

ÉCOLE NATIONALE DES CHARTES

Camille Besse

licenciée ès Spécialiste HES en Information documentaire

Numérisation de masse : vers la création d'un nouvel acteur de l'information

LE PROJET TIME MACHINE

Mémoire pour le diplôme de master
« Technologies numériques appliquées à l'histoire »

2019

Résumé

Ce mémoire a été réalisé en vue de l'obtention du diplôme de Master 2 « Technologies numériques appliquées à l'histoire » de l'École nationale des chartes. Il retrace le travail que nous avons mené lors d'un stage de quatre mois réalisé au sein du Laboratoire d'humanités digitales de l'École polytechnique fédérale de Lausanne, au bénéfice du projet Time Machine. Cette initiative européenne, portée par un consortium d'institutions et de réseaux culturels et patrimoniaux, sous la coordination du laboratoire lausannois, vise entre autres à organiser la numérisation à l'échelle européenne et proposer de nouveaux paradigmes d'accès aux données numérisées par le biais des technologies d'intelligence artificielle, ajoutant par exemple un facteur temporel aux plateformes traditionnelles. Ce mémoire porte sur l'élaboration d'une feuille de route pour l'infrastructure et les opérations du projet, planifiée afin d'atteindre une phase de maturité sous dix ans. Pour mieux appréhender la complexité du travail effectué, il s'agit également d'un rapport sur l'historique, les politiques culturelles, les caractéristiques et les enjeux qui découlent des initiatives de numérisation de masse. Une analyse plus critique exposera les réponses apportées par Time Machine aux différents enjeux de la numérisation, les écueils qu'il reste à éviter et le potentiel impact d'un projet de cette envergure sur le monde politique culturel et patrimonial européen, posant ainsi la question de la place occupée par ces initiatives au sein de cet environnement d'acteurs.

Mots-clefs : accessibilité ; bibliothèques numériques ; droit d'auteur ; Europeana ; *HathiTrust* ; humanités numériques ; *Google Books* ; interopérabilité ; intelligence artificielle ; moteur de recherche diachronique ; numérisation de masse ; partenariats public-privé ; politiques culturelles européennes ; surveillance de masse ; *the Public Library of America* ; valorisation

Informations bibliographiques : Camille Besse, *Numérisation de masse : vers la création d'un nouvel acteur de l'information - le projet Time Machine*, mémoire de master « Technologies numériques appliquées à l'histoire », Thibault Clérice et Frédéric Kaplan, École nationale des chartes, 2019.

Remerciements

Mes remerciements s'adressent d'abord à mon maître de stage, le Professeur Frédéric Kaplan, détenteur de la chaire d'humanités digitales de l'EPFL, pour les échanges vivifiants qui ont rythmé les quatre mois de mon expérience, et surtout pour la confiance qu'il m'a accordée dès les premiers jours. Je remercie les membres du projet Time Machine, François Ballaud pour m'avoir guidée durant mes premières semaines à travers les complexités du projet, ainsi qu'Isabella di Lenardo pour avoir partagé son expertise acquise sur le projet vénitien. Un immense merci à Kevin Baumer, coordinateur de Time Machine, pour m'avoir fait découvrir les coulisses du projet, et partagé mes questionnements de stagiaire. J'ai grandement apprécié son écoute bienveillante et sa constante bonne humeur. Je remercie également chaleureusement mes collègues et autres stagiaires du DHLAB, pour m'avoir accueillie généreusement au sein de leur équipe, m'avoir aidée à comprendre les processus technologiques si délicats de Time Machine et apporté leur soutien indéfectible durant la durée de mon stage. C'est grâce à eux que j'ai pu garder le cap face à la complexité de ma tâche. Je remercie les bibliothécaires du *Rolex Learning Center*, pour m'avoir épaulée face aux subtilités de la gestion des données de la recherche au sein d'un projet d'ampleur européenne, ces moments de partage me furent très précieux.

Je remercie vivement mes professeurs Gautier Poupeau (*data architect* à l'INA) et Clément Oury (Adjoint au chef du service Conservation et Numérisation, Bibliothèque du Muséum d'Histoire naturelle), pour m'avoir prodigué conseils et avis et pris le temps de me recevoir, lorsque j'en ai eu besoin.

Je tiens encore à adresser mes remerciements à Thibault Clérice, directeur du master et tuteur de stage pour m'avoir permis d'intégrer le master et aidée à prendre du recul par rapport à mes activités de stagiaire.

Enfin un merci spécial à ma relectrice de toujours, Danièle Besse, qui n'est jamais effrayée par l'ampleur de la tâche, et à mon compagnon Dimitri Wyss pour m'avoir soutenue durant ces deux années parisiennes.

