Analyse intellectuelle d'un fonds numérisé : obstacles et possibles. Le projet d'Archive Numérique Desanti.

Lucie Chataigné, Najib Fawzi, Maud Ingarao, Pierre-Edouard Portier, David Wittmann.

L'œuvre de Jean-Toussaint Desanti est d'une grande richesse et d'une profonde fécondité. Elle est tout d'abord d'importance du point de vue de l'épistémologie française puisque Les Idéalité mathématiques et La Philosophie silencieuse poursuivent et approfondissent l'œuvre de Bachelard et celle de Cavaillès. La dimension phénoménologique de son travail, présente notamment dans Un destin philosophique ainsi que dans les dialogues avec D-A. Grisoni, outre des concepts absolument centraux comme celui de champ symbolico-charnel, conduit à des réflexions essentielles sur les questions de la subjectivité, de la signification, de la temporalité, de la mémoire ainsi que de l'éthique. Cette œuvre est enfin d'importance pour la qualité exceptionnelle de l'enseignement qui fut celui de Jean-Toussaint Desanti à l'ENS de Saint-Cloud puis à la Sorbonne.

Les archives de Jean-Toussaint Desanti représentent un ensemble évalué à 60000 pages. Il s'agit principalement de documents personnels du philosophe retrouvés à son domicile après sa disparition survenue en 2002. Ces archives ont été confiées par Dominique Desanti conjointement à l'IMEC pour leur conservation et à l'Institut Desanti (ENS LSH, UMR5037) pour leur exploitation intellectuelle et leur édition[1]. L'Institut Desanti et l'IMEC ont élaboré une stratégie de préservation et de valorisation de cet ensemble documentaire, formalisée en 2006 par une convention entre les deux établissements.

Ces archives se composent essentiellement d'un ensemble de cahiers comportant divers types d'inserts (feuillets supplémentaires, lettres etc.) et de plus de deux cents pochettes de taille variable contenant divers types de documents et de notes. Certains manuscrits correspondent à l'oeuvre publiée de Desanti (articles, ouvrages...) d'autres à des projets de livres ou d'articles restés inédits, à des notes de cours, etc.

Qu'est-ce qu'un « double numérique » ?

Le contenu des pochettes est souvent dans un état de désordre qui en rend le déchiffrage fort difficile : une même pochette peut comporter différents projets qui se chevauchent avec des numérotations de feuillets contradictoires, et souvent, les éléments d'un même projet sont dispersés dans plusieurs pochettes et cahiers. Dans ces circonstances procéder à un reclassement physique de l'archive comportait un très grand risque de perte d'information. Il a donc été décidé de numériser entièrement ces archives au plus près de leur état initial (nous verrons ci-après ce que cela implique) et de procéder au reclassement intellectuel de l'archive sur le double numérique ainsi constitué.

De mars 2007 à mars 2009, l'Institut Desanti a donc effectué, en interne[2], la numérisation complète de cet ensemble documentaire, puis restitué les originaux à l'IMEC en vue de leur conservation. Pour chaque pièce de l'archive initiale (cahiers, pochettes, etc.), David Wittmann établit une notice scientifique qui respecte les principes descriptifs de la norme archivistique ISAD(G)[3].

Ces opérations peuvent apparaître d'ampleur mais somme toute classiques. Pourtant,

1 L'institut a déjà réédité ou participé aux rééditions suivantes : Introduction à l'histoire de le la philosophie (PUF, 2006), Une pensée captive (PUF, 2008), Le philosophe et les pouvoirs (Hachette, 2008), Un destin philosophique (Hachette, 2008). Un ouvrage regroupant les principaux articles d'histoire des mathématiques est actuellement en préparation aux presses de l'ENS-LSH.

