Digitised historical sources and non-digital humanists: an interdisciplinary challenge?

Maelle Le Roux[1][0000-0002-0635-6935] and Anna Gasperini[2][0000-0001-5788-1978]

[1] University of Galway, Galway, Ireland
maelle.leroux@universityofgalway.ie
[2] University of Galway, Galway, Ireland
anna.gasperini@universityofgalway.ie

Abstract.

The digitisation of sources has opened new perspectives for humanities scholars. Digitisation allowed a larger access to sources, removing some financial and geographical limits, and the use of digital tools provided new perspectives for humanities scholars, who are able to read the sources differently. However, working with digitised sources also created new challenges that humanities scholars are not always equipped to overcome.

The 'MedIcal Literature and Communication about Child Health' (MILC) project uses historical medical books for a non-specialist audience to analyse discourses on children's health in England, France and Italy between 1850 and 1914. Despite being born a non-digital humanities project, with a focus on manual qualitative analysis and a combination of history and literature methods, it took a digital turn when using digitised sources, with issues of digitisation and Optical Character Recognition (OCR) among others. The team working on the project is composed of three humanities scholars, with limited computer science skills. This required us to find digital humanities and in general IT tools adapted to our skillset, and suited to our needs. These tools did not always fit all our needs, and often presented issues in terms of accessibility and compatibility with the general standards of digital humanities.

Using examples from the issues faced by this project, and from the solutions found, this paper will argue that the challenges encountered by humanities scholars are interdisciplinary, not only because they overcome the traditional disciplinary boundaries inside the humanities, but also because they mirror challenges that computer scientists are working to solve. This paper will also argue that collaboration is a necessity which would benefit both humanities scholars and computer scientists in their work on the improvement and development of new tools, with the help of AI for example. Using the work done by a team of non-digital humanities scholars, it will argue that accessibility is a central issue in digital humanities and in the creation of IT tools, which needs to be addressed.

Keywords: Digital Humanities, Digitised Sources, Accessibility.

1 MILC – a non-digital Digital Humanities project?

1.1 Introduction

MILC - MedIcal Literature and Communication about Child Health is an interdisciplinary Humanities project combining medical history and literature. It performs a transnational comparative analysis of childcare manuals written for a non-specialist audience in French, English and Italian, focusing specifically on how the texts present the themes of breastfeeding, vaccination and physical education. The project focuses on texts published between 1850 and 1914, with 361 books and pamphlets composing the main corpus.

The project's methodology was originally planned to be non-digital, using qualitative methods traditionally adopted in the field of literary analysis, in which the project was grounded. This analysis was to be contextualised with a catalogue of national French laws collected through archival work, and a study of the translations of these popular medical texts, neither task being considered as requiring digital components. This article analyses the challenges the project presented once it took a more digital turn, changing shape slightly while still including all its original aims and methodologies It also outlines the solutions that we adopted as the result of collaboration and communication between digital and non-digital scholars.

1.2 A non-digital methodology

The original, non-digital methodology for MILC, which is still part of the project, envisaged the manual close reading of the texts articulated in three phases. Phase one, recently concluded, was for data gathering, during which we built three "language" corpora – one of texts in English, one in Italian, and one in French – and one corpus of translated texts to help us identify the role of translations in the international circulation of knowledge about child health. The data gathering phase also envisaged two key-intermediate assessment steps (one for the language corpora and one for the translations corpus) to assess the quality and quantity of the material gathered and adjust the literary analysis performed in phase two accordingly.

Phase two, literary analysis, will examine a selection of case studies from each "language" corpus. These texts are not, technically, "literary" in the same way as, for example, novels are: they were handbooks, a series of childcare instructions for parents, guardians, and some professional categories. The project, however, analyses them as literary texts, examining their language and content, as well as the position of the author and that of the reader, against the background of the historical and cultural landscape in which they appeared. Specifically, their content is analysed using a historical social constructivist approach that considers medical knowledge as the product of cultural and social dynamics tied to a specific historical moment, and therefore bound to change over time.[1] The goal is to understand these texts as cultural products that contributed to circulating and creating meaning and knowledge about child health.

