Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper

Toma Tasovac, Sally Chambers, Erzsébet Tóth-Czifra

DARIAH-EU
Digital Research Infrastructure
for the Arts and Humanities

Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper

Version 1.0

1. Introduction

Europe faces significant challenges when it comes to the effective and efficient digitisation of our cultural heritage. With over 80% of Europe's cultural heritage remaining to be digitised (Nauta, van den Heuvel and Teunisse, 2017), there is an urgent need to re-think the European digitisation strategy in order to enable Member States to more effectively work together[1] towards the European Commission's ambitious, but hardly achievable, target of all European cultural heritage being digitised by 2025. Furthermore, a large proportion of the 10-20% of Europe's cultural heritage that has already been digitised is rapidly approaching its 'sell-by' date. Being 'first generation' or 'legacy' digitisation, it is no longer of a high enough quality (Hill and Hengchen, 2019) for analysis using advanced digital humanities methods.

DARIAH, the Digital Research Infrastructure for the Arts and Humanities, enhances and supports digitally enabled research and teaching across the Arts and Humanities. It develops, maintains and operates an infrastructure that sustains researchers in building, analysing and interpreting digital resources. Digital (digitised and born-digital) cultural heritage is fundamental to our research. **Cultural Heritage Data *is* Humanities Research Data**. Without digital cultural heritage, undertaking humanities research with digital methods would be impossible. With an estimated 500,000 humanities researchers in Europe alone (Rossi, 2020), the digital Arts and Humanities community represents a significant community to consider in light of the European Commission's evaluation and possible revision of the *Commission Recommendation of 27 October 2011 on Digitisation and Online Accessibility of Cultural Material and Digital Preservation* (REC 2011/711/EU). It is for this reason that we would like to share a few additional reflections from the DARIAH Community in this position paper.

2. Cultural Heritage Data as Humanities Research Data

Even if we only focus on the digitisation needs of the digital arts and humanities research community in Europe, the opportunities offered by digital technologies for the culture heritage sector are manifold. Digitisation for digital humanities research will facilitate the **digital transformation in Cultural Heritage Institutions (CHIs)**, enabling them to move beyond creating siloed and static digital libraries towards

[1] https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-digitising-cultural-heritage

implementing **sustainable data curation workflows** (Padilla et al., 2019; Candela et al, 2020). Through the **dynamic extraction and curation** of research-driven digital cultural heritage datasets, these 'Collections as Data' can be made accessible online as **Findable, Accessible, Interoperable and Reusable (FAIR) humanities research datasets**. What is needed now, and what the follow-up to REC 2011/711/EU should include, is a recommendation to conduct a thorough evaluation of **the usability of existing digitised cultural heritage collections** for specific digital humanities methods (Van Strien, et al., 2020). In addition, **cultural heritage institutions need guidance on how to improve their digitised collections to meet the needs of 21st-Century digital scholarship** including how to navigate the legal challenges related to text and data mining (White, 2020).

Effective collaboration between the cultural heritage sector, where **cultural heritage professionals are recognised as essential partners in research**, the humanities research community and data, information and computer scientists is an essential condition for ensuring the possibility of excellent research and innovation in the years to come. This is a policy challenge for the European Commission as well considering that Cultural Heritage and Research are traditionally seen as two separate sectors. Yet it will be essential -- and equally beneficial to both sectors -- for the Commission to recognise the **potential for innovation presented by tighter cross-sectoral cooperation and alignment between Cultural Heritage and Research**. Cross-sectoral recommendations and cross-sectoral funding mechanisms should lead the way in reaffirming Europe as a powerful symbolic space in which tradition and innovation are inherently intertwined. Finally, a well-thought-through alignment between a revised REC 2011/711/EU and the flagship initiatives from the Research sector such as the ESFRI Research Infrastructures and the European Open Science Cloud (EOSC) could pave the way towards better interoperability and infrastructural consolidation.

