GUIDELINES FOR DATA SHARING AND DATA CITATION IN SOCIAL SCIENCES AND HUMANITIES JOURNALS PERSPECTIVES AND INSIGHTS FROM THE COST ACTION ENRESSH

Marc VANHOLSBEECK, Tim ENGELS, Andreja ISTENIC STARCIC

Although science has become *data-intensive*, not much research has been conducted about data publication, and data citation in particular[1]. This is particularly true in the social sciences and the humanities (SSH). Hence, considering the centrality of journals and scholarly articles in the production, dissemination and assessment of research, a COST ENRESSH[2] task force decided at the occasion of the RESSH 2017 conference to focus its attention on the data provision, data sharing and data citation policies and guidelines of SSH journals, as well as on authors' actual data citation practices.

This paper aims at reviewing the current state of *Open Research Data* (ORD) in the SSH, by focusing on the role played by European policy makers, researchers and publishers. We will then shortly present the analytical framework through which we are currently running a content analysis of SSH journals, in regards to their data sharing and data citation policies on the one hand, and to authors' data citations actual practices on the other side.

ORD state of play (policy makers, researchers, publishers)

It should be noted from the outset that *Open Research Data* is a multi-dimensional notion, with a variety of possible understandings[3]. Following Borgman[4], we

---

[1] G. Silvello, "Theory and practice of data citation", in *Journal of the Association for Information Science and Technology*, 69. 1, 2018, p. 6-20.

[2] COST ENRESSH Action aims at improving the understanding of social sciences and humanities (SSH) knowledge generation, the scientific and societal interactions in the different SSH disciplines and the patterns of dissemination in the SSH, in the perspective of supporting evidence based SSH policy making and evaluation processes.

[3] In this paper we will use the notion of Open Research Data (ORD) for designating the sharing of data free of charge for the end user, while the FAIR data principle relates to data that have to be Findable, Accessible, Interoperable and Re-usable, and follows the motto "as open as possible, as closed as necessary".

[4] C.L. Borgman, "The conundrum of sharing research data", in *Journal of the American Society for Information Science and Technology*, 63. 6, 2012, p. 1059-1078.

consider that sharing data follows four main rationales: (1) to reproduce or to verify research, (2) to make results of publicly funded research available to the public, (3) to enable others to ask new questions of existant data, and (4) to advance the state of research and innovation. These understandings may differ though by the arguments for sharing, by beneficiaries, and by the motivations and incentives of the diverse stakeholders involved.

European policy makers

While the Open Access movement had been initiated by librarians and researchers in the context of the so-called *Serials Crisis*[5], European policy makers appear to play a significant role in promoting the movement of making accessible the data underlying the research. The European Commission clearly intends to trigger changes in national Open Science policies of the Member States and Associated Countries, which are still mostly in their infancy in regards to ORD, whenever they exist[6].

There is no doubt that making research data accessible, replicable and reusable constitutes the logical extension of the European Research Area (ERA) project which was launched by the European Commission in 2000, aiming at creating an area in which research, scientific knowledge and technology circulate freely. Furthermore, ORD is at the crossroads of the "Digital Agenda for Europe" and the "Innovation Union", two important flagship initiatives constitutive of the Europe 2020 strategy from June 2010. The first one sets out an "open data" policy covering the full range of information produced by public bodies across the European Union (EU), while the second outlines the EU research and innovation policies and programmes.

In July 2012, the Commission integrated to one of its five ERA priorities the "optimal circulation, access to and transfer of scientific knowledge including via digital ERA - to guarantee access to and uptake of knowledge by all"[7]. At the same time, the Commission published its "Scientific Information Package" which followed and updated a communication from 2007 on the access, dissemination and preservation of scientific information in the digital age[8] and another one from

---

[5] J. Schöpfel, "Open access—the rise and fall of a community-driven model of scientific communication", in *Learned Publishing*, 28. 4, 2015, p. 321-325. DOI: 10.1087/20150413.

[6] ERAC SWG Open Science and Innovation, ERAC SWG Open Science and Innovation's assessment of the Amsterdam Call for Action on Open Science, 2018. http://data.consilium.europa.eu/doc/document/ST-1202-2018-INIT/en/pdf

[7] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions (17.07.2012). A Reinforced European Research Area Partnership for Excellence and Growth, 2012.

[8] European Commission, Communication from the Commission to the European parliament, the Council and the European Economic and Social Committee (14.02.2007). Scientific information in the digital age: access, dissemination and preservation, 2007.