The last and final phase envisaged the transnational comparison proper, through case studies. During the project planning phase, it was envisaged that the work should be performed manually, as in, without the aid of digital tools. To an extent, this is still the main method used in MILC. However, as a synergy was created between the digital and non-digital sets of expertise of the different team members, we started realizing the potential digital material has for revealing different types of data than the ones that is possible to gather through manual close reading.

1.3 Using digitized sources

It was decided early in the project to use digitised sources to overcome the hurdle posed by the geographical distance between the different archives. Using digitised sources, in theory, allowed us to access them without the financial or time limitations that accessing and working on the physical versions required. Using online catalogues from major archival institutions in the countries the project focuses on, we identified a large corpus of sources, with 361 overall. We quickly noted the discrepancies in their digitisation.

While the French sources were mostly digitised and available online through the Bibliothèque Nationale de France (BNF) and its digitised sources database, Gallica,[2] few digitised sources were available through the British Library, and even less from the Biblioteca Nazionale Centrale di Firenze (BNCF). We were able to collect digitised versions of some sources in our corpus from other online databases, with the Wellcome Collection especially,[3] but the online databases focused mainly on French and English sources, with very few Italian sources available online. These elements caused the first shift towards a more digital methodology, to help us, first, gather digitized copies of the Italian material and, second, explore the possibilities offered by the analysis of what was immediately available to us: the texts' metadata.

1.4 A digital qualitative analysis

The distant reading analysis of the metadata, especially the titles of the texts, would allow us to better understand the corpus globally, while also providing some information on the themes and audiences of the texts.

Exploring the options available to the group, we selected the Qualitative Data Analysis Software (QDAS) NVivo. NVivo is a commercial software developed by the Lumivero company.[4, 5] Its last version, NVivo 14, which the project uses, was released in 2023. NVivo is structured as a relational database, although it relies on a software-specific vocabulary, sharing few elements with standard relational databases vocabulary. We selected this software because it combined an excellent fit for the qualitative methodology of the project with immediate free access, training, and technical support for all team members through the University of Galway, where the

project is being developed and where NVivo is widely used for qualitative research. The fact that using NVivo does not require specific technical skills further to the training provided made it also especially suited to a team of mixed digital and non-digital scholars, and indeed we were able to proceed immediately, after importing the metadata of the corpus in NVivo from the bibliographical software Zotero, with the 'coding' of the metadata. In this process, the book titles and other elements such as the authors and places of publication were annotated based on pre-determined criteria ('codes') that grouped the texts based on elements such as intended reader, themes, and the vocabulary they used.

Format-wise, NVivo produces projects in a proprietary format, '.nvp', which is not compatible with other databases formats. The results of the analysis are exportable in Excel and CSV formats, which allowed us to conduct some data analysis and create visualisations based on the results of the distant reading of the titles.

2 The digitised sources

2.1 Creating a corpus from digital catalogues

The project intended to create its own corpus to analyse, which was, in itself, a challenge. Indeed, no prior study had catalogued childcare manuals, which required us to identify and collect the sources by researching various online catalogues, namely Worldcat, the BNF, the BNCF, the British Library and the Wellcome Collection.[6]

As noted by Blaney et al., the search process is a part of the methodological process that is difficult to document, and rarely reproducible.[7] The reproducibility of searches was an issue we encountered in the project, whereby we met keyword searches problems of the same kind documented by Hitchcock on the topic, with a potential lack of accuracy leading to a larger number of irrelevant texts.[8] This aspect was further complicated by the project's transnational framework. While we aimed for maximum consistency, translating keywords or finding the closest possible alternatives across the three languages, these translations were not accurate in every context.

