.	83	10	830	100%	-
The local press since the Italian unification	Assoc.	33	17	561	45%	-
History of the Carnival and of the Carts of Marcianise	School	35	10	350	90%	100%
Natural Heritage						
Natural areas	Hetor	42	18	756	94%	-
2018 blue flag beaches	Hetor	54	10	540	86%	100%
Regional forests	School	10	19	190	98%	60%
Seed woods	School	17	19	323	100%	88%
2020 blue flag beaches	Hetor	60	10	600	86%	100%
2021 blue flag beaches	Hetor	61	10	610	85%	100%
2022 blue flag beaches	Hetor	62	11	682	83%	100%
Food and Wine						
Typical products	Hetor	607	15	9105	84%	-

Wines	Hetor	1858	15	27870	91%	95%
Dairies authorized for the production of buffalo mozzarella D.O.P.	School	91	20	1820	100%	96%
Producers at Km 0	School	35	24	840	100%	88%
Farms authorized to produce D.O.P. buffalo mozzarella	School	122	12	1464	100%	92%
Pizzerias in Naples and Caserta	School	49	19	931	100%	98%
D.O.C.G., D.O.C., I.G.P. wines	School	79	16	1264	83%	-
Craft breweries	School	88	27	2376	81%	91%
Coffee roasters	Hetor	107	19	2033	99%	97%
Salerno farmhouses	School	207	15	3105	98%	93%
Slow Food Presidia	School	89	15	1335	100%	-
Social farms	School	19	22	418	98%	95%
Nocera: social farms	Assoc	16	5	80	100%	-
Other (Companies and Geographical Information						
Nocera Inferiore streets	School	245	16	3920	100%	-
Pro Loco	Hetor	580	13	7540	79%	95%
Autonomous Care, Stay and Tourism companies	Hetor	15	11	165	92%	64%
Tourist Boards	Hetor	5	10	50	100%	-
San Nicola La Strada streets	School	163	12	1956	100%	-
ANICAV Companies	School	32	12	384	100%	84%
Companies in Upper Irpinia	School	407	17	6919	78%	100%
Battipaglia & Eboli Companies	School	169	22	3718	99%	100%
Salerno Start up and SMEs	School	153	21	3213	99%	77%
Montoro's fractions	Assoc.	79	11	869	91%	30%
Avellino municipalities	School	118	24	2832	94%	99%
Salerno municipalities	School	158	24	3792	96%	97%

4. Discussion: Potentialities P_x and Limitations L_y

P_1 - **Joint effort.** Since 2016, HETOR has collaborated with three communities, associations, schools, and local PAs, with 27 agencies and 702 users. It demonstrates that the HETOR project is a joint effort of data lovers, experts in the field, citizens, and learners in co-creating content as OD. The biggest community in terms of agencies is the association one, with 14 joining agencies. It involves volunteers, data experts, and data lovers.

P_2 - **Consistent OD co-creation effort.** The HETOR project co-authored 87 datasets concerning CH in the Campania Region since 2016. It is worth noting that the dataset collection presented in this article is a subset of the published datasets as we focus only on local CH in our Region. Looking at Table 1, it is evident that datasets differ in size and topics, covering all the aspects of CH, i.e., tangible and intangible heritage, natural heritage, and food and wine. They also cover other topics relevant for citizens, such as companies, associations, and geographical information in the Campania Region. The same topic is modeled in different areas of the Campania Region,

such as itineraries, and points of interest (POI), to guarantee a wider geographical coverage.
P_3 - **High-quality OD.** As made evident by the CMP. column of Table 1, the completeness percentage of the HETOR datasets is overall very high. Only in 10 out 87 cases, the percentage is lower than 75% of the dataset. It is worth clarifying that the reported percentage count non-empty cells. In some datasets, authors explicitly report *missing information* that does not affect the reported value. Moreover, according to the ACC. column of Table 1, the accuracy score of the geographical information is very high. It is always less than 70% in only 4 out of 87 datasets. It means that published datasets can be considered high-quality data.

