



THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : Sciences du langage Spécialité Informatique et sciences du langage

Arrêté ministériel : 25 mai 2016

Présentée par

Anne VIKHROVA

Thèse dirigée par **Thomas LEBARBE**, ,

préparée au sein du **Laboratoire Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles**
dans l'**École Doctorale Langues, Littératures et Sciences Humaines**

**L'évaluation de la méthode du
crowdsourcing pour la transcription de
manuscrits.**

**Evaluation of the crowdsourcing method for
manuscript transcription.**

Thèse soutenue publiquement le **6 décembre 2017**,
devant le jury composé de :

Monsieur Thomas LEBARBE

PR, Université Grenoble Alpes, Directeur de thèse

Madame Emmanuelle DE CHAMPS

Professeur, Université de Cergy-Pontoise, Rapporteur

Monsieur Marcello VITALI-ROSATI

Professeur agrégé, Université de Montréal, Rapporteur

Madame Claire DOQUET

Professeur des Universités, Université Sorbonne Nouvelle Paris - 3,
Président

Monsieur Michel BERNARD

Professeur des Universités, Université Sorbonne Nouvelle - Paris 3,
Examineur

Acknowledgments

I'd like to express my gratitude to everyone who has supported me throughout this work.

To my research director, thank you for your advice, patience and encouragement, as well as for your thorough readings of my drafts. Your comments and insights have made an immense contribution to the amelioration of this work.

I'd like to thank the scholars and project leaders involved in this project, your shared knowledge enhanced my understanding and appreciation of our objects of study. Thank you also to the daughters of the late author, Benoîte Groult, for kindly permitting this work, which involves the collection of manuscripts left by their mother.

To my family, thank you for your unwavering support, for knowing my strengths and reminding me of them often. And to my extended family, I cannot thank you enough for all the kindness and attention you have shown me, it has made this journey infinitely more lighthearted and serene.

To my sweetheart and best friend, without your encouragement, support, and understanding, this work simply would not have been realised. Particularly since we were going through it all together. Thank you for living not just through one, but two dissertations with me!

Finally, a very special thanks to all the people whose participation and contributions to PHuN 2.0 and PHuN-ET have made this work possible. A complete list is included in Annex C.1 and C.2.

This dissertation is affectionately dedicated to everyone who encouraged me to finish it.

Contents

| | |
|---|--------------|
| List of Figures | xv |
| List of Tables | xix |
| Introduction | 1 |
| 1 Research questions | 3 |
| 2 Thesis plan | 3 |
| Part I Manuscript transcription for Digital Humanists | 5 |
| Chapter 1 Introduction to Digital Humanities | 9 |
| 1.1 Digital Transformations | 10 |
| 1.1.1 Definition of Digital Humanities | 11 |
| 1.1.2 Crowdsourcing | 13 |
| 1.1.3 Citizen science and Citizen Scholarly Editing | 16 |
| 1.2 Manuscript transcription | 18 |
| 1.2.1 Introduction to manuscript transcription | 18 |
| 1.2.2 Transcription as decoding and encoding | 20 |
| 1.2.3 Manuscript transcription in a digital context | 21 |
| 1.2.4 Manuscript transcription in a participative context | 23 |

Chapter 2 Knowledge dissemination: from books to web 29

| | | |
|-------|---|----|
| 2.1 | Chapter Summary | 30 |
| 2.2 | History of print | 31 |
| 2.2.1 | Gutenberg | 31 |
| 2.2.2 | Facsimiles and conservation | 32 |
| 2.3 | Origins of the collaborative web | 33 |
| 2.4 | Contemporary challenges | 34 |
| 2.4.1 | Technological challenges | 34 |
| 2.4.2 | Scientific challenges | 35 |
| 2.4.3 | Participation | 38 |
| 2.5 | Existing projects | 38 |
| 2.5.1 | Citizen Science and Citizen Humanities projects | 38 |
| 2.5.2 | Digital Humanities scholarship projects | 43 |
| 2.6 | Inspiration and next steps | 46 |

Part II Technical foundations 51

Chapter 3 Encoding textual data 55

| | | |
|-------|--|----|
| 3.1 | Chapter Summary | 55 |
| 3.2 | Introduction | 56 |
| 3.2.1 | Metadata | 56 |
| 3.2.2 | XML for encoding content and metadata | 57 |
| 3.3 | XML and interoperability | 59 |
| 3.4 | Reconciling local projects and the TEI | 62 |
| 3.5 | Dynamic Documents | 64 |
| 3.6 | Conclusion | 68 |

| | |
|--|---------------|
| Chapter 4 Transcription tools and architectures | 69 |
| 4.1 Chapter Summary | 69 |
| 4.2 Introduction | 70 |
| 4.3 Tools for converting digital images to machine readable text | 71 |
| 4.3.1 OCR processing for handwritten manuscripts | 71 |
| 4.3.2 WYSIWYG editors for XML transcription | 75 |
| 4.4 Content Management Systems | 77 |
| 4.5 Conclusion | 79 |
| Chapter 5 Interfaces | 81 |
| 5.1 Chapter Summary | 81 |
| 5.2 Introduction | 83 |
| 5.3 Reading and transcription interfaces | 84 |
| 5.4 Understanding user activities | 87 |
| 5.5 General design principles for user interfaces | 88 |
| 5.5.1 Navigation and work flow | 90 |
| 5.5.2 Operations and actions | 91 |
| 5.5.3 Text and data entry | 92 |
| 5.5.4 User guidance | 93 |
| 5.6 Conclusion | 95 |
| Chapter 6 Methods for comparing documents | 97 |
| 6.1 Chapter Summary | 97 |
| 6.2 Comparing documents | 98 |
| 6.3 Measuring differences between texts | 99 |
| 6.3.1 Measuring differences between XML | 101 |
| 6.3.2 Algorithm complexity | 104 |

| | | |
|-------|---|-----|
| 6.3.3 | Pre-processing transcriptions | 105 |
| 6.4 | Clustering techniques | 106 |
| 6.5 | Visualisation | 109 |
| 6.6 | Conclusion | 109 |

Part III Prototype and production implementations 113

Chapter 7 Presentation of PHuN 2.0 117

| | | |
|-------|---|-----|
| 7.1 | Prototype types | 119 |
| 7.2 | Presenting PHuN 2.0 | 121 |
| 7.3 | Editor functionalities | 126 |
| 7.4 | Functionalities for identified users | 128 |
| 7.4.1 | User accounts | 129 |
| 7.4.2 | Page browsing and selection | 130 |
| 7.4.3 | Transcription and revision | 130 |
| 7.4.4 | User comments and discussion | 132 |
| 7.4.5 | User profile | 133 |
| 7.5 | Project Leader functionalities | 133 |
| 7.5.1 | Project creation and corpus integration | 134 |
| 7.5.2 | Project configuration and management | 135 |
| 7.5.3 | Transcription protocols | 136 |
| 7.6 | Discussion on limits and improvements | 138 |
| 7.7 | Conclusions and next steps | 141 |

Chapter 8 Presentation of PHuN-ET 143

| | | |
|-----|---------------------------|-----|
| 8.1 | Chapter Summary | 143 |
| 8.2 | Introduction | 145 |

| | | |
|-------------------|--|------------|
| 8.3 | Premise for the PHuN-ET platform | 146 |
| 8.3.1 | Focus on experimental research | 147 |
| 8.4 | Editor functionalities | 149 |
| 8.5 | Identified User functionalities | 150 |
| 8.5.1 | User accounts | 150 |
| 8.5.2 | Sequential access to pages | 150 |
| 8.5.3 | Transcription instructions | 151 |
| 8.5.4 | Transcription interface | 152 |
| 8.5.5 | Data visualisation and sharing | 152 |
| 8.6 | Conclusion | 153 |
| Chapter 9 | Beyond the Platform- Human considerations | 157 |
| 9.1 | Chapter Summary | 157 |
| 9.2 | Collaboration | 159 |
| 9.3 | Motivations | 162 |
| 9.4 | Communication and Outreach | 166 |
| 9.5 | Skills and Training | 169 |
| 9.5.1 | Training volunteers | 169 |
| 9.5.2 | Online instruction | 170 |
| 9.6 | Chapter Summary and Conclusion | 172 |
| Part IV | Demonstration of experimental results | 173 |
| Chapter 10 | Quality assurance for crowdsourced production | 177 |
| 10.1 | Existing methods of quality assurance | 179 |
| 10.2 | Task-based QA | 183 |
| 10.2.1 | Gold standards | 184 |

| | | |
|-------------------|--|------------|
| 10.2.2 | Worker screening and training | 185 |
| 10.3 | Feedback-based QA | 185 |
| 10.3.1 | Expert feedback | 185 |
| 10.3.2 | Peer feedback | 186 |
| 10.3.3 | Automatic live feedback | 187 |
| 10.4 | Production-based QA | 190 |
| 10.4.1 | Multiple productions | 190 |
| 10.5 | Conclusion | 191 |
| Chapter 11 | Measuring transcription quality | 195 |
| 11.1 | Evaluation of transcription quality | 197 |
| 11.1.1 | Primary conjectures | 198 |
| 11.2 | Experimentation on the Stendhal Corpus | 200 |
| 11.2.1 | Stendhal Experiment 1 | 200 |
| 11.2.2 | Sample description | 200 |
| 11.2.3 | Phylogenetic analysis | 203 |
| 11.2.4 | Digging into the data | 211 |
| 11.2.5 | Study of expert transcriptions | 214 |
| 11.2.6 | Observing effects of page variation and complexity | 218 |
| 11.2.7 | Drawing preliminary conclusions | 219 |
| 11.3 | Experimentation on the Benoîte Groult Corpus | 221 |
| 11.3.1 | Benoîte Groult Experiment 1 | 221 |
| 11.3.2 | Benoîte Groult Experiment 2 | 224 |
| 11.4 | Conclusion | 228 |
| Chapter 12 | Measuring factors of complexity | 233 |
| 12.1 | Introduction | 234 |

| | | |
|--|--|------------|
| 12.1.1 | Design of Experiments (DOE) | 237 |
| 12.1.2 | Benoîte Groult Experiment 3 | 239 |
| 12.2 | Data analysis and results | 242 |
| 12.2.1 | Discussion | 244 |
| 12.3 | Conclusion | 247 |
| Conclusions and perspectives | | 249 |
| 12.4 | Conclusions | 251 |
| 12.5 | Perspectives | 256 |
| Chapter 13 French Summary of the Thesis | | 261 |
| 13.1 | Introduction | 262 |
| 13.2 | Contexte historique | 264 |
| 13.3 | Définitions | 265 |
| 13.3.1 | Les Humanités Numériques | 265 |
| 13.3.2 | Crowdsourcing, Sciences Citoyennes, et Humanités Cito-yennes | 266 |
| 13.3.3 | La transcription de manuscrits | 267 |
| 13.4 | Encodage des textes et outils de transcription | 269 |
| 13.5 | Architectures et interfaces | 271 |
| 13.6 | Mesures de différence entre transcriptions | 272 |
| 13.6.1 | Mesurer les différences entre textes : distance de Levenshtein . . | 273 |
| 13.6.2 | Mesurer les différences entre documents XML | 274 |
| 13.6.3 | Techniques de clustering | 274 |
| 13.6.4 | Visualisation | 278 |
| 13.6.5 | L'ensemble du processus | 278 |
| 13.7 | Types de prototypes | 279 |

| | | |
|---------|---|-----|
| 13.8 | Présentation de PHuN 2.0 | 280 |
| 13.8.1 | Administration des projets | 280 |
| 13.8.2 | Participation aux projets | 281 |
| 13.8.3 | Les fonctionnalités de l'éditeur de transcription | 282 |
| 13.8.4 | Conclusions | 283 |
| 13.9 | Présentation de PHuN-ET | 284 |
| 13.9.1 | Navigation dans la plateforme | 284 |
| 13.9.2 | Administration des projets | 284 |
| 13.9.3 | Participation aux projets | 285 |
| 13.10 | Au-delà des plateformes | 285 |
| 13.10.1 | Collaboration | 286 |
| 13.10.2 | Motivations | 286 |
| 13.10.3 | Communication et sensibilisation | 287 |
| 13.10.4 | Compétences et formation | 288 |
| 13.11 | L'assurance qualité pour le crowdsourcing | 289 |
| 13.11.1 | Assurance qualité fondée sur la tâche | 290 |
| 13.11.2 | Assurance qualité reposant sur la rétroaction | 290 |
| 13.11.3 | Assurance qualité reposant sur le produit | 291 |
| 13.12 | Mesurer la qualité des transcriptions | 291 |
| 13.12.1 | Expérimentation sur Stendhal | 292 |
| 13.12.2 | Expérimentations sur Benoîte Groult | 293 |
| 13.13 | Mesurer les facteurs de complexité | 294 |
| 13.13.1 | Le plan d'expériences | 295 |
| 13.13.2 | L'analyse des données | 297 |
| 13.14 | Conclusion et perspectives | 298 |

| | |
|--|------------|
| Annexes | 311 |
| Appendix A Annex A | 313 |
| A.1 Instructions for Stendhal | 313 |
| A.2 Instructions for Benoîte Groult - online at PHuN 2.0 | 315 |
| A.3 Instructions for Benoîte Groult Workshops | 317 |
| Appendix B Annex B | 319 |
| B.1 Stendhal Experiment 1 - page 2 | 320 |
| B.2 Recovered data for Benoîte Groult Experiment 1 | 322 |
| B.3 Benoîte Groult Workshops | 324 |
| Appendix C Annex C | 327 |
| C.1 PHuN 2.0 Participant list | 327 |
| C.2 PHuN-ET Participant list | 328 |

List of Figures

| | | |
|-----|--|----|
| 1.1 | Word cloud of words used in 40 crowdsourcing definitions | 15 |
| 1.2 | Results of Tesseract-ocr on a handwritten page from Stendhal's collection. | 23 |
| 1.3 | Editorial process with transcriptions as input. | 26 |
| 2.1 | Typology of knowledge production. | 37 |
| 2.2 | Zooniverse's- AnnoTate | 40 |
| 3.1 | XML encoded text represented as a tree | 58 |
| 3.2 | Metadata Types. | 59 |
| 3.3 | Example of annotation. | 60 |
| 3.4 | XML element overlap. | 61 |
| 3.5 | XSLT for converting XML to TEI-XML | 64 |
| 3.6 | Example of a linearised transcription. | 66 |
| 3.7 | Example of a pseudo-diplomatic transcription. | 67 |
| 4.1 | Process overview of a CMS. | 78 |
| 5.1 | Relationship between reading and transcription interfaces | 85 |
| 5.2 | Example of an INKE reading interface | 86 |
| 5.3 | Example of Folger Library's Digital Texts reading interface | 87 |

| | | |
|------|--|-----|
| 5.4 | Text encoding: Bentham transcription desk example. | 93 |
| 5.5 | Annotate's editor. | 94 |
| 6.1 | Observing text differences in two texts. | 98 |
| 6.2 | Measuring text difference. | 100 |
| 6.3 | Tree operations as described by [Chawathe et al., 1996]. | 102 |
| 6.4 | Nierman and Jagadish sub tree operations. | 103 |
| 6.5 | Tree operations as described by [Zhang and Shasha, 1989]. | 104 |
| 6.7 | Cluster linkages: single and complete links. | 109 |
| 6.8 | Schematic flow of transcription analysis. | 110 |
| 7.1 | Illustration of MVC framework. | 122 |
| 7.2 | PHuN 2.0 site architecture. | 123 |
| 7.3 | Benoîte Groult editor configuration example. | 126 |
| 7.4 | Editor configuration. | 129 |
| 7.5 | Transcription work flow in PHuN 2.0. | 131 |
| 8.1 | PHuN-ET's site work flow. | 147 |
| 8.2 | Transcription workflow in PHuN-ET. | 148 |
| 10.1 | Quality assurance through tasks, feedback and production in a crowdsource- ing environment. | 183 |
| 10.2 | Comparison of existing feedback mechanisms | 189 |
| 11.1 | Novice and expert errors compared to an ideal transcription. | 200 |
| 11.2 | Page from Stendhal experiment 1- page 1. | 201 |
| 11.3 | Page from Stendhal experiment 1- page 2. | 202 |

| | |
|---|-----|
| 11.4 Distance matrix with associated phylogenetic tree- raw text. | 206 |
| 11.6 Result of raw text distance measurement on transcriptions for Stendhal's page 1, without marginalia; folio annotations, etcetera. | 210 |
| 11.7 Results of linear regression performed on three experts. | 216 |
| 11.8 Results of linear regression for expert 1 and four novices. | 217 |
| 11.9 Results of linear regression for Expert 1 and Novice 9. | 218 |
| 11.10 Page from the Benoîte Groult Corpus. | 231 |
| 11.11 Phylogenetic tree based on texts. | 232 |
| 11.12 Phylogenetic tree based on XML distance. | 232 |
| 12.1 Three groups of complexity factors. | 234 |
| 12.2 Experiments for the full factorial plan. | 241 |
| 12.3 Regression coefficients | 244 |
| 12.4 Effect of factors on the observed errors. | 245 |
| 12.5 Estimated page classification based on sample. | 246 |
| 13.2 Critères de lien : liens unique et complètes. | 277 |
| 13.3 Flux schématique de l'analyse de la transcription. | 279 |
| 13.4 Flux de transcription dans PHuN 2.0. | 282 |
| 13.5 L'assurance qualité du point de vue des tâches, de la rétroaction, et du produit dans un environnement de crowdsourcing. | 289 |
| A.1 1st instructional figure. | 314 |
| A.2 2nd instructional figure. | 315 |
| A.3 Instructions for transcribing Benoîte Groult, online at PHuN 2.0. | 316 |

B.1 Results of distance analysis on raw text for the Stendhal corpus, page 2. . . 320

B.2 Results of distance analysis on raw text for the Stendhal corpus, page 2. . . 321

B.3 Results for page 011 and page 014, folder 03. 323

B.4 Results of distance analysis on raw text for the Benoîte Groult workshops,
page 01, folder 05. 324

B.5 Results of distance analysis on xml documents for the Benoîte Groult work-
shops, page 01, folder 05. 325

List of Tables

| | | |
|------|---|-----|
| 5.1 | Five web task taxonomies | 89 |
| 6.1 | Complexity of presented algorithms. | 105 |
| 8.1 | Resumé of differences between platforms PHuN 2.0 and PHuN-ET. | 155 |
| 11.1 | Resumé of intra and inter cluster distance averages. | 205 |
| 11.2 | Expert element placements. | 212 |
| 11.3 | Table showing results of element analysis for Stendhal’s page 1. | 213 |
| 11.4 | Comparing expert and novice <i>douteux (doubtful)</i> element placement. | 214 |
| 12.1 | Experiment design plan using two factors. | 240 |
| 12.2 | Distances from novices to expert for each pages of the DOE. The grey column indicates the user’s id in the database. | 243 |

Introduction

Knowledge creation and dissemination have relied on paper-based formats for at least the last several millenia, and have a rich history since their beginnings in ancient Egypt and the Mediterranean. With the introduction of the printing press and, now, digital technology, knowledge dissemination continues to gather momentum even as it branches out into new mediums and accepts new formats.

With new technologies we increasingly observe the implication of the public in activities of knowledge creation and dissemination that otherwise would not be possible. Manuscript transcription in particular is an activity that allows to constitute digital textual data from paper-based formats on a much greater scale than before (the possibility of calling on interested members of the public allows to diminish costs associated with these processes). This, in turn, brings about social and socio-cultural changes, which we began observing with the introduction of the social web, and which we will likely continue to observe in years to come. With greater implication from volunteers, digital textual resources are growing and will inevitably continue to do so in a Big Data kind of way.

Participants taking part in manuscript transcription are people who are not necessarily experienced, but who take on and accomplish tasks proposed by project leaders to engage in activities they are passionate about. Only, what may one expect of their effort? Answers to questions concerning the quality of crowdsourced manuscript transcriptions for purposes of scholarly editing are insufficient in that there are no proposed methods or measures to monitor quality. This makes it difficult to observe the effects of modifications to components making up participative (or contributive) workflows on the quality of documents obtained. For instance, changes to structure or vocabularies of descrip-

tive schemas, changes to instructions or participant training, or challenges pertaining to manuscript objects themselves can significantly affect contributor output. There is still no precise way of studying these to better anticipate needs, and better meet expectations of projects that can benefit from the public's implication in editorial– or transcriberial– processes.

How does one define and manage a transcription task so as to improve the results of participants' efforts? Until now, many projects still do not know what results to expect if they ask inexperienced individuals to transcribe their manuscript collections.

Two viewpoints can be summarised and contested. Firstly, that of a number of experts who consider that transcriptions collected in this way will be inaccurate and of lower quality (and one would not be able to use them in publications or as the basis of scholarly research). And that of others, many of them experts themselves, who think that contributions obtained will be sufficiently accurate and useful, as well as possibly bringing new information to light about the object transcribed. The second viewpoint is illustrated below.

Il serait sans doute démagogique de promettre à tout un chacun qu'il saura déchiffrer séance tenante l'écriture de Pascal ou de Stendhal mais il n'est pas inconcevable que tel amateur de bonne volonté puisse suggérer une lecture pertinente de tel passage difficile, dans lequel la fraîcheur de sa perception aura su distinguer ce que des chercheurs plus aguerris n'avaient pas perçu.
[Leriche and Meynard, 2008].

In the work to be presented here we are interested in exploring, creating, and experimenting with tools and methods that can shed light on these questions, specifically in regards to crowdsourcing manuscript transcriptions. To do this, we have created a digital platform, both an experimental prototype to collect transcription data, as well as a work environment for project leaders and individuals interested in participating in these processes.

1 Research questions

Extended work on the evaluation of crowdsourcing for manuscript transcription is insufficient. If scientific fields have embraced experimentation and evaluation of participative science, the humanities have begun this work only recently. In France, where this doctoral work is being carried out, the previous statement holds even more true. Few humanities scholars undertake large-scale projects that make use of the possibilities offered by public participation, not knowing the potential of contributions from the general public or how to put in place this type of project. This makes it difficult to measure the efficacy of crowdsourcing and its potential for manuscript transcription (to speed up the work of scholars by increasing transcription yield). There is a sense of enthusiasm from the part of scholars about this potential, but there are also questions about whether novices and hobbyists can produce corpora which will be of sufficient quality to use as a basis for research and scholarly publishing [Ghafele et al., 2011 ; Cohn, 2008 ; Franzoni and Sauermann, 2014].

Our work will explore the possibilities of evaluating the efficacy of crowdsourcing for humanities' transcription projects based on work contributed by inexperienced, or novice, transcribers. Using information collected with our digital transcription platform we will show how one can evaluate the results of crowdsourced transcriptions and discuss the potential of these methods to support larger initiatives and the benefits that can be derived therefrom.

2 Thesis plan

This dissertation is composed of four parts. In the first, we present the context in which this work has developed, including definitions of concepts which we will use throughout. We situate participative manuscript transcription as an activity residing within Digital Humanities, which employs a method widely known as Crowdsourcing, and which can be applied to a number of activities in disparate fields. For example, we will explain how

Citizen or Crowd Science has made use of crowdsourcing and also use the term Citizen Humanities to refer to similar initiatives in the humanities. It is within this sphere that participative manuscript transcription fits, as part of what we call Citizen Scholarly Editing activities.

In the second part we present the technical foundations on which we base our work, including how XML metadata can be used in the context of the dynamic web. We also discuss tools used for transcription and environments that have allowed to coordinate contributions from many users. Finally, we present techniques for comparing multiple transcriptions, the objective being to measure data quality.

In the third part we present the digital platform prototypes that we have created and the functionalities that we put in place. We also discuss the difference between production-driven and experimentation-driven prototypes, which are important to apprehend in a Digital Humanities context and in order to learn from the prototyping process.

In the fourth and final part we present the results of our experiments based on the methods we use for analysing contributed data. The first experiment focuses on the Stendhal Corpus and subsequent ones on the Benoîte Groult Corpus. Over the course of our work we also had the opportunity to work with manuscripts of authors such as Michel Butor and Jean-Philippe Toussaint, and the knowledge gained from these experiences will be referred to more generally where appropriate. The numerical analysis that we perform on Stendhal and Benoîte Groult allows us to assess the quality of transcriptions we obtained using the crowdsourcing method and compare them to our expert references.

The knowledge that we are able to gather as a result of our methods of analysis can contribute to an enriched understanding of crowdsourced manuscript transcription on multiple levels. Firstly, knowing where these methods are appropriate and how project leaders can intervene to achieve better results. Secondly, how digital technology and computational techniques can contribute to create smarter ecosystems within which inexperienced transcribers benefit as much as project leaders from the transcription effort.

Part I

Manuscript transcription for Digital Humanists

Part I Summary

This first section consists of two chapters. In the first we introduce the main elements that direct this work. To begin we describe the activity of transcription, and more importantly manuscript transcription. We continue by establishing our definitions of the terms Digital Humanities, Crowdsourcing, and finally Citizen Science and its humanities counterpart, Citizen Humanities.

The second chapter introduces the technological and scientific context within which a growing volume of research and scholarship operate with greater openness to the public. This chapter will introduce examples of projects from a variety of academic spheres that have been an influence within digital humanities and therefore on our work in this dissertation. To conclude the chapter and the first part, we will summarise the contributions made so far to our field of interest and identify what still needs to be done, thereby setting the tone for the work presented in the following chapters.

Chapter 1

Introduction to Digital Humanities

Contents

| | | |
|------------|---|-----------|
| 1.1 | Digital Transformations | 10 |
| 1.1.1 | Definition of Digital Humanities | 11 |
| 1.1.2 | Crowdsourcing | 13 |
| 1.1.3 | Citizen science and Citizen Scholarly Editing | 16 |
| 1.2 | Manuscript transcription | 18 |
| 1.2.1 | Introduction to manuscript transcription | 18 |
| 1.2.2 | Transcription as decoding and encoding | 20 |
| 1.2.3 | Manuscript transcription in a digital context | 21 |
| 1.2.4 | Manuscript transcription in a participative context | 23 |

1.1 Digital Transformations

Innovation in digital technologies has had a significant role in transforming the world of publishing as we imagined it prior. Its effects extend well beyond publishing and into the very fabric that makes up scholarship in the Humanities, precisely because it affects knowledge production and dissemination. Like many other fields, today's humanities scholarship relies on information accumulated over a long history of scholarship. The humanities draws on a vast bank of knowledge, regrouping fields including history, philosophy, anthropology, archeology, classical studies, languages and linguistics, but also literature, politics, art history, and visual and performing arts. These fields are considered foremost as fields of scholarship, and in a secondary way as ones of practice. Methodologies in the humanities are largely distinguishable from experimentation and empirical studies, which are associated with natural, or "hard" sciences. The humanities developed out of scholarly traditions that were based in historical, critical, and comparative analyses of records of information. With the introduction of digital technologies we are indeed observing a shift in the humanities. This shift is said to be changing scholarship in very tangible ways, precisely because it is introducing new practices and new methodologies for research in the humanities. Changes in the world of humanities scholarship are accelerating at the rate of digital innovation and many scholars have witnessed and documented their observations.

Digital technology has engendered a profound transformation of the patterns of production and circulation of content that we have known since the eighteenth century. The web, in particular, has brought about a major upheaval of the very meaning of content: we were in an economy of scarcity, we are today in a superabundance of information. The instances of choice, evaluation and distribution of content were centralized in the hands of certain private or public institutions which were the guarantors; today, legitimisation systems seem absent or unstructured [Vitali-Rosati and Sinatra, 2014]¹.

1. Author's translation from [Vitali-Rosati and Sinatra, 2014]. Original text: Le numérique a engendré une transformation profonde des modèles de production et de circulation des contenus que nous

In light of these important changes to production and dissemination of information, on which humanities scholarship is founded, and which is a product of digital technologies, we will look at how Digital Humanities make use of technological and human elements in contemporary academic contexts. For this, we will first set down some definitions of terms like Digital Humanities, Crowdsourcing, Citizen Science, and our own Citizen Scholarly Editing, which constitute the conceptual landscape within which our work has developed.

1.1.1 Definition of Digital Humanities

Digital Humanities (DH) can be viewed simply as the result of incorporating computing into the humanities, though many scholars would be unsatisfied with this definition [Burdick et al., 2012]. Digital Humanities are not a discipline all to themselves, but should be considered as an approach to practicing research in the humanities [Vitali-Rosati and Sinatra, 2014]. We'd like to take a closer look at the properties commonly attributed to Digital Humanities so as to provide a fitting definition for the ways that Digital Humanities relate to our work.

The Humanities regroup a number of disciplines that focus on that which is generally defined as having an interest for human beings; history, society, culture, and its activities, artifacts, and records. The incorporation of computing, or digital technology, into the humanities allows to scaffold what is proper to humanistic approaches of scholarship by computational methods. Under these conditions a newer generation of humanities scholarship is able to develop.

As [Vitali-Rosati and Sinatra, 2014] propose, today's Digital Humanities have a history in computing for the humanities and social sciences. Some attentive searching will quickly unearth Digital Humanities' antecedent, Humanities Computing [Vitali-Rosati and Sina-

connaissons depuis le xviii^e siècle. Le web, en particulier, a déterminé un bouleversement majeur du sens même des contenus : nous étions dans une économie de la rareté, nous sommes aujourd'hui dans une surabondance d'informations. Les instances de choix, d'évaluation et de distribution des contenus étaient centralisées dans les mains de certaines institutions privées ou publiques qui en étaient les garants ; aujourd'hui, les systèmes de légitimation semblent absents ou déstructurés.

tra, 2014 ; Siemens et al., 2009 ; Svensson, 2009]. Humanities Computing introduced practices of computing, calculation, and data processing to the humanities, bringing with it new ways of working with research materials in humanities disciplines. From Humanities Computing to Digital Humanities the change from suffix to prefix and synonymic shift are subtle and may simply imply fluctuating terminology. However, we'd like to suggest that the change to Digital Humanities reflects an evolution in practicing research in the humanities. This aligns with what are referred to as new modes of scholarship [Burdick et al., 2012]. It is the subtle difference between applying exterior methods to a discipline that doesn't rely on them traditionally, and reflecting how digital technologies have really anchored, or taken root, in the humanities. Interdisciplinarity and collaboration are key constituents in Digital Humanities [Vitali-Rosati and Sinatra, 2014 ; Burdick et al., 2012 ; Fitzpatrick, 2011].

Humanities have as many reasons for collaborating with information technology as any of the hard sciences to ensure its own relevance in the decades to come. This relationship can be questioned and interrogated by seasoned specialists for the purposes of epistemological debate, but students and those entering the field have a stake in this relevance. Like the humanities, digital humanities should maintain its interest in the humanistic. Likewise, those practicing Digital Humanities should be aware that they are operating from a specific perspective, that positions how one thinks humanities disciplines should, or must, react to growing digital technologies, changing research landscapes, and expectations in humanities research [Vitali-Rosati and Sinatra, 2014].

Digital Humanities practices are grounded in data processing and in exploring and creating tools for new ways of conducting research [Vitali-Rosati and Sinatra, 2014 ; Siemens et al., 2009 ; Fitzpatrick, 2011]. The relationship is both creative and analytic. It can allow for other activities, including information retrieval, curating collections, text mining, mapping, data visualisation and a host of others. Digital Humanities allow new practices based on materials that are of interest to humanists, but with the possibility to create new connections between concepts, new perspectives, new interactions, and even new questions for research. An important focus in the Digital Humanities is the work

to open access to data and integrate the public in research projects in meaningful ways [Burdick et al., 2012].

For our purposes, all of these elements, technological practices, people, relationships and collaboration between disciplines, as well as perspectives for conducting research are important in defining Digital Humanities. We will thus define Digital Humanities as a set of practices based in digital data and content, which can include using digital tools to transform objects to create new knowledge and knowledge resources, and should also include applications and methods of sharing knowledge more broadly within humanities disciplines. Acknowledging technological practices for conducting and disseminating research and knowledge in the humanities has been vital in orienting our work, which finds itself at the junction between literary activities of scholarly transcription and computing methods grounded in IT. In our case, doing work in DH has meant both creating tools for transcription and analysing the results.

1.1.2 Crowdsourcing

Despite frequent discussion and definition of the term crowdsourcing, scholars remain unsatisfied with the definitions proposed in scientific discourse [Franzoni and Sauermann, 2014 ; Estellés-Arolas and González-Ladrón-De-Guevara, 2012]. Retracing the term back to its first use by Jeff Howe in an article in 2006 accords an opportunity to mark the starting point from which the term has evolved. The word crowdsourcing has since been applied to a number of different fields, both in industry and research, and has been refined – and redefined – to incorporate the different characteristics particular to each field in which it has been used. Yet, to begin with, the word’s introduction in *Wired*, an American magazine focusing on new technologies’ effects on economy and culture, created a context that is at once general and specifically marked by the technology of the internet. In 2006 Howe defined it as any mode of online production deriving from an open call for participation, whether solicited by private or public institutions [Howe, 2006]. According to Howe, this mode of production was set to change the way people worked all over the world, with significant implications for the world’s economies. What is clear with this

broad stroke definition is the scope of crowdsourcing's potential influence. At the same time, the shade cast by crowdsourcing's definition leaves significant room for ambiguity, as can be observed in the following statements:

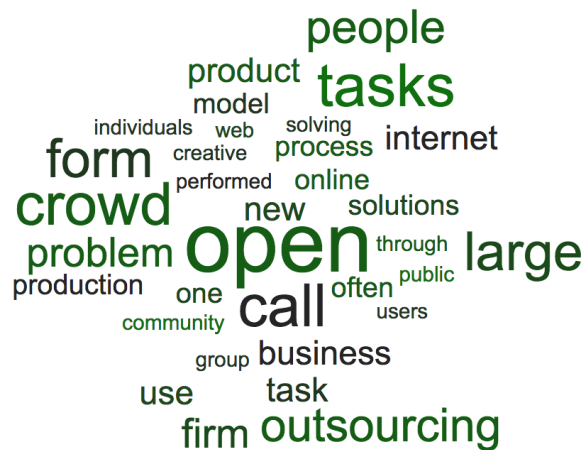
« Crowdsourcing is an ill-defined but common term referring to a set of distributed production models that make an open call for contributions from a large, undefined network of people [Wiggins and Crowston, 2011]. »

« Depending upon the perspective and the definition used, certain initiatives classified by some authors as crowdsourcing are not classified as such by others [Estellés-Arolas and González-Ladrón-De-Guevara, 2012]. »

Crowdsourcing definitions can vary so much that they may not only be divergent but even contradictory [Estellés-Arolas and González-Ladrón-De-Guevara, 2012]. As a result, a number of scholars have sought to clarify this term, adding essential and optional characteristics to outline and to highlight its versatile nature. Notably, Estellés-Arolas and González-Ladrón-De-Guevara (2012) compare various crowdsourcing definitions in order to enhance their understanding of the term and how it is described by practitioners.

One of these definitions incorporates the concept of collective intelligence and the advantages of involving more people in problem solving tasks [Buecheler et al., 2010]. James Surowiecki's book describes this concept. In it, he presents examples of extremely difficult tasks that were accomplished successfully because groups of people collaborated on generating solutions, even privileging groups composed of people considered to have average intelligence over those considered as having high intelligence, as a method of attacking problems from a multitude of perspectives.

In Estellés-Arolas and González-Ladrón-De-Guevara's 2012 article, based on forty definitions from thirty-two articles, at least half of the definitions mention involvement from people. We noted words referring to groups, networked people and individuals (6), to communities (4), to the public (4) and the crowd (7). Some refer to "networks of people" [Howe, 2006], "networked people" [Vukovic, 2009], others to a "general internet public" [Kleemann et al., 2008] or "loosely bound public" [Wexler, 2011], and others still to



"large-scale communities" [DiPalantino and Vojnovic, 2009] and "organized communities" [Chanal and Caron-Fasan, 2008]. Based on this, one can see a general trend beginning with motivated and unrestrained individuals and developing into organized, and potentially vast, collectives.

Also, users and advocates of the method argue that crowdsourcing should be undertaken in as much an open and decentralised system as possible [Ghafele et al., 2011]. The words highlighted by our word cloud seem to indicate this too. But reader beware.

2. We chose this minimum based on the number of times the words "public" appeared in the definition. If we raise the minimum frequency to 10, we are left with only 4 words: open, crowd, tasks, and call.

As many already know and would gladly point out, crowdsourcing is frequently used by private enterprises for their their own ends, which returns often in definitions as well. A question that one may ask is how can crowdsourcing also include public institutions that do not seek to make a profit from the work of contributors, but need the public to improve their products and services? Other authors insist on the importance of a central focus for crowdsourcing, as seen in the description below.

An interface enabling users to (for example) annotate/tag and suggest links without focus is not crowdsourcing; the focus on a shared task or purpose is critical. This relates to an observation that the more closely defined the task is, the more successful it will be. [Dunn and Hedges, 2012]

Howe's original term refers to an activity that has broad applications, without specific reference to industry or discipline, so it may very well be taken up by public institutions also. And it has, with some twists which include introducing other terms: Crowd Science, Citizen Science, and even Citizen Humanities. In the next section we will take one step further and propose another term, Citizen Scholarly Editing (CSE), and explain how it can be appropriate for crowdsourced manuscript transcription³ in the humanities.

1.1.3 Citizen science and Citizen Scholarly Editing

The term *Citizen Science* has been used to refer specifically to scientific research projects that solicit contribution from the public, most often with an online website or platform as an interface between contributing members and scientific experts. In fact, the majority of well-known crowdsourcing projects such as those hosted by Zooniverse are referred to as citizen science projects, specifically because they have a scientific component and because they involve the public. The term *Citizen* connotes a certain degree of involvement within a public community⁴.

Simply put, Citizen Science is the result of public institutions using crowdsourcing

3. A detailed definition and description of manuscript transcription can be found in Section 1.2

4. As opposed to an enterprise or private company.

to collect information that is used for public scientific research. As we discussed in the previous section, crowdsourcing is an open call for participation in a specific activity of production put out by a sponsoring actor. In this case, scientific research institutions are the ones making an open call. Participants that respond to this call, and become involved, contribute to scientific activities that can benefit both the institutions and, inevitably, the people and communities served by these institutions.

Practicing or engaging in citizen science can be understood as participating in a category of activity that benefits scientific research. Like branches in scientific disciplines, these categories can have subcategories or sibling categories. We note that as there exists a distinction between sciences and humanities, other terms may be appropriate when referring to activities wherein crowdsourcing is made use of by institutions in the Humanities. *Citizen Humanities* is a term that already circulates online and in certain communities practicing Digital Humanities [Communities, 2016]. For example, although the Tate Museum’s AnnoTATE project is featured on the Zooniverse website among its many other citizen science projects, AnnoTate can also fit into the *Citizen Humanities* category.

We propose *Citizen Scholarly Editing (CSE)* then to refer more specifically to scholarly editing projects that also make use of crowdsourcing to constitute documents and build corpora in which their scholarly and editorial activities are rooted. Manuscript transcription, which we will introduce in the following section, can be viewed as an integral part of the processes making up CSE.

Throughout this work we often use the terms scientific and scholarly interchangeably to refer to the work of scholarly editors to emphasize the expert dimension of their work. Meanwhile we are clearly situated within the humanities sphere where *scientific* is an adjective used to refer to work requiring expert knowledge and training. We are not referring to scientific work as that which can be situated in fields such as biology, chemistry, physics, and related spheres.

1.2 Manuscript transcription

Manuscript transcription is an important activity among numerous other methods of conservation. Digitization or *numerisation* methods have become a means to working with manuscripts and other forms of artifactual objects for a wide range of disciplines. Digitization allows researchers to work with documents that are otherwise rare and difficult to access. In many cases, transcription is an essential passing stone to other digital processes, including research and editorial processes, which constitute specific areas of practice within Digital Humanities.

1.2.1 Introduction to manuscript transcription

Transcription may be understood in a number of ways, depending on the object or type of data being transcribed, who is transcribing and for what ends. In linguistics and social sciences audio recordings can be transcribed to obtain an associated text recording – often easier to interpret, translate or use for linguistic analysis. Sociologists and political scientists transcribe both audio and video files to study and interpret linguistic acts. Thousands of television series and films are transcribed in order to be translated into dozens of different languages worldwide. Legal proceedings are transcribed by professionals to ensure a written record of statements and events. Some professional writers are historically known to dictate statements to be transcribed by designated secretaries. All of these examples demonstrate a transformation of auditory and or visual information towards a written record of information. Nevertheless, there exists also the notion of transcription from one written document to another and this practice itself has a long history in literary studies.

In literary and textual studies, transcription is an editorial and or genetic practice. Editors aim to constitute editions of text from authors' drafts. Text geneticists (textual scholars or critics depending on the school) work to study the process of the text's creation, from its earliest drafts all the way to known scholarly editions, including conflicting ones, or even possible future ones. The drafts themselves, which it is important to note, often

take form first on paper and are written in the hand of the author him or herself (though perhaps less and less so today with expansive use of word processing softwares).

In order to study authors' writing processes, it is common practice among scholars to transcribe the textual content found in manuscript pages, thus making evident how modifications were carried out (whether by processes of addition or correction) and which textual variants or word choices were supplanted for others. Transcribing documents can make evident the process of working on a text as a sequence of multiple drafts. Transcription may prove to be a task of some complexity since it aims to reconstruct texts while making observable processes of modification or drafting. Thus transcription warrants a closer look. And, in doing, so we may ask what makes this task more or less complex, and what factors may affect resulting transcriptions.

To do this we will follow a basic empirical questioning strategy and describe how answering the five basic questions – who, what where, when and how – can help characterize a transcription task, as well as distinguish a complex transcription task from a simpler one. We will begin by asking who is being transcribed. As it has been shown through projects such as Bentham and Manuscripts de Stendhal, the author is of primary interest. Firstly because the author represents both a time period and a subject or literary genre, factors that contribute to the specificity of the type of writing we are going to transcribe. The next question, 'What' is being transcribed, helps to distinguish the object of study as 19th century English philosophy from 19th century French literary realism. Beyond questions of period and genre, 'what' also helps determine the object of study, travel journal, philosophical treatise, letter, postcard, etcetera. The following two questions, 'where' and 'when' will the transcription process take place, correspond to contextual factors that are incredibly difficult to control, just as where and when reading activities or e-mail messaging can occur. Possible answers are in a library, in a personal study, a busy office, on a train and so on. Transcription can likely occur anywhere where a willing individual can have access to a desktop computer or place a laptop. 'How' is a more technical question as it relates more specifically to the activity itself and requires asking what tools are used, both to view the text and transcribe it, whether transcription conventions are defined and

respected, and according to which guidelines.

The last two questions are a repetition. Firstly, 'who' as it refers to the person transcribing, his or her knowledge of the manuscript, experience with the task and the disciplinary field to which that person belongs, if applicable. Secondly, 'what' as in to what purpose the resulting transcription will be put; whether for research, editorial, comparative or other representational goals. This final factor will have significant impact on the rest. If each of the questions are considered as factors affecting the process of transcription for a project, we can see how sophisticated and complex any one specific transcription project can turn out to be.

1.2.2 Transcription as decoding and encoding

Ambiguous markings, contextual information and enriched scientific commentary are all things that accompany transcriptions of texts. Editors and geneticists, whom we will refer to more generally as textual scholars or scholarly editors, accompany an author's text with information that helps enrich the reader's understanding of it. This information is derived from the text by the scholar's own work of deciphering, translating, cross-referencing and interpreting. We can thus refer to the work of the editor or scholar as an act of decoding. At the same time, scholars follow transcription conventions to annotate the text they reproduce. Transcription conventions are common practice and are even regulated in some disciplinary branches. We can, therefore, also consider it as an act of encoding, much in the same way as the act of writing encodes spoken language according to conventions of an alphabet. In addition, as we will see later, documents can be encoded according to XML conventions (see Chapter 3). At this juncture, parallels between linguistics and computer science are most apparent.

Practices of transcribing texts can extend beyond transcribing textual content, as in the act of encoding and enriching texts according to strictly or loosely defined rules. Specific encoding conventions are put in place in order to address the scholars' needs to study the texts and the editors' needs to represent and structure subsequent scholarly

editions.

Transcription is a way for scholars to study the processes of the creation of a text, not just as a finished product, which we are accustomed to reading, but as a work in progress. Transcribing a text allows to trace out its possibilities. By highlighting modifications to a text, we are bringing into focus all the other possible texts, as well as specific choices that were made (authorial and editorial) that may help complete and even challenge the final versions of that text. As described by Stuart Dunn and Mark Hedges (2012) in the following statement, modifications are crucial to literary transcription:

Transcribing is closely linked to correction and modification, and is currently one of the most high-profile areas of humanities crowd-sourcing, as it addresses directly one of the most fundamental problems with OCR: that handwriting, especially complex and/or difficult to read handwriting, cannot be automatically rendered into machine-readable form using current technology. It can only be transcribed manually with the human eye and, in many cases, with human interpretation [Dunn and Hedges, 2012].

In this paper, the authors place transcription at the top of the list of activities that humanities crowdsourcing projects are concerned with. They also identify the oppositions between the activity of transcription and available technologies for analysing handwriting, before reminding the reader of the essential human dimension in this type of work, and the important place occupied by human perception and interpretation.

1.2.3 Manuscript transcription in a digital context

The changes brought about by digital scholarship have in fact entailed an interest in integrating crowdsourcing as part of the editorial process that involves digitizing literary sources and transcribing them to produce new printed and digital scholarly editions. As a response to the sheer volume of documents and the precision required to process them, technology has been only partially successful in responding to the needs articulated by editors and researchers.

Technology based on optical character recognition (OCR) for example has achieved excellent results on some ancient medieval manuscripts. The task of automatically extracting text from digitized images of these types of manuscripts has been validated by multiple studies, achieving high accuracy scores in the ranges of 80 to 90 percent [Diem and Sablatnig, 2010, 2009]. The successful use of OCR on medieval manuscripts can be attributed to several decisive elements. The creation of medieval documents was a task undertaken by professional scribes of monastic orders— a task exercised with extreme care. Documents were also scrupulously copied from original versions. The act itself is seen as something deliberate and controlled. The resulting pages contain series of symbols that, even if written in an unknown language, can still be made recognisable by machines. Symbols belonging to the same category can be identified and regrouped and problems associated with varying handwriting styles are minimal, compared to those often encountered in more contemporary documents, such as skewed lines, slanted writing, and variations in character size [Espana-Boquera et al., 2011]. The main problems encountered by OCR in deciphering medieval manuscripts are due to factors such as age, deterioration, stains and poor quality images [Diem and Sablatnig, 2010]. Nevertheless, OCR systems have been successfully trained to handle a wide variety of problems associated with ancient and medieval manuscripts.

Unfortunately the same cannot be said for the distinct case of authors' work manuscripts. Work manuscripts have the particularity of being filled with modifications. The same modifications that expose writers' work processes make successful application of OCR difficult for these types of documents. Moreover, even on a document containing few such modifications, the constraints of OCR do not respond well to irregularities observed in more contemporary manuscripts [Espana-Boquera et al., 2011]. In other words, all the aspects that characterise authors' handwritten drafts, make accurate optical recognition of writers' writing very challenging. For example, the application of OCR to a nineteenth century manuscript from the Stendhal collection produces disheartening results, see Figure 1.2.

For the time being, the human element— that is the act of deciphering, reading

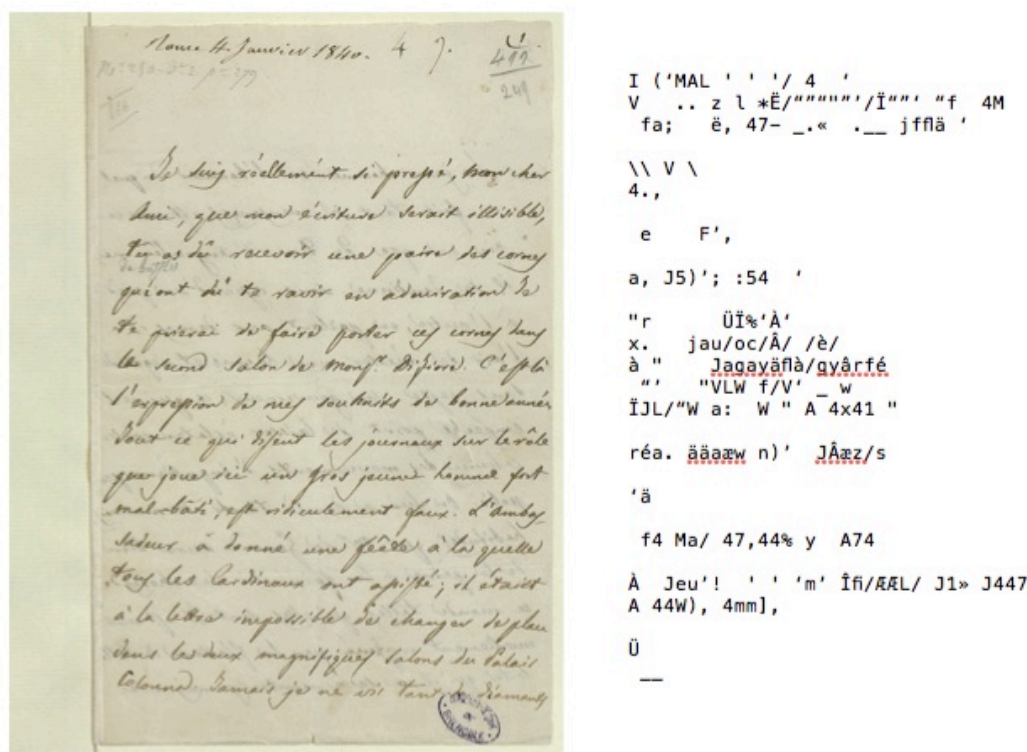


Figure 1.2 – Results obtained using an untrained Tesseract-ocr program on a handwritten page from Stendhal's collection. The page shown was taken from the web at www.manuscris-de-stendhal.org, R. 302, vol. 1, tome 2, feuillet 61, recto.

and interpreting— cannot be substituted by optical recognition technologies for these manuscripts. And at the same time, the sheer volume of documents needing to be processed justifies using greater means in order to facilitate, and yes expedite, these work processes.

1.2.4 Manuscript transcription in a participative context

To understand how to organize transcription in a participative context we must first distinguish it from other tasks that are commonly performed in this context. One can wonder if the act of transcribing is a creative process, and also whether it is a complex

process. The answers to these questions will define how transcription is handled in participative or CSE contexts. Furthermore, we can use this understanding to define what can be expected of participants.

Firstly, projects that crowdsource creative writing, including articles or reviews should be considered. Creative articles or blogs are notoriously difficult to evaluate for quality precisely because of the creative aspect—no two bloggers are alike and the same can be said of the content produced. Manuscript transcription, as opposed to creative writing, has the advantage of working from a source, a digital image that cannot be read by machines, but can indeed be deciphered by human intelligence. Unlike these creative tasks, manuscript transcription is an act of reproduction based on an existing source. This is also supported by its etymological definition. Moreover, if transcription is not a creative process, then one is more likely to accept that there should be a correct answer— or in our case a correct transcription⁵.

Previously, we described transcription as decoding and encoding. This means that there are indeed complex cognitive processes at work when we undertake such tasks. Only, generally speaking it is already widely accepted, and even considered in and of itself, that crowdsourcing and crowd science projects rely on human intelligence [Von Ahn et al., 2008].

Still, let us look at whether transcription is a complex process. Generally, complex processes can be broken down into series or sequences of simple steps, as is often the case in complex problem solving or even project management. When involving people in processes comprising of multiple steps, instructions are provided for each step and checkpoints can be built into the system to help guide workers through subsequent steps. Simple tasks can stand on their own or can be aligned into sequences to make up complex processes. For instance, transcription can be viewed as a step within an editorial process.

[Franzoni and Sauermann, 2014] consider that project tasks are either well-structured

5. Experts are no doubt looking for this ideal transcription. To test what participants produce, we can use a reference transcription provided by an expert, thus using a known solution, and ask multiple people to transcribe the same page.

or ill-structured, leading to either a low complexity task or a task that is accomplished incrementally with the work of contributors allowing "to develop a collective understanding of the problem space and of possible solutions over time".

One may consider it as a problem of scale. That is, one's definition will vary depending on where one is situated in the process. As shown by Figure 1.3, transcription is one step in a process whose objective is to achieve edited and validated documents. Then different procedures can be put in place to work with the resulting content. We borrowed the editorial process flow from the work of [Buard, 2015], and added more detail to account for inputs resulting from the work of transcription. We also qualify the outputs to different formats as part of distribution. What is important to retain here is that transcription itself is just one step within a complex editorial process that builds on the results of work that can be either done in-house or crowdsourced.

We would add that public participation can be applied at any point in the process once a system has been conceived to manage this. Likewise, one can imagine project leaders choosing the extent to which each component is open to public contributions, perhaps being open at one moment in the project's lifecycle and later limited only to project administrators, or vice versa.

Thus, transcription is the entry point to, or component of, a potentially complex editorial process, but itself should not be considered complex. It falls into the category of what are known as data entry and coding tasks, commonly employed in citizen science and citizen humanities projects to collect data and constitute corpora using public participation. However, even when considering the transcription as a simple task, one should not be mistaken about its importance:

One may think that these coding tasks are so simple that their performance should not be considered as a serious scientific contribution at all. However, data collection is an integral part of scientific research and much of the effort (and money) in traditional research projects is expended on data collection... [Franzoni and Sauermann, 2014].

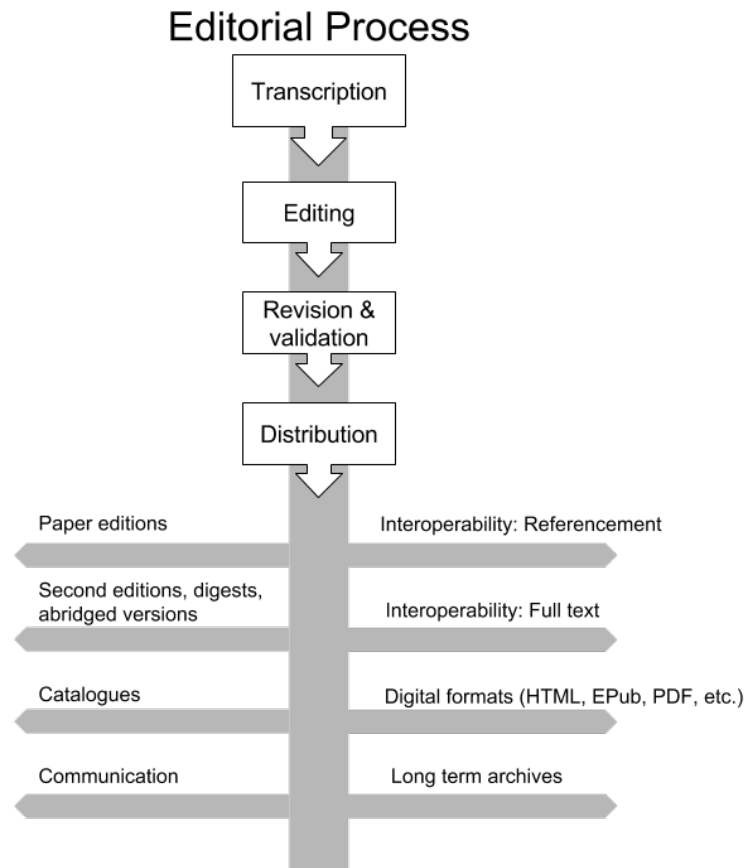


Figure 1.3 – Editorial process with transcriptions as input. The editorial process structure is taken from [Buard, 2015], we only add more detail to steps concerning the process following transcription, and we qualify the outputs to different formats as part of distribution.

And, in particular, its scientific relevance to the study of authors' works [Dufournaud, 2014].

[Franzoni and Sauermann, 2014] furthermore define "task complexity" as the extent to which tasks or subtasks are interdependent within a system. Based on this, it appears to us that it would be in the interest of participative projects to avoid exaggerating the complexity of transcription tasks to avoid discouraging participants from joining in the fun. For these reasons, presenting tasks as simply as possible and privileging the possibility to work independently while contributing to the overall process appears to be

of real interest to project success.

We know that experts look to obtain an optimal transcription and consider that an optimal transcription is possible to obtain. But how do transcriptions resulting from public participation measure up given the high expectations of expert groups? Gaining insight into what kinds of results can be obtained if transcriptions are crowdsourced may benefit organizations and communities that initiate more open scholarly editing projects.

In the next chapter we will provide some background into the history of print and origins of the collaborative web, which we consider to be essential contextual information for our subject of interest. Then, we will also describe contemporary projects that fall into categories of digital humanities, citizen sciences, scholarly editing, and even citizen scholarly editing. Finally, we will consider improvements that can be put in place so that scholarly editing projects can take fuller advantage of possibilities offered by public participation in manuscript transcription.

Chapter 2

Knowledge dissemination: from books to web

Contents

| | | |
|------------|---|-----------|
| 2.1 | Chapter Summary | 30 |
| 2.2 | History of print | 31 |
| 2.2.1 | Gutenberg | 31 |
| 2.2.2 | Facsimiles and conservation | 32 |
| 2.3 | Origins of the collaborative web | 33 |
| 2.4 | Contemporary challenges | 34 |
| 2.4.1 | Technological challenges | 34 |
| 2.4.2 | Scientific challenges | 35 |
| 2.4.3 | Participation | 38 |
| 2.5 | Existing projects | 38 |
| 2.5.1 | Citizen Science and Citizen Humanities projects | 38 |
| 2.5.2 | Digital Humanities scholarship projects | 43 |
| 2.6 | Inspiration and next steps | 46 |

2.1 Chapter Summary

We feel it is important to recount historical events that have shaped the practices of textual scholars specializing in manuscripts, as well as the important role of digital technologies and the web in modernizing these.

What follows is a brief, and by no means exhaustive, narration of these key events. We begin by taking a quick look at the history of manuscripts and print. Then, with the concept of knowledge dissemination in hand we look at how the development of the internet is able to shape most recent practices in manuscript scholarship, but also by association, reading and editorial practices, and finally knowledge management in general.

2.2 History of print

Undoubtedly, the study of manuscripts takes us back to the middle ages and, also, to the development of the first universities. The major details of events are described by Jean Baudet in a historical account of the development of techniques, on which we rely now to transmit the essence of events here.

The Middle Ages were a period which experienced a high demand for books, essentially to increase access to scholars and students. Because of this, by the end of the middle ages books and thus reading practices were no longer only limited to specialists, but emerged as leisure activities as well [Baudet, 2003]. Still, copies of books are painstakingly produced by hand and there aren't yet that many in circulation. At this point it is safe to say that reading is both leisure and luxury. Workshops where scribes— yes there was a time when scribe was a secular activity before it became a religious one— worked to produce handwritten manuscripts in large numbers [Baudet, 2003].

As further described in [Baudet, 2003], specific techniques for creating enlarged letters at the beginnings of texts were adopted by many scribes and, from a technical perspective, it is interesting to note that some even used wooden templates of engraved letters. These were small, but ingenious improvements that made the monotonous work of scribes faster and simpler. However, wooden templates only work well on enlarged letters or drawings; they aren't suitable for smaller script. The same principle of using wood engraving as templates, that is still commonly used in decorative arts today, is the basis for imminent improvements that would transform medieval methods of book-making [Baudet, 2003].

2.2.1 Gutenberg

Baudet also describes how, sometime in the 1450s, Johannes Gutenberg improves on medieval monks' letter templates by making them out of metal instead of wood, which allows him to use much smaller templates than previously possible. The templates are created by carving the tip of a metallic stem made of a relatively soft metal such as steel. The character is carved in relief and in reverse of the way the intended letter will be read,

and the resulting template acts as a stamp. One uses it to hit the surface of a softer metal, such as copper, and the imprint left over will be used for applying ink and pressing onto paper [Baudet, 2003].

Baudet relates a fascinating account of a process that is not only the predecessor of the industrial printing process, but can also be related to the practice of typography, wherein the graphic aspects of fonts are defined. Nowadays typography refers to font-making, but also other aspects such as the arrangement and disposition of characters on a page, which contribute to the overall presentation of a book. For traditional book editors, all aspects of layout, typography, choice of paper, and binding are essential for the creation of books. Just as traditional typography has a digital equivalent— a vast selection of web font libraries exist— so do the other aspects essential for book creation.

Although the creation of paper or digital books is not our primary focus in this dissertation, it is not with indifference that we observe the contribution of Digital Humanities to the work of renewing editorial practices to suit digital contexts of reading and writing.

2.2.2 Facsimiles and conservation

Some manuscript works are so rare or fragile that it is not possible that they be continually accessible to the greater public. Factors such as temperature and humidity in ambient air, the dangers of transferring oils from fingertips to paper, as well as other issues that accompany the handling of documents, are all causes for concern when it comes to preserving valuable manuscripts. The necessity to preserve actually takes documents out of circulation and reduces their accessibility. The possibility of making facsimiles, or copies, of rare documents offers opportunities for restoring access to documents, or rather their facsimiles, while preserving the originals. With progressive photographic techniques well beyond those of 18th century lithography, researchers that need to make detailed observations of documents can have access to high-quality digital copies. With the internet, they can have access to them virtually anywhere they can get a Wi-Fi signal.

2.3 Origins of the collaborative web

Web 2.0 is a term that is easily taken for granted now that major social web applications are officially entering their teen years: Facebook began in 2004, Youtube in 2005, and Google is actually 19 years old, having begun in 1998. At the time when Web 2.0 was still a novelty, at the end of the 90's and just after the turn of the 21st century, definitions used to explain Web 2.0 were as intriguing as the idea itself. For instance, in a 1999 article Darcy DiNucci, who coined the term, writes of the web :

The Web we know now, which loads into a browser window in essentially static screenfuls, is only an embryo of the Web to come. The first glimmerings of Web 2.0 are beginning to appear, and we are just starting to see how that embryo might develop. The Web will be understood not as screenfuls of text and graphics but as a transport mechanism, the ether through which interactivity happens [DiNucci, 1999].

Later, San Murugesan focuses the interest of Web 2.0 for the benefits of individuals and groups :

Web 2.0 is also called the wisdom Web, people-centric Web, participative Web, and read/write Web. Web 2.0 harnesses the Web in a more interactive and collaborative manner, emphasizing peers' social interaction and collective intelligence, and presents new opportunities for leveraging the Web and engaging its users more effectively. Within the last two to three years, Web 2.0, ignited by successful Web 2.0- based social applications such as MySpace, Flickr, and YouTube, has been forging new applications that were previously unimaginable [Murugesan, 2007].

The Web 2.0 is used to signify a web where users are as active in creating the text that is available freely on the internet as they are in reading the freely available text. Thus, collaboration is really a fundamental aspect of the nature of the internet, without which the web and the internet as we know it would not be the same. It is also a fundamental

aspect of new digital environments, which allow for user engagement within different contexts; where interaction with data and creation of content in a collaborative manner has become the norm [Vitali-Rosati and Sinatra, 2014].

Without a widespread and robust network infrastructure, none of the crowdsourcing, citizen science, and citizen humanities projects that exist today would be possible, let alone major digital conservation efforts led by museums and heritage institutions world wide. As a matter of fact, today's users of the internet have access to a majority of tools and services without having to understand much of the complex technical wiring behind them. Today's internet users are actively reading, writing, and participating in non-profit initiatives online. Many of them make up the support bases of successful citizen science projects.

2.4 Contemporary challenges

Certainly, recent technological advances are responsible for some important changes to reading and writing practices. Some of these changes can be disconcerting to older generations. However, they are not necessarily negative, nor do they signal mass intellectual decline [Coady, 2016].

Nevertheless, projects face several challenges when it comes to implementing research activities with greater openness to, and allowing for more involvement from, publics. We have identified three types, which we will discuss in more detail. These types are technological challenges, scientific challenges, and participation.

2.4.1 Technological challenges

A number of successful projects have shown that certain tasks can be entrusted to volunteers [Cohn, 2008 ; Franzoni and Sauermann, 2014]. This is particularly the case for projects that require fieldwork, such as ecological or geographic observations. In these cases volunteers are generally trained by knowledgeable personnel; they are mentored on

the field and taught how to record observations.

In order to be useful, recorded observations by various volunteers should ultimately be centralised, using some form of information management system. A platform or website designed for entering data or uploading documents, usually accompanied by a database or directory for cataloguing, is a common solution. Projects that manage problem solving or other complex tasks require additional elaboration of tools to accompany volunteers in their tasks. In these cases information technology becomes a major instrument for creating the kinds of environments that collaborators will use to model and create complex data objects, as well as serving as repositories [Franzoni and Sauermann, 2014].

The challenge then is to design and devise user interfaces that are capable of doing a number of things. Firstly, to assemble and input information. Secondly, to manage work flows and volunteers. Thirdly, to render accessible a comprehensive and searchable directory of data to researchers. And finally, to render accessible research results to the greater public via published articles, encyclopedias or some other representative form.

