

# Empowering knowledge: the role of open data and open science in contemporary research – insights from the IODH platform

A. Ruoto<sup>a</sup>, A. Romano<sup>a</sup>, V. Santarcangelo<sup>a</sup>, S. Crisafulli<sup>a</sup>, G. Setzu<sup>a</sup>

<sup>a</sup> iInformatica Srl, Matera

## 1. Introduction

In the information age, open data plays a crucial role in promoting transparency, innovation, and knowledge sharing, especially within the scientific community. There is considerable discussion about Open Science and Fair Data, which aim to achieve technology transfer goals related to the deliverables and know-how from industrial research and experimental development projects conducted within the country, advancing the entire scientific community.

Open Science refers to the movement that promotes the accessibility and transparency of various aspects of scientific research. In contrast, Fair Data principles guide the management of data to ensure it is Findable, Accessible, Interoperable, and Reusable. To summarize, Open Science is a broad concept within the realm of science, whereas Fair Data provides specific criteria for making data useful and manageable [1].

To introduce the concepts of Open Data and interoperability, it is essential to discuss the 5-star model, devised by Tim Berners-Lee. This model assesses the quality and availability of public data on the Internet, with stars indicating the level of quality and usefulness. At the base level, data receive one star when they are made available online, regardless of the format. This represents the essential starting point: data are accessible, although they might be presented in unstructured formats such as PDF documents. To enhance usability, data should be organized in a structured, computer-readable format. This second level, represented by two stars, implies that the data are available in formats like Excel spreadsheets that allow for easy manipulation.

The next level, indicated by three stars, requires data to be in a non-proprietary format. This means that the data do not rely on specific software for reading or editing, which increases accessibility. An example of this level is the CSV file, which can be opened and edited with a range of tools without being tied to specific licenses or software. At the fourth level, with four stars, data are linked to other resources via URLs. This step enhances the utility of the data by integrating it with other datasets, thus improving interoperability and allowing the creation of links that enrich the information available.

The highest level, marked by five stars, is achieved when data are linked to other resources in a manner that enables a computer to understand the semantic meaning of those links. This top level of quality involves using technologies such as RDF (Resource Description Framework), which allows data to be interpreted in context and reused in a highly interoperable way [6].

In this context, the open data platform known as IODH (iInformatica Open Data Hub), developed by the private research center iInformatica, serves as a specific case study. The IODH platform aims to provide centralized and structured access to open datasets related to projects and research results. It enables researchers and the scientific community to access organized, high-quality data, fostering interdisciplinary collaboration and the participatory construction of new knowledge. The availability of this open data facilitates independent

verification of scientific results and contributes to creating a more inclusive and participatory research environment. The adoption of open science practices through platforms like IODH demonstrates how collective data sharing and analysis can lead to new discoveries and innovations [2].

## 2. The Caggiulino Project

The IODH platform features an intriguing dataset related to the Caggiulino multimedia project, which was launched during the Covid period immediately after the lockdown. This project aimed to introduce young audiences to the region of Basilicata, tied to the Lucanum project. Due to positive initial feedback, the project was tested in Sicily two months later, linked to the Sicanium project. The dataset provides valuable insights into the number of views obtained by the cartoons on YouTube and Facebook, which were promoted exclusively during the pandemic. It correlates YouTube and Facebook views with dissemination data on news outlets, cartoon structure (including voice-over and number of scenes), reference region, and the number of days since the lockdown began, reflecting the brand's growing popularity.

The Caggiulino project began in Basilicata with the release of the first cartoon on March 19, 2020, in Maratea. On May 17, 2020, the project expanded to Sicily with an episode exploring the Tower of Ligny and the Salt Pans of Trapani. In Sardinia, the project launched on December 26, 2020, featuring the Tomb of the Giants of Su Cuaddu 'e Nixias. The project then made its debut in Campania on February 20, 2021, with a cartoon showcasing the picturesque Sant'Angelo dei Lombardi. Throughout its run, the project experimented with voice-over in 18 cartoons.

Starting on July 8, 2020, a new concept based on postcards was introduced, featuring single scenes without interactions or dialogues with other characters. The first postcard was set in Marzamemi, Sicily, a historic fishing village known for its traditional tuna fishery.

The dataset, available in CSV format on the IODH platform, is suitable for analysis using WEKA, a machine learning software that provides various algorithms for classification, regression, clustering, and feature selection, with a user-friendly interface for pattern analysis and evaluation [3].

