CHeCLOUD—the Cultural Heritage Linked Open Data Cloud

Gabriele Tuozzo[1,*], Maria Angela Pellegrino[1] and Antonio Lieto[1]

[1]*Università degli Studi di Salerno, ITALY*

Abstract

Cultural Heritage (CH) data have become increasingly prominent within the Semantic Web, yet dataset discoverability remains limited due to fragmentation across platforms and a lack of a standard platform to give visibility to published data. This demo paper presents the main features of CHeCLOUD (Cultural Heritage Linked Open Data Cloud): the first domain-specific subcloud of the Linked Open Data Cloud specifically devoted to aggregate and enhance access to Knowledge Graphs (KGs) and ontologies related to CH. CHeCLOUD provides a centralized catalog of 192 curated CH KGs and ontologies and computes a FAIR score for each of them, relying on an automatic and periodic assessment of KGHeartBeat. CHeCLOUD currently offers RESTful APIs, metadata browsing, and interactive graph visualizations to support FAIR evaluation. Additionally, CHeCLOUD features a semi-automated submission pipeline that engages users and maintainers through GitHub-based workflows. CHeCLOUD aims to foster reuse, interoperability, and findability within the CH community, while offering a reusable infrastructure to support similar thematic hubs across other domains.

Keywords

Cultural Heritage, Knowledge Graphs, Linked Open Data, Ontologies, Subclouds, FAIR principles

1. Introduction

In recent years, Cultural Heritage (CH) data have gained increasing relevance within the Semantic Web, as evidenced by the rise of digital platforms offering data as Knowledge Graphs (KGs), Europeana[1], the Digital Public Library of America[2], the Hungarian National Library[3], and the National Library of Australia[4]. While these initiatives have significantly advanced the aggregation of CH data at national or continental levels, the absence of a unified, cross-geographic hub dedicated solely to CH KGs persists. This fragmentation limits dataset discoverability and compels users to navigate multiple repositories, such as the LOD Cloud[5], DataHub[6], GitHub[7], or Zenodo[8], each characterized by heterogeneous interfaces and metadata standards.

A further limitation is that these catalogs rarely provide integrated quality indicators, particularly regarding compliance with the FAIR principles [1], which are increasingly recognized as essential for ensuring Findability, Accessibility, Interoperability, and Reusability. Researchers interested in such assessments must therefore resort to external tools, such as O'FAIR [2], FOOPS! [3], or FAIR-Checker [4], introducing additional complexity into the discovery process.

CHeCLOUD [5] addresses this gap by introducing the Cultural Heritage Linked Open Data Cloud, a thematic sub-cloud within the LOD ecosystem specifically devoted to aggregating, documenting, and exposing CH-related KGs and ontologies. Pronounced "*keh-cloud*" (reflecting the Italian pronuncia-

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

*Corresponding author.

gtuozzo@unisa.it (G. Tuozzo); mapellegrino@unisa.it (M. A. Pellegrino); alieto@unisa.it (A. Lieto)

0009-0004-5108-1995 (G. Tuozzo); 0000-0001-8927-5833 (M. A. Pellegrino); 0000-0002-8323-8764 (A. Lieto)

[1]Europeana: https://www.europeana.eu
[2]Digital Public Library of America: https://dp.la
[3]The Hungarian National Library: https://oszk.hu/en
[4]The National Library of Australia: https://www.nla.gov.au
[5]LOD Cloud: https://lod-cloud.net
[6]DataHub: https://datahub.io
[7]GitHub: https://github.com
[8]Zenodo: https://zenodo.org

tion of che), it builds on the successful experience of other domain-specific sub-clouds, such as the Linguistic LOD Cloud [6] and the Life Sciences LOD Cloud [7], which have demonstrated the value of domain-focused aggregation for enhancing findability and interoperability. Beyond simple aggregation, CHeCLOUD incorporates automated FAIRness evaluations, performed via KGHeartBeat [8], to provide users with immediate and actionable insights into the quality of the indexed resources.