2.4.2 Scientific challenges

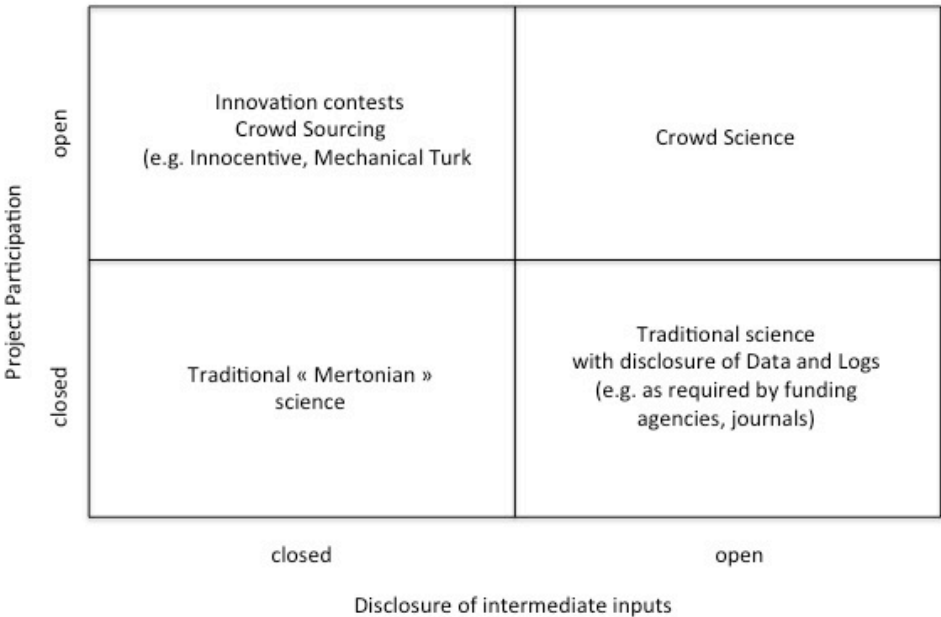
Crowdsourcing has a very specific position with respect to scientific research and scholarship. Authors [Franzoni and Sauermann, 2014] broach this subject diligently in an article that opposes traditional mertonian scientific methods and newer open and collaborative research projects.

The principles of openness that allow projects to solicit contributions from participants outside specific research communities tend to defy the practices that have governed scientific establishments, in which scientific knowledge is accessible to a very select few [Franzoni and Sauermann, 2014]. As already stated, Web 2.0 technologies have already incited changes to this knowledge production paradigm. We should also be inclined to think that changes such as these are ingrained in the very fabric of progress and innovation, as we have also described how, even before the invention of the printing press, a demand for more manuscripts led to greater access to reading materials and resulted in

increased literacy. Here again, situated in the heart of academia, where scientific research takes place, researchers themselves question traditional models of knowledge production. The argument concerns a very important question, that is who should have access to knowledge, and also shouldn't the results of scientific research be made available to the public? Like the writings found in manuscripts have supported literacy over the ages, so has academic research promoted knowledge. Both support knowledge production and we have no reason to believe that these activities should have detrimental effects with increased public participation, particularly since the benefits of knowledge dissemination have been widely repertoried. There is no reason to believe that this should have negative consequences, or otherwise, and worldwide, universities and educational institutions should be called into question.

A comparison of different modes of knowledge production has been proposed by [Franzoni and Sauermann, 2014]. Figure 2.1 shows a rectangle of which the left-most vertical wall represents an axis describing *project participation*. The adjoining horizontal axis represents *disclosure of intermediate inputs*. Both are criteria used to define projects and each can have two states, either open or closed. To allow for combination the rectangle can be subsequently divided into four equal smaller rectangles. Each rectangle represents a category resulting from the interaction of the two defined axes and their two possible states. Existing projects can be said to fit into one of the four resulting categories. According to this model, crowdsourcing does not actually occupy the most open quadrant for the simple reason that when it comes to data disclosure there is no governing principle that compels an organism using crowdsourcing to disclose its data. The most open in terms of project participation and data disclosure is crowd science. It may be interesting to point out that if other intermediate states can be identified between the two extremes of openness and closedness we may observe the emergence of other categories besides the four acutely contrasting categories described.

A notable challenge stemming from this opposition of traditional models of knowledge production and open research is the quality of what is produced. Based on this model it would be simple to assume that crowd science is the most vulnerable to lower quality



C. Franzoni, H. Sauermann, 2013

Figure 2.1 – Typology of different types of knowledge production, with open or closed project participation and open or closed disclosure of intermediate research inputs and results. This figure was taken from Chiara Franzoni and Henry Sauermann’s 2013 article on crowd science.

research, but projects such as Galaxy Zoo have shown that this is not the case. Moreover, as documented by [Franzoni and Sauermann, 2014], Galaxy Zoo and Foldit are examples of crowd science projects that are managed by scientists and credible experts in their respective fields.

2.4.3 Participation

Some of the most obvious difficulties concerning project participation are related to the specificity of each project, and the difficulty in finding contributors. The other major difficulty is related to the technical means and technical skills required to contribute. The third major factor affects is the specialized manuscript reading and analysis skills, which are often difficult to find. All three of these factors affect participation rates for these projects.

Who are the main participants and what kinds of users are they ?

The question lies also in identifying new potential groups of users and thus extending the scope of the projects and producing more transcriptions over a shorter period of time.

In the following sections we will expose a number of existing projects that have made significant contributions to work methods in Digital Humanities' textual scholarship or more broadly in successful application of crowdsourcing for research in the sciences of humanities. Both in what they have achieved and what still remains to be done, they have informed and inspired this work.

2.5 Existing projects

2.5.1 Citizen Science and Citizen Humanities projects

Zooniverse

Zooniverse is an organization that hosts numerous citizen science projects. It began with the creation of an original project, Galaxy Zoo, which aimed to help researchers

process large amounts of data to describe and identify galaxies by soliciting curious individuals. The success of this initial project led to the expansion of the platform and the integration of many other types of projects within fields ranging from astronomy, biology, ecology and the humanities. All of these projects share the principle of recruiting volunteers to constitute data that can be useful to scientists.

The data collected is used to enrich scientific knowledge in various fields of research. Data generated using the help of Zooniverse contributors has been used to produce scientific articles in which volunteers were attributed credit. Such is the case of projects like Galaxy Zoo, Solar Storm Watch, Milky Way Project, Ancient Lives, and Snapshot Serengeti for example¹.

Zooniverse has served as an example for many other projects who share the same goals of including the public in creating data for researchers. By creating a platform to host multiple projects and an API (Ouroboros), Zooniverse has made it easier to create more crowdsourcing projects [Arfon, 2013]. This in itself is a significant contribution as it has rendered project creation and management more accessible to organizations and institutions that do not necessarily have the means to develop their own tools. Zooniverse provides both an infrastructure and a growing community base of participants, from which projects can benefit.

AnnoTate

AnnoTate is one of the projects under the Zooniverse project umbrella. It was developed with the help of Zooniverse-made technology, but adapted to the specific needs of Britain's Tate Archive. The project is particularly relevant as it integrates a transcription editor. Its objective is to collaboratively transcribe artists' journals that the Tate archive has in its collection. By providing an online visualisation and transcription platform, the Tate Archive increases public access to valuable cultural resources housed in their collection.

1. The Zooniverse Website provides links to these publications, organized by project: <https://www.zooniverse.org/about/publications>

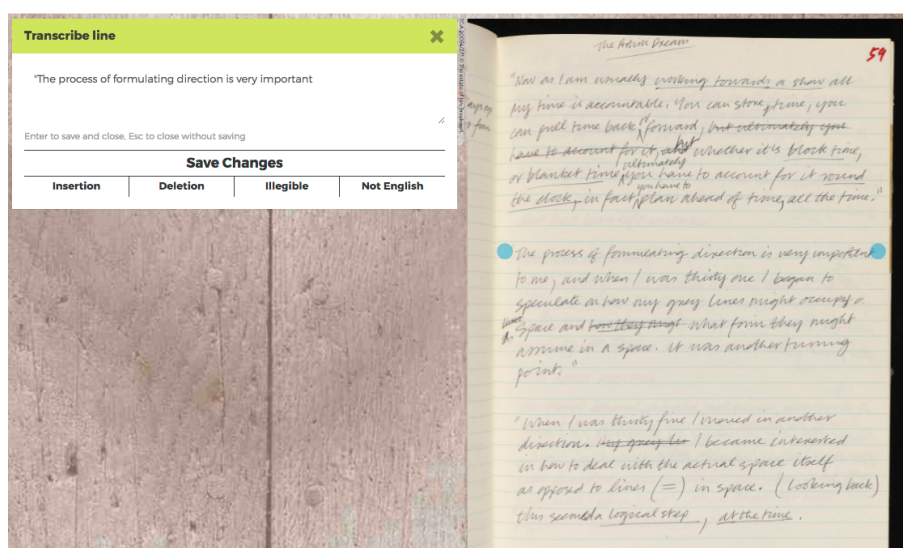


Figure 2.2 – Example of AnnoTate's transcription interface with editor.

The particularity of the transcription approach adopted by AnnoTate is to allow users to transcribe the pages one line at a time. This choice is explained in the transcription guide as a task management strategy: to ensure transcription quality by giving users small simple tasks rather than large complex ones².

Another advantage of using the transcription-by-line approach is that it is easier to accomplish and manage from a technical perspective. The evidence of this is the simplicity of the editor, which contains four buttons (insertion, deletion, illegible, not english) and a fifth to save changes. This means that AnnoTate has chosen as its primary objective the transcription and basic encoding of the main operations and features observed in a manuscript without focusing on the representation of the format or appearance of the manuscript. This renders transcription simple and accessible to many participants and makes it an efficient way of harnessing public interest to convert digitized manuscripts into machine readable text.

Transcribe Bentham

Transcribe Bentham is a project developed at the University College of London for the purpose of transcribing the works of philosopher Jeremy Bentham, a total of 60

2. As described on the website: <https://anno.tate.org.uk/#!/guide/line-by-line>

000 manuscripts [Moirez et al., 2013]. The goals of the project were to accomplish the gargantuan task of transcribing this very large collection of the works of a very prolific writer and influential philosopher. UCL project leaders hoped that transcribing these documents would finally make this archive more widely available for study to the scholars, students, and also the general public who were curious about Bentham’s work [Causer and Terras, 2014]. Beyond preserving and rendering the collection visible and searchable online, the transcriptions contributed by volunteers will be used to continue the work of publishing scholarly editions of the works of Jeremy Bentham [Causer and Terras, 2014]. With this interesting way of involving the public in humanities research this project has been one of the benchmarks for crowdsourcing projects in the humanities. It has also been a pioneer for creating a collaborative online work interface to achieve its goals.

The Bentham Project has created a work environment of which an online transcription editor destined to the public is an integral part. Bentham’s *Transcription Desk* is a custom adaptation of the *MediaWiki* application, which also happens to be one the world’s most widely used software, a factor of accessibility that has certainly benefited the project’s goals [Causer and Terras, 2014].

The project has also furnished considerable effort for the constitution and maintenance of a transcription community, even if this community remains small to this day. Often, sustaining regular contributions from a volunteer community is a complex task, particularly when the object of study requires deciphering handwritten manuscript pages from the previous century, which is certainly not an activity relished by everyone. Many crowdsourcing projects have been met with the same types of difficulties, which are characteristic of this mode of production: a large number of participants make only a very small number of contributions and very infrequently, while a small number of contributors make large contributions [Franzoni and Sauermann, 2014]. This means that to better understand how to manage these types of projects, project leaders must know more about their communities, as well as being able to recognize motivating factors that can impact participation.

One of the main contributions of the Bentham Project to research communities inter-

ested in the crowdsourcing question is the work put forth to recruit users, communicate about the project and animate the website, all done with the goal of motivating participants and keeping people involved in the happenings of the project. The efforts put forth generate and maintain the public's interest are examples to future projects who share the same goals.

Evidence of the kind of success that this project was able to generate is the number of people involved and the number of manuscripts transcribed. In fact, in its first six month testing period Transcribe Bentham had registered 1,222 participants who had transcribed 1,009 manuscripts for the project, 55 percent of the amount were judged to be complete Causer and Terras [2014]. Nevertheless, as the project has reported in statistical studies carried out over the course of several transcription campaigns, aside from a small group of dedicated users the Bentham Project does not retain its users to maintain longterm commitment from them. Some of the factors contributing to these may well be the complexity of deciphering the script of the philosopher, which has discouraged numerous users since the beginning of the project [Moirez et al., 2013].

Another contributing factor may be the project's rudimentary interface, with a transcription editor that puts users face-to-face with xml tags without providing them with helpful syntax colouring. In 2016, during an intervention at the National Archives in Paris, a spokesperson for the project evoked plans of bringing the transcription desk up to date to bridge this existing gap in their user interface.

Crowdcrafting

Crowdcrafting is a crowdsourcing platform that is similar to Zooniverse in many respects. The platform is built on open source software that allows researchers and professionals, but also hobbyists, to create their own citizen science projects [Pellegrini, 2013].

The platform repertoires 180 science projects, 13 projects belonging to economics, 6 belonging to biology, 40 categorised as art projects, as many as 202 social projects, and finally 24 projects in the humanities. Each project has associated statistics, so one can see the total number of tasks, how many have been completed, how many contributors

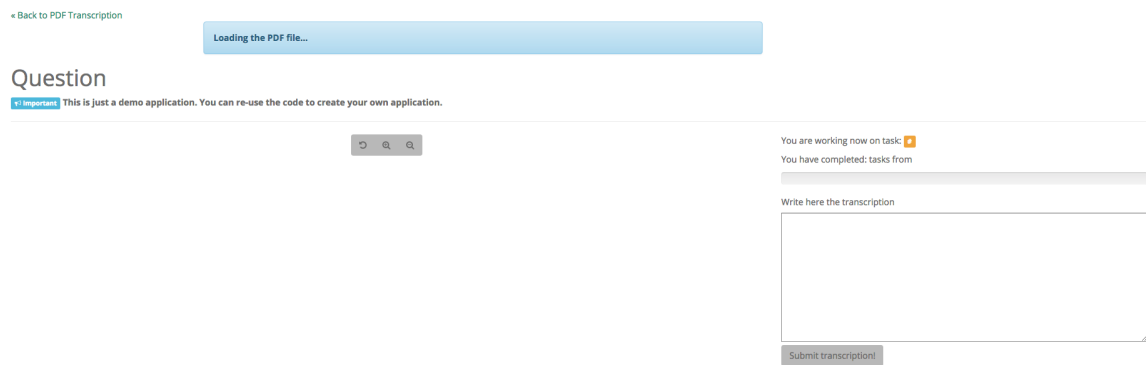


Figure 2.3

(referred to as "crafters") there are, how many tasks are left to do, and also how many of the results have been published.

We observed types of projects listed in the humanities category and found examples focusing on image tagging, translation, digitized document indexing, and even transcription. Unfortunately, the transcription project turned out to be for demonstration purposes only so it was not possible to test out the transcription interface. Figure 2.3 shows the transcription interface; the left portion of the screen should contain an image and the right contains a field for typing in text. Some customization is possible, such as adding more HTML input fields to allow volunteers to fill them with specific data relating to items observed in the documents. However, it was not possible for us to determine the form which the ensuing data would take. Nevertheless, it is possible to provide detailed project descriptions, instructions, reference documents, and even feedback forms using the platform infrastructure. We noted also that tutorials took on a similar format to that of AnnoTate's pop-up dialog windows.

2.5.2 Digital Humanities scholarship projects

Le Centre Flaubert (CEREdI) and *Madame Bovary*

The center for studies on the work of Gustave Flaubert, nineteenth century author famous for his novel *Madame Bovary*, is the result of a collaboration between the municipal

library of Rouen and an interdisciplinary group of researchers at the University of Rouen [Dord-Crouslé, 2010]. There are actually two websites dedicated to studying and editing Flaubert's manuscripts. One site houses the manuscripts and transcriptions from *Madame Bovary* and the second presents the whole documentary collection of Flaubert's unfinished novel *Bouvard et Pecuchet*. We will look closer at the site of *Madame Bovary*.

The manuscripts of Madame Bovary website is in existence since 2001 and makes a complete edition of the work's manuscripts and their accompanying transcriptions accessible online. Its existence was made possible by a doctoral thesis³.

The website is intended to give universal access to the work and working process behind *Madame Bovary*. The site has an interface for viewing each page and its associated transcription. As an integral digital edition, it proposes several types of access to materials. One can either browse materials with a sense of the chronological order in which drafts were established and finalised, or in the order in which they constitute the known edited work. This specialised form of access is highly important to researchers and also highlights the collection's value as an educational resource [Dord-Crouslé, 2010]. Furthermore, as an online resource, it extends this access to the general public, thus allowing all readers of Flaubert to discover the materials and transcriptions.

On the other hand, this collaborative editorial project does not use crowdsourcing. The transcriptions are effectuated by groups of collaborators, most likely remotely and using proprietary software, before being uploaded to the website. The site that is visible to the public does not propose user registration, and users of the resource cannot participate in its creation or maintenance. To general users, the website is primarily intended as a research and reading interface.

3. The doctoral work of Marie Durel holds the title of *Classement et analyse des brouillons de Madame Bovary de Gustave Flaubert*, which analysed the narrative and genetic order of Flaubert's drafts of *Madame Bovary*, and upon completion, proposed a chronological organisation of the collection of 5 000 pages, allowing to apprehend the work in the order that it was written by Flaubert himself [Flaubert, 2017 ; Dord-Crouslé, 2010].

Les Manuscrits de Stendhal

The project of the Manuscripts of Stendhal aims to make available online a vast collection of manuscripts that was left behind by Stendhal and which now resides at the Municipal Library of Grenoble. The collection is composed of many genres: letters, journals, sketches, personal anecdotes, and of course, numerous drafts of novels and plays. Many of these papers have never been formally edited and published.

The digitization project, which was completed in 2009, has successfully transformed the entire manuscript collection into a digital resource [de Stendhal, 2017]. Around the same time, a new database was created to document and organise the collection. As a result of collaboration between Grenoble's Municipal Library and researchers at Stendhal University-Grenoble-3, a website dedicated to Stendhal's Manuscripts was created. This website has made both the author's manuscripts and their transcriptions accessible online for all members of the public, and not only researchers and scholars. According to the project's credo, the work of all collaborating researchers makes it possible to see, read, and understand the works of the author [Meynard and Lebarbé, 2014]. Like the website dedicated to *Madame Bovary*, *Les Manuscrits de Stendhal* is intended as a visual and educational resource, while public participation in the processes of transcription is not currently its intended purpose.

NINES

The NINES project, Networked Infrastructure for Nineteenth-century Electronic Scholarship, has the particularity of not being specifically a crowdsourcing project, but for publishing online peer-reviewed research focusing on nineteenth century British and American studies [Fitzpatrick, 2011]. Nevertheless, NINES has functionalities that benefit from user participation. The project, which began in 2003, has multiple objectives: to peer-review the work of researchers, to oversee and support the creation of digital research materials, and to create innovative software for the digital humanities.

NINES has developed tools for searching through multiple catalogues, repositories, and journals, and it has also integrates a contributive aspect that allows users to tag

and collect items from the catalogues before sharing them with other users. In this way, NINES' environment makes it possible for users to contribute to the creation of new ontologies and establish relationships between independent items found in the collections. This in turn can contribute to resource discoverability for future users of the system.

INKE

At first glance, Implementing New Knowledge Environments (INKE) falls into the category of organizations that advocate for Digital Humanities. As described by its founders, INKE is a large international and interdisciplinary research group created for the study of texts and different kinds of reading environments, particularly within the context of Digital Humanities [Siemens et al., 2012]. INKE is particularly interested in finding and creating successful methodologies for working in the digital humanities to support collaboration between different actors and disciplines, to outline strengths and weaknesses and to identify opportunities for improvement in the ways that research takes place in digital humanities [Siemens et al., 2012]. In reality, its research objectives allow INKE to occupy a very interesting position within the DH landscape. In fact, with collaborations in both research and industry and its occupations with user studies and creating prototypes, INKE not only theorizes on best practices and methods in DH, but also plays a role in creating new tools and interfaces that help shape the existing and future landscape of digital scholarship [Siemens et al., 2012].

2.6 Inspiration and next steps

The projects we have presented are by no means an extensive account of the existing Web 2.0 participative, nor the scholarly editorial, landscape. We focused specifically on those that made use of crowdsourcing: Zooniverse, Transcribe Bentham, Crowdcrafting, AnnoTate. As well as those that represent working in Digital Humanities and editorial fields: Stendhal, Madame Bovary, NINES, INKE and, again, Transcribe Bentham. Some make contributions to scholarly and research practices in digital humanities, while others allow contribution of data to research/editorial and heritage collections. We can see that

the boundaries between many of these projects are permeable, and that several of our examples can actually be categorized under multiple categories.

Observing these examples is useful in a number of ways. Firstly as we observed how projects from different disciplines can be regrouped as a result of crowdsourcing. This is what we observed with Zooniverse and Crowdcrafting in particular.

We were also able to consider how scholarly editorial projects are founded on digitized source materials, which highlights the need to manage all aspects of these projects; from preservation of source objects, to transcription and editorial work processes, and finally to publishing results. Projects that successfully manage these processes in participative contexts can provide valuable information about emerging communities from networked publics. They can also provide insight into the impact of the work of contributors on their work processes and outcomes.

Observing existing projects provide excellent opportunities to consider data management and project management methods. Observations can provide material for reflecting on and proposing improvements where possible. As such, continued improvements to user interfaces may lead to more scholarly editorial projects opening up their work processes to the public. Also, more appropriate functionalities for (a) collaboration and project management tools can be coupled with (b) data (or collections) management tools, as well as (c) transcription and encoding tools. Often, existing infrastructures propose some, but not all necessary components. Whereas missing components require extra customization or development, which inevitably requires technical knowledge and financial resources that projects do not necessarily have.

A simple example of this is Omeka⁴, which allows collection management with the help of a database and online content editing software, but does not have an integrated XML transcription tool, nor user management capabilities to create large-scale contributor communities. Many existing tools that are used in Digital Humanities fields do not wholly address the challenges facing digital scholars; rather, there exist excellent tools, but their implementations provide only partial solutions to problems facing researchers.

4. <https://omeka.org/>

XML editors exist and are accessible (for a price), but they do not manage collections so the user must create his or her own system to do this. Online content editing tools such as Wordpress and Omeka exist and can be relatively easily coupled with databases for data management, but they are not made available with integrated XML editing software. Finally, collaborative and social network platforms such as Twitter, Facebook and Instagram exist and are used by thousands of people, but one cannot use their photo and album sharing modules to the extent required by digital libraries and archives, nor with the same respect for legal rights and obligations associated with the material.

Of course, with significant skill, resources and effort, projects can take disparate pieces of software and use them as bricks to create the type of work infrastructure they require, but most likely, their customizations are highly project specific. Whereas, an improvement that many DH projects would benefit from would come in the form of a tool that integrates these three components, and renders project creation, project management, as well as data management possible.

In addition, projects in DH should continue to be inspired by more generalist crowd-sourcing projects, and those stemming from scientific disciplines. Particularly in order to attract more potential users. In other words, tools that are oriented toward a public of textual scholars are good for experts, but they have the disadvantage of being too specific for generalist users. Proposing tools that inexperienced users will not use, nor appreciate, and at the same time not propose other functionalities (games, social networks, user collections) is an underevaluation of the potential of public interest with regards to these materials. In order to attract more different kinds of users it would be beneficial to also propose tools and activities that may be appreciated by non-expert publics.

Our observation of the projects we discussed in this chapter provided information on which to found our online transcription platform. Since encoding textual data for scholarly research is a process that needs to be planned and managed, we have come up with a way of summarizing our needs in the following list.

- We need to handle encoded textual data.
- We need to create a transcription tool and underlying architecture to support

editorial processes. This also involves creating interfaces for all aspects of the system where users manually intervene in processes or consult data.

- We need to create and implement tools to evaluate transcriptions obtained through crowdsourcing to answer our questions about data quality.

These points require taking a thorough look into the technical means needed to manage editorial processes, handle encoded data, and handle tasks contributed by users in a Web 2.0 environment. We will look at these technical foundations in the following four chapters that make up Part II.

Part II

Technical foundations

Part II Summary

In Part II we assume a technical perspective to explore what it means to work with transcription corpora – particularly in pursuit of ambitious goals of greater public involvement. This will require an exploration of existing technical means for working with manuscripts in a digital scholarly context. We will begin by looking at textual encoding in Chapter 3 and then follow with transcription tools and architectures in Chapter 4. In Chapter 5 we will look at interfaces, a subject of particular interest to Digital Humanities scholars. Then, in Chapter 6 and this part’s final chapter we will describe some methods for comparing documents, on which we will rely for our experimental analysis of crowdsourced documents.

Chapter 3

Encoding textual data

Contents

| | | |
|------------|---|-----------|
| 3.1 | Chapter Summary | 55 |
| 3.2 | Introduction | 56 |
| 3.2.1 | Metadata | 56 |
| 3.2.2 | XML for encoding content and metadata | 57 |
| 3.3 | XML and interoperability | 59 |
| 3.4 | Reconciling local projects and the TEI | 62 |
| 3.5 | Dynamic Documents | 64 |
| 3.6 | Conclusion | 68 |

3.1 Chapter Summary

In this chapter we begin discussing technical aspects of working with digital texts. We begin with the notion of data, describe the processes involved in encoding text, and explain the role of descriptive information about documents or metadata. We speak about the uses of encoding languages, namely XML, and commonly used grammars like XML-TEI for the purposes of interoperability. Finally, we describe what can be done with resulting encoded content.

3.2 Introduction

Scholarly editing projects seeking to establish a corpus of digitized and machine-accessible texts need to consider the technical means at their disposal. Although we can include both automated tools and techniques for extracting content from digitized images and transforming it into machine-readable text, we will focus specifically on those used for encoding data. We will, however, take time also to explain why manual transcription is still the preferred approach in many cases.

The process of converting artifactual objects, such as manuscripts, into digital facsimiles has several steps. The first step is the creation of the facsimile, often requiring the use of powerful digital photography equipment. This creates high fidelity copies of documents, which, once made available online, can subsequently be consulted worldwide by scholars and researchers. The second step involves providing descriptive information known as metadata about the document. Metadata provides descriptive, structural, and administrative information allowing to describe and contextualise documents. It plays an important role in both digital archiving and editorial processes.

3.2.1 Metadata

Descriptive metadata includes descriptive information about the document itself, including provenance, date, author or authors, publishing information, and other details of this order, but does not necessarily refer to the document's written or textual content—the document's data. Structural metadata allows to link resources, in whole or in part, to one another. Administrative metadata is information that is used for managing resources by archives or libraries for example. Administrative metadata regroups technical metadata, preservation metadata, and rights metadata. This information is summed up in a table presented by [Riley, 2017] and which succinctly describes the role of each type of metadata. The table is shown in Figure 3.2.

Beyond descriptive, structural and administrative metadata, [Riley, 2017] also lists "markup languages". This is the type we are concerned with since it concerns the actual

content of documents and includes their "structural or semantic features".

In our work we are interested in the written content contained in manuscripts or, to be more precise, their facsimiles. We are looking to encode the written content of manuscripts, and can be seen as the first step in the process leading up to definitive representations of the text. This first step requires the use of specific descriptive vocabularies for content, which we present in further detail in upcoming sections. However, for now we are not concerned with descriptive, structural, nor administrative metadata (ie. data about the data object). Transcribers focus on transporting the written content of pages and do not necessarily have overall knowledge of a collection, or contextual knowledge of the work, to take care of these other aspects. This information should be provided by archivists preparing the inventories.

For illustration, we provide an example of what a document containing descriptive metadata and content would look like. Figure 3.1 shows a document tree that can be used to represent a text. There are two distinct branches, one for descriptive information and one for textual content itself. Content can be encoded using a specific descriptive vocabulary where encoding elements, or tags, enclose and point out distinguishable features on the page. We will show later on how the vocabulary used to refer to these features can be used to describe both semantic and structural aspects resulting from authors' writing processes.

3.2.2 XML for encoding content and metadata

Scholars and editors specializing in the study of authors' manuscripts overwhelmingly prefer using the XML (eXtensible Markup Language) language for encoding data. XML is a computer language specifically adapted for encoding structured documents for the web. It provides a great degree of flexibility for describing different parts of documents or data sets because it allows creating one's own element names. XML lends itself to the function of encoding or marking up texts for several reasons. Firstly, its tree-like structure reflects structural and semantic components traditionally found in texts. Secondly, one

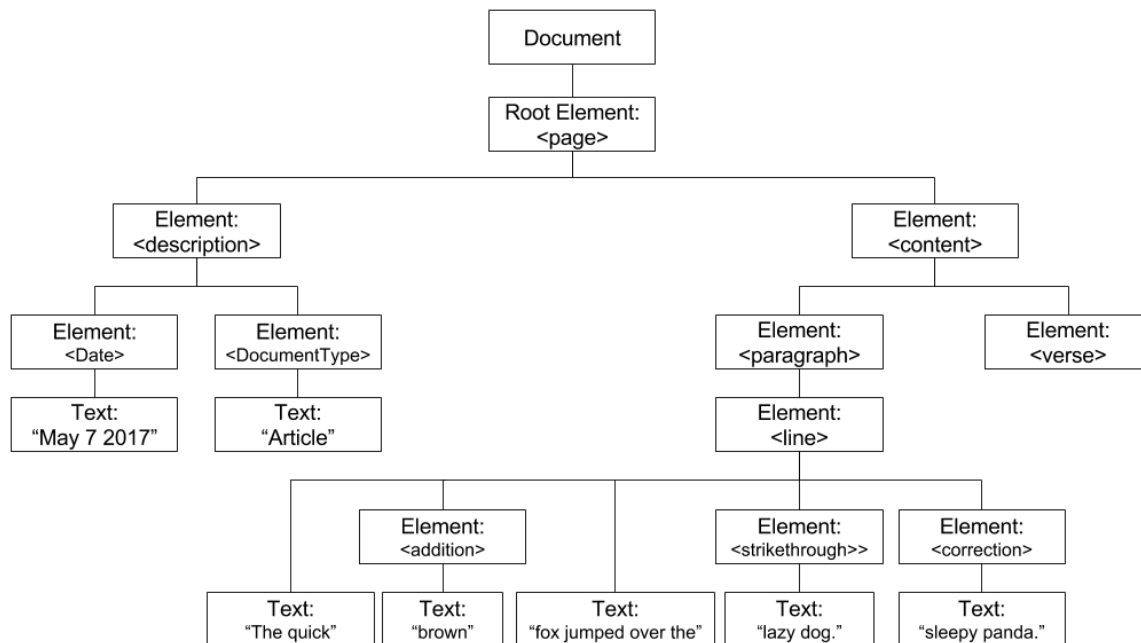


Figure 3.1 – A text can be encoded using XML and the resulting structural components can be represented as a tree.

can define one’s own element names and structure. As long as basic well-formedness rules are respected the XML is well-formed, and if an associated Document Type Declaration (DTD) is respected, the XML is valid according to that DTD. Well-formedness simply means that all elements, unless empty, have both opening and closing tags. A document may be composed of multiple sections containing titles, sub-titles, and paragraphs, which make up its structure. When we talk about document structure we emphasize the relationships between elements and rules we follow to organize them. These rules, which are similar to grammar in natural languages, can be declared for XML documents in a DTD. Whereas names of components can be referred to as vocabulary, which in turn highlights their semantic functions in documents. For instance, in Figure 3.3, the first frame shows raw text, the second shows XML markup and the third the resulting annotated text.

With documents displaying visible evidence of writing and editing processes, corrections and modifications are encoded as semantic features. For example, we may declare

| | |
|---|---|
| Descriptive metadata | For finding or understanding a resource |
| Administrative metadata <ul style="list-style-type: none"> - Technical metadata - Preservation metadata - Rights metadata | <ul style="list-style-type: none"> - For decoding and rendering files - Long-term management of files - Intellectual property rights attached to content |
| Structural metadata | Relationships of parts of resources to one another |
| Markup languages | Integrates metadata and flags for other structural or semantic features within content |

Figure 3.2 – Metadata Types as described by [Riley, 2017].

elements called "correction", "deletion", or "addition". In this way, we can name document features that point to modifications that we observe in documents. Overall, this allows us to describe the various stages of writing and editing that manuscripts are subjected to by their authors. We add for emphasis that vocabularies developed to refer to observed features can be defined by those who use them to encode texts. XML is widely appreciated for this reason.

Some limits to XML actually concern its well-formedness rules, which can be a technical constraint that is not always understood by users. Primarily, this concerns the non-overlap rule wherein no two elements can overlap. That is, if an element is opened inside of another element, its closing tag should always appear before– and not after– that parent element’s own closing tag. Tag overlapping is not permitted in two elements that are next to each other, as shown in Figure 3.4. The *strikethrough* element should enclose only that section of text that is concerned and should not overflow beyond the bounds of the *correction* element, its following sibling.

3.3 XML and interoperability

What we described in the previous section is XML’s vast potential for formalizing different aspects of data objects (and textual data objects in particular) that researchers

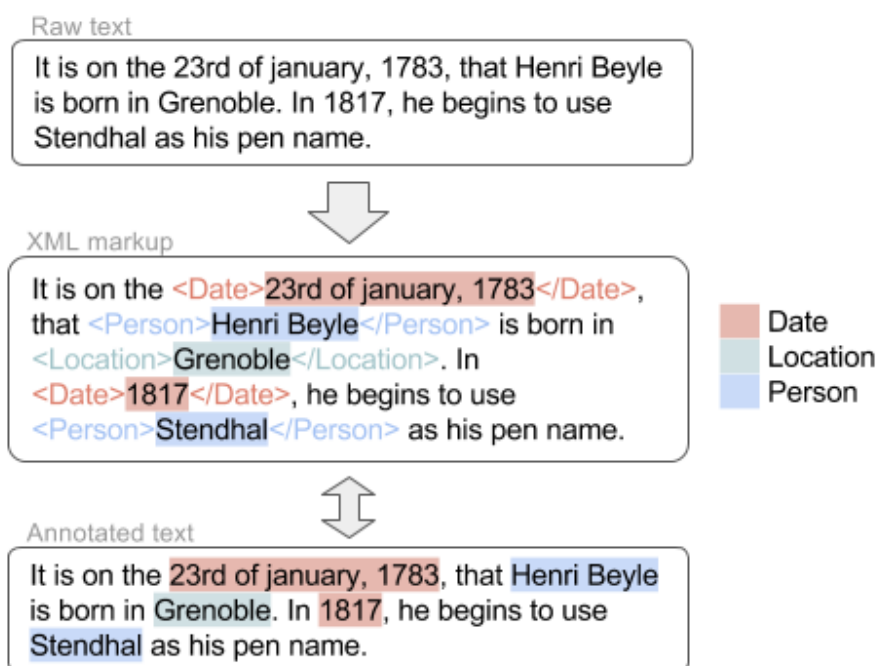


Figure 3.3 – Illustration of text encoded using XML vocabulary for identifying named entities.

and editorial experts seek to represent while adapting to a vocabulary that can be general or very specific. However, XML's potential as a descriptive language in and of itself, does not make it easy for different projects, using their own specific vocabularies, to share and exchange information with other projects, who may also have their own. For interoperability purposes, many institutions handling descriptive data in XML have adopted commonly used schemas in their fields of operation, which have come to be known as standards. What are considered standards in given fields can vary depending on the types of objects being described and the purposes to which the information is put. For instance, an existing encoding standard for working with texts has been developed by the Textual Encoding Initiative (TEI)¹.

TEI-XML is an XML based markup language that is widely adopted by communities of textual scholars because it supports a common vocabulary base for encoding digital text documents. Its applications include prose, verse, transcribed spoken word performances,

1. <http://www.tei-c.org/index.xml>

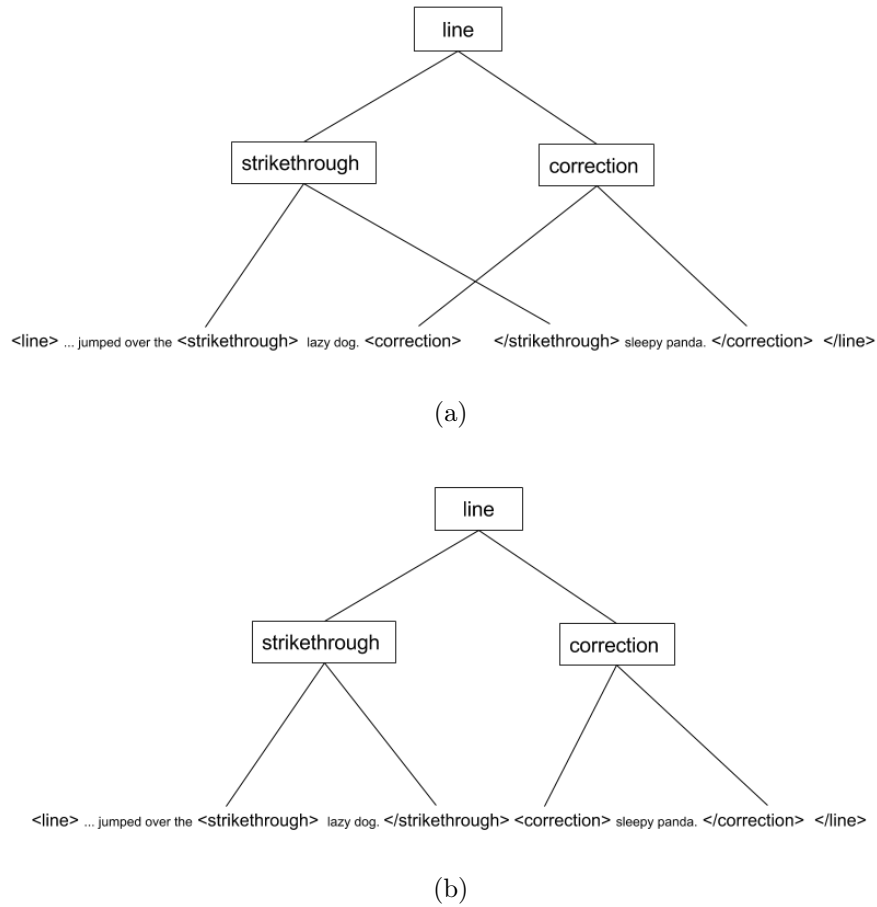


Figure 3.4 – XML element overlap shown in (a) and correct structure in (b).

and manuscripts in particular [Riley, 2017]. The use of TEI-XML across communities of textual scholars makes it easier and simpler to share and exchange documents and datasets between them, thus extending possibilities for disseminating resources and knowledge. As explained in the following excerpt, to benefit from the possibilities of wider dissemination, those who set out to create digital textual resources have an interest in using existing standards.

« [L]a constitution de ressources numériques textuelles exploitables dans les meilleures conditions possibles repose sur la capacité des acteurs à respecter des normes et des standards » [Buard, 2015].

« [T]he constitution of digital textual resources that will be exploited under the best

possible circumstances reposes entirely on the capacity of actors to respect accepted norms and standards [Buard, 2015]. »

This also seems to suggest that using standard markup languages ensures that all encoded information will be transmitted as expected, regardless of which form that information may take.

3.4 Reconciling local projects and the TEI

XML markup standards, such as the TEI, were created to permit exchange between projects and institutions. Choosing to use standards serves the greater community of scholars in ensuring interoperability, however, some scholars consider the benefits derived from its use to have a high learning cost. To give an example, the TEI uses a descriptive vocabulary that may apply to many different kinds of texts, from manuscripts to plays, one can easily encode structural features that are common to all texts. However, there are cases in which scholars from institutions outside those where the TEI is used develop descriptive vocabularies that are different, in part or in totality, from those who adhere to the TEI.

Scholars may have entirely valid and justifiable reasons for using another vocabulary, or even creating another standard, which may indeed be more appropriate for their immediate needs. Although in doing so they should be aware that they are potentially missing out on possibilities of exchanging their data with other groups of scholars, this means both disseminating their knowledge and receiving data from other institutions. If the expense of learning and adopting the TEI is considered higher than obstacles to information exchange then it may indeed be more suitable to privilege local vocabularies for people who will be working most often with the documents. The decision to use, or not use, the TEI belongs to individual projects once they have evaluated their textual encoding needs and defined what they expect to do with their data. Moreover, this should not mean that all is lost, and that their data cannot be shared or used by others. Information technology has enabled to develop and implement effective solutions for this type of problem and

restore the possibility of exchange between communities of scholars.

This may mean establishing relationships of equivalence between series of different vocabularies to allow for converting from one descriptive schema to another, or even to several schemas simultaneously. XSLT (eXtensible Stylesheet Language Transformations) is commonly implemented in editorial processes to convert inputs of structured XML into outputs of other types of XML, TEI-XML, HTML, PDF, and other commonly used formats. Figure 3.5 shows how XSLT may apply to specific elements in an XML document to produce their equivalent in a TEI-XML document. The output elements shown here belong to TEI’s group of core elements [Consortium, 2017]. We have observed in multiple cases that projects prefer using their own vocabularies for these elements and XSLT provides an excellent solution to reconciling these vocabularies with terms used by the TEI. Sometimes, relationships between local terms and TEI terms are precisely one-to-one relationships, wherein an *abréviation* in a local XML is simply an *abbr* in TEI-XML. In others, two or more terms in a local XML, such as *illisible* and *blanc*, translate into a single term, *gap*, in TEI-XML. We add that different XSL stylesheets can be used concurrently to produce different output forms. Editorial processes can integrate XSLT into relatively seamless workflows and output data into desired formats. If projects can manage these transformations effectively, they can use their in-house XML vocabularies and convert them to TEI-XML or other formats when needed.

If there is a cost associated with information loss, which is likely when faced with two very different descriptive schemas, other options may be discussed, including the possibilities of merging schemas or using subsets. TEI does not dictate the use of its entire vocabulary set and many projects effectively use groupes of terms, known as modules, for their specific needs [Riley, 2017]. To gain more information on this subject the reader may find it useful to look at the work of [Buard, 2015], which with its strong inclination in favour of TEI explores formal data conversion and how it can be orchestrated and mastered in contemporary editorial processes.

In our work it was important to allow scholars as much freedom, as technically possible, with regards to their descriptive needs. For those also needing to use TEI for data

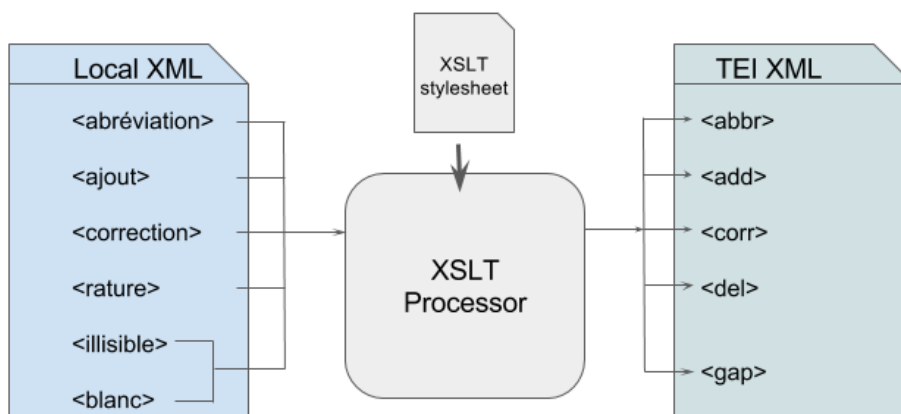


Figure 3.5 – XSLT for converting XML to TEI-XML core elements for transcription.

exchange XSLT conversions should be implemented to meet their requirements. Finally, the editing tool we created allows using TEI vocabulary in whole or in part, by simply defining terms as they are defined by the TEI, and adding other terms if necessary. For projects that only use TEI vocabulary, this may significantly simplify steps to transforming data to TEI-XML at later stages.

3.5 Dynamic Documents

Since we have already spoken about encoding textual content we will now describe how encoded documents can be used as dynamic entities, or how different encoded information can be manipulated as part of editorial processes depending on intended outputs.

As many textual scholars have observed, digital publishing adapts remarkably well to the presentation and study of working manuscripts as it allows to highlight that writing is a process [Leriche and Meynard, 2008 ; Meynard and Lebarbé, 2014]. Transcribing manuscripts allows us to encode features we observe in documents, which result from authors’ writing and editing processes. Part of the process of transcribing documents involves attributing to these observed features semantic elements, or tags, that would allow others to recognise them more easily. Once these features are encoded, those working

with them can also decide how they should be included in, or excluded from, intended output formats or documents. The fact that this can be done to manage intended outputs is what allows us to refer to documents as dynamic rather than static. For comparison, we can refer to print, which inevitably creates a static textual output.

In digital mediums texts can be presented in a number of ways, depending on the goals of their editors. We use an example from Stendhal's manuscripts, *Les Manuscrits de Stendhal*), to illustrate how one encoded XML document can be used effectively to produce two different outputs. Both outputs depicted here are meant to be read online, but the same document can also be used to produce a print version. Figure 3.6 shows an aligned or linearised transcription (*transcription linearisée*) of the original manuscript, intended to produce a text that closely resembles a finished text, what may have been had the text been edited. The second, shown in Figure 3.7 is a pseudo-diplomatic version that reveals and highlights all modifications to the page and exposes the work process, including the author's own ambivalences, deletions, additions, and corrections.

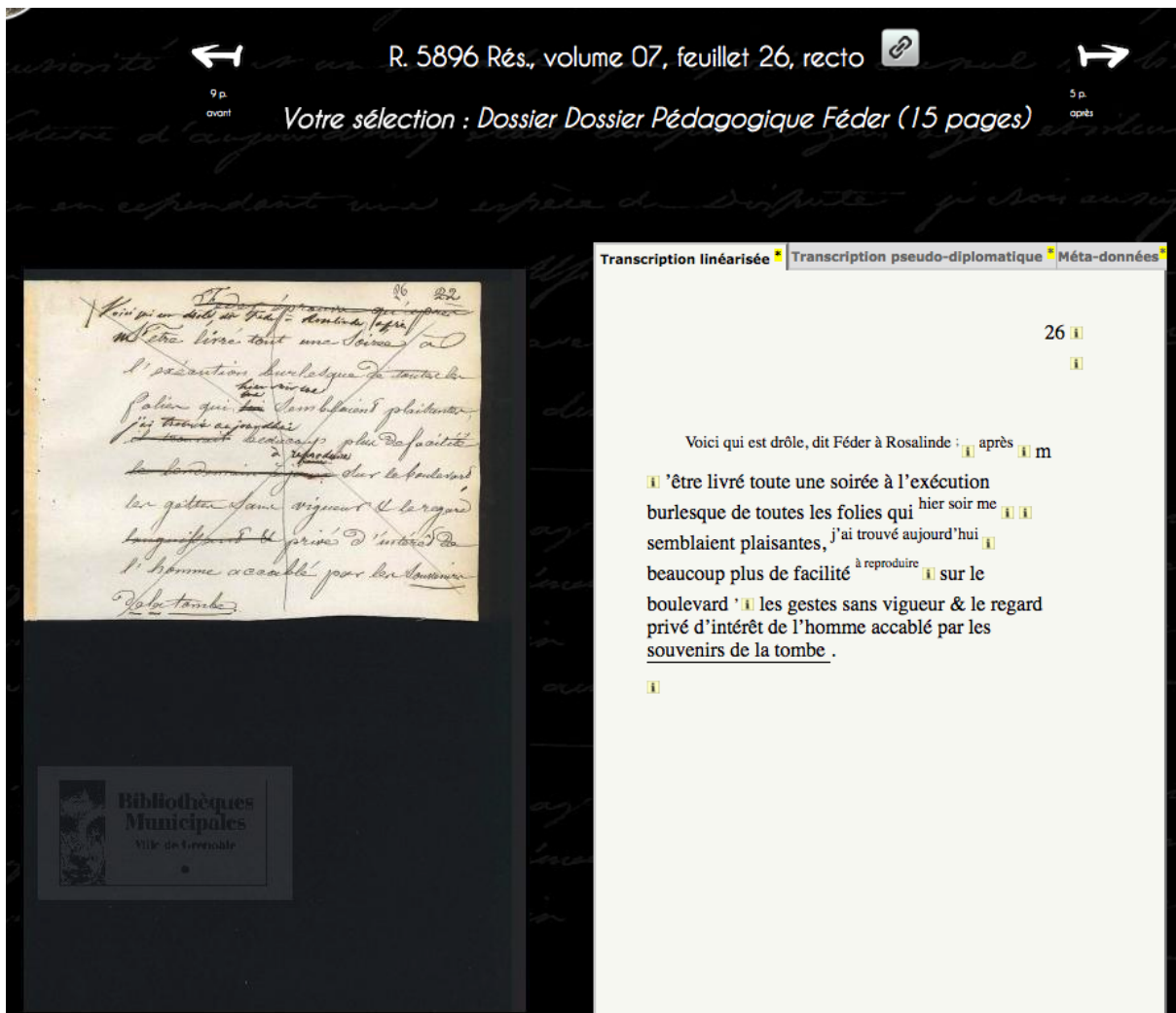


Figure 3.6 – Linearised transcription from the Stendhal online collection of manuscripts. Taken online at www.manuscris-de-stendhal.org, Register 5896, volume 07, leaflet 26, recto. Property of Grenoble's Municipal Library.

R. 5896 Rés., volume 07, feuillet 26, recto

9 p. avant 5 p. après

Votre sélection : Dossier Dossier Pédagogique Fédér (15 pages)

Transcription linéarisée Transcription pseudo-diplomatique Méta-données

26

22

Fédér éprouva qu'après

Voici qui est drôle, dit Fédér à Rosalinde : après

m'être livré tout une soirée à l'exécution burlesque de toutes les

hier soir me

me

folies qui lui semblaient plaisantes,

j'ai trouvé aujourd'hui

il trouvait beaucoup plus de facilité

faire à reproduire

le lendemain à jouer sur le boulevard

les gestes sans vigueur & le regard languissant & privé d'intérêt de l'homme accablé par les souvenirs de la tombe.

Bibliothèques Municipales Ville de Grenoble

Figure 3.7 – Pseudo-diplomatic transcription from the Stendhal online collection of manuscripts. Taken online at www.manuscripts-de-stendhal.org. Register 5896, volume 07, leaflet 26, recto. Property of Grenoble's Municipal Library.

The two examples we have shown of linearised and pseudo-diplomatic transcriptions are produced from a single input XML document. One can generate different outputs from the same original annotated document, with no loss of information to the original.

What can be done with these type of dynamic documents depends on the kinds of information encoded. We have used Stendhal as this corpus effectively illustrates the author's writing process.

3.6 Conclusion

Encoding text using XML markup is a task that researchers themselves are not inclined to do, or are incapable of doing because of technical barriers. Thus, outsourcing transcription to the public is considered a compelling solution. This being said, it requires creating tools and environments for the public. How this can be done, and what it entails, will be our focus in the next chapter.

Chapter 4

Transcription tools and architectures

Contents

| | | |
|------------|---|-----------|
| 4.1 | Chapter Summary | 69 |
| 4.2 | Introduction | 70 |
| 4.3 | Tools for converting digital images to machine readable text | 71 |
| 4.3.1 | OCR processing for handwritten manuscripts | 71 |
| 4.3.2 | WYSIWYG editors for XML transcription | 75 |
| 4.4 | Content Management Systems | 77 |
| 4.5 | Conclusion | 79 |

4.1 Chapter Summary

In the previous chapter we presented the processes of textual encoding and what can be done with digital content. In this chapter it is important to introduce the types of tools and technologies that are used for transforming manuscript facsimiles into digital texts, including the use of manual transcription for encoding texts. We will describe the types of tools that are commonly used to achieve these ends and the types of environments that can be used to manage content in editorial processes.

4.2 Introduction

Recently developing technologies have allowed for specific developments in the treatment of different kinds of information. In our case, as is often the case of projects in the humanities, the information type is most often documents that contain images and texts, and occasionally video and audio records as well. In the case of heritage and historical archives, but also for scientific records, large amounts of data are best managed with the help of database structures and database management systems. In a way, this was already the case of textual collections, whose records were managed by such information systems. The difference being that textual collections used these databases mostly to keep track of metadata regarding objects and their location in physical repositories. With the development of the internet the idea of accessing not just the record of an object to locate it but gaining access to the data itself became a reality. Thousands of file systems have been rendered accessible and linked in this manner. It is thus entirely understandable that projects that seek to render archived collections accessible online use a similar approach, organizing their collections using searchable database systems. It is important to note however that the information that circulates the web is by no means uniform or entirely standardized, information exists in many different forms and its indexation methods are just as variable [Vitali-Rosati and Sinatra, 2014]. In light of the different approaches that exist for structuring data, we will describe some of the particular systems that can be adopted.

In Part 1, Chapter 1 we briefly touched on the link between facsimiles and conservation, wherein copies of rare documents can help preserve originals without preventing the circulation of information. With document digitization (producing high-quality digital facsimiles) the conservation process prevents problems associated with dematerialization and restrictions on physical access. Needless to say that the extent to which these problems can be addressed may depend on a project's available funding or resources. Ultimately, the efforts of conservation are meant to benefit researchers, but also the wider community. The scheme is analogous to the one adopted by crowd science and crowd scholarship (or citizen science and citizen scholarship), all proponents of more open scholarship and

scientific research.

4.3 Tools for converting digital images to machine readable text

Tools and techniques for converting digital images, in our case those of manuscripts, to machine readable text, can fall into two basic categories. In the first category, there are automated and semi-automated processing techniques that involve tools like OCR (Optical Character Recognition) and more recently even HWR (Handwriting Recognition), itself based on OCR techniques. Work in areas of computer science, computational linguistics and NLP (Natural Language Processing) has put forth interesting approaches based in machine learning to improve these automated and semi-automated techniques. In the second category, there are manual techniques incorporating transcription editors and human work.

In this section we will discuss both approaches. We will discuss processing results obtained from OCR (Optical Character Recognition) presented in the literature. Secondly, we will also explain why manual transcription is still a common means of obtaining desired results for both machine readable text and structured XML documents. Moreover, since using manual techniques often implies the use of specialized editors, we will explain why simpler WYSIWYG (What You See Is What You Get) editors can reduce the learning curve associated with manuscript transcription. For manual transcription to be successful with wider audiences, it is important to provide tools that are more straightforward and easier to apprehend for inexperienced users.

4.3.1 OCR processing for handwritten manuscripts

OCR or Optical Character Recognition, as the name suggests, relies on the use of optical technology for the detection of written characters in digitized document. This approach has had extensive testing on digitized documents, both in the case of textual

documents and handwritten manuscripts, with varying degrees of success. There is a long list of factors that have a tendency to affect OCR accuracy, which we will describe in some detail.

Although the technology has achieved incredible improvements in recent years (it has indeed proven to be an adequate solution for automatically processing large quantities of textual documents), many handwritten, and some ancient manuscripts, remain a challenging task for existing OCRs [Diem and Sablatnig, 2010]. Some of these difficulties are attributed to ancient and rare languages that are often difficult to process, while others to complex manuscripts produced by multiple hands. Finally documents which have been damaged or badly preserved still remain a challenge for OCR-based technologies [Cao and Govindaraju, 2007 ; Diem and Sablatnig, 2009]. These cases often demand specific pre-processing and post-processing to achieve results, which makes automating the overall procedure more challenging [Diem and Sablatnig, 2010].

Processing techniques aimed to separate script from a page surface are referred to as binarization [Cao and Govindaraju, 2007 ; Diem and Sablatnig, 2010 ; ?] and are very common among OCR treatments. This is achieved by augmenting levels of contrast in the document, saturating handwritten marks and brightening the page. Using this technique the initial greyscale image resulting from a high fidelity photograph or scan— wherein different shades of ink are distinguishable and the surface of the page itself has stains and shadows— is converted into a black and white image with, ideally, black script and white background [Ntogas and Veintzas, 2008 ; Gatos et al., 2006]. A filtering step can be applied before or after binarization to enhance the image and improve script rendering, including some of the more common means of denoising the image by applying any one or a combination of: median, mean, Weiner, and Gaussian filters [Ntogas and Veintzas, 2008]. Some filters work better when specific paper, ink and degradation conditions are met, while the same conditions can be detrimental to other filters. Finding a balance with the best all around rendering can be a challenge and one solution may be to analyse and sort documents beforehand to determine which filters to apply to the resulting batches.

In general, pre-processing includes the application of specific filters to separate the

text from the page background and accurately delimit the forms of handwritten letters and improve the chances of accurate recognition of characters. Then, after segmentation, the forms are analyzed for specific characteristics and oftentimes they are compared to an existing dictionary of forms for that language or specific corpus for each letter. Having access to references with multiple examples of letter types can improve the probability of accurate recognition. This is why training OCRs is useful, it expands the system's repertoire of comparable forms. Yet with highly variable handwriting, training should be done on very large data sets. Furthermore, erroneous detection is still highly likely to occur. To remedy this, documents can be sorted according to distinguishable similarities in handwriting and OCRs trained on these similar training sets to achieve better detection scores.

Post-processing can include the deletion of superfluous pixels around letters or in the background, which can hamper recognition of letter formats or lead to extra letter or word detections where there are none. Otherwise, post-processing can also mean rereading and validating OCR-ed pages to correct any trailing errors, which, as the reader can imagine, requires human intervention.

Many of the problems with recognition of characters that OCR technologies encounter with handwritten manuscript are precisely the same reasons why typescript was developed. If certain handwriting is difficult for many humans to read, one can imagine the difficulty for a computer program to account for all of the possible variability that can be encountered in human writing. Beyond this, other factors can impact OCR, such as the quality of the manuscript page itself and its level of preservation or deterioration. A manuscript's state of preservation can intensify the challenge of recognizing difficult handwriting, making it almost indecipherable. Some of the difficulties that can cause problems for OCR, thus producing inconsistent results include: spots and stains on pages, shadows caused by poor illumination at the time of digitization, wrinkles, transparent pages, thin pen strokes, broken characters due to light handwriting (pen pressure), poor contrast between text and background, aging paper, and coloured ink. Other issues can be due to image quality itself and related to factors such as image size, resolution, and compression.

All of these, or different combinations of these can produce very different results. As such, many handwritten manuscripts are still no match for OCR technology, even if it has become highly effective for text documents.

We have already shown an example of results obtained on a typical handwritten page in Stendhal's corpus in Chapter 1, Figure 1.2. The particularity with Stendhal's corpus is also due to it being a collection of drafts that span a period of nearly forty years. Over the course of which, not only do multiple hands intervene in the writing process (sometimes within a single page), but the author's writing itself evolves as he ages. With roughly twenty different hands participating in the writing process over such a long period of time, even training an OCR for each of them is unlikely to be useful. Since, for many of them, there simply aren't enough pages to constitute a large enough sample size to be used as training data.

In a way, many of the difficulties for automatic OCR processing are the same as those encountered by a reader or transcriber faced with deciphering a document, only human intelligence is still far superior to artificial intelligence in handling the variability of script and is much more adept at detecting and identifying exterior markers resulting from environmental and time-related deterioration of a page. One approach to improve results is to train programs on new samples of varying types of images associated with possible letters of the alphabet for a language or writing type. Logically, the more training data that can be acquired for a program the better the results as the OCR will be accustomed to a higher degree of variability both for script and backgrounds. As programs are trained and become more intelligent their performance accuracy can be drastically improved.

In terms of putting in place and operating an OCR, some factors should be kept in mind. There are open source OCR programs available online that can run in a number of different languages, though some OCR versions do require proprietary licences and others run using proprietary software applications such as MATLAB[®]¹. Also, technical factors such as the size of files, processing speeds, and overall workflow efficiency when working on large collections should be considered when implementing an OCR. Nevertheless, this

1. <https://www.mathworks.com/products/matlab.html>

technology will continue to seek improvements in the near future to enhance the results obtained on handwritten and ancient manuscripts as well as to automate the process by applying new techniques developed through research in neural networks and machine learning.

4.3.2 WYSIWYG editors for XML transcription

In some cases, having a few spelling or mismatched words in an online manuscript transcription is considered to be a reasonable exchange for a relatively fast and automated solution for manuscript digitization. This may be the case for projects that digitize newspapers or other printed texts. However, since handwritten manuscripts have posed such a significant challenge for OCR technologies many project leaders have decided to stay with manual transcription. In some cases due either to the technical complexity or financial cost of implementing and operating an OCR on a collection, manual transcription remains a more viable option to get the necessary work done. Of course, to obtain a corpus of XML encoded documents using manual transcription requires that human transcribers encode and structure the documents themselves. XML editors exist for this work as well.

In fact, projects in Digital Humanities have made excellent use of available tools. XML editors such as Morphon (not maintained anymore) and Oxygen² (requiring a licence), have made the task of creating and editing structured XML documents possible for many projects, including *Les Manuscrits de Stendhal*, which rely on XML to constitute their corpora [Meynard et al. 2009]. These types of editors supply detailed interfaces specially designed for working with structured documents. In our discussion, we do not consider other common code editors such as Notepad++³ or Emacs⁴ because these are intended for code editing and do not provide users with an "author" mode. To work with these one must manipulate raw code rather than objects, as is permitted by text editors endowed with user interfaces. For contrast, we can consider Microsoft Word, whose interface is

2.

3. <https://notepad-plus-plus.org/>

4. <https://www.gnu.org/software/emacs/>

purely WYSIWYG (meaning *What You See Is What You Get*) and does not allow access to raw encoding. The concept of WYSIWYG editors relies on users manipulating graphical and textual objects with the help of an interface that allows users to see what the end result will look like. The WYSIWYG editor is essentially an interface for encoding content.

Oxygen, for example, allows a user to create an XML document from scratch and ensure that it is valid XML based on a DTD (Document Type Definition). Like both code and WYSIWYG editors, it also provides convenient functionalities, including search and replace, and is able to propose element names (based on the existing DTD) in response to users' partial input. This latter functionality can help reduce document well-formedness errors resulting from typos. Finally, Oxygen's author mode frees scholars from technicalities which they may be unaccustomed to handling, while its raw XML mode is accessible to users when it is necessary to execute fine modifications that the WYSIWYG interface renders difficult.

However, the licensing costs associated with editors like Oxygen make them inaccessible to some users and they also require installation on personal machines. For projects operating on strict budgets, furnishing this type of equipment for an ever-growing community is a complicated undertaking⁵. Furnishing licences to dozens and hundreds of volunteers is often impossible, so free tools are highly desirable. So is a shared infrastructure that can be accessible to everyone involved, in order to make the work process more collaborative and to make collaboration easier and more effective.

The internet offers a solution. Creating an online platform to which all collaborators have access, and which is connected to a database where all sources and documents are stored, is a way of centralizing resources and operations, thus reducing the number of documents being sent back and forth between isolated actors. The work interface itself needs to be online and should support functionalities for less technically oriented users. Creating a WYSIWYG transcription tool means creating a user interface for the task of XML encoding and thus reducing the technicality of the task. Using tools that are

5. In France, organisations like Huma-Num, <http://www.huma-num.fr/>, exist to support projects by providing them with tools and financial resources.

more accessible to users increases chances of having more contributors, and this, in turn, increases the rate at which projects progress toward their digitization and transcription goals.

A transcription tool on its own may be useful to one person or a number of people working independently, but we cannot be sure to fit it into an organized workflow without a supporting architecture for managing content and users. We will consider CMS (Content Management Systems) in the next section as this will lend us the opportunity to position transcription work within editorial processes that involve many users.

4.4 Content Management Systems

Content Management Systems are systems that are used to create and manage digital content. Content can include digital objects such as images, video, music, texts, and as in our case, transcriptions. Some commonly known Content Management Systems are Wordpress, Omeka, and Drupal, but there are many others. Although the word *content* does not necessarily make one think directly of information and knowledge, the CMS is in many ways a modern infrastructure for managing and disseminating knowledge. For many, it is thanks to WordPress that web publishing has become so popular and accessible.

A CMS is not only a system for stocking and publishing content it is also an environment that can support collaboration between multiple users. Sometimes the term UMS (User Management System) is used to refer specifically to the management of users, but in many cases Content Management Systems imply user management as well. Unless otherwise stated and where the distinction is important, we will use CMS to refer to both content and user management. Figure 4.1 presents the general functioning of a CMS. Administrators are in charge of creating layouts for the website and deciding how content will be structured, whereas users contribute by creating or editing the content itself. The CMS application takes on the charge of injecting this content inside the designed layout to create resulting web pages. With respects to user management, certain parts of websites are accessible only if users have sufficient rights within the system.

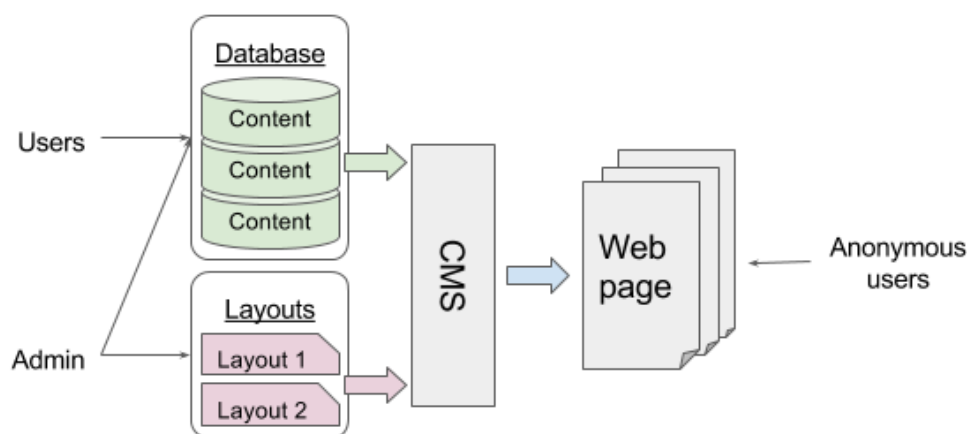


Figure 4.1 – Process overview of a CMS.