Bibliographie

Historique de la numérisation

- ASSOCIATION POUR LE PATRIMOINE NATUREL ET CULTUREL DU CANTON DE VAUD, *Patrimoine numérique, numérisation du patrimoine*, Lausanne, 2012.
- BATTU (Daniel), *L'histoire et l'économie du monde accompagnées par les TIC*, OCLC : 1030367775, London, 2018.
- BEAUDOIN (Joan E.), “Context and Its Role in the Digital Preservation of Cultural Objects”, *D-Lib Magazine*, 18–11/12 (nov. 2012), DOI : 10.1045/november2012-beaudoin1.
- CALHOUN (Karen), *Exploring digital libraries. Foundations, practice, prospects*. OCLC : 894201348, Chicago, 2014.
- CH’NG (Eugene), CAI (Shengdan), ZHANG (Tong Evelyn) et LEOW (Fui-Theng), “Crowd-sourcing 3D cultural heritage : best practice for mass photogrammetry”, *Journal of Cultural Heritage Management and Sustainable Development*, 9–1 (févr. 2019), p. 24–42, DOI : 10.1108/JCHMSD-03-2018-0018.
- Check It Out with Andrew Richard Albanese : How the Google Books Case Could Change Fair Use On Campus*, en, URL : <https://www.publishersweekly.com/pw/by-topic/industry-news/libraries/article/68535-check-it-out-with-andrew-richard-albanese-gsu-and-leval-s-opinion-on-google-books.html> (visité le 03/07/2019).
- Commission Recommendation of 27 October 2011 on the digitisation and online accessibility of cultural material and digital preservation*, en, rapp. tech. 32011H0711, 2011, URL : <http://data.europa.eu/eli/reco/2011/711/oj/eng> (visité le 22/07/2019).
- COOK (Michael), *Google’s Moon Shot : The quest for the universal library*, en-US, URL : <http://www.gutenbergnews.org/20070131/googles-moon-shot-the-quest-for-the-universal-library/> (visité le 04/07/2019).
- CORRADO (Edward M.) et MOULaison SANDY (Heather), *Digital preservation for libraries, archives, and museums*, Second edition, Lanham, Maryland, 2017.
- Council adopts Copyright Directive*, en, URL : <http://era.gv.at/object/news/4678> (visité le 04/07/2019).

- COUTTS (Margaret), *Stepping away from the silos : strategic collaboration in digitisation*, OCLC : ocn952385918, Cambridge, MA, 2017 (Chandos advances in information series).
- COX (Krista), “Authors Guild v. HathiTrust Litigation Ends in Victory for Fair Use”, *ARL Policy Notes blog* (, août 2015).
- COYLE (Karen), “Mass Digitization of Books”, *The Journal of Academic Librarianship*, 32–6 (nov. 2006), p. 641-645, DOI : 10.1016/j.acalib.2006.08.002.
- DCC Curation Lifecycle Model / Digital Curation Centre, URL : <http://www.dcc.ac.uk/resources/curation-lifecycle-model> (visité le 29/06/2019).
- DIEKEMA (Anne R.), “Multilinguality in the digital library : A review”, *The Electronic Library*, 30–2 (avr. 2012), p. 165-181, DOI : 10.1108/02640471211221313.
- DITTRICH (Paul-Jasper), “Balancing ambition and pragmatism for the Digital Single Market”, *Jacques Delors Institut* (, juil. 2017), p. 14, URL : <https://www.delorsinstitut.de/2015/wp-content/uploads/2017/09/BalancingAmbitionandPragmatismfortheDigitalSingleMarket-Dittrich-JDIB-Sept2017.pdf> (visité le 01/07/2019).
- DRENNAN (James), *European Framework for Action on Cultural Heritage*, en, Text, déc. 2018, URL : https://ec.europa.eu/culture/content/european-framework-action-cultural-heritage_en (visité le 12/07/2019).
- DUBIS (Mark), “Web Resources for the Study of New Testament Backgrounds”, *Journal of Religious & Theological Information*, 6–1 (janv. 2003), p. 3-9, DOI : 10.1300/J112v06n01_02.
- Bernadette Dufrêne, Madjid Ihadjadene, Denis Bruckmann, Université Paris Ouest Nanterre La Défense, Université de Paris VIII et Bibliothèque nationale de France (éd.), *Numérisation du patrimoine : quelles médiations ? quels accès ? quelles cultures ?,* OCLC : 859441811, Paris, 2013.
- DUNNING (Alastair), “Digitising the past : Next steps for public-sector digitisation”, *JISC (Joint Information Systems Committee)* (, juil. 2009), URL : <https://core.ac.uk/download/pdf/11890177.pdf> (visité le 01/07/2019).
- ERWAY (Ricky) et OCLC RESEARCH, *Rapid capture : faster throughput in digitization of special collections*, OCLC : 712406431, Dublin, Ohio, 2011, URL : <http://www.oclc.org/research/publications/library/2011/2011-04.pdf> (visité le 04/07/2019).
- “EU Adopts Single-Market Digital Strategy”, *Information Management Journal* (, juin 2015).
- EU Member States sign up to cooperate on digitising cultural heritage*, en, Text, avr. 2019, URL : <https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-digitising-cultural-heritage> (visité le 12/07/2019).
- European Commission report on Cultural Heritage : Digitisation, Online Accessibility and Digital Preservation*, en, Text, juin 2019, URL : <https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-digitising-cultural-heritage> (visité le 12/07/2019).

- single-market/en/news/european-commission-report-cultural-heritage-digitisation-online-accessibility-and-digital (visité le 04/07/2019).
- Salzburg Research Forschungsgesellschaft et Europäische Kommission (éd.), *The Digi-CULT Report : technological landscapes for tomorrow's cultural economy ; unlocking the value of cultural heritage : executive summary*, OCLC : 51714305, Luxembourg, 2002.
- FOULONNEAU (Muriel), "Recherche et numérisation du patrimoine en Europe", *Document numérique*, Vol. 7–3 (2003), p. 179-189, URL : <https://www.cairn.info/reve-document-numerique-2003-3-page-179.htm> (visité le 16/07/2019).
- GEORGE (Carole A.), "Testing the barriers to digital libraries : A study seeking copyright permission to digitize published works", *New Library World*, 106–7/8 (juil. 2005), p. 332-342, DOI : 10.1108/03074800510608648.
- GEORGETOWN UNIVERSITY, RIBES (David), FINHOLT (Thomas) et UNIVERSITY OF MICHIGAN, "The Long Now of Technology Infrastructure : Articulating Tensions in Development", *Journal of the Association for Information Systems*, 10–5 (mai 2009), p. 375-398, DOI : 10.17705/1jaис.00199.
- GERTZ (Janet), "Selection for Preservation in the Digital Age", *Library Resources & Technical Services*, 44–2 (avr. 2000), p. 97-104, DOI : 10.5860/lrts.44n2.97.
- GREEN (Andrew), "Big digitisation : Origins, progress and prospects", *International Journal of Humanities and Arts Computing*, 4–1–2 (oct. 2010), p. 55-66, DOI : 10.3366/ijhac.2011.0007.
- GUPTA (Dinesh K.) et SHARMA (Veerbala), "Enriching and enhancing digital cultural heritage through crowd contribution", *Journal of Cultural Heritage Management and Sustainable Development*, 7–1 (févr. 2017), p. 14-32, DOI : 10.1108/JCHMSD-12-2014-0043.
- HOEKSTRA (Johanna) et DIKER-VANBERG (Aysem), "The proposed directive for the supply of digital content : is it fit for purpose?", *International Review of Law, Computers & Technology*, 33–1 (janv. 2019), p. 100-117, DOI : 10.1080/13600869.2019.1562638.
- HOOLAND (Seth van) et VERBORGH (Ruben), *Linked data for libraries, archives and museums : how to clean, link and publish your metadata*, OCLC : ocn872987681, London, 2014.
- IBEKWE (Fidelia), *European Origins of Library and Information Science*, t. 13, 2019 (Studies in Information), DOI : 10.1108/S2055-5377201913.
- JONES (Elisabeth), "The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative : Assumptions, Intentions, and the Role of the Public", *Information & Culture*, 52–2 (mai 2017), p. 229-263, DOI : 10.7560/IC52205.

