Digital Libraries: supporting Open Science

Report on the 15th Italian Research Conference on Digital Libraries, IRCDL2019

Paolo Manghi, Leonardo Candela, Emma Lazzeri
ISTI - Consiglio Nazionale delle Ricerche
Pisa, Italy
{manghi, candela, lazzeri}@isti.cnr.it

Gianmaria Silvello
University of Padova
Padua, Italy
silvello@dei.unipd.it

1. INTRODUCTION

The Italian Research Conference on Digital Libraries (IRCDL) is the annual Italian forum to discuss research topics on Digital Libraries and related technical, practical, and social issues. Along the years, IRCDL touched several aspects underlying the "Digital Library" domain and promptly adapted to the evolution of the field. Today, the "Digital Library" field includes theory and practices reflecting the evolution of the role of libraries in the scholarly communication domain, and also embracing scholarly communication and open science.

The theme of IRCDL 2019 was "Digital Libraries: Supporting Open Science". Three main reasons motivated this theme: (*i*) science is increasingly becoming digital, meaning that research is performed using data services and digital tools; (*ii*) the results of the research are no longer just traditional scientific publications; (*iii*) the outcomes of science are increasingly encompassing datasets, software, and experiments. As digital artifacts, such products can be shared and re-used together with the article, thus enabling comprehensive research assessment and various degrees of reproducibility of science. Positive consequences of this shift towards Open Science are: accelerating science, optimizing the cost of research, fraud detection, and fully-fledged scientific reward. Digital Libraries are central in the evolution of research outputs by targeting findability, preservation, interlinking, and re-use of research products and by integrating the components of the scholarly communication process.

The conference has been organized in Pisa, and the proceedings are published in the Springer CCIS series Vol. 988 [22]. Pre-print versions, research datasets, and research software relative to the accepted contributions are accessible via Zenodo.org.[1]

[1] https://zenodo.org/communities/ircdl

2. CONFERENCE CONTRIBUTIONS

All submitted contributions were peer-reviewed by three of the thirty-two members of the Program Committee, and twenty-one were accepted, out of which six were short papers. IRCDL comprised of one invited speaker and six sessions.

Invited talk: Citation in the Era of Big Data and Open Source Software.

Prof. Susan B. Davidson, Weiss Professor at the Dept. of Computer and Information Science of the University of Pennsylvania, USA, discussed the most recent developments in data and software citation. Citations are the cornerstone of knowledge propagation in science and the principal means to assess the quality of research as well as to direct investments in science. We are transitioning towards the fourth paradigm of science where data and software are as vital to scientific progress as traditional publications are. Nevertheless, there is no viable computational method for citing data and software. Thus, to recognize the scientific contribution of developers, data scientists, data curators, and data centers and to estimate the value of data. Prof. Davidson presented the main challenges, and the solutions the database and digital library communities are supplying [11].

Open Science and Open Access. This session discussed on the issues originating from enacting Open Access and Open Science principles to the general public and the research world. Lana [19] advocated how Information Literacy needs Open Access, for the citizens to freely access high-quality information. Beamer [5] presented a methodology to optimize the embracing of Open Science practices in academic libraries. Fontanin [14] highlighted the Open Access-related barriers – e.g., technical infrastructures, points of access, digital and cultural di-

vide – making the information potentially available not just to researchers, but to everyone.

Open Science publishing and scientific workflows. The contributions in this session dealt with methodologies, practices, and tools in support of publishing workflows respecting Open Science principles. Latif [20] presented the work on EconStor, ZBW's Open Access Repository, to enrich attribution metadata by linking to external authority data sources. Dosso [12] described the "Learning to Cite" framework, for the creation of citation models to automatically cite XML files and its application with a process of transfer learning in the archival domain. Mizzaro [28] introduced an open-source software solution for the implementation of crowdsourcing Peer Review methodologies. Minelli [24] showcased the practical application of the open scientific life-cycle model proposed by the EcoNAOS (Ecological North Adriatic Open Science Observatory System) project. Bardi [4] illustrated a framework for the description, and peer review of research flows developed in the OpenUp project.

