

# L'édition électronique des dossiers de *Bouvard et Pécuchet* de Flaubert : des fragments textuels en quête de mobilité

Stéphanie Dord-Crouslé

CNRS - LIRE UMR 5611

Emmanuelle Morlock-Gerstenkorn

CNRS - ISH UMS 1798

## Présentation du corpus et de sa complexité particulière

Le 8 mai 1880, Flaubert meurt brusquement d'une attaque cérébrale. Il laisse inachevé le roman intitulé *Bouvard et Pécuchet* dont il avait commencé la préparation en 1872. L'originalité de l'ouvrage est double. Elle tient d'abord à sa portée épistémologique singulière : c'est une « encyclopédie critique en farce » ; elle vient ensuite du fait que la mort de l'auteur a interrompu l'écriture de l'œuvre : un seul des deux volumes projetés a été rédigé (sans être d'ailleurs complètement terminé), et le second volume est resté à l'état de chantier documentaire<sup>1</sup>. Quoique inachevé, le roman a été publié. Mais c'est en fait le seul « premier volume » qui est publié sous le titre de *Bouvard et Pécuchet*, depuis la parution originale posthume de l'ouvrage en 1881, jusqu'aux éditions modernes actuelles. Cependant, si Flaubert était loin d'avoir terminé son « second volume », il en avait déjà rassemblé de nombreux éléments. Ce chantier documentaire est aujourd'hui conservé à la bibliothèque municipale de Rouen où il est connu sous le nom de « dossiers de *Bouvard et Pécuchet* ». Il s'agit de huit gros recueils de documents divers (Ms. g226<sup>1-8</sup>), auxquels il faut ajouter deux recueils plus minces concernant le *Dictionnaire des idées reçues* (Ms g227 et g228), soit un total de 2 300 feuillets.

1. Sur ce dispositif et l'aspect encyclopédique de l'entreprise flaubertienne, voir Stéphanie Dord-Crouslé, *Bouvard et Pécuchet de Flaubert, une « encyclopédie critique en farce »*, Paris, Belin (Belin-Sup Lettres), 2000.

Certains extraits de ces documents ont été publiés dès les années 1880-1890, en particulier par Maupassant dans différents articles. Et à partir des années soixante-dix, toutes les éditions du roman se sont enrichies d'un choix, plus ou moins large, de ces documents présentés en annexe au texte rédigé par Flaubert<sup>2</sup>. Enfin, des ouvrages complets<sup>3</sup> ont tenté la reconstitution conjecturale de ce qu'aurait pu être ce «second volume». Ces tentatives sont à la fois très intéressantes et très insatisfaisantes, car elles ne peuvent parvenir à rendre justice à la complexité et à la singularité du corpus, voire conduisent à en trahir la nature.

Le corpus se compose d'une grande partie de pages entièrement manuscrites au sein desquelles plusieurs scripteurs doivent être distingués : Flaubert évidemment, mais aussi différents collaborateurs réguliers ainsi que des informateurs ponctuels qu'il faut identifier. Les dossiers contiennent aussi des pages imprimées, vierges de toute annotation : des coupures de presse, parfois des journaux entiers, ou bien des publicités et des tracts. Enfin, on trouve dans ces dossiers un grand nombre de pages complexes résultant de collages et comportant à la fois des fragments manuscrits et des fragments imprimés qui sont souvent eux-mêmes annotés. Cette hétérogénéité matérielle recoupe partiellement et introduit à un autre type d'hétérogénéité, cette fois-ci, typologique. En effet, en fonction de la méthode de travail de Flaubert, on peut identifier plusieurs catégories de documents dans ces dossiers conservés à Rouen. Il y a d'abord de la documentation brute ou peu traitée, dont on ne sait si ou comment Flaubert l'aurait effectivement utilisée s'il avait vécu. On trouve aussi des listes de titres issues des recherches bibliographiques menées par le romancier. Il y a des sortes de «fiches de lecture» qui sont les notes que Flaubert a prises à la lecture des ouvrages qu'il a consultés (autour de 1500, écrit-il dans sa correspondance!). On trouve encore des fiches de synthèse disciplinaires qui regroupent par thèmes différents aspects des pages de notes, selon un processus que Flaubert appelait : «coordonner les notes de [s]es notes». Enfin, les dossiers comportent des pages plus ou moins préparées pour le second volume du roman, c'est-à-dire des pages sur lesquelles Flaubert a regroupé des citations extraites de ses notes. Ces pages étaient destinées à devenir la matière première du second volume de son roman.

Un certain nombre de pages préparées dans cette intention existent donc dans les dossiers de Rouen, mais Flaubert est mort avant d'avoir terminé le travail, et surtout avant d'avoir donné sa forme définitive à l'ensemble. Même dans le cas des citations qui semblent les plus conformes à l'état que Flaubert pouvait chercher à atteindre, c'est-à-dire les citations qui sont déjà isolées et classées par sections (comme «Haine des grands hommes» ou «Beautés de la religion»), les catégories se révèlent instables. Les annotations portées par l'écrivain (qui indiquent le lieu probable du classement) sont souvent plurielles. Ainsi, sur un feuillet (Ms. g226<sup>7</sup> f<sup>o</sup> 14), la catégorie initiale «style médical» a été barrée et corrigée en «style rococo». Mais une sous-catégorie subsiste, «dangers du chocolat», qui relève de la logique médicale précédente. Alors, où classer ce fragment ? Une édition papier se trouve dans l'obligation de choisir et donc de mettre un frein à la mobilité de ce fragment. Seul un changement de support pouvait permettre d'inventer autre chose. Aussi notre projet