If our purpose is to create a participative platform for manuscript transcription, we may consider using a CMS as a possible solution. Many existing CMS, including those already mentioned, propose a customizable process for publishing content to the web, in a similar manner as we have presented in Chapter 1, Figure 1.3, that is with the possibility of creating different publishing formats for content. Many existing CMS also propose solutions for managing users. Hierarchical roles are common features built into CMS architectures and can be defined to suit the roles and processes envisioned in the system.

The problem lies in how much control the CMS has over the form, structure, and organization of content once it has been fit into any given CMS. It is a form of control that many scholars are unwilling to give up to what are essentially generic applications. For textual content that precedes the existence of these types of systems by at least a few centuries often generic solutions are insufficient. There can also be concerns about maintenance and portability of content once it has been adapted to suit a particular CMS.

Nevertheless, understanding the architecture and components of a typical CMS is important when one's goal is to implement editorial processes in online environments and implicate many users. We have used the concept of CMS as a way to emphasize the necessity of an architecture to support processes. Furthermore, various components that can be built into an architecture will become essential pieces of the overall puzzle when it comes to providing the public with transcription tools and managing their work. At the

time we were confronted with our research problem, existing CMS had not yet provided us with a simple solution to crowdsourcing transcription and managing them in an editorial workflow.

4.5 Conclusion

Often, Digital Humanities scholars need to adopt the use of separate and heterogeneous tools because existing digital and web based solutions are not adapted to their needs. Many scholars would agree that "[o]ne cannot speak of a single and unique digital solution, but of diverse digital solutions, each adapted to the needs of its designers"⁶ [Leriche and Meynard, 2008]. More broadly, DH scholars insist on the important roles that research and experimentation play in creating new tools and systems. And that these should be rooted in scholars' reading and working practices [Siemens and Meloni, 2010]. We have mentioned examples of this in Chapter 2 with the work of the INKE research group.

We need to continue to find solutions that meet the specific and diverse needs expressed by scholars. Only then can editorial processes become more effective, when they reflect the needs of those who design them. That is, rather than adapting the processes and products of scholarly research to one CMS or another. One thing that is certain is that there is a need for robust architectures that can handle large volumes of data and many users. That are flexible when it comes to adding tools and components and efficient when changes to parts of the system need to be made. Finally, one should be able to access content easily to be able to harvest it for interoperability, digital or print publishing, archiving, or other research purposes.

Much of DH scholars' focus on digital tools actually concerns aspects of interface design, as supported by the following statement, "[i]nterfaces both engage and shape the practices of the research communities they serve" [Crompton and Siemens, 2013].

6. Author's translation of the original text by [Leriche and Meynard, 2008]: On ne peut parler d'une «solution électronique» unique, mais de solutions électroniques diverses et adaptées aux besoins des concepteurs.

Consequently, this will be our focus in the next chapter. We will discuss the importance of designing user interfaces for Digital Humanities scholars in general and for transcription and editorial practices in particular.

Chapter 5

Interfaces

Contents

| | | |
|------------|--|-----------|
| 5.1 | Chapter Summary | 81 |
| 5.2 | Introduction | 83 |
| 5.3 | Reading and transcription interfaces | 84 |
| 5.4 | Understanding user activities | 87 |
| 5.5 | General design principles for user interfaces | 88 |
| 5.5.1 | Navigation and work flow | 90 |
| 5.5.2 | Operations and actions | 91 |
| 5.5.3 | Text and data entry | 92 |
| 5.5.4 | User guidance | 93 |
| 5.6 | Conclusion | 95 |

5.1 Chapter Summary

In this chapter we focus our interest on user interfaces and we also look at types of user activities and how these should be taken into account for designing interfaces and environments. We look at types of interfaces, and the relationship between reading and transcription interfaces. After describing theoretical work on web-based user activities

and the role of user requirements in design, we define four main areas of focus for design: navigation and work flow, user operations and actions, text and data entry, and user guidance.

5.2 Introduction

Digital Humanities scholars are particularly interested in interfaces. This is rightly so because web interfaces are essentially the new pages of books and all knowledge that contemporary readers and writers encounter and produce must inevitably pass through these new pages. Even more so, the new digital page has an untapped potential for organizing, presenting, and linking information that is of great interest to digital scholars. It is not surprising then that "[d]igital research environments, from the e-book to the digital archive, invite scholars to design interfaces that meet, and indeed challenge, scholarly reading and research practices" [Crompton and Siemens, 2013].

Having said this, software and web interfaces are still new to many users and a significant amount of research directed at users' aims to understand precisely how web interfaces can be better designed to suit users' needs. Although it can be a real challenge to meet the needs of all groups, the more information that designers have at their disposal the better equipped they are to make well-informed decisions about structural and design choices that will ultimately affect end users.

With respect to the diversity of users' needs, these can be attributed to several factors, which may originate in the tasks performed, the differences between operating systems, as well as the habits users acquire when using particular devices. Prior knowledge of specific operating systems may also have an effect on user needs and the ways in which these needs are articulated when faced with new interfaces. In other words, software is habit forming and user habits are often a result of the types of devices to which users have access.

Recent changes in user practices have been accompanied by the diversification of personal devices. Users have an ever growing range of choices with respect to the types of devices to use to access internet applications and accomplish various tasks; from desktop personal computers, to laptops, or more recent mobile devices such as smartphones, ipads and notepads. Research into user experience has affirmed that user practices are affected by the type of device they use to access a given platform or internet service because using

a desktop or laptop computer is not the same as using a mobile device to access a website [Kaikkonen and Roto, 2003 ; Cui and Roto, 2008].

5.3 Reading and transcription interfaces

Digital technologies have brought about important changes in the ways in which scholars, and readers in general, interact with texts. Today, many scholars are regularly consulting information and reading texts on digital screens. Furthermore, texts of all forms are increasingly being stored as digital files on digital devices.

Industrial actors have largely embraced the digital medium and have proposed both generic and specific solutions for a new way of reading. From PDF document formats to more recent EPub formats to the hardware (kindles and other electronic reading devices) that allow readers to store thousands of titles in one compact device, the new digital reading public has clearly been targeted by commercial enterprises.

In the wake of new kinds of interfaces that accompany new document formats and new hardware, certain research actors have also grasped the opportunity to influence the ways in which future reading interfaces will be developed. We are led immediately to think of INKE (Implementing New Knowledge Environments), which we have already referred to in Chapter 2, Section 2.0 on page 46, and their work to propose new reading prototypes grounded in an understanding of the history of textual scholarship, research in human-computer interaction and user experience research [Siemens and Meloni, 2010 ; Siemens, 2012 ; Siemens et al., 2012]. Furthermore, INKE's research initiative aims to propose solutions for specific types of readers, that is expert readers and textual scholars, who have specific practices and thus very specific expectations for their digital reading tools.

Reading interfaces can indeed be very specialized types of interfaces, allowing users to collect citations, add annotations, trace references and interrogate the relationships between texts [Siemens and Meloni, 2010 ; Siemens et al., 2012]. This all depends on users' reading practices. Reading interfaces can also be understood more generally as

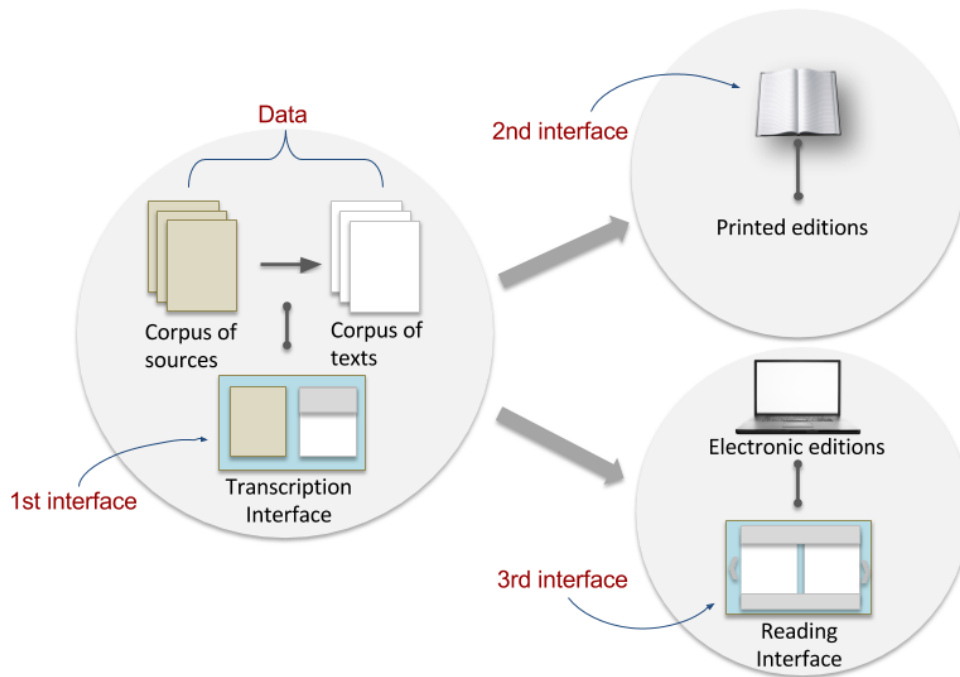


Figure 5.1 – Relationship between transcription and reading interfaces: from corpus to electronic edition.

specialized interfaces for working with texts. This allows us to draw connections between these types of interfaces and transcription interfaces, which are also specialized for working with texts.

Some distinctions that can be drawn are that in the former case the central activity focuses on reading a text (from an already constituted corpus) and in the latter transcription and constitution of a text from a source. The assumption is that transcription will allow to constitute a textual corpus, from which an electronic edition can be derived, and which can later be consulted via a reading interface. We can imagine at least two successive interfaces, one for transcription and another for reading, each proposing specific options for working with texts. Figure 5.1 illustrates this relationship; manuscript objects are transformed into digital texts using a transcription interface, then read as digital texts using a digital reading interface, or otherwise read in print form, which remains a widely preferred interface between readers and texts.

There are also important similarities between reading and transcription interfaces.



Figure 5.2 – An example of an INKE reading interface, which uses a dynamic table of contents. This image was taken from the web at <http://www.artsrn.ualberta.ca/inke/wp-content/uploads/GalleryDynamicTOC.jpg> to illustrate common and existing page layouts for reading interfaces.

Firstly, both should provide access to specific pages of collections and therefore in both cases selection, navigation and search features are important. Also, the display remains faithful (where possible) to common page dimensions, such as those used for printed editions or manuscript pages. In some cases the illusion of a page may be recreated to frame a text. For example, Figure 5.2. The layout itself may allow to view one or two pages at once, although variations are possible and may be dependent on the screen size of the electronic device used for viewing. Left or right arrows may be used to imitate the way readers move backward and forward between pages, or an electronic scrolling functionality may be preferred (compare Figure 5.2 and Figure 5.3).

Although browsing and reading information are tasks that are increasingly performed on mobile devices, depending on the amount of concentration required, or the type of material, these activities are often performed on devices having larger screens (for example

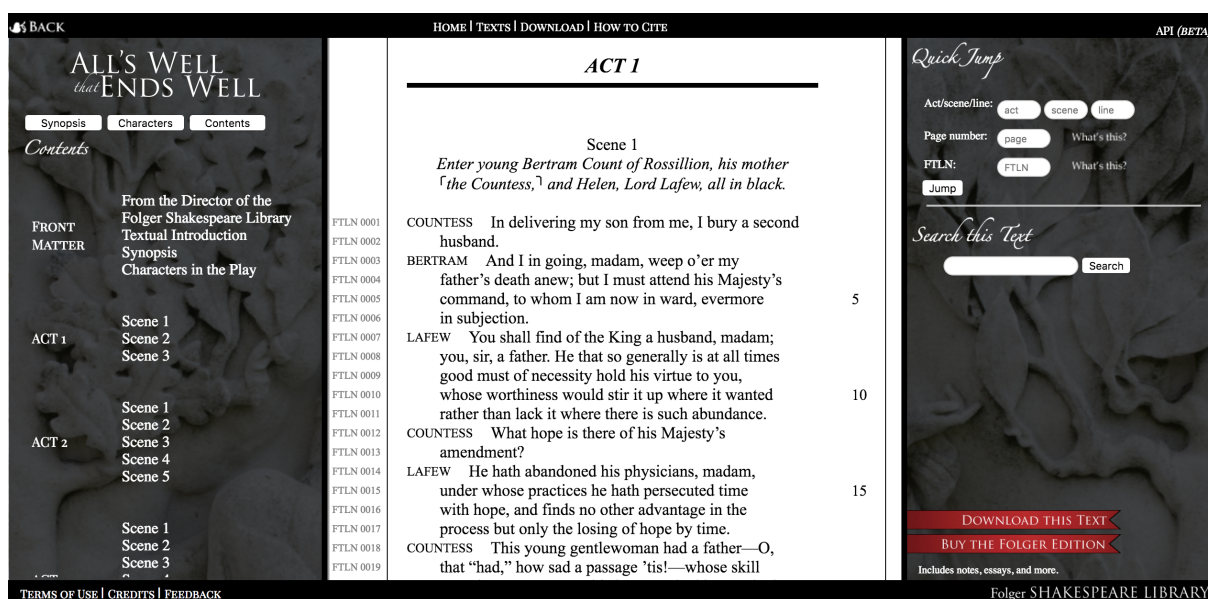


Figure 5.3 – An example of Folger Shakespeare Library’s Digital Texts reading interface. This image was taken from the web at <http://www.folgerdigitaltexts.org>

notepads and ipads over mobile phones). If adding note-taking or other typing activities including e-mailing, stationary devices are still considered preferable to mobile phones [Cui and Roto, 2008]. We thus consider that the same applies to transcription. As transcription requires a fair amount of close reading and encoding and these are actions that are somewhat difficult to perform using tactile mobile keyboards.

5.4 Understanding user activities

In the interest of understanding how users incorporate mobile and desktop devices for the web, a number of studies have been realised with the goal of creating a taxonomy of user activities. The studies that focused on better understanding how mobile devices affect user activities are grounded in prior research that focused on describing how people use desktop devices to accomplish what are referred to as stationary web tasks [Cui and Roto, 2008]. According to the literature, there are at least five research studies between 1995 and 2006 that aimed to identify the main types of user web tasks. Aside from variations in the methods used to obtain results and variations in the terms attributed to the activities,

the categories listed show a relative amount of consensus among researchers [Cui and Roto, 2008]. The first two studies named categories that focus on finding information and differentiate between casual browsing and purposeful searching. Later studies identify a category for communicating with other users and also take into account forms of exchange, which is described as transacting. To give an overview of the categories and highlight their similarities and differences we have represented them in the following table (see Table 5.1). With these five taxonomies we can see the overlap in terms used to define user activities. Cells left blank mean some terms were not identified by those particular authors. For instance, [Morrison et al., 2001] groups both browsing and finding into one and does not make the distinction between unintentional and intentional acts, which other authors make. We can also see where later works build on earlier works to extend the range of activities that are considered as web tasks. This also reflects the turn taken by web technologies between the 90s and early 2000s, and the more active role that web users have acquired. We can see this in particular with [Sellen et al., 2002] and [Kellar et al., 2006] who move beyond browsing, fact finding, and information gathering to include communicating, transacting and housekeeping/maintenance. The latter is specifically related to user accounts and implies the active role of users in managing their personal virtual spaces.

Understanding activity or task types constitutes an important basis for user interface design from desktop to mobile devices. We will see in the following section (5.5) how some general design guidelines for both software and web can apply to transcribing manuscripts online.

5.5 General design principles for user interfaces

« To the extent that information systems support human users performing defined tasks, careful design of the user-system interface will be needed to ensure effective system operation [Smith and Mosier, 1986]. »

Guidelines for designing user interfaces are necessary to support the work of even the

| N°. | Catledge (1995) | Choo et al. (1998) | Morrison et al. (2001) | Sellen et al. (2002) | Kellar et al. (2006) |
|-----|---|--|---------------------------|--------------------------|--------------------------|
| #1 | browsing (serendipitous/ general-purpose) | conditioned/ unconditioned viewing | | browsing | browsing |
| #2 | searching | formal/ informal searching | finding | finding | fact finding |
| #3 | | | comparing/ choosing | information gathering | information gathering |
| #4 | | | understanding | | |
| #5 | | | | communicating | communicating |
| #6 | | | | transacting | transacting |
| #7 | | | | housekeeping | maintenance |

Table 5.1 – Five taxonomies of web tasks as described by Catledge (2005), Choo et al. (1998), Morrison et al. (2001), Sellen et al. (2002), and Kellar et al. (2006). Analysis based on article by Cui and Roto [Cui and Roto, 2008].

most knowledgeable designers [Smith and Mosier, 1986]. And of course, some information can be oriented specifically at web applications, or mobile web, while others focus mainly on software, or specifically on learning environments.

In many cases across software, web, and mobile web common categories are recognizable as focusing on user tasks or activities, even if terms can vary as we saw in Section 5.4, and in many cases systems should (i) make evident users' actions and their effects on screen and (ii) provide confirmation and feedback where appropriate on actions taken.

Also continued research focusing on the way users interact with computer systems helps inform new guidelines and sometimes question pre-existing ones [Law et al., 2009]. The overall aim is to ensure that web tools remain accessible to users and that this acces-

sibility can be improved where possible and where it is most needed. Existing guidelines concern topics such as navigation of workflow, user actions, user guidance, and specific considerations for data and text entry. We will look at each of these in more detail in the following subsections.

5.5.1 Navigation and work flow

Web sites, particularly large ones, are often a challenge to organize in ways that are evident for users, but organization is vital as otherwise it reflects negatively on user orientation and motivation [Webster and Ahuja, 2006]. Thus, navigation has direct influence on the usability of systems.

There are generally accepted rules about site architecture or structure and how one should organize information presented in web environments in order to make navigation more intuitive for users. Not all guidelines are adopted by web designers as there are legitimate differences that can be identified between web, software and mobile web environments, but these guidelines serve to improve and maintain usability. Since the 90's, the number of commercial and professional websites that employ usability guidelines has risen in order to attract more users [Webster and Ahuja, 2006].

Even though some guidelines may be challenged in mobile environments [Kaikkonen and Roto, 2003], for the most part, considering them is useful for defining areas of focus. For instance, [Kaikkonen and Roto, 2003] cite [Nielsen, 1999] to define minimal navigation as a guideline for general user interface design. General knowledge about web design seems to support this in suggesting that users should be able to access content in three clicks or less, but investigations into this rule have also shown that as long as users find what they are looking for they will not be dissatisfied, even if it takes more than three clicks [Porter, 2003].

Regardless of the type of interface, information should be organized and focused so that users are able to locate where they are in the structure of a site [Nielsen, 1999]. Many guidelines enforce flatter architectures over highly hierarchical structures because simpler

structures are easier to navigate and faster for finding information [Webster and Ahuja, 2006]. It seems to hold true that large sites will use complex architectures for structuring large volumes of information, while small sites used for presentation will have flatter architectures. To illustrate one may imagine a large website as a typical university website. Different information is organized according to topics; from programs and their class offerings to syllabi, to admission information, student resources, and curricular activities. The university may also link to its affiliated research institutions, journals, magazines, and libraries. All of these different topics can be structured in a system of menus and submenus and their organisation will determine the ease with which students, personnel, and visitors will find their way around the site. For a small website, such as a personal portfolio, this navigation system will be much simpler and may be composed of four or five distinct items without necessarily having subitems, which implies a flatter or more linear architecture. Authors of [Webster and Ahuja, 2006] associate the latter with simple navigation systems and the former with global navigation systems.

Some advantages to navigation systems that privilege constant visual representation of site structure, such as those made possible by global navigation systems, are also attributed to web design guidelines. In particular, visual representations that rely on recognition of information rather than relying on users' memory are considered effective [Smith and Mosier, 1986 ; Webster and Ahuja, 2006]. Similarly, it is important to keep elements in navigation systems and workflows consistent for the same reasons and to avoid disorienting users.

5.5.2 Operations and actions

Concerns with user operations and actions are derived from the types of tasks that users will perform in an online environment, which may indeed be very specific. Just as users need well structured navigation to know where to go on a website, the actions available to them once they have arrived on any specific page should be evident.

A rule that applies specifically to web environments, but which the authors consider

important, is the simple identification of clickable items [Krug, 2000]. There are other guidelines that the authors question in mobile environments, but which are generally accepted in web environments.

In some high precision environments it may be important to confirm actions before completing them and saving the changes. Confirmation is reassuring to users on particular actions, such as deleting items and they find it preferable to confirm an action before effectuating it rather than having the option to undo an action. We will now focus specifically on text and data entry operations that users are most likely to encounter in online editing and transcription work flows.

5.5.3 Text and data entry

We consider this type of operation separately from other operations that users can perform in online work environments because it relates directly to transcription and editorial activities. For more general data entry, software guidelines reported in [Smith and Mosier, 1986] indicate that users should be able to enter information once and the system should in turn be able to access previously entered data, thus preventing the inconvenience of having to re-enter information multiple times and the danger of entering conflicting information. Likewise, when users are working with text or entering data, all actions they perform with a mouse or keyboard should be reflected in the interface. Also, the interface should provide the possibility to cancel actions or return to a previous work state.

Text encoding and annotation, or transcription, can be a highly detail-oriented task requiring a significant degree of attention from users. Oftentimes, text encoding environments can be challenging for users who do not fully grasp the technical aspects of encoding. For example, Transcribe Bentham's work environment, as shown again in Figure 5.4 may be visually disconcerting for users who are inexperienced in text encoding. For this reason, an interface having more distinctive markers for elements, or even syntax highlighting, may be fitting. Including fewer encoding options may also help avoid confusion; the Bentham editor has approximately fourteen buttons and AnnoTate's editor has

Editing JB/035/214/001

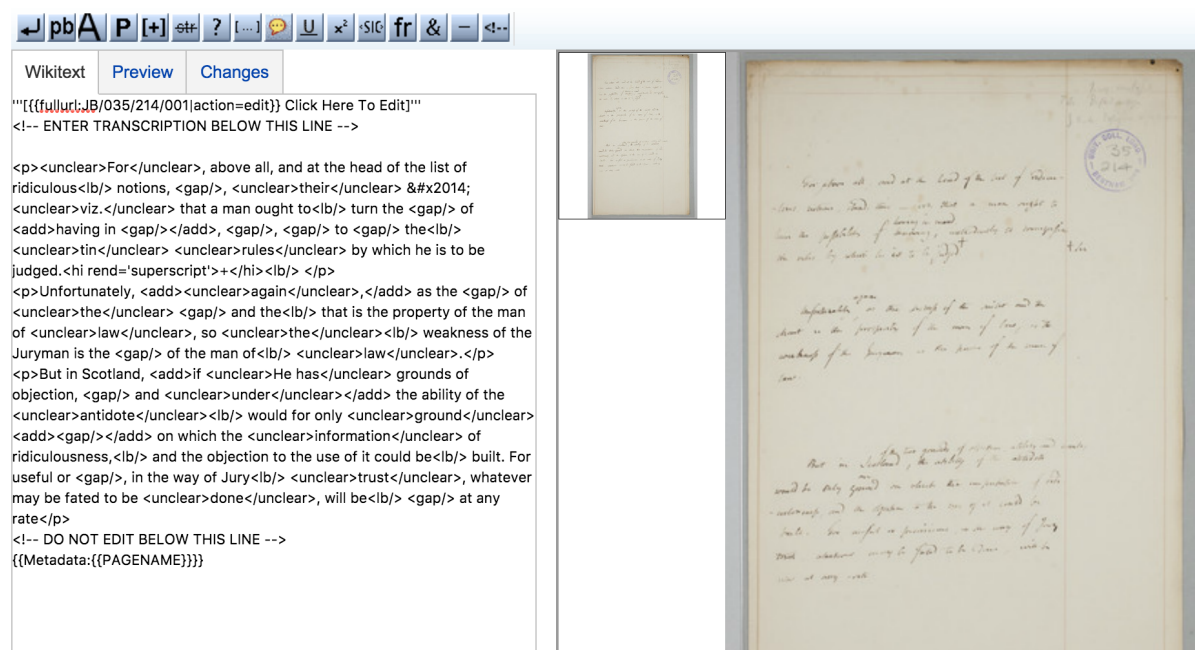


Figure 5.4 – Text encoding: Bentham transcription desk example.

a total of four buttons for encoding.

It is worthwhile to consider how screen space is used and whether text entry fields are sufficient for kinds of text entry being performed. In general, long texts require displays with a minimum of 20 lines [Smith and Mosier, 1986]. Though we have observed how seemingly large text entries can be segmented into lines, thus reducing the area needed for working, as we have seen with AnnoTate’s transcription environment (Figure 5.5).

5.5.4 User guidance

The web is an environment that provides a high degree of what in educational psychology is referred to as "learner control" and positive experiences while using a system will determine whether users will strive to master new skills, such as navigating through a system or accomplishing tasks [Eveland Jr and Dunwoody, 2016]. Guiding novice users throughout this experience is important. This is why learners in new environments need structure and advice about what they are doing in order to accompany their decisions

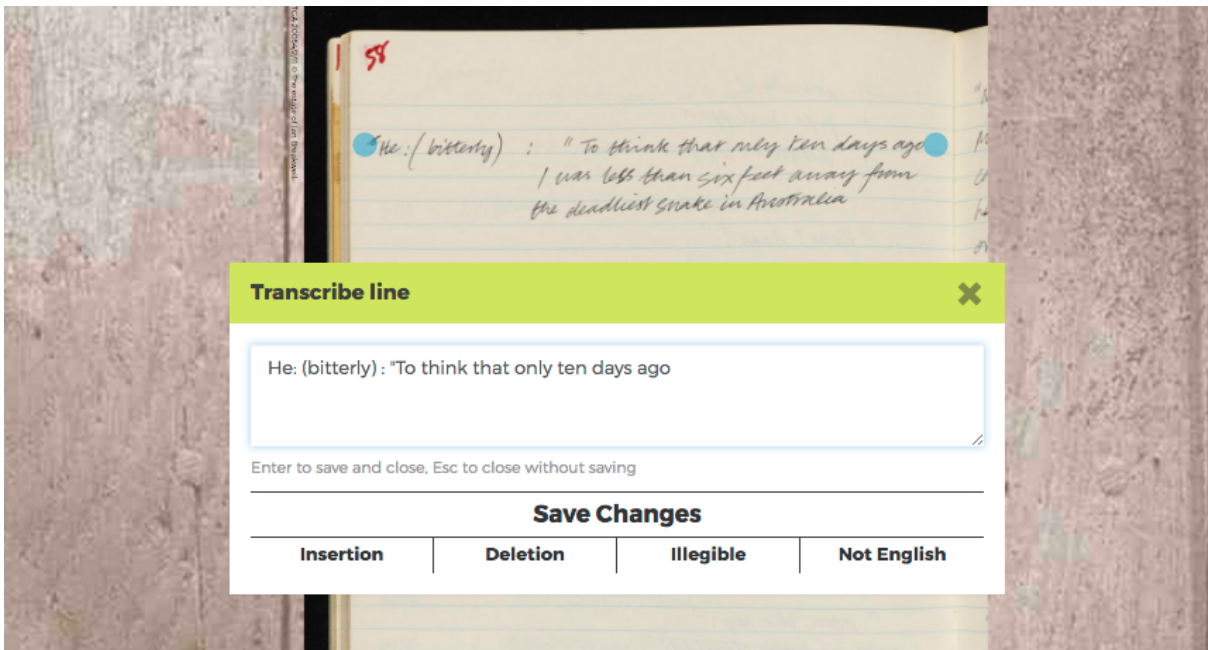


Figure 5.5 – AnnoTate’s transcription editor is intended for line by line transcription, drastically reducing the minimum text entry field required.

while navigating and working in a system that is new for them [Eveland Jr and Dunwoody, 2016].

Some basic guidelines for accompanying users in a web environment or workflow include providing feedback and status information for user actions [Smith and Mosier, 1986]. For example, much like when users are working with text or data manipulation and they should see the effects of their actions on the screen, the system should provide confirmation of users’ changes to data. Once again, this is where visual markers or syntax highlighting is useful. In learning environments, user guidance also often includes adapted feedback based on user input. We discuss user feedback in more detail as part of systems in Chapter 10.

For designing instructions, information and advice should be presented in a concise manner, including aids, FAQs, and documentation. Also, affirmative statements should be used and instructions should be presented so as to call on underlying understandings or users’ existing reading and text-processing skills [Smith and Mosier, 1986]. Finally, considering limited menu options may be appropriate for work environments intended for

novice users [Webster and Ahuja, 2006].

5.6 Conclusion

User interface design, like web design in general, is an iterative process. During the process of designing a user interface, interaction between designers and users is incredibly important as it allows to evaluate designs based on user needs and then implement changes based on these evaluations. This process is iterative because it is repeated several times before finalising the design of a user interface.

We used knowledge about web design and user interface design guidelines to produce an environment where users would easily find what they need to participate in activities and would not be disoriented in apprehending site structure or individual pages. The goal was of course to have participants want to continue their involvement in the site's activities.

So far in this methodological part, we have looked at data, encoding, and underlying processes for transforming digital facsimiles into dynamic texts. In this chapter we focused on understanding and describing some basic user activities and needs and using this information for designing interfaces. We looked at interfaces because these are the first points of contact that users have with systems. We have also shown how in current digital contexts user interfaces are used as work environments for transforming information and producing data. We now need to consider how to analyse the data produced through transcription. That is, how documents can be compared amongst themselves and also to expert transcription references.

In the next chapter we will look at the kinds of tools that we can use for evaluating the products of crowdsourced transcriptions. To do so we will look at existing methods for comparing documents and measuring differences between them.

Chapter 6

Methods for comparing documents

Contents

| | | |
|------------|--|------------|
| 6.1 | Chapter Summary | 97 |
| 6.2 | Comparing documents | 98 |
| 6.3 | Measuring differences between texts | 99 |
| 6.3.1 | Measuring differences between XML | 101 |
| 6.3.2 | Algorithm complexity | 104 |
| 6.3.3 | Pre-processing transcriptions | 105 |
| 6.4 | Clustering techniques | 106 |
| 6.5 | Visualisation | 109 |
| 6.6 | Conclusion | 109 |

6.1 Chapter Summary

Having looked at data encoding, tools for doing so, and what can be done with results, we now direct our focus on methods for analysing data quality. This chapter presents techniques for measuring differences between texts as well as XML documents using distance metrics. We will also describe the usefulness of clustering techniques, commonly used as methods of classification, for observing similarities and differences between documents.

| | |
|---|---|
| 1 - (1 | 1 + Laure 4. Janvier 1840. 4 1) |
| 2 - 4 J. | 2 + Je suis réellement si pressé , mon cher |
| 3 - N.° 250 - T. e 2. - p = 279 | 3 + ami, que mon écriture serait illisible |
| 4 - 666 | 4 + tu as du recevoir une paire des cornes |
| 5 - 499 | 5 + qui ont du te ravir en admiration. Je |
| 6 - 249 | |
| 7 - Rome | |
| 8 - 4 Janvier 1840 . | |
| 9 - Je suis réellement si pressé, mon cher | |
| 10 - Ami, que mon écriture serait illisible, | |
| 11 - Tu as dû recevoir une paire des cornes | |
| 12 - de buffles qui ont dû te ravir en admiration. Je | |
| 13 - te prierai de faire porter ces cornes dans | 6 te prierai de faire porter ces cornes dans |
| 14 - le second salon de mons. r Difiore. C'est là | 7 + le second salon de monf. Dijioré . C'est là |
| 15 - l'expression de mes souhaits de bonne année | 8 + l'expression de mes souhaits de bonnes annés |
| 16 - Tout ce que disent les journaux sur le rôle | 9 + Tout ce que disent les journeaux sur le rôle |
| 17 que joue ici un gros jeune homme fort | 10 que joue ici un gros jeune homme fort |
| 18 - mal bâti, est ridiculeusement faux. L'ambas | 11 + mal-bâti est ridiculement faux. L'ambos |
| 19 - sadeur a donné une feête à la quelle | 12 + saveur à donné une poêle à la quelle |
| 20 - tous les Cardinaux ont assisté ; il était | 13 + tous les cardinaux ont assisté ; ils étaient |
| 21 à la lettre impossible de changer de place | 14 à la lettre impossible de changer de place |
| 22 - dans les deux magnifiques Salons du Palais | 15 + dans les deux magnifiques salons du palais |
| 23 - Colonna. Jamais je ne vis tant de diamants | 16 + Colouna . Jamais je ne vis tant de diamants |

Figure 6.1 – Observing text differences in two texts using a text comparison interface.

6.2 Comparing documents

Document comparison is a task that is well entrusted to computer programs, which can detect changes between two versions of a document, making this task less burdensome than using manual comparison.

Typically, document comparison software has an interface that indicates to users where modifications occur in documents. Text is often highlighted to represent what text has been removed and what text has been added between the two documents. Many versions of this type of software exist, and many of them are accessible online. For example, in Figure 6.1, we can see how one such interface can be used to quickly detect the difference between two transcriptions of a page from the Stendhal corpus. The interface is divided vertically to show both texts being compared. Text highlighting is used to show which parts of the text on the left were deleted (in red) and which were added (in green) to obtain the text on the right. This visual support is very useful for detecting details that would otherwise take much longer to trace when performing close readings of documents.

We will present and explain the principals on which text comparison is based in order to better understand how the underlying algorithms can be applied in our case for comparing and measuring differences between contributed transcriptions.

6.3 Measuring differences between texts

Text comparison interfaces give us a visual support for observing the differences between texts. However, when we have to compare more than two, or three texts, even with the help of text highlighting the task can quickly become unmanageable. We need to discern the differences between all texts in a set of similar texts in some other way which would allow us to observe overall differences. In other words, to express the differences between texts we need to obtain quantifiable measures of difference.

We explored a number of options for measuring differences between transcriptions using string metrics. Those we used included basic online text-diff editors that, beyond highlighting deletions and additions, also quantified each operation. We also used PHP scripts to do the same, and finally, we implemented several useful Python libraries for the same purpose.

Common string metrics express difference between strings as *distance*. Among these, we can cite several that we have come across over the course of our work including Hamming distance and Levenshtein distance.

Hamming distance

This algorithm consists in counting the number of positions where characters differ. Its main drawback is that it requires that two strings be the same length in order a for comparison to be possible. Since we cannot be certain that all transcriptions contain the same number of words, this algorithm is not possible to use in our case.

| | |
|---|---|
| Deletions: 11 Additions: 14 | |
| Rome 4 Janvier 1840. | Laure H. Janvier 1840. 4 |
| Je suis réellement si pressé, mon cher | Je suis réellement si pressé, mon cher |
| ami, que mon écriture serait illisible, | Ami, que mon écriture serait illisible, |
| tu as dû recevoir une paire des cornes | tu as dû recevoir une paire des cornes |
| de buffles qui ont dû te ravir en admiration. | de buffles qui ont dû te ravir en admiration. |
| Distance = 25 characters | |

Figure 6.2 – Measuring text difference.

Levenshtein Distance

Levenshtein Distance is often applied in linguistics and computer science to measure the difference between strings [Levenshtein, 1966]. This difference is attributed a measure using the quantifiable distance between two texts. The operations allowed include additions, deletions, and finally substitutions at the character level [Levenshtein, 1966]. In our case, it can be considered as a measure of the minimum number of corrections necessary to go from one transcription to another. This is our chosen method of comparing transcriptions. The formula can be given as follows:

$$Distance_{i,j} = additions_{i,j} + subtractions_{i,j} \quad (6.1)$$

As shown in formula 13.1, the distance between two texts i and j can be obtained by calculating the sum of the number of additions and subtractions necessary to transform text i into text j .

On text, Levenshtein distance calculations are performed at the character level so that each character of each word (including spaces) that is added or subtracted is counted to obtain an overall *Distance* measurement. Figure 6.8 on page 110 shows an example of this¹. The result of text comparison is made evident to the user by highlighting the deletions and additions. We have added a tally of these operations on the text and the total *Distance*.

Levenshtein distance is also known as *edit distance* and a number of useful algorithms

1. We used an online text difference tool for this example, available at <http://www.diff-online.com/fr>

have been adapted to monitor the operations of addition, deletion, and substitutions. Most importantly, calculating edit distance between multiple documents can allow us to determine those that are most similar to one another as well as those that are least similar. Edit distance can be performed not only on strings, but also on structured documents [Zhang and Shasha, 1989].

6.3.1 Measuring differences between XML

Comparing XML is more challenging than comparing textual strings. Simply put, an XML file contains encoded, or structured, information and determining the differences between XML documents requires looking at the differences in the structure of elements that compose these documents [Nierman and Jagadish, 2002]. A number of algorithms have been described for computing changes to XML documents, including Chawathe et al.'s algorithm [Chawathe et al., 1996], Nierman & Jagadish's algorithm [Nierman and Jagadish, 2002], and Zhang & Shasha's algorithm [Zhang and Shasha, 1989]. We will briefly present these and explain their differences.

Chawathe edit distance metric

Chawathe et al. [1996] suggest a method to detect changes as well as move operations when comparing XML documents. Changes can include additions and deletions of elements, and elements can also be moved from one part of the document tree to another.

The authors also express the idea that this type of operation is challenging because, even in hierarchically structured information, sentences or paragraphs do not have key identifiers.

With Chawathe's edit distance metric, the operations that are performed are summarized as following: node addition, node deletion, node update, and finally subtree move. The first three operations are the XML equivalent of insertions, deletions and substitutions. The fourth responds to the hierarchical problem described by the authors. To illustrate, in Figure 6.3 we show a simplified representation of these operations to transform T1 into T2. As stated by [Chawathe et al., 1996] the algorithm works on or-

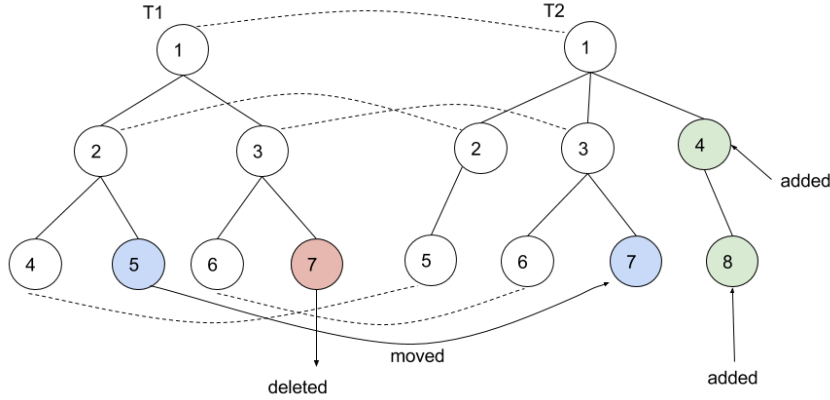


Figure 6.3 – Tree operations as described by [Chawathe et al., 1996].

dered trees, thus we have numbered the nodes in the trees shown. When the trees are compared, the process begins by matching nodes with equal values in trees T1 and T2; matched nodes are connected with dotted lines. Then the algorithm takes steps to identify which nodes have been deleted from the first tree, which have been added to the second tree, and which have changed position between the two. These operations of addition, deletion, and movement are summarised in the illustration. Node updates, however, are not shown; these are specific functions used in the algorithm to update the values of nodes themselves.

Nierman and Jagadish edit distance metric

The authors describe their approach in [Nierman and Jagadish, 2002] as basically the same as the one described in [Chawathe et al., 1996], except that the Nierman and Jagadish algorithm allows sub tree insertions and deletions. This means that the algorithm detects multiple nodes that make up a sub tree within a document and counts this insertion or deletion as one operation on a sub tree rather than several operations on a series of nodes. Since the other main operations of addition and deletion are the same as described in [Chawathe et al., 1996], Figure 6.4 focuses only on subtree operations and illustrates how these would occur between two trees T1 and T2.

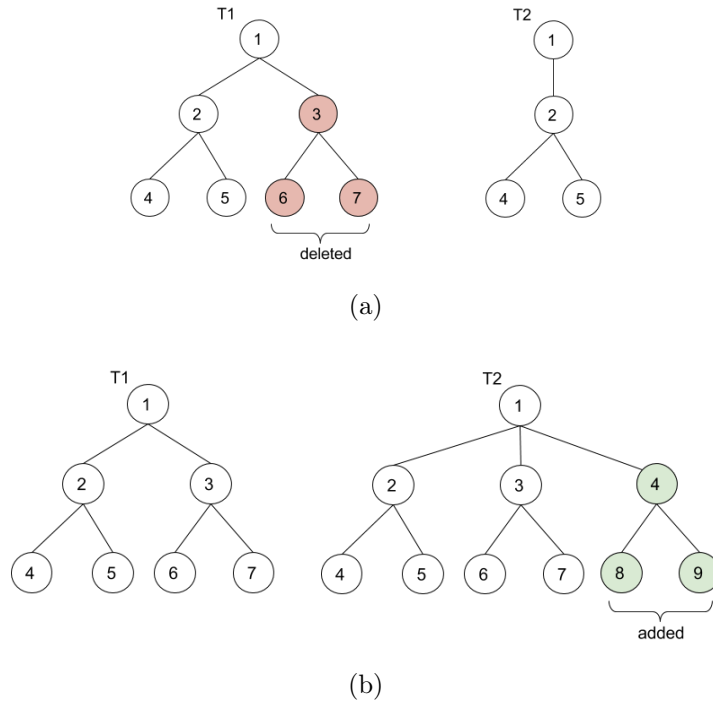


Figure 6.4 – Deletion of an entire sub tree is shown in (a) and addition of another subtree is shown in (b).

Zhang & Shasha edit distance metric

The Zhang & Shasha algorithm is also able to determine the distance between two XML trees based on the number of operations, including additions, deletions, and modifications, necessary to transform one tree into another. This algorithm is characterised as allowing additions and deletions of any single element in the tree, regardless of its location in the tree. When this happens, the child elements of the node are first attached to the parent element of the node, then the node is deleted. Figure 6.5 shows this procedure on T1, then shows how the same node can be added to another place in T2. This is what is meant by the modification of an element. If we use the analogy of the document tree, then we can say that additions and deletions can concern branches of the tree and not only its leaves. The hierarchy of the tree can be altered without loss of dependent leaf elements. Unlike the Nierman and Jagadish method, this algorithm does not allow for subtrees to be added or deleted in one single step [Nierman and Jagadish, 2002]; doing this would

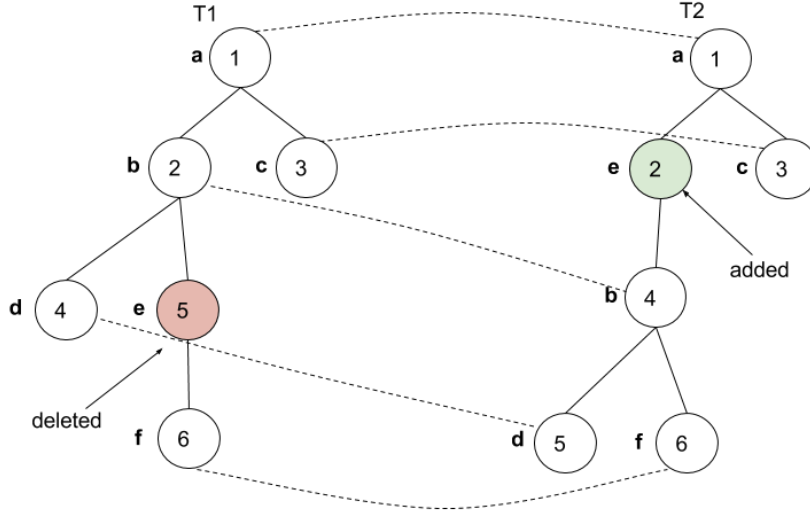


Figure 6.5 – Tree operations as described by [Zhang and Shasha, 1989]. Element **e** is deleted from T1 without affecting its children nodes, then added to another spot, resulting in its position in T2.

require multiple operations, which would inevitably raise the associated distance measure and affect the algorithm’s execution speed. However, it may also more accurately reflect the amount of manual operations that would be necessary to correct an erroneous XML file.

The main advantage of using Zhang and Shasha’s algorithm over the other two is the simplicity with which it can be implemented². This algorithm’s role in our overall analysis method is represented in Section 6.6, Figure 6.8.

6.3.2 Algorithm complexity

We report in Table 6.1 the complexity of the 3 algorithms we presented. We see that the complexity is $O(|T1||T2|)$ for [Chawathe et al., 1996] and [Nierman and Jagadish, 2002] and in the case of [Zhang and Shasha, 1989] has the potential of being higher as a complexity of $O(|T1||T2|depth(T1)depth(T2))$ takes into account the depth of the tree. The factor $depth(T1)depth(T2)$ will increase the computational cost of the algorithm

2. We use the Zhang-Shasha module, written in Python

when the number of tree levels increases. For instance, in our Figures 6.3, 6.4, and 6.5, we show trees having a 3-level hierarchy, composed of a document root, parent elements, and children elements. If trees T1 and T2 had a depth of 10, the computational cost of the algorithm would be 100 times higher than using the other two algorithms.

| Algorithm | Complexity |
|------------------------------|---------------------------------|
| [Chawathe et al., 1996] | $O(T1 T2)$ |
| [Nierman and Jagadish, 2002] | $O(T2 T2)$ |
| [Zhang and Shasha, 1989] | $O(T1 T2 depth(T1)depth(T2))$ |

Table 6.1 – Complexity of presented algorithms. The notation $|T|$ is used to denote the number of node in the tree T and $depth(T)$ is the number of edges from the the root node to the deepest possible node.

6.3.3 Pre-processing transcriptions

Our transcriptions are received as XML documents. If we want to calculate the string distance between transcriptions, we have to transform them into raw texts by removing all XML encoding elements. To do so, we first retrieve the XML tree containing the text itself (stored inside a content tag). As transcribers may or may not have added breaklines, we convert all breaklines to spaces. Then, we trim all trailing spaces to obtain the batch of raw texts that will be compared amongst themselves.

When performing the measurement on XML document trees we need to keep the structure of the document along with all associated elements. We apply the Zhang & Shasha algorithm directly on XML files to obtain tree edit distance.

6.4 Clustering techniques

The distance values we obtain from *string edit distance* and *tree edit distance* allow us to quantify the differences between multiple texts. Now we need to organise these findings using what are known as clustering techniques.

Clustering is used in Computer Science to organize documents according to defined characteristics such as terms or keywords. It is useful for creating intelligent search systems and can be helpful in improving the organisation of collections. Clustering is well-known as a relevant technique for organizing large corpora. Justifiably, it can even be useful in classifying documents according to themes or subjects. It is common practice to use clustering to organize closely related documents together and distinguish these from unrelated documents [Huang, 2008]. Clustering is considered to be particularly effective on large and heterogenous data sets. Using this technique allows to group objects according to their similarities or dissimilarities.

Similarity between objects is often expressed as proximity. Typical representations of clusters are based on measuring the distance between objects, in order to determine if they belong in one group or separate groups. One object A is said to be more similar to an object E compared to an object B if the distance from A to E is lower than the distance from A to B. This situation is depicted in Figure 6.6. In this collection of objects A and E are closer compared to the other objects. The ovals represent clusters resulting from hierarchical classification.

In our case, the objects are transcriptions. To measure similarity between objects we use the notion of distance, which can either be taken literally as a metric distance between objects in space, as in the example, or be assigned a value based on the quantity of operations, or errors, separating two objects, as we described with Levenshtein distance. The units we use for texts are characters, whereas for XML we use element nodes that constitute the XML tree.

Depending on the units used, distance values will vary. This means that results will not necessarily correlate. One needs to understand that the values one is looking at

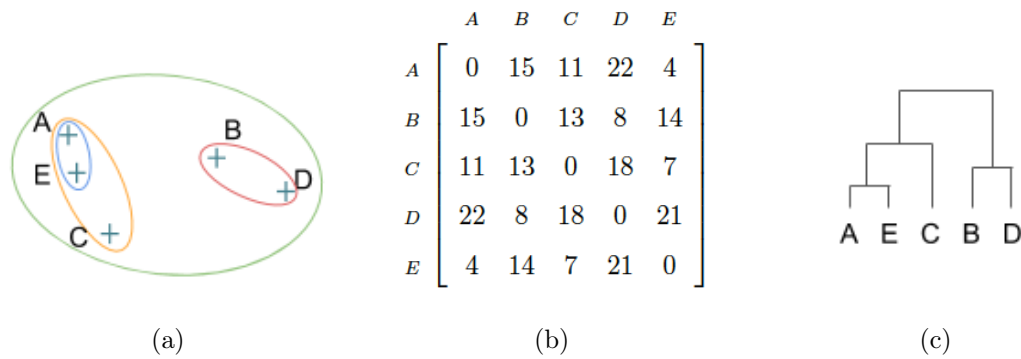


Figure 6.6 – Clusterization of objects in space is shown in (a). A distance matrix representing distance measurements (au) between objects in (b), and (c), a phylogenetic tree can be used to represent the objects based on distance values.

may represent different types of operations. This choice will define the distance between objects. For example, in a text, a distance value of 3 can represent the deletion or addition or addition of three characters, which can constitute a part of a word or a whole word if it is short. This same value of 3 in the case of XML, represents three operations on entire elements. This can represent two deleted elements and one added element in a tree. Since elements can contain several characters, words, or even lines, the same number of changes in an XML may actually represent many more characters when looking purely at text. This explains why distance values are often higher for the same documents, depending on whether one is observing text distance or XML distance.

In our case, we use what is known as agglomerative hierarchical clustering. It consists in iterating through the data set to find the closest pairs of objects and forming them into clusters, then we merge these to form bigger and bigger clusters, until finally, obtaining the overall cluster. If we consider Figure 6.6 as a hierarchical clustering process, it would consist of the following steps:

1. Objects A and E are the closest, they are joined together to form the blue cluster (A,E).
2. Objects B and D are merged to form the red cluster (B,D).

3. Cluster (A,E) and object C are merged to form the cluster yellow ((A,E),C).
4. Finally, clusters (B,D) and ((A,E),C) are merged together to form the largest green cluster.

These steps are the equivalent of an algorithmic process for grouping objects based on their proximity. For our purposes clustering is a useful way to sort transcriptions and visualise results. Without this technique it would not be possible to make observations we describe in Chapter 11.

The way clusters will be formed will depend on the linkage criteria used. Popular criteria are single-linkage and complete-linkage. These two linkage criteria produce different clustering results. With single-linkage, in order to determine which groups of objects will constitute clusters, we find the two closest objects of two different groups and link their associated groups [Everitt et al., 2001 ; Manning et al., 2009]. With complete-linkage as a criterion, we use the maximal distance between objects of two different groups, which means that the similarity of two clusters is determined by their most dissimilar objects [Everitt et al., 2001 ; Manning et al., 2009]. In the example we give for Figure 6.6, the yellow cluster (A,E,C) and red cluster (B,D) are merged to form the green cluster. Depending on whether we use single-linkage or complete linkage, we will rely on different points to create the green cluster. Figure 6.7 shows examples of single-linkage and complete linkage for this cluster set. In the example shown, regardless of which linkage we use, we obtain our green cluster, however, depending on whether other objects or clusters are present, the result could be very different. In our case, we rely on complete-linkage to cluster transcriptions, because the complete-linkage criterion is not local and implicates entire structures to compose clusters [Manning et al., 2009]. For us, this is a better way of determining coherent groups of transcriptions.

The sorting operations that allow for the formation of clusters are executed on a matrix of distance values, which are obtained from the comparison of pairs of objects. This matrix is then converted into a notation format that is a machine representation of the proximity of objects. We then use existing libraries to process these to visualise results.

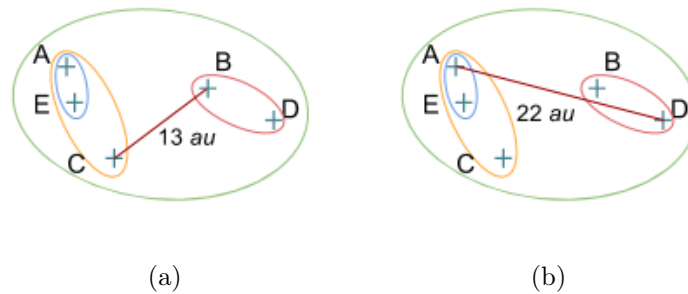


Figure 6.7 – Cluster linkages: (a) shows single-linkage and (b) shows complete-linkage.

6.5 Visualisation

Phylogenetic trees can be used as tool for visualising the relationships between clusters, as we have shown in Figure 6.6, where (c) shows a phylogenetic tree drawn based on cluster groups represented in (a) and their distance values represented in (b).

Phylogenetic representations are commonly used for cluster analysis and a number of functions exist for this purpose in different languages. We have come across jsPhyloSVG³, which is a Javascript library for visualising phylogenetic trees, and have implemented Python’s Seaborn library⁴ for statistical data visualisation.

To accompany phylogenetic visualisation we can generate heat maps, which are also based on distance values. Heat maps are created by associating colours with numerical values. Low distance values map to soft colours that gradually intensify as distance values rise. Heat maps can also allow to identify cluster formations and their boundaries.

6.6 Conclusion

To compare transcriptions based on their similarities or differences, we can apply the methods that we have presented here. To do so we have created a document processing

3. <http://www.jsphylosvg.com/>

4. <https://seaborn.pydata.org/>

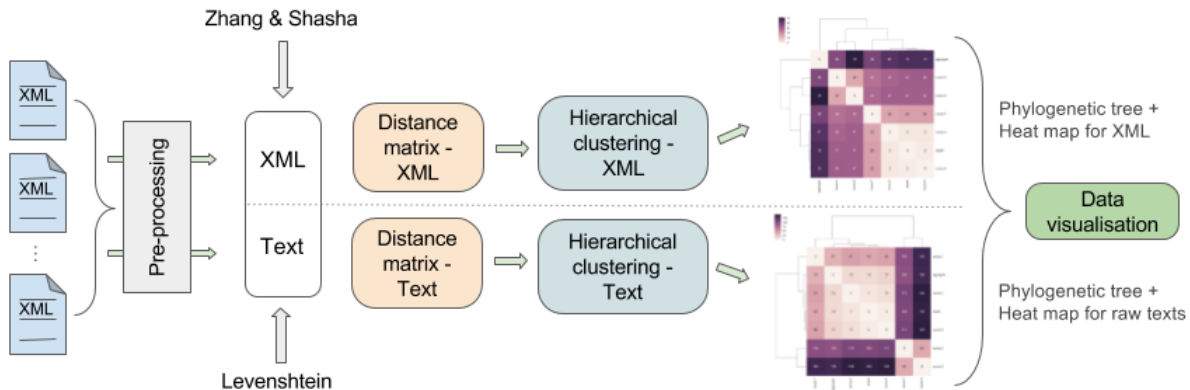


Figure 6.8 – Schematic flow of transcription analysis. Transcriptions are pre-processed to extract either raw text or xml structure. Then, the distance matrix is computed which is used to compute the hierarchical classification. Finally, the result can be seen using either a phylogenetic tree or a heat map.

pipeline, which we represent in Figure 6.8. The overall process is as follows. We begin by comparing a batch of transcriptions created from the same manuscript object. To compare raw texts we remove all elements and apply the Levenshtein distance metric. For XML, we apply the Zhang & Shasha algorithm. We obtain distance values for all transcription pairs, which we record into a matrix. We do this for both text and XML, resulting in two matrices. From these, we then use a hierarchical clustering algorithm to draw phylogenetic trees. Python’s *Seaborn* library is good for drawing both trees and heat map representations that allow to visualise these clusters⁵.

In this chapter we explained the processes that can be used to compare documents, measure similarities between them, and hierarchically determine which groups of documents are more similar. Many clustering-based applications are used to group documents based on keywords, but we use the values obtained by measuring Levenshtein-type operations on transcriptions to obtain representations of similarities and investigate the distributions of our results. In other words, we use this method to analyse experimental data collected using our transcription platforms. In the following two chapters we describe

⁵. More examples of this form of visual representation can be seen in Annex B.2. We present more elaborate representations with our findings in Chapter 11.

these platforms in greater detail.

Part III

Prototype and production implementations

Part III summary

Burdick, Drucker, Lunefeld, Presner, and Schnapp (2012) insist that prototyping be accepted as an important part of research in Digital Humanities:

a production based endeavor in which theoretical issues get tested in the design of implementation, and implementations are loci of reflection and elaboration” [Barber, 2016 cites [Burdick et al., 2012]].

In Part Three we introduce the results of our prototyping process, which has led to the creation of an online transcription platform. In Chapter 7 we describe a production implementation of PHuN 2.0 and in Chapter 8 we present an experimental variation of the platform which we used to collect data. Finally, in Chapter 9 we discuss the human element that plays essential roles in collaborating, motivating, and communicating in order to constitute and maintain virtual communities. We also discuss the importance of skills and training in such environments.

Chapter 7

Presentation of PHuN 2.0

Contents

| | | |
|------------|--|------------|
| 7.1 | Prototype types | 119 |
| 7.2 | Presenting PHuN 2.0 | 121 |
| 7.3 | Editor functionalities | 126 |
| 7.4 | Functionalities for identified users | 128 |
| 7.4.1 | User accounts | 129 |
| 7.4.2 | Page browsing and selection | 130 |
| 7.4.3 | Transcription and revision | 130 |
| 7.4.4 | User comments and discussion | 132 |
| 7.4.5 | User profile | 133 |
| 7.5 | Project Leader functionalities | 133 |
| 7.5.1 | Project creation and corpus integration | 134 |
| 7.5.2 | Project configuration and management | 135 |
| 7.5.3 | Transcription protocols | 136 |
| 7.6 | Discussion on limits and improvements | 138 |
| 7.7 | Conclusions and next steps | 141 |

In Chapter 7 we focus on our experimentation in prototyping and creating a transcription platform. Through various components, ideas, and issues that we encountered we are

able to reflect on the design and implementation of user tools and environments. We discuss the ways that the platform can reflect working ecosystems and relationships between users. We also consider how user knowledge and skills can fit into such environments. We conclude by addressing improvements we deem necessary for the existing system.

7.1 Prototype types

In this section we will discuss prototyping as a way of introducing the development work of creating our transcription platform. We will first present three different categories of prototypes and relate how each can be useful under different circumstances or in order to respond to particular needs.

We wanted to collect crowdsourced transcriptions for our analysis and thus we created a transcription tool that could be used in an online environment. For our expected participants we made an online work environment, with functionalities for browsing manuscripts, selecting pages to work on, and viewing completed transcriptions. We also created the possibility to access their completed transcriptions from users' personal accounts. Since there was a high chance of having remote participants, it was also important that they have access to instructions and some way to initiate discussions with other participants.

Starting from scratch, we knew that the tools we created and implemented would need to evolve in order to achieve expectations. Consequently, as many DH scholars would support, prototyping proved to be an essential part of the research process [Galey and Ruecker, 2010 ; Ruecker, 2015].

In an article focusing on prototypes Stan Ruecker presents three distinctive categories. These categories are production-driven, experimental, and *provotypes*¹ or provocative prototypes. Unlike predecessors whom he mentions as having introduced interesting taxonomies, Ruecker suggests classifying prototypes based on types of projects that they are intended for [Ruecker, 2015].

For Ruecker, Production-driven prototypes are meant to achieve a working version of a product or system at the end of a given period of development. This form of prototype will eventually be introduced to the public after undergoing a series of successive improvements in the form of iterations or versions. The ultimate goal is to take an initial prototype and implement improvements on it in order to achieve a robust functioning model intended for use.

1. A term Ruecker borrows from Boer and Donovan in [Boer & Donovan, 2012]

Experimental prototypes are different from production-driven prototypes in that they are not necessarily intended to become independent working systems. "The goal is not to create a product but instead to produce a kind of generalized knowledge about an idea that the prototype embodies" [Ruecker, 2015, p. 2]. In this way, an experimental prototype is used simply to test an idea, which may develop into another idea or even multiple other ideas requiring the creation of more prototypes. The development of experimental prototypes allows for exploration. The prototype may also undergo multiple iterations, as with production-driven prototypes, but the result may branch out into new research questions and possibilities [Ruecker, 2015].

The third and final category described by Ruecker is the provocative prototype, which aims neither to develop a working system nor directly address any research questions but, as its name suggests, aims to provoke a reaction from users so as to ultimately challenge the ways that people or society approach certain subjects. These types of prototypes are often of a more creative nature as they intend to introduce previously untapped subjects of inquiry into a dominant structure or discourse [Ruecker, 2015].

Prototype categories reflect the process of scholarly rationalization on the subject, which indicates that Digital Humanities do more than just create or provide tools for humanities research. Processes of experimentation and creation are accompanied by reflection and analysis, which are also key in acquiring new knowledge. This can also help to to harmonize tensions between humanities and computing, particularly those arising from questions regarding the value of what each brings to the relationship.

With these three categories in mind, to which do our prototypes belong? To answer this question we will need to delve deeper into the intentions behind our project, the development of functionalities for PHuN 2.0, and its evolution as PHuN-ET over the course of our work.

7.2 Presenting PHuN 2.0

PHuN2.0 is most closely defined as a production-driven prototype as its aim is to create a working environment for researchers and the public. Its architecture and functionalities were implemented with the intention of creating a robust system for many users, but also for accomodating multiple projects and many data objects. We will outline the functionalities included in the system and also point out necessary improvements.

The decision to develop the very first version of the prototype, which was written in simple PHP²/HTML³/CSS⁴/Javascript⁵, using Symfony⁶ was largely motivated by the understanding that as the system grew and developed it would be increasingly difficult to maintain. We used PHP's Symfony Framework to have access to an active community of developers and recent documentation. Using the Symfony framework gave us immense flexibility in creating an architecture that reflects our data while applying best practices based on an MVC (Model-View-Controller) pattern [Peltier, 2011]. This type of framework allows to separate the database from the logic that operates on data and the views that present it in web pages. Figure 7.1 illustrates this architecture as a simple relationship between model, view, controller, and ultimately the user. The MVC model makes it easy to present data in different views without inherently modifying the model [Gamma et al., 1995]. It also allows us to take advantage of a large collection of existing components, known as bundles, that can relatively easily be implemented to add functionalities to the existing system.

We wanted the system to be able to seamlessly handle integration of manuscript images. As the platform developed we imagined the possibility of stocking large volumes of images from different collections and their associated transcriptions.

User account security was also an important matter, and Symfony's User bundle provides functional code for handling account registration, sign in and password reset. We

2. <http://php.net/manual/en/intro-what-is.php>

3. <https://www.w3.org/html/>

4. <https://www.w3.org/standards/webdesign/htmlcss>

5. https://www.w3.org/wiki/The_web_standards_model_-_HTML_CSS_and_JavaScript

6. <https://symfony.com/>

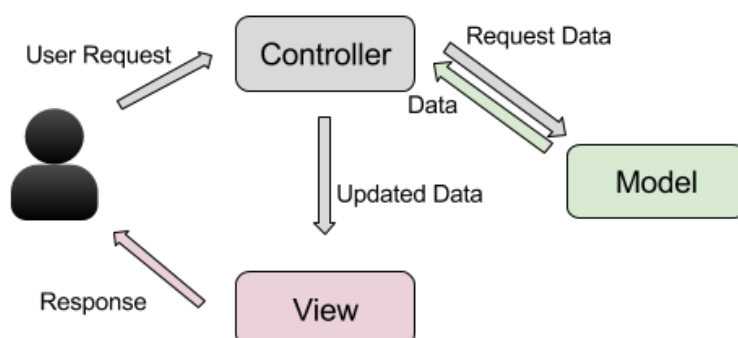


Figure 7.1 – Illustration of relationships between components that make up an MVC framework: model, view, and controller, and users of a web application.

were able to successfully implement numerous improvements to PHuN 2.0 since launching it online and its MVC pattern has consistently simplified maintenance.

Developing the first production platform, PHuN 2.0, was a crucial part of the project. The platform provided an accessible work space for participants. Work flow was organized into transcription, editing, and revision, and each phase was open to all users. Participants could also interact with others by posting their observations or questions on a discussion list; each page has its own. The platform serves as proof of concept and functioning model for an editorial space focusing on transcription that may include many different projects. This platform's creation has played a vital role in the development and study of participative and crowdsourcing methods for manuscript transcription. An earlier version of the platform has also been adapted for working on scholarly editions at the University of Paris Diderot.

This online work environment was created for different kinds of users, where anyone can sign up and begin working on available projects. To begin, roles are clearly defined within the system, resulting in a hierarchical structure. Firstly, projects are created and maintained by project leaders, then transcriptions are solicited from contributors through loosely defined channels that can be defined by the project's ties to cultural,

heritage or academic institutions, and its geographical location. In Figure 7.2 we show the main components of the platform’s architecture. Purple shaded areas indicate processes accessible to project leaders and blue areas show those accessible to users having accounts. The rest of the site is accessible without an account. Arrows indicate how one would access particular areas of the site, including those intended for project leaders or account-holding participants. Simple lines indicate page hierarchies, for example, to show which pages are accessible from the project management menu and which are accessible from each project’s catalogue.