2009 about ICT infrastructures for the e-science[9]. The 2012 package included the communication "Towards better access to scientific information: Boosting the benefits of public investments in research"[10]. The latter text clarifies the ORD understanding of the Commission: "The vision underlying the Commission's strategy on open data and knowledge circulation is that information already paid for by the public purse should not be paid for again each time it is accessed or used, and that it should benefit European companies and citizens to the full. This means making publicly-funded scientific information available online, at no extra cost, to European researchers and citizens via sustainable e-infrastructures, also ensuring long-term access to avoid losing scientific information of unique value." The package also included a recommendation on access to and preservation of scientific information[11] in which the European Commission recommends that Member States ensure that "datasets are made easily identifiable and can be linked to other datasets and publications through appropriate mechanisms, and additional information is provided to enable their proper evaluation and use", while underlining that "participants in multi-stakeholder dialogues" at national, European and/or international level "should in particular look at ways of linking publications to the underlying data."

This recommendation, in turn, has been updated in April 2018[12], focusing on the new developments in Open Science and Open Research Data, such as research data management, FAIR data (i.e. data that is Findable, Accessible, Interoperable and Re-usable), Text and Data Mining (TDM) and technical standards that enable re-use incentive schemes. This update is to be considered in the perspective of the European Open Science Cloud, a new major ORD e-infrastructure through which the Commission now intends to create "a trusted environment for hosting and processing research data to support EU science in its global leading role"[13]. In the *European Open Science Cloud* (EOSC) Declaration of October 2017, the Commission even calls for "considerable cultural change" towards opening research data and following the FAIR principles[14].

---

[9] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions (05.03.2009). ICT infrastructures for e-science, 2009.

[10] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions (17.7.2012). Towards better access to scientific information: Boosting the benefits of public investments in research, 2012.

[11] European Commission, Recommendation (17.7.2012) on access to and preservation of scientific information, 2012.

[12] European Commission, Recommendation (25.4.2018) on access to and preservation of scientific information, 2018.

[13] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud, accessed on 4 September 2018.

[14] https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf#view=fit&pagemode=none, accessed on 4 September 2018.

In the meantime, in 2016, the DG Research and Innovation of the Commission published Commissioner Carlos Moedas' vision for EU research and innovation, which reinforces EU support to Open Science, one of its five lines of potential policy action consisting in "mainstreaming and further promoting open access policies as regards both research data and research publications"[15]. The same year, the role of data citation in the bigger Open Science scheme had also been explicitly recognized by the Council in its Conclusions on the transition towards Open Science, emphasizing that "[proper data citation] will assist both the assessment of researchers and their projects and help to implement the findability, accessibility, interoperability and reusability of research data."[16]

Furthermore, as a research funder, the Commission has introduced an ORD pilot in Horizon 2020 in January 2017 and intends to make ORD mandatory – with possibilities of opting out for reasons relating to security, privacy or IPR – in the next framework programme, Horizon Europe (2021-2027).

Finally, it should be mentioned that the European Commission has also launched different expert groups in charge of providing policy advice in Open Science. In particular, the *Open Science Policy Platform* (OSPP) has been launched in 2016 with all the relevant actors involved in science and research in Europe. OSPP has recommended in April 2018 that "data resulting from publicly funded research must be made FAIR and citable, and be as open as possible, as closed as necessary."[17]

Researchers

Even though data archiving and data sharing have been a usual practice for a long time in some disciplines such as astronomy and genomics, in most others – and in a majority of SSH disciplinary fields - it is still not often that data are curated adequately so that they may be sustainably available for other researchers to be replicated or reused.

Current developments have to be taken into account though. At a collective level, scholars participate – together with other stakeholders like librarians, funders or publishers – in initiatives in support to ORD. Most noticeable and recent ones are CODATA (ICSU Committee on Data for Science and Technology), FORCE11, DataCite and Research Data Alliance (RDA). Those interest groups all seek to improve research data policy standards, data linking and citation, and request that

---

[15] Directorate-General for Research and Innovation (DGRTD), *Open Innovation, Open Science, Open to the World, A Vision for Europe*, Luxembourg, 2016. DOI: 10.2777/061652.

[16] Council, *Council conclusions (27/05/2016). The transition towards an Open Science system*, 2016.

[17] Directorate-General for Research and Innovation (DGRTD), *Open Science Policy Platform Recommendations*, 2018. DOI: 10.2777/958647.

the datasets be assigned the scholarly status of scientific reference while requiring proper citation standards[18].