2. CHeCLOUD

CHeCLOUD is a domain-specific subcloud within the LOD ecosystem that indexes linked datasets, KGs, and ontologies relevant to the CH domain, covering tangible, intangible, and natural heritage in accordance with UNESCO's definition [9]. Its construction followed a three-phase methodology inspired by Systematic Literature Reviews, designed to ensure transparency and reproducibility. The process included: (i) the structured identification of CH datasets from the LOD Cloud[9] and complementary repositories, (ii) a FAIRness evaluation of the selected datasets using a mapping framework between data quality dimensions and FAIR principles, operationalized through KGHeartBeat [8], and (iii) a maintenance strategy involving periodic updates and user feedback mechanisms.

Out of the $1,658$ datasets originally listed in the LOD Cloud snapshot, 147 were initially identified as potentially relevant to CH. This collection was subsequently refined and expanded to 192 datasets through expert validation, manual resolution of broken links, and targeted searches in additional repositories (e.g., DataHub, Zenodo, and institutional portals). With this scale, CHeCLOUD currently ranks as the third-largest subcloud in the LOD Cloud ecosystem, following the Life Sciences and Linguistic subclouds.

While Lieto et al. [5] provide a detailed account of the construction methodology and present an in-depth FAIRness assessment of the datasets, this demonstration paper focuses on showcasing the practical features and interfaces of CHeCLOUD, illustrating its utility for different stakeholders, such as data providers, curators, and researchers. The platform is publicly accessible at http://isislab.it:12280/CHe-cloud , offering a navigable catalog enriched with FAIRness indicators. Its source code is openly available on GitHub at https://github.com/GabrieleT0/CHe-CLOUD, ensuring transparency and reproducibility, and it provides programmatic access through RESTful APIs[10], enabling integration with external applications, workflows, and third-party services.

2.1. Coarse-grain Exploration

CHeCLOUD offers a coarse-grained exploration interface that provides a high-level overview of the KGs and ontologies included in the catalog. Implemented as a visual **dashboard**, it aggregates metadata and FAIRness assessment results, supporting maintainers in monitoring the overall health and structure of the cloud. The dashboard displays a variety of statistical insights, such as the number of datasets per CH subcategory, the availability of SPARQL endpoints and RDF dumps, the most commonly adopted ontologies, license types, media formats for data dumps, and the availability of access mechanisms (e.g., SPARQL endpoints, RDF dumps), enabling analyses similar to those performed on the broader LOD Cloud [10, 11]. In addition, a visual synthesis of data quality shows FAIR principles scores via boxplots and details scores at the dataset level via an interactive tabular representation.

Figure 2 shows a snapshot of the dashboard, illustrating that CHeCLOUD is composed predominantly of datasets, with only 10% being ontologies. Approximately 70% of the datasets relate to tangible heritage, while natural heritage is the least represented category—an imbalance aligned with trends observed in related studies [12, 13]. FAIR scores range from 1 to 4, with a median above 2.5. Among the FAIR principles, accessibility shows the most variability, while findability is the most consistently curated. All principles, except interoperability, tend to reach high scores. This high-level view not only supports maintainers in identifying trends but also facilitates comparison and benchmarking. For

[9]LOD Cloud: https://lod-cloud.net
[10]CHeCLOUD APIs: https://github.com/GabrieleT0/CHe-CLOUD/tree/main/WebApp