Text mining. Text mining techniques play a crucial role in Digital Libraries to automatically extract information used to serve user's needs better. Serra [26] proposed an approach to keyphrase extraction via an Attentive Model, a neural network designed to focus on the most relevant parts of data. Carducci [7] presented a system combining standard and semantic learning for automatically annotating bibliographic records. Pandolfo [25] described how they built the semantic layer of the Pisudski Institute of America digital archive. Ferilli [13] described the work performed to extend the BLA-BLA tool for learning linguistic resources by adding a Grammar Induction feature based on the advanced process mining and management system WoMan. Petrocchi [9] presented a study performed on Google Shopping to showcase how large search engines apply query steering depending on the user's profile.

Research Communities and Research Data. Research communities and the way they manage research data are increasingly becoming critical elements of digital libraries. Witt [31] presented the Repository Finder tool, designed to help researchers in the domain of Earth, space, and environmental sciences at finding the thematic repository they need based on a user-friendly wizard. Vezzani [30] presented TriMED, a digital library of terminological records designed to satisfy the information needs of different categories of users within the healthcare field. Castro described the results of two exploratory studies: in [27] the authors adopt a researcher-curator collaborative approach involving researchers in metadata description and discussing the use of generic and domain-oriented metadata; in [17] the authors analyze a data deposition workflow in CKAN using a Dublin Core metadata model for non-expert users. Luzi and Ruggieri [21] presented the OpenUp project pilot on research data sharing, validation, and dissemination in Social Sciences, intending to investigate the applicability of peer review and/or Open Peer Review to datasets in disciplines related to Social sciences.

Information retrieval and discovery. The relationship between information retrieval and discovery with digital libraries is long-standing. Fabris [1] presented a study exploring the relationships between SIGIR Information Retrieval articles from 2003 to 2017 with topics in the Digital Library domain. The goal is to identify trends and synergies between the two research fields. Amelio [2] showcased a study of the CAPTCHA usability which analyses the predictability of the solution time, also called response time, to solve the Dice CAPTCHA and suggested strategies towards the achievement of the "optimal" CAPTCHA. Tardelli [10] introduced on-demand tools provided by the `SoBigData.eu` research infrastructure for user-driven monitoring of Twitter data and publishing of the results as research data. Hast [16] described a training-free word spotting algorithm to mine images of digitized historical handwritten material to enable text search across the collection. Metilli [23] presented a case-study based on the Wikidata knowledge base exploring techniques to improve search functionalities by semi-automatically extracting narratives.

Applications. The last session included contributions about four application use-cases. Mannocci [18] presented DOIBoost, a version of the CrossRef metadata collection enriched with ORCID and the Microsoft Academic Graph, and Unpaywall made public in Zenodo.org, together with the software required to generate it. Foufoulas [15] presented user interfaces included in the Research Community Dashboard service of OpenAIRE enabling users to fine-tune text mining algorithms over a 10M full-texts corpus. Bellotto and Bettella [6] illustrated the experience of extending the metadata model of the Phaidra repository (University of Wien) towards the MODS data model. Firmani and Nieddu [3] reported on the Codice Ratio project, deliver-

ing a system taking advantage of character segmentation to support paleographers with tools for the minimal-effort transcription of large medieval manuscripts from the Vatican Secret Archives.

3. CONCLUSION AND PROSPECT

The research activities and results presented at IRCDL2019 give a clear indication of how active and multifaceted Digital Library research is.

A panel of experts[2] was organized to start a dialogue aiming at identifying research directions. Digital Libraries have always supported two phases of science, namely sharing of "mature" research products and discovery of published research products. Open Science has *de facto* revolutionized this model that conceptually separated the production of science from the publishing of science. For example, Research Infrastructures offer services constituting the "digital laboratory" where scientists are executing their experiments while accessing and sharing their intermediate results with others.