L_1 - **Tabular OD.** All the authored datasets are published as CSV. They are the best way to publish independent datasets, not yet interlinked. Modeling data as tables forces the data publisher to represent all the entries with the same structure, causing empty values for not applicable columns or the use of lists in a single cell. By exploiting the Semantic Web technologies, any entry can be modeled with an arbitrary number of relations.
L_2 - **No uniform schema.** The datasets differ for schema, in terms of the amount and the type of modelled columns, and lack a uniform terminology in the column headers. Before modeling a unified schema, it is suggested to carefully check the datasets' content to avoid modeling columns that are declared as headers, but contain no data.
L_3 - **Inaccurate values due to manual input.** The datasets are manually curated. Hence, typos, improper use of apostrophes as accents, and misspelled words are common errors. It causes the deficiency observed in the datasets accuracy. Moreover, string facets in Open Refine detected non-uniform use of lower and upper-case, switched letters, wide use of acronyms, and improper usage of apostrophes and accents.

5. Conclusions and Future directions

Since 2016, the HETOR project co-create OD with different communities (P_1) to digitize CH in the Campania Region. This effort resulted in 87 high-quality Open Datasets freely available on CKAN (P_2, P_3). Topics span from tangible and intangible CH, natural heritage, gastronomic curiosities, and information of public interest. This remarkable result is attributable to the effort of the HETOR project to propose structured activities built around the collaborative platform SPOD and a meticulous search for the data to be modeled to digitize CH of the Campania region. All the datasets are published as CSV attached to the Creative Commons License. Since different communities author them over time, they have no uniform schema (L_1, L_2). Published datasets might take advantage by proposing a uniform schema, such as an ontology, for each dataset group. Moreover, datasets are manually curated (L_3). Hence, they contain inaccurate values that can be easily corrected by automatic data quality approaches, such as clustering approaches to detect and correct typos, or by reconciling values with the ones published in well-known Knowledge Graphs, such as Wikidata. Further effort should be invested in quantifying the coherence and the coverage with respect to the covered topics.

References

[1] Open Knowledge Foundation, Defining open data, 2013. https://blog.okfn.org/2013/10/03/defining-open-data, [Online, Last access November 2022].

[2] A. Trentini, S. Scaravati, Raising curiosity about open data via the 'physiradio' musicalization iot device, Data Science Journal 19 (2020) 39. doi:10.5334/dsj-2020-039.

[3] L. Van Audenhove, W. Van den Broeck, I. Mariën, Data literacy and education: Introduction and the challenges for our field, Journal of Media Literacy Education 3 (2020) 1–5. doi:10.23860/JMLE-2020-12-3-1.

[4] T. Coughlan, The use of open data as a material for learning, Educational Technology Research and Development 68 (2020) 383–411. doi:10.1007/s11423-019-09706-y.

[5] K. Shamash, J. P. Alperin, A. Bordini, Teaching data analysis in the social sciences: A case study with article level metrics, Open Data as Open Educational Resources (2015) 49.

[6] E. Tovar, N. Piedra, Guest editorial: open educational resources in engineering education: various perspectives opening the education of engineers, IEEE Transactions on Education 57 (2014) 213–219. doi:10.1109/TE.2014.2359257.

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

[8] C. Z. Dib, Formal, non-formal and informal education: concepts/applicability, in: AIP conference proceedings, volume 173, American Institute of Physics, 1988, pp. 300–315. doi:10.1063/1.37526.

[9] N. Piedra, J. Chicaiza, J. López, E. T. Caro, A rating system that open-data repositories must satisfy to be considered OER: Reusing open data resources in teaching, in: Global Engineering Education Conference, 2017, pp. 1768–1777. doi:10.1109/EDUCON.2017.7943089.

[10] J. Álvarez Otero, M. Lázaro, M. JesusG, A cloud-based GiScience learning approach to spanish national parks, European Journal of Geography 9 (2018) 6–20. URL: http://hdl.handle.net/10612/10756.