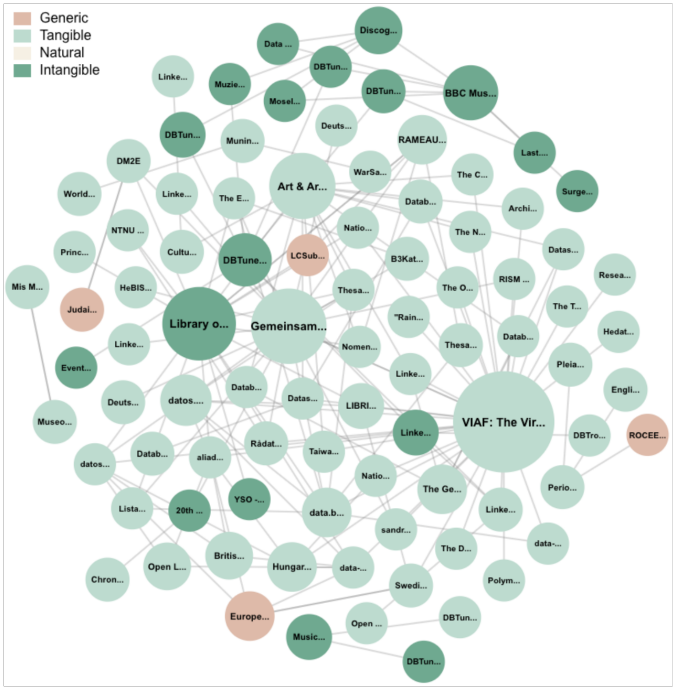

Figure 1: CHeCLOUD landing page

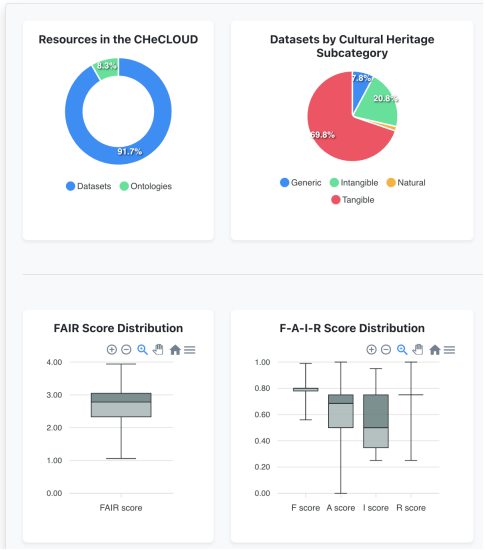

Figure 2: Snippet of the dashboard

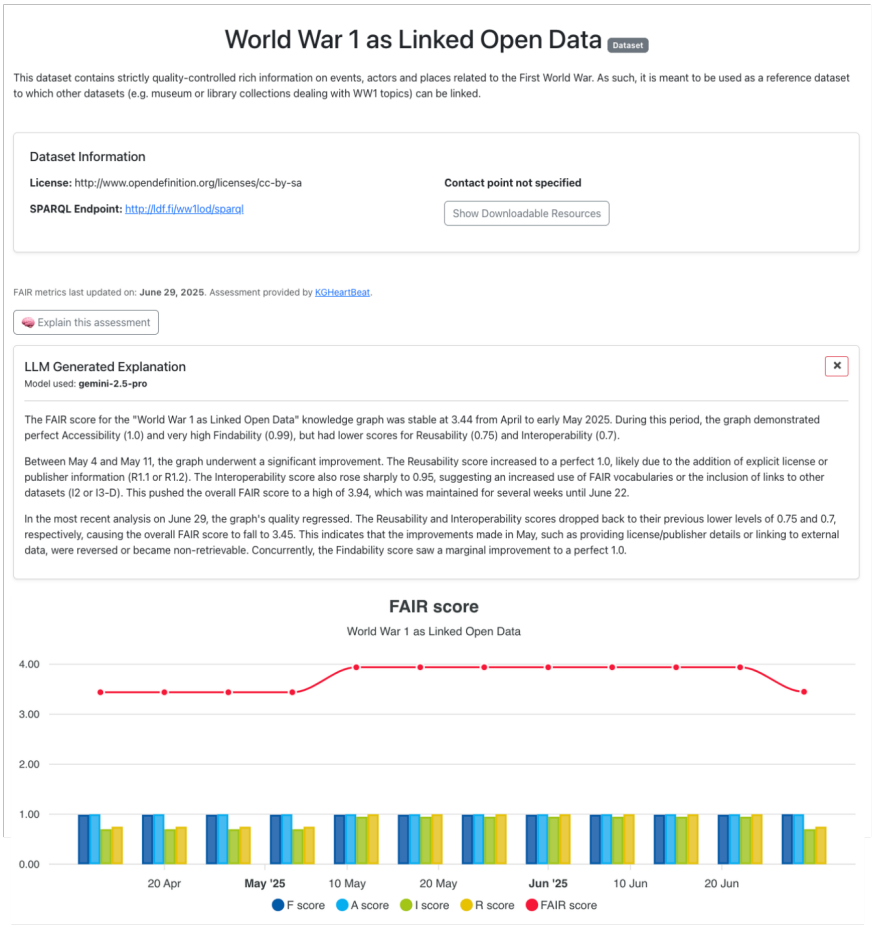

Figure 3: Metadata and temporal FAIRness scores for the *World War 1 as LOD*.

example, data providers can compare their datasets against others in the same thematic area to assess alignment with community standards or improve FAIRness. Likewise, cloud consumers can explore and compare datasets of potential interest based on topical relevance or specific data needs for their tasks.

2.2. Fine-grained Exploration

Resembling the well-known visualization provided by the LOD Cloud, the **CHeCLOUD landing page** displays catalogued datasets as an interactive graph. Figure 1 illustrates only the connected component, while isolated datasets, i.e., those not linked to others, are accessible in the online version. Nodes are color-coded according to CH subcategories (generic, tangible, intangible, natural), following the UNESCO classification, and their size reflects the number of incoming links (in-degree). In the online interface, hovering over a node highlights its connections and displays a tooltip with dataset details, e.g., its full name and in-degree. Moreover, users can locate specific datasets through a **keyword-based search**, as in Figure 4 (1), which matches terms found in the dataset's name, ID, description, or keywords.

Once a dataset is selected, users are directed to a metadata visualization page that includes both descriptive **metadata and FAIRness** metrics, automatically updated on a weekly basis by KGHeartBeat [8]. This page provides details such as dataset type (KG or ontology), license, description, access endpoints, and contact information for the maintainer or author. The FAIRness evaluation is presented, including (i) the overall FAIR score, (ii) individual scores for each FAIR principle, and (iii) a radar chart visualizing compliance across sub-principles (as in Figure 4 (2)). Tooltips allow users to view the full names and exact scores of each sub-principle, while an automatically created explanation