A further, if opposite, issue emerged when we used the categories available in the online catalogues to identify texts which supposedly were related to the project. Counterintuitively, searches by category returned very few texts, which we could note due to the absence of key-texts we found through other modes of research. Little information exists on the definition of these categories in online catalogues, with limited accessibility to the details behind the search engine and to the original metadata of the texts, making it impossible to verify the accuracy of our search.

The solution we adopted to overcome these challenges was using a manual combing method through the keywords, combined with the use of historiography and secondary readings to identify various other sources we had missed in the first version of the corpus. The result was a corpus composed of 97 sources in the English, 159 in the French and 105 in Italian, which we determined to be balanced and sufficient for the

analysis we intended to conduct. This combination of traditional search methods, secondary readings, manual combing of the catalogue, and keyword searches, had the added benefit of allowing us to identify the key-texts in the corpus, as in, the most important specimens of childcare manuals produced in the time span examined.

2.2 Digitised sources and OCR

Besides the matters related to research through digital catalogues, the other main challenge we faced in MILC was related to Optical Character Recognition (OCR). OCR issues, as in, whether or not texts are machine-readable, are common for humanities scholars working with digitized sources. Blaney et al. noted the variety of standards available with digitized texts, depending possibly on the goal of the institutions digitizing the texts, with some focusing on their preservation and their availability to a larger number of readers, and others focusing on the compatibility of these newly digitized texts with new standards in digital texts availability.[9] Since MILC had not been planned as a digital project, we had to work with comparatively limited resources when it came to solving OCR issues. The resource that proved most effective among the ones we could access was the ABBYY PDF FineReader 15 software, on which we heavily relied for some countries, less for others, based on the differences in digitization policies across the archives.

The French sources – an OCR-oriented digitization policy

The French sources were mostly pre-OCRed, as they were made available by Gallica. The institution uses a combination of internal and non-specified external systems to OCR its sources, and has been involved in the research and development of new OCR tools.[10] A sample testing on the texts showed that their OCR was more efficient than other tools we had access to, and we did not modify them in any way.

Gallica, in its OCR policy, indicates aiming for its OCR to have an efficiency rate of 96 per cent on texts, although it acknowledges that this rate does not apply to all the texts due to some being digitised prior to this policy, and due to this policy not being applied to ‘numbers, tables, unreadable sections, adverts’.[11] Gallica also states that while a high rate of accuracy might be indicated, this rate might have been calculated excluding pages where the OCR would have encountered difficulties, making the reading of the document less efficient than announced. Furthermore, the quality of the OCR has an impact on the search function and possibly impacted the corpus itself. Indeed, Chiron et al. noted the higher error rates of Gallica’s OCR on named entities, and indicated that this impacted the search function for users searching for proper nouns, such as people’s names, while being difficult to clean post-OCR due to the absence of some of these names in traditional dictionaries.[12] While searching for specific texts, we encountered similar issues, as the search engine did not retrieve some of the texts we were searching for, despite us knowing of their presence on Gallica through exterior sources. We were able to find the texts through the authors’ pages,

indicating that while the text had a correct metadata, the search engine seemed to focus on the OCR of the texts which had not read the authors' names correctly.

These issues had a limited impact on the project itself, beyond suggesting the potential existence of more digitised texts that could have been added to the corpus; had we proceeded to scale up the project with a social network analysis, which we considered at one stage of the project before discarding the idea due to time constraints, it would have been a different matter. Overall, while noted, these OCR issues did not impact the quality of the corpus, which we deemed sufficient for the distant and close reading work we intended to do.

The English sources – dealing with limited OCR within an image preservation-oriented digitization policy

Most of the English texts we collected were digitised by the Wellcome Collection, a private library and museum focusing on the theme of medicine which is part of the Wellcome Trust charity. Unlike the texts available through Gallica, the texts we collected through this database were not pre-OCRed. Even though it does mention OCR a few times in their digitization policy,[13] from the document emerges that the main goal of digitisation at the Wellcome Collection is to preserve a copy of the historical text and to make it accessible online to users who cannot access the physical library. Both the policy and the presentation of the digitised sources would indicate that each page of the texts is perceived as an image, rather than as text.