- KELLY (Kevin), “AR Will Spark the Next Big Tech Platform—Call It Mirrorworld”, *Wired* (, févr. 2019), URL : <https://www.wired.com/story/mirrorworld-ar-next-big-tech-platform/> (visité le 03/07/2019).
- KIM (Dorothy), RUSSWORM (Treaandrea), VAUGHAN (Corrigan), ADAIR (Cassius), PAREDES (Veronica) et COWAN, “Race, Gender, and the Technological Turn : A Roundtable on Digitizing Revolution”, dir. Everett et Guisela Latorre, *University of Nebraska Press*, *Frontiers : A Journal of Women Studies* 39–1 (), p. 149-177.
- L'Europe après 60 ans = : Europe 60 years later*, Paris, 2017 (Revue d'économie financière, no 125 (2017)), URL : <https://www.cairn.info/revue-d-economie-financiere-2017-1-page-239.htm>.
- LAMPERT (Cory), “Ramping up : Evaluating large-scale digitization potential with small-scale resources”, *Digital Library Perspectives*, 34–1 (févr. 2018), p. 45-59, DOI : 10.1108/DLP-06-2017-0020.
- ANGLEY (Adam) et BLOOMBERG (Dan S.), “Google Books : making the public domain universally accessible”, dans *Document Recognition and Retrieval XIV*, 2007, t. 6500, 65000H, DOI : 10.1117/12.710609.
- LAULAN (Anne-Marie), “Diversité culturelle et mondialisation”, *Hermès, La Revue*, 80–1 (2018), p. 168-174, URL : <https://www.cairn.info/revue-hermes-la-revue-2018-1-page-168.htm>.
- LOPATIN (Laurie), “Library digitization projects, issues and guidelines”, *Library Hi Tech* (, avr. 2006), DOI : 10.1108/07378830610669637.
- MATTELART (Armand), *Histoire de la société de l'information*, OCLC : 1082200705, 2018.
- MEISSNER (Dennis) et GREENE (Mark A.), “More Application while Less Appreciation : The Adopters and Antagonists of MPLP”, *Journal of Archival Organization*, 8–3-4 (juil. 2010), p. 174-226, DOI : 10.1080/15332748.2010.554069.
- Numérisation*, août 2013, URL : <https://www.enssib.fr/le-dictionnaire/numerisation> (visité le 11/07/2019).
- ORGANIZATION (National Information Standards), *Special Information Standards Quarterly Issue on the Evolution of Bibliographic Data Exchange Published by NISO*, URL : <https://www.niso.org/press-releases/2013/12/special-information-standards-quarterly-issue-evolution-bibliographic-data> (visité le 04/07/2019).
- Orphan Works DB*, URL : <https://euipo.europa.eu/orphanworks/> (visité le 16/07/2019).
- Henriette Roued-Cunliffe et Andrea Copeland (éd.), *Participatory heritage*, OCLC : ocn971483437, London, 2017.
- SALAÜN (Jean-Michel), *Vu, lu, su : les architectes de l'information face à l'oligopole du Web*, Paris, 2012 (Cahiers libres).
- SCHROFF (Simone) et STREET (John), “The politics of the Digital Single Market : culture vs. competition vs. copyright”, *Information, Communication & Society*, 21–10 (oct. 2018), p. 1305-1321, DOI : 10.1080/1369118X.2017.1309445.

- STUART (David), *Facilitating access to the Web of data a guide for librarians*, OCLC : 776688108, London, 2011.
- The Next Steps for the Digital Single Market : From Where do We Start ?, URL : https://ecipe.org/publications/the-next-steps-for-the-digital-single-market-from-where-do-we-start/* (visité le 04/07/2019).
- THYLSTRUP (Nanna Bonde), *The politics of mass digitization*, Cambridge, MA, 2018.
- VIOLA (Roberto) et BRINGER (Olivier), “Vers un marché unique numérique : faire de la révolution numérique une opportunité pour l’Europe”, *Revue d’économie financière*, 125–1 (2017), p. 239, DOI : 10.3917/ecofi.125.0239.
- WARWICK (Claire), “Studying users in digital humanities”, dans *Digital Humanities in Practice*, dir. Claire Warwick, Melissa Terras et Julianne Nyhan, 1^{re} éd., 2012, p. 1-22, DOI : 10.29085/9781856049054.002.
- WEISS (Andrew), *Using massive digital libraries : a LITA guide*, First edition, Chicago, 2014 (LITA guides).
- WU (Michelle), “Building a Collaborative Digital Collection : A Necessary Evolution in Libraries”, *Georgetown Law Faculty Publications and Other Works*. 699 (, 2011).
- XIE (Iris) et MATUSIAK (Krystyna K.), *Discover digital libraries : theory and practice*, OCLC : 907120360, Amsterdam Boston Heidelberg, 2016.
- ZEINSTRA (Maarten), *Research : Orphan Works Directive does not work for mass digitisation*, en-US, févr. 2016, URL : <https://www.communia-association.org/2016/02/16/orphan-works-directive-does-not-work/> (visité le 16/07/2019).