2. Voir, par exemple, Bouvard et Pécuchet, avec un choix de scénarios du Sottisier, *l'Album de la Marquise et le Dictionnaire des idées reçues*, Claudine Gothot-Mersch (éd.), Paris, Gallimard (Folio), 1979 ; ou Bouvard et Pécuchet, avec des fragments du «second volume» dont le Dictionnaire des idées reçues, éd. mise à jour de Stéphanie Dord-Crouslé, Paris, Flammarion (GF), 2008.
3. Voir en particulier : *Le second volume de Bouvard et Pécuchet*, Geneviève Bollème (éd.), Paris, Denoël (Dossier des Lettres Nouvelles), 1966 ; Bouvard et Pécuchet, *Œuvre posthume augmentée de la Copie*, t. 5 et 6 des *Œuvres complètes*, éd. nouvelle établie, d'après les manuscrits inédits de Flaubert, par la Société des Études littéraires françaises..., Paris, Club de l'Honnête homme, 1972 [éd. de Maurice Bardèche] ; *Le second volume de Bouvard et Pécuchet, le projet du Sottisier, reconstitution conjecturale de la «copie» des deux bonshommes d'après le dossier de Rouen*, Alberto Cento et Lea Caminiti Pennarola (éd.), Naples, Liguori, 1981 ; et plus récemment : *Universalenzyklopädie der Menschlichen Dummheit. Ein Sottisier*, Hans-Horst Henschen (éd.), Frankfurt am Main, Eichborn, 2004.

d'édition en ligne a-t-il pour ambition de répondre aux différents défis que présente ce corpus. À terme, tous les visiteurs du site devront pouvoir expérimenter une ou plusieurs reconstitution(s) conjecturale(s) de l'organisation du second volume de *Bouvard et Pécuchet*, intégrant – entre autres – les problèmes spécifiques posés par le *Dictionnaire des idées reçues*.

Pour mener à bien ce projet<sup>4</sup>, une équipe scientifique s'est constituée, autour de l'UMR lyonnaise LIRE<sup>5</sup>, avec des spécialistes de Flaubert et plus largement du XIX<sup>e</sup> siècle. Elle réunit aujourd'hui une trentaine de chercheurs répartis en quatre pôles géographiques : France, Italie, Japon et États-Unis. Il s'agit de chercheurs littéraires classiques, formés à la logique de l'édition papier, sans aucune connaissance de la TEI ni aucune appétence particulière pour les questions de balisage... Pour parvenir au but qu'elle s'est fixé, l'équipe avait besoin des images des documents : d'abord pour les traiter scientifiquement, puis, à terme, pour les diffuser. La numérisation en haute définition de l'ensemble des pages du corpus était donc souhaitable. De fait, pour des raisons complexes, il n'a pas été possible d'établir un réel partenariat avec l'institution de conservation, la bibliothèque municipale de Rouen. En revanche, celle-ci nous a gracieusement autorisés à diffuser l'ensemble des images numérisées des microfilms de sauvegarde que nous avons acquis, ainsi que les trois cents images haute définition qu'elle a accepté de réaliser pour que nous disposions d'un échantillonnage représentatif des documents.

Pour le traitement scientifique des manuscrits (leur transcription et leur annotation), il était nécessaire que l'équipe scientifique dispose rapidement d'un accès aisé aux images. Cela a été rendu possible grâce au site<sup>6</sup> conçu et maintenu par le Service d'ingénierie documentaire (SID<sup>7</sup>) de l'ISH. Pour l'instant, le site est réservé uniquement aux collaborateurs ; à terme, il sera accessible à l'ensemble des internautes. Il repose sur une base de données relationnelle, créée par Raphaël Tournoy, qui articule une base d'images et une base de transcriptions. Ces transcriptions sont créées par les membres de l'équipe scientifique dans un traitement de texte ; elles sont transformées et archivées au format PDF pour procurer une transcription ultradiplomatique. Simultanément, les transcriptions sont intégrées au format HTML dans la base de données et sont utilisées par l'actuel moteur de recherche.

Cependant, la gestion des transcriptions au niveau de la page ne permet pas d'atteindre le but visé, à savoir conserver la mobilité des énoncés. Pour pouvoir insérer les fragments dans différents contextes de classement, il faut pouvoir y accéder directement. La base de données doit donc préalablement passer de l'échelle de la page à celle du fragment. Ce changement d'échelle implique d'identifier et d'enregistrer dans la base de données l'ensemble des citations et extraits sélectionnés par Flaubert. Pour réaliser cette opération, l'une des pistes suivies<sup>8</sup> par le projet passe par l'utilisation de l'encodage TEI<sup>9</sup> qui est avant tout une opération qui délimite chaque unité jugée signifiante d'un texte. Ce balisage devant

4. Depuis son lancement en 2006, le projet a bénéficié d'un soutien financier spécifique du CNRS (appel d'offres « ATIP Jeunes chercheurs » 2006 du Département Sciences humaines et sociales) ; de l'Agence Nationale de la Recherche (appel à projets « Corpus et outils de la recherche en Sciences humaines et sociales » du programme Sciences humaines et sociales 2007) ; et de la Région Rhône-Alpes (allocation doctorale allouée au projet dans le cadre du Cluster de recherche n° 13 « Culture, patrimoine, création »).