3. Strengthening Europeana's Role in Research

Europeana, which REC 2011/711/EU recognises as "Europe's digital library, archive and museum," already collaborates with the Arts and Humanities research community. The roots of this collaboration go back to 2013 and the Europeana Cloud project which did extensive investigations around the needs of Humanities and Social Science research communities and looking at the research behaviour and practices (see, for instance, Benardou, et al., 2014). Europeana Cloud also served as the prototype infrastructure for Europeana Newspapers, a project that produced a first transnational full-text collection of digitised historic newspapers from Europe, greatly appreciated by the scholarly community[2] . The Europeana Cloud project also led to the development of Europeana Research and eventually a Europeana Research Community with the aim of developing a deeper insight into meeting researchers' needs regarding the use of digital cultural heritage.

Any follow-up to REC 2011/711/EU should recognise the importance of research communities for the future development of Europeana and encourage funding opportunities for shared activities with

[2] See e.g. https://reviews.history.ac.uk/review/1894 and
http://www.europeana-newspapers.eu/category/interviews-with-researchers/

research infrastructures such as DARIAH, CLARIN etc. It should also, in agreement with the 2018 Evaluation of Europeana (COM/2018/612 final), highlight **the urgency of continuously improving the quality of aggregated metadata *and* addressing the lack of actual content in Europeana as a major drawback for both researchers and the general public.** The revised REC 2011/711/EU should emphasise both of these aspects with an understanding that metadata, while being crucial for identifying resources, should never be perceived as the ultimate horizon of digitisation. Users need 'actual content' and Europe should lead the way in providing easy access to it.

4. Cultural Heritage as Open Data

Collaboration with cultural heritage institutions and professionals and establishing as open as possible data exchange protocols are cornerstones of responsible research workflows in the arts and humanities. This cooperation, however, is often constrained by structural, legal and technical challenges. A significant number of collections throughout Europe have been digitised but not made openly accessible due to unclear licensing conditions. To make things worse, some digitised materials are still not openly available to the public - despite being in the public domain (see, for instance Wallace and Euler 2020; and the Passenger Pigeon Manifesto: "a call to public galleries, libraries and museums to liberate our cultural heritage.") We recognise **the clearing of sharing rights and licensing best practices as an absolute priority to fulfil the promises of REC 2011/711/EU**. Institutions should clearly indicate in which cases they serve as the right holders and should follow the principle '**open by default, closed when necessary**' in their licensing policies. In addition, the revised recommendation should take into account the impact on digital cultural heritage of the General Data Protection Regulation (GDPR) which came into force following the publication of REC 2011/711/EU.

Several European organisations and consortia (APEF, CLARIN, Europeana, E-RIHS, IPERION) involved in the cultural heritage domain are already working with DARIAH on a **Heritage Data Reuse Charter** to establish principles and mechanisms for improving the use and re-use of cultural heritage resources by researchers. The follow-up to REC 2011/711/EU should support the creation of **a common environment based on the core principles of Reciprocity, Interoperability, Citability, Openness, Stewardship and Trustworthiness,** to help all the relevant actors to work together to connect and improve access to heritage data. For this to happen, it is essential for the Commission to address the **existing imbalance between the qualitative and the quantitative targets of the European digitisation efforts**. In this respect, recognising the diversity of European CHIs in terms of infrastructural and technological maturity and providing support to them to increase their capacities in this respect (for instance, implementing PID policies) is crucial.

We see another major conflict in the current policy landscape that seriously limits the impact of cultural holdings on scholarship and society: the Directive 2003/98/EC on the re-use of public sector information, known as **the PSI Directive, is quite permissive regarding the revenue models based on charging fees for the reuse of cultural heritage data**. The Directive includes specific provisions which

allow museums, archives and libraries to charge higher fees for the re-use of their holdings and allow them to conclude exclusive agreements for the digitisation of their material. **This is in sharp conflict with the objectives of both REC 2011/711/EU and the increasingly prevalent European and national open data mandates.** While recognising the importance of establishing stable and diversified sources of funding for European CHIs, we recommend that the existing business models be re-evaluated in view of the fundamental question as to whether the generated income from pay-to-access models matches their administrative and implementation costs. In accordance with the general European commitment to open data, the follow-up to REC2011/711/EU should recommend exploring alternative business models, and recognise that public-private partnerships do not always lead to transparent agreements and free, unlimited access to end users.

