European Commission

EUROPEAN

POLICYBRIEF

NEWS EYE

https://www.newseye.eu/ I @NewsEyeEU I newseye-communication(at)ml.univ-lr.fr

July 2020

INTRODUCTION: VALUING DIGITISED NEWSPAPERS AS DATA

Newspapers collect information about cultural, political and social events in a more detailed way than any other public record. Since their beginnings in the 17th century, they record billions of events, stories and names, in almost every language, every country, every day. Newspapers have always been an important medium for the dissemination of public and political opinions, literary works, essays and art. This thematic wealth sets them at the centre stage for anyone interested in European cultural heritage.

The importance of newspapers as cultural heritage is thus irrefutable, but whilst some progress has been made concerning *digitising* newspapers, it is their potential as *data* which opens up new possibilities for their exploration and analysis using digital methods.

The NewsEye project has prototyped an integrated platform of data, digital tools and methods for the exploration and analysis of digitised historical newspapers. It thus demonstrates the potential of developing a European-wide ecosystem for large-scale, transnational and multilingual analysis of digital cultural heritage. NewsEye offers a unique combination of state-of-the-art artificial intelligence applied to a rich and linguistically diverse historical newspaper data corpus with a set of humanities-friendly tools. This opens up the possibility for the use of digital methods by anyone seeking to use historical evidence to understand major social and cultural debates (policy, industry, civil society) and significantly deepens our understanding of European culture and history.

NewsEye paves the way for future research to be undertaken in the European Commission's Horizon Europe and Digital Europe programmes, bridging the gap between computer science, cultural heritage and digital humanities (and their funding streams).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299

What digital cultural heritage and digital humanities now need is in-depth interdisciplinary collaboration with state-of-the art computer science. This is no longer a nice-to-have, but essential. The development of the NewsEye project has proven the value and necessity of progressing toward opening the utility of historical newspaper data as a concerted effort combining expertise in digital cultural heritage, digital humanities and computer science.

POLICY RECOMMENDATIONS

The achievements of the NewsEye project in its first two years clearly demonstrate the strength of its cutting-edge interdisciplinary research collaborations, a model according to which Europe can remain a leader in digital scholarship in the humanities and computer science alike. Yet, these new developments also show how such approaches are only in their infancy and require funding for new (and coordination of existing) structural research programmes to progress towards maturity. Based on NewsEye's research outcomes to date, the strategic policy directions are recommended in the below sections.

POLICY AREA 1: CULTURAL HERITAGE DIGITISATION: 2D STILL MATTERS

Digitisation is not a 'once and for all' process. 'First generation' or 'legacy' digitisation is no longer of a high enough quality for analysis using advanced digital humanities methods. In addition to this, with over 80% of Europe's cultural heritage still to be digitised, the European Commission's ambitious goal of all European cultural heritage being digitised by 2025 remains a distant target. Before moving on to strategic priorities which focus on 3D technologies and tools (which we agree is extremely important) **we need to first revalue and reassess where we are with 2D digitisation.** Text is a privileged data format: written records give us access to the actual provenance of ideas and actions, allowing us to get beyond the technological and interpretive layers added when data is recreated. The NewsEye project has already demonstrated that we can do much better with historical newspapers; and thus it can be inferred that digitisation of other types of 2D heritage can be significantly improved as well. There is an urgent need to re-think the European digitisation strategy for the coming years, i.e. **first reassessing what we have and then improving the quantity and quality of that.**

Moreover, the current state of digitisation in Europe is still not sufficiently quantified. 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. Furthermore, the European Commission is currently undertaking an evaluation of *Recommendation (2011/711/EU) Digitisation and online access of cultural material and digital preservation*, the results of which are expected in late 2020.

NewsEye Contribution (Evidence & Analysis): NewsEye has been working with the digital newspaper collections of three national libraries as a test bed for where we are with historical text corpora in three different European countries, four different languages and a range of levels of resource and technology to demonstrate how advanced digitisation can be applied. Our work has proven both the value of this approach and the urgent need to sustainably prolong it. Whilst initiatives and projects like NewsEye are already taking place in several countries (in the UK for example who are working on building a national collection), what is lacking on a European level is a transnational corpus of historical texts and adequate methods to deal with them. Essentially, we need a European platform for analysing European historical cultural data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299

Recommendation: Work on historical text is not finished. If Europe is going to stay ahead in digital humanities, we still need to do work in regard to textual resources. **Ongoing sustainable investment in digitisation and the use of digitised material is needed** and we urge for the undertaking of a survey to provide important knowledge for evidence-based policy development for digitisation in Europe. More structural support is required to enable research which will push forward digitisation and digital literacy to the level needed. Finally, **the technology** we have developed within the NewsEye project **should be scaled up and applied to all national libraries as well as other cultural heritage institutions in Europe**, contributing to a European historical cultural data platform.