of the FAIR assessment and suggestions for improvement are available on demand (as in Figure 4 (3)). Additionally, a temporal line chart visualizes the evolution of FAIR scores over time, accompanied by an automatically generated textual summary created using large language models (LLMs) (Figure 3). Visualizing the evolution of the FAIR score over time makes it possible to understand how updates to the data or metadata of the resource affect its overall FAIRness. This functionality is particularly valuable for dataset producers and maintainers, as it enables them to monitor potential issues introduced during dataset's updates.

2.3. Contribute to the Cloud

Contributors can support CHeCLOUD (i) by proposing the inclusion of new datasets, or (ii) by suggesting updates to the metadata of datasets already indexed in the catalog.

To streamline the **integration process**, CHeCLOUD employs a semi-automated submission and validation workflow. New datasets can be submitted through a dedicated metadata form. Upon submission, a LLM (currently Gemini 2.5 Pro, though interchangeable) automatically evaluates the dataset's relevance to the CH domain and assigns it to the appropriate subcategory. This classification is based on an analysis of the dataset's title, description, and keywords. A human reviewer subsequently validates both the metadata and the assigned classification to ensure accuracy. Once approved, the system creates a new branch and pull request (PR) in the CHeCLOUD GitHub repository, ensuring transparency and traceability. Contributors can track progress and engage with maintainers via the PR interface. The only manual step required of maintainers is the approval and merging of the request, after which the dataset is automatically synchronized and integrated into the catalog. Newly accepted datasets are also connected via REST APIs to KGHeartBeat, which calculates weekly quality metrics. FAIRness scores for all datasets, including new ones, are updated accordingly. Note that due to the weekly evaluation cycle, newly added datasets may appear in CHeCLOUD with their FAIRness scores after up to one week.

