Collaboration and Transparency: A User-Generated Documentation for eScriptorium

Alix Chagué, Floriane Chiffoleau, Hugo Scheithauer

▶ **To cite this version:**

Alix Chagué, Floriane Chiffoleau, Hugo Scheithauer. Collaboration and Transparency: A User-Generated Documentation for eScriptorium. DH2024 Reinvention & Responsibility, Alliance of Digital Humanities Organizations, Aug 2024, Washington D. C., United States. hal-04594142

HAL Id: hal-04594142
https://hal.science/hal-04594142v1

Submitted on 30 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Short Paper Proposal to DH2024: Reinvention & Responsibility

Collaboration and Transparency: A User-Generated Documentation for eScriptorium

Alix Chagué[1,2,3], Floriane Chiffoleau[1,4], and Hugo Scheithauer[1]

[1]ALMAnaCH - Automatic Language Modelling and Analysis & Computational Humanities, Inria, Paris, France
[2]UdeM - Université de Montréal, Montréal, Canada
[3]EPHE - École Pratique des Hautes Études, Paris, France
[4]Le Mans Université, Le Mans, France

December 2023

Abstract

We use the example of the user-generated documentation created for eScriptorium to investigate the benefits and limitations of such contributions to open-source software. The new documentation offers a solution to a scattered, hard to maintain landscape of documentation on eScriptorium. Its design favors future collaborations across user groups and languages.

1 Introduction

eScriptorium is one such tool born from a research project pertaining to the domain of the Digital Humanities, like Transkribus (Muehlberger et al. 2019), Voyant Tools (Sinclair and Rockwell 2016) and TXM (Heiden 2010). It was designed and developed in the context of the SCRIPTA-PSL research project[1] as an open-source web application to conduct automatic text recognition (ATR) campaigns (Stokes, Kiessling, et al. 2021). ATR is now an essential technology in the toolbox of patrimonial institutions and researchers in (digital) humanities: it enables users to swiftly and seemingly effortlessly obtain transcriptions of printed or manuscript documents, making them compatible with a spectrum of computational investigation techniques—from basic string matching processes to sophisticated

[1]See https://scripta.psl.eu/.

information extraction relying on natural language processing. However, ATR workflows are complex and involve several steps, making eScriptorium, like the other tools mentioned above, an "expert software," offering a large range of functionalities which constitutes a challenge for newcomers who need to familiarize themselves with a substantial amount of information.

The success of software extends beyond its ability to meet a specific need; it must also be welcoming to new users, which can be ensured by several means: the design of the interface (UX), the compatibility of the tool with the rest of the software environment (for example thanks to the adoption of standard for input and output), but also the availability of reliable documentation. In the case of eScriptorium, given the limitations in the size of the team responsible for developing the application (eScripta), no official extensive documentation was created. Instead, most of the available documentation was user-generated content scattered across the web and tailored to the specific needs of the user group which generated it.

At the beginning of 2023, the ALManaCH team from Inria Paris gathered a small group of expert users and, with the approval of the eScriptorium development team, worked on a solution to create a centralized documentation that could be a single point of reference for all user groups. The motivation for our initiative was two-fold. First, as the authors of the first extensive tutorial for the application (in French) and as the administrators of one of the largest instance of eScriptorium, we are frequently asked to either update the French tutorial or share our expertise with new users. Since the initial tutorial was published on a restricted, project-specific Hypothesis blog, we needed to design a new publication pipeline for our documentation. Second, we wanted to take the opportunity of this reconfiguration to find a solution to the dispersion of the pre-existing documentation, in a way that would contribute positively to the open-source and the scientific community around eScriptorium.

During the DH2024 conference, we would like to use the case of the documentation created for eScriptorium as an occasion to explore the ways in which a documentation can be created by a group of people outside of the team in charge of developing a software, and the conditions for this to succeed. Additionally, we want to interrogate the role that such an initiative can play in accelerating the integration of open-source, project-generated software to larger infrastructures.