POLICY AREA 2: ACCELERATING ACCESS TO AND USE OF CULTURAL HERITAGE THROUGH ARTIFICIAL INTELLIGENCE (AI)

The application and advancement of AI built to work with the complexities of cultural data has an immense, **unrecognised (and therefore largely untapped) innovation potential** which could act as a catalyst for the digital transformation of Europe's cultural heritage. It is also a source of technological innovation in itself by creating challenging and complex use cases. Europeana is currently conducting a survey on the application of AI in relation to galleries, libraries, archives, and museums (GLAMs), which we believe is very necessary.

The European Commission's White Paper on Artificial Intelligence (February 2020) describes a basis for developments in AI which we wholeheartedly support. However, this report omits **cultural heritage and digital humanities as an application and innovation area with significant potential.** Due to the richness, diversity and multilingual nature of Europe's cultural heritage, its complexity as data and richness in context, investment in AI research and innovation related to this application area will boost Europe's competitive advantage and significantly contribute to ensuring that Europe is fit for the digital age.

NewsEye Contribution (Evidence & Analysis): NewsEye has shown the advantages of bringing machine learning, AI and state-of-the-art computer science closer to the digital humanities and cultural heritage. However, recent advances in machine learning and knowledge extraction have widened the gap between the technological possibilities available and their practical implementation. The Europeana project has been essential for stimulating the opening up of digital collections of Europe's rich cultural heritage. This valuable work is only a first, yet important step, towards extracting and analysing the knowledge still buried deep inside our cultural artefacts. Digitised newspapers, while an excellent use case to explore the potential of data science to unlock these embedded semantic layers, are only the tip of the iceberg when it comes to extraction of knowledge from our historical documentary heritage, whether it be written on clay tablets, papyrus, parchment or most recently on paper. In particular, machine readability of textual documents is an area where significant additional development is required: NewsEye has proven this case through its ground-breaking work starting with historical newspapers, but additional challenges and application areas yet to be tackled are numerous, including ones beyond textual documents.

Recommendation: We strongly urge the consideration and **inclusion of the unique challenges and opportunities provided by cultural data in AI research and policy**. This will bring cultural heritage up to a level of technological advancement that is required to bolster the semantic knowledge we talk about in the next section, and to ensure AI can meet the kinds of nuanced requirements these artefacts of the human record present.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299

POLICY AREA 3: DEEPENING CULTURAL UNDERSTANDING THROUGH SEMANTIC KNOWLEDGE EXTRACTION

In order to fully use the potential of digital technologies, we need to move on from basic digital text analysis of digital cultural heritage resources (e.g. newspapers, periodicals, literary works, etc.) towards enabling a deeper semantic understanding of European culture and its history (e.g. language, emotions, discourses and memory) through the application of advanced analytic methods such as historical sentiment mining, temporal-spatial analysis and linguistic processing, discourse mapping, event detection and big data visualisations.

NewsEye Contribution (Evidence & Analysis): Digitisation does not stop at the scanning of documents. The digitised historical documents need to be 'pre-processed' before any interpretation of them as social or cultural data can be undertaken. NewsEye has demonstrated significant results in the area of **Automatic Text Recognition, Layout Analysis and Article Separation**, which enrich the digitised newspapers with a foundational layer, including full text transcripts, along with text block and article separation information. Beyond this, NewsEye has significantly advanced the state of the art in **Semantic Text Enrichment** for historical documents**,** producing semantic annotations such as the cross-lingual identification, disambiguation and linking of mentions of named entities (e.g., person, locations and organisations), and the detection of events. This is used to improve document access and allow the advanced systematic analysis of the newspaper collections. In addition to this the project's work on **Dynamic text analysis** has shown how we can produce methods to automatically find topics, trends, viewpoints, and themes in the corpus being studied, both within a specified context and in comparison between contexts.