Research has been conducted too in regards to individual scholars' perception of data sharing. Internationally there is an increased willingness to share data[19]. Formal data citation in the reference section of scholarly articles is considered by researchers across the sciences and the social sciences as the proper way to credit dataset creators, while citation counting is viewed as the most useful measure of impact, appropriate metadata being deemed as essential[20]. A survey of Global Environment researchers showed that funder policies are also considered as a crucial policy-related motivator[21].

Constraints and enablers of data sharing vary across disciplines though. SSH researchers, in particular, may have the feeling of missing computing, budgetary and personal resources[22]. ORD e-infrastructure (repositories) are not as developed in SSH as in STEM, although disciplinary archives, like the UK-based Archaeology Data Service (ADS), or non-specialist ones - like Dataverse or Figshare - are at the disposal of SSH researchers. At an epistemological level, there is a broader diversity of conceptions on what constitutes data in the SSH – and more particularly in the digital humanities -, in regards to disciplines, methods, equipment and scientific topics[23]. In science and SSH alike, researchers need a clearer perception of the benefits coming from data sharing and share similar

---

[18] See for ex. the Joint Declaration of Data Citation Principles from FORCE11 which has been endorsed by Wiley and Elsevier among many others (https://www.force11.org/datacitationprinciples).

[19] C. Tenopir, E.D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock and K. Dorsett, "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide", in *PLOS Open*, 10. 8, 2015. https://doi.org/10.1371/journal.pone.0134826; D. Stuart, G. Baynes, I. Hrynaszkiewicz, K. Allin, D. Penny, M. Lucraft and M. Astell, *PRACTICAL CHALLENGES FOR RESEARCHERS IN DATA SHARING White Paper*, 2018. https://doi.org/10.6084/m9.figshare.5971387.

[20] C. Tenopir, S. Allard, K. Douglass, A. Aydinoglu, L. Wu, E. Read, M. Manoff and M. Frame, "Data Sharing by Scientists: Practices and Perceptions", in *PLoS One*, 6. 6, 2011. https://doi.org/10.1371/journal.pone.0021101; J.E. Kratz and C. Strasser, "Researcher Perspectives on Publication and Peer Review of Data", in *PLoS One*, 10. 2, 2015. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117619.

[21] B. Schmidt, B. Gemeinholzer and A. Treloar, "Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey", in *PLoS One*, 11. 1, 2016. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0146695.

[22] J. Schöpfel, *Vers une culture de la donnée en SHS : Une étude à l'Université de Lille. Rapport de recherche*, 2018.

[23] J. Edmond, "Will Historians Ever Have Big Data? Theoretical and Infrastructural Perspectives", in *Computational History and Data-Driven Humanities: Second IFIP WG 12.7 International Workshop*, CHDDH 2016, Dublin, Ireland, May 25, 2016, revised Selected Papers 2, 2016, p. 91-105.

concerns about significant technological and operational barriers[24]. Researchers from all disciplines are also still unsure about the organization of data "in a presentable and useful way", copyrights, licensing and which repository to use[25].

Such an ambivalence in the perception of ORD is somehow mirrored in the individual scholars' practices. While there may be a general increase in actual data sharing behaviours[26], 40% of the 2300 respondents to the global Digital Science State of Open Data survey, which included PhD candidates and academics from a diversity of disciplines, still responded that they either "rarely" or "never" share their data[27]. Similarly, in the field of biodiversity science respondents appear to be unwilling to share primary data before publishing[28]. Furthermore, even if SSH researchers have manifested their interest for replication through data sharing in political science, economics, psychology, and quantitative sociology[29], ORD is still at a less advanced stage in SSH than in other sciences[30].

Publishers

Publishers have a central role to play in regards to advocating for – and encouraging - good data practices that take into account the different disciplinary usages, implementing open data policies, providing authors with guidelines and relevant information as well as with credit mechanisms which relate to data management and include data citation and linking[31]. Through ad hoc data validation or data reviewing processes, publishers may also assure the quality of the data underlying the published articles and adapt peer reviewers' guidelines accordingly.

Already in 2006, the Association of Learned and Professional Society Publishers (ALPSP) and the International Association of Scientific, Technical & Medical Publishers (STM) declared in a joint statement that "as a general principle, data sets

---

[24] X. Huang, B.A. Hawkins, F. Lei, G.L. Miller, C. Favret, R. Zhang and G. Qiao, "Willing or unwilling to share primary biodiversity data: results and implications of an international survey", in *Conservation Letters*, 5. 5, 2012, p. 399-406.