[11] K. Charvat, O. Cerba, D. Kozuch, M. Splichal, Geospatial data based environment in INSPIRE4Youth, Procedia Computer Science 104 (2017) 183–189. doi:10.1016/j.procs.2017.01.101.

[12] R. R. Kurada, Y. Ramu, S. Pattem, Lessoning geospatial visualizations on real-time data, in: 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2021, pp. 1–6. doi:10.1109/CSITSS54238.2021.9683776.

[13] F. Windhager, E. Mayr, G. Schreder, M. Smuc, Linked information visualization for linked open government data. a visual synthetics approach to governmental data and knowledge collections, JeDEM-eJournal of eDemocracy and Open Government 8 (2016) 87–116. doi:10.29379/jedem.v8i2.436.

[14] R. De Donato, M. Garofalo, D. Malandrino, M. A. Pellegrino, A. Petta, Education meets knowledge graphs for the knowledge management, in: Methodologies and Intelligent Systems for Technology Enhanced Learning, 10th International Conference. Workshops, Springer International Publishing, Cham, 2021, pp. 272–280. doi:10.1007/978-3-030-52287-2_28.

[15] J. E. Weishart, Democratizing education rights, William & Mary Bill of Rights Journal 29 (2020) 1.

[16] I. Susha, Å. Grönlund, M. Janssen, Driving factors of service innovation using open government data: An exploratory study of entrepreneurs in two countries, Information polity 20 (2015) 19–34. doi:10.3233/IP-150353.

[17] I. Safarov, A. Meijer, S. Grimmelikhuijsen, Utilization of open government data: A systematic literature review of types, conditions, effects and users, Information Polity 22 (2017) 1–24. doi:10.3233/IP-160012.

[18] E. G. Martin, G. M. Begany, Opening government health data to the public: benefits, challenges, and lessons learned from early innovators, Journal of the American Medical Informatics Association 24 (2017) 345–351. doi:10.1093/jamia/ocw076.

[19] C. Baldwin, Using public sector open data to benefit local communities, Computer Weekly (2014) 17–20.

[20] M. Gascó-Hernández, E. G. Martin, L. Reggi, S. Pyo, L. F. Luna-Reyes, Promoting the use of open government data: Cases of training and engagement, Government Information Quarterly 35 (2018) 233–242. doi:10.1016/j.giq.2018.01.003.

[21] C.-P. Chen, J.-L. Shih, Y.-C. Ma, Using instructional pervasive game for school children's cultural learning, Journal of Educational Technology & Society 17 (2014) 169–182. URL: https://www.jstor.org/stable/jeductechsoci.17.2.169.

[22] A. Dickinson, M. Lochrie, P. Egglestone, Datapet: Designing a participatory sensing data game for children, in: Proceedings of the British Human-Computer Interaction Conference, 2015, p. 263–264. doi:10.1145/2783446.2783602.

[23] I. Vargianniti, K. Karpouzis, Using big and open data to generate content for an educational game to increase student performance and interest, Big Data and Cognitive Computing 4 (2020). doi:10.3390/bdcc4040030.

[24] M. Saddiqa, L. Rasmussen, R. Magnussen, B. Larsen, J. M. Pedersen, Bringing open data into danish schools and its potential impact on school pupils, in: Proceedings of the 15th International Symposium on Open Collaboration, 2019, pp. 1–10. doi:10.1145/3306446.3340821.

[25] M. Saddiqa, B. Larsen, R. Magnussen, L. L. Rasmussen, J. M. Pedersen, Open data visualization in danish schools: A case study, in: Proceedings of International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG), 2019. URL: http://hdl.handle.net/11025/35629.

[26] A. Antelmi, M. A. Pellegrino, Open data literacy by remote: Hiccups and lessons, in: Proceedings of the Symposium on Open Data and Knowledge for a Post-Pandemic Era (ODAK), BCS Learning & Development, 2022, pp. 1–5. doi:10.14236/ewic/ODAK22.7.