To overcome this obstacle, we used ABBYY PDF FineReader 15 to OCR the sources available to us from this collection. The software, and this specific version, was chosen because it is recognised as a standard in the field of non-professional OCRing historical sources,[14, 15] but also for its convenience as we had access to a licence of the software and we could rely on previous experience in using it to OCR digitised historical sources. Our choice of version was also guided by availability of funds. Since ABBYY PDF FineReader is a commercial proprietary software and the latest version, FineReader 16, was released in 2023 on a subscription model, we decided to use version 15, released in 2020. The OCR model of the software is not accessible due to its proprietary nature, and few details are available on it. The software itself emphasises its use of Artificial Intelligence (AI), with Machine Learning especially, and indicates the advances made by each new version using the latest research in the field.[16]

The use of a previous version of the software to OCR the sources in this project can be considered an issue, as it means that we did not use the latest technology available and that potentially we could have produced better OCRed documents using the latest version. However, there were multiple reasons for choosing this version: first, our analytical approach did not require perfect accuracy in words recognition by the software, as would have had a qualitative approach such as corpus linguistics; second, we did not intend to share the texts of the full digitised corpus for copyright reasons. These first two elements allowed us a certain degree of flexibility in opting for a lower degree of accuracy in the OCR. Finally, as with NVivo, we had to account for the fact

that the digital element was integrated in the project at a later stage, which made budget a key-element to take into account in our choice. As with NVivo, ABBYY PDF FineReader combined cost efficiency with suitability to our purposes, as one of the team member had access to a licence. Therefore, we OCRed the texts using mainly automatic settings in ABBYY PDF FineReader with limited manipulation.

The Italian sources – dealing with limited digitization

The Italian sources presented the greatest challenge regarding OCR. The BNCF, like other national Italian libraries, has a digitization policy and regularly collaborates with institutions and projects to digitize and make accessible some of its collection.[17, 18] A governmental report in 2016 indicates that the BNCF intends to pursue this digitization work to improve accessibility of its collection, while following the current standards of the field.[18] Due to the quantity of material in their collection, they had not been able to digitize most of the texts in our corpus, possibly due to their lack of popularity amongst researchers and readers, compared to other documents. We started the process of having them professionally digitized and, in the meanwhile, we proceeded with manual digitization to be able to start the close reading work.

The documents were photographed page by page by members of our team visiting the BNCF using a phone camera. Blaney et al. notes that this is a common method of digitization amongst historians, as due to the necessary selectivity of institutions in digitizing sources, as well as financial constraints from these same institutions and from researchers in the case of digitized documents being behind a paywall, it is common for historians not to have access to a digitized version of the documents they intend to use.[18] The photographs were directly transformed into a PDF using OneDrive, as the program organized and backed up the documents. This method, while cost-effective, was not without its pitfalls. First, the manual photography method for digitization is time consuming for the researcher, and with the lack of proper photography equipment, can represent a physically difficult task. Secondly, this digitisation method also presents problems in terms of long-term preservation of the data, as the most popular photographs formats are formats such as JPEG or PDF, and not TIFF, which is the recommended format for long-term preservation of digitized sources. Finally, and most importantly, this method of digitization produces images whose quality may vary significantly.

A researcher without specialist skills in photography will hardly perceive this difference in quality while taking the pictures; however, it will emerge and potentially present challenges during the OCRing process. These difficulties are of three kinds: blurriness; variety of frames and light; and quality of the original material. Images captured during our research trips were at times blurry and, while still readable by human eyes, they were partially unreadable by the OCRing software. As Taş and Müngen noted, pre-processing the images can improve the OCR results for historical sources.[19] Image processing through OCRing process in ABBYY PDF FineReader 15 allowed us to an extent to correct the blurriness of the photographs and to improve their overall quality. In our case, pre-processing was further complicated because,

compared with professionally digitized sources, the photographs taken manually had variable frames, orientations and lighting.