5. Monitoring the Digitisation of European Cultural Heritage

The follow-up to REC 2011/711/EU should make a strong push towards the democratisation of digitisation across Europe. CHIs should be encouraged to move beyond the idea of the digitisation of cultural heritage as a national project: **Europe needs cooperative, transnational, multilingual cultural heritage digitisation policies**. This is a matter of both principle *and* practice. As European citizens, we can truly reap the benefits of digitisation only if our digital collections are as diverse as Europe itself. Yet in practical terms, it still remains a challenge to know what has and what has not been digitised, what should be digitised, and by whom. Even within individual Member States, an exclusive focus on nationally recognised masterpieces and canonical works can lead to the suppression of other important regional and local materials, private collections and lesser-known resources in less-resourced languages. **Effective collaboration between Member States cannot be achieved without the participation of transnational research communities such as DARIAH and without a clear understanding of the transnational impact of digitised cultural heritage.**

This also means that a future European digitisation strategy will need to be rooted in the assessment of what we have and an examination of what needs to be done to improve both the quantity and quality of digitisation in the future. **The current state of digitisation in Europe is still not sufficiently quantified.** To what extent are the current digital collections truly *representative* of European cultural heritage as a whole? Without reliable statistical data, it will be practically impossible to work towards improving the representation of diverse linguistic and cultural communities in digital collections. In previous years, the Numeric (2007–2009) and Enumerate (2011–2014) projects were set up 'to create a reliable baseline of statistical data about digitisation, digital preservation and online access to cultural heritage in Europe'. Europeana subsequently continued this work, with the latest report published in Summer 2017, but this work needs to be continued and expanded. We believe that **effective collaboration between Member States in this field can be achieved only if more effort is invested in monitoring digitisation across Europe**. The revised REC 2011/711/EU should provide a mechanism for monitoring the actions taken in response to the Recommendation. The revised Recommendation itself should also be reviewed on a regular basis.

6. 2D Digitisation Remains Crucial

Digitisation has an expiry date. Increasingly 'First generation' or 'legacy' digitisation is no longer of a high enough quality for analysis using advanced digital humanities methods. While DARIAH fully supports the need for an increased emphasis on the application of 3D technologies in the area of cultural heritage, it is essential that the **revision of REC 2011/711/EU evaluates the current status of 2D digitisation.** While significant progress has been made in the digitisation of Europe's historical textual cultural heritage, particularly thanks to advances in Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR), continued investment in improving access to historical texts across the full range of Europe's languages (especially the lesser-resourced languages) is crucial. While the revised recommendation should certainly prioritise 3D technologies, **it is essential not to lose sight of the ongoing efforts needed to provide full-text access to Europe's documentary heritage.** Finally, we also need to remain critical of equating virtual and physical manifestations and experiences of cultural heritage by retaining the careful balance of being 'digitally-enabled and analogue-aware'.

7. Advanced Technologies: Artificial Intelligence

The European Commission's White Paper on Artificial Intelligence (February 2020) outlines the importance of a robust European strategy for the development and deployment of AI. However, it fails to recognise the significant innovation potential of AI for Europe's digital cultural heritage. The richness, diversity and multilingual nature of Europe's cultural heritage data is an ideal application area for both the advancement of AI as a research domain, and the acceleration of advanced access to Europe's digital cultural heritage. Major initiatives outside Europe, such as the Library of Congress' *Computing Cultural Heritage in the Cloud Project* and the UK's Living with Machines project are already demonstrating the potential of such advanced technologies for the Cultural Heritage Domain. For Europe to retain a leading position in the area of AI, the revision of REC 2011/711/EU needs **a strong emphasis on cultural heritage as a key application area for the advancement of AI**.