2 Description of the proposed documentation

Prior to our initiative, information about eScriptorium's features was scattered across various media:

1. Developer-oriented documentation is available alongside the source code;[2]

[2]The manual and Docker-based installations are described here: https://gitlab.com/scripta/escriptorium/-/wikis

2. A series of short videos showcased a selection of essential features and was published on a blog edited by eScripta;

3. A galaxy of tutorials in English,[3] French,[4] German[5] or Polish[6] written outside of the eScripta team was published on various websites but rarely updated;[7]

4. Numerous hands-on workshops, usually tailored for beginners, were conducted by eScripta[8] or by user groups sharing a specific interest in the application.[9]

This situation posed various challenges: locating tutorials or videos could be cumbersome, and they might not comprehensively cover all aspects, especially if the content had not been updated when new features were added to the application. Additionally, for the authors of the tutorials, updating the content could be difficult because of the chosen formats, the lack of stake to do so or simply the unavailability of individuals to perform the updates.

Considering the difficulty for eScripta to propose a comprehensive reference documentation, we proposed to create a dedicated website designed to fulfill two essential criteria: it needed to be easily maintainable since eScriptorium has yet to achieve a stable official release; and it had to be open to external contributions while supporting multilingualism.

The key to creating documentation that is effortlessly maintainable lies in its modularity, use of a lightweight markup language, and commitment to transparency. We operationalized this vision by adopting the "continuous documentation" paradigm through ReadTheDocs[10] (RTD), with the source code openly accessible on GitHub (Chagué et al. 2023). This documentation follows a versioned structure, composed of multiple Markdown files, each addressing specific aspects of the software (such as export, training, prediction, shortcuts, etc). Whenever there is an update to the source code,[11] RTD activates Mkdocs, the tool responsible for constructing web pages from the Markdown files, and publishes the resulting website at a dedicated URL: escriptorium.readthedocs.io (See Fig. 1). As of April 2023 (eScriptorium v0.13.6), the RTD-hosted eScriptorium documentation supplanted the older English tutorial on the application homepage.

[3] See https://lectaurep.hypotheses.org/documentation/escriptorium-tutorial-en.

[4] See https://lectaurep.hypotheses.org/documentation/prendre-en-main-escriptorium.

[5] See https://ub-mannheim.github.io/eScriptorium_Dokumentation/Nutzungsanleitung_eScriptorium.html.

[6] See https://github.com/pjaskulski/escriptorium_tutorial/blob/master/escriptorium_tutorial.md.

[7] For example, the English tutorial was published in February 2021, and was referred to on the homepage of eScriptorium. However, while new features appeared on the application, it was never updated

[8] One such workshop occurred during the DH2022 conference (Stokes and Stökl Ben Ezra 2022). Another recent workshop was held at the University of Pennsylvania (Stokes 2023)

[9] For example, the HTRomance project organized a series of three workshop at the French National Library (BNF) at the end of 2023 (https://bnf.hypotheses.org/35711).

[10] Refer to https://docs.readthedocs.io/.

[11] GitHub's Pull Request system allows for editorial control of proposed updates.