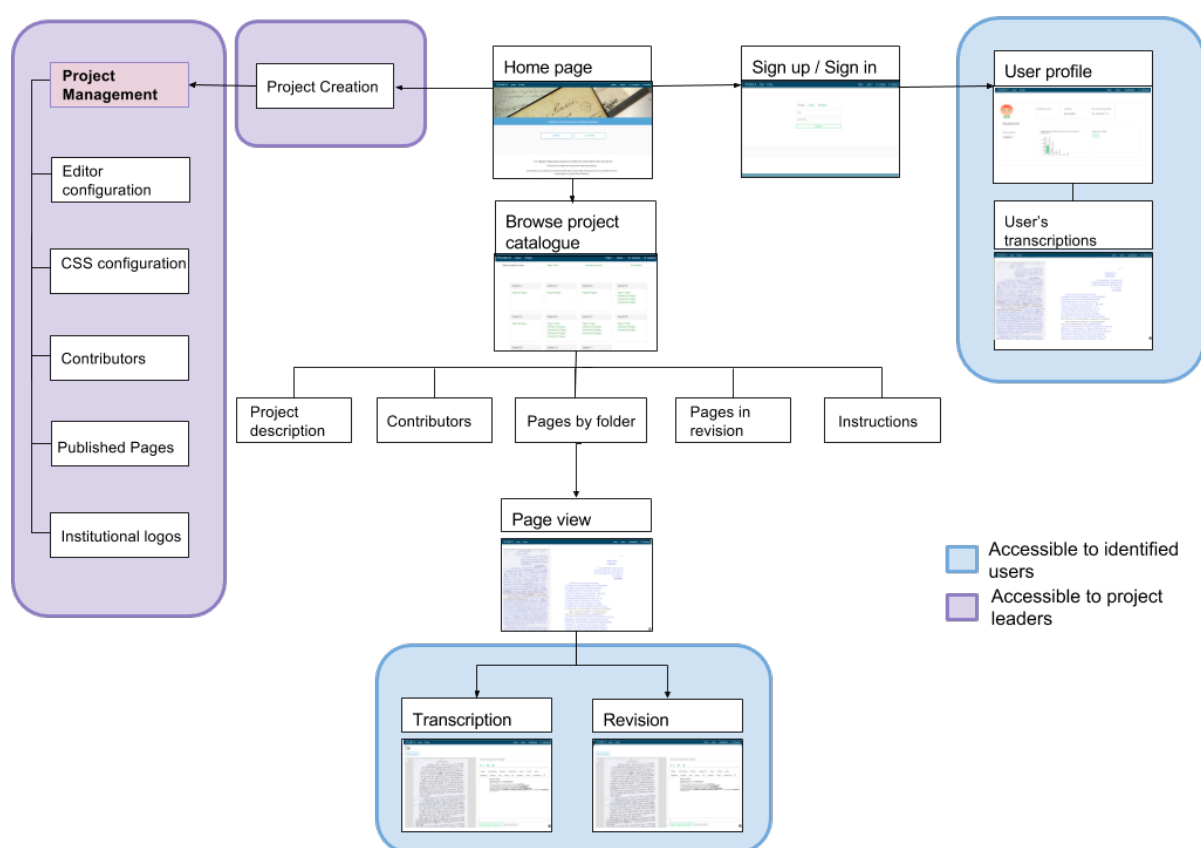


Figure 7.2 – PHuN 2.0 site architecture.

In order to simplify the process of project creation as much as possible, we have taken example from projects such as Zooniverse, who propose online project generators [Arfon, 2013]. In the case of PHuN2.0 interested project administrators make contact with platform administrators via an online contact form. They indicate their interest in beginning a new project and give a brief description, leaving their contact information. This allows

for platform administrators to contact interested parties and establish a protocol for the transfer of image files. This part of the project creation process has not been automated as file uploading is potentially a heavy task, which may be unsuitable via an online form and may be better handled by direct SQL injection of data fixtures. This also allows platform administrators to have more direct contact with potential project leaders, ensuring that the interest is real and avoiding potential problems.

For project leaders, the project creation process is broken down into two steps. The first is mainly to fill out a project creation form and upload a project description with the necessary project files: a cover image, as well as a project DTD and CSS – two of the only technical documents that project administrators must be able to supply themselves, and according to provided guidelines. The second step involves the configuration of a WYSIWYG editor that reflects the project's chosen XML schema outlined in the DTD. The project administrator confirms the creation of WYSIWYG toolbar element names that correspond to their XML elements and decides on the organization of these elements within the editor, either accessible directly in the toolbar or located in one of the editor's dropdown menus. This ensures the compactness of the editor and the ease of use of the interface. The configuration step also allows adding XML elements that were not originally present in the DTD, project leaders should be vigilant in making sure that they update the corresponding CSS to apply a presentation style for new elements. The project administrator must save the settings to confirm the generation of the editor and the associated project. He or she can return at any time to the administrator menu to modify the configuration of the editor by adding or deleting elements, without having to modify their DTD schema. Once the settings are configured the project is ready to start and the project administrator may involve collaborators to participate.

The project management menu integrates a certain number of essential functionalities to manage a project. These include the possibility to update the CSS, to consult a list of contributors and change their roles, and to consult both transcriptions in-review and published ones. Project leaders can also de-validate published transcriptions that they judge incomplete or erroneous, which sends these documents back into the transcription

and editing cycle. The functionalities included in this menu have been developed progressively as the project evolved and as needs for greater manoeuvrability and control became apparent. Project leaders can create projects, and then their role is to oversee these projects, to describe their objectives, and to implicate users. They must verify that transcriptions that achieve published status are accurate and complete, and take measures to correct them if necessary. They may also need to change the encoding schema of their project and should be able to update the editor. Finally, they should be able to upgrade users who can help them in their administrative roles and demote users who no longer fulfill these roles. The system has been created to be able to incorporate these actions and as the project evolves the platform will certainly see new functionalities added to complete and improve it.

If the platform, with its possibility of creating numerous projects, resembles other existing infrastructures, it nevertheless incorporates an innovative aspect with respects to its configurable editor. That is, each project's editor can be specifically configured to reflect that project's own XML encoding vocabulary. The WYSIWYG editor can have as many or as few elements as necessary, and only those elements that are specifically decided upon by project leaders themselves. Figure 7.3 shows an example of an editor, which was configured for transcribing the Benoîte Groult corpus. It contains a toolbar of unique terms used by the project and four menus that regroup other related terms. In this specific example we used icons to represent some of the terms required by the project leaders to solicit users' visual recognition of their functions. The tags that are created when these buttons are pressed correspond to the project's own XML schema. For example, the strikethrough and underline buttons produce *<rature>* and *<souligne>* elements respectively. Users can see these corresponding terms when they hover a cursor over the buttons.

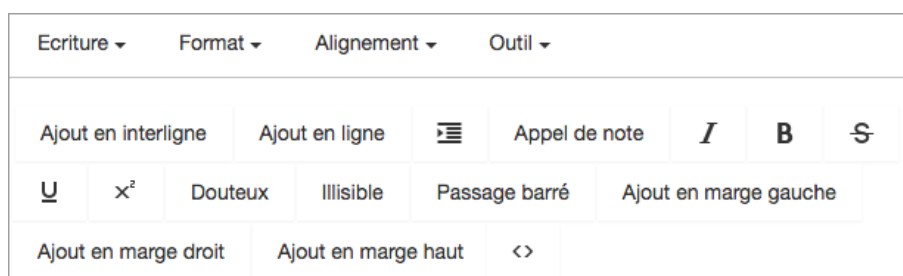


Figure 7.3 – Benoîte Groult editor configuration example.

7.3 Editor functionalities

Before presenting functionalities specifically geared toward transcribers or project leaders we present the functionalities associated with PHuN 2.0’s editor and common to both types of users.

The editor contains some basic functionalities that are intended to make it more easily accessible for inexperienced transcribers. Below is a list of these functionalities.

- Each button of the editor corresponds to an element belonging to a project’s XML vocabulary as defined by project administrators.
- Each button has a description explaining the corresponding element’s function. This description becomes visible when hovering the cursor over the element.
- Elements can be presented either directly in the toolbar or regrouped as menu items.
- We added one of TinyMCE’s existing plugins that allows users to access encoding if and when necessary.

The editor allows to structure content by simply selecting wanted text using a cursor and then clicking on buttons that correspond to elements that should be placed around that selection of text.

We initially explored several options for editors, including CKEditor⁷ and TinyMCE⁸. TinyMCE proved simple to create custom plugins that corresponded to entire XML vocab-

7. <https://ckeditor.com/>

8. <https://www.tinymce.com/>

ularies. Also TinyMCE is supported by extensive documentation and a large community of developers, making it relatively straightforward to find solutions when problems arise.

The editor relies on a configuration file that is written at the time of project creation and finalized by project administrators, who decide on button names, placement, and corresponding CSS. The CSS, which we discuss in greater detail in Section 7.5 on page 133, is responsible for the visual representation of text assigned to each of the existing elements in a project's XML vocabulary or schema.

Creation of the editor

TinyMCE is a commonly used WYSIWYG editor for the purposes of editing HTML. It also provides the possibility to create customized plugins for specific needs. Plugins are a very generic tool to add metadata to text inside the editor. For instance, the *bold* plugin allows to put two elements `` and `` to wrap a selection of text. However, this functionality is not automatic. Our adaptation of TinyMCE within the Symfony environment also automates the creation of custom plugins based on projects' DTDs. The plugins will determine the buttons contained in the transcription editor and the terms used will be reflected in the elements produced. For instance, an element in the DTD named *addition* will create a plugin of the same name, meaning the editor will contain a button named *addition* and this button will wrap selected text with a pair of `<addition>` and `</addition>` tags. Throughout the life span of a project, leaders can adjust their editor and create new plugins and buttons by adding new terms or deleting unneeded terms.

To create a custom plugin, we need to create a javascript file containing information about the plugin. This information includes the element name, its description, and code that defines its behavior when it is triggered by the user. Thus, we created a table in the database called *Plugin* composed of 3 columns: *name*, *description* and *container*. The *container* column simply refers to where the plugin can be found in the editor, either inside the toolbar or inside one of the prescribed menus⁹. There is no behaviour column

9. We used a predefined list of menu names, but this too can be rendered adjustable by creating a new linked table in the database and a form to allow project leaders to define their own menu names.

for we want all of our plugins to behave the same way. The corresponding database entity is shown in the left part of Figure 7.4.

The editor configuration file is responsible for the organisation of elements within the editor, including its toolbar and menus. The configuration file lists names of plugins as defined by project administrators and so that TinyMCE can load the corresponding plugin files into appropriate menus or directly into the toolbar. Other miscellaneous options are also defined in this file, such as the theme of the editor. In our case, we have at times included default plugins such as *code*, which allows users to view raw XML code, and also *remove format*, used for removing XML tags from text without losing the text itself.

As already mentioned, administrators configure the editor and its elements at the time of project creation or during a subsequent adjustment. When this occurs, the appropriate controller is triggered and a new configuration is created, or changes are made to an existing configuration. This means that project leaders can make changes to their editors at any moment after initial setup, although they need to keep in mind that previously transcribed pages may need to be re-edited to include updated vocabulary. The controller handles both the creation of plugin files as well as the editor configuration file. The creation of a configuration file was actually an improvement upon an earlier version of the system wherein elements were loaded into an editor dynamically. With this new procedure, because the editor loads an existing configuration file, the computational cost associated with loading an editor instance is decreased, thus reducing latency for users when they open a transcription interface.

7.4 Functionalities for identified users

PHuN 2.0 has three different levels of users; there are unidentified visitors to the site, identified users having accounts, and project leaders. Unidentified users have no specific privileges other than viewing pages on display and transcriptions contributed by other users. They cannot create new transcription or participate in the editing process, which are privileges reserved for account holding users of the site. Project leaders create projects,

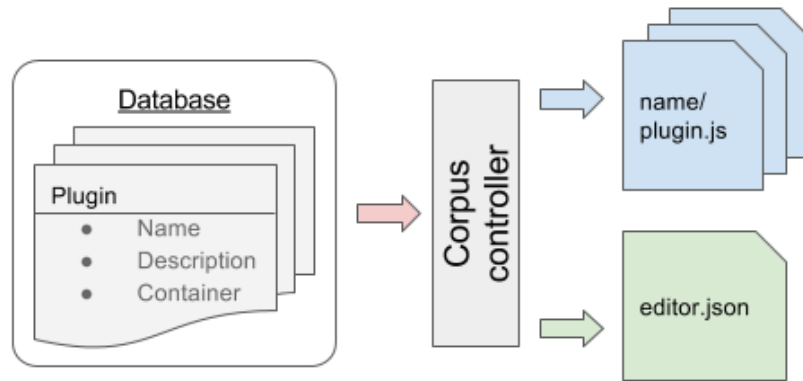


Figure 7.4 – Editor configuration with parameterized plugins.

coordinate editorial processes, and promote users; they have all the same privileges as regular identified users, but also hold those pertaining specifically to project management. In this section we will present the platform’s functionalities for identified users.

7.4.1 User accounts

At this time, user accounts are not specific to projects, which means that once an account is created its holder can contribute to any project listed on the site¹⁰. Project leaders may provide specific instructions regarding participation via the project’s description.

Once enrolled, users can log in to the platform using their chosen user name and password. If one finds him or herself locked out, it is possible to reinitiate one’s password by entering one’s user name and submitting the reinitiation form. The system will send an automatic e-mail to the user’s recorded e-mail address with a link to replace the old password with a new one.

10. Further development can allow to create more options for creating open, semi-open, or closed circuit projects. This in turn can be used to examine how to manage contributions from crowds and groups.

7.4.2 Page browsing and selection

From an identified user's perspective the platform interface is simple. All listed manuscript pages belonging to a project can be viewed regardless of whether a transcription exists for a given page or if a user intends to contribute a transcription. In PHuN 2.0 a catalogue browsing interface was created, which allows users to see how many documents exist in a collection and to browse by folder before selecting individual pages. Identified users can then choose a page to work on, or intervene on a page already begun by someone else. Unidentified users cannot participate in transcribing or editing, but they can view pages and transcriptions contributed by others.

7.4.3 Transcription and revision

If a transcription exists for a given page, it is visible to all users including those identified, unidentified and project leaders. However, it does not attain published status until it has been submitted to be revised and received a certain number of revisions. We define a revision as either a reading of the transcription to confirm its accuracy or, should the case be necessary, its correction. At the time of implementation, we established a system that requires three revisions before a transcription is validated and attains published status (and can no longer be modified). We based this decision on discussions with our project leaders. With further development this requirement can be rendered more flexible, with project leaders deciding on the quantity of revisions necessary before validation. This said, project leaders always have the possibility to unpublish a document, putting it back into circulation with other working transcriptions if they consider the document to be inaccurate or incomplete. All created XML documents are stored in the database system and XML files are written to the server. Transcribers have access to all transcriptions they have created or upon which they have intervened from their personal user account.

In Figure 7.5a, the transcription structure in the database is shown. It is composed of a user that owns the transcription, a content attribute to store the transcription itself in

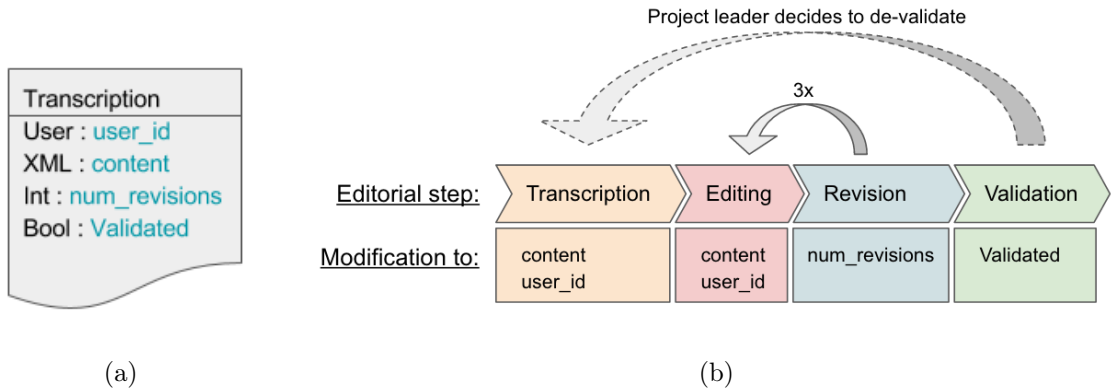


Figure 7.5 – Illustration of transcription work flow in PHuN 2.0. (a) Transcription representation in the database. (b) Editorial flow implemented within the platform.

XML format, an integer to store the number of revisions whose initial value is null, and finally a boolean to indicate whether or not the transcription has been validated. The editorial work flow leading up to validation and potentially the publication of a transcription is shown in Figure 7.5b. The rectangles below show which attributes of a transcription entity are modified at any given step. To begin, a transcription is created by a user and saved in the system. Then, other users can edit the transcription and modify its content. Once a user considers the transcription ready, he or she can send it into revision. During the revision process three different users must confirm that the transcription is accurate and or make improvements. At the end of revision the transcription is automatically validated by the system. At this point it is published and can no longer be modified by ordinary users. It appears in the project leaders' list of published transcriptions, but the process does not necessarily stop here. Project leaders can de-validate transcriptions and bring them back into the editorial circuit if they consider that they still need improvements.

Before opening a page for viewing, a user has access to basic information concerning the page, including its number, completion status, the name of the last person having worked on the page, and the latest intervention date. This basic information can be helpful to users deciding which page they will transcribe or review.

Once selected for transcription, pages can be opened in transcription mode wherein

users have access to an editor containing buttons corresponding to an XML vocabulary, as described in Section 7.3. Users transcribe according to instructions put together by project leaders. We tested several different protocols and more detail on this is given in Section 7.5.3. To save their work users have three options, which are listed below.

- A temporary save function, which allows users to return to their work if they close the browser or log out before completing a transcription, but does not write a file to the server.
- An official save and submit file, which writes a file and which allows other users to access and edit the submitted transcription.
- Checking the "Envoyer en relecture" (Submit for revision) and clicking the save and submit button signals that a transcription should be revised and sends it into the revision cycle. Subsequent interventions on transcriptions are considered as official revisions.

Pages sent into revision will appear in the revisions section and their status will be visible to browsing users. When users intervene on a page in review they contribute directly to the revision process leading to the publication of a page.

7.4.4 User comments and discussion

During the transcription process users can also comment on a particular page. The comment button is found at the top-right corner of the editor and clicking on it will open a discussion list for the page. The discussion section may be useful to users wishing to interact with other contributors, ask questions specific to the page they are working on, or about the transcription protocol more generally. This section was created as a way to further engage users in the transcription process and to provide outlets for connecting with other more knowledgeable transcribers. Similar ideas of creating forums and discussion lists have been widely implemented in Web 2.0 and more specifically by DH projects, including Manuscrits de Stendhal, Transcribe Bentham, Ancient Lives, TROVE et ARCHIVE [Moirez et al., 2013], but also for crowd science projects such as Polymath

[Franzoni and Sauermann, 2014]. It is true that many transcription and editorial projects use the data itself as a means of communication. For instance, by inserting *comment* tags, or using *illegible (illisible)*, *gap*, or *uncertain (douteux)* tags. Transcribers let each other know which words or passages may need particular attention. Project leaders have also made use of these meaningful elements in descriptive schemas used by the *Benoîte Groult* project and *La Réticence*. This encoded form of communication can be very effective, but it does not allow to generate discussion or occasions to socialise around common subjects of interest. Page discussion lists have the potential for being a complementary space for communicating with others about the work of transcribing, but also about the objects themselves. That is, they can be spaces for animating transcription activities through discussion about objects, authors and writing more generally. Finally, projects can also use this list to provide background information or details they think may be useful to others as a way to encourage participation and animate the community.

7.4.5 User profile

A series of other specific functionalities for users have been developed. These include a user profile space where users can choose and change their profile image, which is visible on the website. This space also gives the user access to all the transcriptions which they have contributed so that they can easily find and access these pages.

7.5 Project Leader functionalities

As we have already mentioned in Sections 7.3 on page 126 and 7.4.3 on page 130, project leaders create projects, configure transcription tools, and oversee transcription processes. We will describe the functionalities that allow them to do so in the following sections.

7.5.1 Project creation and corpus integration

Keeping in mind that some collections can be of variable sizes, from those composed of a few hundred pages to others, such as Stendhal or Bentham containing from thirty to over forty thousand pages, the task of integrating these materials into a working infrastructure needs to be planned correctly. This is all the more true when considering how to manage a service for project leaders who are located remotely and to whose collections we do not have direct access. When conceiving a platform infrastructure for the deposit of images into the platform database and server, large file sizes and voluminous collections, can rapidly become an issue.

We use Symfony's Data Fixtures component to handle uploading large volumes of images to the database. The process is relatively fast, allowing to upload several hundred images in a matter of a few short minutes. There are a few key rules to keep in mind for the code implemented to work correctly.

The image names must follow the following pattern: `w_x_y_z.ext` where *w* is the name of the collection, *x* its folder, *y* its sub-folder, *z* the page number and finally *ext* is the file extension. If more than three underscores are found in the path (that is the file name contains more than four units), we remove the first underscore iteratively until there are only three left. If there are less than three underscores, we create an unnamed sub-folder and possibly an unnamed folder. The system requires that there be at least one underscore¹¹.

Despite its rigidity, using our method of automatic data handling allows us to avoid uploading files manually and introducing errors into database records.

11. The pattern we adopted for image names is quite strict. However, project owners may want to have a more flexible hierarchy. One possible solution would be to allow projects to define their own container hierarchy and map it to each unit (a number or series of numbers separated by an underscore) found in image names.

7.5.2 Project configuration and management

Project leaders can monitor their projects from an administrator menu. The current project administrator menu has the following components or views:

- **Editor configuration:**

Project administrators can modify an editor's configuration here by defining element names, deciding on their placement in the editor, and adding descriptions of their functions to guide transcribers.

- **Editor CSS styles:**

From here project administrators can make adjustments to the CSS stylesheet they uploaded at the time of project creation and which controls the visual presentation of editor elements. To clarify, the editor allows to encode textual content as XML and the interface is handled by an associated CSS.

- **Project Description:**

Project leaders can edit the descriptions used to present their projects to the public.

- **Institutional logos:**

Project leaders can upload logos of partnered and participating institutions.

- **Contributor list:**

Project leaders have access to a list of users having contributed or intervened on one or more transcriptions for a specific project. They can access this list to promote other transcribers to project admin status (to help manage revision and validation procedures).

- **Published Transcriptions:**

This view displays all published transcriptions for a project. Project leaders can consult transcriptions from this list and devalidate or unpublish transcriptions considered incomplete.

Besides managing these aspects of projects, project leaders can also be involved in the transcription process itself by transcribing and revising transcriptions from less experienced contributors. Finally, having access to all the same functionalities as identified

users, they can start discussions for specific pages and use the commentary list as a space to share advice with other less experienced contributors.

7.5.3 Transcription protocols

Transcription protocols are instructions project leaders create for transcribers. Transcription protocols are a necessary support for unexperienced contributors and ensure a certain degree of uniformity in the results obtained from tasks performed by many different individuals. In general, providing transcription instructions is an important part of project management as participating transcribers appreciate having access to resources that outline expectations and detail how a task is to be carried out. A protocol should clearly explain the nature of the task and outline the implicated steps, being careful to address ambiguities, but also leaving out extraneous information that may demotivate inexperienced transcribers. If a certain degree of interpretation is expected of the work, then the user should be made aware of this, so as to minimize confusion or hesitation, which can inadvertently modify behaviour and lead to unwanted results.

When managing projects, protocols are common and advised, when handling scientific experiments they are absolutely necessary. In the course of this doctoral project we were necessarily exposed to both types of situations. Project leaders developed protocols to guide the work of their contributors, with the goal of furnishing the most clear instructions and obtaining the highest quality transcriptions possible. At the same time, experiments were organized to test the usability of the platform as well as the quality of obtainable results from users, here too well-articulated and clear instructions were necessary to help users better understand the work they were performing.

In many ways the first protocols were based on the documents and manuals from long-lived projects like *Les Manuscrits de Stendhal*¹², which provides a detailed reference manual for the XML vocabulary used by this project. The functions and uses of each XML element are explained, which also helps disambiguate certain elements that may appear

12. http://stendhal.msh-alpes.fr/wordpress/?page_id=91

to serve similar purposes (*ajout*¹³ and *ajout en interligne*¹⁴ *inline addition* for instance), and clarify others whose titles do not provide sufficient indication as to their intended purpose (*surcharge*¹⁴ and *codé*¹⁵ for example). The original integral manual used for the Stendhal project is a lengthy document that would be challenging to integrate into a crowdsourcing project, let alone expect casual users to read, which directly undermines its usefulness to the inexperienced contributors that it is meant to benefit most.

When the PHuN platform was subjected to its first round of user tests in the spring of 2016 for the Michel Butor manuscripts, a *clé-en-mains* document of instructions was elaborated by the scholar, Cécile Meynard, leading the experiments with her group of Master students. This short document of approximately five pages was given to students to quickly read through before opening a work session that lasted approximately two hours. It summarized the functions of available XML elements, or buttons in the WYSIWYG editor. Concurrently, it was meant to walk students through the process of using the new platform and its editor, since this type of transcription interface was entirely new to most, if not all, participants. At the time of this first experiment, this was still a rather lengthy document, and the information within could be effectively condensed further to make it more easily accessible to participants from disciplines other than literature and the humanities.

We have seen from examples found in related literature that communication with volunteers and volunteer training can be organized in a number of different ways. However, taking care to design instructions for volunteers is of crucial importance to a project's success as it may affect results Cohn [2008] ; Wiggins and Crowston [2011]. Researchers have established a connection between task phrasing and the kinds of results obtained [Brown and Allison, 2014]. Because of this existing link, instruction sets should be tested to determine if they yield reliable data when executed by participants; it is important to ensure that tasks are not too complex [Cohn, 2008].

In our case, instructions were prepared with the goal of explaining how transcribers

13. addition

14. A word or letter corrected directly by writing over top of it.

15. A coded word used by the author to signify another word.

were expected to use the editor and which types of features in the manuscript were relevant to annotate. For simplicity and concision the transcription process was presented as a sequence of steps and explanations were accompanied by supporting images. In the tutorial initially created on the PHuN 2.0 platform for the Benoîte Groult corpus we used the principle of an online powerpoint presentation. The slides are included in Annex A.2.

In fact it would be interesting to know how many projects actually measure the effects of their communication on their publics. This could be an interesting area of inquiry for those seeking to further develop the study of quality in participative digital humanities projects. Furthermore, it stresses the necessity to evaluate the quality of documents obtained. In Chapter 6 we presented methods that can be used by projects to evaluate the contributions they collect. The potential of these methods will be described in more detail in Chapter 11.

7.6 Discussion on limits and improvements

Improvements to the platform were implemented in an ongoing manner in response to issues related by users. A few of the more pertinent ones (related to user accounts, project administration, transcription validation, and the editor) are related in this section.

Transcription editor

The addition of configurable button descriptions at the place of tooltips to help users understand button function was an important improvement that had the benefit of simplifying transcription instructions. With element descriptions in place users have quicker access to a reference manual. An improvement on paper versions or even digital reference documents. These explanations should not be considered as extensive; their purpose is to give inexperienced users keys to understanding the purposes of element terms that may not be familiar to them. An improvement on this functionality would be to create a more complete frequently asked questions (FAQ) page where project leaders address commonly asked questions about their manuscripts and explain how to use the provided editor to

encode manuscript features.

We added a list of custom elements in the editor configuration file, which improved the way TinyMCE handles projects' custom descriptive elements. This solved a number of page formatting errors that were problematic in earlier versions of the platform.

User accounts

The user profile is a landing page from which users can access their transcriptions. They can also see when others have intervened on their transcriptions so as to keep them informed of the editorial process. Users can also choose to upload an image to represent them on the site.

Arguably, some important elements are missing from this space. Adding functionalities to allow users to track their own activities or auto-evaluate their progress would constitute positive improvements for this type of user space.

Project administration

Project leaders can promote users to their own level, but at this time there is no way for them to manage user groups. A useful functionality would be to include the possibility to create groups, invite users, and manage the visibility of the activities of these groups.

Creating channels for feedback would be beneficial to participants. Therefore, establishing some way for project leaders to be able to contact participants directly may be appropriate. Transcribe Bentham project staff provide very detailed feedback to participants, which is appreciated by volunteers [Dunn and Hedges, 2012].

Intellectual validation

The validation process begins with a transcription being sent into revision. From revision to validation the system requires that three different individuals either confirm that a transcription is accurate to the best of their knowledge or edit it and save the changes. As a final security measure project administrators can consult published pages and devalidate them if necessary. The current cycle imposes a number of limits on projects, which may benefit from greater flexibility.

Firstly, the revision process is secured by requiring revisions from three different individuals. The system does not require that these individuals be experienced or trusted members of the project, just that they not be the same person. A possible improvement could be to allow project leaders to decide on the type of hierarchy they want when they configure their project. With greater hierarchy, revision would be a task specifically managed by project administrators, which would include reading, editing, and correcting transcriptions before validation. While with lesser hierarchy, any one can revise transcriptions. Of course, this option is greatly contingent on the number of persons involved in the project. With a less hierarchical revision process, we may gain more participants, although these may indeed be less experienced.

Another improvement could be to allow for project leaders to decide on the number of revisions necessary before transcriptions are considered complete. Projects based on easier-to-decipher documents or a pared-down XML schema may not need three revisions as two or even one may be enough. If this can be decided and configured like the editor itself, the validation process may better reflect individual projects and their editorial needs.

Likewise, some useful feedback loops can be put in place to better accompany the revision process and ensure that transcriptions are revised in a thoughtful and conscientious manner.

Technical validation

TinyMCE does not handle XML DTD validation in the way that specialized XML editors like Oxygen do. XML editors can rely on a DTD to dictate which elements are allowed within which other elements. Although TinyMCE allows to define custom allowed elements within a document structure based on terms taken from a DTD, it does not control hierarchies based on this DTD. Arguably, TinyMCE is more flexible because it was originally intended for HTML and the web. For instance, if *addition* or *deletion* elements are allowed inside a paragraph, these elements themselves are also allowed *additions* or *deletions* as children. For documents that have deletions within additions this is quite acceptable and convenient— the contrary would be too restrictive. However, this does

mean that TinyMCE does not guide users through the document hierarchy while they transcribe, and it does not point out errors as an editor like Oxygen would. This also means that if users are not careful, they may place multiple elements of the same type side-by-side or stacked within one another. In worst case scenarios, they may delete an opening or closing tag, leaving its intended match on its own. Though, in this case, TinyMCE recognizes and deletes pairless elements and the harm is that the particular feature of the manuscript is not encoded. Also, when extra pairs of empty elements are present in the document, they generally do not affect content, besides producing extra spaces or lines in the document. These can be filtered (or cleaned up) in an additional post-processing step.

7.7 Conclusions and next steps

The issues described and improvements proposed in this section can serve as a basis for future requirement specifications as part of ongoing improvements to PHuN 2.0's work environment.

In making the production-driven platform prototype we were more concerned with creating a work environment that had functionalities and features that are comparable to existing digital work environments. The questions this prototyping process raised were indeed relevant and extensive, as they concerned challenges associated with handling encoded data, managing work flows, making customizable tools, and also creating interfaces for users. For us, what was missing was the actual experimental and analytical component with regards to what kinds of data crowdsourced users produced. We created a platform to experiment with crowdsourcing transcriptions and we still had no way of evaluating how these crowdsourced transcriptions compared to those of specialists or trained contractors. To resolve this issue it was necessary to create a second prototype, a trimmer version of the original PHuN 2.0 platform, which we called PHuN-ET (Plateforme des Humanités Numériques - Espace Transcription). This second prototype is the subject of the next chapter. In it, we will present the experimental prototype's functionalities and

discuss how these allowed us to achieve our goal of collecting experimental data. At the same time, we used the opportunity that comes with working on a new prototype to gain new knowledge about computing and also about our users' experiences of the platform.

Chapter 8

Presentation of PHuN-ET

Contents

| | | |
|------------|---|------------|
| 8.1 | Chapter Summary | 143 |
| 8.2 | Introduction | 145 |
| 8.3 | Premise for the PHuN-ET platform | 146 |
| 8.3.1 | Focus on experimental research | 147 |
| 8.4 | Editor functionalities | 149 |
| 8.5 | Identified User functionalities | 150 |
| 8.5.1 | User accounts | 150 |
| 8.5.2 | Sequential access to pages | 150 |
| 8.5.3 | Transcription instructions | 151 |
| 8.5.4 | Transcription interface | 152 |
| 8.5.5 | Data visualisation and sharing | 152 |
| 8.6 | Conclusion | 153 |

8.1 Chapter Summary

In this chapter we present PHuN-ET, the experimental platform we created for the purposes of collecting crowdsourcing manuscript transcriptions for comparative quality

analysis. We describe the platform's functionalities in relation to users and to data collection goals.

8.2 Introduction

Initially, PHuN 2.0 was intended to produce transcription data that would constitute the basis of this doctoral thesis. As the project developed, and as we continued to learn from our prototyping efforts, the working environment of our users itself required more in-depth consideration. We made a number of adjustments to meet requirements of both simple users and researchers for whom the platform was being developed. Still, after long months of development we saw that further focus on improvements would not guarantee having experimental data to work with. The production platform had to be set aside in order to address our initial research questions. That is, the questions focusing on quality evaluation of crowdsourced transcriptions.

PHuN-ET was developed on the foundation of the already existing functioning model of PHuN 2.0 in response to the need to address specific research questions. This prototype thus integrates specific modules for comparing, graphing and visualizing recovered data. It is intended to serve first and foremost as a tool for collecting experimental data.

This new version is essentially a copy of the older prototype, minus some of the functionalities of the original, but one that incorporates a series of data analysis modules and interfaces intended for the exploration of the original research questions at the basis of this work. PHuN-ET is thus an experimental prototype developed in parallel to the original PHuN 2.0 production-driven prototype, created for purely research-oriented objectives. The existing prototype architecture made it easy to duplicate and the two platforms can coexist without infringing on each other's functions. Throughout this chapter, we will take care to indicate the similarities and differences between the two prototypes.

PHuN-ET distinguishes itself from PHuN 2.0 in its approach for collecting transcriptions. The approach is based on a crowdsourcing model, which accumulates contributions from multiple users. This *contributive* model differs from the *collaborative* model chosen by the research team for whom PHuN 2.0 was built. The collaborative model did not allow for multiple contributions from multiple users, but opted rather that once a transcription was created subsequent users intervened on the same document so that the document was

constituted collaboratively by multiple individuals and would thus be credited collaboratively. This approach ensures an economy of users' efforts, since the efforts of numerous users aren't used to constitute the same— and thus potentially competing— transcription. However, since our experimental intentions required comparing multiple contributions to obtain a maximum of information about the contributions we could expect from the crowdsourcing method, we needed an environment that would allow us to collect multiple transcriptions for each page or manuscript object. Our experiment-oriented prototype offered this solution.

8.3 Premise for the PHuN-ET platform

Our primary need was to maintain access to pages to as many users as possible and obtain transcriptions that are produced in one uninterrupted sitting. The first condition more closely resembles our intended crowdsourcing conditions. The second helps limit effects of variability that are tied to changes in concentration, fatigue, or changing environments that can accompany working on a transcription over several sessions. It is also easier to evaluate experimental data if we impose that each contributed transcription is done from beginning to end in one sitting.

Our secondary purpose for altering the original platform was to put in place simpler navigation, which more closely resembles existing crowdsourcing project like Zooniverse.

Our goal was to maximally reduce the number of distracting steps between user registration and transcription tasks. By adapting the site's architecture, we give priority to transcription protocol, the transcription task itself, and also user accounts, where users can revisit their transcriptions and also track their progress. The revision and validation process is replaced by an analysis of all contributions.

Changing the site's architecture has also led to replacing the browsing interface, which allows users to see whole collections and select pages, but this component is easily put back in place once experimental objectives are met. Figure 8.1 represents the PHuN-ET's architecture as a sequential diagram showing the order in which users access each page

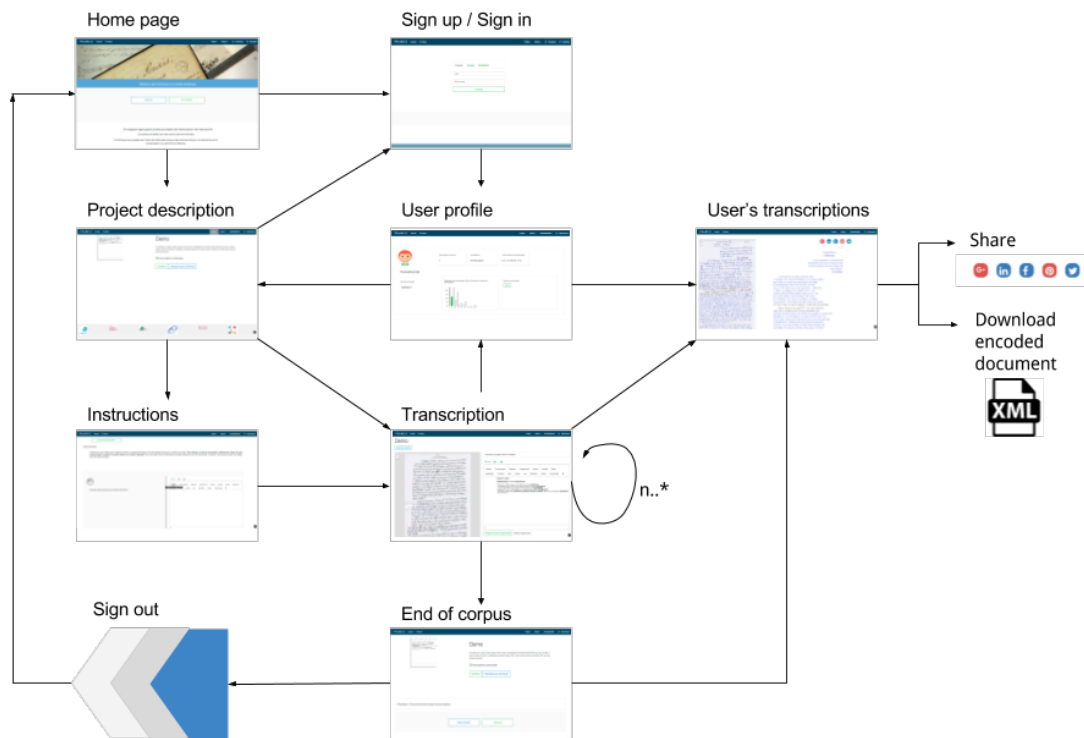


Figure 8.1 – PHuN-ET's site work flow.

when participating in transcription activities.

8.3.1 Focus on experimental research

Even with the production environment in place, we still had to address our questions regarding crowdsourcing. For this, new experiments needed to be done and we wanted to adapt the architecture and user interfaces to facilitate this. Certain functionalities created for the production platform were potential sources of problems. Notably, the catalogue for browsing and selecting pages narrowed the likelihood that multiple users transcribe the same page. This is indeed what was observed in our first experiment on Benoîte Groult's corpus. In this experiment, data was recovered from the production platform and although over a hundred documents were collected only 2 were transcribed by multiple users (5 users), which we had to accept despite this low participation count.

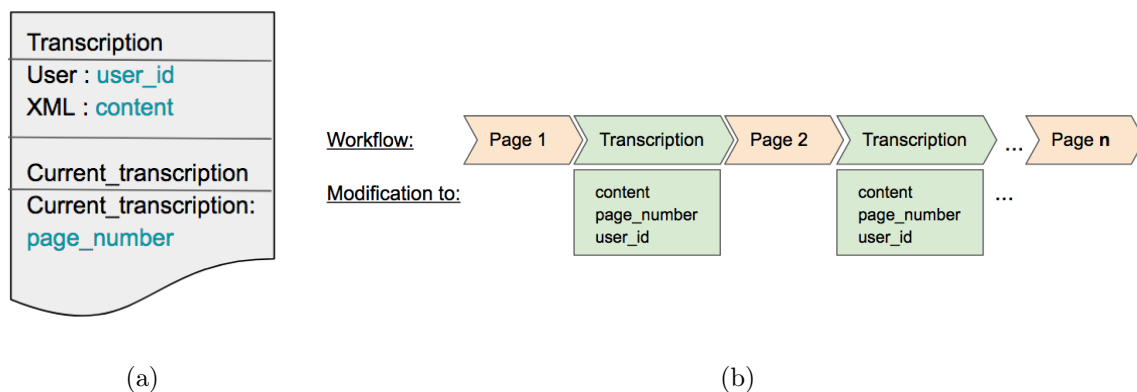


Figure 8.2 – Illustration of transcription workflow in PHuN-ET. (a) Transcription representation in the database and pointer to user' current transcription (*current_transcription*). (b) Users' transcription flow from page 1 to page n of a corpus.

We recount this experiment in Chapter 11 Section 11.3.1. The decision to replace the page catalogue, which certainly has its advantages in a production environment, was to ensure adequate participation on monitored pages.

Replacing the catalogue also simplified navigation for users who were expected to create accounts before starting on a transcription. With navigation to a minimum, we were more certain that users would more quickly land on the transcription interface and not get distracted or lose interest in the activity before starting.

We also wanted to ensure that once a user submitted a transcription it could not be edited or modified by other users. In a production setting subsequent interventions are likely to produce positive improvements, but we were interested in seeing what transcribers accomplished in one sitting, without subsequent editing. Figure 8.2 shows the process as it is provided for in PHuN-ET: users transcribe a given page before moving on to the next. With each intervention on a page, we collect a new transcription for that page. This was a way to limit the introduction of other unknowns into the process. Crowdsourcing already having a significant number of these, we were not concerned with preventing users from using the web as a resource, or interacting with others, but simply to observe the results produced by users in crowdsourcing conditions. No editing also meant no need for a scientific validation circuit. We wanted to observe the transcriptions one can obtain

before expert intervention. We would then be able to hypothesize how that compares to expert-made transcriptions and perhaps how much effort would be required to correct work produced by crowdsourced transcribers. We needed quantifiable data to be able to evaluate if crowdsourcing is an appropriate method for manuscript transcription and the PHuN-ET platform was the solution to acquiring this data.

Finally, components put in place for the purposes of project management, which concerned specifically project administrators in the production platform, were maintained in PHuN-ET. These made it possible to perform all the necessary steps in setting up a project, including configuring the editor and controlling CSS presentation of editor elements. During experimentation we did not make use of the other functionalities initially created for the production platform, since managing contributors and revising published transcriptions were not our primary goals in this case.

8.4 Editor functionalities

The editor used in our experiments included the same functionalities as those described previously in 7.3. For Experiment N°2 on Benoîte Groult we incorporated the XML vocabulary and CSS rules introduced by our experts. In Experiment N°3 we introduced our own XML descriptive schema that focuses on modifications¹ in the pages and a number of other visible features².

The possibility to edit raw code was removed to ensure that transcriptions were not copy-pasted from other sources. We also added a button that allowed users to remove XML elements, without having to go into raw code view (raw code view was included in PHuN 2.0). The "Remove Formatting" button exists among TinyMCE's set of default plugins and only requires being called in the editor configuration in order to be used. We made use of this functionality to improve users' experience of the transcription editor,

1. This include additions, deletions, and corrections.

2. These include mainly citations, abbreviations, names, places, chapter titles, paginations, doubtful, and illegible elements.

allowing them to undo or change annotations without losing what was already typed in the editor.

8.5 Identified User functionalities

This section retains essential functionalities from the original PHuN 2.0 platform, with user registration and associated profile that gives access to all the user's transcriptions. The main difference between the two is the visibility of the user and his or her work to other users. PHuN 2.0 set out to create a communal environment where transcriptions contributed by users can be edited and improved by others and the results are viewable by all and the efforts are credited to each participating user. PHuN-ET is a more private environment, because users create transcriptions to which only they themselves have access from their user accounts. To compensate for this, and to provide users with the possibility of sharing their work, albeit in a more discrete manner, another sharing functionality was added. From the user's own account he or she may choose any transcription to be shared to four major social networks (LinkedIn, Facebook, Twitter and Google+, depending on user preference).

8.5.1 User accounts

User accounts basically function in the same manner as for PHuN 2.0. Once enrolled, users have access to all projects listed on the platform. Since the two platforms are independent from one another, users previously enrolled on PHuN 2.0 need to create another account on PHuN-ET to participate in transcription activities on the experimental platform.

8.5.2 Sequential access to pages

In PHuN-ET we replaced PHuN 2.0's project catalogue with sequential access to pages. What this means is simply that users access pages in the same order as they appear in the

database. For a given project, the first page in the database will be the first to be proposed to transcribers, followed by the second, and so on until the last listed database record. To keep track of users' transcriptions, we created a table that handles users' transcription indices. Each time a user is enrolled, and for each existing project, the system creates a record initializing the user on the first pages for each existing project. Each time a user submits a transcription, their index is increased by one and they are given access to the following page.

We were inspired by Zooniverse and similar crowdsourcing platforms, where users are proposed data objects one by one. In the same way, we decided to make it possible for users to skip pages that didn't interest them, by clicking a "Passer à la page suivante" (Next) button. Clicking this button has the simple effect of increasing the user's transcription index and thus bringing up a page corresponding to the following index.

The decision to use this sequential transcription flow had the advantage of simplifying navigation within the platform. Moreover, it turned out to be an effective way for us to acquire the data we wanted more quickly.

8.5.3 Transcription instructions

Transcription protocol or instructions is a vital aspect of the platform as it may be the only way to transmit the necessary how-to instructions to remote users. In creating the instructions our primary concern was concision and clarity. The transcription process was broken down into its main elementary steps and listed in order. The text is accompanied by a short video sequence showing the main transcription steps from beginning to end. The main aspects users should understand is the order in which operations should be performed, notably the text should be typed first and then selected with a cursor before clicking on the available buttons in the editor.

A secondary user aid is included in transcription editor itself, which lists some useful keyboard shortcuts. These include known shortcuts to undo actions (Cmd+Z or Cntrl+Z) and their redo counterparts, as well as a few others to help users have a better handle on

the transcription editor.

8.5.4 Transcription interface

The transcription interface consists of two main panels, one containing a zoomable manuscript image or facsimile and the other a transcription editor. The user can choose to switch the placement of these two panels with the help of a button. Doing so will exchange the panels moving the editor from right to left or vice-versa, depending on user preferences. For some users this adaptation may be of very little consequence, but we chose to include it to improve the flexibility and usability of the interface.

8.5.5 Data visualisation and sharing

Since the very first versions of PHuN 2.0 we considered it important that users have access to the result of their work. This is why access to one's transcriptions from one's user profile was put in place and kept. As a result of discussions and input from users we implemented several improvements that concern user's visualisation of their transcriptions. These improvements are listed below.

- **Data retrieval:**

Transcriptions can be downloaded in XML format, reflecting the project's chosen XML vocabulary. A second button also allows to simply visualise the transcription as an XML tree in a separate window of the browser.

- **Data sharing:**

Users can share links to their transcriptions in the same manner as web content is shared on social networks. For this purpose, we implemented five social network buttons, including Facebook, Google+, LinkedIn, Pinterest, and Twitter.

Allowing users to recover files they create on the platform opens up new opportunities for digital scholarship and creation. It makes it possible for example for users to constitute their own collections or corpora, which may be used in editorial processes or as educational

resources. Allowing users to download data or share it was also a way to concretize their activity on the platform, helping many of our participants gain a better understanding of what they were doing, as well as what could be done, with their contributions. It was also our way of disclosing intermediate inputs [Franzoni and Sauermann, 2014].

In an auxiliary way, these improvements were important for experimentation because we initially had very few participants. We hoped that if some users shared their work on social media they would attract others to participate.

Likewise, more direct access to social networks maximizes on the platform's potential to be incorporated into users' more habitual Web 2.0 practices. This may also be another way to include transcription– and reading transcriptions– into existing "communications circuits", which according to many scholars can help bring readers out of isolation [Fitzpatrick, 2011]. We can imagine that a linked transcription can accompany a post on any number of social networks and include the poster's reaction to the text, commentary, ideas, and a call to others to discuss both the text and the poster's reading of it.

8.6 Conclusion

We implemented these changes separately from the production platform for experimental purposes as discussed in this chapter. A number of these functionalities can be introduced into the production platform with minimal effort. Table 8.1 summarizes the primary differences between the two platforms. Each of the functionalities listed for PHuN-ET can be reintegrated into PHuN 2.0 if seen fit. Symfony's MVC framework makes this relatively simple, as discussed in Chapter 7, Section 7.2 on page 121. Mainly, it would require adjustments to the Controller and in some cases to the Model (or database), whereas the views can simply be reused. More specifically, improvements to the user space should be considered in production, as user functionalities can produce observable effects on user motivation and implication.

As already mentioned, PHuN-ET allowed us to simplify navigation and process flow for users of the platform. Other architectural modifications can be studied to determine

which are most effective in a production environment. These aspects may constitute rich subjects of study in the fields of information architecture and human-computer interaction.

In this chapter we have described the functionalities introduced in response to experimental needs. Drawing on Ruecker's idea about experimental prototypes we built one that was used to generate both "generalized knowledge about an idea" [Ruecker, 2015, p. 2] and data on which we base the rest of our analysis in this work. We will look more closely at the data we collected with this platform in Chapters 11 and 12. In the following chapter we consider how human beings are implicated in crowdsourcing initiatives. We discuss how to support the various aspects of collaboration that extend beyond digital environments and require human implication, including motivations, communication, and competences.

| Functionalities | PHuN 2.0 | PHuN-ET |
|---------------------------------------|--|--|
| User Accounts | Date back to initial stages of development: Access to users' transcriptions. | Includes improvements: Access to transcriptions; tracks user progress; simple user ranking; average transcription time |
| Transcription Instructions | Yes | Yes |
| Page browsing / selection | Yes | No. Access to pages is predetermined. Work flow is organized to lead users more directly to the transcription interface. |
| Editorial / validation process | Yes. See Figure 7.5 and Section 7.4.3 on page 130 | No. Multiple transcriptions are collected for each page. See Figure 8.2 |
| Discussion list | Yes. Includes a discussion list for each page. | No. Suspended for experimental purposes. |
| XML download / sharing | No | Yes |

Table 8.1 – Résumé of differences between platforms PHuN 2.0 and PHuN-ET.

Chapter 9

Beyond the Platform- Human considerations

Contents

| | | |
|------------|---|------------|
| 9.1 | Chapter Summary | 157 |
| 9.2 | Collaboration | 159 |
| 9.3 | Motivations | 162 |
| 9.4 | Communication and Outreach | 166 |
| 9.5 | Skills and Training | 169 |
| 9.5.1 | Training volunteers | 169 |
| 9.5.2 | Online instruction | 170 |
| 9.6 | Chapter Summary and Conclusion | 172 |

9.1 Chapter Summary

In this chapter we take a step away from the technical environment that makes up the platforms to discuss how human beings are implicated in crowdsourcing initiatives. We discuss collaboration in collectives, motivation, and the role of communication in getting projects necessary exposure to publics and in order to constitute and maintain virtual

communities. We also discuss the importance of skills and training in such environments.

Virtual environments play a crucial role in the organisation of digital materials, work flows, and contributions. With online tools project leaders can take fuller advantage of the potential of open participation. Unmistakeably, digital and web technologies play an important role in defining human work practices. Still, human beings have a considerable amount of influence on the goings on of projects, which extends far beyond simply defining and performing tasks using tools and virtual environments. We should also look at how social actions and influences of both individuals and teams can shape successful initiatives of crowdsourcing, crowd science, and also citizen scholarly editing projects.

We will consider aspects belonging to three main themes: collaboration, communication and outreach, and finally, skills and training. This will allow us to address questions that concern individuals and project teams directly, so as not to forget who is really behind the technologies that make Digital Humanities projects possible.

9.2 Collaboration

As already mentioned, investing in collaborative software is not enough to put in place effective and long-lasting practices of collaboration. Yet collaboration is an important component of crowdsourcing projects. This has been shown by scientifically-inclined projects such as Fold It, and others for which collaboration among participants has proven to be vital to finding solutions to complex intellectual challenges [Franzoni and Sauer-
mann, 2014]. It is further supported by [Prestopnik and Crowston, 2011], who consider citizen science projects as "as a form of social-computational system".

Yet it is also important to go beyond considerations of collaboration at the contributor level, it is also vital to create the appropriate conditions for collaboration between all individuals and entities having a stake in the project. These can include researchers, developers, partnered financing institutions, and archive collections. There are some specific challenges to making this happen, particularly in a Digital Humanities context.

Traditionally and historically, humanities scholars and researchers work alone and do not engage in expansive collaboration with other scholars, much less on an interdisciplinary

level [Siemens et al., 2011]. This of course is something that is being encouraged to change on a disciplinary level; for the Digital Humanities in particular, interdisciplinary collaboration has been worked into the very tissue of the discipline. This is not to say that the process to making collaboration in DH work is simple, it is not, but increasingly publicized work on collaboration in a large number of disciplines makes information on functional examples more accessible to all disciplines, and this should be taken advantage of. Moreover, a number of successful digital scholarly editing projects advocate for greater emphasis on collaborative work methods, which is encouraging [Siemens et al., 2011 ; Leriche and Meynard, 2008 ; Causer and Terras, 2014].

There is some discussion on distinguishing terms that are often used interchangeably in literature and practice focusing on crowdsourcing. For instance, *collaboration*, which means working with one or more individuals¹ and *contribution*, meaning to give support for a common purpose² Several studies actually looked at the crowdsourcing phenomenon by examining social actions as belonging to I-mode or we-mode collective intentions, wherein I-mode is seen as personal and independent intention and we-mode is group-oriented interdependent intention [Bagozzi, 2000 ; Shen et al., 2014]. The [Shen et al., 2014] study on Wikipedia participants collects empirical evidence to support that both I-mode and we-mode intentions impact contributive behaviour and that the main difference is with respects to relational factors of trust and commitment, which appear to impact we-mode intentions only. While contribution is possible both in I-mode and we-mode, this study leads us to suggest that looking at relational factors would be important to address the differences between collaboration and contribution. For instance, while both collaboration and contribution can be spontaneous and short-lived, many projects are interested in long-term commitment from participants. Thus, emphasizing relational factors like commitment and trust when looking at participative models may be effective in placing intended focus more accurately and more transparently on what aspects are

1. From Latin *com-* together and *labōrāre* to work, according to <https://www.collinsdictionary.com/dictionary/english/collaborate>.

2. From Latin, *contribuere* to collect, from *tribuere* to grant or to bestow, according to <https://www.collinsdictionary.com/dictionary/english/contribute>.

desirable in crowdsourcing projects, particularly when there is a high tendency to use the terms *collaborate* and *contribute* interchangeably.

We can say, for instance, that lasting and effective relationships are sought, or, on the contrary, that one is looking to organize punctual and spontaneous efforts.

Even in cases where collaboration is not rooted in virtual environments, it is important to recall that people are of foremost importance as agents of collective practices. This highlights the importance of interpersonal skills, organisation and flexibility, and to some degree, creativity and imagination as well [Siemens, 2012]. Collaborations can also extend to include those between institutions and organisations— and once again there are always motivated human actors who make collaboration possible— thus sharing knowledge and expanding networks of common practices [Siemens, 2012].

Let us consider how individuals can form groups to meet common goals, like making immense tasks such as the digitization and transcription of 40,000 manuscripts more feasible. Collaboration can also include the matching of disciplines and individuals with diverse and complementary skills, thus optimising productivity and increasing the likelihood of finding creative and appropriate solutions to complex problems [Surowiecki, 2005]. Finally, it seems that an advantage of effective collaboration is, not surprisingly, more collaboration, which supports the idea that collaborative practices can take root and become the norm [Siemens, 2012].

For collaboration between project leaders and the public, or project leaders that support and oversee collaboration between members of the public, creating the environment itself is not the only step involved. To develop on this, Dunn and Hedges [2012] cite Trevor Owens, as will we:

Most successful crowdsourcing projects are not about large anonymous masses of people. They are not about crowds. They are about inviting participation from interested and engaged members of the public. These projects can continue a long standing tradition of volunteerism and involvement of citizens in

the creation and continued development of public goods³.

From this, we understand that involvement goes beyond achieving greater visibility from large anonymous groups of people online. The real achievement of collaboration for crowdsourcing between projects and crowds is to attract specific groups of contributors that will ultimately translate into long-lasting community involvement. This is far from a random process and requires a significant amount of planning, thought and coordination. In the next section we will look at different aspects of motivation and how these can affect user participation in collective efforts.

9.3 Motivations

Individuals' intentions to participate in collective actions are thought to be regulated by three main factors: cognitive, motivational, and social-relational [Cho et al., 2010]. We will look mainly at motivation, and also identify where cognitive and social-relational factors impact on individual motivation. To recall, a key challenge for crowdsourcing projects is attracting interested users [Shen et al., 2014].

Motivation is a factor of involvement that has been studied by scholars interested in the complex social mechanisms that animate crowdsourcing projects [Franzoni and Sauermann, 2014], but not only. It is also a vital ingredient identified in the behaviours and attitudes of successful students or entrepreneurs [Ryan and Deci, 2000]. The questions behind what motivates people to take part in certain activities and overlook others are indeed a complex set, deeply grounded in human psychology. We will take a look at how these aspects of human psychology play a role in crowdsourcing environments.

Franzoni and Sauermann [2014] consider the problem of motivation first and foremost from an economics perspective. Based on this, we may be tempted to ask where one would find "contributors who are willing to exert effort without pay, potentially allowing projects to take advantage of human resources at lower financial cost than would be

3. <http://www.trevorowens.org/2012/05/the-crowd-andthe-library>

required in traditional science" [Franzoni and Sauermann, 2014]. They suggest that what contributors get in return for their involvement replaces basic financial compensation, by a series of benefits or "pecuniary pay-offs" that are coveted by those participants. And while most projects cannot propose monetary compensation to participants, this is the first extrinsically motivating factor that people tend to think of. Therefore, projects must find other forms of rewards. These rewards often take the form of social status, networking, and crediting [Dunn and Hedges, 2012].

To connect human psychology to economics, these benefits we speak of act as extrinsic motivators, a type of motivation that has long been studied in relation and opposition to intrinsic motivation. What many scholars of psychology conclude is that extrinsic motivation is not nearly as effective as intrinsic motivation, that it is behaviour that positions the individual within a social construct where he or she is the subject of social demands, or an actor in the process of acquiring goods of instrumental value [Ryan and Deci, 2000]. If the actor stands to lose something from not accomplishing a task, whether it be social status or economic value, we can see how extrinsic motivation has the potential to become a negative force on an individual [Ryan and Deci, 2000].

There is no danger of this happening when an individual is intrinsically motivated because they only stand to gain in enjoyment and personal satisfaction. Intrinsically motivated individuals perform activities because they enjoy them or because they feel challenged and they derive a sense of satisfaction upon completion of a task [Franzoni and Sauermann, 2014 ; Ryan and Deci, 2000]. This makes intrinsically motivated individuals whose focus falls within the field of activity proposed by a particular crowdsourcing project the optimal scenario, as it is one in which everyone involved stands to benefit from the exchange. Furthermore, studies show that intrinsically motivated people are more likely to succeed in the field that motivates them and that in an academic setting for example this translates into better grades and better quality work from students [Ryan and Deci, 2000]. Once again, this kind of involvement can have great advantages for crowdsourcing.

There are also factors according to [Ryan and Deci, 2000] that can have an impact on intrinsic motivation, which can either enhance or hamper individual attitudes and

behaviours. Positive feedback plays an important role in helping individuals maintain intrinsic motivation and increase it. Some examples of this type of positive feedback are observed in many project environments, either as part of the virtual framework that compensates invested participants with points, rewards, or status within the community. Some examples for this in citizen science projects include, once again, Fold It, which with its gaming environment succeeds in making challenging and difficult tasks intrinsically motivating and rewarding [Franzoni and Sauermann, 2014 ; Prestopnik and Crowston, 2011].

There are some rewards users receive in virtual environments that may at first appear to function as part of an extrinsic motivation pattern, but this is not necessarily the case. Firstly, because a user who becomes involved in a project rarely does so for the simple joy of receiving gold stars or points. Secondly, because these forms of recognition exist only in a virtual environment and have no actual impact on the social environment of the user [Ryan and Deci, 2000]. To participants who are intrinsically motivated, these rewards actually play the role of positive feedback and help maintain their motivation.

Another form of positive feedback that enhances intrinsic motivation is the individual's consciousness of autonomy or freedom while, or as a result of, engaging in an activity [Ryan and Deci, 2000]. This can translate into something like seeing one's skills improving over time and thus gaining more autonomy, which is gratifying. In a more general way, crowdsourcing frameworks should try to install a balance between structure and liberty. Individuals should have a high degree of liberty in the tasks they undertake, the degree to which they contribute, and to which they interact with others, as well as the amount of time that they contribute to these activities. Although some projects do manage to operate with a certain level of control on the degree to which individuals contribute. For instance, in the case of Marine Lives, the project requires that participants commit to working for three hours a week, for fourteen weeks [Dunn and Hedges, 2012]. This has the potential of infringing on participants' sense of freedom, and thus, directly impacting their motivation. Nevertheless, this system appears to work, as project managers reciprocate by taking direct responsibility for bolstering the participants' motivation over the course

of their engagement [Dunn and Hedges, 2012].

From a scientific and organizational standpoint, [Dunn and Hedges, 2012], describes projects based on the content type or "asset type", the "task type", the "process type" and finally the "output type". These categories can help describe a project objectively based on the type of content it proposes, what is done with it by volunteers, what is then done with the work of volunteers at the project level, and finally what products are derived from the activity at the end of the project. Having analysed the project in this way, it may be possible to gain better understanding where its weaknesses may be in terms of gaining contributions. It may be reasonable to suggest that if a project is having difficulty it may be traced to a problem with one of these factors. For instance, the proposed task is not interesting or, inversely, too complicated. Perhaps the resulting product does not have an audience, or is not perceived to be useful by the public. By connecting the output or product directly to research needs, or concrete and desired ends, organizers may better succeed at motivating appropriate publics to help them achieve their goals.

Finding intrinsically motivated individuals may indeed be the key to the success of a project. For crowdsourcing initiatives that propose interesting content, or stimulating or challenging tasks this should not be a problem. In much the same way, tasks that are geared at deciphering pages full of elusive handwriting from previous centuries may have their particular target audience. This is a situation that projects like Transcribe Bentham have already faced and tackled by capitalising on the high intellectual value and philosophical merit of this English philosopher's work. The most difficult problem to circumvent is if the manuscripts themselves are not appealing to audiences, it will likely be difficult to attract contributors in this case.

However, it is also important that media communication about projects be able to impart the relevance of what they aim to accomplish to potential audiences. This, of course, would include project objectives and community benefits. In other words, how these ends may in turn positively affect those very same contributors if they become involved. In the next section we will look at the role of communication as well as outreach

in attracting participants.

9.4 Communication and Outreach

In much the same way as with collaborative software, it is not enough to create a crowdsourcing platform and expect participants to gather in multitudes to discover proposed projects. As with much of the traffic on the web, it is not uncommon for websites to exist without drawing any worthwhile attention to themselves simply because there is not a sufficiently large community of people that has shown interest. And on the other hand, there are websites out there that manage to generate so much traffic– Facebook, Youtube, Twitter, etcetera– that they have rapidly become household names. In some cases the number of likes and views is enough to propel these websites to success, but for the majority of high scoring candidates a considerable amount of effort is required to achieve these results. Of course, there are Search Engine Optimisation (SEO) techniques which can help increase a site’s visibility thanks to keywords and indexation. However, platforms looking to crowdsource digital labour cannot be sure that there are participants already looking to give away their time, particularly when the compensation provided is little more than self-satisfaction for the participant. For many crowdsourcing initiatives effective communication and outreach campaigns beyond the platform are fundamental for success.

Communication about the project should go beyond its platform. This has several advantages, it allows multiplying the intended message across other existing platforms and social networks that already have a stable base of followers, which can help get the message out there faster, and with relatively minimal effort. It is not uncommon for projects to have multiple representative sites on various social platforms: a main website, a dedicated facebook page, a twitter account, and a wiki page for example. Each one of these pages helps to extend the sphere of influence of the project, increasing the chances that potential contributors come across the website and decide to contribute. Of course, cases of these one-hit participants are many, but what projects really hope for is that new

participants become regular contributors. Here, crowdsourcing projects are somewhat similar to consumer websites, whose goal beyond making sales is retaining customers to build a more solid and diverse customer base in the longterm. Therefore, crowdsourcing communications initiatives may indeed find important overlap with consumer marketing campaigns. Perhaps even so far as to consider using some of the more traditional communications and marketing techniques to spread the word about their crowdsourcing campaigns. Project attractiveness can be positively enhanced with clever or creative titles that stay in users' memories.

Traditional modes of communications such as newspapers, magazines, or radio may be effective in extending the scope of a project. These media usually propose advertising space or can publish an article or interview, exposing some of the main motivations behind the project and inviting people to get involved. Notably, this technique was used by *Transcribe Bentham*, which thanks to articles in the *New York Times*, the *Sunday Times* and through various radio communications was able to drastically augment, and later even sustain, interest from the general public [Causer and Terras, 2014 ; Dunn and Hedges, 2012]. The Bentham project owes much of its success, and almost 6,000 transcriptions over the course of three years, to very clever handling of its communications with the help of mainstream media [Causer and Terras, 2014].