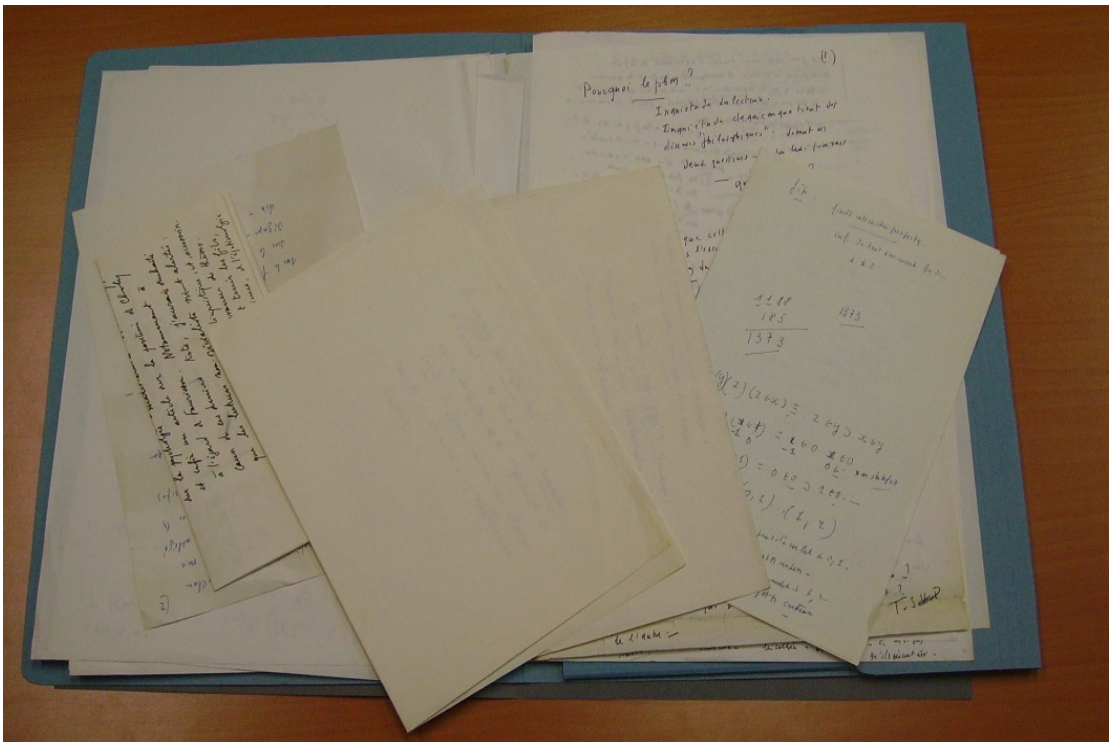
2 Matériel de numérisation : CopiBook I2S de l'ENS LSH.

3 Une notice comprend des éléments de datation, une description intellectuelle, un sommaire et des indexes. Le catalogue des notices ainsi constitué, qui sera accessible en ligne début 2010 (au format EAD), permet une première exploration du fonds et assure le lien entre l'archive physique et l'archive numérique.

l’objectif de numériser l’archive au plus près de l’état dans lequel elle a été découverte a en réalité ouvert d'emblée de nombreuses perspectives et posé dans le même mouvement plusieurs questions méthodologiques.

Tout d'abord, vouloir numériser l'archive « au plus près de son état initial » a posé question. Le système de nommage des fichiers choisi devait notamment permettre de conserver très précisément l’information relative à l’imbrication des feuillets les uns dans les autres, tout en restant maniable informatiquement. Nous avons choisi de représenter ces niveaux d'insertion par une structure arborescente de répertoires informatiques[4].

Malgré des règles de nommage précises, les chercheurs ont très vite constaté que le travail sur un double numérique posait des problèmes de visualisation et de perte d'information, du fait notamment des très nombreuses manières différentes dont les documents peuvent être imbriquées les uns dans les autres (cf. photo). De ce fait, la personne qui numérise a, dans des cas assez nombreux, une certaine marge d'appréciation quant à la manière concrète de réaliser la capture. Les premières règles de numérisation et de nommage ont donc été doublées par un jeu de règles supplémentaires permettant de désambigüiser l'écriture graphique de l'archive. Mais dans le même temps, il fallait s'interdire de multiplier les cas particuliers, car trop d'exceptions auraient anéanti toute possibilité ultérieure de traitement automatique de l'information[5].



1: Un exemple rencontré très fréquemment est celui de feuillets A3 pliés en liasse A4 et insérées dans un ensemble de plus haut niveau avec la "reliure" (la pliure) située tantôt vers l'extérieur, tantôt vers l'intérieur.

En somme, dans ce travail sur le nommage des fichiers, il est important de retenir que les

4 L'image nommée 0125/48_5/54_1/0009.tif correspond par exemple à la 9ème page (recto du 5ème feuillet) du premier groupe de feuillets insérés entre les pages 54 et 55 du 5ème groupe de feuillets insérés entre les pages 48 et 49 de la pièce n°125

5 Une règle supplémentaire très simple a ainsi été adoptée : une image de numéro pair doit correspondre au verso de la page à laquelle correspond l'image de numéro impair qui précède. Cette simple règle permet aux personnes qui numérisaient de trancher dans la quasi-totalité des cas entre plusieurs façons possibles d'effectuer la capture. A l'autre bout de la chaîne, cette règle est aussi un guide pour les chercheurs qui consultent l'archive lorsqu'une structure leur semble ambiguë. Le « coût » d'une telle règle (en temps et en espace de stockage) est qu'elle oblige à numériser des pages ne comportant aucune inscription manuscrite.

choix se sont fait en tenant compte de trois paramètres en tension les uns avec les autres :

- qualité informatique des noms de fichiers (régularité, portabilité) ;
- temps de numérisation supplémentaire et risques d'erreurs induits par les interventions sur le générateur des noms de fichiers au moment de la capture ;
- efficacité en termes de restitution de l'information aux chercheurs.