5. Voir <http://lire.ish-lyon.cnrs.fr/>.

6. Voir <http://dossiers-flaubert.ish-lyon.cnrs.fr/>.

7. Voir <http://sid.ish-lyon.cnrs.fr/>.

8. Une piste complémentaire, reposant exclusivement sur des outils de reconnaissance automatique d'images, est explorée parallèlement par Vincent Malleron (voir <http://www.malleron.info/>) dans le cadre de sa thèse (codirection LIRE-LIRIS), financée par une allocation doctorale de recherche attribuée au projet Bouvard par la Région Rhône-Alpes (Cluster 13).

9. La TEI est un standard d'encodage XML de textes, de plus en plus répandu dans les projets d'édition critique sur support électronique. Elle reflète l'organisation hiérarchique de l'information contenue dans un document et repose sur le principe de la séparation du contenu et de la mise en forme. Elle vise à faciliter la création, l'échange et l'intégration des données textuelles informatisées. Voir <http://www.tei-c.org/>.

être réalisé pour l'ensemble du corpus à des fins d'édition<sup>10</sup>, d'exploitation en recherche<sup>11</sup>, d'échange et de pérennisation, il nous offre par là même une méthode efficace de découpage.

## Le fragment en question

Mais les fragments sont-ils aisément identifiables et délimitables ? Au cours du travail de rédaction du guide d'encodage TEI spécifique au projet<sup>12</sup>, on s'est aperçu que l'on désignait par le terme de « fragment » des unités de nature différente : des portions de texte, des zones dessinées sur les images fac-similé du manuscrit, et des morceaux de pages manuscrites ou imprimées, provenant de coupures de presse ou de pages intermédiaires découpées et recollées par morceaux sur d'autres pages. En vue de définir une stratégie de balisage précise, homogène et cohérente avec les exploitations souhaitées, il était nécessaire de définir précisément cette unité. Avant de le faire, reprenons les différentes acceptions concurrentes du terme.

Jusqu'à présent, on a parlé de « citations » et d'« extraits » pour désigner les passages recopiés ou insérés par collage dans les pages préparées pour le second volume, à partir des notes de lecture prises par Flaubert. L'écrivain les a sélectionnées par étapes pour donner naissance à un volume dans lequel les parties narratives laisseraient le devant de la scène aux emprunts littéraires et scientifiques. Dans cette perspective, les « fragments » sont des unités textuelles structurées de manière régulière : les énoncés exogènes sont le plus souvent accompagnés d'un renvoi bibliographique indiquant leur source (roman, traité philosophique, article de quotidien, etc.). Sur les pages utilisées pour la composition du roman, cette structure canonique se voit complétée par des annotations de la main de Flaubert : catégories de classement, commentaires, ou croix de sélection signalant que l'unité doit être copiée sur une nouvelle page... Cependant, une autre approche du fragment a été concurremment mise en œuvre dans le projet. Profitant de la capacité croissante des ordinateurs à afficher des images haute résolution, et de procédés de numérisation toujours plus performants, les éditions électroniques de sources textuelles font dorénavant la part belle aux images fac-similé. Aussi la mise en relation de la transcription avec les images du manuscrit fait-elle partie des objectifs prioritaires du projet. Au niveau des pages, le but est déjà atteint ; en ce qui concerne les fragments, l'essentiel reste à faire, même si, pour un certain nombre de pages, notamment les pages préparées pour le second volume ou celles du *Dictionnaire des idées reçues*, les fragments semblent visuellement assez faciles à repérer et à isoler, qu'il s'agisse de morceaux imprimés ou manuscrits. En tout cas, dans cette perspective, la notion de « fragment » désigne une région sur le manuscrit dont les contours peuvent être délimités à l'aide d'un outil d'édition graphique permettant de tracer des formes polygonales – et pas seulement rectangulaires –, pour mieux épouser les développements irréguliers de l'écriture manuscrite dans l'espace de la page.

Mais l'examen des pages complexes résultant de collages successifs nous a amenés à utiliser parfois la notion de « fragment » dans une troisième acception, encore différente :

- 
10. On entend ainsi produire les versions diplomatique et linéarisée du texte à partir de la même source XML.
  11. Les requêtes dans un document XML peuvent porter à la fois sur les contenus (les mots du texte) et leur structure (la nature des éléments utilisés pour le balisage).
  12. La TEI n'est pas une DTD ni un format qui peut s'utiliser directement. Il s'agit au contraire d'un système modulaire, formulé sous forme de recommandations très génériques. C'est à partir de ces *guidelines* que chaque projet doit définir sa « personnalisation ». Celle-ci comprend généralement une stratégie globale (choix de ce que l'on décide d'encoder et de la structure hiérarchique principale), un schéma de validation et un guide d'encodage définissant des règles d'utilisation spécifiques au projet et aux encodeurs. Les recommandations sont consultables en ligne sur le site Web du consortium TEI : TEI Consortium (éd.), *Guidelines for Electronic Text Encoding and Interchange*, <http://www.tei-c.org/P5/>.

les fragments sont alors les découpes de papier qui ont été collées par Flaubert sur une page. Le critère de délimitation est ici purement matériel, et seules les images couleur haute définition permettent de distinguer clairement que deux, trois voire quatre couches de papier se trouvent parfois superposées.