Even large projects such as Zooniverse have put a significant amount of importance on this aspect of getting the word out about their various projects. In fact, by subscribing to an existing Zooniverse project participants may also choose to receive information about new initiatives and invitations to test out recently created projects and also to give user feedback about said project that could be of potential use to project leaders and Zooniverse itself. Zooniverse has effectively optimized its relationship with its users. In essence, this technique is no different from those used by marketing campaigns for consumer products and services, who use these techniques to build and support a stable customer base. The communication can be an invitation to special events hosted by the company or a newsletter for sales or promotions of certain products. With Zooniverse, it takes the form of regular e-mails inviting subscribed members to try out recently launched projects.

Reminders of this sort keep members virtually linked to the platform and encourage them to remain actively involved in its operations.

Building a network within a given sector of activity can also help diffuse information about existing projects, as registries of similar initiatives become regrouped it may be worthwhile to create bridges between project websites. If this happens, it may become easier to discover new projects from existing project websites. Of course, since many projects are linked directly to an institution, a university, library or research center, it may only be possible to be cited if one's project belongs to a particular institution.

As proposed by Chignard [2012] the strategies used to promote open data initiatives should include animation, promotion and quantification ("animer, valoriser, mesurer"), but the project organisers can in many cases resort to third party organisations to help with promotion, animation and generally spreading the word. Furthermore, the networking technique among several related projects and initiatives can be even more effective in engaging potential audiences. These third party promoters can be organisations that regroup regional or federal initiatives. Yet, there also exist a certain number of online hosting sites that seek to collect various projects belonging to crowdsourcing or citizen science in one registry, which may help augment the visibility of these sites. A good example of this for the digital humanities is the Connected Communities site⁴ which regroups crowdsourcing projects belonging to humanities disciplines and also SciStarter⁵ for citizen science projects. In France in particular, the organisation specialising in supporting work in digital humanities is Huma-Num⁶.

It is particularly important to be aware that there are, in many cases, costs associated with communication and outreach. Community managers, social influencers, and scientific mediators are professional positions that can make up project leading teams. That is, besides scientists and or scholars themselves. One may ask how these may be included in research environments where these competences, logic, and more particularly, post profiles, are not necessarily accounted for. One may be inclined to look toward the

4. <https://connected-communities.org>

5. <https://scistarter.com/>

6. <http://www.huma-num.fr/>

DH Manifesto, which anticipates the evolution of research professions and custodians of knowledge to more inclusive and engaged practices [Schnapp et al., 2009].

9.5 Skills and Training

Acquiring new digital skills and training are important for working in novel and collaborative ways in the DH research context. Training for certain skills and competences may be particularly difficult to put in place, since many teams composed of scholars of a particular discipline have acquired skills considered pertinent for that specific field, before complementary (and often computer-based) skills were considered a necessary component. In a significant manner, Digital Humanities aim to change this and not only for younger cohorts of scholars, but also for more experienced members. Technical training for all members of DH teams should be available so as to facilitate the transitions from one manner of working to another and also from one technology to the next, as this aspect will certainly continue to evolve.

9.5.1 Training volunteers

With respects to training of contributors a variety of practices exist. Some of these put researchers and participants in close contact through face-to-face or group training sessions. Examples of this are cited in the works of [Cohn, 2008] for citizen science projects in the field of ecology. Researchers having a great stake in the quality of the results of the work of volunteers will go to great lengths to assure that tasks are well formulated, the equipment is well calibrated and the volunteers themselves know how to gather appropriate data. They have obviously already considered how the quality of volunteer-contributed data may affect their research findings, in some cases going so far as creating groups of volunteers who are overseen by knowledgeable staff during data gathering activities [Cohn, 2008]. This can certainly transform the work dynamic into something that resembles the professional workforce more closely, where interns are overseen by trained colleagues during a process that ultimately leads to the interns acquiring the same (or at

least partial) knowledge of the tasks performed. There is significant reason to consider that this form of hands-on training can be an excellent way to supplement contemporary educational programs, where experts' knowledge is diffused voluntarily to motivated individuals outside of any rigorous educational or professional framework.

The benefits will be all the more worthwhile if novice volunteers are given opportunities to acquire knowledge that can later be transferred to other activity sectors, or to a professional activity of their choice. Once again, to recall what was said in Section 9.3, acquiring useful skills and knowledge can be both intrinsically motivating for many individuals as well as being an opportunity to build positive feedback loops of extrinsic motivation [Ryan and Deci, 2000]. Finally, it may be a way for volunteers who seek to use acquired knowledge to enter a particular sector of activity, but who do not have the means to invest in academic training [Cohn, 2008].

Similarly, very involved training practices exist in the humanities, but are associated with the training of paid work by interns or specific contract positions. In France, individuals to whom these types of contracts are attributed are called *vacataires* and they are required by their contracts to perform a certain amount of work in a limited amount of time. They also receive specific training for the tasks they undertake. Projects like *Les Manuscrits de Stendhal* have long operated with the help of *vacataires* to mutual benefit; the *vacataires* receive training and enhance their professional portfolio while working to help the project achieve its transcription goals. Albeit, this practice greatly depends on available funding and in most cases only one or two people can be attributed part-time contracts at one time. Thus, projects can employ contract workers to increase their progress, but their rapidity is still not as high as may be expected with a few dozen volunteers.

9.5.2 Online instruction

Increasingly, with crowdsourcing projects that use online platforms we are seeing more and more autonomous training and protocols. Users can access these at their leisure and

use them to enhance or supplement their knowledge as they engage in various crowdsourcing activities online. This practice is developing almost in parallel with practices that diffuse knowledge in open source formats, through online tutorials and MOOCs for example. In a new mode of "hacking an education" [LaPlante, 2013], for those delivering knowledge for what could previously be acquired only through select certifying or licensing institutions, or through programs of "distance learning" organised by these same institutions hoping to improve access to their teaching services, online tutorials are replacing face-to-face learning and training. In the case of institutions providing training for volunteers, depending on the task, the skill-level required, the supporting skillset, and the clarity of instructions, this method of training may produce variable results.

Crowd science and humanities projects are good candidates for providing autonomous instructions to participants for accomplishing tasks. These often take the form of written instructions accompanied by supporting images in a sequence of pop-up dialogue windows (Zooniverse, AnnoTate, and Crowdcrafting are three examples that use this form of online tutorial). More detailed instructions or supporting documentation can be included in the form of *wikis*, such as in the case of Transcribe Bentham.

When projects rely on volunteers, the principle of writing clear protocols is of utmost importance, but so is defining tasks of appropriate levels of difficulty to ensure the accuracy of resulting data [Cohn, 2008]. Effective protocols are a combination of clear and concise communication about reasonably practicable tasks that users can carry out in work environments created for that purpose. Preparing protocols that use different media to communicate expectations, including through video, images, and audio, can help increase users' understanding of what is expected. Designing protocols for tasks of various degrees of complexity can also provide for an excellent terrain of study on the efficacy of using online and autonomous training for crowdsourcing. Furthermore, this should contribute to developing more extensive knowledge on the evaluation of crowdsourcing results, with the goal of optimizing the quality of instructions provided to participants.

9.6 Chapter Summary and Conclusion

In this chapter we have seen the important role that human actors play in crowd science and also citizen humanities or citizen scholarly editing. We have identified their involvement as being largely decisive for successful collaboration, effective communication, and outreach.

We have also addressed the need to support skills and training in Digital Humanities; from enhancing volunteer skills and valuing volunteer involvement, to taking advantage of autonomous instruction. All are ways of supporting skills and training for human actors in these fields.

In the next chapter we will look at the work produced within crowdsourcing environments with the intention of assuring quality work from participants. Throughout this following chapter we will see how methods that support training, behaviour, and work quality intercept with tasks, feedback, and products to ensure successful and productive crowdsourcing environments.

Part IV

Demonstration of experimental results

Part IV summary

[L’humanisme numérique permet d’éviter de penser la technique comme quelque chose qui s’oppose à l’humain, allant au-delà du cliché d’un conflit entre l’homme et la machine, pour penser, au contraire, une convergence entre technique et culture [Vitali-Rosati and Sinatra, 2014].

So far in our dissertation we have presented the different elements that constitute our research subject. We have presented the theoretical and methodological reasons for our work on manuscript transcription within an increasingly digital context. We have presented our data object and exposed its formal components and the processes that govern its transformations. We have also looked at architectures and interfaces that create work environments for opening these processes up to inexperienced and motivated publics. The following part of this dissertation looks closer at methods for assuring quality of data obtained through crowdsourcing. Namely, in Chapter 10 we present existing methods of quality assurance and how these are applied to different aspects of work, that is by focusing on tasks, feedback, and products. By paying attention to their interactions, we can create work environments that are more beneficial to participants. In Chapter 11, we will describe our crowdsourcing experiments and evaluate the data that was collected using our method of comparative quality analysis based on expert reference transcriptions. Finally, in Chapter 12, we expose different factors that contribute to the complexity of transcription tasks and present the results of an experiment that investigates two such factors.

Chapter 10

Quality assurance for crowdsourced production

Contents

| | |
|---|------------|
| 10.1 Existing methods of quality assurance | 179 |
| 10.2 Task-based QA | 183 |
| 10.2.1 Gold standards | 184 |
| 10.2.2 Worker screening and training | 185 |
| 10.3 Feedback-based QA | 185 |
| 10.3.1 Expert feedback | 185 |
| 10.3.2 Peer feedback | 186 |
| 10.3.3 Automatic live feedback | 187 |
| 10.4 Production-based QA | 190 |
| 10.4.1 Multiple productions | 190 |
| 10.5 Conclusion | 191 |

In this chapter we discuss existing methods that can be part of a broad strategy of quality assurance used by crowdsourcing projects. We consider these methods from three perspectives: pertaining to task, pertaining to feedback, and finally to productions themselves. We then present our chosen method for evaluating transcriptions contributed

by novice transcribers.

10.1 Existing methods of quality assurance

Since soliciting work from non-expert publics has gained in popularity, questions about quality control and assurance have become central to discussions on crowdsourcing. In reality, although the number of projects that use crowdsourcing has increased, the research and scientific literature on the efficacy of crowdsourcing and the quality of data produced is still insufficient [Franzoni and Sauermann, 2014]. Research on quality control and implementation of defined quality assurance practices is more prevalent in industrial contexts, where participants are paid workers. Whereas, academic environments are more hesitant to put in place similar modes of operation, perhaps for fear of alienating volunteers in a context where the distinction between worker and volunteer becomes increasingly ambiguous. Meanwhile, numerous techniques have been tested by industrial crowdsourcers, such as the use of gold standard training data, various forms of feedback, and having the same work performed by multiple workers [Le et al., 2010]. These techniques present a number of interesting solutions to the question of quality raised in an industrial crowdsourcing context. Particularly in light of evidence that crowdsourced workers tend to produce mediocre rather than exemplary work [Callison-Burch, 2009 ; Downs et al., 2010].

Of course, a predominant number of the methods employed by industrial crowdsourcers entail a considerable level of technical complexity, which needs to be mastered. However, there are also commonly known and used techniques that have proven effective, and which can be implemented relatively simply in a crowdsourcing workflow. For comparative purposes we can consider programmatic gold standard techniques or periodic screening and feedback, which are both techniques that make use of training data in slightly different ways [Oleson et al., 2011 ; Downs et al., 2010]. We can compare these two techniques to peer and expert review [Dunn and Hedges, 2012], which are arguably simpler to put in place from a technical perspective, but do require continuous user involvement.

Peer and expert review are also a fundamental part of scholarly publishing processes, which have historically strived to achieve scientific excellence through critical examination of scholars' work [Fitzpatrick, 2011]. Although peer and expert review are associated with

a longstanding tradition of quality assurance within scientific and scholarly disciplines, there appears also to be much room for criticism of these practices. Scholarly publishing is also an industry, one that resides within the sphere of scholarship and academic research, but an industry nonetheless. As Fitzpatrick argues, the process of evaluation by peers has also historically been proven to be a process of censorship, less intended to ensure quality control of information that circulates within the academic sphere than to boost editorial expertise [Fitzpatrick, 2011]. Based on information collected, Fitzpatrick summarizes her take on the position and the role of peer review in academic establishments:

On the one hand, peer review has its deep origins in state censorship, as developed through the establishment and membership practices of state-supported academies; on the other, peer review was intended to augment the authority of a journals' editor rather than assume the quality of a journal's products. Given those two disruptions in our contemporary notions about the purposes of peer review, it may be less surprising to find that the mode of formalized review that we now value in the academy seems not to have become a universal part of the scientific method, and thus of the scholarly publishing process, until as late as the middle of the twentieth century [...] The history of peer review thus appears to have been both longer and shorter than we may realize. And yet, because of the role that it has played in authorizing academic research—because we ourselves, as Biagioli suggests, are both the subject and the object of its disciplining gestures—it has become so intractably established that we have a hard time imagining not just a future without it, but any way that it could conceivably change [Fitzpatrick, 2011].

Furthermore, it is important to consider that the academic context may have a deliberately different position on the employer-worker relationship. Simply put, few scholarly project leaders wish to put in place work environments comparable to Amazon's Mechanical Turk. Whether for fear of alienating participants by recreating an environment that has often been criticized for openly exploiting underskilled workers. Or, for fear of disseminating mediocre quality data in a scientific and scholarly research context. Nevertheless,

questions concerning quality have been evoked both in industrial and academic contexts. Both have clearly defined reasons for expecting quality work and both would like to avoid low quality contributions.

For these reasons, it becomes interesting, despite our clear position in a scholarly context, to investigate some solutions put in place by large-scale industrial crowdsourcers, and at least consider their applicability to a scientific context. Once again, we find ourselves right in the center of participative activities. It is thus an excellent opportunity to consider different techniques that can be used to resolve questions on quality assurance, and how these can be applied in a context where participants are volunteers and not paid workers.

At the same time, a number of projects in the humanities have successfully implemented quality assurance methods that can be good options for projects with modest technical means. Moreover, learning from these projects may allow to lay the groundwork for further improvements. Quality assurance has its place within the context of citizen scholarly editing just as it does within the broader context of crowdsourced production. An investigation of common and existing practices from a wide range of areas will allow to expound a certain number of available options. We will also see that a number of approaches jointly rely on forms of peer or expert review, and also on expert feedback [Dunn and Hedges, 2012]. While others rely on comparative algorithms to determine the best of multiple contributions. Provided that both approaches can be useful and enriching for processes and people involved, it may be worthwhile to explore how techniques can be combined to achieve desired goals.

In many cases these techniques require putting in place complex technical environments with specific focus either on task or data processing, on group management or a combination of both. Implementation of various components within the overall system is meant to increase its degree of intelligence and obtain better quality results across the whole system. When focusing on tasks the components involved will concern instructions and the way they are communicated to workers. When focusing on productions the components will deal with processing, comparing and evaluating data. When focusing

on group management, the components will handle behavioural aspects of group work and the manner in which participants receive feedback from other implicated actors. Figure 10.1 illustrates the relationship between the conceptual components involved and the various procedures that have been put in place for assuring work quality.

Let us take a moment to define some related terms: quality assurance, quality control, and quality assessment. Quality assurance (QA) generally refers to a broad plan for maintaining quality of all aspects of a program or process. It can include a combination of processes, including managerial ones, geared at ensuring and maintaining quality. Quality control (QC) refers to specific steps taken to determine whether procedures or components within a system are valid, as part of a broader plan. Then, quality assessment (QAssessment) is an appraisal or evaluation that can take place at various stages of a process to determine outcomes based on the controls put in place. Quality assessment can also refer to the appraisal of an overall outcome with respects to goals established at the outset. To give an example from the Benoîte Groult corpus, transcription comparison and scientific validation are two components that fit into a quality assurance plan for the project. Once a certain number of pages are transcribed, a quality assessment could be carried out to evaluate how well goals were met for that particular set of pages.

Quality assessment of hundreds, or thousands, of pages is difficult to imagine, but putting in place efforts and procedures to assure quality is. In our case, we refer to the evaluation transcription quality as part of a broader plan for quality assurance. When we use the abbreviation QA, we are referring to quality assurance, and procedures that fit into this plan. Otherwise, we employ the full terms to refer specifically to quality assessment and quality control.

Procedures that are put in place to assure quality can concern different aspects of the process. We have identified three of these areas of influence on the overall system, as shown in Figure 10.1. We name them as related to tasks, feedback, and product; we consider that a quality assurance plan can be organized by taking into account the relationship between tasks, feedback, and production. Within this system, various quality control methods can be put in place. In this figure, we have positioned procedures within

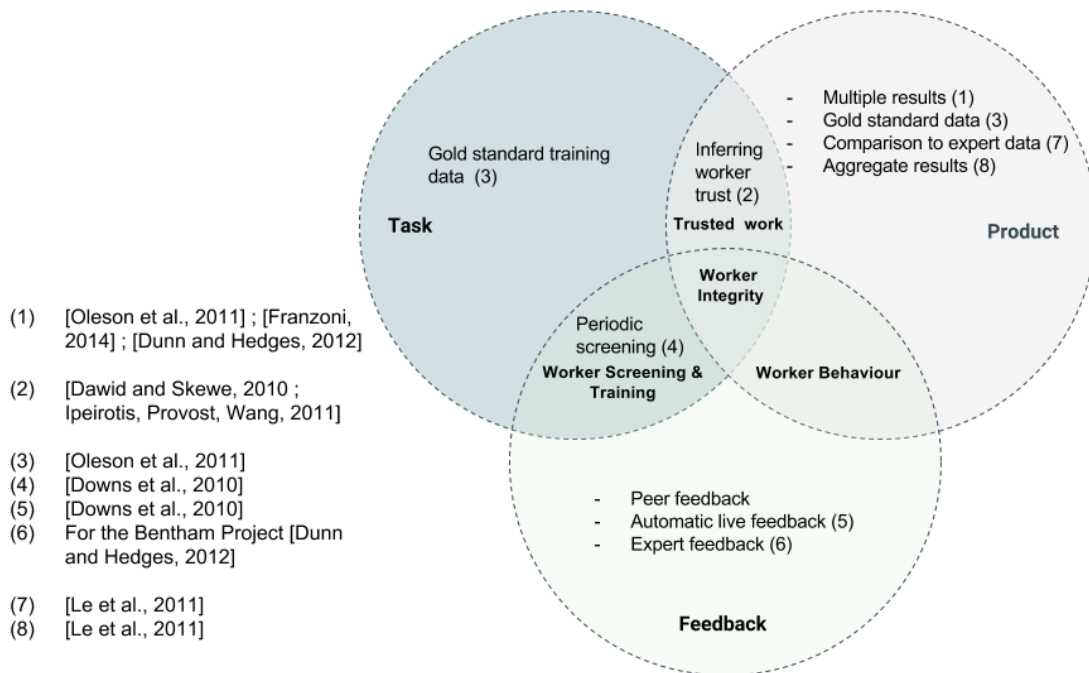


Figure 10.1 – Illustration of the relationship between tasks, feedback and production in a crowdsourcing environment.

their main components of influence, but a few lie within areas of overlap, which result from the interaction of two components. These areas show how components combine to produce effects on workers and their work. For example, worker screening and training involves assigning tasks and providing feedback upon accomplishment. Worker behaviour is deduced from product, but is also influenced by feedback received. Finally, trusted work results from tasks being accomplished so as to achieve desired product or output. The convergence of these three areas of overlap can be considered as achievement of worker integrity, based on training, exemplary behaviour and trusted work. In the following sections, we describe these components and present associated quality control methods in more detail.

10.2 Task-based QA

To understand what is involved in task-based QA, we need to consider the tasks

that participants set out to accomplish. We have considered that the task component interacts with the other two components in the following manner. The worker is given a task, which he or she must accomplish according to given instructions. The worker may receive feedback in various forms, which will impact his or her understanding of the task. Depending on a certain number of factors, including outset skills, understanding of the task, and type of feedback received, the worker will produce a work output.

Quality control that is centered on the task actually involves a number of different activities that can be implemented partially or in parallel. The first is screening based on various factors that are contingent on participant skills and motivation relative to the task. The second is training of individuals that are considered to be a good fit for specific tasks. Task-based QA intercepts with the feedback component to accomplish worker training through periodic screening, even while the worker is already producing an output based on the task he or she is working on. Meanwhile, the work produced is subjected to quality assessment that conjointly evaluates the workers skill and accuracy based on the work produced, thus establishing a worker profile [Oleson et al., 2011].

10.2.1 Gold standards

One method that has been used to ensure the quality of data produced by crowdsourcing involves inserting what is referred to as gold standard data into regular data sets. Gold standard data is actually training data, for which correct responses are known and the ability of participants to respond correctly determines their eligibility to continue working on a given assignment. The results from training data are used to infer an estimated level of quality for the rest of the data produced by a particular worker [Le et al., 2010 ; Oleson et al., 2011]. Using gold standard data also allows to provide feedback on common errors and thus administer ongoing training to individuals [Oleson et al., 2011]. The primary downside of this practice is the need to manually create gold standard data sets and solutions, which is an expensive and time-consuming process [Oleson et al., 2011]. Another downside is the discoverability of gold standard data sets within regular data sets, which makes scamming the system easier for participants who achieve high scores

only on training data, but otherwise produce subpar work [Oleson et al., 2011].

10.2.2 Worker screening and training

Experiments have been conducted to evaluate the process of worker training and incorporate both periodic screening and feedback with the goal to train individuals while they are completing tasks. The method used by [Downs et al., 2010] seeks to effectively screen individuals who apply to contribute work in order to retain only those who are qualified and conscientious. The study was used within the environment of Amazon’s Mechanical Turk.

10.3 Feedback-based QA

Now positioned in the lowest circle in Figure 10.1, we will look at feedback-based QA. Feedback-based procedures may include expert and peer feedback on the task and work output (thus its central position). Within complex (intelligent) systems, feedback can be automated and delivered live or while the worker is active. Feedback that has an effect on the worker’s understanding of the task comprises worker training, during which time the worker learns what is expected of him or her. Further feedback on the work product will have the effect of upholding worker behaviour, including preventing subpar work and scamming of the system.

10.3.1 Expert feedback

One of the simplest means of putting in place quality control in a transcription project is to have productions reviewed by an expert group, who would then provide feedback to transcribers. This method also functions as a training method, allowing experts to advise transcribers on the errors they make and how to correct or avoid them. This method may be more or less automated. In the former case, a system can be built to detect common types of errors and provide automatic feedback to correct and advise transcribers. The

latter less automated option resembles more of a student-mentor relationship, in which experts overlook transcribers' work and guide them in acquiring the necessary skills to do the tasks as they would themselves. In this case putting in place individualized expert feedback is an excellent opportunity to train transcribers as well as establish relationships between members of the project. Expert feedback is generally appreciated by novice participants, who do not yet feel confident in the choices they make in their work.

Transcribe Bentham provides expert feedback, which has proven both helpful and motivating to participants [Dunn and Hedges, 2012 ; Causer and Terras, 2014].

10.3.2 Peer feedback

Peer feedback is like expert feedback. The main difference arises from a change in who provides feedback to whom. Systems that employ peer feedback are likely to be less hierarchical in organisation and closer to Wikipedia's model of production [Dow et al., 2011].

We can also note a difference between direct *person-to-person* feedback and indirect *modification-in-document* feedback. The former is more beneficial to users if they actually receive notification of modifications made to documents they create. This does not replace personal feedback such as can be provided by mentors, but it allows users to observe reactions to their work. Person-to-person feedback allows to establish contact between different users of the system, but it can have a high productivity cost if we consider the effort required to write feedback to users versus the effort required to make modifications directly. As such, notifications of modifications can be a good compromise.

As in the case of more hierarchical models, peer feedback also allows to establish contact between differently ranked users of the system. The more traditional student-mentor axis is replaced by the possibility to receive feedback from all workers regardless of their status. Anyone may be in a position to notice errors and propose worthwhile corrections. This encourages situations where co-learning, or cooperative learning, can take place.

Furthermore, integrating peer feedback into a system requires allowing users to occupy multiple roles. When roles are performed simultaneously, such as transcribing and providing feedback to other transcribers, we are talking about overlapping roles [Dow et al., 2011]. Work-feedback overlap exists in its simplest form in systems like Wikipedia. However an increased level of hierarchy can be established if workers acquire feedback roles as a result of being promoted for quality work [Dow et al., 2011]. Hierarchy within a system is not necessarily unwanted, so long as role flexibility and opportunities for progress are implemented into the system. That is, so long as workers can interact, learn and evolve within said system. Indeed, these are all benefits for workers, without which demotivation [Ryan and Deci, 2000] and sub-par work performance may become obstacles [Dow et al., 2011 ; Downs et al., 2010].

Within PHuN2.0, for example, we put in place a way for transcription verification to be administered by peer transcribers and not only expert transcribers. A difficult and largely questioned decision by project leaders, but one that aims to liberate the transcription workflow from a heavily hierarchical constraint that puts the bulk of verification tasks on a select few. To rebalance the system in light of this decision, project leaders maintain the right to devalidate unsatisfactory transcriptions and push them back into the workflow if deemed necessary.

10.3.3 Automatic live feedback

Amazon's Mechanical Turk has been criticized for its lack of timely feedback mechanisms for workers, or its asynchronous feedback [Dow et al., 2011]. The work of [Dow et al., 2011] investigates how feedback mechanisms can be put in place within crowdsourcing infrastructures to distribute automatic feedback to workers. The authors describe a system¹ for visualising crowdsourced work and distributing feedback to workers. It is a clever use of available technological means.

According to the authors of [Dow et al., 2011], using synchronous feedback, as opposed

1. The system is called Shepherd.

to postliminary feedback, provides interactive support to users, which enhances users' experience of the virtual work environment. The investigations led by the authors of the system they created point to improvements not only in work quality but also worker management and engagement. This can be seen as a support mechanism for quality assurance.

As shown in Figure 10.2, a feedback system can take into consideration multiple parameters. Dow et al. [2011] identify and explain five of these in particular:

- **timeliness:** feedback can either be delivered synchronously or asynchronously to the worker when he or she completes a task. It is preferable to limit the time elapsed between tasks and feedback, as this will be more effectively assimilated by workers.
- **specificity:** rather than binary responses, feedback can be adapted to types of tasks and specific user input. Generic, but adjustable response templates can help users gain a better understanding of how they can improve their work.
- **source:** feedback can come from different sources, either experts themselves or other workers. Diversifying the sources of feedback can be beneficial, since, experts sometimes do not perceive the difficulties in their work and do not explain concepts in terms that are understandable to workers. The benefits are similar to peer feedback as discussed in Section 10.3.2
- **format:** feedback can come in different formats, such as text, images, video, and audio, although most systems currently distribute it only in text form.
- **ratio of work to feedback:** the work to feedback ratio or relationship can be unique; either there can be multiple feedbacks for one task, or there can be one general response to multiple tasks. Managing this ratio can be an effective way to manage the effort required to write feedback to users.

Any one system may address some or all identified parameters (others can surely be identified). Typically, the more detailed the feedback required, the more user involvement is necessary to ensure sufficient activity and benefit. This entails a number of actions, including designating responsible members, creating necessary infrastructure and elabo-

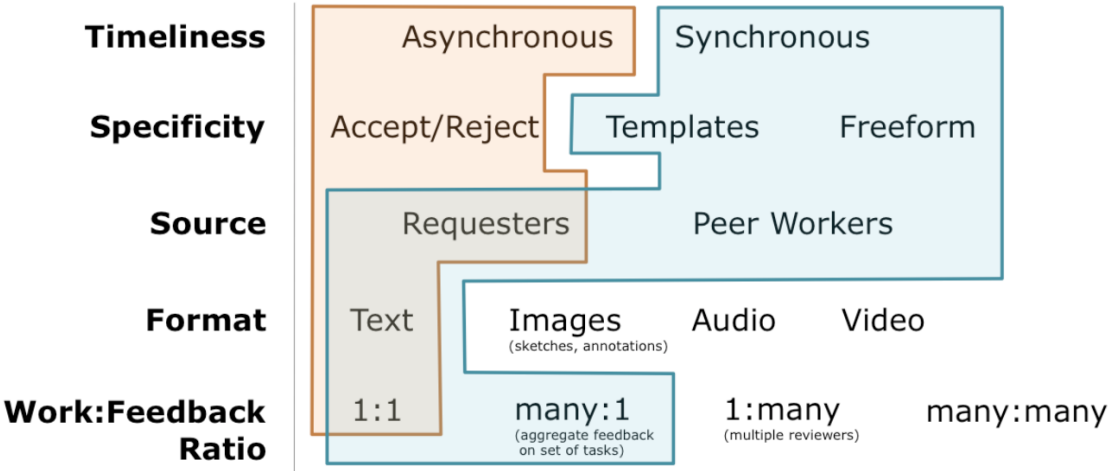


Figure 10.2 – Authors Dow et al. (2011) compare existing feedback mechanisms (shown in orange) with Shepherd’s adapted feedback system (shown in blue). The authors illustrate different aspects of feedback and areas of overlap between existing systems and their innovative one.

rating more detailed expectations.

In our case, our responsible members can be experts, and later on, experienced users. Instructions can be elaborated to communicate expectations. Lastly, creating an infrastructure that distributes suitable feedback to users can be a constituent part of a broader plan to assure quality in crowdsourced transcription environments.

10.4 Production-based QA

Situated within the product circle of Figure 10.1 on page 183, production-based QA involves evaluating the work output itself. This component can interact with the task component to simultaneously or in parallel create a worker profile based on work produced relative to a given task. Methods that evaluate product output can involve comparison of multiple productions between themselves, comparison of productions with regards to known correct responses (or expert data), and finally using multiple responses to aggregate information or results.

10.4.1 Multiple productions

Testing a set of worker output against another group is a way of normalising the output. This method has been largely implemented in crowd work settings, particularly in micro-task platforms and also in research cases implemented within the Mechanical Turk environment [Downs et al., 2010]. Citizen science projects like Galaxy Zoo and citizen humanities projects like Marine Lives use multiple productions as a way to verify data and minimize errors [Franzoni and Sauermann, 2014 ; Dunn and Hedges, 2012]. In the case of Galaxy Zoo, multiple user responses can be weighed to filter errors by relying on most frequently submitted responses [Franzoni and Sauermann, 2014]. Now, using this method on classification tasks or multiple choice responses is simpler than on transcriptions, where each data set can be highly variable. However, even variabilities in submitted texts can be weighed against one another to determine which words or characters were chosen most frequently by contributors. This method of aggregating texts can be effective in filtering errors by relying on multiple transcriptions. For example, this can be used on a batch of transcriptions containing a difficult word in a manuscript. If we can rely on users to correctly recognize the word in question the majority of the time, then we can generate a transcription that contains the correctly spelled word, and filter erroneously spelled variants.

We have already discussed how gold standard data can be used as part of training in

crowd work environments, but it can also be used to verify the quality of data produced against known references [Oleson et al., 2011 ; Le et al., 2010]. Combining reference data with multiple contributions can be used to select texts that most closely match the reference. In Chapter 11, we present a technique based on these principles to measure the quality of crowdsourced transcriptions.

10.5 Conclusion

As shown in the sections outlining a crowdsourcing system, the work produced by contributors is actually a system of interacting components that can be divided into task, feedback and product (see Section 10.1 and Figure 10.1).

Some scholars consider that expert review is « more related to censorship than to quality control » [Fitzpatrick, 2011]. And while peer review can indeed be considered an important component of professional scientific and scholarly practice it is more appropriately applied to critical readings of authors' work. While the work of constituting digital primary resources for critical scholarly work should be controlled for accuracy, especially as it constitutes the basis of scholarly work, methods other than peer review exist and should be considered for the purpose of assuring and controlling quality of public contributions.

Industrial systems that have been built for crowdsourcing data encoding work that is arguably more similar to transcription work than other types of content creation. Techniques developed and used in these contexts should be considered in order to create more support for editorial processes that can assist and supervise contributions from motivated inexperienced contributors as well as experienced ones.

Technologies can facilitate this change and accompany a wider group of contributors in the production of quality work. For this to happen we need once again to consider the benefits of existing methods. For example, in Sections 10.3.2 and 10.3.1 we speak of peer and expert feedback rather than review. These forms have been put in place for crowdsourcing in both industrial and academic contexts. It is important to consider the

positive role that interaction with peers and experts plays in training and accompanying contributors. Yet other methods exist and have been more or less widely implemented in micro-task environments and generally in crowdsourcing environments. In general, many scholars would agree that both scientific and scholarly publishing would benefit from a combination of both "editorial authority" and "modern technology" [Kolowich, 2011].

Our consideration of the digital editorial system has brought us also to consider the need to look beyond just peer and expert review and into how other methodologies can contribute to improving it, particularly with respects to manuscript transcription. Without overlooking the other components and their interaction (Figure 10.1), we should look closer at the product component. The product component refers to actual transcriptions with which scholars work to produce scholarly editions. Existing industrial methods for evaluating productions include comparison of multiple productions (or units [Oleson et al., 2011] of output). We focus on these methods for our evaluation of crowdsourced transcriptions for scholarly editing using known distance measurement techniques as they are commonly applied to texts, and which we described in Chapter 6. This will allow us to compare different transcriptions between themselves and, where possible, in relation to expected output as defined by expert groups.

Furthermore, it is important to address the likelihood that a correlation exists between the complexity of a page and the quality of transcriptions that non-experts produce within a complex system (task, feedback, product). If we use a distance measurement technique, such as those frequently employed for document comparison, we can compare expert transcriptions to non-expert transcriptions and articulate the similarities or differences in terms of document distance. Performing this analysis on pages of varying complexity would allow us to study the correlation between page complexity and transcription quality and allow us to characterize it if indeed it is present.

In particular, if a correlation can be characterized, we can respond to questions such as the following: Are complex pages more likely to produce non-expert transcriptions of low or insufficient quality? What degree of complexity produces satisfactory results for non-expert transcriptions? And other questions of this order. We investigate these

questions in particular in Chapter 12.

More generally, regardless of the type of source material (manuscript page or otherwise), we can use this comparative approach of gathering multiple productions (or transcriptions) in order to evaluate what the public can contribute to scholarly editing. The quality of contributed transcriptions is a primary indicator of the type of material that scholars will work with within a larger editorial workflow. Subsequently this would allow for a better organisation of editorial processes, in terms of time, invested effort as well as task planning.

It is therefore of great interest to use available technologies and consider existing, albeit largely industrially-implemented methods, to evaluate crowdsourced transcriptions. In the following, Chapter 11, we will demonstrate through our experiments how comparing multiple transcriptions can be used to both observe variability in the work of inexperienced transcribers, and also as a method of monitoring transcription quality.

Chapter 11

Measuring transcription quality

Contents

| | |
|---|------------|
| 11.1 Evaluation of transcription quality | 197 |
| 11.1.1 Primary conjectures | 198 |
| 11.2 Experimentation on the Stendhal Corpus | 200 |
| 11.2.1 Stendhal Experiment 1 | 200 |
| 11.2.2 Sample description | 200 |
| 11.2.3 Phylogenetic analysis | 203 |
| 11.2.4 Digging into the data | 211 |
| 11.2.5 Study of expert transcriptions | 214 |
| 11.2.6 Observing effects of page variation and complexity | 218 |
| 11.2.7 Drawing preliminary conclusions | 219 |
| 11.3 Experimentation on the Benoîte Groult Corpus | 221 |
| 11.3.1 Benoîte Groult Experiment 1 | 221 |
| 11.3.2 Benoîte Groult Experiment 2 | 224 |
| 11.4 Conclusion | 228 |

In this chapter we present our experiments of quality evaluation of crowdsourced manuscripts. We focus on results collected for Stendhal and Benoîte Groult. Our results demonstrate how variability in the work of inexperienced transcribers can be observed

using computational methods. Furthermore, comparing multiple transcriptions can be used as the basis of a method of quality assurance. To do so we established the underlying assumptions of our work, accumulated data for analysis, and established tools and methods for evaluating results. To bring this penultimate chapter to a close we consider some possible applications.

« Digital Humanities infrastructures encourage PROTOTYPING, generating new projects, beta-testing them with audiences both sympathetic and skeptical, and then actually looking at the results. [Burdick et al., 2012] »

11.1 Evaluation of transcription quality

To evaluate the quality of what can be produced by contributors we need to be able to translate observables into quantifiable terms. In the previous section, Chapter 10, we referred to methods that make use of gold standard data [Le et al., 2010 ; Oleson et al., 2011] to evaluate (and, if we look at the overall system, assure) the quality of content produced within a system. In our case, with access to a perfect transcription, or a set of these, we should be able to compare the transcriptions produced by inexperienced transcribers, or non-experts, and thus obtain a score reflecting the level of quality of each contributor to our defined reference.

However, in most cases of manuscript transcription, gold standard or target data does not exist. Firstly, the nearest we can get to target references are transcriptions made and validated by experts. Taking into account that even experts are subject to committing errors, we will show that we can nevertheless use their productions (both text and XML) as ground truth. To do so, in the first section (Section 11.1.1), we will introduce the distance measurements used to compare transcriptions and we will present a series of simple assumptions about the work of both experts and non-experts. Doing this will help the reader understand why work produced by experienced transcribers, or experts, can be used as reference even if they make mistakes.

In the sections that follow, we will discuss the results obtained from our experimentations. We conducted a total of four experiments and the first three are described in this chapter.

The first one was based on an existing XML editor commonly used for transcription and which we have already mentioned previously, Oxygen Author. Our goal was to verify our primary hypotheses about how the work of non-experienced transcribers held up

against that of experienced transcribers. The results in this experiment were obtained using a transcription interface that is widely used for encoding XML and is, in other words, a high-functioning and effective tool.

The second experiment was conducted using the platform we made and allowed to observe the quality obtained with the new tool.

The third experiment is the result of transcriptions collected over the course of two workshops as well as some third-party participation on the crowdsourcing platform.

Finally, the fourth and final experiment focuses on the possibility to correlate the quality of the obtained transcriptions with the an estimated complexity of the page (itself a component of the overall task complexity). In particular, we followed an experiment design to observe how two identified factors may or may not affect transcription results. We present and discuss these in Chapter 12.

11.1.1 Primary conjectures

We based our experimentations on a number of assumptions that we formalize here. Of course, we began with the question of what crowdsourcing could contribute to manuscript transcription. In general, we observed crowdsourcing projects that attracted contributors having diverse knowledge bases and interests, with a tendency of having particular interest in literature and authors' manuscripts. With an activity like manuscript transcription, crowdsourced transcribers are very likely to be new to the activity of transcription, even if they are also highly likely to be avid readers, and perhaps even writers themselves. The first assumptions we had about this activity and what can be produced this way can be summed up in the following list :

- **A₁**: Expert transcribers are very good at their work. They produce few errors compared to a hypothetical ideal transcription. That is, few corrections would need to be made to obtain a publishable transcription.
- **A₂**: Novice transcribers will produce work that is different from expert transcribers. Novices will likely commit the same errors (at the same places in the

text) as experts but they will also commit errors which expert transcribers will not. The relationship is inclusive.

- **A₃**: We wonder if a novice whose work resembles that of an expert more closely may avoid making errors made by an expert on the same page.

As already stated in Chapter 6, Sub- section 6.3, we used *Levenshtein Distance* to measure the difference between strings. To recall, here is a simple expression of the formula:

$$Distance_{i,j} = additions_{i,j} + subtractions_{i,j} \quad (11.1)$$

As shown, the distance between two texts i and j can be obtained by calculating the sum of the number of additions and subtractions necessary to transform text i into text j .

To understand this further, let us note t_{ideal} as the ideal transcription, t_{expert} an expert's transcription, and t_{novice} a non-experienced user's transcription for the same page. The first assumption **A₁** states that the distance between an ideal and an expert transcription, $d(t_{ideal}, t_{expert})$, is a low value. Figure 11.1 represents this distance roughly as the shortest double sided arrow. The second assumption, **A₂**, states that the number of errors made by a novice is equal to the number of errors made by an expert plus a value. It can be translated with the following formula:

$$d(t_{ideal}, t_{novice}) = d(t_{ideal}, t_{expert}) + d(t_{expert}, t_{novice}) \quad (11.2)$$

Or, if we look back to the Figure 11.1, the longest arrow, representing the total distance between the novice and the ideal, $d(t_{ideal}, t_{novice})$, is the sum of the shortest arrow, $d(t_{ideal}, t_{expert})$, and the mid-sized arrow representing the distance between expert and novice, $d(t_{expert}, t_{novice})$. In reality, we cannot know the distance represented by $d(t_{ideal}, t_{expert})$, so we will concern ourselves only with the distance between experts and novices.

Our assumptions should be verified. In the following section we describe an experi-

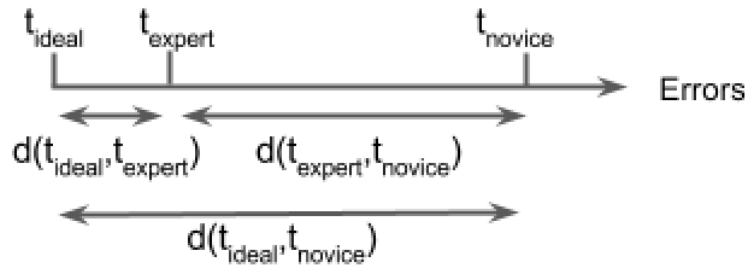


Figure 11.1 – Novice and expert errors compared to an ideal transcription.

ment that was performed to compare expert transcriptions to the work of inexperienced transcribers.

11.2 Experimentation on the Stendhal Corpus

11.2.1 Stendhal Experiment 1

An initial experiment was conducted on a sample from the Stendhal Corpus. For our study we selected two pages from the Stendhal corpus. We chose two pages for which there was no existing (or no available) expert-validated transcription at the time. Also, we sought out pages with a distinguishable (or quantifiable) difference in complexity; one which would appear easier and simpler to transcribe and another more difficult. We did not, however, want that either of the pages be overwhelming to our inexperienced transcribers so as not to discourage them. Figures 11.2 and 11.3 show the first and second page, respectively.

11.2.2 Sample description

Both pages contain only script and no tables or diagrams. The first page contains a total of sixteen lines (not counting lines used for marginalia), with an average of eight words per line. There is only one minor modification to the body of the text (for more

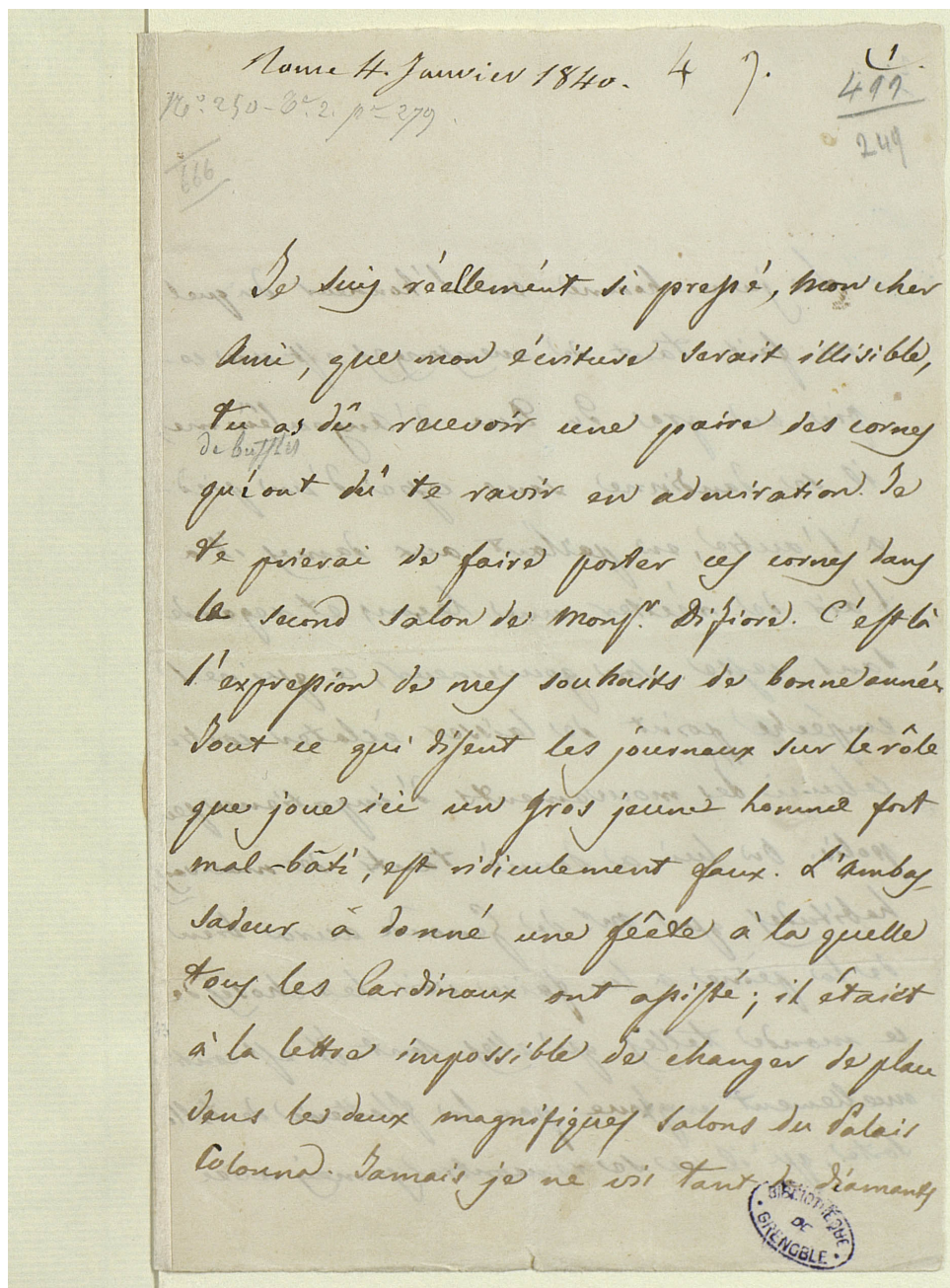


Figure 11.2 – Page from Stendhal experiment 1- page 1. Images are the property of the Grenoble Municipal Library.

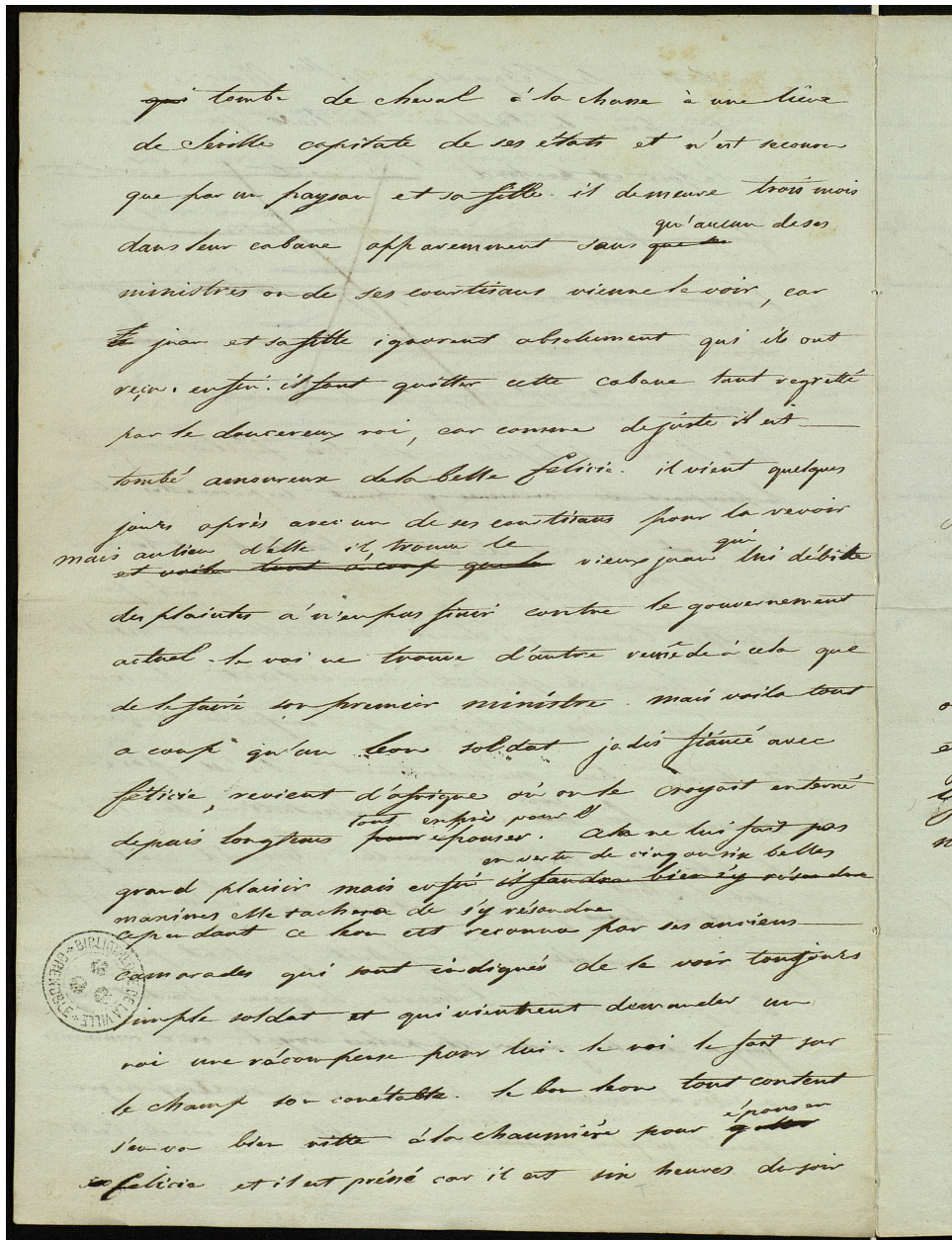


Figure 11.3 – Page from Stendhal experiment 1- page 2. Images are the property of the Grenoble Municipal Library.

details refer to 11.2.4). The second page contains more writing and is thus more dense, with smaller script (important), and with twenty-five lines of text (not counting words inserted in between the lines), averaging ten words per line. This page contains deletions and additions and other types of modifications described by our expert annotation schema. Based on these observations we intuitively considered this page to be more complex. We identified our distinguishing factors as the following:

- word density,
- modifications (deletions, additions, corrections, and other features that can be identified and described using our expert annotation schema),
- script size and inclination,
- multiple mediums (ink, pencil, etc.)

Having identified our sample and validated the XML schema with our expert contributors, we asked them to transcribe the two pages using Oxygen XML Author¹, which they are accustomed to using. We then asked ten other individuals (non-experts) to repeat the same exercise, based on instructions taken from the manual, and which describe each of the element's function. Over the course of approximately one month, our participating non-experts performed the transcription task in their spare time using the software indicated, which they installed on their personal computers.

11.2.3 Phylogenetic analysis

The transcriptions collected from our expert and non-expert groups were compared with the goal of observing the differences between all individuals and between the two groups. We found that quantifying the differences allowed us to do several things. Firstly, to observe each of the individual contributions relative to one another and our multiple expert references. Secondly, to get a general overview of the distribution of individuals as well as groups that formed.

Based on the results obtained using *Levenshtein Distance* we constructed a matrix of

1. https://www.oxygenxml.com/xml_author.html

values detailing the distance score of each text to every other text. This matrix $D = [Distance_{i,j}]$ is called a distance matrix and it has the following properties:

1. D is symmetric: $Distance_{i,j} = Distance_{j,i}$.
2. The diagonal elements of D are equal to 0: $Distance_{i,i} = 0$.

Once obtained, this matrix is used to compute a hierarchical classification of transcriptions. Finally, we visualize the result using a phylogenetic tree as seen in figure 11.4. Each leaf of the tree represents a transcription, and leaf length represents the level of dissimilarity between transcriptions. The closer two transcriptions appear to be in the tree, the more similar they are. One can also refer to their associated numerical values, which are the result of a series of *Levenshtein* operations for each pair of texts amounting to a distance value measured in characters (*chars*). We have also shown the distance matrix as a heatmap, with softer colours representing lower values and more intense colours representing higher values, or greater distances. Moreover, the indexes (labels) identifying each transcription have been aligned so that each row or column can be associated to a particular leaf.

We observed the formation of distinct clusters: one containing our experts and another containing only novices. The first cluster contains the three experts but also contains two novice transcribers that we did not define as experts for the activity². The average distance of the novice cluster to the expert cluster is given as the average of all individual distances observed in the novice cluster in relation to individuals in the expert cluster. We call this the average inter-cluster distance and it amounts to 95,3 characters. It corresponds to values that are visible in the top-right and bottom left corners of the matrix and do not include *novice 2*, which we considered as its own outlying cluster-leaf.

We observe that the average inter-cluster distance between novices and experts is higher than the average intra-cluster distance of the expert cluster, which itself amounts to 62,2 characters. Within it, experts obtain minimal values between them, when considered relative to the whole matrix.

2. These individuals *had* received prior training directly from experts

Finally, the average intra-cluster distance for the novice cluster is 55,1 characters, which is more dense and suggests that novices between them obtain similar results. This contrasts somewhat to the lower density of the expert cluster, which, although it contains novices, displays higher novice-expert distance values than novice-novice values in the neighboring cluster. The novices that appear in the expert cluster are roughly 22 - 35 characters away from the average novice residing in the novice cluster. Table 11.1 resumes this information.

| Intra-expert cluster | Inter-novice-expert cluster | Intra-novice cluster |
|----------------------|-----------------------------|----------------------|
| 62,2 | 95,3 | 55,1 |

Table 11.1 – Résumé of intra and inter cluster distance averages.

In our case, to use the transcriptions for comparison, we performed distance measurements first on raw text files (generated from XML and stripped of all tags) and secondly on the XML itself with Zhang-Shasha’s tree edit distance algorithm [Zhang and Shasha, 1989]. Firstly, we will look at the raw text.

a) Raw text analysis

We consider the cluster containing our experts to be closest to the target text. It should be noted that the two novices (9 and 10) sharing the cluster with experts (1, 2 and 3), seen in the top left corner of Figure 11.4, received training from the experts themselves. Whereas, the other novices had just received instructions to follow. Initially, these experienced novices were considered simply as novices. It was after looking at the results that it became clear that the prior instruction they received surely had some impact on the results we were able to observe.

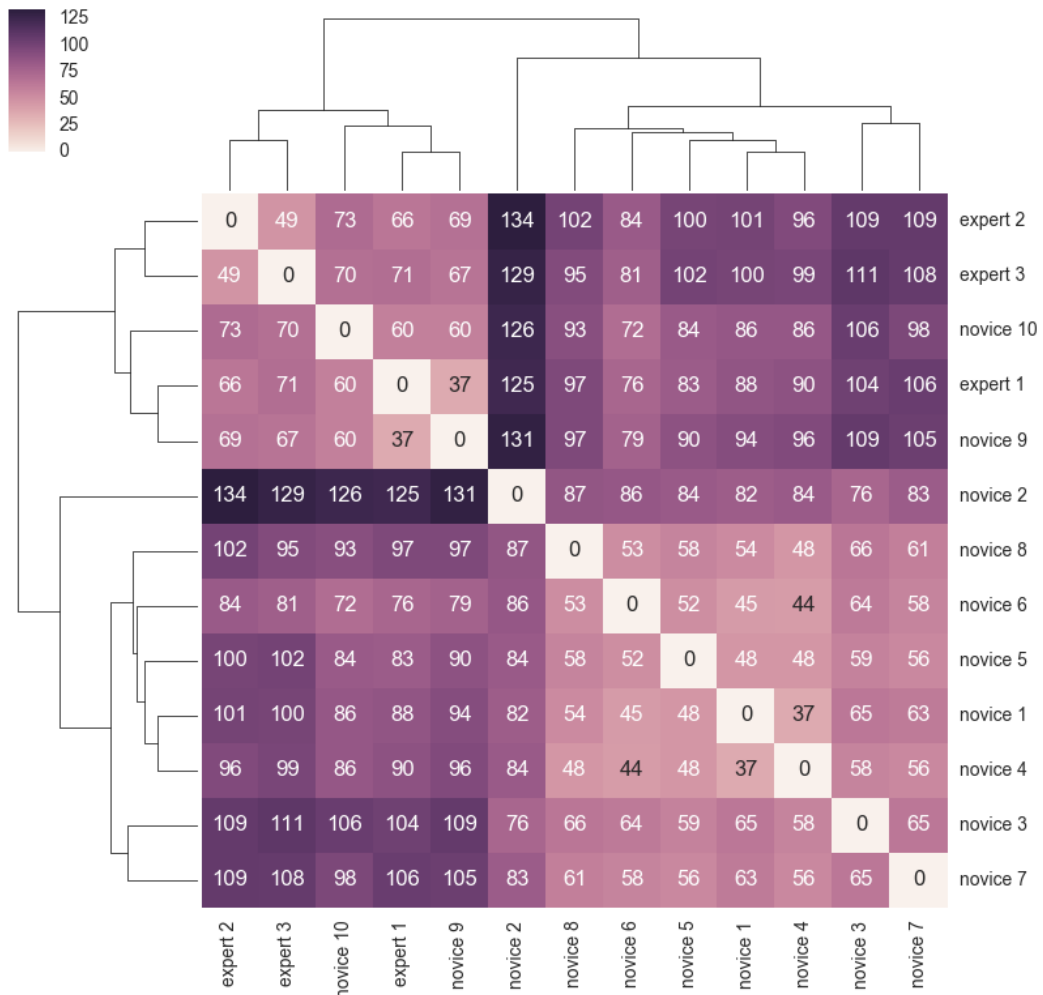


Figure 11.4 – Representation of the distance matrix for Stendhal page 1 based on Levenshtein distance, for raw text. Darker colours represent greater distance from one text to an other. The phylogenetic representation is shown on the top and left sides.

b) XML analysis

After performing our distance measurement technique on text, we followed up with an analysis of the XML. The results that we obtained are shown in Figure 11.5, which also shows the results obtained on text already seen in Figure 11.4. The two cluster graphs are shown alongside each other to facilitate comparison. Mainly, we note the discrepancy in the results. We see a reordering of the clusters of novice transcribers and experts' positions with respects to novices and one another. This is because we are now looking at the XML's structural properties.

We obtain a cluster with very similar results for novices. This cluster can then be divided further to see which novices are nearest to one another. Remarkably, we obtain a cluster that clearly separates our strongest transcribers from the other participants. If we look at the far right of the tree, and at the bottom-right corner of the matrix, we see a cluster that groups experts and the strongest novice, *novice 9*, separately from all other transcriptions. We have reason to think that it is knowledge of the XML document, with its structural and semantic properties that differentiates knowledgeable and inexperienced transcribers.

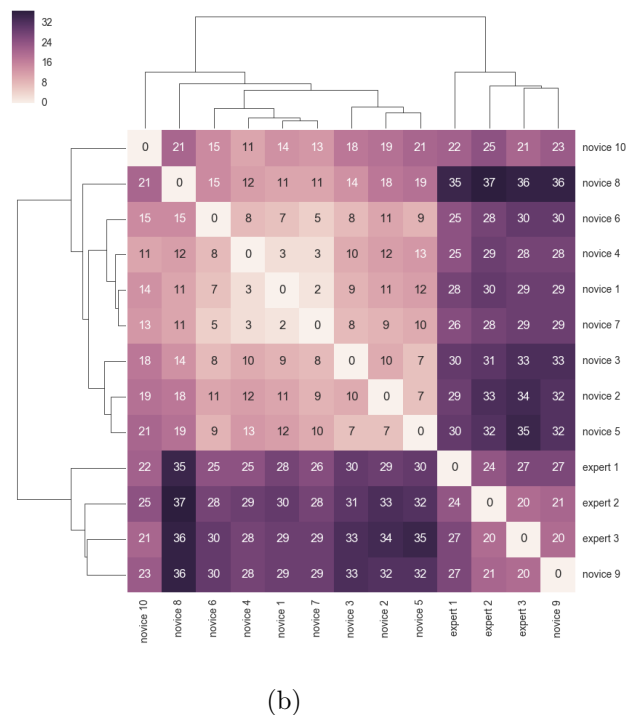
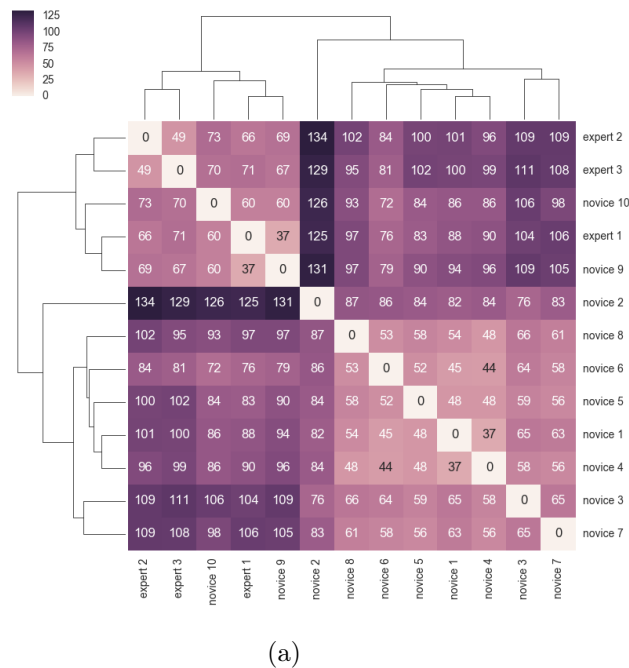


Figure 11.5 – Comparative of two distance matrices generated for the same set of transcriptions of Stendhal page 1, with (a) representing results obtained for raw text and (b) those for XML documents. As in preceding graphics, darker colours represent greater distance from one text to an other and the phylogenetic representation is shown on the top and left sides. Figure (b) reveals a telling cluster that includes our three experts and novice 9, the highest scoring transcriber that was considered as non-expert.

There are nevertheless some observable differences between expert transcriptions themselves. Therefore, it is necessary to look more closely at expert-produced XML to determine where dissimilarities occur.

The main differences we observed between expert transcriptions may be attributed to differential uses of elements *foliotation* (*folio annotation*), *marginal* (*marginalia*), and *pagination* (*pagination*)³.

The differences observed between experts and novices are more important. In particular, because although both experts and novices were given the same descriptive schema and the same transcription tool, in truth experts make more complete use of the vocabulary set at their disposal. They place elements and attributes that novices do not think of using, including both structural elements such as *texte* (*text*) and elements for scholarly or scientific commentary, *commentaire_scientifique*. On the contrary, as already stated, experts seldom place *douteux* (*doubtful*)⁴ in the text. The XML documents produced by experts are structurally more complete and also more complex.

With this information, we are persuaded that certain structural information is generally overlooked by novice transcribers. Moreover, with the likelihood that vocabulary use will remain limited to specific and unambiguous features, it may be sensible to ask less from contributors by providing a narrower descriptive vocabulary with fewer elements.

Since the majority of novice transcribers did not transcribe or identify marginalia, we ran our distance analysis again without them. More specifically, we removed all text and elements that pertained to marginalia, such as folio annotations, paginations, titles or subtitles, which were identified by experts, but not novices. We then ran a distance calculation and raw text clustering algorithm on the new set of files. We observed a drop in average error. Table 11.6 shows the results obtained on raw text. This table should be compared with the one obtained originally, shown in 11.4 on page 206.

3. Folio annotations refer to sheet numbering, paginations refer to page numbering, and marginalia are annotations in margins. However, to the casual observer, all three appear in the margins of pages.

4. These elements are used by transcribers to signal to others that their guess about a word or phrase should be verified as they are not certain to have transcribed it correctly.