The same workflow applies to **metadata updates** for indexed datasets where the form is pre-filled with current metadata, allowing contributors to selectively edit the information, as in Figure 4 (3).

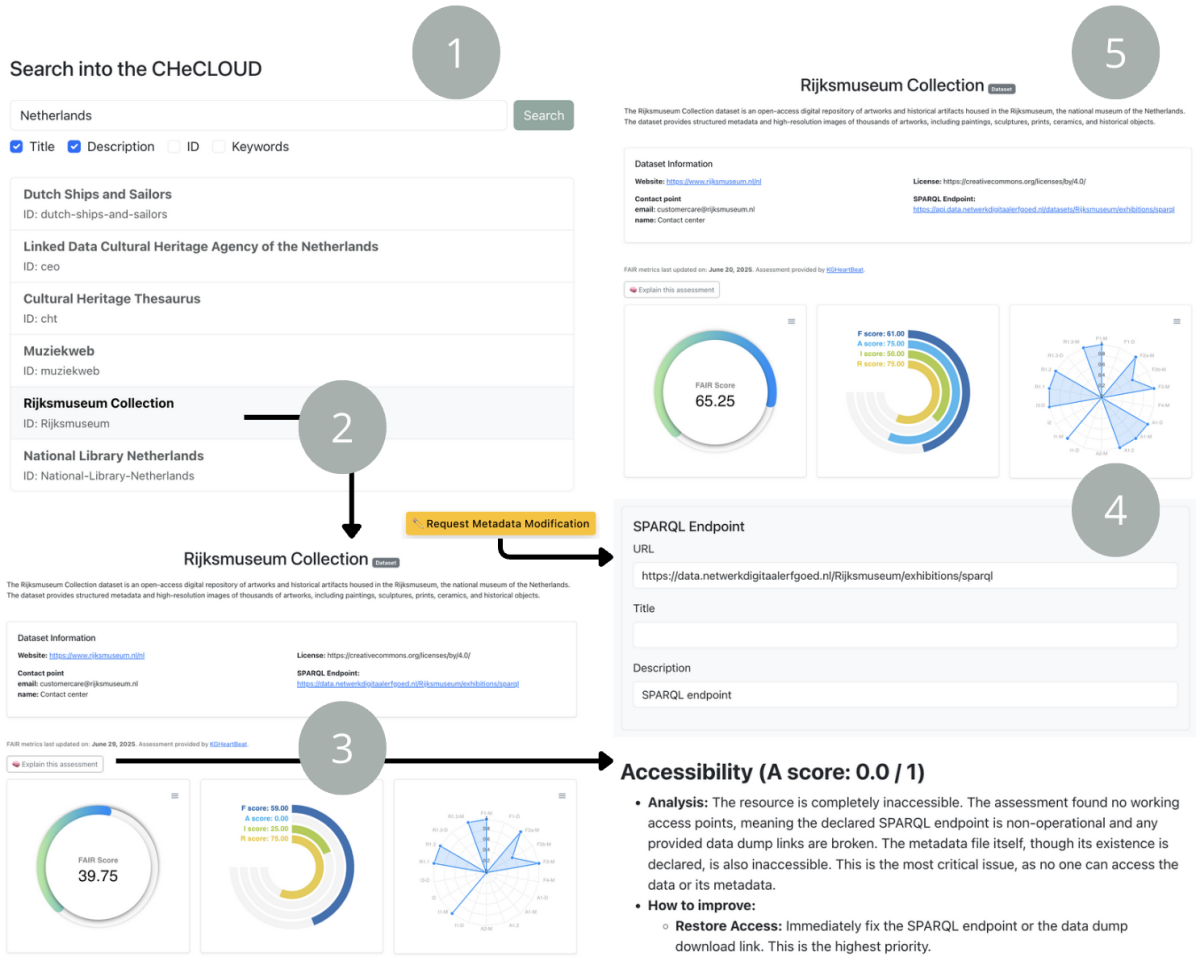

Figure 4: Console round-trip of a demonstration case. (1) The user searches for a dataset of interest, for example, the `Rijksmuseum Collection`, a well-known Dutch museum. (2) Upon selection, the user reviews its associated FAIRness scores. (3) Using Gemini 2.5 Pro, the user can request an explanation of these scores and is informed of a critical accessibility issue caused by a broken SPARQL endpoint. (4) The user manually submits a correction for the SPARQL endpoint. (5) Once the modification is approved and the next periodic assessment is executed (within one week), the updated FAIR score is made available.

3. Demonstration and Use Case

This section illustrates a sample workflow that will be demonstrated during the live session, showcasing how a user interacts with the CHeCLOUD platform. A recording of the demonstration is also available at: https://shorturl.at/6o0EZ.

Imagine the data steward of the Amsterdam Museum preparing a grant proposal to publish or reuse Linked Data describing their institution and its collection. To strengthen the proposal, the institution seeks to align with established best practices and reference successful examples in the domain. CHeCLOUD provides a valuable resource throughout this process:

- *Identifying relevant datasets.* The steward begins by accessing the **CHeCLOUD dashboard** (Figure 2) to explore datasets related to cultural heritage. By consulting the FAIR score table, they can perform a coarse-grained comparative analysis of datasets based on their overall FAIR score and individual principle scores. This allows them to discover high-quality datasets in similar domains, providing concrete references for reuse, alignment, or inspiration.
- *Evaluating their own dataset.* Returning to the **graph-based visualization** (Figure 1), the user searches for `Rijksmuseum` or a broader keyword such as `Netherlands` via the **search bar** (Figure 4 (1)). Upon selecting the dataset, the **detail page** (Figure 4 (2)) reveals that it currently has a