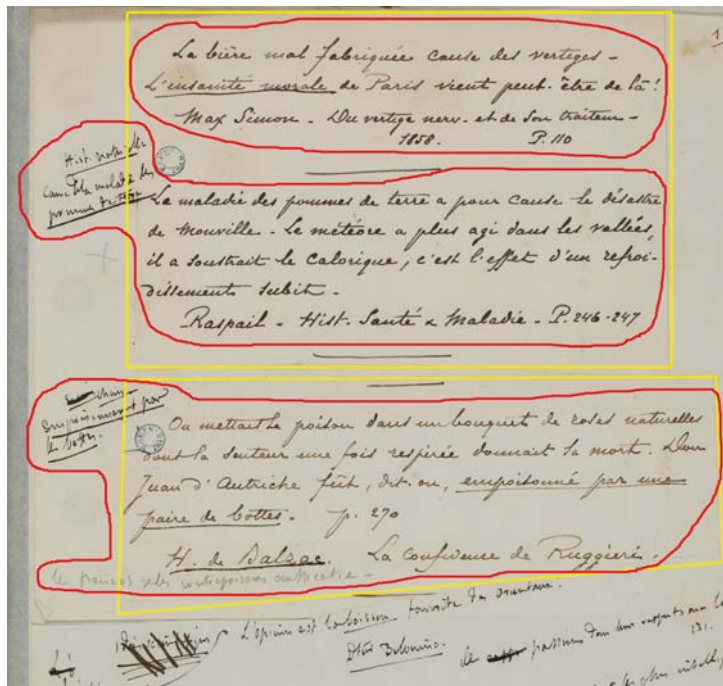


Figure 1 – le conflit des découpes et des citations: un exemple extrait du Ms. g226° f° 126. Les découpes sont représentées par des contours jaunes. Un tracé rouge entoure chacune des trois premières citations (« Collections bibliothèque municipale de Rouen – photographie Thierry Ascencio-Parvy »).

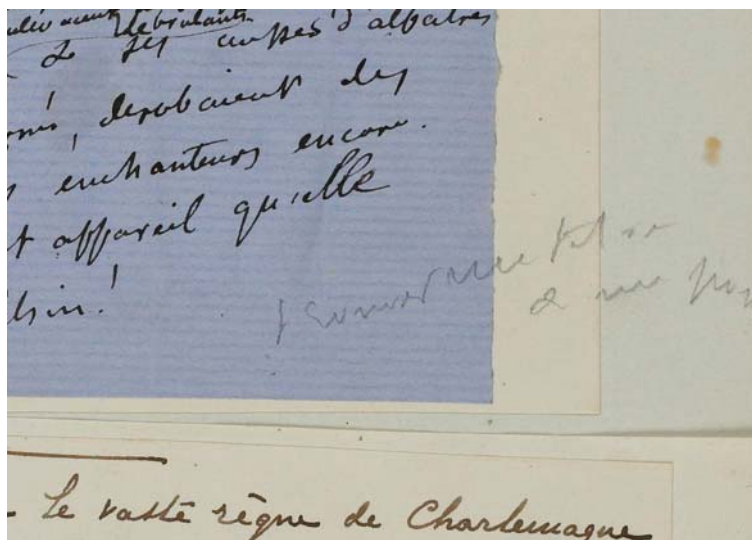


Figure 2 – le conflit des découpes et des citations: un autre exemple extrait du Ms. g226° f° 12 (« Collections bibliothèque municipale de Rouen – photographie Thierry Ascencio-Parvy »).

Comme le montrent les figures 1 et 2, plusieurs citations peuvent coexister sur une même découpe, tandis qu'inversement, une citation peut se développer à cheval sur plusieurs découpes. Il y a donc conflit. En ce qui concerne la délimitation des zones visuelles sur

l'image fac-similé, la prééminence du niveau physique sur le niveau logique aurait pour conséquence des lacunes dans l'encodage.