Two decades ago, the DELOS Grand Vision of Digital Libraries challenges focused on "[. . . enabling] any citizen to access all human knowledge anytime and anywhere, in a friendly, multimodal, efficient and effective way, by overcoming barriers of distance, language, and culture and by using multiple Internet-connected devices" [29]. The advent of Open Science, together with the natural evolution towards digital science, has profoundly impacted on this vision. IRCDL2019 conference has widely proven this statement, by highlighting strong interests in connecting digital library methods, tools, and services with thematic services for science and Open Science challenges. The current scenario, although addressing the urgent requirements of digital science (e.g. big research data, data-intensive science, multi-disciplinarity), suffers from the downsides arising when solutions originate from spontaneous initiatives rather than overarching engineering. The scholarly record is today kept in highly distributed and poorly connected sources, operated by publishers, research infrastructures, and institutions, adhering to heterogeneous publishing workflows, publishing best practices, and standards.

As remarked by Dr. C. Thanos in the final conference panel on the Future of Digital Library research, digital library research should envision "a world in which all scientific literature, data and other research outcomes are on-line, open and interoperable [. . . and seek for . . .] the creation of discipline-specific and interdisciplinary interconnected scholarly information spaces [. . . altogether forming a global . . .] Scholarly Record". Literature, datasets, software, and other digital assets of science should reside in resource-specific digital libraries (archives, repositories, databases), intended as active nodes in scholarly infrastructures [8]. To this aim, Digital Libraries should act as critical elements of Research Infrastructures and Open Cyber-Scholarly Communication Infrastructures, therefore flexibly adapt to support scientific communities at performing and publishing science by managing any research asset. In summary, Digital Libraries have upgraded their vest, their original intent, and are evolving to serve different actors. They should ambitiously act as an enabling service between scientists performing science, scientists publishing science, scholars, and scientists discovering scientific results, innovators accessing science for industrial benefits, and officers in need of monitoring science.

[2] https://ircdl2019.isti.cnr.it/?page_id=371

Acknowledgments

We desire to thank all those who contributed to the event, especially our colleagues Costantino Thanos and Maristella Agosti for their advice. Special thanks are also due to the members of the program committee whose research experience largely contributed to making this conference an attractive venue and valuable experience. Our sincere gratitude goes to all participants, invited speakers and authors, whose enthusiasm and vision constitute the soul of this conference. This event was supported by OpenAIRE-Advance project (EU H2020 research and innovation program, grant agreement No. 777541).

4. REFERENCES

[1] M. Agosti, E. Fabris, and G. Silvello. On synergies between information retrieval and digital libraries. In Manghi et al. [22], pp 3–17.

[2] A. Amelio, R. Janković, D. Tanikić, and I. R. Draganov. Predicting the usability of the dice captcha via artificial neural network. In Manghi et al. [22], pp 44–58.

[3] S. Ammirati, D. Firmani, M. Maiorino, P. Merialdo, and E. Nieddu. In codice ratio: Machine transcription of medieval manuscripts. In Manghi et al. [22], pp 185–192.

[4] A. Bardi, V. Casarosa, and P. Manghi. Foundations of a framework for peer-reviewing the research flow. In Manghi et al. [22], pp 195–208.

[5] J. E. Beamer. Digital libraries for open science: Using a socio-technical interaction network approach. In Manghi et al. [22], pp 122–129.

[6] A. Bellotto and C. Bettella. Metadata as semantic palimpsests: The case of PHAIDRA@UNIPD. In Manghi et al. [22], pp 167–184.

[7] G. Carducci, M. Leontino, D. P. Radicioni, G. Bonino, E. Pasini, and P. Tripodi. Semantically aware text categorisation for metadata annotation. In Manghi et al. [22], pp 315–330.

[8] D. Castelli, P. Manghi, and C. Thanos. A vision towards scientific communication infrastructures - on bridging the realms of research digital libraries and scientific data centers. *Int. J. on Digital Libraries*, 13(3-4):155–169, 2013.

[9] V. Cozza, V. T. Hoang, M. Petrocchi, and R. De Nicola. Transparency in keyword faceted search: An investigation on google shopping. In Manghi et al. [22], pp 29–43.