low FAIR score, particularly due to a missing score in the Accessibility dimension.

- *Identifying gaps and justifying improvements.* By using the **on-demand textual explanation** to understand the root causes of the dataset's FAIRness (Figure 4 (3)), and by inspecting the metadata directly within CHeCLOUD, the steward identifies a key issue: a missing or incorrect SPARQL endpoint. They propose a correction via the **metadata update workflow** (Figure 4 (4)). After approval by a maintainer and subsequent re-evaluation by KGHeartBeat, the updated FAIR score becomes visible (Figure 4 (5)). This not only improves the dataset's quality but also provides evidence of active curation, supported by collaborative contributions. Such transparency can strengthen the proposal's impact and help users make informed decisions.

This scenario demonstrates how CHeCLOUD supports stakeholders in benchmarking, improving, and promoting Linked Data datasets, ultimately contributing to better data quality and more informed decision-making within the CH sector.

4. Conclusion, Limitations and Future directions

This demo overviews CHeCLOUD, a catalog of CH KGs and ontologies, designed as a thematic subcloud within the LOD Cloud. It offers tailored interfaces for producers, consumers, and contributors, enabling FAIRness monitoring, metadata updates, and cloud exploration via dashboards and graph views. The infrastructure, available on GitHub, is reusable for creating subclouds in other domains. FAIRness assessments are automated through integration with KGHeartBeat.

Limitations and Future Work. At present, no comparative studies have been conducted to demonstrate superior performance and enhanced usability of this cloud in relation to other similar resource aggregators, such as the LOD Cloud. Furthermore, the entire process of FAIRness assessment currently relies exclusively on KGHeartBeat, and no comparison has been carried out with alternative tools that enable the evaluation of resource FAIRness.

Dataset discovery is currently semi-automated, relying on LLMs and metadata, currently limiting the platform's scalability and its long-term sustainability. Future plans include content-based classification, advanced search with adjustable FAIR weights, and improved explainability with corrective suggestions.