As an example, a data mining task performed with WEKA on this dataset involved extracting potential correlations. The initial step was to remove the cartoon name label, as it was deemed irrelevant. The Ranker algorithm was then used to analyze the dataset, with "YouTube views" defined as the dependent variable ( $y=f(x_1, x_2, x_3, \dots)$ ). This analysis revealed the significance of four independent variables: dissemination, presence of voice-over, reference region, and number of scenes. Contrary to expectations, there was no direct correlation with the number of days since the lockdown (which could indicate the timing of cartoon launches and brand notoriety), cartoon duration, or cartoon type (postcard vs. cartoon).

The number of YouTube views was categorized into five "bins": below 493 (34 videos), between 493 and 944 (12 videos), between 944 and 1394 (3 videos), between 1394 and 1845 (3 videos), and above 1845 (2 videos) [4]. Notably, the presence of a voice-over appears to significantly impact view counts, as illustrated in Figure 1.

No.	Label	Count	Weight
1	'(-inf-493.6]'	34	34
2	'(493.6-944.2]'	12	12
3	'(944.2-1394.8]'	3	3
4	'(1394.8-1845.4]'	3	3
5	'(1845.4-inf)'	2	2

Class: Voce fuori campo (Nom) Visualize All

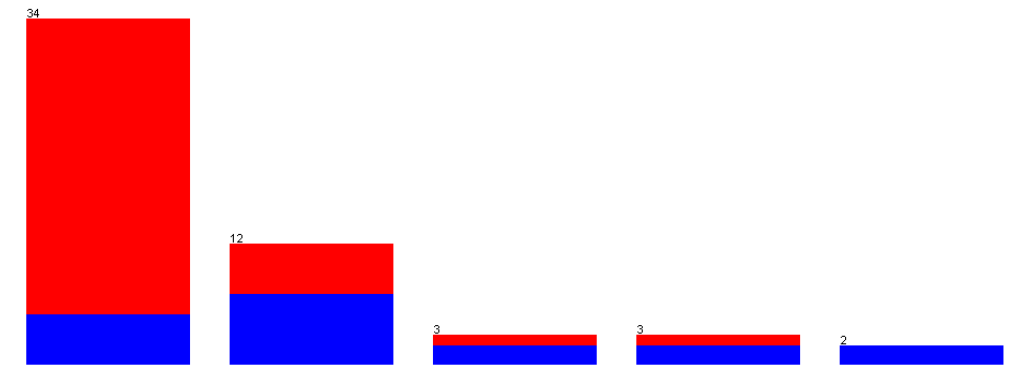


Figure 1 – Weka

### 3. Application developments

We also asked generative artificial intelligence (ChatGPT) to evaluate the novelty of the Caggiulino project and the potential of the associated dataset [5]. ChatGPT's response highlighted the following points:

The “Caggiulino” project, launched during the COVID-19 lockdown, is notable for its originality and innovative approach to cultural dissemination aimed at children. Let’s delve into the novel aspects and potential of the dataset linked to this project.

One of the key innovations of the “Caggiulino” project is its contextualization and adaptation to the COVID-19 era. Created to address the need for educational and entertaining content during a time of widespread isolation, the project introduced a novel method of using cartoons to educate young children about Italian geography and culture. This approach was particularly innovative as it merged digital storytelling with educational objectives in a historically unprecedented context.

Another distinctive aspect of the project is its regional approach. “Caggiulino” did not confine itself to exploring just one area but gradually expanded its narrative across several Italian regions, including Basilicata, Sicily, Sardinia, and Campania. Each region was showcased through its iconic landmarks, offering a narrative that highlights local cultural and natural heritage. This approach helped foster a sense of belonging and identity among young viewers, allowing them to discover the unique characteristics of various Italian regions.

The project's multichannel dissemination strategy also played a crucial role in its success. Distributing content through platforms such as YouTube and Facebook during the lockdown was particularly effective, enabling the project to reach a broad audience. These channels, already popular among children and families for consuming content, proved to be ideal for ensuring maximum visibility for the project.

From a narrative perspective, "Caggiulino" has explored various storytelling techniques to engage its young audience. The introduction of voice-overs in 18 episodes has added depth to the narrative, while the shift to single-scene postcards without dialogues represents an innovative approach to storytelling. These strategies have successfully captivated children's attention, offering them fresh and engaging content.