[25] Stuart et al., *Practical challenges*.

[26] Tenopir et al., "Changes in data sharing"; Stuart et al., *Practical challenges*.

[27] M. Hahnel, J. Treadway, B. Fane, R. Kiley, D. Peters and G. Baynes, *The State of Open Data Report 2017*, figshare, 2017. https://doi.org/10.6084/m9.figshare.5481187.v1

[28] X. Huang et al., "Willing or unwilling".

[29] S. Gherghina and A. Katsanidou, "Data availability in political science journals", in *European Political Science*, 12, 2013, p. 333-349; J. Ishiyama, "Replication, Research Transparency, and Journal Publications: Individualism, Community Models, and the Future of Replication Studies", in *Political Science & Politics*, 47. 1, 2014, p. 78-83; S. Vlaeminck and L.K. Herrmann, "Data Policies and Data Archives: A New Paradigm for Academic Publishing in Economic Sciences?", in B. Schmidt and M. Dobreva (ed.), *New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust*, Amsterdam, 2015, p. 145-155.

[30] A.M. Pienta, G.C. Alter and J.A. Lyle, *The enduring value of social science research: the use and reuse of primary research data*, 2010. http://hdl.handle.net/2027.42/78307; Tenopir et al., "Changes in data sharing"; Tenopir et al., "Data Sharing by Scientists"; Stuart et al., *Practical challenges*.

[31] M. Hahnel et al., *The State of Open Data Report 2017*.

[…] should wherever possible be made freely accessible to other scholars. We believe that the best practice for scholarly journal publishers is to separate supporting data from the article itself, and not to require any transfer of or ownership in such data or data sets as a condition of publication of the article in question"[32]. Since then, most major publishers have developed data sharing and data citation policies, such as the American Association for the Advancement of Science – *Science* getting an open data policy as soon as 2011 -, Springer Nature – with its standardized research data policies, research data support helpdesk and recommended repositories list -, Wiley, Taylor & Francis or Elsevier. Leaders in Gold Open Access publishing were soon to propose their ORD policies too, like PLOS, F1000 Research or BioMed Central. A Publishers Early Adopters Expert Group which included Elsevier, Springer Nature, PLOS, eLife Sciences Publications, Wiley and EMBO Press even developed a data citation roadmap for scientific publishers in 2017[33].

While it appears that there is no correlation between journals being Open Access and ORD policies, it has been shown that a higher impact factor correlates to open data and code policies in computational sciences[34], biomedicine[35] and Open Access journals[36].

Data journals have been launched too, dedicated to the exclusive publication of contextualized datasets, one of the first to be launched being *Scientific Data* in May 2014, by Nature Publishing Group. In comparison to other disciplines – and in particular as compared to health and life sciences – only a few SSH data journals are currently available[37]. Let's mention here the *Research Data Journal for the Humanities and Social Sciences*, published by Brill in collaboration with DANS, which publishes data papers describing the datasets and putting the data in context. The *Journal of Open Psychology Data* and the *Journal of Open Archaeology Data* from Ubiquity Press publish peer reviewed data papers in their respective fields. Combinations of data overlay certification platforms and peer- or community- review processes

---

[32] https://www.stm-assoc.org/2006_06_01_STM_ALPSP_Data_Statement.pdf, accessed on 4 September 2018.

[33] H. Cousijn, A. Kenall, E. Ganley, M. Harrison, D. Kernohan, F. Murphy, P. Polischuk, M. Martone and T. Clark, "A data citation roadmap for scientific publishers", in *bioRxiv*, 2017. https://doi.org/10.1101/100784.

[34] V. Stodden, P. Guo and Z. Ma, "Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals", in *PloS One*, 8. 6, 2013. https://doi.org/10.1371/journal.pone.0067111.

[35] N.A. Vasilevsky, J. Minnier, M.A. Haendel and R.E. Champieux, "Reproducible and reusable research: are journal data sharing policies meeting the mark?", in *PeerJ*, 5. https://doi.org/10.7717/peerj.3208.

[36] E. Castro, M. Crosas, A. Garnett, K. Sheridan and M. Altman, "Evaluating and promoting open data practices in open access journals", in *Journal of Scholarly Publishing*, 49(1), 2017, p. 66-88.