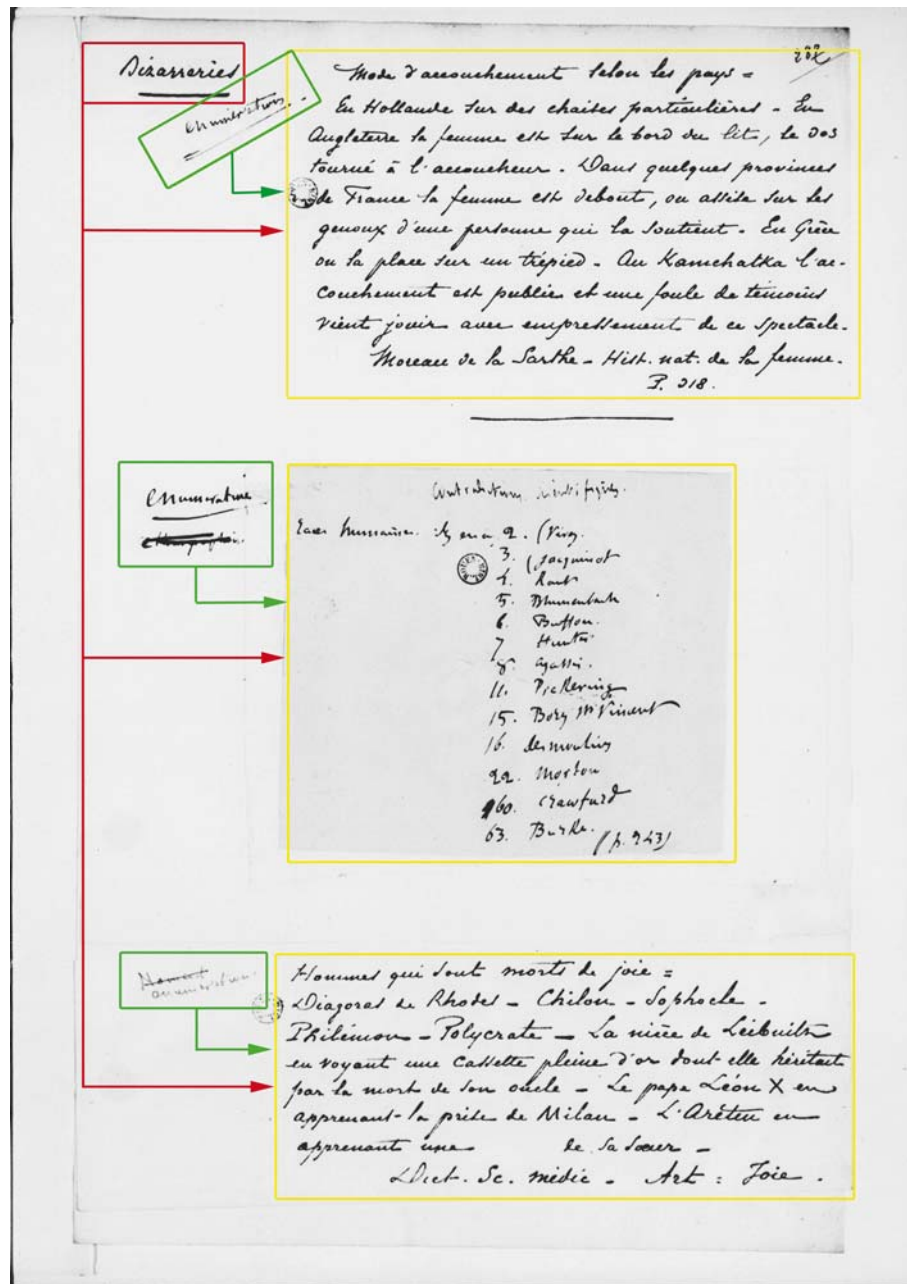


Figure 3 – structures implicites sur le Ms g2261 f° 287 («Collections bibliothèque municipale de Rouen»).

Dans le Ms. g2261 f° 287 (figure 3), la catégorie de classement «Bizarries» n'est inscrite qu'une seule fois en haut de la page. Mais elle s'applique aux trois extraits présents sur le feuillet et non au seul extrait qui lui est contigu. Or un découpage reposant uniquement sur un zonage visuel de l'image ne pourrait intégrer directement cette relation entre les deux citations du bas et la vedette excentrée qui les qualifie pourtant toutes deux. La mise en page induit des effets de structuration implicites qu'il faut repérer.

Les trois acceptions concurrentes du terme «fragment» sont donc inconciliables: chacune produit des unités de granularité et de contenu différents. Et comme chaque principe de

découpage correspond à un type de structuration distinct, il est impossible de combiner ces trois perspectives dans une hiérarchie TEI unique<sup>13</sup>. Parallèlement, la structure canonique du fragment<sup>14</sup> offre une sorte de grille d'analyse qui peut être appliquée à tous les types de page créés ou composés pour le projet de roman<sup>15</sup>. Elle constitue une aide pour l'encodeur, car elle lui permet, quand un composant attendu se révèle manquant (comme par exemple une vedette en marge) de le rechercher dans des formes de structures implicites, induites par exemple par la mise en page.

Cette question de la prise en compte des structures implicites se pose avec encore plus d'acuité quand il s'agit d'extraire les fragments de leur contexte. Toute l'information doit en effet être embarquée lors du découpage sans qu'il soit nécessaire d'y revenir ultérieurement. Seule la perspective textuelle que l'on qualifiera de « logique » (par opposition à la perspective visuelle et à la perspective matérielle des superpositions de collages) permet de distinguer des unités discrètes délimitées par le biais de l'encodage TEI, et autorisant l'intégration automatique<sup>16</sup> des fragments à la base de données. On doit donc se placer au niveau logique et non pas au niveau physique si l'on veut pouvoir identifier des unités textuelles agencables et comparables<sup>17</sup>.

Jusqu'à présent, nous avons raisonné à partir des pages préparées pour le second volume. Qu'en est-il lorsque d'autres types de pages sont concernés : la distinction reste-t-elle opératoire ? Pour répondre à cette question, on va suivre le parcours génétique d'une citation, en s'appuyant sur l'édition du second volume établie par Lea Caminiti. Le fragment numéroté 6 figure dans la section intitulée « Nomenclatures et bizarreries » de cette reconstitution conjecturale<sup>18</sup>. La citation reprend une liste d'« hommes qui sont morts de joie », que Flaubert a découverte dans l'article « Joie » du *Dictionnaire des sciences médicales*. Lea Caminiti a trouvé trois occurrences de ce passage dans le corpus. D'abord pris en note dans la fiche de lecture consacrée à l'article « Joie » (Ms. g226<sup>7</sup> f° 116 v°), le fragment semble avoir été ensuite recopié par Flaubert dans une page de récapitulation d'extraits intitulée « Curiosités médicales » (Ms. g226<sup>7</sup> f° 154). Enfin, dans la page préparée pour le second volume intitulée « Bizarreries », le fragment est recopié par l'ami et collaborateur de Flaubert, Edmond Laporte, accompagné d'une annotation marginale indiquant une seconde catégorie de classement : « énumération » (Ms. g226<sup>1</sup> f° 287). Au fil du processus génétique, la logique médicale s'efface ainsi progressivement, au profit d'une perspective littéraire qui correspond au regard critique – et ironique – de Flaubert.