The dataset available on the IODH platform provides substantial opportunities for data analysis across several domains. It facilitates marketing and cultural promotion analyses by examining how factors such as media dissemination, the presence of voice-overs, and the number of scenes affect views and audience engagement. This information is vital for optimizing marketing strategies not only for "Caggiulino" but also for similar projects.

Analyzing correlations between various variables in the dataset can shed light on which narrative and technical elements enhance content effectiveness and engagement. This insight can guide the creation of future educational content, improving its quality and impact. Additionally, the dataset offers the chance to study the evolution of "Caggiulino" brand awareness over time. Although preliminary analysis did not show a direct correlation between the number of days since the lockdown and views, further investigation could uncover trends related to brand awareness, especially with its gradual regional expansion.

The dataset also holds potential for developing predictive models using machine learning techniques. Once enriched and further processed, it could enable predictions about the future performance of new episodes based on the characteristics of the cartoons and historical project variables.

Insights from the dataset analysis could contribute to research in communication and media, offering new perspectives on digital content consumption during crisis situations and how narrative adaptations can influence project effectiveness. The "Caggiulino" project is an innovative initiative that addresses the educational and cultural needs of children during a critical period. The dataset, with its rich information, has great potential to enhance the understanding of the project's success and provide valuable insights for future digital cultural dissemination initiatives.

When reviewing the results obtained from WEKA, ChatGPT highlighted several key considerations. The analysis appears well-structured and consistent, though it is crucial to confirm the accuracy of the correlations. The choice of variables (media dissemination, the presence of voice-overs, reference region, and number of scenes) seems appropriate as they likely influence YouTube views significantly.

However, the lack of significant correlation with variables such as the number of days since the lockdown, cartoon duration, and cartoon type suggests that these factors may not directly impact video success or that their effects might be mediated by other variables not included in this analysis.

Validating the correlations is essential, as multicollinearity (where independent variables like the reference region could affect both media dissemination and the number of scenes) might distort results. Using the Ranker algorithm to prioritize variables by relevance is a good start, but additional statistical techniques, such as regression analysis, should be employed to ensure the robustness of observed correlations.

The discretization of views into five bins appears reasonable, yet it's important to ensure this choice does not introduce distortions or lose significant details. The placement of cut points between bins can affect results, particularly if data is unevenly distributed. Additionally, exploring potential confounders not considered in the initial analysis, such as production quality, other promotion channels, and audience demographics, could offer further insights.

In summary, while the current analysis is well-structured and the correlations seem valid, further investigation with additional statistical tools and consideration of potential confounders could provide a more comprehensive understanding of the dataset.

## 4. Conclusions

The paper also highlights how the use of open data enhances scientific research, making a wider dissemination of knowledge possible and encouraging the growth of a dynamic and sustainable research ecosystem. Furthermore, the benefits of participatory construction of new knowledge are discussed, underlining the importance of collaboration and transparency in scientific production processes. In conclusion, the IODH platform represents a significant example of how open data can transform the scientific research landscape and support informed decision-making. The integration of open data and open science emerges as a key factor for scientific progress and for the adoption of policies based on concrete evidence, demonstrating how transparency and active participation are essential to address contemporary challenges in an effective and sustainable way [7].

## References

1. Kitchin, Rob. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.
2. Begany, Grace M., and J. Ramon Gil-Garcia. "Understanding the actual use of open data: Levels of engagement and how they are related." *Telematics and Informatics* 63 (2021): 101673.
3. Kulkarni, Eshwari Girish, and Raj B. Kulkarni. "Weka powerful tool in data mining." *International Journal of Computer Applications* 975 (2016): 8887.
4. Singhal, Swasti, and Monika Jena. "A study on WEKA tool for data preprocessing, classification and clustering." *International Journal of Innovative technology and exploring engineering (IJITEE)* 2, no. 6 (2013): 250-253.
5. Xing, Yixun. "Exploring the use of ChatGPT in learning and instructing statistics and data analytics." *Teaching Statistics* 46, no. 2 (2024): 95-104.
6. Murphy, Fiona. "Open access, open data, FAIR Data and their implications for life sciences researchers." *Emerging Topics in Life Sciences* 2, no. 6 (2018): 759-762.
7. Neumaier, Sebastian, Axel Polleres, Simon Steyskal, and Jürgen Umbrich. "Data integration for open data on the web." *Reasoning Web. Semantic Interoperability on the Web: 13th International Summer School 2017, London, UK, July 7-11, 2017, Tutorial Lectures 13* (2017): 1-28.