After processing the PDF in the software and conducting a first OCRing of the document, we proceeded to more targeted interventions on pages in which the software had been unable to read the text. The software allowed us to process images individually or per document; the variable quality of our pictures made us opt for processing per individual image. The most common issues encountered were defocus blurs, especially side-ways due to the book format making the surface uneven, and making lines of text askew, which was often fixable with the software's tool to align the text lines, or to reorientate the page in the photograph. In instances where blurriness of the picture affected only a small portion of text, we simply corrected the OCR text by hand. There were cases instead in which some or all of the text was fully unreadable, either because of the blurriness of the picture, or because of the quality of preservation of the original document. Indeed, a major issue for OCRing is the quality of the original paper and ink, associated with potential preservation issues creating stains in the paper.[8, 9] While this is usually an issue associated with historical newspapers, we can assume that childcare texts, which aimed to be financially accessible to a relatively large public, were sometimes made with lower quality materials, resulting in these issues. Furthermore, accidental flooding of the archives in the middle of the twentieth century caused damages to some of the books in our corpus, causing further readability issues.

Overall, while this manual version of the Italian corpus is digitized and OCRed, it is only partially, and while we were able to solve some of the issues, we faced limitations that we did not have the resources to overcome.

The aim of MILC is to analyse these texts comparatively, with a transnational perspective, to understand the nuances in child health discourses they present based on context. Since the methodology for MILC always intended to focus on case studies and did not necessarily require using all the texts in a full comparative approach, the issues we encountered with the digitized sources had a limited impact on the project. However, the potential impact of the issues encountered when working with digitized sources on a project should be noted, especially in a context of comparative approaches, in a transnational or global history perspective.

2.3 Open-source alternatives for OCR and non-digital humanists

As discussed in the previous section, we decided to use ABBYY PDF FineReader 15 to OCR the texts in our corpus for multiple reasons, one of them being that we had access to a license through a member of the team. This caused some concerns regarding the long-term access to OCR methods, and overall the reproducibility of the research. Indeed, since ABBYY FineReader is a commercial proprietary software, we do not have the details of its OCR, making it impossible for other researchers to reproduce the methodology unless they themselves had access to the same version of the software. Furthermore, the use of a license associated to an individual member of the team could

cause some issues in the long term, if the team member were to stop working for the project and the project required other texts to OCR. This would therefore require finding a different OCRing tool and defining a new process. With these issues in mind, we considered other OCRing tools that could be used by all the team members and by other researchers without constraints of cost or technical skills.

As we looked into open-source OCR software, we noted that studies indicated Tesseract to be the main open-source alternative in the field and decided to experiment with it, to see if it would be a suitable alternative.[15] Despite fitting our requirements for being open-source and free, it did require technical skills, rendering it difficult to use by our team of non-digital humanists. Tesseract was developed at HP before being released under an Apache license.[20] The software can be installed directly on a computer or run through an Application Programming Interface (API). We did not experiment with API in this project and worked exclusively with Windows OS. The installation of the software is possible through files in Github or through an installer developed by the Mannheim University Library (UB Mannheim), which is the process we decided to follow as an installation through GitHub files required more technical skills.[21] As noted in the documentation, Tesseract does not have a graphical user interface (GUI), which meant that there was no front-end visual as support for the user. This required us to use a command line interface, which is a complex tool to use for non-digitally trained researchers. Indeed, command line interface requires a good understanding of the logic involved with computer languages, which is rarely ever part of non-digital humanists training. While it is possible to find documentation and tutorials online which explain how to use some basic functions, use of this software remains, mostly, entirely inaccessible without extensive specialist training. We attempted to find a suitable GUI through the ones developed by third parties,[22] but this attempt had limited success, with multiple GUIs being difficult to install or difficult to use without specific technical skills.