Pour permettre leur comparaison ultérieure, l'encodage de ces trois occurrences d'une même citation doit s'opérer de manière similaire, c'est-à-dire en reprenant la même

13. La TEI reposant sur le formalisme XML, elle oblige à structurer le contenu selon une hiérarchie unique ou à défaut, prioritaire. Plusieurs méthodes de contournement existent et sont documentées dans les *Guidelines*, mais aucune n'est entièrement satisfaisante. Un groupe d'intérêt spécifique (ou SIG, Special Interest Group) traite spécifiquement de ces questions de chevauchements de structures au sein du consortium. Voir : <http://wiki.tei-c.org/index.php/SIG:Overlap/>.

14. On peut aussi parler de « modèle abstrait ». Voir Stéphanie Dord-Croulé et Emmanuelle Morlock-Gerstenkorn, « Le "modèle abstrait" du corpus Bouvard : première approche » ; à paraître dans les actes de la journée d'étude du 12 mars 2009 « Constitution et exploitation de corpus issus de manuscrits – Lectures, écritures et nouvelles approches en recherche documentaire » organisée à Grenoble par Cécile Meynard et Thomas Lebarbé (article disponible en ligne : <http://halshs.archives-ouvertes.fr/halshs-00368044/fr/>).

15. On laisse de côté les documents bruts insérés dans le corpus ou les pages associées à la composition d'autres œuvres de Flaubert, utilisées pour *Bouvard et Pécuchet* à titre documentaire, comme certaines pages de brouillons de *L'Éducation sentimentale*.

16. Des scripts informatiques permettront d'extraire les données de la source XML et de les reformater selon la structure de la base de données.

17. Le processus de comparaison est inhérent à toute entreprise de reconstitution conjecturale du second volume de *Bouvard et Pécuchet*. Ainsi, dans son édition *Le second volume de Bouvard et Pécuchet, le projet du Sottisier*, Lea Caminiti indique toutes les occurrences d'un même fragment, mais n'en donne que la version la plus récente dans le processus d'élaboration génétique.

18. *Le second volume de Bouvard et Pécuchet, le projet du Sottisier*, p. 16.

structure<sup>19</sup>. Ici, la prédominance de la structuration logique sur la simple analyse de la disposition topologique des zones amène à analyser la première ligne de la troisième occurrence (Ms g226<sup>1</sup> f° 287, bas de la figure 3): « hommes qui sont morts de joie », non comme l'élément liminaire de l'énoncé, mais comme une véritable vedette, accidentellement extraite de sa localisation usuelle en marge présente dans les deux occurrences antérieures. Grâce à cet encodage précis, vedettes et énoncés pourront ainsi être comparés terme à terme au cours du processus d'analyse génétique.

Au terme de cette quête du « fragment », tout concourt donc à prouver qu'il faut privilégier la perspective logique. Aussi le définira-t-on dorénavant comme une unité de nature essentiellement textuelle. Il renvoie à un passage d'une source imprimée ou manuscrite exogène et peut connaître plusieurs matérialisations dans le corpus. Il correspond à une entité de la base de données et est en relation avec un élément de la transcription TEI et une ou plusieurs zones de l'image<sup>20</sup>. Ainsi, sans délaisser complètement les autres perspectives existantes, on a placé résolument au cœur du projet la dimension logique du fragment, de manière à savoir où concentrer les efforts sans risquer de se perdre...

Comment cette structuration du fragment sera-t-elle exploitée dans l'interface du site? Ce balisage nous permet-il de conserver et de transférer dans la base de données, toutes les informations nécessaires à la recontextualisation des fragments dans le cadre des reconstitutions? C'est ce que nous avons voulu vérifier en construisant une maquette schématique de cette interface.

### Une préfiguration fonctionnelle de l'interface de reconstitution conjecturale

La maquette papier d'une interface Web offre la représentation statique de la future application; elle liste les éléments présents, leur localisation, leur appellation... Généralement, l'utilisation de ce type d'outil de conception sert à tester la compréhension globale de la navigation, de l'organisation de l'information et du vocabulaire. Dans notre contexte, il s'agit tout d'abord de vérifier la pertinence et la complétude des éléments que l'on doit transférer de l'encodage à la base de données.

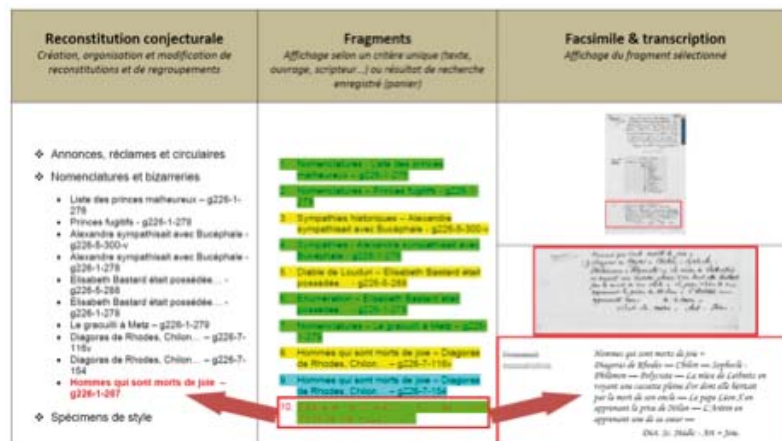


Figure 4 – la maquette de l'interface de reconstitution conjecturale.