[37] L. Candela, D. Castelli, P. Manghi and A. Tani, "Data journals: A survey", in *Journal of the Association for Information Science and Technology*, 66. 9, 2015, p. 1747-1762.

have been experimented in the context of the Episcience project[38]. The OpenUP project which is currently running includes a pilot project of data journal in the humanities[39].

ORD state of play

In spite of the rising consideration of policy makers, researchers and publishers for opening research data, it is argued that there are still important problems of discoverability of datasets[40] and even that data citation systems are still in their infancy[41]. Journals research data policies, in particular, are "in critical need of standardization and harmonization"[42]. For example, the abovementioned *Science* ORD policy constitutes an "improvement over no policy, but [is] currently insufficient for reproducibility", findings being reproducible only for 26% of the sample[43]. Similarly, a majority of datasets in ecology and evolution journals with a strong ORD policy are incomplete or archived in a way that partially or entirely prevent reuse and reanalysis[44].

SSH journals with solid ORD policies are still in a minority[45]. In the field of archaeology, journals' editorial policies lack adequate enforcement. Although most of the data available at repositories are licensed to enable flexible reuse, only a small proportion of the data are stored in structured formats for easy reuse[46].

There are still no proper career incentives for researchers to cite and share data, the focus of research evaluation remaining on the articles and – to a lesser extent - books. Since citations to data are not commonly taken into account in bibliometric indicators, there is also a lack of incentives for publishers to engage in stronger

---

38 L. Romary, M. Mertens and A. Baillot, "Data fluidity in DARIAH – pushing the agenda forward", in *BIBLIOTHEK Forschung und Praxis*, 39. 3, 2016, p. 350-357.

39 E. Toli, E. Sifacaki, N. Manola, Y. Ioannidis, T. Ross-Hellauer, E. Görögh, M. Vignoli, V. Banelytė, P. Manghi and S. Woutersen-Windhouwer, "SIG Proceedings Paper in word Format", in *Proceedings of The 14th International Symposium on Open Collaboration, Paris, France, August 2018 (OpenSym'18)*, 2018. https://doi.org/10.1145/3233391.3233528.

40 H.A. Piwowar and T.J. Vision, "Data reuse and the open data citation advantage", in *PeerJ*, 1, 2013. https://peerj.com/articles/175.

41 G. Silvello, "Theory and practice of data citation".

42 L. Naughton and D. Kernohan, "Making sense of journal research data policies", in *Insights*, 29. 1, 2016, p. 84–89. DOI: http://doi.org/10.1629/uksg.284.

43 V. Stodden, J. Seiler and Z. Ma, "An empirical analysis of journal policy effectiveness for computational reproducibility", in *Proceedings of the National Academy of Sciences*, 115. 11, 2018, p. 2584-2589.

44 D.G. Roche, L.E. Kruuk, R. Lanfear and S.A. Binning, "Public data archiving in ecology and evolution: how well are we doing?", in: *PLoS biology*, 13. 11, 2015. https://doi.org/10.1371/journal.pbio.1002295.

45 M.B. Nuijten, J. Borghuis, C.L.S. Veldkamp, L.D. Alvarez, M.A. van Assen and J. Wicherts, "Journal Data Sharing Policies and Statistical Reporting Inconsistencies in Psychology", in: *Collabra: Psychology*, 3. 1, 31. DOI: http://doi.org/10.1525/collabra.102 2017.

46 B. Marwick and S.E.P. Birch, "A Standard for the Scholarly Citation of Archaeological Data as an Incentive to Data Sharing", in *Advances in Archaeological Practice*, 6. 2, 2018, p. 125-143.

ORD and data citation policies. Indeed, data citations do not contribute to the Impact Factor of journals, while the less prestigious ones may even consider data sharing as an extra burden to be put on their authors' shoulders.

ENRESSH work on ORD

Taking into account this current state of play of ORD in the SSH, we are conducting a quantitative content analysis of data related journals' guidelines – relating to data provision, sharing and citation – as well as a quantitative content analysis of actual authors' data citations. After the completion of a pilot study in the field of educational technology, we will address a diversity of SSH disciplines. In both cases, the sample is provided by the JUFO Finnish list of journals, excluding journals that are non-peer reviewed or non-empirical, without online information or written in a language the task force does not master.

The independent variables of our content analysis relate to journal description including its Impact Factor, if any. The dependent variables include a score for the data provision, data citation and data sharing policy of the journal. We also look at the provision of guidance or guideline on how to link the article to the related dataset(s).

The pilot study is planned for publication in the autumn of this year, while the interdisciplinary broader study will be conducted during 2019.