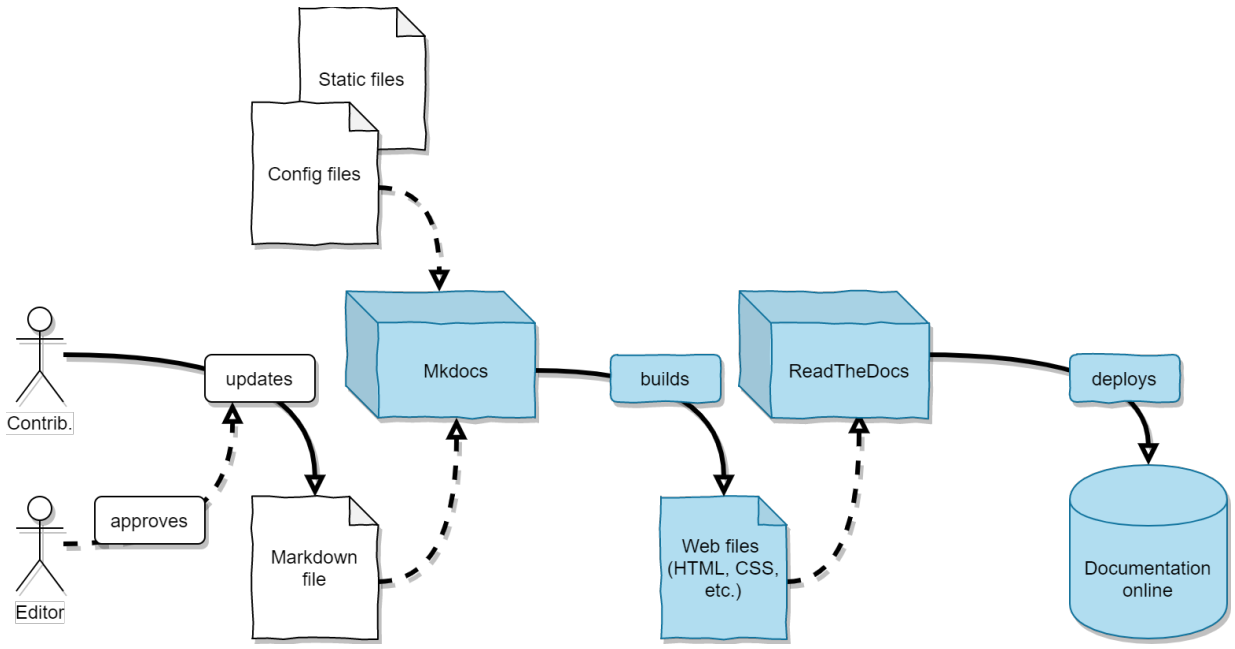

Figure 1: Modelization of the editorial workflow to update the content of the documentation: manual editions are only made on local Markdown or static files (images), while Mkdocs and ReadTheDocs automatically take on the building and the deployment of the updated website.

3 Discussion

Our proposed solution comes with inherent limitations. Similar to the resources mentioned earlier, it is susceptible to partial obsolescence as the developer team integrates new features. Also, as primarily users of eScriptorium ourselves, the content we propose may initially be exposed to blind spots. Thus, a question worth exploring is that of the trustworthiness of a documentation generated by users. The transparency of the process and its openness to any contributions are the keys to remediate these limitations.

Our proposition successfully solved our initial issue: the necessity to redesign the existing documentations for French- and English-speaking users, which had become impossible to maintain. Our efforts focused on a first proposition written in English only, but its compatibility with multilingualism makes it possible to imagine adding a French version later, or even integrating the German and Polish tutorials. Additionally, we realized that contributions of open-source software can come in diverse forms, including in the form of rationalizing its documentation.

To conclude with a few elements of answer to our initial question: the creation of a comprehensive reference documentation for project-generated open-source software, such as the eScriptorium documentation initiative discussed here, can serve as a catalyst for accelerated integration into larger infrastructures. Firstly, by facilitating knowledge diffusion and enhancing accessibility, (well-)documented projects break down entry barriers, ensuring that a broader audience can understand and engage with

the software. Secondly, the collaborative nature of documentation projects has the potential to foster community engagement, which could lead to the creation of a network of users actively involved in the software's development and adoption. Lastly, clear documentation can also attract developers and organizations looking for software solutions that align with their existing systems.

References

Chagué, Alix et al. (Oct. 2023). *eScriptorium Documentation (Source Code)*. (Visited on 12/10/2023).

Heiden, Serge (2010). "The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme". In: *24th Pacific Asia Conference on Language, Information and Computation*. Vol. 2. Institute for Digital Enhancement of Cognitive Development, Waseda University, pp. 389–398. (Visited on 12/09/2023).

Muehlberger, Guenter et al. (Jan. 2019). "Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study". In: *Journal of Documentation* 75.5, pp. 954–976. ISSN: 0022-0418. DOI: 10.1108/JD-07-2018-0114. (Visited on 04/09/2021).

Sinclair, Stéfan and Geoffrey Rockwell (2016). *Voyant Tool*. http://voyant-tools.org/.

Stokes, Peter Anthony (Dec. 2023). *How to Transcribe a Million Manuscripts with eScriptorium*. https://www.library.upenn.edu/events/escriptorium. (Visited on 12/09/2023).

Stokes, Peter Anthony, Benjamin Kiessling, et al. (2021). "The eScriptorium VRE for Manuscript Cultures". In: *Classics@ Journal*. Ancient Manuscripts and Virtual Research Environments 18. Ed. by Claire Clivaz and Garrick V. Allen. (Visited on 08/31/2021).

Stokes, Peter Anthony and Daniel Stökl Ben Ezra (July 2022). "Hands-on Introduction to eScriptorium, an Open-Source Platform for HTR (WT-20)". In: *DH2022*. Tokyo, Japan. (Visited on 12/09/2023).