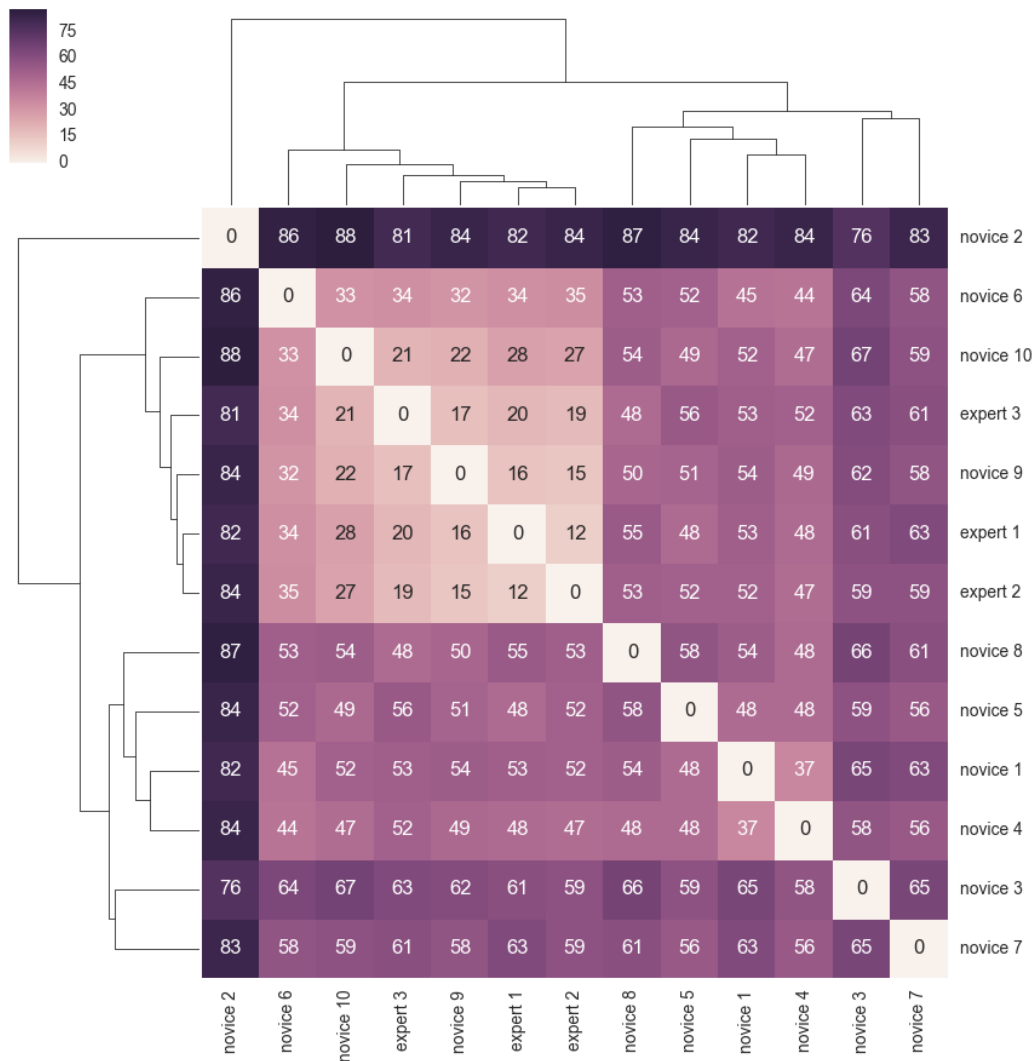


Figure 11.6 – Result of distance measurement on transcriptions for Stendhal's page 1, without marginalia; folio annotations, etcetera.

Again, we observed that our three expert transcriptions obtained distance measurements that placed them in close proximity to one another. Only this time our distance between experts ranges from 12 being the lowest and 20 being the highest. This is significantly different from the values and average we described in Section 11.2.3. Again, *novices 9* and *10* occupy the expert cluster, but this time they are also accompanied by *novices 6* and *3*. Average novice error drops from 93.497 to 48.565 *chars* and our best scoring novices see their distances to experts halved as well.

11.2.4 Digging into the data

When looking at the data, we remark several things. The first is related to the types of elements that novices annotate among options that they can choose from. To perform the activity, everyone received the same instructions and these are included in Annex A. When transcribers encode a transcription, their work involves deciding, depending on the content encountered, which elements correspond best to the content. For example, they place elements around words that have been added or striken-through. Novice transcribers successfully placed some of these elements, which we identified in our study samples. The following tables present information that was collected about expert and novice element use for our Stendhal sample. Table (11.2) shows the number of placements by element for our *experts 1, 2* and *3*. When looking at Table 11.3, column one lists the elements, column two shows the number of novices having comparable element placements to experts, and column three presents the success rate, expressed as percentage, with which these placements were correct as compared to experts. Finally, in column four, the information is relativized to account for our total number of novice participants, of which a considerable number did not make use of these elements. This information presented concerns only page 1 of our sample.

There is a discrepancy between experts on the *lieu (place)* element. Two out of three experts placed *lieu* elements twice and one placed it only once. We thus considered all novices that placed either 1 or 2 elements. Unfortunately, we could not attribute a percentage to these for two reasons. In the first case, the novice accurately placed

| Element | no. of placements for expert | | |
|------------------------|------------------------------|----|----|
| | 1 | 2 | 3 |
| ligne (line) | 21 | 21 | 22 |
| ajout (addition) | 0 | 1 | 1 |
| biffe (deletion) | 1 | 1 | 0 |
| date (date) | 1 | 1 | 1 |
| lieu (place) | 2 | 2 | 1 |
| douteux (doubtful) | 0 | 4 | 0 |
| exposant (exponent) | 3 | 3 | 2 |
| illisible (illegible) | 0 | 0 | 0 |
| interligne (interline) | 1 | 0 | 0 |
| souligne (underline) | 2 | 1 | 2 |

Table 11.2 – Number of placements of each of the listed elements by each of our three experts. Shown for purposes of comparison.

the first element, but inaccurately transcribed the text, and was not able to place the second element. In the second case, the first element was accurately placed and the text transcribed correctly, whereas the second element was identified correctly but placed on only one of the two words (*Palais* was identified instead of *Colonna* in the segment *Palais Colonna*).

We also notice that the *biffe* (*deletion*) element was not placed by our novice transcribers and only two out of 3 experts had indeed placed it. When looking at the page we remark that this element is notably difficult to spot, as it concerns an accent *grave* on the letter *a*, likely due to a spelling error.

All four correctly identified the date, but two of the four included an extra word or number as part of the date element. To be rigorous we considered that extra words or numbers, including those that could potentially belong to another element, constituted an error of element placement.

| Element | N° novices | Placement accuracy (%) | % of Total |
|------------------------|------------|------------------------|------------|
| ligne (line) | 1 | 100% | 10% |
| ajout (addition) | 2 | 100% | 20% |
| biffe (deletion) | 0 | 0% | 0% |
| date (date) | 4 | 50% | 20% |
| lieu (place) | 2 | 0% | 0% |
| douteux (doubtful) | 5 | - | - |
| exposant (exponent) | 3 | 50% | 15% |
| illisible (illegible) | 0 | 0% | 0% |
| interligne (interline) | 1 | 100% | 10% |
| souligne (underline) | 1 | 100% | 10% |

Table 11.3 – This table shows the N° of novices (out of 10 participants) having correctly placed elements as compared to experts. It also gives a corresponding accuracy measurement (expressed as %), as compared to experts, and presents this accuracy relative to the total number of participants . Resulting discrepancies are explained in the text.

Two out of ten novice transcribers identified a maximum of 2 *exposant* (*exponents*) (one of them identified only 1 *exposant* (*exponent*)), whereas two out of three experts identified 3 and one identified only 2. We noticed that two of the exponents which posed difficulty to novices (and one of our experts) concern marginalia. Meanwhile, both novices correctly identified the exponent present in the body of the text.

The number of *douteux* (*doubtful*) elements that were placed by our five novices ranges from a minimum of 2 and a maximum of 15. We note that five out of ten novices did not use this element and we cannot know with certainty if this is intentional or due to omittance. Also, taking into account the element's function (marking parts of a text that a transcriber is not certain to have read and transcribed correctly), it is not untenable that experts use this element less frequently than novices. However, because the element can be placed on any part of a text, we cannot use it as a gauge to measure accuracy

between experts and non experts as done with other elements in table 11.3. Nevertheless, we can perform a more extensive comparison based on the data collected because one of our experts placed 4 *douteux* (*doubtful*) elements. In the interest of finding similarities (if present) in what an experts and novices find difficult in this sample, we have compared our one expert’s element placements with our best scoring novice and the results are shown in Table 11.4. Column four shows our novice’s two placements corresponding closely to positions 3 and 4 used by our expert. These segments were later verified against other experts and are shown in the third column.

We noted that no other novice transcribed the marginalia for page 1. In fact, the majority of *douteux* (*doubtful*) elements were placed in the body of the text, suggesting that inexperienced transcribers encountered more difficulty in transcribing the core text and that this task itself was sufficiently challenging, regardless of finer details such as various marginalia present on the page.

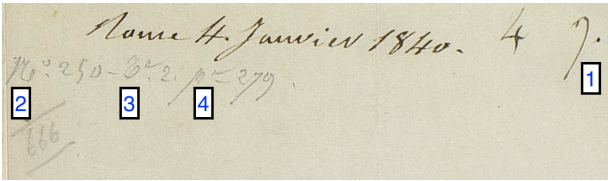
| N° | Doubted | Confirmed | Novice placed | Source |
|----|----------------|----------------|----------------|--|
| 1 | 7. | 7. | - |  |
| 2 | R.° | N.° | - | |
| 3 | T.e | T.e | 2 ^e | |
| 4 | p ^e | p ^e | n° | |

Table 11.4 – This table compares four text segments for which an expert placed *douteux* (*doubtful*) elements (column 2) with two novice-transcribed segments also identified as *douteux* (*doubtful*) (column 4). Expert confirmed segments are shown in column 4. Each of the four placements concerns marginalia.

11.2.5 Study of expert transcriptions

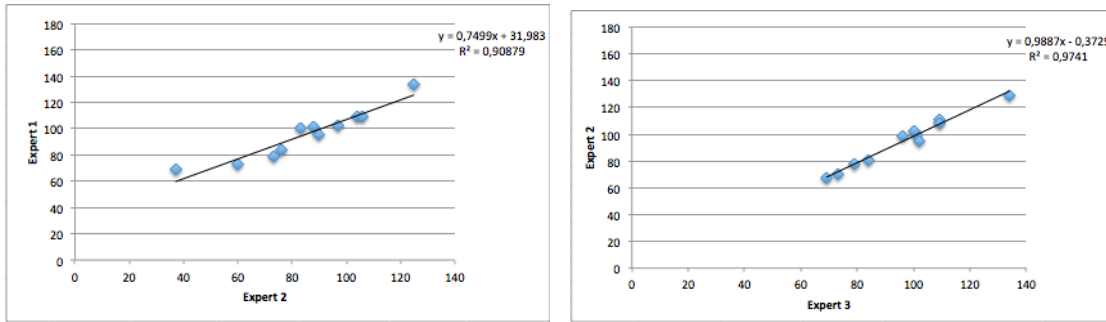
In the previous section we assume that experts can be used as references, and their transcriptions as targets, for novices. We will justify this assumption in the following explanation.

Merriam-Webster defines an expert as someone having knowledge or skill based on specific training or experience. We use this widely accepted definition, much like numerous other projects both within DH and other scholarly fields, as grounds to accept the authority of experienced transcribers when it comes to evaluating transcription quality. We have repeatedly stated that experts can be used as references because they produce transcriptions that correspond to expectations of quality. However, further investigation is necessary to demonstrate why this statement is valid. We use the data we acquired in our first experiment to demonstrate how, thanks to an expert transcription, we can easily identify other quality transcriptions.

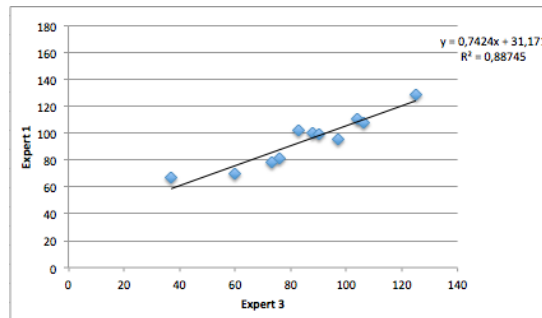
We are fortunate to have three experts that were able to provide us with transcriptions of the same pages, allowing us to observe how these three expert transcriptions measure in relation to one another. We observe a pattern that emerges in the results, which is founded on the minor differences we observed in their work, already discussed in Section 11.2.4. We can plot the results of our 10 novice transcriptions in relation to two of our experts on the basis of our observation that experts produced very similar transcriptions. Figure 11.7 demonstrates the pattern observed between experts. What we see between *experts 1* and *2*, *2* and *3*, and *1* and *3* is an apparent tendency toward a linear relationship between all novice transcriptions and any two given expert transcriptions. These figures were produced using linear regression analysis on the data obtained for page 1. For each figure, resulting trend lines describe a clean linear distribution.

Furthermore, having more than one expert allows to perform a linear regression because each individual can be used as an axis against which all novice transcriptions are plotted. In other words, if *expert 1* is y and *expert 2* is x we can plot the distance of each novice transcription relative to *expert 2* and *2*.

When we observed all our individuals, we were faced with a set of relative distances or points that were scattered at random. We were observing distances between all of our different transcriptions, without the slightest suggestion that the way in which the points were scattered could tell us more about the nature of the relationship between the transcriptions produced. Worst of all, is actually observing the results spread out



(a) Distribution as compared to experts 1 and 2 (b) Distribution as compared to experts 2 and 3



(c) Distribution as compared to experts 1 and 3

Figure 11.7 – Results of linear regression performed on three experts (a) (b) (c). Trend-lines indicate highly linear distributions for all three references and a relatively high R^2 , ranging from $R^2 = 0.88$ to $R^2 = 0.97$.

over two large clusters. If two clusters form, and within each we have transcriptions that demonstrate smaller and greater distances to one another, how can we possibly know which cluster contains the results we want? Indeed, when comparing novice transcriptions against other novice transcriptions, the distributions of distance values observed can be puzzling. Figure 11.8 shows what this type of result looks like for page 1.

If, however, we pair an expert and novice and observe a pattern similar to one produced when two experts are paired, we can deduce that within our selected expert-novice pair resides a novice who produces excellent work. In other words, work that correlates closely to that of an expert. This example is illustrated by Figure 11.9.

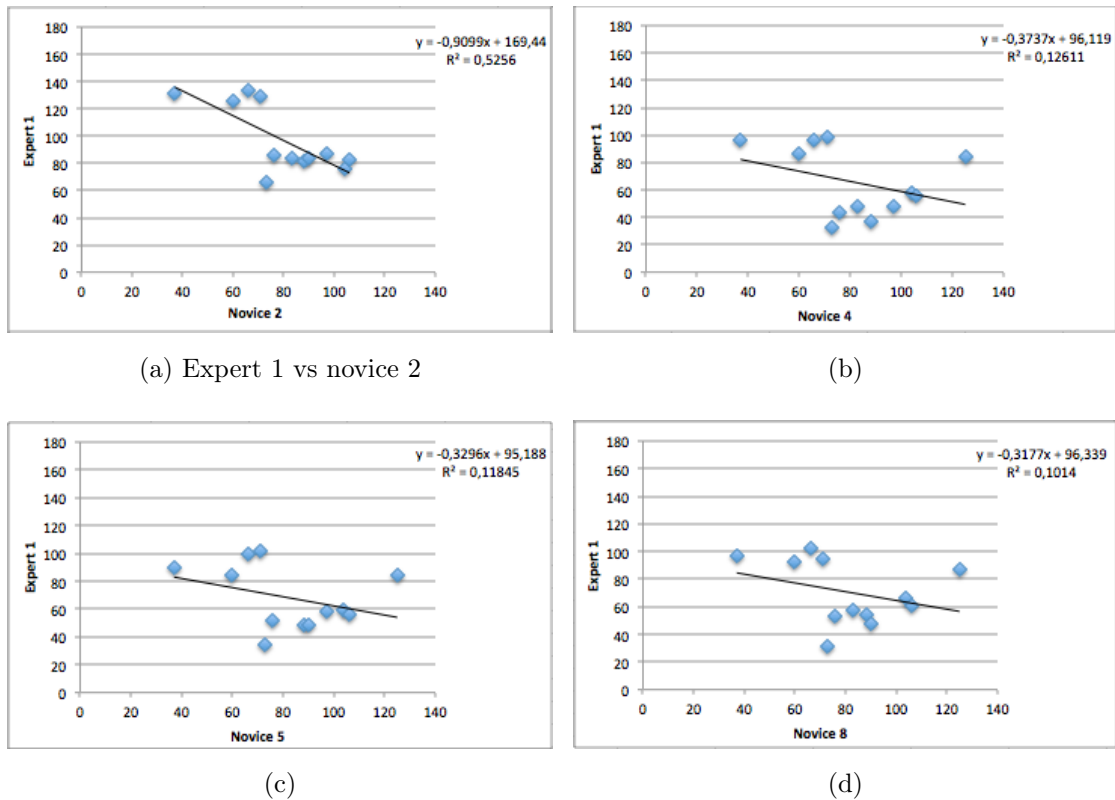


Figure 11.8 – Results of linear regression performed on expert 1 paired with (a) novice 2, (b) novice 4, (c) novice 5, and (d) novice 8. As opposed to results based on pairs of experts, these distributions indicate poor linear relationships and a low R^2 , with a range from $R^2 = 0.10$ to $R^2 = 0.52$.

With an expert reference in hand, we can perform this type of analysis on a series of incoming transcriptions and should be able to identify transcriptions that stand out from the patchwork. Even without any closer analysis of the work, project leaders can use this method to select transcriptions that contain fewest errors and in theory can be validated without major modifications. If individuals regularly obtain this type of result on expert-referenced transcriptions then perhaps their contributions can be considered more reliable and some form of recognition can be dispensed by the system and its administrators. Working linear regression analysis into the system could be used to identify contributors that consistently achieve highly linear correlations when paired with an existing expert reference. Perhaps these individuals can even be considered as new experts themselves?

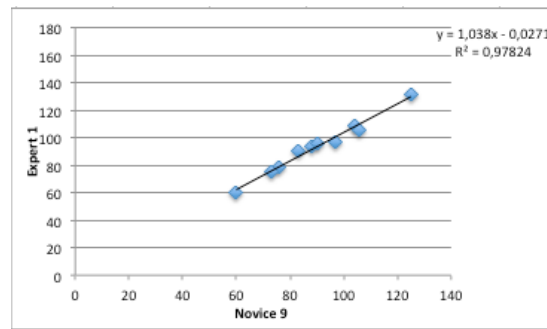


Figure 11.9 – Results of linear regression performed with expert 1 and novice 9 as reference. All points are novice transcriptions. The trendline indicates a highly linear relationship, with an $R^2 = 0.978$. Results are very similar to those obtained for a pair of experts and indicate Novice 9’s proximity to the references.

These are just a few of the possible outcomes of this type of analysis. The following steps would be to consider how to put in place an procedure to promote users based on outstanding work.

Likewise, a closer look at the transcriptions themselves, to see how these distinguished transcribers differ from others, may be helpful in implementing useful feedback, instructions, and FAQs. Motivated novice transcribers will then be able to use these to improve their results, gain confidence, and produce work with a higher degree of reliability.

11.2.6 Observing effects of page variation and complexity

Our first intuitive observations of pages of different degrees of complexity suggested that differences in pages may affect transcription results. We inferred page 2 to be more difficult than page 1 based on the reasoning described in Section 11.2.1 and this was confirmed by the results obtained on page 2. As a matter of fact we observed a greater dispersion among novices for the second page, which confirmed our thinking that it presented a more complex transcription task than the first. The results obtained on text and XML for page 2 can be found in Annex B.1, in Figure B.2. The distance values observed for raw text are significantly higher as the results of novices are more dispersed. XML distances are relatively similar to page 1, suggesting that the results concerning encoding

are comparable. Measuring text becomes a good indicator of variability because the greatest differences in text are often caused by the fact that although transcribers encode the same features, they do not place XML elements in the same order. Then, when the files are parsed as text, encoded sequences are not restituted in the same order, producing large discrepancies even though the text transcribed is actually the same. To minimize this phenomenon, project leaders can include indications of the order in which elements should be placed in their instructions. For instance, *correction* elements should always be placed after *deletion* elements as we consider that this accurately reconstitutes the chronological order in which the writing was modified. Then, if contributors follow these indications, these kinds of discrepancies will likely diminish.

For this experiment, we concentrated simply on the observation of contributing factors of page complexity. In Chapter 12, Section 12.1.2 we implement a design plan for studying possible correlation between factors of page complexity and resulting transcriptions.

11.2.7 Drawing preliminary conclusions

Based on the small group of contributors that we were able to solicit for this experiment, we were able to conduct a preliminary test case for crowdsourcing transcriptions from novices. We made our primary observations concerning page complexity and, also, our primary assumptions regarding how more difficult pages may affect the quality of contributions. These were confirmed in the differences in results obtained for pages 1 and 2. On these premises we were also able to consider other fundamental conditions that would need to be met in order to perform subsequent tests.

Firstly, we should consider the importance of having our expert transcriptions. These were essential to establishing a reference against which to evaluate novice work, and played the role of *gold standard* data [Oleson et al., 2011]. This first experiment allowed us to obtain data to affirm that our expert group produces transcriptions that have a verifiable degree of similarity. We consider this to be the case due to the low *Distance* variation observed. Furthermore, the fact that we observed few variations for both pages, regardless

of a change in level of difficulty, also suggests that using experts as a reference would be a reproducible measure if further experiments on this corpus were conducted.

Also, we were able to observe how novices performed on XML element placement as compared to experts. This gives us an idea of how well an untrained novice who only receives basic instructions performs a transcription for which there are clearly defined expectations.

Our results tend to suggest that projects can use simpler descriptive schemas for work intended for inexperienced transcribers. Notably, we observed improvements in results when more ambiguous or subtle elements were disregarded during analysis.

In conclusion, the benefits of this experiment can be summed up in the following list:

1. Expert transcriptions are essential for establishing a reference, or target transcription, and a reproducible measure of comparison with other transcriptions.
2. Experts produce transcriptions that have a high degree of similarity, or low edit distance.
3. Untrained and inexperienced transcribers who receive only instructions can be used as participants.
4. Simpler descriptive schemas reduce variability both between experts and inexperienced transcribers.
5. Having roughly 10 participants for a given page creates a situation where results are well distributed, ranging from very close to very far away from our target transcriptions.
6. It would be significantly easier to repeat this experiment if participants have access to an online transcription platform.

In this experiment we observed that inexperienced transcribers' results were spread out and rather disparate, suggesting a high variability in potential results. This variability can be explored further and to do so we must look at different factors that may influence transcription results.

11.3 Experimentation on the Benoîte Groult Corpus

Journalist, novelist, and militant feminist, Benoîte Groult has left behind a collection of drafts of her autobiography, *Mon Evasion*, among other papers. These are currently conserved at the Feminist Archives Center (Centre d'Archives du Feminisme - CAF) at the University of Angers. This small collection of digitized pages amounts to a little over 450 pages.

From a general point of view, Benoîte Groult's handwriting can be described as being uniformly legible. Nevertheless, it may pose some specific challenges for transcribers. These include interpreting the variable manner in which she accentuates letters, which by no means conforms precisely to schoolbook examples. However, her writing habits are sufficiently repetitive so that a keen observer should be able recognize patterns and use them to deduce the intended accentuated character. An example of a typical page, and one that was used for our second experiment, is shown in Figure 11.10. With Benoîte Groult's writing, it is safer for a novice transcriber to rely on one's knowledge of French spelling rules to disambiguate words. We therefore consider that to perform these tasks with accuracy transcribers must have good and extensive knowledge of french spelling rules. Nevertheless, enthusiastic readers of the french language will have what it takes to transcribe this corpus.

11.3.1 Benoîte Groult Experiment 1

Our first experiments produced XML files whose structures required a significant amount of manual correction. This was directly related to TinyMCE's configuration. At first glance, the transcription work appeared to have been in vain. However, we were actually squarely in the middle of a prototyping phase, and anticipating improvements. Structural errors that were observed in the files allowed us to understand which changes were necessary to obtain desirable results.

Indeed we could not use this data for quality analysis on the XML produced, a disappointing realization initially. Nevertheless, further examination led to the realization

that structural problems in the XML were not extended to the text produced. Since our first experiments on Stendhal’s corpus simply measured distance between texts (and XML tags were filtered), we were able to use these text results to perform a partial analysis (without performing an XML element count or analysing placement).

Between February 2016 and April 2017, we obtained a total of 203 transcriptions, for which we only had 2 expert-validated transcriptions. We recall that PHuN 2.0 uses a validation process, and since experts were involved, they revised several of the pages themselves. For each of the expert references we were only able to obtain 5 novice transcriptions.

We applied the same method of analysis and generated distance matrices and phylogenetic trees, which are presented in Annex B. Like in previous cases, we were able to observe the formation of clusters. The results can also be attributed to the fact that all transcriptions were complete and not partial. Had we included partially completed files we would surely observe greater distribution in the data. In both cases, distances were less significant because files were created as a result of subsequent editing from multiple users and not independently. The benchmark file that we use as expert reference is actually the file that was validated in the system.

We also use our 5 novice transcriptions to create a median transcription, which we refer to as an aggregate transcription. The median or aggregate transcription is created by weighing the text produced from multiple transcriptions, in our case 5 of these, and writing the text that appears most frequently in the transcriptions. Due to the fact of having fewer participants, we did not observe distributions that could be considered representative of the kinds of results that can be expected and we did not pursue further analysis of these documents. Nonetheless, this experiment allowed us to consolidate the analysis workflow, identify necessary improvements to the platform, and plan subsequent experimentations that would be executed using the PHuN-ET platform. What follows is an account of the instructions used for this phase of experimentation.

a) Instructions

Transcription instructions for transcribing the Benoîte Groult corpus were intended to serve as a step-by-step manual. The protocol used was elaborated in french and descriptive images were included to enhance understanding. The instructions were validated by an expert before being uploaded to the platform. The instructions were used as a resource by participating transcribers, but we had no control mechanism in place to know whether transcribers consulted these instructions or whether they started to transcribe based on intuitive personal understandings of the activity of manuscript transcription or other previous knowledge. Since participants were not accompanied during the activity we could not observe them nor obtain their reactions to instructions. Also, having few and infrequent participants we were not able to gather user experience nor impact of instructions on the documents produced. Future studies geared at acquiring more relevant data on user experience of protocol would be beneficial to developing further understanding on this topic. A possible method would be to constitute multiple groups and provide variants on protocol to observe how different types of protocols may impact users' understandings and results. The instructional slides that were used are included in Annex A.

b) Conclusions

Based on the overall experiment we initially thought that the files created by novice transcribers collected in this experiment would need an important number of structural corrections, without which these transcriptions should not be validated for publishing. Furthermore, we have low confidence in that novice transcriptions can be validated without some form of expert review and correction. The reason for this is simple: it is our experts that have the most pertinent knowledge of the project's descriptive schema.

As a result of this experiment we were able to identify improvements that were necessary at the level of our TinyMCE editor. Notably, in future experiments these improvements would ensure that we are able to avoid unwanted structural changes to the XML. These changes most often concerned child inline elements being rejected from their parent

elements because they had not been declared within TinyMCE's configuration file⁵

We also noted the importance of instructions in starting novice transcribers on a transcription task. Unfortunately it was not possible to observe the effects of protocol changes on the results obtained from novice transcriptions. In all the cases that were observed, novices reproduced the texts in their entirety.

11.3.2 Benoîte Groult Experiment 2

Our second experiment on the Groult corpus involved two groups of voluntary participants. This experiment recounts three different types of information that we were able to gather. This includes information that was gathered concerning users' perceptions of transcription, information regarding transcription instructions and how to improve them, and finally data from the transcriptions performed, which we analyse using our methodology. Finally, with the help of user feedback we were also able to implement improvements to three specific areas of the platform. One of these includes the implementation of a new image zoom plugin, based on the same tool used by Google Maps and which solved two specific issues observed by our users. The second is a minor bug fix concerning page order, and finally the third improvement concerns the XML schema itself and its visual representation (CSS rules) in the user interface.

5. The configurable editor is presented in more detail in Chapter 7, Section 7.4. Changes to elements occurred at the time of writing data to files. These changes were responsible for differences observed between the moment a document was edited and after it had been saved. Most frequently, affected elements were in-paragraph or in-line elements such as *rature*, *ajout_en_interligne* and others. The error produced unwanted new lines, throwing undeclared elements and their text contents onto separate lines from the rest of the original text. Often, the error would result in three lines: the first containing the text before the element, the second containing the element itself and the third any trailing or remaining text that came after the element.

a) Workshop Context

Two workshops were organized in May of 2017 and new transcriptions were collected during the two sessions. Our public consisted of librarians, students at the master and PhD level, as well as researchers in literature and social sciences. All participants were new to our expert-defined XML vocabulary and were given access to concisely formulated instructions. These were prepared for the workshop and are shown in Annex B.3.

The transcription workshops were organised in one of the computer rooms of the university library. This way our participants were sure to have access to a desktop computer if they didn't have a personal laptop. The sessions lasted approximately two hours each, during which time our group of novice transcribers engaged in a transcription activity and replied to survey questions. Everyone was given access to the same instructions and we noted users' reactions to instructions to anticipate future improvements.

The workshop context is, in itself, specific. It puts participants in situations that are not necessarily equivalent to conditions under which crowdsourcing commonly takes place. During a workshop, contributors are in the same room and can interact with one another, and with the workshop facilitator, more directly. There are also obvious time constraints that are not the same as those that participants deal with independently; there is a time frame allotted to activities. The workshop may have a schedule and the facilitators may intervene at regular intervals, which may affect or alter participants' attention. Nevertheless, workshops are an effective way of introducing activities that are unfamiliar, creating interest in new projects, and ultimately, finding new participants.

A list of survey questions was elaborated. The questions focused on participants' perceptions of the task. Notably, we were interested in how participants viewed transcription in a crowd context. Also, since the task required volunteering time, participants' responses were valuable in gathering information of how they viewed their work with respects to factors such as time management. Other questions asked them to consider the social benefits of contributing transcriptions and whether they felt that what they were doing was helpful or useful. Finally, they were asked if they would consider sharing their completed

work with others and if they would tell others about the platform.

The responses were informative. Users appeared concerned about the quality of their work. They were able to estimate the time required for transcribing a page and based this on their own personal knowledge or experience. Many responses revealed that participants did not necessarily see transcription as a social task. Nor, one that they would share with their circles. This confirmed that, like many activities, transcription is seen as one demanding a high degree of intrinsic motivation from participants [Dunn and Hedges, 2012 ; Franzoni and Sauermann, 2014 ; Ryan and Deci, 2000].

b) Gathering data

Having users perform transcription activities while in a workshop setting allowed us to gather data while users were using the transcription tool. Our intention was to observe users in this situation, solicit and record their reactions to the activity we proposed as well as the tools and instructions we presented. We were also able to observe difficulties they encountered, which allowed us to identify future improvements to the system as well as gain a better understanding of how users perceive the task.

c) User instructions and reactions

Instructions were informed by the nature of our experts' XML schema. The schema called for global elements that would identify either a manuscript (*manuscrit*) or printed text (*imprimé*), within which one could identify other features, which are markers of modifications to the text and which may be attributed tool elements to identify colours (for example: *stylo_bleu*, *stylo_noir*, *crayon* for blue pen, black pen, crayon). If a manuscript was written in black pen then these two elements (*manuscrit*, *stylo_noir*) would be wrapped around the text. Corrections inserted in blue pen, called additions *ajouts*, would be encapsulated by the two elements *ajout*(child element) and *stylo_bleu* (parent element).

For this to work at the level of the existing user interface, it was necessary to use a

rather inflexible formula or otherwise very procedural instructions that were to be followed as a list. The list not being very long, it contained a total of 5 steps (excluding the final step of pressing the *Enregistrer et Fermer* or Save button. The instructions used can be seen Annex A, A.3. Everyone attending the workshop was given the same instructions and the same explanations.

Prevailing reactions from participants confirmed our impression that many of the pages in this collection are very easy and accessible. Even though the writing is dense it remains highly legible and there are few marked changes to passages or words. Transcribers' attention is mainly occupied by correctly reproducing the intended document structure, as well as noting changes in ink colour, which may be attributed to successive drafts. Even though pages are dense, for the most part the author's handwriting remains uniform. We cannot be certain as to which aspects of the manuscripts may be more or less challenging for transcribers.

A short video also accompanied the instructions, intended as a visual aid for users, it showed the beginning of a transcription and basic gestures such as placing the cursor, typing the text and selecting it to attribute chosen elements from the toolbar.

d) Analysing results

We analysed the data collected in the same way as was done for the Stendhal experiment. Figure 11.11 shows a crosscut of the matrix and phylogenetic tree obtained on transcribed text. Figure 11.12 shows a crosscut of the matrix and phylogenetic tree obtained on XML documents. The full figures can be found in Annex B.3.

When looking at Figure 11.11, which depicts our text analysis, we observe three main clusters. One which contains the expert reference and five novices. Of these, *novice 20* was a participant in the second workshop. A second cluster which contains the majority of participants of both workshops, as well as other contributors to the platform. This cluster can be broken down into many other densely packed clusters. Novices sharing a cluster with the expert occupy a cluster that is quantifiably less dense than the all-novice

clusters. This is because even though novices achieve good scores in relation to experts, they still bear more similarities to other novices, explaining their location.

For the XML, we can see the formation of four main clusters, as shown in Figure 11.12, and the large central cluster can also be divided into many smaller clusters. We observe that the expert transcription is contained within a small cluster of only three leaves. As for the other clusters, we observe a relatively even distribution between participants of the first workshop, the second workshop, and contributions that were collected ulteriorly. To us, this suggests that there was not a significant difference between sessions. Both sessions had individuals that scored very closely to the reference; for instance, *novices 9* and *15* for the first session, and *novices 21* and *17* for the second.

11.4 Conclusion

The assessment of quality is of primordial importance for projects eager to obtain the best from non-expert transcribers. As seen in Chapter 10, the advantage of having gold reference data allows to enable screening, ranking and, possibly, feedback. Another advantage is to be able to measure the effects of a changes in the way the contributors are prepared for tasks. For example, we would be able to answer questions such as, "can changes to instructions affect the quality observed?"

In the experimentation on Stendhal, we had already assessed if transcriptions made by experts could be used as references. We concluded that even though they can make mistakes, we can still use them as gold references. We then observed, thanks to hierarchical clustering, that people who received training for the task of transcription obtained results that were closer to experts. We also found that certain structural information is generally overlooked by novice transcribers who do not have the knowledge and experience with specific collections to recognize the intended meanings of certain features. Or, if they should be associated with particular elements even if they are available in the editor. Thus, it is an indication for project leaders to strive to include essential elements only, and limit schema complexity. As such, crowdsourcing can be used as a way to obtain

partially enriched data, which can subsequently be enhanced by experts at later stages in the editorial process.

For Benoîte Groult, we found that the descriptive vocabulary aimed at novice transcribers was not the same as the one used by experts, but that novices consistently recognized and used appropriate elements to encode structural features, such as *titles* (*titres*) and *paginations*, modification derived features, such as *additions* or (*ajouts*) and *deletions* or (*raures*), and even information related to changes to writing tools based on color, such as (*stylo_bleu*, *stylo_noir*, *etc.*). Even though experts used richer vocabularies to which our novices did not have access for the exercise, it was simple to perform some batch processing to normalize elements accross files so that we were able to compare them on the basis of, firstly, the text that was transcribed, and secondly, the XML descriptive vocabulary used to encode and structure the documents. This allowed us to accurately determine whether novice transcribers recognized the main visible features in authors drafts and were able to encode them accordingly.

Instructions for this type of activity play an important role for users in outlining procedures and expectations, as well as providing useful advice. Nevertheless, instructions should be concise and to the point, so as not to confuse or discourage potential participants. Conversely, they can also be designed as a way to screen participants.

We found that organizing a workshop was helpful to attracting participants. However, beyond the two sessions that were organized we cannot be sure that participants will continue to return to the site, without some further incentive or encouragement, as discussed in Chapter 9. The contributions that were received outside workshop settings seem to indicate that, provided instructions, participants are capable of working independently. Our results show that transcriptions contributed by those who did not attend the workshops were still roughly competitive with those provided by workshop attendees; *novices* 7 and 8, for example, are only 2 and 3 points behind *novices* 15 and 9. Given the particular conditions under which crowdsourcing often takes place, participants are often remote. For these reasons, particular attention should be paid to ensuring participants' understanding of the task from the instructions provided online. Finally, workshops should be

seen as a way of stimulating interest in new participants, proposing additional support, and for purposes of community management.

Experimentations were also vital for determining necessary improvements to tools, environments, processes, instructions, and descriptive schemas themselves.

The method described in this chapter can be implemented into a crowdsourcing system for two applications. Firstly, to allow faster and automatic organization of multiple contributions. Project leaders can use this method as a sorting tool to help them determine which transcriptions to edit or review, for example. Multiple transcriptions can also be used to correct errors, by relying on commonly agreed upon content, in a way similar to Galaxy Zoo [Franzoni and Sauermann, 2014].

Secondly, if projects can provide a number of expert transcriptions at the outset of a project, then these references can be used in the same way as *gold standard* data [Oleson et al., 2011]. This can help determine which transcribers consistently contribute according to project expectations and identify trusted contributors. Then, a system can be developed to reward or promote these individuals to tasks having more responsibilities and, ultimately, help develop more experts to support the editorial process.

To put this type of evaluation into practice, it is important to be able to implement it on two levels; taking into account both textual accuracy, as well as XML markup.

In the following chapter we look at some factors of complexity that can affect the quality of work contributed by inexperienced users. We also choose to focus on two factors relating specifically to page complexity. We design an experiment according to a Design of Experiments (DOE) method to determine their effects on transcription results.

CHAPITRE VII
CHER PAUL

p. 1

"Il est totalement inutile que les
femmes écrivent leurs inepties. Cela
ne fait qu'embrouiller les choses les
plus claires."

Strindberg

J'ai 30 ans et je n'écris toujours pas
d'inepties, mon vieux Strindberg. Je n'y saug même
pas. Mais j'ai repris mes mauvaises habitudes
et recommencé à tenir mon journal. Je note
tout, en me disant qu'un jour peut-être... Mais pour
l'instant, je savorise ma liberté. Enfin... une
liberté ornée de deux petites filles qui ont 1 an
et 2 ans et demie. Une blonde et une brune.
Et le matin par mon travail à la Radio : j'écris
(6 heures par jour) les bulletins d'information diffusés toutes les heures
sur Paris Inter, l'Antenne pépère... j'écris même...
Mais j'habite à nouveau (mon petit appartement
de la rue Reynoard). J'ai pu raccrocher les
portraits de Pierre dans mon alcôve et je garde une photo de
Georges sur la bibliothèque, à cause de mes filles, me
dis-je, mais c'est surtout parce que je suis encore sensible
à sa séduction. Chaque fois qu'on nous sommes
revenus au cours des années j'ai été tentée pendant les
premières minutes de retomber amoureuse. Comment
avons-nous réussi à ne jamais être heureux ensemble?
Le # qui est-ce la faute? J'espère, mon Georges, que tu
auras trouvé le bonheur dans ton mariage éternel
mais j'en envoie rien su : tu ne m'as jamais livré la
moindre clé pour ta précieuse personne... ou bien je
n'ai jamais su m'en servir.

Pourtant je n'avais rien à te reprocher que
d'avoir été un homme comme les autres. Mon mariage
ne m'avait pas paru une prison, il ressemblait à la
(la plupart)

Figure 11.10 – Page from the Benoîte Groult Corpus. Conserved at the university library of Angers.

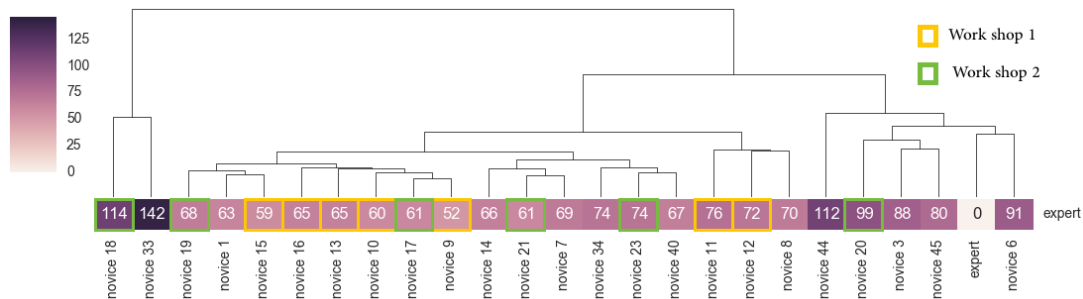


Figure 11.11 – Phylogenetic tree based on text distance. The row shows the distance of participants as compared to the expert and the contoured squares identify workshop participants.

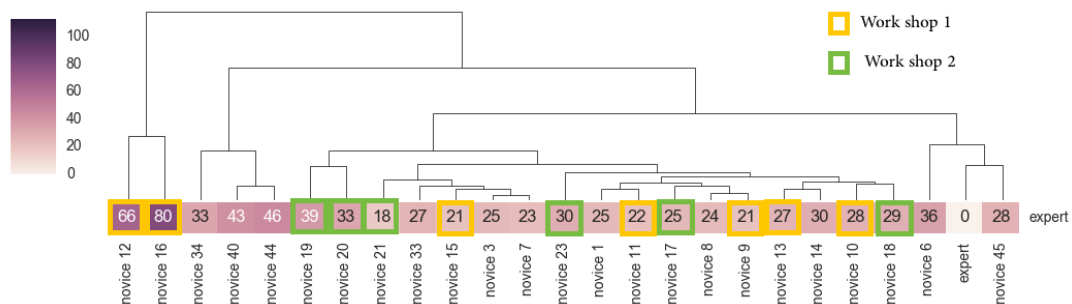


Figure 11.12 – Phylogenetic tree based on XML distance. The row shows the distance of participants as compared to the expert and the contoured squares identify workshop participants.

Chapter 12

Measuring factors of complexity

Contents

| | |
|---|------------|
| 12.1 Introduction | 234 |
| 12.1.1 Design of Experiments (DOE) | 237 |
| 12.1.2 Benoîte Groult Experiment 3 | 239 |
| 12.2 Data analysis and results | 242 |
| 12.2.1 Discussion | 244 |
| 12.3 Conclusion | 247 |

In this chapter we focus on exploring factors that could affect results of crowdsourced transcriptions.

12.1 Introduction

Over the course of our work and experimentation we encountered factors, which we refer to as factors of complexity, and which may be attributed to different areas related to the activity of transcription. Notably, to describe factors which contribute to render a transcription task more or less complex, three categories or families of factors have been identified. Each of these comprising a group of factors, and each contributing in some way to the complexity of a transcription task. Figure 12.1 introduces these three groups and we will describe each in more detail in the following paragraphs.

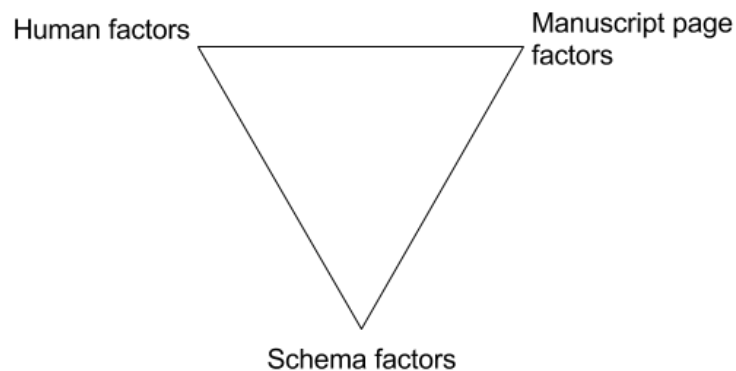


Figure 12.1 – Three groups of complexity factors: human, manuscript, and schema related.

The first family of factors concerns transcribers themselves. Factors relating to individual transcribers can include the following: age, experience with transcription, experience with manuscripts, experience in XML editing, experience using online editors, french language competence, motivation, free time, professional or educational background, work attitude, and other factors that can be described as belonging to social and socio-cultural, socio-economical and psychological spheres. However, studying these factors with respect to individuals, and in particular in an online context, we are quickly subject to ethical, moral and legal issues. Having been advised against these types of studies for the purposes of our work we thus decided not to pursue this path of inquiry, which requires considering much more sophisticated methods of study and having access to much larger groups of

participants. Furthermore, in a study that focuses on the impact of crowdsourced transcriptions within DH manuscript transcription projects, these types of studies focusing on individuals are indeed auxiliary to the questions that we are concerned with. It would indeed be interesting to researchers to also use crowdsourcing environments for the basis of more sophisticated studies, such as those involving cohort analysis. However, all and any such studies should be constructed within an appropriate ethical and legal framework and preferably be carried out on large population samples and over long periods of time (decades) [Glenn, 2005].

The second family of factors belongs to the XML schema or vocabulary used to transcribe a manuscript. As we have seen in Section 7.2 on page 125, when projects can create their own transcription vocabularies, one can imagine their potential for more thorough description. However, a large and subtly nuanced vocabulary, as well as the ways in which XML elements are intended to interact with one another, can be sources of complexity. Factors that contribute to overall complexity of transcription belonging to this axis can include, but are not limited to the number of elements used and the intended hierarchical relationships between these elements. Other factors can be identified, but these are likely to actually be determined by projects themselves. We therefore consider that the second family is directed, if not by an XML standard, then by a project-based XML schema, and thus by a project itself.

The third relates to the manuscript page itself. Observation of a manuscript allows to identify a certain number of factors, which, rather than being qualitative and subjective, can be described objectively. Although the third family of factors is a component for evaluating the overall complexity of a transcription task, alone it can be used to evaluate the complexity of a manuscript page. Furthermore, we can much more easily affect quantitative measures to identified factors, allowing for more objective evaluation of transcription results in relation to manuscript page complexity.

Over the course of our observations of the pages of Stendhal, Benoîte Groult, Jean-Philippe Toussaint and even (although in a lesser way) Michel Butor, we identified a certain number of factors. A number of these are related to the size and the inclination

of the script and also the use of special characters. Others are linked to the types of writing (manuscript or print), the tools used to write (pencil, pen or marker). We also have factors such as the number of additions and deletions (that contribute to decreased decipherability). Finally we can also note the presence of figures, which, as non-linear elements, can break up lines and add to the difficulty of transcribing a text.

To gain a better understanding of how these factors may affect the act of reading, deciphering or transcribing, more detailed explanations of each are necessary. Based on our observations of pages from three different corpora, we were able to identify nine factors, which we describe in the following list.

1. **Number of lines:** We expect the number of errors to increase with the number of lines the page contains.
2. **Number of additions:** We have observed that additions are often smaller than the main text, resulting in decreased readability.
3. **Number of deletions:** Deleted or crossed out writing can be difficult to read for an inexperienced user, particularly because of added pen strokes that conceal the original letters.
4. **Number of writing tools used:** Some tools greatly increase the difficulty to read a page such as a marker or crayon.
5. **Number of types of writing (manuscript or print) :** A printed page will be easier to transcribe than a handwritten one.
6. **Size of script:** One can expect that the smaller the writing is, the harder it is to decipher and transcribe accurately.
7. **Angle of inclination of script:** Like size, inclination can be a factor that affects the readability of a page.
8. **Number of figures present:** Figures being non-linear elements, they increase the difficulty of transcription since they may affect the orientation of writing, including word placement or the way that words wrap around figures.
9. **Number of special characters present:** Special characters including mathemat-

ical symbols or short-hand symbols adopted by a particular author aren't always directly accessible in users' keyboards making it difficult for these elements to be transcribed.

Identification and study of these factors to describe the pages of a given corpus presented an interesting opportunity for analysis as we considered the possibility of correlation between transcription results and page complexity.

12.1.1 Design of Experiments (DOE)

Having been introduced to experiment design, we understood that the greater the number of factors to be studied the greater the number of experiments required. For example, to design an experiment based on full factorial design, that is taking into account all of our identified factors, it would be necessary to use an approach known as 2^k , introduced by R.A. Fischer [Fisher, 1937]. Using this approach, the number of factors, represented by k , would give us a total number of experiments, on the basis of 2^k . With nine factors our resolution is based on 2^9 number of experiments, or 512.

There can also be factors stemming from the interactions between the n number of factors in the two or even three of the families. An interaction can be something like a transcriber's understanding of an XML schema, which depends on multiple competences of the transcriber him or herself in relation to an XML schema that he or she is unfamiliar with. Into the mix, can be thrown individual auto-evaluation scales that are subject to high degrees of subjectivity, thus increasing the likelihood of unpredictable results.

To give an idea of how to represent this problem, we derive a formula to measure the overall complexity of a transcription task, which we may identify as C_{total} . Let us consider that this total can be obtained from the sum of all identified and observed factors belonging to each of the three families we described. We consider also that there are factors that we cannot identify, nor can we predict all probable interactions between

them. Nonetheless, abstractly, our formula would look like this:

$$C_{total}(x_1, \dots, x_k) = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^{k-1} \sum_{j>i}^{k-1} \beta_{i,j} x_i x_j \quad (12.1)$$

Here, x_i represents a factor of complexity, belonging to any one of the three families that we listed. β_i is the weight associated with a factor x_i , whereas β_0 represents a constant value of the model. $\beta_{i,j}$ is the weight of the interaction between two factors x_i and x_j . The purpose of the design of experiments is to find the value of all unknowns in the model (this means β_0 as well as all β_i and $\beta_{i,j}$) using a minimum number of experiments. The model takes into account all factors x_i and all possible interactions $x_i x_j$. Having said this, considering the substantial number of factors we have identified, a comprehensive study including all factors with the goal of obtaining C_{total} for the purposes of measuring task complexity is unrealistic.

In order to explain why, let us imagine that we identified 9 factors for each of our three categories: human-related factors, page-related factors, and schema-related factors. This would give us a total of 27 factors. With 2^k or 2^{27} we'd be looking at a complete design plan requiring more than 134 million¹ experiments. With over 134 million x 10 (number of participants), we simply do not have that many participants at our disposal. Nor are we certain to have the necessary number of pages containing the appropriate combinations of characteristics. Finally, it would be a challenge to find a representative sample of project schemas to put in place this monumental study.

This said, even if we only consider the factors that we were able to identify here; 11 for transcribers, 9 for pages, and only 2 for project schemas, we would still be in for 2^{22} or more than 4 million² experiments. This is enough to understand that the problem in its entirety cannot be resolved here.

In order to approach the problem using what we have learned from experiment design, we have selected two factors to study and created a complete factorial plan, which we

1. precisely 134 217 728.

2. precisely 4 194 304

implemented on the Benoîte Groult corpus. We present this experiment in the following section

12.1.2 Benoîte Groult Experiment 3

The design plan that we created for our third experiment on pages from the Benoîte Groult corpus intends to observe the potential effects of two complexity factors, which relate to manuscript pages, on transcription results. Since our experiment focuses on two factors, we have a design plan of $2^2 = 4$ experiments. To realize our experiment we need to select 4 pages from the corpus, and each page should embody a particular combination of the two factors being studied. Our two chosen factors are:

- Number of modifications ($X1$): We consider a modification as any one of the following: addition, subtraction, correction.
- Script area (height x width in pixels) ($X2$): We calculate this based on the dimensions of the letter "e", considered as the most frequent letter in the French language.

Using the Yates³ order method ([Goupy and Creighton, 2013]) to organize a plan using our 2 chosen factors we can see the combination of conditions that our four pages need to meet in order to satisfy experimental conditions. Table 12.1 presents these combinations, where $X1$ represents the number of modifications, and $X2$ represents script area. This table presents the four possible combinations resulting from our two chosen factors, in which -1 represents a low value for the factor and 1 a high value. The maximal and minimal real values for each factor are also shown. In other words, we needed to find four types of pages:

1. A page containing few modifications and large writing
2. A page containing many modifications and large writing

3. Frank Yates, a statistician, designed a technique for ordering factors for experiment design to exploit all possible combinations of factors and derive the minimum number of experiments.

3. A page containing few modifications and small writing⁴
4. A page containing many modifications and small writing

| Experiment N° | N° of modifications X1 | Script Area: X2 |
|---------------|---------------------------|-----------------|
| 1 | -1 | -1 |
| 2 | 1 | -1 |
| 3 | -1 | 1 |
| 4 | 1 | 1 |
| Niveau -1 | 1 | 600 pixels |
| Niveau 0 | 15 | 1450 pixels |
| Niveau +1 | 30 | 2300 pixels |

Table 12.1 – Experiment design plan using two factors and also showing real values associated with factors' high and low levels.

a) Data collection

We analysed a random batch of pages to find four that matched the appropriate combinations. Figure 12.2 shows our complete factorial plan based on 2 factors, with each of the four pages corresponding to one of four possible combinations.

4. It being a real challenge to find such cases in this corpus, we took measurements from modifications, which does present cases of very small writing.

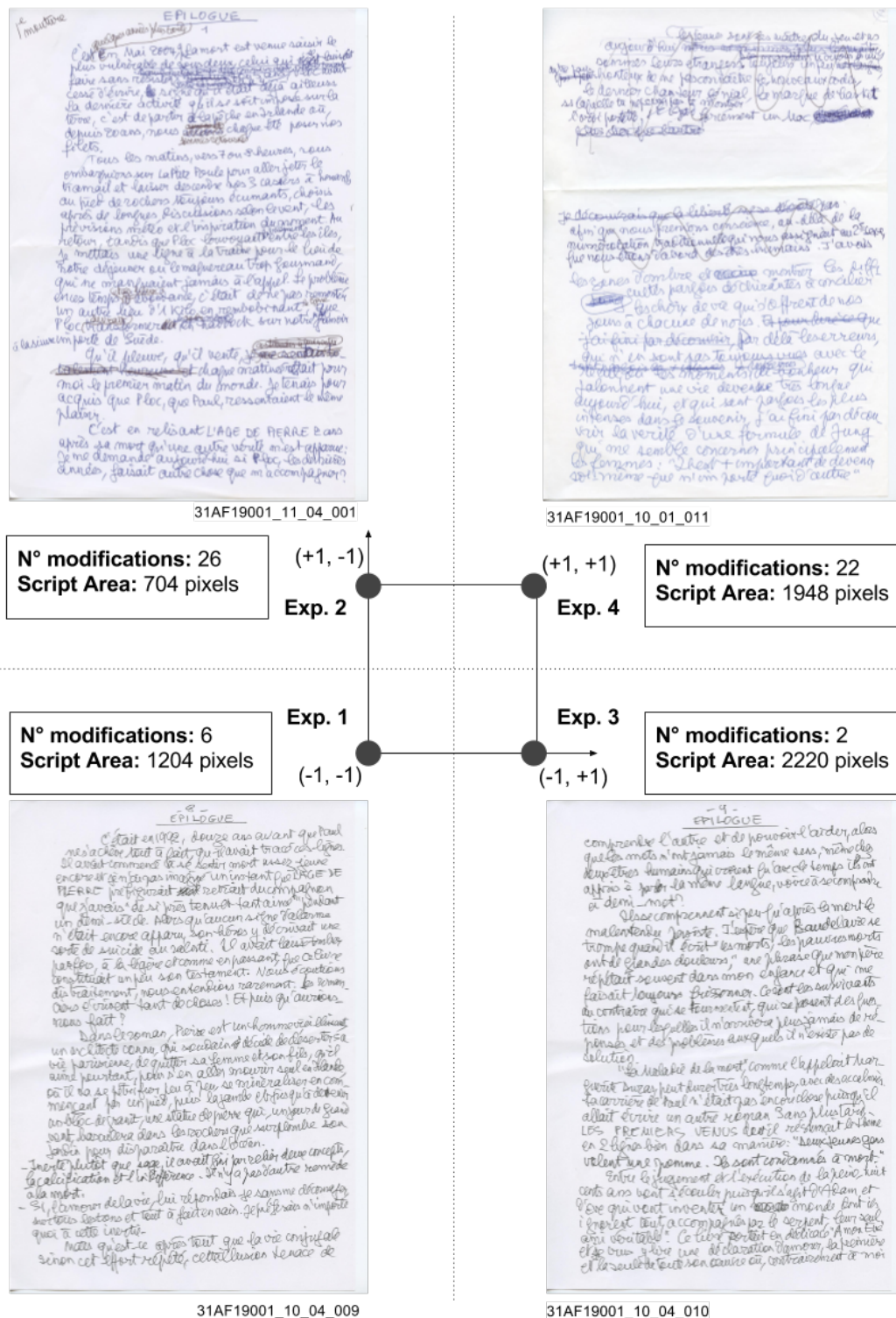


Figure 12.2 – Experiments for the full factorial plan.

We ran the experiment asking participants to transcribe all four pages at their own rate and in their spare time. A total of 57 transcriptions were collected from 15 participants, however some participants did not transcribe all four pages. After sorting the contributions we were left with 10 participants who had correctly completed the activity. We also asked an expert to transcribe the same 4 transcriptions, which would be used as references.

12.2 Data analysis and results

At this point, for each of our 4 pages, we have 11 transcriptions made by our novices and 1 provided by an expert. We can thus measure, for each participant, his or her distance to the expert. In the Design of Experiments (DOE), we use this value as a measure of the complexity itself: the higher the distance, the more complex the page is to transcribe. The results we obtained are shown in table 12.2. The left-most column lists participating novices, with their unique identifiers. The other four columns correspond to each of the pages they transcribed, identified by one of the four combinations explained previously in Figure 12.2. Each value given is the distance in characters of the corresponding novice to our expert reference. We can thus observe how individuals score given each of the four types of pages. The sharp increases observed in the second and fourth columns are an indicator that modifications influence the results and that they can be considered a significant page complexity factor. The first and third columns, on the other hand, repertory ranges of numbers that we have already observed in previous experiments.

We then run the DOE analysis itself. In our case, the DOE consists in determining the coefficients β_0 , β_1 , β_2 and $\beta_{1,2}$ of our linear system:

$$C_{total}(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2 \quad (12.2)$$

To recall, when we refer to a coefficient, we refer to the weight of a factor on the system. That is to say, how much of an effect it has on our study domain. These values

| Novice | Exp. 1 (-1,-1) | Exp. 2 (+1,-1) | Exp. 3 (-1,+1) | Exp. 4 (+1,+1) |
|--------|-----------------------|-----------------------|-----------------------|-----------------------|
| 3 | 63 | 263 | 45 | 250 |
| 26 | 61 | 254 | 47 | 273 |
| 29 | 52 | 161 | 94 | 207 |
| 34 | 63 | 201 | 46 | 501 |
| 35 | 65 | 233 | 35 | 249 |
| 36 | 55 | 268 | 50 | 241 |
| 37 | 84 | 250 | 65 | 222 |
| 40 | 63 | 318 | 35 | 234 |
| 48 | 70 | 160 | 43 | 271 |
| 49 | 71 | 224 | 86 | 247 |
| 51 | 59 | 189 | 39 | 183 |

Table 12.2 – Distances from novices to expert for each pages of the DOE. The grey column indicates the user's id in the database.

can be used as input to calculate overall complexity, or as we recall from Section 12.1.1, C_{total} . To determine the weight of our coefficients, we used the software MODDE 12⁵ to perform the regression and we obtained the coefficients depicted in Figure 12.3.

The plot shows the two factors studied; the number of modifications ($N^{\circ}m$) and script area (Scr) and also their interactions ($N^{\circ}m \times Scr$). We can see from looking at the plot that the dominating factor is modification number. Whereas, effects of script area are less evident, with values much closer to 0. There is also minimal interaction between the two factors. In term of coefficient values, we obtain $\beta_1 = 86.7$, $\beta_2 = -1.3$ and $\beta_{1,2} = 4.2$.

In fact, as the Figure 12.3 indicates, the highest value is occupied by the $N^{\circ}m$ factor and it is responsible for 94% of the observed error, the script area to 1.4% of the error and finally, their interaction contributes to 4.5% of the total error. MODDE also calculates the confidence interval associated with each coefficient, which estimates a possible range

5. <http://umetrics.com/kb/modde-12>

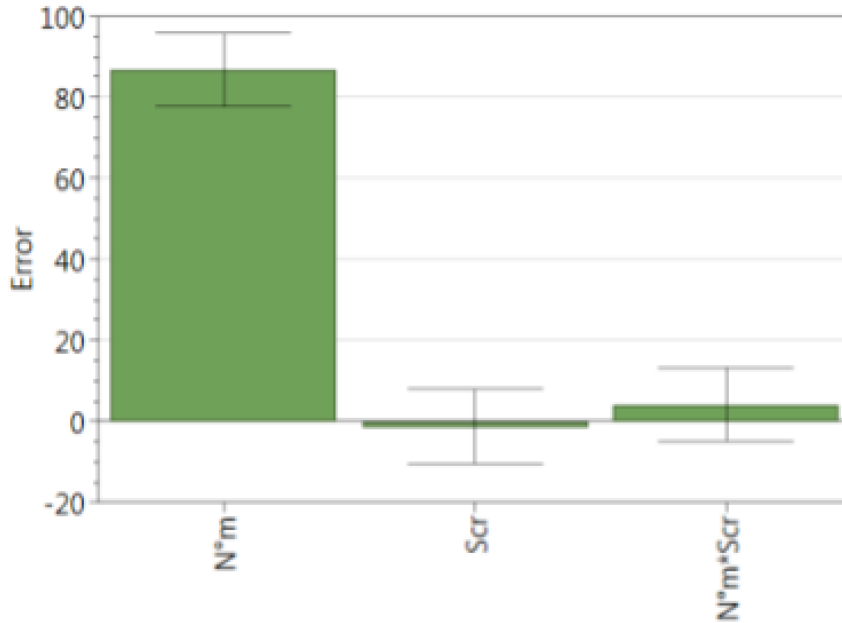


Figure 12.3 – Regression coefficients for factors: modification number ($N^{\circ}m$), script area (Scr), and their interactions ($N^{\circ}m * Scr$).

for the coefficient. For instance, for number of modifications, this range is approximately 76 to 96. For Scr and $N^{\circ}m \times Scr$ these values can also fall below zero.

When a coefficient is a positive value, this indicates that its effect on errors will increase as the factor increases. When a coefficient is a negative value, such as for script area, this indicates that errors will decrease as script area increases, which is what one would expect. As shown in figure 12.4a, an increase of $N^{\circ}m$ from -1 to +1 means an increase in the number of modifications, which will lead to an increase in the number of errors, as shown in the second and fourth columns in Table 12.2. The increase of Scr from -1 to +1 will mean that writing gets larger, which leads to a fewer errors. However, with a β_2 of -1.3, compared to a β_1 of 86.7, the effect of script area is negligible.

12.2.1 Discussion

The β coefficients give us the effects of the two factors on the resulting errors of non-expert transcribers. Performing experiments such as these allows us to determine which

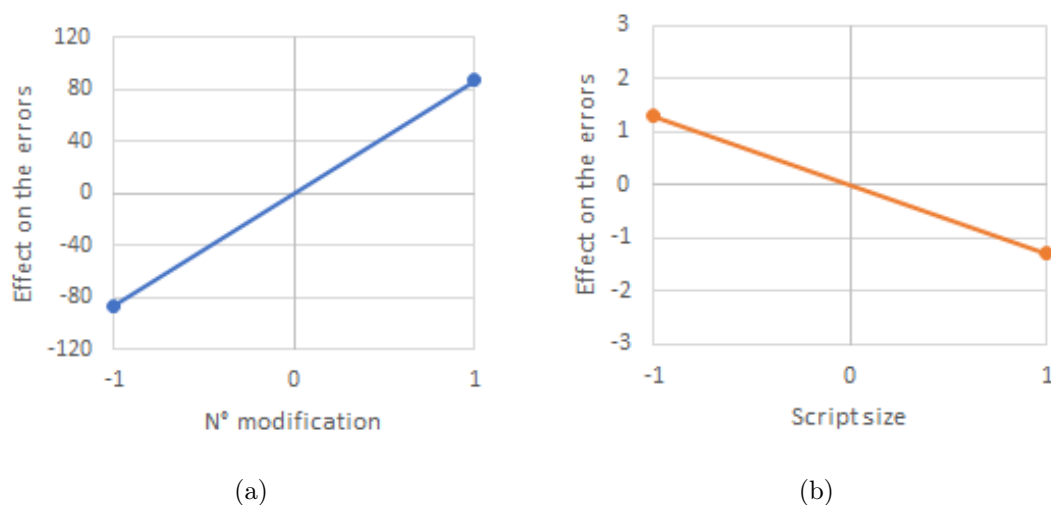


Figure 12.4 – Effect of factors on observed errors. (a) As the number of modifications rises, errors rise. (b) As script size size rises, errors drop.

factors are most prominent for a particular corpus of pages, or a set of pages within a corpus. Using a DOE can allow us to obtain reliable results with a minimum number of experiments.

After performing this DOE on multiple participants, we can expect that in the Benoîte Groult sample one of the two tested factors, that of script or letter size, has minimal to no effect on results. This can be explained by two reasons. Firstly, Benoîte Groult's writing tends to be uniform throughout the corpus, as we said in Chapter 11, in Section 11.3 on page 221, and is considered an easy corpus to transcribe by our experts. Secondly, having means of zooming in on a page effectively minimizes the problems associated with writing size, as a zoom increases words and letters by several sizes.

The factor that did affect transcription results significantly was modification number. Greater numbers of modifications contributed to the difficulty of reading sentences and deciphering the order in which phrase sections were intended by the author, thus making transcribing specific portions of texts highly error prone.