Recommendation: We call for the **move beyond static digital libraries towards a new way of creating computational analytical tools for digital data exploration**. More research and practice is needed in the domain of semantic knowledge exploration in order to develop our understanding of cultural heritage content more deeply.

POLICY AREA 4: DEMOCRATISING DIGITAL LITERACY FOR DIGITAL TRANSFORMATION

Until now, the digital transformation of European cultural heritage has been more of an evolution than a revolution. However, the latest advancements in artificial intelligence and data science will require to an ever-increasing extent that **discrete disciplines and areas of expertise be combined to deliver a coherent and multidisciplinary approach to digital transformation**. Therefore, NewsEye very much welcomes the joining up of innovation, research and culture in the portfolio of Commissioner Mariya Gabriel as a necessary, and comprehensive three-pronged approach.

NewsEye Contribution (Evidence & Analysis): Transnational research is constantly being undertaken and advocated for by NewsEye, bringing humanities and computer science groups together to do cross-lingual, transnational comparative research. The NewsEye project has championed inter- and cross-disciplinary methods in several ways, which has bolstered our understanding of the importance of such approaches. Firstly, with the incorporation of **digital humanities hackathons**, we have changed the way humanities scholars and computer scientists come together to co-create knowledge using digital tools. These innovative 'lab-like' environments, which can and should increasingly be opened up within Europe, model an effective workflow for building digital literacy across cultural actors to mutually enrich results. GLAM labs have started to develop all over the continent, and across GLAM sectors, a development that is crucial for our work in this field for cross learning, skills and knowledge transfer. GLAM labs generally present a strong methodology for the digital transformation of cultural heritage and digital humanities.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299

Secondly, we are creating **Jupyter Notebooks** on coding for analysis of newspaper data within the project acting as a scientific and pedagogical tool, even within disciplines where coding is not yet an established methodological norm. This, in addition to cross-institutional user workshops and study trips between partners to understand each other's work, has proved crucial for sharing of knowledge and skills.

> *"COMING FROM A TRADITIONALLY DISCOURSE-DRIVEN RESEARCH AREA THE COLLABORATION IN THE PROJECT FORCED ME TO THINK CLOSELY ABOUT WHAT THE DIGITAL TRANSFORMATION MEANS IN OUR FIELDS OF STUDY. THE NEWSEYE PROJECT HAS ALLOWED ME TO UNDERSTAND HOW IMPORTANT THE HUMANITIES' PARTICIPATION IN THE DEVELOPMENT OF AI IS – AND HOW IMPORTANT IT IS, TO HAVE TOOLS THAT SHOW HOW MACHINES ARE WORKING IN A VERY SIMPLE AND UNDERSTANDABLE WAY SO THAT THESE CAN BE MADE AVAILABLE FOR WIDER RESEARCH AND USAGE."*
>
> PROF. DR. EVA PFANZELTER - UNIVERSITY OF INNSBRUCK

Recommendation: In order to deliver the inter- and trans-disciplinary capacity that the emerging digital society will require, and which has been tested within the NewsEye project, **Europe will need to increase its investment in broad cross-disciplinary capacity building**, ideally via innovative environments such as GLAM labs. Funding that crosses sectoral, disciplinary and other traditionally siloed approaches to knowledge can be a cornerstone for this development. On the one hand, we need more capacity for implementing leading-edge computer science in cultural heritage institutions, on the other hand we need a hybrid approach where the 'humanistic' and the 'digital' meet on the same level.

POLICY AREA 5: INNOVATIVE OPEN SCIENCE: SHARING 'FAIR' RESEARCH DATA

The data, which results from the digital historical text analysis pipeline is a final step to close the virtuous circle of feedback and improvement that connects digital humanities and computer science. We need to make sure that cultural heritage data can be made available for research, in spite of the many necessary restrictions required to ensure its provenance and protect the human subjects that created or were the subject of it. This complex challenge is beginning to be approached by initiatives such as the Social Sciences and Humanities Open Cloud project (SSHOC) providing humanist-friendly access to the European Open Science Cloud (EOSC). There is still significant work to be done, however.