Enjeux de la numérisation

- BERMÈS (Emmanuelle), ISAAC (Antoine) et POUPEAU (Gautier), *Le Web sémantique en bibliothèque*, OCLC : 866574281, Paris, 2013.
- Thierry Claerr et Isabelle Westeel (éd.), *Manuel de la numérisation*, Paris, 2011 (Bibliothèques).
- CONSORTIUM (IIIF), *Home — IIIF / International Image Interoperability Framework*, URL : <https://iiif.io/> (visité le 18/05/2019).
- COÜASNON (Bertrand), *Accès par le contenu aux documents manuscrits anciens numérisés*, document pdf, 2019.
- *Documents numériques Documents textuels*, document pdf, 2019.
- *Recherche dans des images iconographiques*, document pdf, 2019.
- DELESTRE (Nicolas), MALANDAIN (Nicolas) et BUSSI (Michel), *Du Web des documents au Web sémantique*, OCLC : 974379992, 2017.
- ESCHENFELDER (Kristin R.), SHANKAR (Kalpana), WILLIAMS (Rachel D.), SALO (Dorothea), ZHANG (Mei) et LANGHAM (Allison), “A nine dimensional framework for digital cultural heritage organizational sustainability : A content analysis of the LIS literature (2000–2015)”, *Online Information Review*, 43–2 (avr. 2019), p. 182-196, DOI : 10.1108/OIR-11-2017-0318.
- EUROPEANA, *Linked Open Data - What is it ? on Vimeo*, URL : <https://vimeo.com/36752317> (visité le 18/05/2019).
- INSTITUT NATIONAL D'HISTOIRE DE L'ART, #LundisNum / 11 février 2019 - Gautier Poupeau : rassembler les métadonnées des collections de l'INA, URL : https://www.youtube.com/watch?list=PLs18NWzVv6T2CQFtBOfn1A_EKLFeCFSUG&time_continue=52&v=KY0zoRPks8Q&fbclid=IwAR3QgiLH5rEPiDGFx-PPdIMVsH_83H3mJuHnn8jeJBVi2_01-jVWJa4hZW8 (visité le 18/05/2019).
- JURIK (Bolette Ammitzbøll), BLEKINGE (Asger Askov), FERNEKE-NIELSEN (Rune Bruun) et MØLDRUP-DALUM (Per), “Bridging the gap between real world repositories and scalable preservation environments”, *International Journal on Digital Libraries*, 16–3-4 (sept. 2015), p. 267-282, DOI : 10.1007/s00799-015-0152-4.
- KOWALCZYK (Stacy T.), *Digital curation for libraries and archives*, Santa Barbara, California, 2018.
- MANNING (Patrick), *Big Data in History*, London, 2013, DOI : 10.1057/9781137378972.

MAUREL (Lionel), “Quel modèle économique pour une numérisation patrimoniale respectueuse du domaine public ?”, dans *Communs du savoir et bibliothèques*, Paris, 2017 (Bibliothèques), p. 73-84, DOI : 10.3917/elec.dujo.2017.01.0073.

MINISTÈRE DE LA CULTURE ET DE LA COMMUNICATION : DIRECTION DES ARCHIVES DE FRANCE, *Ecrire un cahier des charges de numérisation du patrimoine*, rapp. tech., URL : https://francearchives.fr/file/bf50d8fa5f554586dbf18fdc862d25970a1da0a7static_4132.pdf.

MOATTI (Alexandre), “Bibliothèque numérique européenne : de l’utopie aux réalités”, *Réalités Industrielles* (, nov. 2012), p. 43-46.

MONS (Barend), *Data Stewardship For Open Science : implementing FAIR principles*, Boca Raton, 2018.

OURY (Clément), *Les données bibliographiques : Modélisation et structuration*, document pdf, 2019.

PAVLOVIC (Aleksandra), “The Serpent in the Garden of Eden : Intellectual property in the Digital Millennium”, *Academia.edu* (, déc. 2011), URL : https://www.academia.edu/34127494/_The_Serpent_in_the_Garden_of_Eden_Intellectual_property_in_the_Digital_Millennium (visité le 04/07/2019).

POMERANTZ (Jeffrey), *Metadata*, Cambridge, Massachusetts ; London, England, 2015 (The MIT Press essential knowledge series).

POUPEAU (Gautier), *Modèles conceptuels et ontologie*, document powerpoint, 2019.

— *Open Data, Big data, Data Mining*, document powerpoint, 2019.

RASMUSSEN PENNINGTON (Diane) et CAGNAZZO (Laura), “Connecting the silos : Implementations and perceptions of linked data across European libraries”, *Journal of Documentation*, 75–3 (mai 2019), p. 643-666, DOI : 10.1108/JD-07-2018-0117.

SHANKAR (Kalpana) et ESCHENFELDER (Kristin R.), “Sustaining Data Archives over Time : Lessons from the Organizational Studies Literature”, *New Review of Information Networking*, 20–1–2 (juil. 2015), p. 248-254, DOI : 10.1080/13614576.2015.1111699.