The fact that we can say with certainty that modifications really determine difficulty,

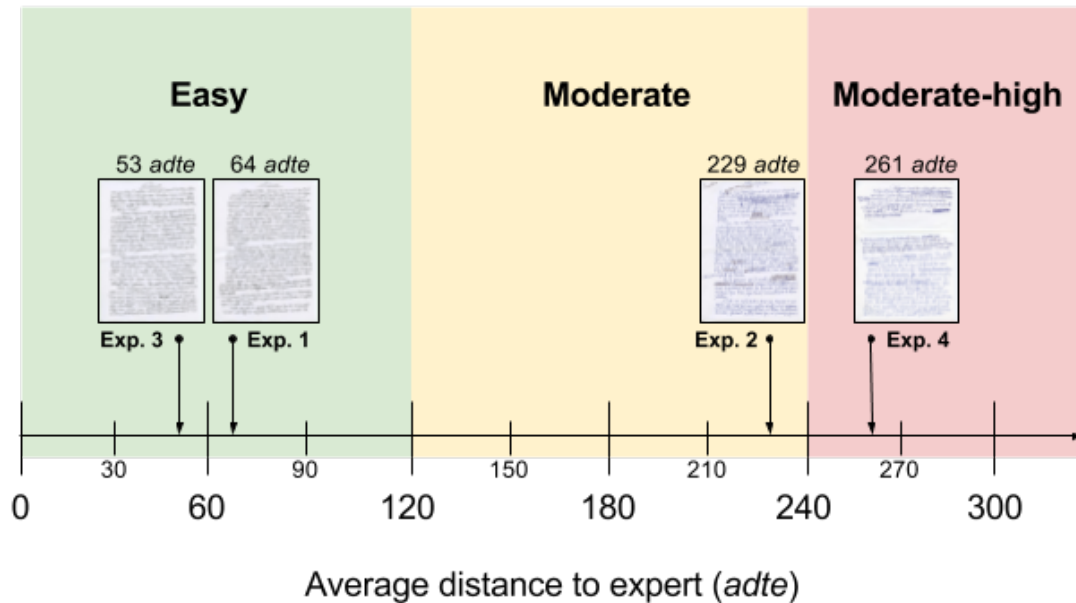


Figure 12.5 – Estimated page classification based on sample.

even on a corpus that is considered to have very accessible handwriting, is very promising. The results suggest that sorting pages according to the number of modifications may be a way to grade page complexity for the benefit of users with different levels of experience. One could create a weighted classification system of complexity that includes other investigated factors and use it to propose pages that are more likely to be well-handled by users. This would be a way to ensure higher quality of crowdsourced transcriptions. For example, based on the factors studied and the distance values obtained from multiple users for each of the four pages, we can estimate their levels of difficulty and propose a classification system. We show what such an estimation would look like in Figure 12.5; each page's category is determined by calculating the average distance to experts of our participants. Other factors can also be studied in this way. The tests can be run on a small representative sample, and then project leaders can sort the rest of the corpus intuitively on the basis of shared characteristics found in pages of the sample.

12.3 Conclusion

Knowing that modifications are an important cause page complexity is not trivial. Project leaders also design descriptive schemas based on the kinds of features they observe in a manuscript batch or collection. If a certain number of pages contain many modifications and many different types of modifications (this means beyond those we assume as additions, deletions, corrections, but extends to marginal and folio annotations, etcetera), this also results in more complex descriptive schemas. This in turn will likely create an interaction between factors at the level of page complexity and those of descriptive schema, but to study these one would need to compare manuscripts from two different collections or two different authors.

At present, the information gathered in this experiment suggests that the number of modifications can be used to categorize different batches of pages within a single collection. Those containing high levels for this factor, or many modifications, can be reserved for more experienced users. As is the case with Transcribe Bentham for example. Or, they may simply be excluded from a crowdsourcing work flow and reserved for experts only. Then, expert-transcribed examples can also be published to the website so that community members can also access these legible representations of texts.

Conclusions and perspectives

12.4 Conclusions

Today's digital editorial processes are increasingly innovative with regards to ways that information and knowledge circulate through different channels before finally arriving in the hands of readers. The practice of manuscript transcription is a vital stepping stone to accessing these processes, particularly for many unedited documents and artifacts that, otherwise, simply would not be accessible to readers.

The opportunities presented by digital tools for research and editorial processes have some very important implications for the production of textual resources. Our subject of research has brought us to look at these in more detail. We have looked at manuscripts as objects of study; the different ways that they can be described and what scholars may look for when transcribing them. Their evolving practices will be crucial also in redefining research and academic professions, as well as those of editors.

We have discussed textual encoding using XML, a markup language that suitably accommodates the descriptive potential of manuscript objects. Transcription, or the work involved in transforming digital facsimiles into machine-readable and exploitable texts, also involves encoding these texts using some form of descriptive markup. Then, depending on the markup, various output formats can be conceived. Moreover, output formats can also be anticipated by choosing to work with specific markup or managing the processes that govern how transformations between formats take place within editorial chains [Buard, 2015].

Even before one can anticipate the use of textual and or encoded material in editorial processes, transcription should be considered as a process itself. Between a virtually accessible archive of images and a collection of machine-readable, dynamic texts, a series of crucial steps must take place. Not all of which are entirely mastered to perfection. Manual transcription allows to bridge the existing technological gap created by the insufficiencies of OCR, and in 2017 it is still one of the most reliable ways to extract written content from digitized manuscripts. Inevitably, the process involves the work of experts to determine how best to represent different kinds of manuscript objects using structural and semantic

languages such as XML. We observed cases that justify using specific vocabularies and have described solutions that would allow to translate documents into other formats.

We also investigated tools and environments for manual transcription and encoding data. Our goal was to create an editor that could easily be made accessible to online publics, while allowing researchers and scholars to wholly decide on descriptive vocabularies used to encode and structure objects they choose to work with. WYSIWYG (What You See Is What You Get) editing for XML provides an interface between users and XML encoding, thus alleviating some of the more technical aspects associated with the work. The process of creating tools for transcription and encoding inevitably led to considerations of larger editorial environments, and then, even to interfaces destined for users of these environments.

With web users increased involvement in activities of Web 2.0, much discussion is focused on the potential of the crowd to support public research activities. Crowdsourcing is the term used for general open calls for participation regardless of the domain of application [Howe, 2006 ; Surowiecki, 2005 ; Estellés-Arolas and González-Ladrón-De-Guevara, 2012].

With respects to the term crowdsourcing, we have shown, based on numerous existing definitions collected by [Estellés-Arolas and González-Ladrón-De-Guevara, 2012], that the notion of openness was consistently present in definitions. The second major notion concerned involvement of people. Crowdsourced participants can actually be described as ranging from motivated and independent individuals, to organized collectives of various sizes. Internet-based networks are crucial in connecting and organizing individuals into these collectives. However, in Chapter 9, we also discuss the role of human actors to continue to support and motivate participation.

In our discussion on the subject of crowdsourcing participation, we have also called on terms like *Citizen Science* to refer to public participation in data collection for scientific research. For activities focused on humanities research, this is echoed with *Citizen Humanities*. Finally, for purposes of scholarly editorial activities, we have proposed the term *Citizen Scholarly Editing*.

Our intentions were to connect transcription activities to a potentially larger sphere of participants than is currently possible for many scholarly editing projects. What we were particularly interested in observing were transcription tasks that were opened up to people who did not necessarily have transcription experience and we investigated what crowdsourced contributors could bring to this type of scholarly work.

Our explorations of interfaces were informed by a combination of theoretical literature on interface design for the humanities and web design guidelines widely circulating in manuals and online. We experimented, and this allowed us to reflect on the ways in which digital tools change research in digital humanities. We developed prototypes and received feedback from users on how to improve them. The activity of prototyping has become important in digital humanities practice, particularly because experimentation is an indispensable part of the research endeavour [Burdick et al., 2012]. Creating these prototypes allowed us to test different aspects of the subject that motivated our research on crowdsourced manuscript transcriptions. The experimentations ultimately led to proposing new methods for monitoring and assessing the work produced by inexperienced transcribers.

With the first prototype (PHuN 2.0) we put in place a specific transcription and editorial workflow that implicates multiple users. With the second, (PHuN-ET), we directed our focus on the quality of transcriptions produced by inexperienced volunteers. Throughout, we also used experimentation as an opportunity to make improvements to aspects of the user interface, or to suggest further improvements. We prototyped a user space that can be used by users to track their own activities. We also gave users access to data they produce and the possibility to share it with others as a way to extend projects' potential for reaching new publics.

We iteratively added functionalities over the course of prototyping and development that led to improvements, which can be integrated into a single system for crowdsourced transcription and scholarly editing, or put in place other experimental environments for gathering more information about crowdsourcing under different conditions. Further development could introduce more flexible features to support the needs of many more projects.

The transcriptions we collected from our contributors were our research data. We used methods of analysis that are rooted in document comparison. We used Levenshtein's edit distance to perform textual analysis and the Zhang-Shasha tree edit distance algorithm to perform XML analysis on our collected transcriptions. The former calculates distance between documents based on the number of modifications (additions, subtractions, and substitutions) necessary to transform one document into another. While the latter, performs the same types of operations specifically on document structural components, or "nodes". Since we are dealing with XML this concerns the XML elements themselves. The total distance between two documents gives us an indication of their similarity. With text, our unit of measurement is the character unit, while with XML, it is the element. We used this method to compare the different productions collected from contributors having worked on the same manuscript object.

The divergences we observed between different contributions gave us a sense of the variability that can be expected when soliciting the general public for this type of activity. Using distance measurement allowed us to observe the overall distributions for transcriptions of specific pages, which is much faster than analysing and comparing results manually. Hierarchical clustering is a powerful visualisation tool that allowed us to rapidly get a sense of the distributions of contributions, identify outliers, as well as strong correlations in contributions. Having many contributors for an activity such as this consistently produces variable results and this is something that project leaders need to be aware of if they intend to use this method of contribution.

To determine the quality of contributed transcriptions we had to measure them in relation to transcriptions that could be considered as references, or benchmarks. An ideal transcription is an elusive target, but the closest one can expect to get to this are expert transcribers. Therefore, we ran analyses on batches of contributions that contained expert transcriptions. This also compared each and every single one of our non-expert contributions to an expert reference for a given page. The results provided us with information that we could not have obtained otherwise and can be used to study transcription outcomes for multiple participants under different circumstances. We learned that there can

be a great number of sources of variation for outcomes and that these can be investigated in greater detail by pursuing the collection of contributions from volunteer participants.

Data collection from as large a number of users as possible can permit scholars to observe variations in results contributed by users. When observed in relation to specifically identified factors, these observations can help scholars enrich their understanding of the kinds of determinants that can affect their inexperienced transcriber communities.

For example, with our first experimentations on *Stendhal*, we were able to observe distributions between participants that proved accurate in relation to prior experience that they had in the work of transcription and encoding. The phylogenetic trees we obtained from distance measurements actually regrouped transcribers who had received prior training in the same cluster as identified expert transcribers. We can say that our trained transcribers passed a major test! And so did our experts, whose similar work showed that it could be used as a reference for subsequent studies.

In later experiments on the Benoîte Groult corpus, and using the platform environment, we obtained further results that attested to variability amongst participants and quantifiable differences between their work and that of experts. We considered how different characteristics of pages, factors that were observed since the experiments on *Stendhal*, could affect results obtained from novice transcribers.

We showed a case in which using a design of experiments plan can be used to study the impact and interaction of identified factors that affect results obtained through crowdsourcing. We studied just two observable factors associated with pages themselves. However, other types of factors can arise from types of tasks⁶ as well as from XML schemas. If further experimentations of this type can be run on factors that are of particular interest or concern to projects, then new knowledge can be made available to communities. This information can include how best to handle specific aspects of projects that have been shown to impact results. For example, this can concern how to describe tasks effectively and study the impacts of different approaches on results obtained from users. Or, it can

6. These can include who defines the tasks, to whom, and how, all of which need to be carefully controlled.

be used to determine which XML schemas have more success with users and produce more consistent results.

Based on transcription comparison, project leaders can consider putting in place other forms of support for users— whether through feedback, instructions, or more guidance in choosing pages. They can also use the principle of measuring distances between novice transcriptions and expert reference transcriptions to identify talented or prodigious contributors. These individuals can be given greater responsibilities in supporting the activities of a given project or group of projects. For projects where overseeing operations is particularly difficult for experts, this would allow promoting motivated transcribers to help out with managerial roles.

Our findings have shown that variations can be observed in the contributions of inexperienced transcribers, and also that these variations can be studied in relation to different factors. This can be used to evaluate contributions and, over time, assure quality in projects, such as the ones presented. Furthermore, we would suggest that in taking steps to observe and evaluate what all contributors are capable of doing, projects can involve more people, not less. We show that what novices and enthusiasts produce can be improved. We would also suggest that projects interested in harnessing the full potential of crowd transcribers should take their results in stride and work to help them to improve their understanding and results. The information we have drawn on, the methods we have presented, and the observations we have collected can be furthered in order to create more intelligent environments that support contribution in more open and beneficial ways to greater numbers of participants.

12.5 Perspectives

Problems are solved progressively. What we have found over the course of our research and experimentations are methods to support the work of editors by opening processes to motivated and enthusiastic publics. If we spoke of knowledge dissemination at the beginning of our work, and echoed it at different points throughout, it is because in

the transcription process, the relationship between readers and writers is omnipresent. However, it also always implies— at least in traditional knowledge dissemination channels—the presence of editors. That is, those who prepare and put into proper form, and those who ultimately render more accessible that information, which is sent into circulation.

We have presented how digital technologies are shaping this process into one that is increasingly more open to participants outside of prescribed knowledge communities. Through Web 2.0, crowdsourcing, and public-oriented science and humanities projects that are anchored in web environments and platforms, has arisen the idea that research and its products can not only be shared more widely with curious and inquisitive publics, but also implicate such individuals in processes of creating data and knowledge for the benefit of others.

We have proposed an approach to enrich our understanding of the potential of participative transcription for scholarly editing. Our approach was particularly grounded digital humanities practices that view experimentation as a vital component of research and theory. Therefore, we have explored and developed tools to test our hypotheses about this subject through experimentation. And also, we have continually kept in mind the possibilities that further development could mean for the types of data objects we worked with and the types of prototypes we built:

"Future digital storytelling might utilize ebooks and mobile devices as convenient and powerful contexts for multimedia narratives created from “publicly created contributions” (Adams, 2009, p. 239). The implications and importance of these opportunities may be interesting, especially as they promote new ways of teaching and learning as well as creating, critiquing, and consuming humanities research and scholarship" [Barber, 2016].

This requires continued improvement of user interfaces so as to make online manuscript transcription more accessible to larger publics, and, yes, even inexperienced transcribers. Furthermore, environments for collaboratively managing processes should be considered as integral parts of systems. And, within these there should be particular attention to

user support. We have mentioned feedback and its role in encouraging users in Chapter 9, and then again as part of systems that support user development in Chapter 10. When we can evaluate what users produce to identify where they can improve or where they are performing exceptionally well, we have the keys to reward and encourage them further.

The information obtained from hierarchical clustering of contributions can be developed further into tools for project leaders. They can be used to select the most complete contributions. For instance, when selecting documents for further editorial processing or validation, project leaders can have at once a clustered overview of all contributions and access to text and XML documents. They can quickly see if it is best to accept an entry from the group of most similar transcriptions, or look to see if more complete entries exist in secondary clusters. They can also identify outliers and provide them with appropriate feedback for improvement. Or, they can have access to functionalities allowing them to generate median texts by aggregating multiple texts. Projects like Old Weather use multiple transcriptions to verify data and Galaxy Zoo use multiple answers to minimize errors [Dunn and Hedges, 2012 ; Franzoni and Sauermann, 2014]. However, they do not describe their methods, making it difficult to put them to greater use.

Beyond using this technique as support for sorting and selecting contributions, hierarchical clustering can help to observe recurrent patterns in contributions from users. It can be used to recognize consistency in contributions from participants, who may later take on greater responsibilities that support project advancement.

At the same time, this requires considering that multiple individuals may be interested in looking at, reading, and transcribing the same documents, and if so, then why deny them the pleasure of doing this? We have yet to hear of an art gallery or museum that closes its exhibits to visitors because there is already someone inside enjoying the objects on display. Transcription allows enthusiasts to interact with manuscripts in a more involved manner than just viewing or reading them. By transcribing pages one can also indulge in an act of solving intellectual puzzles [Causer and Terras, 2014]. There is no reason why digital environments cannot accommodate contributions from more interested participants. Particularly if project leaders have tools to monitor all contributions, both

at a glance and in detail.

We have looked at factors that can affect results and presented a method that would allow us to continue studying these factors in greater detail. Further applications would be to study the effects of different types or forms of instructions on transcriptions contributed by inexperienced users. The similarities and differences between users' results may provide further keys to understanding how information is perceived and its effect on outcome.

As well, identifying major characteristics of complex pages would allow categorizing them in appropriate ways. Here, different projects can share what they learn from their own corpora with others. We can anticipate playful interfaces where users can apply filters to obtain propositions of pages based on complexity. Or, another possibility would be to ask users to categorize pages based on intuitive perceptions of their complexity.

Based on different tests run using benchmark expert references, we can identify users at different levels of experience. We can then go further in creating appropriate feedback mechanisms, adapted to users at different levels. If using mechanisms is unwanted, then project leaders can provide appropriate feedback to individuals more directly within the platform. If there are many participants, then feedback can be directed at groups, taking the form of informative blog entries, or used to update pages of frequently asked questions (FAQs), which participants can then use to get information quickly.

We can also anticipate new research implementations in other disciplines and for other types of objects. By extending public participation into other genres of heritage projects for example and other fields in social sciences and humanities. Our tools for describing and structuring data can be adapted to activities rooted in disciplines like archeology and art history for the description of artifacts and iconographies. Or, for linguistic analysis of speech and writing based on textual corpora collected in different learning situations. In this way, new practices for data creation, assisted by methods of evaluation, can continue to support research and knowledge dissemination.

Further work would allow creating an environment where user experience is enhanced for both project leaders and participants. This means further development and improve-

ment of tools and interfaces for users. It means enhancing flexibility for configuring projects and managing work flows, as well as reinforcing comprehensive monitoring of results, and providing channels for giving and receiving feedback. It also means putting in place more experimentations to understand the effects of individual competences, schemas, and objects (manuscript pages) on contributed transcriptions. The knowledge gathered from these efforts will undoubtedly inform, encourage, and sustain better practices within scholarly editing involving the public, and perhaps crowdsourcing more generally.

Chapter 13

French Summary of the Thesis

Contents

| | |
|---|-----|
| 13.1 Introduction | 262 |
| 13.2 Contexte historique | 264 |
| 13.3 Définitions | 265 |
| 13.4 Encodage des textes et outils de transcription | 269 |
| 13.5 Architectures et interfaces | 271 |
| 13.6 Mesures de différence entre transcriptions | 272 |
| 13.7 Types de prototypes | 279 |
| 13.8 Présentation de PHuN 2.0 | 280 |
| 13.9 Présentation de PHuN-ET | 284 |
| 13.10 Au-delà des plateformes | 285 |
| 13.11 L'assurance qualité pour le crowdsourcing | 289 |
| 13.12 Mesurer la qualité des transcriptions | 291 |
| 13.13 Mesurer les facteurs de complexité | 294 |
| 13.14 Conclusion et perspectives | 298 |

13.1 Introduction

Les nouvelles technologies numériques jouent un rôle important dans la dissémination des savoirs tel que nous la connaissons aujourd'hui, avec ses nouveaux médias et formats de lecture.

La transcription de manuscrits est un travail important pour l'étude de manuscrits ainsi que pour la constitution de matériaux destinés à l'édition. Cette pratique se voit affectée par l'introduction de nouvelles technologies, qui permettent d'impliquer de plus en plus des publics motivés à participer à ces processus.

Ces publics n'ont pas nécessairement une grande expérience dans les domaines concernés et s'impliquent souvent dans des projets par passion. Néanmoins, l'apport de leur contribution est souvent mis en doute à cause de leur faible expérience du domaine concerné. Ainsi, les méthodes qui ouvrent la participation au plus grand nombre, ou qui utilisent des personnes non-spécialistes de la transcription sont questionnées quant à leur capacité à avoir une plus-value pour les processus éditoriaux fondés sur la transcription.

Néanmoins, beaucoup de chercheurs souhaitent découvrir les méthodes qui se fondent sur les contributions du plus grand nombre – les méthodes de mise en réseau des foules – ou de crowdsourcing, et sont, pour la plupart, enthousiasmés par l'apport potentiel de l'implication des amateurs :

Il serait sans doute démagogique de promettre à tout un chacun qu'il saura déchiffrer séance tenante l'écriture de Pascal ou de Stendhal mais il n'est pas inconcevable que tel amateur de bonne volonté puisse suggérer une lecture pertinente de tel passage difficile, dans lequel la fraîcheur de sa perception aura su distinguer ce que des chercheurs plus aguerris n'avaient pas perçu. [Leriche and Meynard, 2008].

Nous portons notre intérêt sur l'exploration de ces méthodes contributives de crowdsourcing ainsi que sur l'expérimentation avec des outils permettant sa mise en oeuvre, plus particulièrement pour la transcription de manuscrits. Ainsi, nous avons créé des pro-

totypes de plateformes numériques, pour l'expérimentation et pour la collecte de données, ainsi que pour envisager des environnements de travail pour les individus et porteurs de projet qui souhaitent se lancer dans ce type de production en ligne.

a) Questions de recherche

Il est difficile de mesurer l'apport de la méthode de crowdsourcing et son efficacité vis-à-vis des sources manuscrites des auteurs. Nous avons donc choisi de nous concentrer sur l'évaluation des résultats des transcriptions produites par des participants non-initiés pour la plupart.

L'enjeu est de savoir si des personnes novices et amateurs pourront produire des matériaux de qualité suffisante pour permettre aux projets de recherche et d'édition de s'appuyer sur leurs efforts. Nous avons donc comparé leurs contributions au regard des transcriptions d'experts, qui sont considérées comme des références, afin d'observer ce qu'il est possible d'attendre des participants. Cette démarche permettra de mieux comprendre comment la méthode de crowdsourcing pourrait bénéficier aux projets travaillant sur les sources manuscrites.

b) Plan de thèse

Cette dissertation est présentée en quatre parties. Dans la première, nous présentons le contexte de la thèse et les définitions des termes importants.

Dans la deuxième nous présentons les fondations techniques sur lesquelles nous nous appuyons dans notre travail. Nous présentons les principes de l'encodage des données en XML, les différents types de metadonnées, et comment les données encodées se traduisent en contenus dynamiques du web. Nous présentons également les outils de transcription et les outils de gestion de contenu (CMS) et des utilisateurs (UMS). Enfin, nous présentons les techniques que nous utilisons pour comparer de multiples contributions des internautes afin de mieux comprendre la qualité qu'il est possible d'avoir en utilisant le crowdsourcing.

Dans la troisième partie nous présentons les deux prototypes de plateforme numérique

que nous avons créés ; le premier destiné à la production collaborative et le second destiné à l'expérimentation et à l'évaluation des données. Nous menons une réflexion aussi sur la dimension humaine qui est cruciale pour la réussite des projets de crowdsourcing, y compris la gestion des aspects de motivation, la communication, et le développement des compétences des utilisateurs.

Dans la quatrième et dernière partie, nous présentons les résultats des expérimentations menées sur les corpus des manuscrits de Stendhal et sur celui de Benoîte Groult. Nous utilisons les méthodes d'analyses décrites dans le Chapitre 6 pour évaluer la qualité des données obtenues des participants inexpérimentés. Nous comparons leurs contributions à des transcriptions de référence. Nous examinons les facteurs de complexités qui influencent les résultats et proposons des applications pour l'analyse, par exemple pour trier les pages des corpus par niveaux de difficulté.

Finalement, pour conclure cette thèse, nous résumons les connaissances acquises ainsi que les résultats obtenus lors de cette étude. Nous discuterons de l'apport des méthodes mises en place pour les projets de transcription contributives et comment elles peuvent bénéficier à la fois aux participants et aux porteurs de projets.

13.2 Contexte historique

Les méthodes de diffusion de l'information ont beaucoup évolué depuis les besoins croissants des universités au moyen âge [Baudet, 2003]. À cette époque, les techniques pour copier des livres ont d'abord été améliorées par les copistes, qui ont introduit des outils et des techniques de gravure afin de reproduire des lettres ornées plus facilement [Baudet, 2003].

Avec l'introduction de l'imprimerie par Gutenberg dans les années 1450, le même principe de copier les lettres a été appliqué avec des outils et matériaux plus adaptés pour reproduire des lettres de plus petite taille ; Gutenberg introduisit l'utilisation des poinçons en métal pour produire des lettres individuelles, permettant ensuite de les arranger dans des lignes et des pages entières [Baudet, 2003].

Quelques siècles plus tard, l'internet a permis d'accélérer encore les moyens de diffusion des savoirs. Le Web 2.0, autrement connu sous le nom de web collaboratif, a permis de mettre l'utilisateur au centre des activités effectuées sur internet. Au lieu d'être simplement une archive d'information indexée, le Web 2.0 implique des internautes pour créer les contenus qui sont disponibles sur la toile [Murugesan, 2007 ; Vitali-Rosati and Sinatra, 2014]. Cela implique, bien sûr, des infrastructures robustes et des outils adaptés pour accueillir les informations et les internautes. Sans cela, des projets de crowdsourcing contemporains ne seraient pas possibles, ni les efforts de conservation menés par les musées et les institutions patrimoniales.

13.3 Définitions

Pour aller plus loin, nous présentons quelques définitions de termes qui sont au centre de notre sujet de recherche : les humanités numériques, le crowdsourcing et les sciences/humanités citoyennes, l'édition érudite citoyenne, ainsi que la transcription de manuscrits.

13.3.1 Les Humanités Numériques

Nous définissons les humanités numériques comme un ensemble de pratiques fondées sur le contenu et les données numériques, qui pourraient inclure l'utilisation d'outils numériques pour manipuler et transformer des objets numériques afin de créer de nouvelles informations et de nouvelles ressources. Elles incluent aussi les méthodes de partage des connaissances entre chercheurs en humanités et, plus largement, des personnes qui s'intéressent aux sciences humaines.

13.3.2 Crowdsourcing, Sciences Citoyennes, et Humanités Citoyennes

En 2006, Howe a défini le crowdsourcing comme tout mode de production en ligne provenant d'un appel ouvert à participer, qu'il soit sollicité par des institutions privées ou publiques [Howe, 2006]. Selon Howe, ce mode de production était destiné à changer la façon dont les gens travaillaient dans le monde entier, avec des implications importantes pour les économies mondiales.

Aujourd'hui, les définitions de crowdsourcing peuvent varier grandement et peuvent non seulement être divergentes mais aussi contradictoires [Estellés-Arolas and González-Ladrón-De-Guevara, 2012]. En conséquence, un certain nombre de chercheurs ont tenté de clarifier ce terme, en ajoutant des caractéristiques essentielles, ainsi soulignant sa nature polyvalente. Notamment, Estellés-Arolas et González-Ladrón-De-Guevara (2012) comparent diverses définitions du crowdsourcing afin d'améliorer la compréhension du terme. Dans ces définitions, tirées de trente-deux articles différents, il est possible de voir qu'une définition du crowdsourcing comporte la notion de l'individu, soit en tant que personne libre et motivée, soit ayant le potentiel de s'organiser dans des collectifs d'individus très vastes. Le crowdsourcing dépend des réseaux et implique des infrastructures nécessaires à la mise en relation des personnes avec des activités et avec d'autres personnes.

Le terme *Citizen Science* est utilisé pour désigner spécifiquement les projets de recherche qui sollicitent la contribution du public, le plus souvent avec un site ou une plateforme en ligne comme interface entre les membres contributeurs et les experts du domaine. La majorité des projets de crowdsourcing bien connus, tels que ceux hébergés par Zooniverse, sont appelés projets scientifiques citoyens, car ils ont une composante scientifique et parce qu'ils impliquent le public. Le terme *citoyen* connote un certain degré d'implication dans une communauté publique¹. Dans le cas des sciences citoyennes, les établissements de recherche sont ceux qui font un appel ouvert. Les participants qui répondent à cet appel et s'impliquent contribuent à des activités scientifiques qui peuvent

1. Contrairement à une entreprise ou une entreprise privée.

bénéficier à la fois aux institutions et, inévitablement, aux personnes et aux communautés desservies par ces institutions.

Dans la mesure où il existe une distinction entre les sciences et les sciences humaines, d'autres termes peuvent être appropriés lorsque l'on se réfère à des activités dans lesquelles le crowdsourcing est utilisé par les institutions des sciences humaines. *Citizen Humanities* est un terme qui circule déjà en ligne et dans certaines communautés pratiquant les humanités numériques. Des exemples de tels projets peuvent inclure l'AnnoTate de la Galerie Tate en Angleterre. Nous proposons un autre terme s'appliquant spécifiquement à l'édition scientifique qui sollicite des participants : l'Edition Erudite Citoyenne (Citizen Scholarly Editing).

13.3.3 La transcription de manuscrits

La transcription de manuscrits est une activité importante parmi de nombreuses autres méthodes de conservation. La numérisation de manuscrits, ou d'autres formes d'objets, permet aux chercheurs de travailler avec des documents qui sont rares et difficiles d'accès. Dans de nombreux cas, la transcription est un moyen d'accéder à d'autres processus numériques, y compris des processus de recherche et d'édition, qui constituent des domaines de pratiques spécifiques dans le domaine des humanités numériques.

Dans les études littéraires et textuelles, la transcription est une pratique de l'édition critique tout comme de l'édition numérique. Les éditeurs ont pour objectif de constituer des éditions de textes issus des projets d'auteurs. Les généticiens du texte (chercheurs textuels ou critiques selon l'école) travaillent à étudier le processus de création du texte, en commençant par des premières ébauches et qui mènent parfois à des éditions connues du grand public. Les brouillons étudiés prennent souvent forme d'abord sur papier et sont écrits par la main de l'auteur lui-même ou elle-même, ou par des scribes sous la dictée de l'auteur.

La transcription peut se révéler être une tâche complexe, car elle vise à reconstruire les textes tout en mettant en évidence les modifications et/ou le processus de rédaction.

Grâce à ce travail, des chercheurs peuvent étudier l'écriture des auteurs à l'aide des indices misent en lumière par la transcription.

Cette activité peut être vue comme un processus de décodage et d'encodage. La lecture est un processus cognitif qui réceptionne et interprète les marques sur le papier ; il s'agit du décodage de l'information. Ensuite, en reproduisant ou transcrivant l'écriture selon les conventions de transcription adaptés (cela pourrait impliquer des conventions d'un langage informatique aussi), nous parlons alors de l'encodage.

Nous abordons le processus de transcription dans un contexte de plus en plus numérique, il est donc important d'aborder le sujet du point de vue numérique et voir comment ce contexte influence le travail de transcription. Par exemple, en réponse au volume de documents et à la précision requise pour les traiter, la technologie n'a été que partiellement capable de répondre aux besoins exprimés par les éditeurs et les chercheurs. Même sur des documents contenant peu de modifications, les technologies de Reconnaissance Optique de Caractères (OCR) ne répondent pas bien aux irrégularités observées dans les manuscrits contemporains d'auteurs [Espana-Boquera et al., 2011]. Ainsi, la transcription, même dans un contexte numérique, est encore une tâche manuelle qui est majoritairement réalisée par des experts. Généralement, leurs objets d'études sont les textes qu'ils transcrivent.

Afin d'organiser les processus de transcription pour un public plus large de contributeurs, la transcription doit être introduite en tant que tâche pour ces contributeurs. Il y a différents types de tâches pouvant être sollicitées du public, y compris celles d'écriture créative ou de création du contenu créatif. La transcription diffère de l'écriture créative et du contenu créatif car c'est un acte de reproduction qui se fonde sur une source existante, la page manuscrite elle-même. Avec la transcription de manuscrits nous considérons qu'il est possible d'avoir un contenu attendu ainsi qu'un contenu qui répond aux attentes des experts. Nous considérons donc qu'il est possible d'évaluer ce que produisent des contributeurs inexpérimentés en les comparant avec ce qui est produit par des transpositeurs expérimentés.

Nous savons que les experts cherchent à obtenir une transcription optimale et consi-

dèrent qu'il est possible de l'obtenir. La question est de savoir si les transcriptions résultant de la participation du public sont à la hauteur, compte tenu des attentes élevées des groupes d'experts. Avoir un aperçu des résultats qu'il est possible d'obtenir avec des transcriptions crowdsourcées, peut bénéficier aux organisations et aux communautés qui lancent des projets de rédaction s'appuyant sur la participation du public.

13.4 Encodage des textes et outils de transcription

Les projets d'édition scientifiques et savants qui visent à établir un corpus de textes numérisés doivent tenir compte des moyens techniques à leur disposition. Dans cette section nous présentons les moyens utilisés afin d'encoder des données textuelles.

Le processus de conversion d'objets historiques, tels que les manuscrits, en facsimiles numériques se fait en plusieurs étapes. La première est la création du fac-similé, nécessitant souvent l'utilisation d'un équipement de photographie numérique. Cette étape permet la création de copies de documents en haute fidélité qui, une fois disponibles en ligne, peuvent ensuite être consultées dans le monde entier par des chercheurs. La deuxième étape consiste à fournir des métadonnées sur le document. Les métadonnées fournissent des informations descriptives, structurelles et administratives permettant de décrire et de contextualiser les documents. Ils jouent un rôle important, tant dans l'archivage numérique que dans les processus éditoriaux.

Les métadonnées descriptives incluent des informations sur le document lui-même, y compris la provenance, la date, l'auteur ou les auteurs, et d'autres informations concernant l'édition, mais ne se réfère pas nécessairement au contenu écrit ou textuel du document – les données du document. Les métadonnées structurelles permettent de relier les ressources entre elles. Les métadonnées administratives sont des informations utilisées pour la gestion des ressources par des archives ou des bibliothèques par exemple. Les métadonnées administratives regroupent les métadonnées techniques, les métadonnées de conservation et les métadonnées des droits.

Dans notre travail, nous sommes intéressée par le contenu écrit des manuscrits. Nous

cherchons à coder le contenu écrit des manuscrits et le considérons comme la première étape du processus menant à des représentations définitives du texte. Cette première étape nécessite l'utilisation de vocabulaires descriptifs spécifiques pour le contenu. Il ne s'agit pas ici de travailler avec les métadonnées administratives ni descriptives du document, car cette tâche relève plus souvent du domaine de compétence des spécialistes qui travaillent plus étroitement avec l'objet manuscrite lui-même, ou des personnes responsables de la numérisation de l'objet.

Le XML se prête particulièrement bien à l'encodage des textes pour plusieurs raisons. Tout d'abord, sa structure arborescente reflète des composantes structurelles et sémantiques traditionnellement trouvées dans les textes. Deuxièmement, on peut définir ses propres noms et structure d'éléments. Tant que les règles fondamentales de structuration sont respectées, le XML est bien formé et, si une Déclaration de Type de Document (DTD) associée est respectée, le XML est valide selon cette DTD. Bien formé signifie simplement que tous les éléments, à moins qu'ils soient vides, ont à la fois des étiquettes d'ouverture et de fermeture. Un document peut être composé de plusieurs sections contenant des titres, des sous-titres et des paragraphes qui forment sa structure. Lorsque nous parlons de la structure du document, nous soulignons les relations entre les éléments et les règles que nous suivons pour les organiser. Ces règles, qui sont similaires à la grammaire dans les langues naturelles, peuvent être déclarées pour les documents XML dans une DTD. Les noms des composants peuvent être qualifiés de vocabulaire et mettent en évidence leurs fonctions sémantiques dans les documents.

Le codage du texte à l'aide du balisage XML est une tâche qu'une partie des chercheurs, notamment en sciences humaines et sociales, ne sont pas enclins à faire, ou ne sont pas capables de faire à cause de barrières techniques. Ainsi, l'externalisation de tâches de transcription au public est considérée comme une solution convenable. Cela étant dit, il faut créer des outils et des environnements pour le public permettant de transcrire et encoder des textes provenant des sources manuscrites.

Les éditeurs XML existants, tels que Morphon et Oxygen ont permis à beaucoup de projets d'entreprendre des tâches de création de documents structurés en XML. Certains

de ces éditeurs, et Oxygen en particulier, ont une option d'interface WYSIWYG (*What You See Is What You Get*), qui permet d'encoder les documents sans devoir manipuler du code brut. Ce type d'interface est plus accessible pour des personnes n'ayant pas de compétences spécialisées en informatique et permet d'éviter des erreurs d'encodage et de structuration. Néanmoins, des outils tel que Oxygen ne sont pas accessibles à tous les projets pour des raisons financières. De plus, ce ne sont pas des logiciels qui permettent d'éditer des textes en ligne, car il faut impérativement avoir une copie du logiciel sur son propre ordinateur. Il est donc intéressant de proposer une solution permettant aux projets de disposer d'un éditeur de transcription en ligne, qui est accessible à tous les participants des projets.

De plus, afin de pouvoir gérer les transcriptions contribuées par des internautes, il est nécessaire également de considérer comment l'éditeur de transcription est relié avec un système de gestion de contenu (CMS). Nous décrivons l'architecture d'un tel système dans la section suivante et nous discutons aussi de l'importance des interfaces pour les utilisateurs.

13.5 Architectures et interfaces

Les systèmes de gestion de contenu sont des systèmes utilisés pour créer et gérer des contenus numériques. Le contenu peut inclure des objets numériques tels que des images, des vidéos, de la musique, des textes et, dans notre cas, des transcriptions. Certains systèmes de gestion de contenu communément connus sont Wordpress, Omeka et Drupal, mais il y en a bien d'autres. Bien que le mot *contenu* ne fasse pas nécessairement penser directement à l'information et au savoir, le CMS, est à bien des égards, une infrastructure moderne pour la gestion et la diffusion des connaissances. Pour beaucoup, c'est grâce à WordPress que l'édition Web est devenue aussi accessible et répandue.

Un CMS n'est pas seulement un système de stockage et de publication de contenu, c'est aussi un environnement capable d'accompagner la collaboration entre plusieurs utilisateurs. Parfois, le terme UMS (*User Management System*) est utilisé pour se référer

spécifiquement à la gestion des utilisateurs, mais dans de nombreux cas, *Content Management Systems* implique également une gestion des utilisateurs. En général un CMS fonctionne de la manière suivante. Les administrateurs sont chargés de créer des mises en page pour le site Web et de décider comment le contenu sera structuré, alors que les utilisateurs contribuent en créant ou en éditant le contenu lui-même. L'application CMS prend en charge l'injection de ce contenu dans la disposition conçue pour créer des pages Web.

De nombreux CMS proposent également des solutions pour la gestion des utilisateurs. Les rôles hiérarchiques sont des fonctionnalités communes intégrées aux architectures CMS et peuvent être définis en fonction des types d'utilisateurs et des processus envisagés dans le système.

Pourtant, un problème rencontré avec des CMS concerne la maintenance et de la portabilité du contenu, car, une fois qu'il a été adapté à un CMS particulier, un changement de plateforme demande souvent d'importants changements structurels aux données.

Il est important de continuer à chercher des solutions qui répondent aux besoins spécifiques et multiples exprimés par les chercheurs. Les processus éditoriaux peuvent devenir plus efficaces lorsqu'ils reflètent les besoins de ceux qui les conçoivent. Plutôt que d'adapter les processus et les produits de la recherche scientifique à tel ou tel CMS, il y a un besoin d'avoir une architecture robuste pouvant gérer de gros volumes de données et de nombreux utilisateurs, mais qui est également flexible en termes d'outils et de composants proposés. Enfin, le contenu doit être facilement accessible pour qu'il soit utilisé pour l'édition web, l'édition papier ainsi qu'à d'autres fins archivistiques ou de recherche.

13.6 Mesures de différence entre transcriptions

Les logiciels de comparaison de textes comportent une interface qui permet aux utilisateurs de repérer les modifications qui ont dû être apportées au premier texte afin d'arriver au deuxième texte. Les modifications sont souvent surlignées afin de montrer les parties de texte supprimées et celles ajoutées. Ce support visuel est utile pour facilement repé-

rer les détails qui, sinon, peuvent échapper l’œil humain quand il est face à des grandes quantités de texte.

13.6.1 Mesurer les différences entre textes : distance de Levenshtein

Afin d’obtenir une mesure quantitative de la différence entre deux textes nous utilisons des métriques se fondant sur les chaînes de caractères, qui expriment la différence entre deux chaînes en tant que *distance*.

La distance de Levenshtein est une métrique connue et souvent appliquée en linguistique et en informatique afin de mesurer la différence entre deux chaînes de caractères [Levenshtein, 1966]. Les opérations autorisées incluent des ajouts, des suppressions et enfin des substitutions au niveau du caractère [Levenshtein, 1966]. Dans notre cas, il peut être considéré comme une mesure du nombre minimum de corrections nécessaires pour passer d’une transcription à l’autre. C’est la méthode que nous avons choisie pour comparer les transcriptions. La formule peut être donnée comme suite :

$$Distance_{i,j} = ajouts_{i,j} + suppressions_{i,j} \quad (13.1)$$

La distance entre deux textes i et j peut être obtenue en calculant la somme du nombre d’ajouts et de suppressions nécessaires pour transformer un texte i en un texte j .

La distance de *Levenshtein* est aussi connue sous le nom de *distance d’édition* (*edit distance* en anglais), et un certain nombre d’algorithmes utiles ont été adaptés pour surveiller les opérations d’addition et de suppression. Plus important encore, le calcul de la distance d’édition entre plusieurs documents (comparaison de chacune des paires de documents) peut nous permettre de déterminer ceux qui sont les plus similaires les uns aux autres ainsi que ceux qui sont les moins similaires. La distance d’édition peut être effectuée non seulement sur les chaînes, mais aussi sur les documents structurés [Zhang and Shasha, 1989].

13.6.2 Mesurer les différences entre documents XML

La comparaison de XML est plus complexe que la comparaison de chaînes textuelles. Autrement dit, un fichier XML contient des informations codées ou structurées et la détermination des différences entre les documents XML exige d'examiner les différences dans la structure des éléments qui composent ces documents [Nierman and Jagadish, 2002]. Un certain nombre d'algorithmes ont été décrits pour calculer les modifications apportées aux documents XML, y compris l'algorithme de [Chawathe et al., 1996], de [Nierman and Jagadish, 2002], et de [Zhang and Shasha, 1989]. Chacun de ces algorithmes consiste à détecter des opérations d'ajout, de suppression, de substitution et parfois d'autres types d'opérations sur les arbres de documents XML. Chacun a ses particularités. Par exemple, [Chawathe et al., 1996] détecte les ajouts et suppressions, mais détecte également quand des éléments ont été déplacés dans l'arbre. L'algorithme de [Nierman and Jagadish, 2002] est très similaire à [Chawathe et al., 1996], mais aussi des ajouts et suppressions de sous-arbres (et non pas seulement des noeuds). Enfin l'algorithme de [Zhang and Shasha, 1989] permet des ajouts et suppressions des éléments comme les autres, mais il permet aussi d'ajouter ou supprimer des noeuds n'importe où dans l'arbre, même si celui-ci a des noeuds-enfants, sans supprimer ses noeuds-enfants. L'opération consiste à rattacher d'abord les noeuds-enfants au parent du noeud à supprimer, puis supprimer le noeud en question.

Nous avons choisi l'algorithme de [Zhang and Shasha, 1989] pour sa simplicité d'implémentation, ce qui a rendu possible d'analyser les ensembles de documents XML contribués par des transcripateurs volontaires.

13.6.3 Techniques de clustering

Les valeurs de distance que nous obtenons de la distance d'édition de chaînes (*string edit distance*) et de la distance d'édition des arbres XML (*tree edit distance*) nous permettent de quantifier les différences entre plusieurs textes. Maintenant, nous devons organiser ces résultats en utilisant les techniques de regroupement ou de *clustering*.

Le *clustering* est utilisé dans les traitements informatiques afin d'organiser des documents selon des caractéristiques définies telles que des termes ou des mots-clés. Il est utile pour créer des systèmes de recherche intelligents et peut être utile pour améliorer l'organisation et l'accessibilité des collections. Cette méthode peut être utilisée pour classer les documents selon des thèmes ou des sujets. Il est courant d'utiliser le clustering pour organiser des documents étroitement liés et les distinguer des documents sur d'autres sujets [Huang, 2008]. Le clustering est considéré comme particulièrement efficace sur des ensembles de données larges et hétérogènes. L'utilisation de cette technique permet de regrouper les objets en fonction de leurs similarités ou de leurs dissimilarités.

La similarité entre objets est souvent exprimée en tant que proximité. Les représentations typiques des clusters sont fondées sur la mesure de la distance entre les objets, afin de déterminer s'ils appartiennent à un groupe ou à des groupes distincts. On dit qu'un objet A est plus similaire à un objet E comparé à un objet B si la distance de A à E est inférieure à la distance de A à B. Cette situation est représentée dans la Figure 13.1. Dans cette collection d'objets A et E sont plus proches que les autres objets. Les ovales représentent des *clusters* résultant d'une classification hiérarchique.

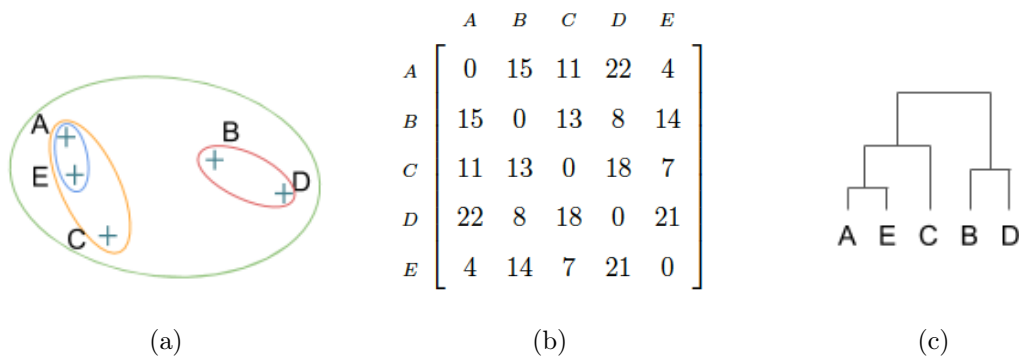


FIGURE 13.1 – La clusterisation des objets dans l'espace est montré dans (a). Dans (b), une matrice de distances représente les mesures de distance (*au*) entre objets, et (c), montre un arbre phylogénétique qui représente les objets selon leurs valeurs de distances relatives.

Dans notre cas, les objets sont des transcriptions. Pour mesurer la similarité entre les

objets, nous utilisons la notion de distance qui peut être prise littéralement comme une distance métrique entre objets dans l'espace, comme dans l'exemple, ou être affectée d'une valeur reposant sur la quantité d'opérations ou d'erreurs séparant deux objets, comme nous l'avons décrit avec la distance de Levenshtein. Les unités que nous utilisons pour les textes sont des caractères, alors que pour XML nous utilisons des nœuds d'éléments qui constituent l'arbre XML.

Dans notre cas, nous utilisons la méthode de regroupement hiérarchique agglomératif. Elle consiste à parcourir l'ensemble des données pour trouver les paires d'objets les plus proches et les former en *clusters*, puis nous les fusionnons pour former des *clusters* de plus en plus grands, jusqu'à obtenir finalement le cluster global. Si nous considérons Figure 13.1 comme un processus de clusterisation hiérarchique, il comprend les étapes suivantes :

1. Les objets A et E sont les plus proches, ils sont réunis pour former le cluster bleu (A, E).
2. Les objets B et D sont fusionnés pour former le cluster rouge (B, D).
3. Le cluster (A, E) et l'objet C sont fusionnés pour former le cluster jaune ((A, E), C).
4. Enfin, les clusters (B, D) et ((A, E), C) sont fusionnés pour former le plus grand cluster vert.

Ces étapes sont l'équivalent d'un processus algorithmique pour regrouper des objets en fonction de leur proximité. Pour nos besoins, le regroupement est un moyen utile de trier les transcriptions et de visualiser les résultats.

La façon dont les *clusters* sont formées dépend des critères de lien utilisés. Les critères les plus utilisés sont le lien unique et le lien complet. Ces deux critères de couplage produisent différents résultats de regroupement. Avec un lien unique, afin de déterminer quels groupes d'objets constitueront des clusters, nous trouvons les deux objets les plus proches de deux groupes différents et lions leurs groupes associés [Everitt et al., 2001 ; Manning et al., 2009]. Avec le lien complet comme critère, nous utilisons la distance maximale entre les objets de deux groupes différents, ce qui signifie que la mesure de

similarité de deux groupes est déterminée par leurs objets les plus dissemblables [Everitt et al., 2001 ; Manning et al., 2009]. Dans l'exemple que nous donnons pour Figure 13.1, le cluster jaune (A, E, C) et le cluster rouge (B, D) sont fusionnés pour former le cluster vert. Selon que nous utilisons un lien unique ou un lien complet, nous nous appuyons sur des points différents pour créer le cluster vert. La Figure 13.2 montre des exemples de liaison unique et de liaison complète pour cet ensemble de *clusters*. Dans l'exemple illustré, quel que soit le lien que nous utilisons, nous obtenons notre cluster vert et selon que d'autres objets ou clusters soient présents, le résultat pourrait être très différent. Dans notre cas, nous nous appuyons sur une liaison complète pour organiser en clusters des transcriptions, car le critère de liaison complète n'est pas local et implique des structures entières pour composer les clusters [Manning et al., 2009]. Pour nous, c'est une meilleure façon de déterminer des groupes cohérents de transcriptions.

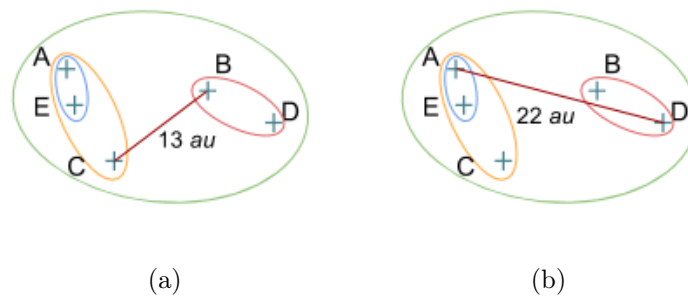


FIGURE 13.2 – Critères de lien : (a) montre un exemple de lien unique et (b) montre un exemple de lien complet.

Les opérations de tri qui permettent la formation de clusters sont exécutées sur une matrice de valeurs de distance, qui sont obtenues à partir de la comparaison de paires d'objets. Cette matrice est ensuite convertie en un format de notation qui est une représentation machine de la proximité des objets. Nous utilisons ensuite les bibliothèques existantes pour les traiter et visualiser les résultats.

13.6.4 Visualisation

Les arbres phylogénétiques peuvent être utilisés comme outil pour visualiser les relations entre les clusters, comme montré dans la Figure 13.1 page 275, où (c) montre un arbre phylogénétique tiré en fonction des groupes de clusters représentés en (a) et leurs valeurs de distance représentées en (b).

Les représentations phylogénétiques sont couramment utilisées pour l'analyse de clusters et un certain nombre de fonctions existent pour cela.

Pour accompagner la visualisation phylogénétique, nous pouvons générer des cartes thermiques, qui s'appuient également sur des valeurs de distance. Les cartes thermiques sont créées en associant des couleurs à des valeurs numériques. Les valeurs de faible distance correspondent aux couleurs douces qui s'intensifient graduellement à mesure que les valeurs de distance augmentent. Les cartes thermiques peuvent également permettre d'identifier les formations de clusters ainsi que leurs limites.

13.6.5 L'ensemble du processus

Pour comparer les transcriptions en fonction de leurs similarités ou différences, nous pouvons appliquer les méthodes que nous avons présentées ici. Pour ce faire, nous avons créé un flux de traitement de documents, que nous représentons dans la Figure 13.3. Le processus global est comme suite :

1. Nous commençons par comparer un lot de transcriptions créées à partir du même objet manuscrit. Pour comparer des textes bruts, nous supprimons tout le balisage XML et appliquons la métrique de distance de Levenshtein. Pour le XML, nous appliquons l'algorithme de Zhang & Shasha.
2. Nous obtenons des valeurs de distance pour toutes les paires de transcriptions, que nous enregistrons dans une matrice. Nous le faisons pour le texte et le XML, ce qui donne deux matrices.
3. Nous utilisons ensuite un algorithme de classification hiérarchique à partir des ma-

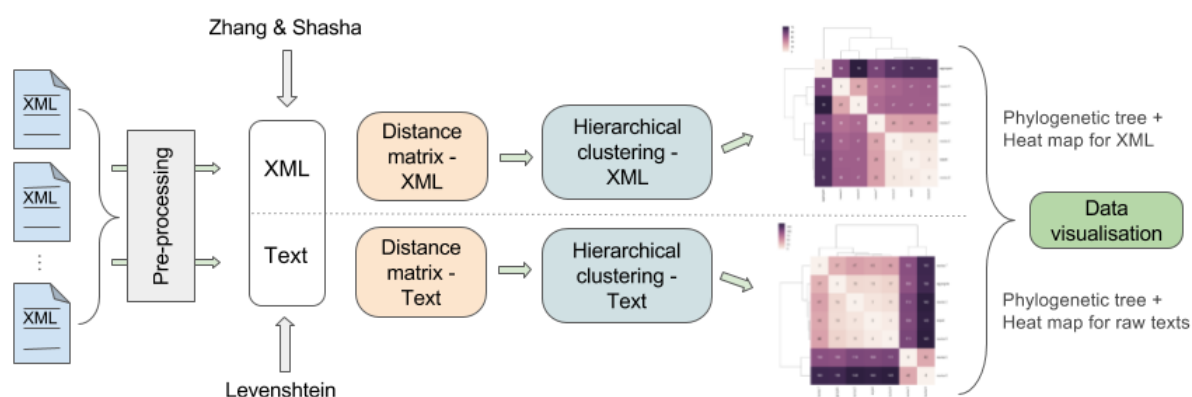


FIGURE 13.3 – Flux schématique de l’analyse de la transcription. Les transcriptions sont pré-traité afin d’extraire soit le texte brut soit la structure xml brut. Puis, la matrice de distance est calculée, et ensuite utilisée pour déduire une classification hiérarchique des textes analysés. Enfin, le résultat peut être représenté avec un arbre phylogénétique et une carte thermique.

trices. Puis, nous générons des arbres phylogénétiques et des représentations de cartes thermiques qui permettent de visualiser les clusters.

Maintenant que nous avons les outils pour évaluer la qualité des transcriptions produites, nous présentons les plateformes que nous avons créées pour tester le flot d’éditeurs et récupérer des données.

13.7 Types de prototypes

Dans un article sur les prototypes, Stan Ruecker présente trois catégories de prototypes, chacune étant utile dans différentes circonstances ou pour répondre à des besoins particuliers. Les catégories sont les suivantes : axée sur la production, axée sur l’expérimentation, et des *provotypes*, ou prototypes axés sur la provocation.

Les prototypes axés sur la production visent à obtenir une version fonctionnelle d’un produit ou d’un système à la fin d’une période de développement donnée [Ruecker, 2015]. Un prototype expérimental est utilisé pour tester ou explorer une idée de recherche. Fi-

nalement, le prototype de provocation a pour objectif de provoquer une réaction des utilisateurs afin de contester la façon dont des personnes ou la société abordent certains sujets.

13.8 Présentation de PHuN 2.0

Notre objectif avec la plateforme PHuN 2.0 (Patrimoine et Humanités Numériques) était de créer un environnement de travail pour des porteurs de projets et des participants. Les fonctionnalités qui ont été développées visent à coordonner les processus de transcription et d'édition dans un cadre collaboratif. La plateforme a été développée à l'aide du framework Symfony, une architecture de Modèle – Vue - Contrôleur (MVC), qui permet de facilement organiser les relations entre la base de données (modèle), les opérations sur les données (contrôleur), et la présentation de ces données (vue).

La plateforme prévoit une structure hiérarchique des rôles des utilisateurs : les porteurs de projets, qui ont un rôle d'administrateurs, les utilisateurs connectés, qui peuvent contribuer aux projets en transcrivant des pages, ainsi que les éditer et les réviser et enfin les visiteurs non-connectés au site qui peuvent apercevoir les pages manuscrites des collections et lire les transcriptions publiées, sans pouvoir transcrire ni modifier les transcriptions.

13.8.1 Administration des projets

Les porteurs de projets peuvent créer des projets de transcription au sein de la plateforme. La création d'un projet implique plusieurs étapes importantes :

- Le dépôt des images des manuscrits à l'administrateur du site, pour que ces images puissent être chargées dans la base de données associée à la plateforme.
- La création d'un projet, avec titre et description.
- Le versement d'une Document Type Description (DTD) qui décrit le schéma d'encodage utilisé par le projet et un Cascading Style Sheet (CSS) associé à la repré-

sentation des éléments du schéma.

- Le paramétrage d'un éditeur de transcription sur la base des éléments comprises dans la DTD et/ou d'autres éléments que les porteurs de projets souhaitent ajouter à l'éditeur.
- Le lancement du projet et l'implication de nouveaux participants dans les activités de transcription et de l'édition/révision.

Au cours d'un projet, les porteurs peuvent également suivre son avancement, soit en relisant et révisant eux-mêmes les transcriptions, soit en étant attentifs aux nouvelles transcriptions publiées, qui peuvent être dévalidées si les porteurs du projet considèrent que la transcription n'est pas assez juste vis-à-vis des attentes. Ils peuvent aussi modifier ou ajuster l'éditeur de transcription pour rajouter ou supprimer des éléments. Enfin, les porteurs de projet peuvent promouvoir des transcripteurs de confiance au même rang qu'eux afin de partager les responsabilités avec ces personnes.

13.8.2 Participation aux projets

La participation à un projet de transcription requiert la création d'un compte utilisateur. Les personnes ayant un compte peuvent sélectionner des pages manuscrites pour commencer des transcriptions et aussi éditer ou réviser les transcriptions faites par d'autres personnes. Le processus de transcription avec ses étapes de transcription, d'édition, de révision et enfin de validation est présenté à la Figure 13.4. Pour commencer, une transcription est créée par un utilisateur et enregistrée dans le système. Ensuite, d'autres utilisateurs peuvent éditer la transcription et modifier son contenu. Une fois qu'un utilisateur considère la transcription prête, il peut l'envoyer en révision. Au cours du processus de révision, trois utilisateurs différents doivent confirmer que la transcription est exacte et / ou apporter des améliorations. A la fin de la révision, la transcription est automatiquement validée par le système. À ce stade, elle est publiée et ne peut plus être modifiée par les utilisateurs ordinaires. Il apparaît dans la liste des transcriptions publiées par les porteurs de projet, mais le processus ne s'arrête pas nécessairement ici. Les chefs de projet peuvent dévalider les transcriptions et les ramener dans le circuit éditorial s'ils considèrent qu'elles

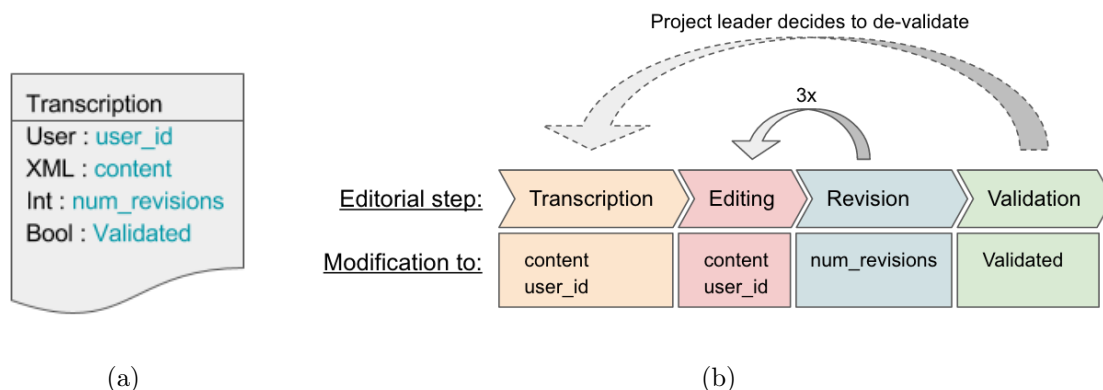


FIGURE 13.4 – L’illustration de la flux de transcription dans PHuN 2.0. (a) La représentation de l’entité de la transcription dans la base de données. (b) Le flux éditorial implémenté dans la plateforme.

ont encore besoin d’améliorations.

Les transcrip-teurs ont aussi accès à leurs pages depuis leurs espaces personnels, qui leur permet de suivre l’évolution d’une transcription qu’ils ont commencée et qui a pu être améliorée par d’autres participants.

Les transcrip-teurs ont aussi accès à une liste de discussion pour chaque page de manuscrit, qui est accessible depuis l’interface de transcription d’une page. Cet espace de discussion sert notamment à poser des questions sur la transcription ou partager des connaissances et des astuces. Ceci peut également être un espace utilisé par des porteurs de projets pour donner plus d’informations contextuelles sur des pages spécifiques et ainsi motiver les participants amateurs de la génétique de texte.

13.8.3 Les fonctionnalités de l’éditeur de transcription

L’éditeur de transcription propose quelques fonctionnalités spécifiques afin de rendre la tâche d’encodage XML plus accessible pour des utilisateurs non-expérimentés. Nous avons adapté un éditeur WYSIWYG (What You See Is What You Get) qui est souvent utilisé pour éditer du HTML. TinyMCE rend possible la création des extensions spécifiques

afin de rajouter à l'éditeur des fonctionnalités particulières. Dans notre cas, il permet d'ajouter des boutons qui seront utilisés pour placer des éléments XML autour d'un texte sélectionné. Par exemple, se fondant sur une entité dans une DTD telle que *ajout*, le système crée un bouton *ajout* et la fonction de ce bouton est d'entourer un texte sélectionné d'une paire de balises `<ajout>` et `</ajout>`. La liste suivante résume les fonctionnalités de l'éditeur :

- Chaque bouton de l'éditeur correspond à un élément XML du schéma de description du projet (ou le vocabulaire), tel qu'il a été défini par des porteurs de projets.
- Chaque bouton a une description expliquant la fonction de l'élément correspondant.
- Les éléments peuvent être organisé soit directement dans la barre d'outils de l'éditeur, soit regroupés dans un menu en haut de l'éditeur.
- Nous avons ajouté une extension existante de TinyMCE, qui permet d'accéder au code brut du fichier si nécessaire.

L'éditeur permet d'encoder et structurer le contenu simplement en sélectionnant le texte voulu avec le curseur et ensuite en cliquant un des boutons qui correspond à l'élément que l'utilisateur souhaite placer autour du texte.

13.8.4 Conclusions

En développant le prototype de la plateforme axée sur la production, nous étions davantage intéressée par la création d'un environnement de travail doté de fonctionnalités et de caractéristiques comparables aux environnements de travail numériques existants. Les questions soulevées par ce processus de prototypage étaient en effet pertinentes et importantes car elles concernaient les défis associés au traitement des données encodées, à la gestion des flux de travail, à la création d'outils personnalisables et à la création d'interfaces pour les utilisateurs. Pour nous, ce qui manquait était la composante expérimentale et analytique des types de données qu'il est possible de produire en utilisant le crowdsourcing. Nous avons créé une plateforme pour expérimenter avec les transcriptions issues du crowdsourcing et nous n'avions toujours aucun moyen d'évaluer comment ces transcrip-

tions se situent par rapport à celles de spécialistes ou de personnes formées. Pour résoudre ce problème, il a fallu créer un deuxième prototype, une version adaptée de la plateforme PHuN 2.0 d'origine, nommée PHuN-ET (Plateforme des Humanités Numériques - Espace Transcription).

13.9 Présentation de PHuN-ET

Notre besoin était de maintenir l'accès aux pages au plus grand nombre d'utilisateurs possible et d'obtenir des transcriptions produites en une séance ininterrompue. La première condition ressemble plus étroitement à des conditions de crowdsourcing. La seconde aide à limiter les effets de la variabilité liée aux changements de concentration, de fatigue ou d'environnement qui peuvent accompagner le travail d'une transcription au cours de plusieurs sessions. Il est également plus facile d'évaluer les données résultantes si chaque transcription est effectuée du début à la fin en une seule séance par un unique transcripateur.

13.9.1 Navigation dans la plateforme

Dans PHuN-ET nous avons remplacé le catalogue des projets de PHuN 2.0 avec un accès séquentiel aux pages du projet. La décision d'avoir un accès séquentiel a l'avantage de simplifier la navigation pour les participants, qui trouvent plus rapidement et plus facilement l'interface de transcription pour un projet en cours. Notre objectif était de minimiser le nombre de pas qu'un utilisateur doit effectuer entre son inscription et l'activité de transcription.

13.9.2 Administration des projets

Le système d'administration de projets a été repris de la plateforme PHuN 2.0, permettant de configurer entièrement l'éditeur utilisé dans les expériences menées.

13.9.3 Participation aux projets

Comme dans PHuN 2.0, la participation aux projets dans la plateforme PHuN-ET nécessite que l'utilisateur soit inscrit sur la plateforme. Avoir un compte d'utilisateur permet aux participants de transcrire les pages qui leur sont proposées par la plateforme. Pour les besoins de l'expérimentation, une fois que l'utilisateur enregistre sa transcription il ou elle ne peut plus la modifier, elle est enregistrée dans le système. Cependant, l'utilisateur peut consulter ses transcriptions effectuées dans son espace personnel. L'utilisateur a accès à la vue de ses transcriptions ainsi qu'aux documents XML produits, qui sont téléchargeables depuis la plateforme ou consultable en tant qu'arbre XML dans une nouvelle fenêtre. L'utilisateur a également la possibilité de partager ses transcriptions avec d'autres personnes en envoyant un lien vers la vue d'une page depuis un des réseaux sociaux proposés (Facebook, LinkedIn, Twitter, Google+, et Pinterest). L'intérêt est de faciliter le partage de leur travail et de diffuser la plateforme auprès d'un plus grand nombre de personnes avec comme objectif d'acquérir plus de participants.

13.10 Au-delà des plateformes

Les environnements virtuels jouent un rôle crucial dans l'organisation des matériaux numériques, des flux de travail et des contributions. Les technologies numériques et web jouent incontestablement un rôle important dans la définition des pratiques de travail humaines. Pourtant, les êtres humains ont une influence considérable sur les évolutions des projets, ce qui va bien au-delà de la simple définition et exécution de tâches à l'aide d'outils et d'environnements virtuels. Nous considérons l'importance de la collaboration, de la communication et de la sensibilisation, et enfin, des compétences et de la formation, aux projets de crowdsourcing.