Overall, the OCRing of the texts was the biggest challenge we encountered in the digital aspect of the project, as it had multiple ramifications and our methodology depended on having access to a corpus of OCRed texts in three various languages, digitized in different contexts and with different methods. While we aimed to be consistent in the OCRing process, and we aimed to find a solution to the non-reproducibility associated with the use of commercial proprietary software, there were none suitable for a team of non-digital humanists.

3 Integrating the FAIR principles into a non-digital native project

3.1 From corpus to dataset

The Findable Accessible Interoperable and Reusable (FAIR) data principles are central in research, and especially in digital humanities and computer science due to the creation and use of datasets.[23] The MILC project was planned with these principles in mind, and the introduction of the digital element gave a different inflection to its FAIR approach. Indeed, while the corpus originally intended to be a non-digital object, the use of various digital tools to catalogue it and analyse it transformed it into a potential dataset, which led us to reflect on how to integrate this new component in the project.

In the original planning of MILC, the catalogues were to be made accessible at the end of the project. These catalogues would have been tables with the metadata of the texts composing the corpus, but also of the French laws and their relevant texts. This took a different turn when we started using NVivo to use the metadata of the corpus into the distant analysis, but also when the catalogue of the French laws became more complex and we decided to organise it as a relational database using Access. With the change in methodology for both catalogues came the issue of the sources themselves, either institutionally digitised or privately digitised, which were directly associated with the items in the catalogues. Indeed, to avoid issues of copyright and rights of use of the digitised sources, it was decided early on not to release the sources themselves. As these sources have various origins, with various legislative frameworks, not releasing any source was easier than releasing only a portion of the corpus. Therefore, while the metadata of the dataset is releasable, the full corpus will not be released.

3.2 From NVivo to FAIR-compatible data?

Since the work conducted on the corpus through NVivo became an important part of the analysis, it was necessary to consider how we could make this data potentially available to researchers in a FAIR-compatible format. FAIR-compatible format here refers to interoperable formats usable by most software and Operating Systems, such as XML here in the case of the corpus. NVivo, as a commercial software, has its own proprietary formats, which are not interoperable. This caused us some issues when we attempted to extract the full dataset with its encodings from the software for data analysis, as this was not an option it offered.

This led us to find that NVivo projects could be exported in an interoperable format compatible with other QDAS, the REFI-QDA format, which is based on, and easily convertible to, the XML format.[24] This format is mainly used in transferring a project from a QDAS to another similar software, and so far the tests we conducted with open source alternatives did not bring satisfactory results. Another issue is that the format, and the XML file resulting, are structured around the texts used to create the project.

Therefore, making available this data in XML or other formats would necessitate making the full corpus available, which is not an option due to copyrights issues.

This issue shows the limits encountered by the project in fitting with the FAIR principles, since the corpus and part of the analysis were not expected to become potentially sharable data.

4 Conclusion – Learning from non-native Digital Humanities project for the future of the field

4.1 Learning from MILC – is there such a thing as a non-Digital Humanities project? Digital Humanities and the accessibility question

The issues the MILC team encountered are fairly normal for a Digital Humanities project. The real challenge was that they were not anticipated, because the project was perceived as non-digital, leading to limitations in the resources available to face these challenges. Despite these challenges, we were able to create datasets in the project which fit the standards of the field, and partly fit the FAIR principles.

In the early phases of the project, as a team we reflected on whether MILC should be called a Digital Humanities project, or a project using Digital Humanities tools, which led us to question the difference between these two definitions. We concluded that, even though it does not lead to any significant innovation in Digital Humanities, MILC is a Digital Humanities project because it uses digital sources, and uses digital tools to clean, structure and analyse them.