STOBO (Victoria), PATTERSON (Kerry), ERICKSON (Kristofer) et DEAZLEY (Ronan), “I should like you to see them some time : An empirical study of copyright clearance costs in the digitisation of Edwin Morgan’s scrapbooks”, *Journal of Documentation* (, janv. 2018), DOI : 10.1108/JD-04-2017-0061.

THE DUTCH TECHCENTRE FOR LIFE SCIENCES (DTL), *Vision on Open Science*, fr, URL : <https://vimeo.com/162062013> (visité le 18/05/2019).

VARNIENĖ-JANSSEN (Regina) et KUPIRIENĖ (Jūratė), “Authenticity and Provenance in Long-Term Digital Preservation : Analysis of the Scope of Content”, *Informacijos mokslai*, 82 (déc. 2018), p. 131-160, DOI : 10.15388/Im.2018.82.9.

YEATES (Robin) et GUY (Damon), “Collaborative working for large digitisation projects”, *Program*, 40–2 (avr. 2006), p. 137-156, DOI : 10.1108/00330330610669262.

ZENGENENE (Dydimus), “Global interoperability and linked data in libraries”, *New Library World*, 114–1/2 (janv. 2013), p. 84-87, DOI : 10.1108/03074801311291992.

ZHAROVA (Anna), “Influence of the principle of interoperability on legal regulation”, *International Journal of Law and Management*, 57–6 (nov. 2015), p. 562-572, DOI : 10.1108/IJLMA-07-2014-0044.

Recherche en humanités numériques

- ASSOCIATION FRANCOPHONE DES HUMANITÉS NUMÉRIQUES, *Humanistica*, fr-FR, URL : <http://www.humanisti.ca/> (visité le 15/06/2019).
- BOULAIRE (Cécile), BOURGATTE (Michaël), CARABELLI (Romeo), CARTIER (Aurore), CHAGNOUX (Marie), DUCROS (Jérémy), FABRE (Chloée), FONIO (Filippo), GRANDI (Elisa), HAUTCŒUR (Pierre-Cyrille), *et al.*, *Expérimenter les humanités numériques : Des outils individuels aux projets collectifs*, OCLC : 1089419502, Montréal, 2018, URL : <http://books.openedition.org/pum/11091> (visité le 09/06/2019).
- CARACO (Benjamin), “Les Digital humanities et les bibliothèques”, *Bulletin des bibliothèques de France (BBF)*—Bulletin des bibliothèques de France (BBF) (2012), p. 69-73.
- CLAVERT (Frédéric) et SCHAFER (Valérie), “Les humanités numériques, un enjeu historique”, *Quaderni*–98 (févr. 2019), p. 33-49, DOI : 10.4000/quaderni.1417.
- DARIAH, *Digital Research Infrastructure for the Arts and Humanities*, en-GB, URL : <https://www.dariah.eu/> (visité le 17/06/2019).
- DH2019, *Digital Humanities conference 2019 – Utrecht*, en-US, URL : <https://dh2019.adho.org/> (visité le 15/06/2019).
- DHBENELUX 2019, *DHBenelux 2019*, en-US, URL : <http://2019.dhbenelux.org/> (visité le 15/06/2019).
- EVANGELISTA (Sandy) et BROUET (Anne-Muriel), “Time Machine dans la course au FET Flagship européen” (, févr. 2018), URL : <https://actu.epfl.ch/news/time-machine-dans-la-course-au-fet-flagship-europe/> (visité le 20/06/2019).
- GEFEN (Alexandre), “Des humanités numériques en 2017”, *Mélanges de la Casa de Velázquez. Nouvelle série*–47-2 (nov. 2017), p. 315-318, DOI : 10.4000/mcv.7957.
- MEUNIER (Jean-Guy), “Le paradoxe des humanités numériques”, *Quaderni*–98 (févr. 2019), p. 19-31, DOI : 10.4000/quaderni.1407.
- MOUNIER (Pierre), *Lou Burnard : Du Literary and linguistic computing aux Digital Humanities : retour sur 40 ans de relations entre sciences humaines et informatique*, fr-FR, Billet, URL : <https://leo.hypotheses.org/3764> (visité le 20/06/2019).
- SCHREIBMAN (Susan), SIEMENS (Raymond George) et UNSWORTH (John), *A new companion to digital humanities*, OCLC : 953120034, 2016.

THATCAMP, *The Humanities and Technology Camp*, en, URL : <http://thatcamp.org/> (visité le 15/06/2019).

WARWICK (Claire), TERRAS (Melissa M) et NYHAN (Julianne), *Digital humanities in practice*, OCLC : 836872277, London, 2012, URL : http://www.123library.org/book_details/?id=92814 (visité le 10/06/2019).