13.10.1 Collaboration

La définition du mot collaboration provient du latin *com-* pour dire ensemble et *labōrāre* de travailler². Ceci est différent du mot contribuer, qui vient du latin *contri-**buere* collecter, de *tribuere* d'accorder³.

Plusieurs études se sont intéressées au phénomène de crowdsourcing en examinant les actions sociales comme appartenant à des intentions collectives en mode-je ou en mode-nous, où le mode-je est considéré comme une intention personnelle et indépendante et le mode-nous est une intention interdépendante orientée vers le groupe [Bagozzi, 2000 ; Shen et al., 2014]. L'étude de [Shen et al., 2014] sur les participants de Wikipédia recueille des preuves empiriques pour soutenir que les intentions en mode-je et en mode-nous ont un impact sur le comportement contributif et que la principale différence concerne les facteurs relationnels de confiance et d'engagement, qui impactent seulement le mode-nous. Bien que la contribution soit possible à la fois en mode-je et en mode-nous, cette étude nous amène à suggérer qu'il serait important d'examiner les facteurs relationnels pour aborder les différences entre la collaboration et la contribution. Par exemple, alors que la collaboration et la contribution peuvent être spontanées et de courte durée, de nombreux projets sont intéressés par l'engagement à long terme des participants. Ainsi, en mettant l'accent sur des facteurs relationnels tels que l'engagement et la confiance lors de l'examen des modèles participatifs, il est possible d'orienter plus précisément et de façon plus transparente les aspects souhaitables dans les projets de crowdsourcing, en particulier lorsque les termes *collaborer* et *contribuer* sont souvent utilisés de manière interchangeable.

13.10.2 Motivations

Sachant que la motivation est l'un des facteurs qui impactent les intentions et les actions des individus, il est intéressant de regarder de plus près les différents types de

2. Selon la source : <https://www.collinsdictionary.com/dictionary/english/contribute>.

3. Selon la source : <https://www.collinsdictionary.com/dictionary/english/contribute>.

motivation et considérer leurs effets sur des participants. Les deux types de motivation qui ont été identifiés sont la motivation de type intrinsèque et la motivation de type extrinsèque. La motivation extrinsèque influence les actions des individus sur la base d'une compensation, souvent matérielle, pour un service rendu. Cependant, les volontaires, qui participent aux projets de crowdsourcing, ne sont pas forcément compensés. Cela tend à montrer que les bénéfices extrinsèques ne sont pas nécessairement des facteurs clés de motivation. Il existe d'autres types de récompenses extrinsèques que des récompenses monétaires, elles peuvent prendre la forme de statut social, de réseautage, et d'attribution de crédit. Pourtant, il y a aussi de la motivation intrinsèque, qui peut amener des individus à participer à des activités sans récompense, car ces individus sont motivés par un intérêt pour l'activité en question et le plaisir et la satisfaction qu'ils gagnent en participant. Les études montrent que les personnes intrinsèquement motivées sont plus susceptibles de réussir dans le domaine qui les motive et dans un contexte académique, par exemple, cela se traduit par de meilleures notes et un travail de meilleure qualité par les étudiants [Ryan and Deci, 2000]. Il est intéressant donc pour des projets d'avoir des participants qui sont intrinsèquement intéressés par le sujet dans lequel s'ancre le projet. Cela a plus de chance de produire de meilleurs résultats, mais aussi de constituer des communautés de plus longue durée.

13.10.3 Communication et sensibilisation

Pour de nombreuses initiatives de crowdsourcing, des campagnes de communication et de sensibilisation efficaces au-delà de la plateforme sont fondamentales pour la réussite du projet. Les projets peuvent considérer les avantages d'avoir plusieurs sites de représentation sur différentes plateformes de réseaux sociaux. Ceci peut permettre au projet d'étendre sa sphère d'influence et augmenter les chances que des personnes intéressées croisent le projet via ces différents réseaux. Des modes de communication plus traditionnels peuvent aussi être bénéfiques : des journaux, des magazines, ou la radio peuvent aider à exposer le projet à un public plus large. Il y a des projets, comme Zooniverse, qui se servent de leur base d'utilisateurs existants pour leur proposer de participer à de nouveaux

projets récemment lancés. Avec leurs emails réguliers, Zooniverse ne se laisse pas facilement oublier par ses publics. En effet, certaines techniques de communication utilisées par des projets de crowdsourcing ressemblent étroitement aux campagnes de marketing des entreprises. Il y a donc un intérêt de maîtriser certaines de ces techniques afin d'atteindre un public plus large. Enfin, les projets de sciences citoyennes et humanités citoyennes se regroupent souvent en réseaux, avec des sites qui accueillent plusieurs projets tels que les sites *Connected Communities*⁴ et *SciStarter*⁵.

13.10.4 Compétences et formation

L'acquisition de nouvelles compétences numériques et la formation sont importantes pour travailler collaborativement dans le contexte des humanités numériques. Ceci devrait inclure les chercheurs ainsi que des participants volontaires. Proposer les formations aux volontaires permet de s'assurer de la qualité du travail. De plus, les connaissances acquises peuvent permettre aux volontaires d'améliorer leur portefeuille professionnel ou accéder à des formations auxquelles ils n'auraient pas accès autrement.

Le crowdsourcing étant souvent organisé via des plateformes et sites web, il est important de prendre conscience que la formation de volontaires peut elle aussi prendre place en ligne. Lorsque les projets reposent sur des volontaires, écrire des protocoles clairs est de la plus haute importance, tout comme la définition des tâches de niveaux de difficulté appropriés pour assurer l'exactitude des données résultantes [Cohn, 2008]. La même attention doit être accordée à des instructions en ligne que dans d'autres cas, d'autant plus que l'information est transmise à distance. La conception de protocoles pour différents types de tâches peut être un excellent terrain d'étude sur l'efficacité de l'utilisation de la formation en ligne et le travail en autonomie dans le cas du crowdsourcing.

4. <https://connected-communities.org>

5. <https://scistarter.com/>

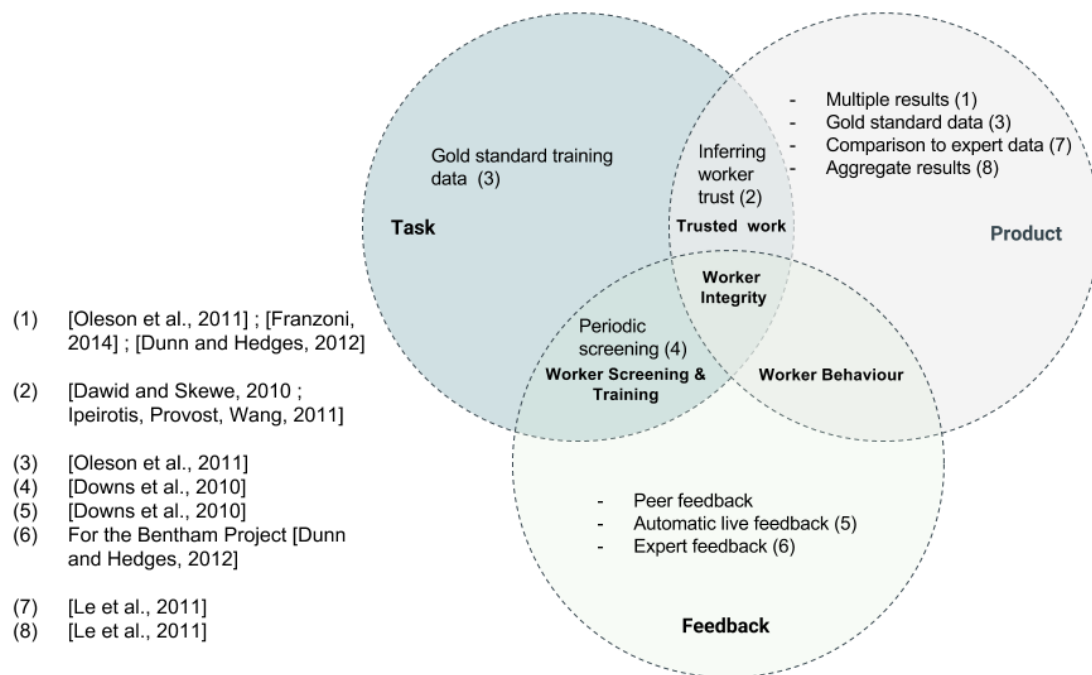


FIGURE 13.5 – L'illustration de la relation entre les tâches, la rétroaction, et le produit dans un environnement de crowdsourcing.

13.11 L'assurance qualité pour le crowdsourcing

Depuis que la participation de publics non-experts aux projets de recherche gagne en popularité, des questions sur le contrôle et l'assurance qualité deviennent aussi centrales au sujet de crowdsourcing. De nombreuses techniques ont été testées par les crowdsourcers industriels, telles que l'utilisation de données de référence dans la formation de participants, de diverses formes de rétroaction (feedback), et le principe de reproduction d'une même tâche par plusieurs personnes [Le et al., 2010]. Ces techniques présentent un certain nombre de solutions intéressantes au problème de la qualité dans un contexte de crowdsourcing industriel.

Nous abordons le sujet de l'assurance qualité, en regardant des méthodes centrées sur les tâches, celles centrées sur la rétroaction et finalement celles centrées sur le produit. Nous avons représenté chacun des trois aspects et leurs interactions dans la Figure 13.5. Dans cette figure, nous avons positionné les procédures d'assurance qualité dans leurs principales composantes d'influence (tâche, rétroaction, et produit), mais quelques-unes

se situent dans des zones de chevauchement résultant de l'interaction de deux composantes. Ces zones montrent comment les composants se combinent pour produire des effets sur les bénévoles et leur travail. Par exemple, la procédure de sélection et la formation des personnes impliquent d'assigner des tâches et de fournir des commentaires sur l'accomplissement de ces tâches. Le produit, ou le résultat rendu, permet de s'assurer que le comportement du participant est adéquat vis-à-vis de la tâche, mais ce comportement est également influencé par la rétroaction qu'il ou elle reçoit. Enfin, le participant rend un produit de confiance quand l'activité est accompli conformément aux attentes. Les trois régions de chevauchement, c'est-à-dire la formation, un comportement exemplaire et un travail de confiance, garantissent l'assurance de l'intégrité des participants.

13.11.1 Assurance qualité fondée sur la tâche

L'assurance qualité fondée sur la tâche implique plusieurs procédures : la sélection de participants et la formation de participants. Ces procédures reposent sur l'utilisation de données de test dont les réponses sont connues (les données étalons). Ces données de référence permettent, dans le cas de la sélection de participants, d'évaluer la capacité des participants à accomplir correctement les tâches demandées. En ce qui concerne la formation des participants, elle peut être organisée en s'appuyant sur ces données, ce qui permet de mettre en place des systèmes de corrections automatiques pour guider l'utilisateur.

13.11.2 Assurance qualité reposant sur la rétroaction

L'assurance qualité fondée sur la rétroaction peut inclure une rétroaction délivrée par des experts ou des pairs sur les tâches effectuées par les participants. La rétroaction peut être automatisée et livrée au bénévole lorsque celui-ci est actif, ou bien, en réponse à ses actions. La rétroaction peut être distribuée au cours de la formation des participants, pour qu'ils apprennent comment répondre correctement aux tâches. La rétroaction, dans un autre cadre que celui de la formation, peut aider à maintenir la qualité des transcriptions

fournies par les participants, y compris prévenir le travail subalterne ou l'escroquerie du système.

13.11.3 Assurance qualité reposant sur le produit

L'assurance qualité fondée sur le produit interagit avec le composant de la tâche. L'objectif est de créer, pour un contributeur, un profil établi sur les transcriptions qu'il est capable de produire. Les méthodes qui évaluent la production peuvent comprendre la comparaison de productions multiples entre elles, la comparaison de productions par rapport à des réponses correctes connues (ou des données d'experts / étalons), et enfin l'utilisation de réponses multiples afin d'agréger les résultats. La comparaison aux données étalons est finalement similaire aux méthodes déjà évoquées et pourra permettre une meilleure connaissance des compétences des participants. C'est aussi une façon d'évaluer les résultats produits sous différentes conditions qui peuvent être contrôlées par des projets. Par exemple, cela permettrait de mieux comprendre quelles instructions produisent de meilleurs résultats, ou quelles types de pages sont plus ou moins difficiles à transcrire. Enfin, l'usage de multiples productions qui peuvent être comparées entre elles permet d'agréger les résultats dans le but de produire la réponse la plus probable ou celle qui reflète le mieux les réponses des participants. Cette méthode d'agrégation de textes peut être efficace pour filtrer les erreurs en s'appuyant sur de multiples transcriptions. Par exemple, cela peut être utilisé sur un lot de transcriptions contenant un mot difficile dans un manuscrit. Si nous pouvons compter sur les utilisateurs pour reconnaître correctement le mot en question la majorité du temps, nous pouvons générer une transcription qui contient le mot correctement orthographié, et filtrer les variantes orthographiées à tort.

13.12 Mesurer la qualité des transcriptions

Afin de pouvoir évaluer la qualité des transcriptions qui ont été produites par des transpositeurs inexpérimentés nous avons utilisé les méthodes décrites dans le Chapitre 6, et ici dans la Section 13.6 page 272, et nous avons réalisé des expériences sur le corpus

de Stendhal et celui de Benoîte Groult.

13.12.1 Expérimentation sur Stendhal

Une expérience initiale a été réalisée sur un échantillon du Corpus Stendhal. Pour notre étude, nous avons sélectionné deux pages de ce corpus. Ces pages sont visibles dans les Figures 11.2 page 201 et 11.3 page 202.

Après avoir identifié notre échantillon et validé le schéma XML avec nos collaborateurs experts, nous leur avons demandé de transcrire les deux pages en utilisant Oxygen XML Author⁶. Nous avons ensuite demandé à dix autres personnes (non-experts) de répéter le même exercice, en s'appuyant sur les instructions tirées du manuel.

Nous avons procédé à l'analyse phylogénétique des résultats en utilisant la distance de Levenshtein sur des textes extraits des transcriptions ainsi que sur le XML produit. Les résultats ont été organisés hiérarchiquement en clusters et des arbres phylogénétiques ont été générées. Nous avons observé le regroupement de nos experts dans un même cluster et notamment avec les novices 9 et 10 pour l'analyse du texte et avec seulement novice 9 pour l'analyse du XML. Sachant que novice 9 a été formé par les experts eux-mêmes, nos résultats confirment que les experts produisent des résultats différents des transpositeurs inexpérimentés, et aussi que les transpositeurs formés se rapprochent beaucoup plus des experts. Nous avons aussi pu constater les raisons clés de la différence des résultats, que nous attribuons à la non-reconnaissance, par des transpositeurs inexpérimentés, des composants des pages tels que les éléments marginaux (foliotations, paginations, marginaux). Nous proposons donc que les porteurs de projets prennent en considération que leurs besoins d'encodage peuvent être partiellement remplis par des participants inexpérimentés (le texte ainsi que l'encodage des modifications et features bien distincts) et ensuite complétés par des experts (foliotations et autres éléments marginaux qui demandent une connaissance plus précise du manuscrit).

Nous avons aussi pu constater que les experts peuvent être mis à contribution pour éta-

6. https://www.oxygenxml.com/xml_author.html

blir des données étalons, puisqu'ils produisent des résultats très similaires. Enfin, il serait considérablement plus facile de continuer des tests avec une plateforme de transcription en ligne.

13.12.2 Expérimentations sur Benoîte Groult

La première expérience sur le corpus de Benoîte Groult a permis de prendre des marques et faire les améliorations au système et à l'éditeur de transcription, ainsi qu'aux instructions fournies aux transcrip-teurs. La deuxième expérience a permis de collecter des transcriptions de 24 participants différents, dont 14 lors de deux ateliers et le reste venant des interventions non encadrées sur le site. L'analyse phylogénétique a été effectuée sur les textes et les documents XML. Nous n'avons pas observé des distinctions particulières entre les contributions des participants d'un des ateliers ou venues des participants libres. Comme pour l'expérimentation précédente de Stendhal, la méthode s'est avérée appropriée pour observer la distribution de variabilité entre contributions, ainsi que de repérer celles qui sont les plus proches de notre référence expert. Les résultats nous suggèrent aussi, compte tenu de la présence d'instructions appropriées, que cette activité est très propice à être réalisée à distance.

Nous avons également constaté que les instructions pour ce type d'activité jouent un rôle important pour les utilisateurs puisqu'elles décrivent les procédures et les attentes de l'activité. Néanmoins, les instructions doivent être concises et précises afin de ne pas embrouiller ou décourager les participants potentiels. À l'inverse, elles peuvent également être conçues comme un moyen de filtrer les participants.

Nous avons constaté que l'organisation d'un atelier était utile pour attirer les participants. Cependant, au-delà des deux sessions organisées, nous n'avons pas de moyens pour nous assurer que les participants continueront à revenir sur le site, sans encouragement supplémentaire.

13.13 Mesurer les facteurs de complexité

Au cours de notre travail nous avons rencontré des facteurs de complexité qu'il est possible d'attribuer à différents domaines de l'activité de la transcription. Notamment, pour décrire les facteurs qui contribuent à rendre une tâche de transcription plus ou moins complexe, trois catégories ou familles de facteurs ont été identifiées. Chacun de ceux-ci comprenant un groupe de facteurs, et chacun contribuant d'une certaine façon à la complexité d'une tâche de transcription. Ces trois catégories de complexités sont celle liée au transcripateur lui-même, celle liée au schéma descriptif utilisé, et celle liée à la page manuscrite.

La première catégorie peut inclure des facteurs tels que : l'âge, l'expérience de la transcription, l'expérience des manuscrits, l'expérience de l'encodage XML, l'expérience de l'utilisation des éditeurs en ligne, la compétence en français, la motivation, le temps libre, la formation professionnelle ou l'aptitude professionnelle, ainsi que des facteurs liés aux sphères sociales, socio-culturelles, ou psychologiques des participants. Cela comprend beaucoup de facteurs qui ne peuvent pas être identifiés en totalité, ni étudiés ici pour des raisons d'éthique.

La seconde catégorie concerne le schéma descriptif, ou schéma XML utilisé pour encoder les manuscrits. Lorsque les porteurs de projets ont la possibilité d'utiliser leurs propres vocabulaires, il est possible d'imaginer d'atteindre un niveau plus élevé de description. Pourtant, des éléments de vocabulaire plus fins ou plus nuancés peuvent également ajouter à la complexité de la tâche de transcription. On peut imaginer le nombre d'éléments ou les relations hiérarchiques entre eux comme étant des sources de complexité.

La troisième catégorie de complexité provient de la page manuscrite elle-même. Nous avons pu observer certains de ces facteurs lors des expérimentations sur le corpus de Stendhal et de Benoîte Groult. Ces facteurs peuvent inclure des éléments de la liste suivante.

1. **Nombre de lignes** : Le nombre d'erreurs peut augmenter avec le nombre de lignes contenues dans la page.
2. **Nombre d'additions** : Les ajouts sont souvent plus petites que le texte principal,

reduisant ainsi leur lisibilité.

3. **Nombre de suppressions** : L'écriture raturée peut être plus difficile à déchiffrer.
4. **Nombre d'outils d'écriture** : Certains outils augmentent la difficulté de lecture (feutre, crayon).
5. **Nombre de types d'écriture (manuscrit ou imprimé)** : Une page imprimée sera sans doute plus facile à transcrire qu'une page manuscrite.
6. **La taille de l'écriture** : Plus petit sera l'écriture, plus difficile elle sera à déchiffrer.
7. **L'angle de l'inclinaison de l'écriture** : Comme la taille, l'inclinaison du script peut affecter la lisibilité d'une page.
8. **Nombre de figures** : Les figures sont des éléments qui cassent la linéarité dans une page, elles peuvent rendre plus difficile l'identification de lignes et leur structuration.
9. **Nombre de caractères spéciaux** : Certains caractères spéciaux ne sont pas présents dans les claviers des utilisateurs, les rendant plus difficile à représenter.

L'identification et l'étude de certains de ces facteurs ont présenté une opportunité d'expérimentation sur le corpus de Benoîte Groult en utilisant un plan d'expérience, nommé en anglais *Design Of Experiments* (DOE).

13.13.1 Le plan d'expériences

Pour étudier l'effet de ces facteurs, on utilise les plans d'expériences (*Design of Experiments* ou DOE). Cette méthode repose sur des expériences qui permettent de mettre en lumière l'importance des facteurs étudiés. Si nous notons k le nombre de facteurs étudiés, le nombre d'expériences à réaliser est égal à 2^k . Par exemple, s'il y a 9 facteurs à étudier, il faut effectuer 2^9 expériences, c'est-à-dire 512 expériences. Ce nombre donne le nombre minimal d'expériences à effectuer pour obtenir un maximum de résultats fiables [Fisher, 1937].

Une fois tous les facteurs sont identifiés, nous pourrons aussi prendre en compte des interactions entre deux ou trois différents facteurs. Ensuite, nous pouvons mesurer l'effet

total des facteurs, que l'on noterait C_{total} . La formule complète utilisée dans les plans d'expérience ressemble à ceci :

$$C_{total}(x_1, \dots, x_k) = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^{k-1} \sum_{j>i}^{k-1} \beta_{i,j} x_i x_j \quad (13.2)$$

Dans cette formule, x_i représente un facteur de complexité, appartenant à l'une des trois catégories identifiées. β_i est le poids associé avec un facteur x_i , et β_0 représente une valeur constante du modèle. $\beta_{i,j}$ est le poids de l'interaction entre deux facteurs x_i et x_j . Le principe du plan d'expériences est de trouver les valeurs de tous les facteurs inconnus du modèle, y compris β_0 et tous les β_i et $\beta_{i,j}$, en mettant en place un minimum nombre d'expériences. Le modèle prend en compte aussi tous les facteurs x_i ainsi que toutes les interactions possibles $x_i x_j$.

Nous avons choisi seulement deux facteurs à étudier afin de mettre en place un nombre réalisable d'expériences sur le corpus Benoîte Groult. Notre plan d'expérience est donc composé de $2^2 = 4$ expériences. Afin de le réaliser, nous avons sélectionné 4 pages du corpus Benoîte Groult et chaque page aura une combinaison particulière des deux facteurs suivants :

- Nombre de modifications($X1$) : Nous considérons une modification comme étant un ajout, une suppression, ou une correction.
- La taille de script (hauteur x largeur en pixels) ($X2$) : Nous le calculons sur la base des dimensions du "e" dans les pages, considérée comme la lettre la plus fréquente dans la langue française.

Pour réaliser le plan nous avons donc besoin de 4 pages qui correspondent aux critères suivants :

1. Une page ayant peu de modifications et une grande écriture.
2. Une page ayant beaucoup de modifications et une grande écriture.
3. Une page ayant peu de modifications et une petite écriture.
4. Une page ayant beaucoup de modifications et une petite écriture.

13.13.2 L'analyse des données

Après avoir trouvé les quatre pages qui correspondent aux critères évoqués, nous avons demandés à une dizaine de personnes de transcrire chacun les quatre pages et ensuite nous avons analysé les résultats. Les résultats montrent l'écart de chaque individu par rapport à la transcription d'expert pour chacune des pages. Nous avons constaté des valeurs plus faibles pour les expériences 1 et 3, tandis que les expériences 2 et 4 présentent des valeurs élevées. Ceci correspond aussi aux pages ayant le plus grand nombre de modifications.

Nous avons analysé les résultats afin d'extraire les coefficients β_0 , β_1 , β_2 et $\beta_{1,2}$, en utilisant le logiciel d'analyse MODDE⁷.

Les coefficients β nous donnent les effets des deux facteurs sur les erreurs résultantes des transpositeurs non-experts. La réalisation d'expériences telles que celles-ci nous permettent de déterminer quels sont les facteurs les plus importants pour un corpus de pages particulier, ou un ensemble de pages dans un corpus. L'utilisation d'un DOE peut nous permettre d'obtenir des résultats fiables avec un nombre minimum d'expériences.

Après avoir effectué ce DOE sur plusieurs participants, on peut s'attendre à ce que dans l'échantillon de Benoîte Groult, l'un des deux facteurs testés, celui de la taille du script, ait un effet minimal ou nul sur les résultats.

Le fait que l'on puisse dire avec certitude que les modifications déterminent réellement la difficulté, même sur un corpus considéré comme très accessible, est très prometteur. Les résultats suggèrent que le tri des pages en fonction du nombre de modifications peut être un moyen de classer la complexité de la page pour le bénéfice des utilisateurs ayant différents niveaux d'expérience. Nous avons ainsi pu estimer la difficulté relative des 4 pages testées.

7. <http://umetrics.com/kb/modde-12>

13.14 Conclusion et perspectives

Les processus éditoriaux numériques d'aujourd'hui sont de plus en plus innovants en ce qui concerne les moyens qui permettent à l'information et à la connaissance d'arriver dans les mains des lecteurs. La transcription de manuscrits est une étape essentielle pour transmettre des écrits, et en particulier pour de nombreux documents et artefacts non édités, qui autrement ne seraient tout simplement pas accessibles aux lecteurs.

Les opportunités offertes par les outils numériques pour la recherche et les processus éditoriaux ont des implications considérables pour la production de ressources textuelles. Nous avons discuté du codage textuel en utilisant le XML, un langage de balisage qui s'adapte de manière appropriée au potentiel descriptif des objets manuscrits. La transcription, ou le travail de transformation des fac-similés numériques en textes exploitables et lisibles par machine, implique également l'encodage de ces textes à l'aide d'une forme de balisage descriptif. Ensuite, en fonction du balisage, différents formats de sortie peuvent être conçus.

Nous avons également étudié des outils et des environnements pour la transcription manuelle et l'encodage des données. Notre objectif était de créer un éditeur facilement accessible aux publics en ligne, tout en permettant aux chercheurs de décider entièrement des vocabulaires descriptifs utilisés pour encoder et structurer les objets avec lesquels ils choisissent de travailler. Un éditeur WYSIWYG (*What You See Is What You Get*) pour XML fournit une interface entre les utilisateurs et le codage XML, ce qui permet d'atténuer certains des aspects plus techniques associés aux tâches de transcription. Nous avons aussi développé un environnement de transcription qui permet de gérer l'effort collaboratif (PHuN 2.0) et un environnement d'expérimentation qui permet d'acquérir plus de données de crowdsourcing tout en faisant attention à l'expérience de l'utilisateur sur la plateforme (PHuN-ET).

Finalement, nous avons mené des investigations pour mesurer la qualité des transcriptions contribuées par des publics inexpérimentés. Ceci dans le but de comprendre comment mieux organiser ce genre de démarche et afin de connaître les différents facteurs

qui contribuent à la variabilité des résultats.

Les perspectives s'avèrent très prometteuses pour l'implication de nouvelles personnes dans l'activité de la constitution de ressources et de textes encodées. Nous pouvons imaginer la création de systèmes plus intelligents pour encadrer des participants et assurer la qualité de leur contributions, mais aussi de fournir un moyen pour les porteurs de projets de s'impliquer dans l'évaluation des contributions de toutes ces personnes à l'aide d'outils plus puissants. Nous envisageons aussi l'amélioration continue des interfaces proposées aux utilisateurs de ces systèmes, ce qui aura sans doute un impact positif sur leurs résultats et leurs réactivité vis-à-vis de ces projets. Enfin, nous pouvons anticiper des nouvelles implémentations dans d'autres genres de projets et pour d'autres disciplines, tels qu'en histoire de l'art, en archéologie, ou en sciences du langage. Le développement et l'amélioration des outils permettant de travailler avec des textes numériques pourra impliquer des progrès tels qu'améliorer la flexibilité des configurations de projets, la gestion des flux de données, le renforcement du suivi et de l'évaluation des résultats, et l'intégration de plus de mécanismes de rétroaction. Finalement, ceci impliquera aussi plus d'expérimentation afin de mieux comprendre les effets des compétences individuelles, des schémas, et des objets manuscrits sur les résultats de contributions. Les connaissances ainsi recueillies informeront, encourageront et soutiendront sans doute de meilleures pratiques en édition scientifique qui impliquent le soutien du public, et peut-être en crowdsourcing de manière plus générale.

Bibliography

- Arfon (2013). Making the zooniverse open source.
- Bagozzi, R. P. (2000). On the concept of intentional social action in consumer behavior. *Journal of Consumer research*, 27(3):388–396.
- Barber, J. F. (2016). Digital storytelling: New opportunities for humanities scholarship and pedagogy. *Cogent Arts & Humanities*, 3(1):1181037.
- Baudet, J. (2003). *De l’outil à la machine: histoire des techniques jusqu’en 1800*. Vuibert.
- Brown, A. W. and Allison, D. B. (2014). Using crowdsourcing to evaluate published scientific literature: methods and example. *PloS one*, 9(7):e100647.
- Buard, P.-Y. (2015). *Modélisation des sources anciennes et édition numérique*. PhD thesis, Université de Caen.
- Buecheler, T., Sieg, J. H., Füchslin, R. M., and Pfeifer, R. (2010). Crowdsourcing, open innovation and collective intelligence in the scientific method-a research agenda and operational framework. In *ALIFE*, pages 679–686.
- Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., and Schnapp, J. (2012). *Digital_Humanities*. Mit Press.
- Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using amazon’s mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics.

- Cao, H. and Govindaraju, V. (2007). Template-free word spotting in low-quality manuscripts. In *Proceedings of the 6th International Conference on Advances in Pattern Recognition*, pages 135–139.
- Causser, T. and Terras, M. (2014). Many hands make light work. many hands together make merry work: transcribe bentham and crowdsourcing manuscript collections. *Crowdsourcing Our Cultural Heritage*, pages 57–88.
- Chanal, V. and Caron-Fasan, M.-L. (2008). How to invent a new business model based on crowdsourcing: the crowdspirit® case. In *Conférence de l'Association Internationale de Management Stratégique*, pages 1–27.
- Chawathe, S. S., Rajaraman, A., Garcia-Molina, H., and Widom, J. (1996). Change detection in hierarchically structured information. In *ACM SIGMOD Record*, volume 25, pages 493–504. ACM.
- Chignard, S. (2012). *L'open data: comprendre l'ouverture des données publiques*. Fyp.
- Cho, H., Chen, M., and Chung, S. (2010). Testing an integrative theoretical model of knowledge-sharing behavior in the context of wikipedia. *Journal of the Association for Information Science and Technology*, 61(6):1198–1212.
- Coady, L. (2016). *Who Needs Books?: Reading in the Digital Age*. University of Alberta.
- Cohn, J. P. (2008). Citizen science: Can volunteers do real research? *BioScience*, 58(3):192.
- Communities, C. (2016). Connected communities website.
- Consortium, T. (2017). P5: Guidelines for electronic text encoding and interchange: Representation of primary sources. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html>. Accessed: 2017-09-05.
- Crompton, C. and Siemens, R. (2013). Introduction: Research foundations for understanding books and reading in the digital age: Text and beyond. *Scholarly and Research Communication*, 3(4).

- Cui, Y. and Roto, V. (2008). How people use the web on mobile devices. In *Proceedings of the 17th international conference on World Wide Web*, pages 905–914. ACM.
- de Stendhal, E. M. (2017). Les manuscrits de stendhal.
- Diem, M. and Sablatnig, R. (2009). Recognition of degraded handwritten characters using local features. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 221–225. IEEE.
- Diem, M. and Sablatnig, R. (2010). Recognizing characters of ancient manuscripts. In *IS&T/SPIE Electronic Imaging*, pages 753106–753106. International Society for Optics and Photonics.
- DiNucci, D. (1999). Fragmented future. *Print*, 53(4):32.
- DiPalantino, D. and Vojnovic, M. (2009). Crowdsourcing and all-pay auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 119–128. ACM.
- Dord-Crouslé, S. (2010). Vers une édition électronique des dossiers de bouvard et pécuchet.
- Dow, S. P., Bunge, B., Nguyen, T., Klemmer, S. R., Kulkarni, A., and Hartmann, B. (2011). Shepherding the crowd: managing and providing feedback to crowd workers. In *In Ext. Abstracts CHI 2011, ACM Press*, pages 1669–1674.
- Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. (2010). Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2399–2402. ACM.
- Dufournaud, N. (2014). Des humanités aux données. *Les Cahiers du numérique*, 10(3):73–88.
- Dunn, S. and Hedges, M. (2012). Crowd-sourcing scoping study. engaging the crowd with humanities research. *Centre for e-Research, King's College London*.
- Espana-Boquera, S., Castro-Bleda, M. J., Gorbe-Moya, J., and Zamora-Martinez, F. (2011). Improving offline handwritten text recognition with hybrid hmm/ann models. *IEEE transactions on pattern analysis and machine intelligence*, 33(4):767–779.

- Estellés-Arolas, E. and González-Ladrón-De-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200.
- Eveland Jr, W. P. and Dunwoody, S. (2016). Users and navigation patterns of a science world wide web site for the public. *Public understanding of science*.
- Everitt, B., Landau, S., Leese, M., and Stahl, D. (2001). Cluster analysis. 2001. *Arnold, London*.
- Fisher, R. A. (1937). *The design of experiments*. Oliver And Boyd; Edinburgh; London.
- Fitzpatrick, K. (2011). *Planned obsolescence: Publishing, technology, and the future of the academy*. NYU Press.
- Flaubert, C. (2017). Les manuscrits de madame bovary: édition intégrale sur le web.
- Franzoni, C. and Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy*, 43(1):1 – 20.
- Galey, A. and Ruecker, S. (2010). How a prototype argues. *Literary and Linguistic Computing*, 25(4):405–424.
- Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1995). *Design patterns: elements of reusable object-oriented software*. Pearson Education India.
- Gatos, B., Pratikakis, I., and Perantonis, S. J. (2006). Adaptive degraded document image binarization. *Pattern recognition*, 39(3):317–327.
- Ghafele, R., Gibert, B., and DiGiammarino, P. (2011). How to improve patent quality by using crowdsourcing. *Innovation management, Sept*.
- Glenn, N. D. (2005). *Cohort analysis*, volume 5. Sage.
- Goupy, J. and Creighton, L. (2013). *Introduction aux plans d'expériences-5e éd.: Toutes les techniques nécessaires à la conduite d'une étude*. Dunod.
- Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, 14(6).

- Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand, pages 49–56.
- Kaikkonen, A. and Roto, V. (2003). Navigating in a mobile xhtml application. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 329–336. ACM.
- Kellar, M., Watters, C., and Shepherd, M. (2006). A goal-based classification of web information tasks. *Proceedings of the Association for Information Science and Technology*, 43(1):1–22.
- Kleemann, F., Voß, G. G., and Rieder, K. (2008). Un (der) paid innovators: The commercial utilization of consumer work through crowdsourcing. *Science, technology & innovation studies*, 4(1):PP–5.
- Kolowich, S. (2011). Killing peer review. *Inside Higher Ed*, 19.
- Krug, S. (2000). *Don't make me think!: a common sense approach to Web usability*. Pearson Education India.
- LaPlante, L. (2013). Hackschooling makes me happy. *TEDx Talks*, (1).
- Law, E. L., Roto, V., Hassenzahl, M., Vermeeren, A. P., and Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach.
- Le, J., Edmonds, A., Hester, V., and Biewald, L. (2010). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, pages 21–26.
- Leriche, F. and Meynard, C. (2008). Introduction. de l’hypertexte au manuscrit: le manuscrit réapproprié. *Recherches et Travaux*, 72:9–36.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

- Manning, C., Raghavan, P., and Schütze, H. (2009). Introduction to information retrieval/christopher d.
- Meynard, C. and Lebarbé, T. (2014). Donner à voir, à lire, à comprendre: destinataires et finalités d’une édition polymorphe des manuscrits de stendhal. In *Corpus littéraires numérisés: la place du sujet lecteur et usager*, number 9, pages 97–115. Université de Savoie.
- Moirez, P., Moreux, J. P., and Josse, I. (2013). Etat de l’art en matière de crowdsourcing dans les bibliothèques numériques. *Livrable L-4.3*, 1.
- Morrison, J. B., Pirolli, P., and Card, S. K. (2001). A taxonomic analysis of what world wide web activities significantly impact people’s decisions and actions. In *CHI’01 extended abstracts on Human factors in computing systems*, pages 163–164. ACM.
- Murugesan, S. (2007). Understanding web 2.0. *IT professional*, 9(4).
- Nielsen, J. (1999). *Designing web usability: The practice of simplicity*. New Riders Publishing.
- Nierman, A. and Jagadish, H. (2002). Evaluating structural similarity in xml documents. In *webdb*, volume 2, pages 61–66.
- Ntogas, N. and Veintzas, D. (2008). A binarization algorithm for historical manuscripts. In *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*, number 12. World Scientific and Engineering Academy and Society.
- Oleson, D., Sorokin, A., Laughlin, G. P., Hester, V., Le, J., and Biewald, L. (2011). Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human computation*, 11(11).
- Pellegrini, C. (2013). La plateforme crowdcrafting met les connaissances des citoyens au service de la science. <https://www.unige.ch/communication/communiques/2013/cdp130422/>. Accessed: 2017-08-30.
- Peltier, M. (2011). Développement d’applications web avec le framework php symfony 2.

- Porter, J. (2003). Testing the three-click rule. *User Interface Engineering*.
- Prestopnik, N. and Crowston, K. (2011). *Gaming for (citizen) science: Exploring motivation and data quality in the context of crowdsourced science through the design and evaluation of a social-computational system*, pages 28–33.
- Riley, J. (2017). Understanding metadata: What is metadata, and what is it for?: A primer.
- Ruecker, S. (2015). A brief taxonomy of prototypes for the digital humanities. *Scholarly and Research Communication*, 6(2).
- Ryan, R. M. and Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67.
- Schnapp, J., Presner, T., Lunenfeld, P., et al. (2009). The digital humanities manifesto 2.0. Retrieved September, 23:2012.
- Sellen, A. J., Murphy, R., and Shaw, K. L. (2002). How knowledge workers use the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–234. ACM.
- Shen, X.-L., Lee, M. K., and Cheung, C. M. (2014). Exploring online social behavior in crowdsourcing communities: A relationship management perspective. *Computers in Human Behavior*, 40:144–151.
- Siemens, L. (2012). Understanding long-term collaboration: Reflections on year 1 and before. *Scholarly and Research Communication*, 3(1).
- Siemens, L., Cunningham, R., Duff, W., and Warwick, C. (2011). A tale of two cities: Implications of the similarities and differences in collaborative approaches within the digital libraries and digital humanities communities. *Literary and linguistic computing*, 26(3):335–348.
- Siemens, L., Duff, W., Cunningham, R., and Warwick, C. (2009). “it challenges members to think of their work through another kind of specialist’s eyes”: Exploration of the

- benefits and challenges of diversity in digital project teams. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–14.
- Siemens, R. and Meloni, J. (2010). Implementing new knowledge environments: Building upon research foundations to understand books and reading in the digital age.
- Siemens, R., Siemens, L., Cunningham, R., Galey, A., Ruecker, S., and Warwick, C. (2012). Implementing new knowledge environments: Year one research foundations. *Scholarly and Research Communication*, 3(1).
- Smith, S. L. and Mosier, J. N. (1986). *Guidelines for designing user interface software*. Mitre Corporation Bedford, MA.
- Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.
- Svensson, P. (2009). Humanities computing as digital humanities. *Digital Humanities Quarterly*, 3(3).
- Vitali-Rosati, M. and Sinatra, M. E. (2014). *Pratiques de l'édition numérique*. Les Presses de l'Université de Montréal.
- Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. (2008). re-captcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468.
- Vukovic, M. (2009). Crowdsourcing for enterprises. In *Services-I, 2009 World Conference on*, pages 686–692. IEEE.
- Webster, J. and Ahuja, J. S. (2006). Enhancing the design of web navigation systems: The influence of user disorientation on engagement and performance. *Mis Quarterly*, pages 661–678.
- Wexler, M. N. (2011). Reconfiguring the sociology of the crowd: exploring crowdsourcing. *International Journal of Sociology and Social Policy*, 31(1/2):6–20.

- Wiggins, A. and Crowston, K. (2011). From conservation to crowdsourcing: A typology of citizen science. In *System Sciences (HICSS), 2011 44th Hawaii international conference on*, pages 1–10. IEEE.
- Zhang, K. and Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262.

Annexes

Appendix A

Annex A

A.1 Instructions for Stendhal

Introduction

Regardez l'image manuscrite et recopiez le texte aussi précisément que possible en utilisant les outils disponibles dans l'éditeur XML. Faites attention à la structure du texte et à la mise en page (les mots raturés, les ajouts, les marginaux, les fins des lignes, etc.). Suivez les instructions énumérées par la suite pour installer le logiciel et faire une transcription.

Télécharger le logiciel Oxygen Author et faire une transcription

1. Allez sur le site officiel d'Oxygen et télécharger une version d'essai (gratuite) du logiciel Oxygen Author. http://www.oxygenxml.com/download_oxygenxml_author.html
2. Choisissez votre système d'exploitation (Windows / Mac) et la version du celui-ci (Windows : 32-bit / 64-bit ; Mac OSX 10.6+ / 10.8+).
3. Suivez les instructions d'installation.
4. Ouvrez un des documents XML fournis dans le dossier (docs) dans Oxygen Author et choisissez le mode « Author », parmi « Text », « Grid », « Author » en bas de la fenêtre d'aperçu du document.

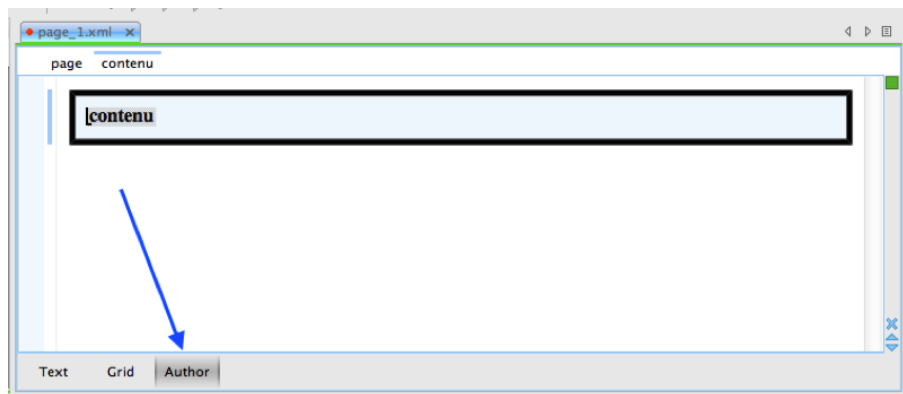


Figure A.1 – 1st instructional figure.

5. Dans la fenêtre d’aperçu centrale vous allez voir votre fichier XML en mode « Author » , ce qui vous affiche un bloc de texte avec l’entête contenu Figure A.1. Pour commencer, mettez le curseur dessus. Dans la fenêtre à gauche vous allez voir la structure arborescente du fichier. Vous allez commencer votre transcription dans l’élément racine de cette structure, en occurrence l’élément `<contenu>`. Dans la fenêtre à droite, et une fois que vous avez mis le curseur sur `<contenu>` vous allez aussi voir tous les éléments qui sont disponibles pour enrichir votre transcription. Sélectionnez l’onglet `<X> Elements` si vous ne le voyez pas. Ces éléments correspondent aux différents éléments que vous pouvez trouver dans la page manuscrite, mais ils permettent aussi de contrôler la structure du document. Vous allez vous servir de ces éléments pour transcrire et mettre en forme le texte.

- Pour commencer, sélectionner l’élément `<texte>`, pour désigner que vous créez un texte. Remarquez que la sélection d’éléments disponible est maintenant changée. Figure A.2
- Pour continuer, sélectionner l’élément `<paragraphe>` pour créer un nouveau paragraphe dans votre texte.
- À l’intérieur du paragraphe créer une ligne avec l’éléments `<ligne>`, vous avez maintenant à votre disposition tous les éléments les plus utilisés pour annoter le texte (ex : du texte raturé `<biffe>`, du texte souligné `<souligne>`, du texte rajouté en haut d’une ligne de script `<ajout>`, etc).
- Pour plus d’informations sur l’ensemble des éléments et leurs fonctions, vous avez

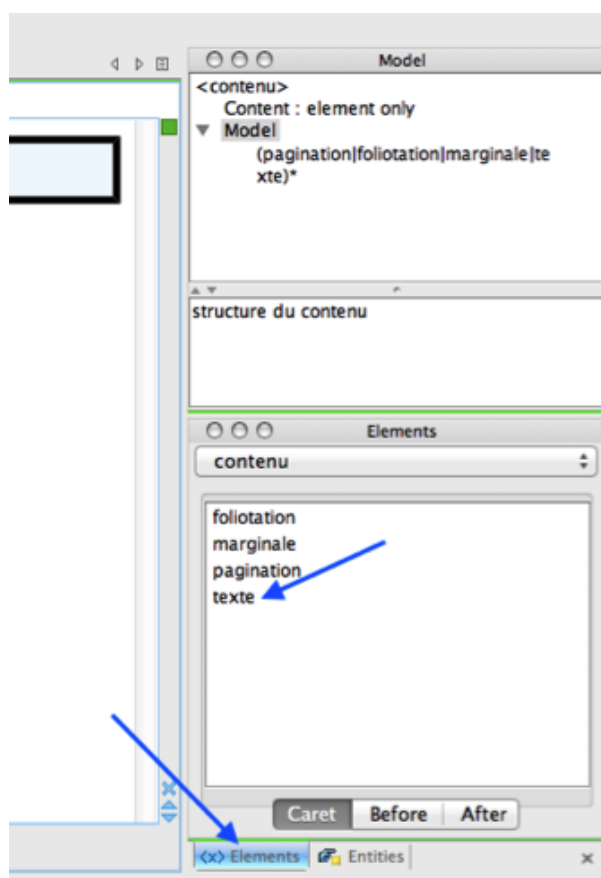


Figure A.2 – 2nd instructional figure.

à votre disposition une fiche descriptive des éléments (dictionnaire_elements.docx).

6. Complétez vos transcriptions en utilisant les outils proposés dans l'éditeur Oxygen Author. Une fois que vous avez terminé et que vous êtes satisfait(e) de la mise en forme et des annotations sauvegardez le document .xml dans le même dossier (docs), zippez-le et renvoyez-le à l'adresse mail fourni.
7. N'oubliez pas de remplir le document d'observation Google Docs pour chacun des documents transcrits.

Vous avez terminé !! Bravo !

A.2 Instructions for Benoîte Groult - online at PHuN 2.0



Figure A.3 – Instructions for transcribing Benoîte Groult, online at PHuN 2.0.

A.3 Instructions for Benoîte Groult Workshops

A.4

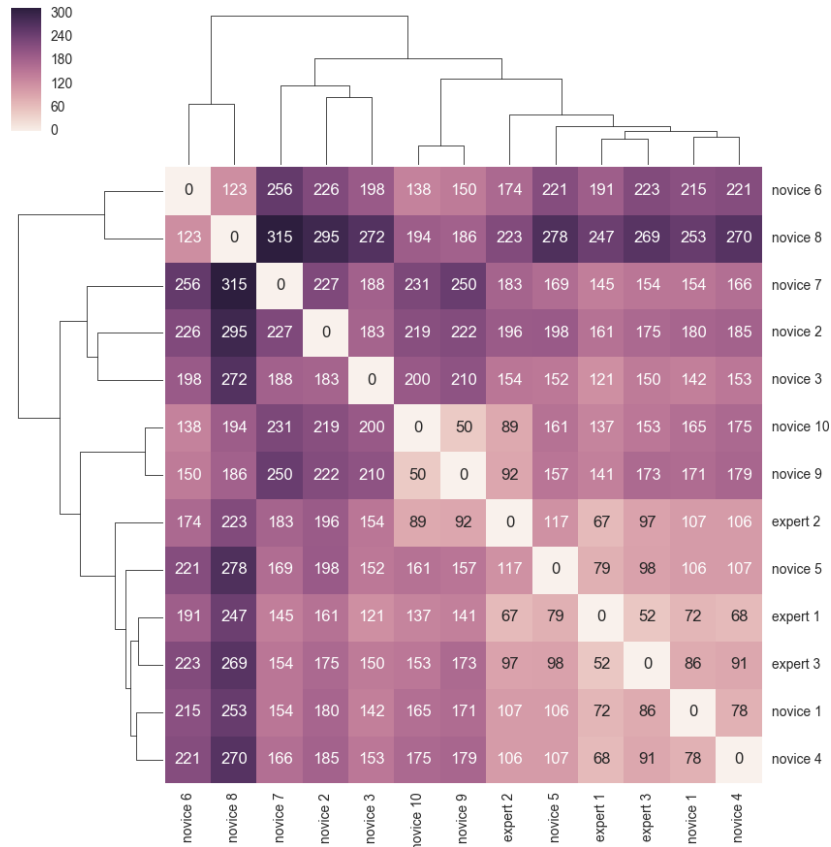


Figure A.4

Appendix B

Annex B

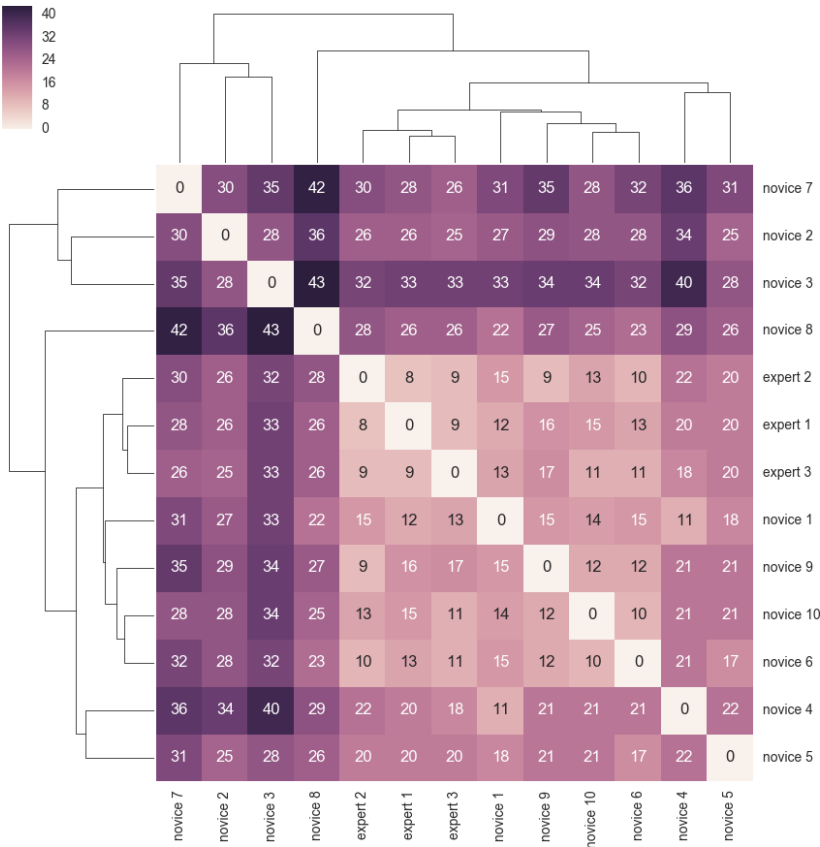
B.1 Stendhal Experiment 1 - page 2



(a)

(b)

Figure B.1 – Distance results on raw text for Stendhal's page 2.

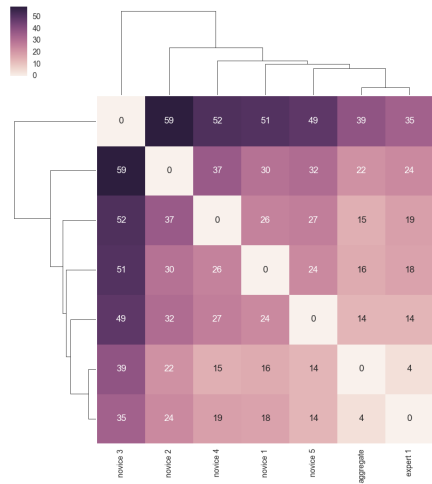


(a)

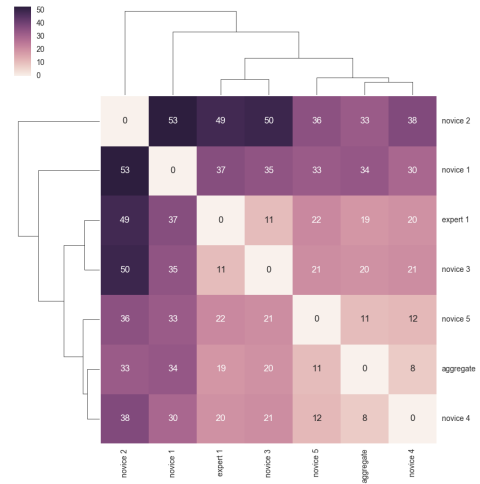
(b)

Figure B.2 – Distance results on raw text for Stendhal’s page 2.

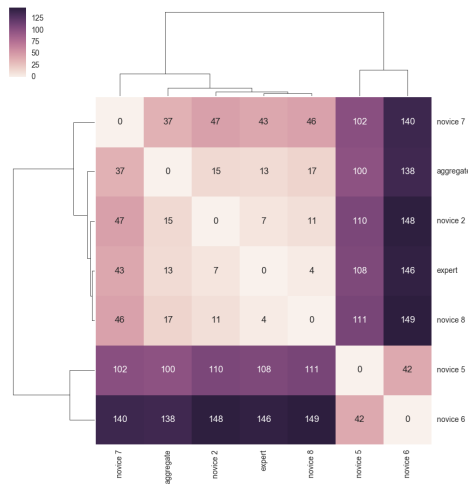
B.2 Recovered data for Benoîte Groult Experiment 1



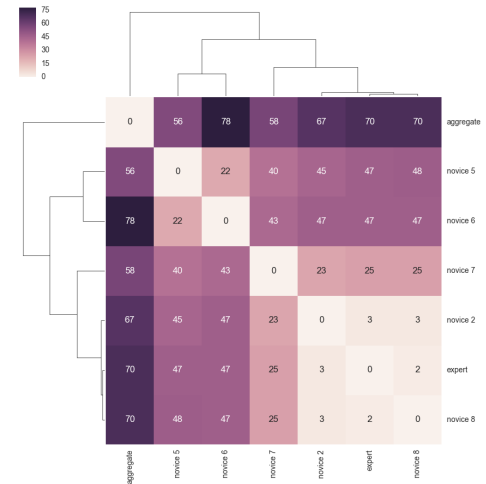
(a)



(b)



(c)



(d)

Figure B.3 – Distance results on (a) raw text and (b) xml for page 011, folder 03. Distance results on (c) raw text and (d) xml for page 014, folder 03.

B.3 Benoîte Groult Workshops

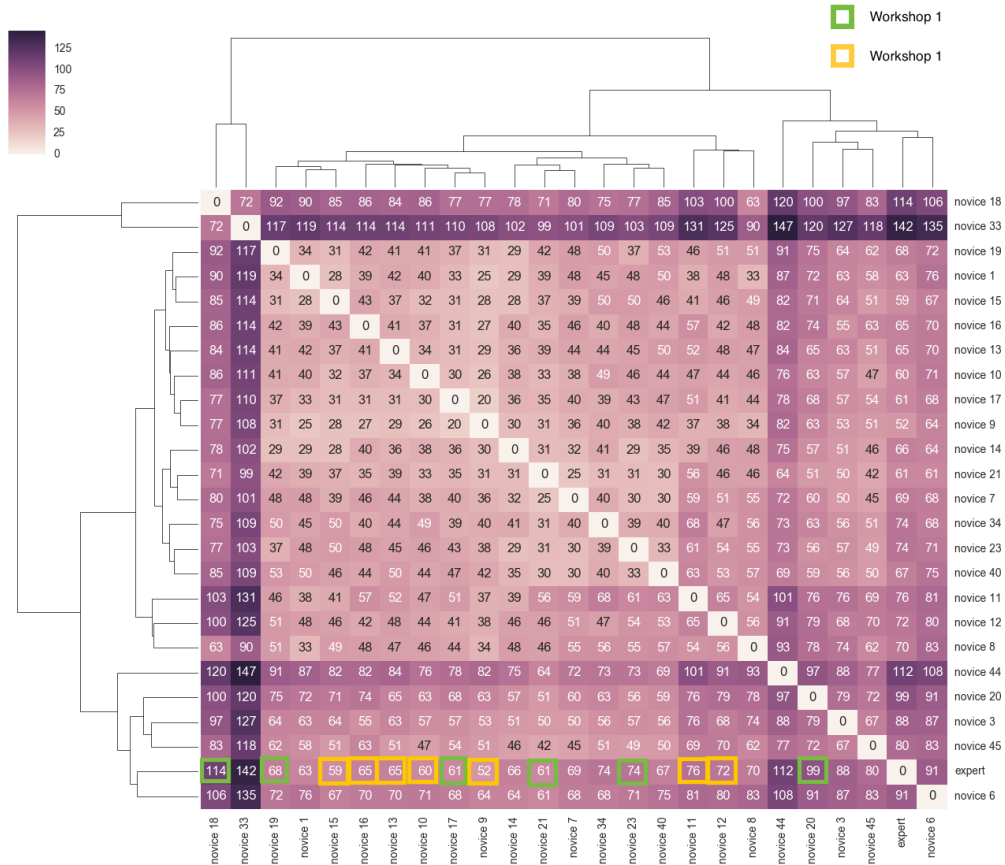


Figure B.4 – Distance results on raw text for Benoîte Groult workshops, page 01, folder 05.

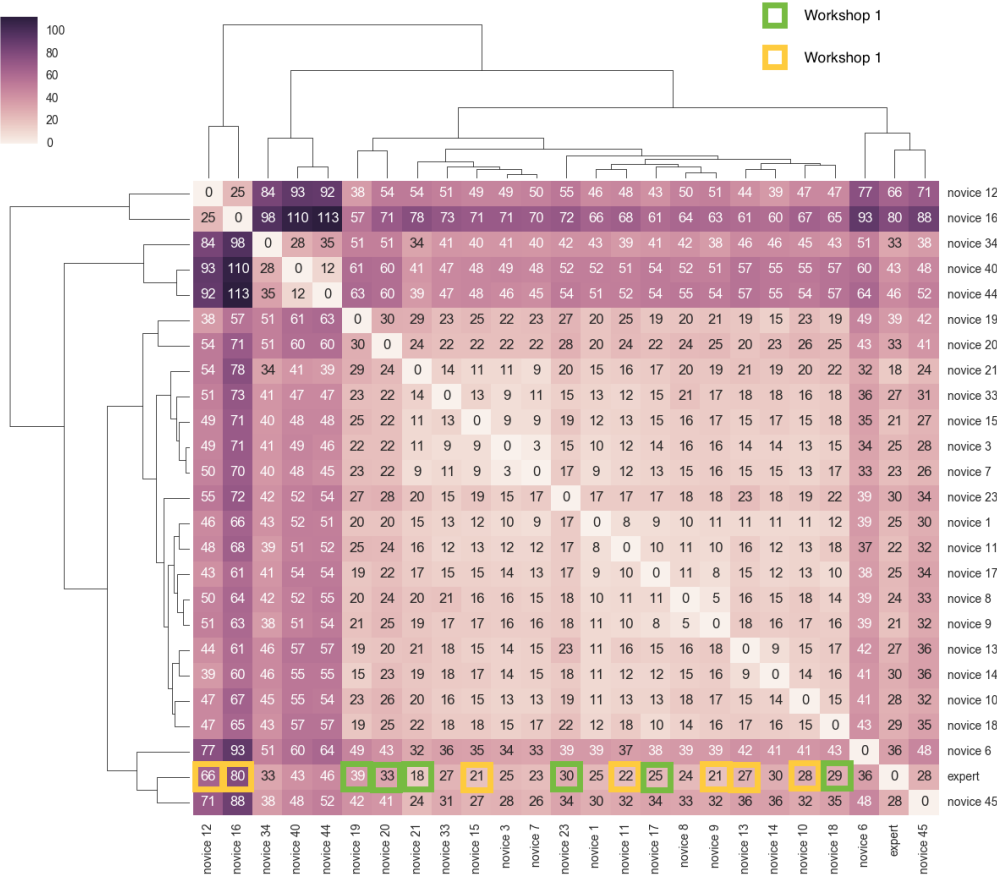


Figure B.5 – Distance results on xml for Benoîte Groult workshops, page 01, folder 05.

Appendix C

Annex C

C.1 PHuN 2.0 Participant list

| | | | |
|-------------------|--------------------|---------------|----------------|
| Ballester | Elmonbo89 | Loris | SabrinaG |
| cdessaint | Erisdar | lucia0682 | Sanda |
| Cécile Meynard | EtudiantHN | maimonem | sandvic |
| Cécile Meynard et | FrPeyre | Marine Salès | shzhang |
| Elisabeth Greslou | GillesRusse | marochereau | Thomas Lebarbé |
| ClaireW | GrandePerrine | Marteau.A | Vigo |
| dbourrion | GuillaumeB | mdessartre | vinz32 |
| Dissler | Haimo Wang | oceaneb73 | Volodia |
| dorian.bellanger | hanSolo | Olivier | willy |
| E.Melnikova | Jean-Baptiste Bre- | Pauline | Yazhu XU |
| efe.acu | ton | pauline.evrat | |
| Elisabethg | Julien L | Rougeaux | |
| eliseagniel | khadija_napoli | Sabah | |

C.2 PHuN-ET Participant list

| | | |
|-----------------|------------|-----------------|
| Amelie | blabla | oschneider |
| annegf | rouffiam | Pascale |
| brigitte | rousseti | RocketScientist |
| bibnum | lschneider | Roxanne |
| clairehugonnier | lucieM | ounoughs |
| ClaireW | Lucien | SaraMazziotti |
| galottac | MC | schneiderj |
| claude | Marta | pauline-s |
| Bidule | mplafitau | |
| eleclerc | habran | |

Résumé

Les projets en humanités numériques utilisent de plus en plus des méthodes de collaboration axées sur le public, telles que le crowdsourcing pour atteindre les objectifs de recherche, de conservation et d'édition scientifique en sciences humaines et sociales. Par exemple, le crowdsourcing représente une opportunité pour accélérer les projets de transcription pour des communautés de chercheurs qui travaillent traditionnellement dans des circuits-fermés. Certaines questions importantes soulevée par les chercheurs et les érudits concernent notamment l'intérêt de la méthode, et en particulier la qualité des résultats obtenus avec cette méthode. En outre, l'efficacité du crowdsourcing pour les humanités numériques n'est pas documenté dans la littérature. Se pose ainsi la question de savoir si le public peut produire du matériel pouvant être par la suite utilisé pour des éditions scientifiques, aux quels cas, pour quel type de projet et combien de post-traitement ou corrections seront nécessaires.

Cette thèse de doctorat examinera le potentiel apport du crowdsourcing des transcriptions pour les projets d'édition scientifique en humanités numériques. Pour cela, nous allons premièrement explorer les technologies et les techniques disponibles pour produire les transcriptions sous format XML en ligne. Deuxièmement, ayant développé et testé une plateforme internet de transcription que nous présenterons, nous pourrons examiner les besoins des utilisateurs vis-à-vis des environnements de travail collaboratifs fondées sur les retours des utilisateurs et les environnements de crowdsourcing industriels existants. Troisièmement, les données récoltées seront soumises à une analyse numérique qui permettra de comparer les productions des experts et celle des non-experts en s'appuyant sur les mesures de distances entre documents. Les résultats obtenus permettront de déterminer le potentiel apport du crowdsourcing pour les projets d'édition numérique scientifique. Enfin, le travail se terminera avec une discussion sur les implications des travaux actuels et présentera des opportunités pour des recherches futures sur le terrain.

Mots-clés: crowdsourcing, transcription, manuscrits, édition scientifique, évaluation, prototypage, humanités numériques

Abstract

Projects in digital humanities increasingly employ public-oriented collaboration methods such as crowdsourcing to achieve objectives that include research, conservation and scholarly editing in the humanities and social sciences. For example, crowdsourcing presents an opportunity to quicken the pace of progress for transcription projects for research communities that have traditionally operated within closed circuits. Some important questions raised by researchers and scholars concern the benefits of using this method and in particular the quality of results that can be obtained. Meanwhile, literature that evaluates the efficacy of crowdsourcing for digital humanities projects is insufficient. Questions as to whether the public can produce material that can be used for scholarly editions, in which cases, for which types of projects, and how much post-processing or corrections will be required, continue to occupy discussions on the matter.

This doctoral thesis will examine the potential benefits of crowdsourced transcription for scholarly editing projects in the digital humanities. Firstly, by exploring the technologies and techniques available to render online transcription in XML possible. Secondly, by developing and testing an online transcription platform, which will allow to examine user needs for collaborative work environments based on user responses and existing industrial crowdsourcing environments. Thirdly, the data collected will be subjected to digital analysis to compare the productions of non-expert transcribers to those of expert transcribers on the basis of document distance measurements. The results will be interpreted to determine the potential benefits of crowdsourcing for digital scholarly editing projects. Finally, the work will conclude by discussing the implications of current work and presenting opportunities for future research in the field.

Keywords: crowdsourcing, transcription, manuscripts, scholarly editing, evaluation, prototyping, digital humanities