The newly indexed datasets and their updated metadata are not yet reflected in the LOD Cloud. However, this integration can be readily achieved, as CHeCLOUD supports the export of the entire catalog as a JSON file that adheres to the same structure as the original LOD Cloud, thereby facilitating seamless future incorporation.

Acknowledgments

This work was partially supported by the European Alliance NEOLAiA (Project 101124794: "NEOLAiA – Transforming Regions for an Inclusive Europe").

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT in order to: Grammar and spelling check.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The FAIR guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9. doi:10.1038/sdata.2016.18.

[2] E. Amdouni, S. Bouazzouni, C. Jonquet, O'FAIRe: Ontology fairness evaluator in the agroportal semantic resource repository, in: P. Groth, A. Rula, J. Schneider, I. Tiddi, E. Simperl, P. Alexopoulos, R. Hoekstra, M. Alam, A. Dimou, M. Tamper (Eds.), The Semantic Web: ESWC 2022 Satellite Events, Springer International Publishing, Cham, 2022, pp. 89–94. doi:10.1007/978-3-031-11609-4_17.

[3] D. Garijo, O. Corcho, M. Poveda-Villalón, FOOPS!: An ontology pitfall scanner for the FAIR principles 2980 (2021). URL: http://ceur-ws.org/Vol-2980/paper321.pdf.

[4] A. Gaignard, T. Rosnet, F. De Lamotte, V. Lefort, M.-D. Devignes, FAIR-Checker: supporting digital resource findability and reuse with knowledge graphs and semantic web standards, Journal of Biomedical Semantics 14 (2023) 7. doi:10.1186/s13326-023-00289-5.

[5] A. Lieto, M. A. Pellegrino, G. Tuozzo, The FAIRness of CHeCLOUD, the Cultural Heritege Linked Open Data Cloud, Semantic web (2025). URL: https://www.semantic-web-journal.net/content/fairness-checloud-cultural-heritage-linked-open-data-cloud.

[6] C. Chiarcos, S. Hellmann, S. Nordhoff, Linking Linguistic Resources: Examples from the Open Linguistics Working Group, 2012, pp. 201–216. doi:10.1007/978-3-642-28249-2_19.

[7] A. Hasnain, S. Sana E Zainab, M. Kamdar, Q. Mehmood, C. Warren, Jr, Q. Fatimah, H. Deus, M. Mehdi, S. Decker, A roadmap for navigating the life sciences linked open data cloud, 2014. doi:10.1007/978-3-319-15615-6_8.

[8] M. A. Pellegrino, A. Rula, G. Tuozzo, KGHeartBeat: An Open Source Tool for Periodically Evaluating the Quality of Knowledge Graphs, in: International Semantic Web Conference, Springer, 2024, pp. 40–58. doi:10.1007/978-3-031-77847-6_3.

[9] UNESCO, Cultural heritage, 2009. URL: https://uis.unesco.org/en/glossary-term/cultural-heritage, [Online, Last access April 2025].

[10] J. Debattista, C. Lange, S. Auer, D. Cortis, Evaluating the quality of the LOD cloud: An empirical investigation, Semantic Web 9 (2018) 859–901. doi:10.3233/SW-180306.

[11] T. Gabriele, Navigating the LOD Subclouds: Assessing Linked Open Data Quality by Domain, in: Companion Proceedings of the Web Conference, Association for Computing Machinery, New York, NY, USA, 2025. doi:10.1145/3701716.3717569.

[12] D. Monaco, M. A. Pellegrino, V. Scarano, L. Vicidomini, Linked open data in authoring virtual exhibitions, Journal of Cultural Heritage 53 (2022) 127–142. doi:10.1016/j.culher.2021.11.002.

[13] M. A. Pellegrino, V. Scarano, C. Spagnuolo, Move cultural heritage knowledge graphs in everyone's pocket, Semantic Web 14 (2023) 323–359. doi:10.3233/SW-223117.