Contexte du Projet Time Machine

- ABBOTT (Alison), “The ‘time machine’ reconstructing ancient Venice’s social networks”, *Nature News*, 546–7658 (juin 2017), p. 341, DOI : 10.1038/546341a.
- AMSTERDAM TIME MACHINE, *Amsterdam Time Machine*, en-US, URL : <https://amsterdamtimemachine.nl/> (visité le 23/06/2019).
- ANTWERP (University of), *Antwerp Time Machine*, URL : <https://www.uantwerpen.be/en/projects/antwerp-time-machine/> (visité le 23/06/2019).
- ARCHiVe *Analysis and Recording of Cultural Heritage in Venice*, it-IT, URL : <https://www.cini.it/istituti-e-centri/archive-analysis-and-recording-of-cultural-heritage-in-venice> (visité le 23/06/2019).
- ARTE, *Tomography - The Digitisation of the Future (3-8) - Venice Time Machine*, en, URL : <https://www.arte.tv/en/videos/075631-003-A/tomography-the-digitisation-of-the-future-3-8/> (visité le 18/05/2019).
- *Venice Time Machine - History and Big Data (1/8)*, en, URL : <https://www.arte.tv/en/videos/075631-001-A/venice-time-machine-history-and-big-data-1-8/> (visité le 18/05/2019).
 - *Venice Time Machine - History and Big Data (2/8) - Preserving the Past*, en, URL : <https://www.arte.tv/en/videos/075631-002-A/venice-time-machine-history-and-big-data-2-8/> (visité le 18/05/2019).
 - *Venice Time Machine - History and Big Data (4-8) - Making Sense of Archives*, en, URL : <https://www.arte.tv/en/videos/075631-004-A/venice-time-machine-history-and-big-data-4-8/> (visité le 18/05/2019).
 - *Venice Time Machine - History and Big Data (5/8) - Four-Dimensional History*, en, URL : <https://www.arte.tv/en/videos/075631-005-A/venice-time-machine-history-and-big-data-5-8/> (visité le 18/05/2019).
 - *Venice Time Machine - History and Big Data (6/8) - History Through Pictures*, en, URL : <https://www.arte.tv/en/videos/075631-006-A/venice-time-machine-history-and-big-data-6-8/> (visité le 18/05/2019).
 - *Venice Time Machine - History and Big Data (7/8) - What Future For Time Machines ?*, en, URL : <https://www.arte.tv/en/videos/075631-007-A/venice-time-machine-history-and-big-data-7-8/> (visité le 18/05/2019).

- ARTE, *Venice Time Machine - History and Big Data (8/8) - Interpreting History*, en, URL : <https://www.arte.tv/en/videos/075631-008-A/venice-time-machine-history-and-big-data-8-8/> (visité le 18/05/2019).
- DIGITAL HUMANITIES ORGANIZATIONS (Alliance of), *About*, URL : <http://adho.org/about> (visité le 15/06/2019).
- DUBUC (Damien), “Venice Time Machine, un canal à remonter le temps” (, déc. 2017), URL : https://www.lemonde.fr/tant-de-temps/article/2017/12/13/venice-time-machine-un-canal-a-remonter-le-temps_5229068_4598196.html (visité le 18/05/2019).
- EPFL, *Chiffres clés*, fr-FR, URL : <https://www.epfl.ch/about/overview/fr/presentation/chiffres-cles/> (visité le 13/06/2019).
- *CROSS – Collaborative Research on Science and Society*, fr-FR, URL : <https://www.epfl.ch/schools/cdh/fr/recherche/cross/> (visité le 13/06/2019).
 - *EPFL Fribourg*, fr-FR, URL : <https://www.epfl.ch/about/campus/fr/epfl-fribourg/> (visité le 13/06/2019).
 - *EPFL Valais Wallis*, fr-FR, URL : <https://www.epfl.ch/about/campus/fr/valais-fr/> (visité le 13/06/2019).
 - *History of EPFL*, en-GB, URL : <https://www.epfl.ch/about/overview/overview/history-of-epfl/> (visité le 13/06/2019).
 - *Innovation Park*, en-US, URL : <https://www.epfl-innovationpark.ch/community/companies/> (visité le 13/06/2019).
 - *La vision du CDH : POLY-perspective*, fr-FR, URL : <https://www.epfl.ch/schools/cdh/fr/la-vision-du-cdh-poly-perspective/> (visité le 13/06/2019).
 - *Research data at EPFL*, en-GB, URL : <https://researchdata.epfl.ch/> (visité le 18/05/2019).
- EPFL.DHLAB, *2016 – 2020 ScanVan (FNS)*, en-GB, URL : <https://dhlab.epfl.ch/page-96350-en-html/page-152926-en-html/> (visité le 23/06/2019).
- *2017-2020 Impresso (FNS)*, en-GB, URL : <https://dhlab.epfl.ch/page-96350-en-html/page-150782-en-html/> (visité le 23/06/2019).
 - *DH Research*, URL : <https://www.epfl.ch/schools/cdh/research-2/dhi/dh-research/> (visité le 13/06/2019).
 - *Home / TimeMachineBox*, URL : <https://www.timemachinebox.eu/> (visité le 18/05/2019).
 - *Introduction — dhSegment documentation*, URL : <https://dhsegment.readthedocs.io/en/latest/intro/intro.html#use-cases> (visité le 20/06/2019).
 - *DH Seminar Lecture 2019 - Prof. Frédéric Kaplan*, EPFL, mars 2019, URL : <https://tube.switch.ch/videos/9410934a> (visité le 18/05/2019).
- EUROPEANA, *Europeana Collections*, en, URL : <https://www.europeana.eu/portal/?locale=en> (visité le 23/06/2019).