By adopting this broad definition of Digital Humanities, and considering the fast-paced development of new tools and the advancement of the digitisation process, we can envisage that most, if not all, humanities project will take a digital turn, similarly to MILC. Learning from this non-native Digital Humanities project, and the challenges we encountered in its transformation to the digital, Humanities researchers must plan for projects that might need higher digital skills than anticipated, and therefore account for the necessary interdisciplinary collaborations.

The question of accessibility was central in our methodological process, especially when it came to the question of selecting software. As classically-trained humanities scholars, we had to select software and tools that were easily accessible to us. While training was an option, we relied on training that was easily available to us, and that would allow us to quickly become proficient in using the tool. Since this was a minor aspect of the methodology, we did not have the resources and time necessary to focus on general computer science training.

These accessibility and training matters necessarily limited our choice, resulting in the use of tools that would complicate the process to make the data FAIR. The open-source or FAIR compatible alternatives did not fit our skills or the project's needs. Therefore, we would say that there is a need for more accessible Digital Humanities software, as an increasing number of researchers in the field come from non-Digital

native background and do not have the technical skills that most open-source software require at the present.

4.2 Creating accessible tools - the importance of collaboration

MILC, while interdisciplinary, was born as a Humanities project, without a digital aspect. Its conception as a non-Digital Humanities project made collaboration with the field of computer sciences more difficult, as no computer scientists were involved, and no resources were allocated to such a collaboration, which was not envisaged in the initial layout of the project. However, MILC was thought as a stepping stone project, leaving the possibility for future researchers to scale it up as a Digital Humanities project. By using the latest advances in Artificial Intelligence and Machine Learning applied to OCR the texts, therefore creating a more accurate dataset, researchers could then apply various tools to the corpus in order to provide a broader analysis using both qualitative and quantitative methods. The data we have accumulated would benefit from applying a social network analysis method, and the multilingual corpus would be an excellent source for language analysis through corpus linguistics and Natural Language Processing, with named entities analysis especially.

Digital Humanities is by its nature an interdisciplinary field, and collaboration is central. First and foremost, it is central in the production of more accessible tools, as this accessibility can only happen by a communication process between non-digital native humanists and computer scientists. These non-digital researchers are part of the future of Digital Humanities, and they are the potential users of the tools that computer scientist researchers are developing. Making accessibility a central part of this development and facilitating a dialogue between computer scientists and non-digital scholars would ensure that the innovative tools supporting Digital Humanities research are strategically tailored to its needs.

Another important step would see archival institutions, humanities scholars and computer scientists collaborating to establish standards in the process of digitising sources and making them accessible, including through retrospective work on sources digitised using outdated methods and tools. As AI and Machine Learning progress, computer scientists are creating tools that are used by some archival institutions in their digitisation process, but not by all, and archives tend not to provide information about this aspect. Consequently, humanities scholars do not have sufficient data to ascertain whether the quality of the digitised source they intend to use is sufficient for their purpose.

The development of standards for Digital Humanities must be done collaboratively, to ensure that they are both sustainable and accessible by researchers without formal computer science training, while still promoting good practices for the long-term contribution to knowledge and research. For example, the achievement of the FAIR principles, amongst other standards promoting the release of a fully reusable dataset to other researchers, may be out of reach for scholars with limited computer science training, due to technical difficulties, legal complexities, or the absence of accessible compatible tools. A reflection on these issues, in collaboration with archival institutions

and researchers in computer science and Humanities is necessary to promote these standards of open-data in the Humanities and help the field of Digital Humanities navigate this turn of non-native Digital Humanities project.

Acknowledgments.

The research conducted in this publication was funded by the Irish Research Council under award number IRCLA/2022/3921.

Disclosure of Interests.