In terms of humanities research data, how 'open' can open science be? While cultural heritage metadata is generally open for reuse, the underlying historical textual data may still be protected under copyright and other legal restrictions. **We should be striving for cultural data to become Findable, Accessible, Interoperable, and Reusable (FAIR) throughout this continuum**, but within the limits that its hybrid nature as both cultural heritage and research data require. To take an example, Association of European Research Libraries (LIBER) recently published (May 2020) a paper on supporting text and data mining. LIBER is actively advocating for researchers to be able to use text and data mining for their research, without fear of legal ramification. Similarly, **cultural heritage institutions need to be confident as to what is legally possible with their data**.

NewsEye Contribution (Evidence & Analysis): All tools, services and datasets developed within NewsEye are made available on the following sites and will be sustained beyond the project duration:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299

- project website newseye.eu/
- GitHub repository: github.com/newseye
- publications and data sets: zenodo.org/communities/newseye/
- NewsEye Demonstrator platform platform.newseye.eu/
- social media and interactive platforms: Podcasts, YouTube and Twitter

In addition to this, NewsEye team members have embedded their work firmly within the landscape or European and international cultures to make digital cultural heritage visible and usable: this is an effort that needs to be scaled up. The NewsEye project is already linked with Europeana Newspapers, Impresso, Living with Machines, Oceanic Exchange, IMPACT, READ-coop, OCR-d, Embeddia, amongst others. The project also seeks to draw up associated partnerships with national libraries and research communities therein to develop its work and make it sustainable and accessible. For these leads to be built upon, we need an effective platform to coordinate programmes of work and to facilitate more efficient knowledge sharing and cooperation. This would be a mechanism with which to address gaps between needs of researchers and institutions with regards to the production and use of cultural heritage data.

The communication between these groups and linking to ongoing and upcoming research is key, for instance with the EOSC, Digital Europe, and European research infrastructure consortiums (ERIC) for language resources and technology (CLARIN) and for digital research for arts and humanities (DARIAH). **Ensuring that these initiatives are joined up is the next step in a European wide data access-friendly approach**. We need an effective platform to reduce inefficiencies in programmes of work and to facilitate more effective knowledge sharing and cooperation between them. This would be an effective mechanism with which to address the many gaps between researcher and institutional needs with regards to the production and use of cultural heritage data.

Recommendation: Although a few European countries already have copyright exceptions for text data mining, many still do not. The currently tricky labyrinth of legalities of copyright still needs to be made clear for cultural heritage institutions and digital humanities professionals. **We call for exceptions for research of digitised newspaper data or clarity on what is possible within those limitations**. Additionally, we need training in how to deposit such materials in open access platforms and then continued investment in such a resource to make data available for reuse. Finally, funding is needed for cross sharing and 'fairing' of data between the various strands and national efforts of European cultural heritage open access initiatives. We suggest setting up a platform as a room for exchange and to coordinate and resolve issues of shared concern regarding access.

CONCLUSION

NewsEye's contributions to the progress of European cultural heritage initiatives and the application of advanced digital methods in humanities research through interdisciplinary collaboration have given the team behind this project a privileged position from which to observe both the current state of the art and the potential futures for this research. Continued investment on European level and policy support, as detailed in this brief, are now needed more than ever in order to make our work impactful; we owe that to European cultural heritage.

PROJECT IDENTITY

Project Name: NewsEye: A Digital Investigator for Historical Newspapers

Project Number/ Grant agreement ID: 770299

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299

Coordinator: University of La Rochelle, France
Primary contact: Antoine Doucet antoine.doucet@univ-lr.fr

Consortium:
University of La Rochelle – ULR – France
Austrian National Library – ONB – Austria
University of Helsinki – UH – Finland
Department of computer science
Helsinki Centre for Digital Humanities
National Library of Finland
University of Innsbruck – UIBK – Austria
Institute of Contemporary History
Digitisation and Digital Preservation Group
Bibliothèque nationale de France – BNF – France
University of Rostock – UROS – Germany
University Montpellier III Paul Valery – UPVM – France
University of Vienna – UNIVIE - Austria

Funding scheme:
Horizon 2020
Type of action: Research and Innovation action
Call: Understanding Europe, promoting the European public and cultural space
Topic: European cultural heritage, access and analysis for a richer interpretation of the past
Call/Topic IDs: H2020-SC6-CULT-COOP-2017/CULT-COOP-09-2017

Duration: May 2018 – April 2021 (36 months)

Budget: EU contribution: 2 998 565 €

Website: https://www.newseye.eu/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299