- Formations offertes au CDH, fr-FR, URL : <https://www.epfl.ch/schools/cdh/fr/formations-offertes/>* (visité le 15/06/2019).
- GMBH (Klokan Technologies), *Time Machine Atlas*, URL : <https://timemachineatlas.eu> (visité le 18/05/2019).
- HUNGARICANA, *Budapest Time Machine*, URL : <https://hungaricana.hu/en/budapest-idogep/> (visité le 23/06/2019).
- ICONEM, *Iconem*, URL : <http://iconem.com/en/> (visité le 23/06/2019).
- INSTITUT NATIONAL D'HISTOIRE DE L'ART, *#LundisNum / 8 janvier 2018 - Isabella Di Lenardo et Benoit Seguin : projet Replica*, URL : <https://www.youtube.com/watch?v=JxFMEAokjTM> (visité le 18/05/2019).
- KAPLAN (Frédéric), *La cartographie en quatre dimensions Propos recueillis par Marc Frochaux*, fr, mai 2018, URL : <https://www.espazium.ch/fr/actualites/la-cartographie-en-quatre-dimensions> (visité le 18/05/2019).
- KAPLAN (Frédéric) et LENARDO (Isabella di), “Big Data of the Past”, *Frontiers in Digital Humanities*, 4 (mai 2017), DOI : 10.3389/fdigh.2017.00012.
- “L’Europe doit construire la première Time Machine” (, déc. 2016), URL : <https://www.letemps.ch/opinions/leurope-construire-premiere-time-machine> (visité le 18/05/2019).
- La crise révélée par la Time Machine*, fr, URL : <https://www.bilan.ch/opinions/fabrice-delaye/la-crise-revelee-par-la-time-machine> (visité le 22/06/2019).
- “Les villes d’Europe prêtes à remonter le temps” (, mars 2018), URL : https://www.lemonde.fr/sciences/article/2018/03/06/les-villes-d-europe-pretes-a-remonter-le-temps_5266443_1650684.html (visité le 18/05/2019).
- MORISCL, *FET Flagships*, en, Text, déc. 2013, URL : <https://ec.europa.eu/programmes/horizon2020/en/h2020-section/fet-flagships> (visité le 22/06/2019).
- NATURE VIDEO, *A virtual time machine for Venice*, URL : https://www.youtube.com/watch?time_continue=147&v=uQQGgYPRWfs (visité le 18/05/2019).
- Recognition and Enrichment of Archival Documents / Projects / H2020 / CORDIS*, URL : <https://cordis.europa.eu/project/rcn/198756/factsheet/en> (visité le 17/06/2019).
- Reconnaissance d’entités nommées*, fr, Page Version ID : 144628341, janv. 2018, URL : https://fr.wikipedia.org/w/index.php?title=Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es&oldid=144628341 (visité le 23/06/2019).
- RTS, *L’invité de La Matinale - Frédéric Kaplan, professeur en humanités digitales à l’EPFL - Radio*, fr, URL : <https://www.rts.ch/play/radio/linvite-e-de-la-matinale/audio/linvite-de-la-matinale-frederic-kaplan-professeur-en-humanites-digitales-a-lepfl?id=10247166> (visité le 18/05/2019).

RTS, *Remonter le temps à Venise - Vidéo*, fr, URL : <https://www.rts.ch/play/tv/mise-au-point/video/remonter-le-temps-a-venise?id=9020497> (visité le 18/05/2019).

- *Coup de frein de l'UE au projet de recherche "Time Machine" de l'EPFL*, fr, info-Sport, mai 2019, URL : <https://www.rts.ch/info/sciences-tech/technologies/10435599-coup-de-frein-de-l-ue-au-projet-de-recherche-time-machine-de-l-epfl.html> (visité le 22/06/2019).

S.L (Factum Arte), *Factum Arte*, en, URL : <http://www.factum-arte.com/aboutus> (visité le 20/06/2019).

TED, *Frederic Kaplan : How I built an information time machine*, URL : https://www.youtube.com/watch?time_continue=296&v=2-Ev4rU27HY (visité le 18/05/2019).

TIME MACHINE, *Diamond*, URL : <https://diamond.timemachine.eu/> (visité le 18/05/2019).

- *Introduction*, en, URL : <https://timemachine.eu/> (visité le 18/05/2019).
- *Local Time Machines / Time Machine Europe*, URL : <https://timemachine.eu/time-machines/> (visité le 23/06/2019).

WORLD.MINDS, *Frédéric Kaplan : The Venice Time Machine (2017 WORLD.MINDS Annual Symposium)*, URL : https://www.youtube.com/watch?time_continue=666&v=6brInBZ-jLk (visité le 18/05/2019).

Exemples de projets de numérisation

CLAVERT (Frédéric) et NOIRET (Serge), *L'histoire contemporaine à l'ère numérique Contemporary history in the digital age*, OCLC : 894315404, 2013, URL : <http://site.ebrary.com/id/10819452> (visité le 25/06/2019).

DIGITAL PUBLIC LIBRARY OF AMERICA, *Becoming a Service Hub*, en, URL : <https://pro.dp.la/prospective-hubs/becoming-a-service-hub> (visité le 25/07/2019).

- *History*, en, URL : <https://pro.dp.la/about-dpla-pro/history> (visité le 25/07/2019).
- *Membership Program*, en, URL : <https://pro.dp.la/hubs/membership-program> (visité le 25/07/2019).
- *Metadata Application Profile*, en, URL : <https://pro.dp.la/hubs/metadata-application-profile> (visité le 25/07/2019).
- *Our Supporters*, en, URL : <https://pro.dp.la/about-dpla-pro/our-supporters> (visité le 25/07/2019).
- *Our Values*, en, URL : <https://pro.dp.la/about-dpla-pro/our-values> (visité le 25/07/2019).
- *Projects / DPLA*, URL : <https://pro.dp.la/projects> (visité le 25/07/2019).
- *Strategic Plan*, en, URL : <https://pro.dp.la/about-dpla-pro/strategic-plan> (visité le 25/07/2019).
- *Understanding Copyright*, en, URL : <https://pro.dp.la/projects/understanding-copyright> (visité le 25/07/2019).

EUROPEANA, *Breathing new life into the Europeana Aggregator Forum*, en-GB, URL : <https://pro.europeana.eu/post/breathing-new-life-into-the-europeana-aggregators-forum> (visité le 24/07/2019).

- *Europeana Collections*, URL : <https://www.europeana.eu/portal/en> (visité le 24/07/2019).
- *Europeana Publishing Guide*, en-GB, URL : <https://pro.europeana.eu/post/publication-policy> (visité le 04/07/2019).
- *Everything you need to publish in Europeana*, en-GB, URL : <https://pro.europeana.eu/services/data-publication-services/existing-provider> (visité le 24/07/2019).
- *Linked Open Data*, en-GB, URL : <https://pro.europeana.eu/page/linked-open-data> (visité le 24/07/2019).