19. La stratégie de balisage consiste à appréhender ces pages comme des listes de fragments. On utilise donc les balises < list > < item > et éventuellement < label >, pour le cas où des commentaires marginaux ont une fonction d'étiquetage, d'indexation ou de régie du fragment. À l'intérieur de chaque < item >, les citations sont encodées avec les balises < cit >, < quote > et < bibl >. D'autres formes de commentaires se rapportant soit à la citation entière, soit à l'une de ses composantes, sont encodées dans des < notes >.
20. Il s'agit ici de préserver la compatibilité de la démarche avec le logiciel de segmentation automatique développé par Vincent Malleron dans le cadre de sa thèse.

L'utilisateur doit pouvoir accéder, dans un même espace de travail, à toutes les informations nécessaires aux opérations d'agencement, de comparaison et d'analyse des fragments. Pour répondre à ce besoin, l'écran principal se présente sous la forme d'une fenêtre divisée verticalement en trois parties. L'espace central permet de visualiser des listes de fragments constituées selon différents critères<sup>21</sup>, comme par exemple l'ensemble des fragments associés à une vedette. La zone de droite est consacrée à la visualisation du fragment sélectionné dans la partie centrale : elle en affiche la transcription et l'image fac-similé et propose aussi une vue de la page fac-similé complète dont il est extrait. Si un tracé polygonal a été associé à l'image, il apparaît en superposition sur l'image, indiquant ainsi la position du fragment courant dans la page. Le processus de reconstitution conjecturale se déroule dans la zone de gauche. On y appelle les sections et catégories de classement que l'on souhaite utiliser pour composer un agencement. Puis on y classe les unités disponibles au centre par un simple glisser-déposer. Les espaces latéraux se synchronisent suivant l'unité sélectionnée au centre : à gauche, l'agencement affiche en surbrillance le fragment sélectionné ; à droite, la fenêtre est actualisée avec les transcriptions et images fac-similé correspondantes.

La création de cette maquette vise à résoudre des cas difficiles et à repérer des oublis dans notre stratégie d'encodage. Par exemple, dans les premiers essais de zonage graphique des feuillets, les titres de pages de notes de lecture étaient systématiquement identifiés comme des « fragments ».

Comme le montre clairement la figure 5, le titre du périodique *L'Artiste*, accompagné de l'indication « année 1839 », centré dans l'en-tête de la page, se détache visuellement et incite fortement l'encodeur à l'identifier comme un fragment. Mais jamais un titre de page comme celui-ci ne sera appelé dans le cadre d'une reconstitution conjecturale. Ce n'est donc pas un fragment dans le sens où nous avons défini cette entité : le « test de la maquette » permet de le prouver définitivement.

De même, on ne savait pas comment encoder certains renvois bibliographiques allusifs faisant suite à une citation et souvent introduits par la mention « voy. » ou « voyez ».

Dans la figure 6, la note de régie inscrite au crayon par Flaubert lors d'une campagne de relecture (« voy. la citation de Michel Raymond ») doit-elle être balisée comme une annotation du fragment précédent sur laquelle elle apporterait un complément d'information ? Ou constitue-t-elle, à elle seule, une citation – ce qui, en dépit de son caractère lacunaire, la placerait au même niveau hiérarchique que les autres citations comportant un énoncé et une référence ? Pour trancher la question, nous avons utilisé la maquette pour « tester » la mobilité que requerrait l'élément. Or cette référence bibliographique tronquée et dépourvue d'énoncé renvoie à un fragment complet, présent ailleurs dans le corpus<sup>22</sup>. Elle a donc toute légitimité à figurer dans le module de reconstitution et doit donc être balisée comme un fragment.

En outre, construire une maquette de l'interface nous a permis de nous rendre compte qu'on ne pouvait pas se contenter, comme on le pensait au début, de désigner les fragments par une dénomination numérique du type « g226\_1\_116\_v\_\_f\_1 » formée à partir d'une combinaison de la cote patrimoniale et d'un numéro de fragment : le chercheur a besoin d'identifier plus précisément le fragment pour pouvoir se repérer dans son entreprise de reconstitution. On envisage donc de donner accès aux vedettes associées au fragment ainsi qu'aux premiers mots de l'énoncé. En outre, on fournira, grâce à un code couleur,

21. On a également prévu une fonction « panier » qui permettra à l'utilisateur de constituer des regroupements personnalisés fragment par fragment ou ensemble de fragments par ensemble de fragments, de les enregistrer et de les rappeler à loisir.
22. Il s'agit du Ms. g226<sup>3</sup> f° 18 où l'on trouve un fragment complet (énoncé et référence bibliographique) dont le rapprochement avec le fragment portant sur Fléchier cité sur le Ms. g226<sup>3</sup> f° 25 est tout à fait pertinent : « Bossuet, Fléchier, Massillon, dont nous n'avons jamais eu le bonheur de comprendre la haute éloquence, et qui seront toujours pour nous des hommes médiocres en fait de bon sens (p. 117, Michel Raymond, *Les intimes*) ».

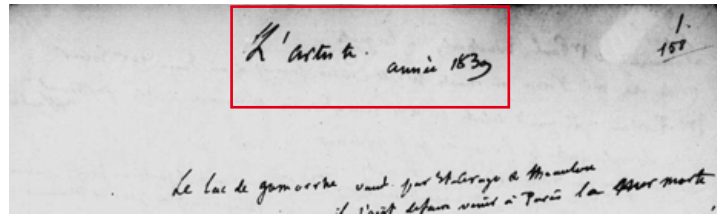


Figure 5 – un titre : l'exemple du Ms. g226<sup>1</sup> f° 158 (extrait)  
 (« Collections bibliothèque municipale de Rouen »).