Le projet Desanti illustre ainsi, après d'autres, en quoi il peut être fécond d'envisager la numérisation comme constitutive d'un projet scientifique d'exploitation de corpus, plutôt que comme une étape technique préalable ou annexe, et de permettre aux chercheurs de s'impliquer dans la planification et le suivi de la numérisation, afin que l'opération serve au mieux les spécificités du corpus.

Annotation collaborative et multistructure.

Au-delà du classement au sens archivistique, il devient en outre possible sur une archive numérique d'effectuer autant de classements concurrents qu'on le souhaite en fonction d'objectifs divers. Se posent alors des questions de représentation informatique de l'archive afin de rendre exploitable ces classements multiples qui, à la fois peuvent être exclusifs les uns des autres, et à la fois doivent pouvoir s'alimenter et s'enrichir mutuellement.

C'est ainsi que le projet d'archive numérique J-T. Desanti donne lieu depuis novembre 2007 au travail de thèse en informatique de Pierre-Edouard Portier (LIRIS, UMR CNRS 5205) sur l'annotation collaborative de corpus et la création de documents multistructurés[6].

Le désordre initial de l'archive, tant du fait de J-T. Desanti que des mains par lesquelles elle passa, intéresse la thématique très actuelle de l'annotation collaborative — nous pensons aux outils disponibles en ligne tels Youtube, Flickr, Delicious, etc. et à leurs mécanismes de mots-étiquettes ("tags") ou de fils de commentaires — en tant que les besoins de reclassement qu'il implique signalent les limites de l'expressivité de ces systèmes et appellent la construction de modèles mieux adaptés.

Ainsi, un chercheur peut proposer pour une partie de l'archive un reclassement dont la représentation formelle est une structure hiérarchique de pages. Chaque élément de cette structure, page ou groupement de page, est éventuellement l'objet d'annotations. Cependant, un autre chercheur (ou peut-être le même) propose un second classement de la même partie de l'archive (ou d'une partie qui partage certaines de ses pages avec la première) auquel est aussi associée une structure hiérarchique. Maintenant, une page, en tant qu'elle appartient de manière bien déterminée à l'un des classements, doit pouvoir être annotée distinctement ; mais également, pour deux versions d'une même page physique dans des classements différents, il est nécessaire de toujours conserver l'origine commune afin de satisfaire à la précision que requièrent les études philologiques. Autrement dit, l'objet numérique, ici la page, doit toujours non seulement porter la trace de son contexte de création, mais encore se présenter au sein d'un complexe d'autres objets qui favorisent la richesse des interprétations que l'utilisateur en fera.

C'est encore ce souci d'un usage non réducteur des capacités de calcul et de mémoire des systèmes d'information pour aider au déroulement de processus herméneutiques qui mettent en jeu plusieurs utilisateurs, qui nous a amené à reconsidérer la problématique

6 Cette thèse est financée dans le cadre du Cluster régional "Culture, Patrimoine et Création", dont l'objectif central est de coordonner les recherches pluridisciplinaires portant sur des projets qui possèdent une dimension d'ordre culturel et patrimonial en Rhône-Alpes.

des "documents multistructurés". Cette dernière a une origine seulement technologique : l'impossibilité pour le formalisme de représentation de données dominant, XML, de décrire des éléments qui se chevauchent. Or, plusieurs usages peuvent impliquer des structurations hiérarchiques concurrentes, par exemple, si une zone de texte qui chevauche deux pages est identifiée par un utilisateur comme présentant un intérêt particulier, alors il est impossible, sauf à utiliser des techniques qui rendent le document très lourd à manipuler, de représenter dans un même document XML la structuration physique des pages et cette zone d'intérêt. Après avoir choisi de résoudre la dimension technique du problème au moyen d'une méthode bien connue ("stand-off markup") de séparation des structures et du contenu, nous avons proposé une méthodologie pour la création de tels documents. Cette dernière facilite la multiplication des structures documentaires et la mise en relation de leurs auteurs en tant qu'ils utilisent des structures similaires[7].

7 P.E. Portier, S. Calabretto. *Multistructured documents: beyond overlapping hierarchies* DocEng 2009, 15-18 septembre 2009, Munich, ISBN 978-1-60558-575-8. pp. 181-184