[10] S. Cresci, S. Minutoli, L. Nizzoli, S. Tardelli, and M. Tesconi. Enriching digital libraries with crowdsensed data. In Manghi et al. [22], pp 144–158.

[11] S. B. Davidson, P. Buneman, D. Deutch, T. Milo, and G. Silvello. Data Citation: A Computational Challenge. In *Proc. of the 36th ACM PODS 2017*, pp 1–4, 2017.

[12] D. Dosso, G. Setti, and G. Silvello. Learning to cite: Transfer learning for digital archives. In Manghi et al. [22], pp 97–106.

[13] S. Ferilli and S. Angelastro. Towards a process mining approach to grammar induction for digital libraries. In Manghi et al. [22], pp 291–303.

[14] M. Fontanin and P. Castellucci. Water to the thirsty reflections on the ethical mission of libraries and open access. In Manghi et al. [22], pp 61–71.

[15] T. Giannakopoulos, Y. Foufoulas, H. Dimitropoulos, and N. Manola. Interactive text analysis and information extraction. In Manghi et al. [22], pp 340–350.

[16] A. Hast, P. Cullhed, E. Vats, and M. Abrate. Making large collections of handwritten material easily accessible and searchable. In Manghi et al. [22], pp 18–28.

[17] Y. Karimova, J. A. Castro, and C. Ribeiro. Data deposit in a ckan repository: A dublin core-based simplified workflow. In Manghi et al. [22], pp 222–235.

[18] S. La Bruzzo, P. Manghi, and A. Mannocci. Openaire's doiboost - boosting crossref for research. In Manghi et al. [22], pp 133–143.

[19] M. Lana. Information literacy needs open access or: Open access is not only for researchers. In Manghi et al. [22], pp 236–247.

[20] A. Latif, T. Borst, and K. Tochtermann. Collecting and controlling distributed research information by linking to external authority data - a case study. In Manghi et al. [22], pp 331–339.

[21] D. Luzi, R. Ruggieri, and L. Pisacane. The openup pilot on research data sharing, validation and dissemination in social sciences. In Manghi et al. [22], pp 248–258.

[22] P. Manghi, L. Candela, and G. Silvello, editors. Proc. of the *15th Italian Research Conference on Digital Libraries, IRCDL 2019*, CCIS 988, Springer, 2019.

[23] D. Metilli, V. Bartalesi, C. Meghini, and N. Aloia. Populating narratives using wikidata events: An initial experiment. In Manghi et al. [22], pp 159–166.

[24] A. Minelli, A. Sarretta, A. Oggioni, C. Bergami, and A. Pugnetti. A practical workflow for an open scientific lifecycle project: Econaos. In Manghi et al. [22], pp 209–221.

[25] L. Pandolfo, L. Pulina, and M. Zieliński. Exploring semantic archival collections: The case of piłsudski institute of america. In Manghi et al. [22], pp 107–121.

[26] M. Passon, M. Comuzzo, G. Serra, and C. Tasso. Keyphrase extraction via an attentive model. In Manghi et al. [22], pp 304–314.

[27] J. Rodrigues, J. A. Castro, J. R. da Silva, and C. Ribeiro. Hands-on data publishing with researchers: Five experiments with metadata in multiple domains. In Manghi et al. [22], pp 274–288.

[28] M. Soprano and S. Mizzaro. Crowdsourcing peer review: As we may do. In Manghi et al. [22], pp 259–273.

[29] C. Thanos and V. Casarosa. The key role of the DELOS Network of Excellence in establishing Digital Libraries as a research field in Europe. *LIBEER Quarterly*, 26(4):296–307, 2017.

[30] F. Vezzani and G. M. Di Nunzio. Computational terminology in ehealth. In Manghi et al. [22], pp 72–85.

[31] M. Witt, S. Stall, R. Duerr, R. Plante, M. Fenner, R. Dasler, P. Cruse, S. Hou, R. Ulrich, and D. Kinkade. Connecting researchers to data repositories in the earth, space, and environmental sciences. In Manghi et al. [22], pp 86–96.