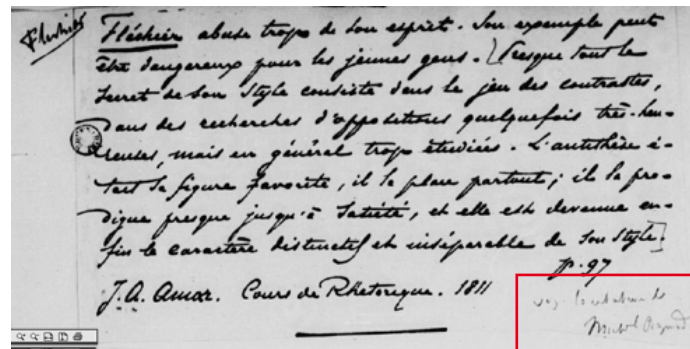


Figure 6 – les renvois bibliographiques allusifs : un exemple extrait  
 du Ms. g226<sup>3</sup> f° 25 (« Collections bibliothèque municipale de Rouen »).

l'appartenance typologique (note de lecture, récapitulation d'extraits, page préparée pour le second volume...) du feuillet où apparaît chaque fragment.

La maquette montre enfin que l'interface, initialement conçue pour proposer un mode d'édition du second volume plus pertinent qu'une édition papier, peut trouver des usages de recherche allant bien au-delà des reconstitutions conjecturales de l'œuvre. Elle permet de créer toutes sortes de regroupements, de types génétique (parcours de fragments), thématique ou chronologique. Sur un corpus aussi complexe et imposant, l'analyse ne peut procéder que par regroupements successifs. Effectuer toutes ces opérations du côté de la base de données présente l'intérêt de permettre de multiples modifications sans qu'il soit nécessaire d'intervenir sur les fichiers TEI. La question de l'intégration ultérieure des annotations scientifiques dans le fichier TEI se pose néanmoins, puisqu'il serait regrettable de ne pas les intégrer au corpus encodé. Nous aurons donc un nouveau défi à relever : produire une « deuxième version TEI » qui fusionnera la transcription délimitée avec les analyses scientifiques<sup>23</sup>.

## Les étapes du processus d'encodage global

Avant de conclure, rappelons les cinq étapes du processus global d'encodage :

1. les transcriptions des feuillets permettent de générer un fichier XML TEI minimal, organisé par page, sans balisage et ne récupérant que les sauts de ligne ;

23. Nous avons testé l'usage de l'élément <graph> des *guidelines* de la TEI (voir <http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/GD.html#GDGR>) pour représenter les relations génétiques d'un regroupement de fragments. Cet usage fait également partie des recommandations données dans les recherches du groupe de travail consacré aux éditions génétiques et qui devrait prochainement proposer l'intégration d'un nouveau module dédié (voir [http://wiki.tei-c.org/index.php/Genetic\\_Editions](http://wiki.tei-c.org/index.php/Genetic_Editions)).

2. la structuration des pages est réalisée manuellement à l'aide d'un éditeur XML<sup>24</sup> selon le schéma et les recommandations d'encodage définis spécifiquement pour le projet<sup>25</sup>. Les fragments sont alors délimités et caractérisés selon une perspective logique ;
3. un script parcourt les arborescences ainsi créées et extrait les fragments des pages pour alimenter la base des fragments dans la base de données ;
4. la transcription TEI est enrichie d'annotations sémantiques sans remettre en cause la délimitation des fragments: identification des noms de personnes ou de personnages, des événements historiques, des contenus épistémologiques convoqués et constitution des index... Parallèlement, le travail d'exploration du corpus et l'analyse scientifique se poursuivent du côté de la base de données. Ils donnent lieu à une ou plusieurs reconstitutions conjecturales de l'organisation du second volume, à la datation des fragments<sup>26</sup>, à l'identification de la source à partir de laquelle ils ont été copiés (source endogène: autre fragment du corpus; ou exogène: livre, journal, etc.);
5. à la fin du projet, le savoir scientifique enregistré dans la base de données sera « converti » en TEI. Ainsi formalisé, il sera réinjecté dans le document XML obtenu à l'issue de l'étape précédente. On obtiendra ainsi une version TEI complète, comprenant l'identification de chaque fragment, le balisage sémantique et la partie formalisable du savoir critique élaboré dans le cadre du projet. L'enjeu est alors la stabilisation d'un état du savoir scientifique sur ce corpus, sa pérennisation ainsi que sa mise à disposition d'autres projets.

## Conclusion

Répondre aux différents défis posés par l'édition des dossiers de *Bouvard et Pécuchet* a conduit l'équipe éditoriale à éprouver d'une part l'interdépendance des choix de structuration et d'encodage, et d'autre part celle du design des interfaces de lecture et de consultation.

Pour parvenir à se défaire de la fixité de l'édition papier et rendre aux fragments leur mobilité, il nous a préalablement fallu passer par une mise à plat de la TEI qui seule pouvait permettre de fixer une structure de base. La possibilité offerte par l'édition électronique de représenter les textes « en variance » repose donc paradoxalement sur une structuration initiale d'autant plus forte.

24. Le logiciel choisi est Oxygen.

25. Sans entrer dans le détail de ce travail, on précisera cependant qu'on utilise un fichier ODD (*One document does it all*) permettant de formaliser à la fois le schéma (le choix des balises et les listes d'attributs prédéfinis) et sa documentation. C'est l'utilisation des *guidelines* recommandée par le consortium.

26. Soit de manière absolue, soit de manière relative par rapport à un autre fragment dont on peut dire s'il est antérieur ou postérieur.