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Lupton, D.: *Medicine as Culture. Illness, Disease and the Body*. 2nd edn. Sage, New York (2012).
2. Gallica Homepage, https://gallica.bnf.fr/, last accessed 2023/12/09
3. Wellcome Collection Homepage, https://wellcomecollection.org/, last accessed 2023/12/09
4. NVivo product page, https://lumivero.com/products/nvivo/, last accessed 2023/12/09
5. Jackson, K., Bazeley, P., Bazeley, P.: *Qualitative Data Analysis with NVivo*, 2nd edn. Sage, New York (2018).
6. WorldCat Homepage, https://search.worldcat.org/, last accessed 2023/12/09
7. Blaney, J., Winters. J, Milligan, S., Steer, M.: *Doing Digital History*, Manchester University Press, Manchester (2021).
8. Hitchcock, T., 'Confronting the Digital: Or How Academic History Writing Lost the Plot'. *Cultural and Social History* 10(1), 9-23 (2013).
9. Pardé, T., Jacquot, O., 'Les humanités numériques à la Bibliothèque nationale de France'. Paris, *Culture et recherche* (2016), hal-01379908.
10. Gallica, 'Mode texte et OCR', https://gallica.bnf.fr/edit/und/consulter-les-documents#Mode%20texte%20et%20OCR, last accessed 2023/12/09
11. Chiron, G., Doucet, A., Coustaty, M., Visani, M., and Moreux, J.-P., 'Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information,' *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Toronto, ON, Canada, 2017, pp. 1-4, doi: 10.1109/JCDL.2017.7991582.
12. Wellcome Collection Digitisation Strategy, 2020-2025, https://wellcomecollection.cdn.prismic.io/wellcomecollection/0047856d-bba9-4ab2-81b6-a270f887a8fb_WC+Digitisation+Strategy+2020-2025.pdf, last accessed 2023/12/09
13. Volk, M., Furrer, L., Sennrich, R., 'Strategies for Reducing and Correcting OCR Errors'. In: Sporleder, C., van den Bosch, A., Zervanou, K. (eds) *Language Technology for Cultural Heritage. Theory and Applications of Natural Language Processing*. Springer, Berlin, Heidelberg, 3-22 (2011). https://doi.org/10.1007/978-3-642-20227-8_1
14. Tafti, A.P., Baghaie, A., Assefi, M., Arabnia, H.R., Yu, Z., Peissig, P. 'OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym'. In: Bebis, G., et al. *Advances in Visual Computing. ISVC 2016. Lecture Notes in*

Computer Science(), vol 10072. Springer, Cham, 735-746 (2016). https://doi.org/10.1007/978-3-319-50835-1_66
15. 'ABBYY FineReader PDF: Powered by AI', https://pdf.abbyy.com/blog/finereader-powered-by-ai/, last accessed 2023/12/09
16. BNCF website, 'Categorie delle Risorse: Collezioni digitalizzate', https://www.bncf.firenze.sbn.it/categoria_risorse/collezioni-digitalizzate/, last accessed 2023/12/09
17. Lucarelli, A.: 'Web dei dati alla Biblioteca nazionale centrale di Firenze'. *DigItalia*, 10(1/2), 30–39 (2016), https://digitalia.cultura.gov.it/article/view/1471
18. Taş, İ. Ç., Müngen, A. A., 'Using Pre-Processing Methods to Improve OCR Performances of Digital Historical Documents,' *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Elazig, Turkey, 1-5 (2021) doi: 10.1109/ASYU52992.2021.9598972.
19. Tesseract Documentation, https://tesseract-ocr.github.io/tessdoc/Installation.html, last accessed 2023/12/09
20. Tesseract at UB Mannheim, https://github.com/UB-Mannheim/tesseract/wiki, last accessed 2023/12/09
21. GUIs and Other Projects Using Tesseract OCR, https://tesseract-ocr.github.io/tessdoc/User-Projects-–-3rdParty.html, last accessed 2023/12/09
22. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
23. Müller, A., 'From QDA to XML: The REFI-QDA project exchange standard', https://methodos.hypotheses.org/1707 last accessed 2023/12/09