8. Digital Transformation through Capacity Building

The COVID-19 pandemic has clearly demonstrated how essential digital skills are. Within DARIAH, we recognise that our high-level of digital literacy has been a valuable asset for the resilience of our research community. Indeed, the COVID-19 pandemic has emphasised the importance of enabling the cultural heritage sector to seize the opportunities provided by digital technologies, for example, with CHIs extending the remote access to in-copyright digitised materials or developing digitisation-on-demand programmes. However, this digital transformation requires a sustained effort, which should not be underestimated. The **revision of REC 2011/711/EU should highlight the need for continued investments in capacity building in the area of digital skills for CHIs,** ideally in collaboration with the training and education offerings of digital humanities initiatives such as DARIAH-Campus.

DARIAH-EU
Digital Research Infrastructure
for the Arts and Humanities

References

Benardou, A., Dallas, C., & Dunning, A. (2014). From Europeana Cloud to Europeana Research: The challenges of a community-driven platform exploiting Europeana content. In Euro-Mediterranean Conference (pp. 802-810). Springer, Cham.

Candela, G., Sáez, M. D., Escobar Esteban, Mp., & Marco-Such, M. (2020). Reusing digital collections from GLAM institutions. Journal of Information Science. https://doi.org/10.1177/0165551520950246

European Commission, Directorate-General for Communications Networks, Content and Technology (2018) Report from the Commission to the European Parliament and the Council on the evaluation of Europeana and the way forward. COM/2018/612 final. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2018:612:FIN

Hill, M. J. and Hengchen, S. (2019) "Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study", Digital Scholarship in the Humanities, https://doi.org/10.1093/llc/fqz024

Mahey, M., Al-Abdulla, A., Ames, S., Bray, P., Candela, G., Chambers, S., Derven, C., Dobreva-McPherson, M., Gasser, K., Karner, S., Kokegei, K., Laursen, D., Potter, A., Straube, A., Wagner, S-C. and Wilms, L. with forewords by: Al-Emadi, T. A., Broady-Preston, J., Landry, P. and Papaioannou, G. (2019) *Open a GLAM Lab.* Digital Cultural Heritage Innovation Labs, Book Sprint, Doha, Qatar, 23-27 September 2019. https://glamlabs.io/books/open-a-glam-lab/

Nauta, G.J, van den Heuvel, W. and Teunisse. S. (2017) *D4.4. Report on ENUMERATE Core Survey 4*, Europeana DSI 2– Access to Digital Resources of European Heritage.

Padilla, T., Allen, L., Frost, H., Potvin, Sarah, Russey Roke, E., & Varner, S. (2019). *Final Report --- Always Already Computational: Collections as Data*. http://doi.org/10.5281/zenodo.3152935

Rossi, G. et al. (Ed.) (2020) *Supporting the Transformative Impact of Research Infrastructures on European Research*. Report of the High-Level Expert Group to Assess the Progress of ESFRI and Other World Class Research Infrastructures Towards Implementation and Long-Term Sustainability. Luxembourg, Publications Office of the European Union.

Van Strien, D., Beelen, K., Coll Ardanuy, M., Hosseini, K., McGillivray, B. and Colavizza, G. (2020), Assessing the Impact of OCR Quality on Downstream NLP Tasks (2020) ARTIDIGH 2020, Artificial Intelligence and Digital Heritage: Challenges and Opportunities, 22 - 24 February, 2020 - Valletta, Malta,

within the 12th International Conference on Agents and Artificial Intelligence - ICAART 2020: http://www.insticc.org/node/TechnicalProgram/icaart/presentationDetails/91690

Wallace, A., Euler, E. Revisiting Access to Cultural Heritage in the Public Domain: EU and International Developments. IIC 51, 823–855 (2020). https://doi.org/10.1007/s40319-020-00961-8

White, Ben. (2020). Research Libraries: How You Can Support Text and Data Mining. https://doi.org/10.5281/zenodo.3801114 and https://libereurope.eu/?s=TDM