- Europeana Bot (@EuropeanaBot) / Twitter*, en, URL : <https://twitter.com/europeanabot> (visité le 24/07/2019).
- GOOGLE, *About Google Books – Google Books*, URL : <https://www.google.com/googlebooks/about/> (visité le 22/07/2019).
- HARPER (Sarah Fletcher), “Google Books Review”, *Journal of Electronic Resources in Medical Libraries*, 13–1 (janv. 2016), p. 2-7, DOI : 10.1080/15424065.2016.1142835.
- HATHI TRUST DIGITAL LIBRARY, *Charting HathiTrust’s Strategic Directions* / www.hathitrust.org/charting-hathitrusts-strategic-directions (visité le 23/07/2019).
- *HTRC Analytics*, URL : <https://analytics.hathitrust.org/> (visité le 23/07/2019).
 - *Zephir* / www.hathitrust.org/zephir, URL : <https://www.hathitrust.org/zephir> (visité le 23/07/2019).
- HOFFMANN (Anna Lauren), “Google Books, Libraries, and Self-Respect : Information Justice beyond Distributions”, *The Library Quarterly*, 86–1 (janv. 2016), p. 76-92, DOI : 10.1086/684141.
- IOANA ROIU, CRISTINA, “From Family Memories to Digital Archives – Europeana and the Educational Value of its Digital Collections” (), p. 137-143.
- LAU-SUCHET (Soline), *HathiTrust : une bibliothèque numérique éléphantesque...* fr-FR, Billet, févr. 2014, URL : <https://bulac.hypotheses.org/967> (visité le 23/07/2019).
- LEETARU (Kalev), “Mass book digitization : The deeper story of Google Books and the Open Content Alliance”, *First Monday*, 13 (oct. 2008), DOI : 10.5210/fm.v13i10.2101.
- O’BRIEN (Kevin), “Large-Scale Book and Journal Digitization Projects and Interlibrary Service : Opening the Discussion”, *Journal of Interlibrary Loan, Document Delivery & Electronic Reserve*, 25–1-2 (mars 2015), p. 39-42, DOI : 10.1080/1072303X.2016.1150380.
- RightsStatements.org*, URL : <https://rightsstatements.org/fr/> (visité le 24/07/2019).
- Ten years of Europeana : bringing Europe’s cultural heritage into the digital age*, en, Text, sept. 2018, URL : <https://ec.europa.eu/digital-single-market/en/news/ten-years-europeana-bringing-europes-cultural-heritage-digital-age> (visité le 04/07/2019).
- THELLE (Mikkel) et BONDE THYLSTRUP (Nanna), “Persuasive territories in European cultural politics : critical and controlled knowledgescapes”, *Library Hi Tech*, 29–4 (nov. 2011), p. 573-585, DOI : 10.1108/07378831111189705.
- Viewpoint - What’s Next for Google Books*, URL : <http://www.infotoday.com/IT/sep16/Quint--Whats-Next-for-Google-Books.shtml> (visité le 03/07/2019).

WEISS (Andrew), “Examining Massive Digital Libraries (MDLs) and Their Impact on Reference Services”, *The Reference Librarian*, 57–4 (oct. 2016), p. 286-306, DOI : 10.1080/02763877.2016.1145614.

WEISS (Andrew) et JAMES (Ryan), “Assessing the coverage of Hawaiian and Pacific books in the Google Books digitization project”, *OCLC Systems & Services : International digital library perspectives*, 29–1 (févr. 2013), p. 13-21, DOI : 10.1108/10650751311294519.

WILLEMS (Marieke) et ATANASSOVA (Rossitza), “Europeana Newspapers : searching digitized historical newspapers from 23 European countries”, *Insights the UKSG journal*, 28–1 (mars 2015), p. 51-56, DOI : 10.1629/uksg.218.

Glossaire

Apprentissage automatique, fr, Page Version ID : 160226362, juin 2019, URL : https://fr.wikipedia.org/w/index.php?title=Apprentissage_automatique&oldid=160226362 (visité le 23/06/2019).

Artificial neural network, en, Page Version ID : 901207463, juin 2019, URL : https://en.wikipedia.org/w/index.php?title=Artificial_neural_network&oldid=901207463 (visité le 23/06/2019).

Base de données, fr, Page Version ID : 160423450, juin 2019, URL : https://fr.wikipedia.org/w/index.php?title=Base_de_donn%C3%A9es&oldid=160423450 (visité le 18/07/2019).

Big data, fr, Page Version ID : 159931315, juin 2019, URL : https://fr.wikipedia.org/w/index.php?title=Big_data&oldid=159931315 (visité le 23/06/2019).

Citizen science, en, Page Version ID : 906579059, juil. 2019, URL : https://en.wikipedia.org/w/index.php?title=Citizen_science&oldid=906579059 (visité le 24/07/2019).

Creative Commons, en, Page Version ID : 906058802, juil. 2019, URL : https://en.wikipedia.org/w/index.php?title=Creative_Commons&oldid=906058802 (visité le 24/07/2019).

Curation de contenu, fr, Page Version ID : 160878957, juil. 2019, URL : https://fr.wikipedia.org/w/index.php?title=Curation_de_contenu&oldid=160878957 (visité le 30/07/2019).

Deep learning, en, Page Version ID : 902979782, juin 2019, URL : https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=902979782 (visité le 23/06/2019).

Inference engine, en, Page Version ID : 893894240, avr. 2019, URL : https://en.wikipedia.org/w/index.php?title=Inference_engine&oldid=893894240 (visité le 23/06/2019).

Intelligence artificielle, fr, Page Version ID : 160343358, juin 2019, URL : https://fr.wikipedia.org/w/index.php?title=Intelligence_artificielle&oldid=160343358 (visité le 23/06/2019).

Knowledge Graph, fr, Page Version ID : 138894172, juil. 2017, URL : https://fr.wikipedia.org/w/index.php?title=Knowledge_Graph&oldid=138894172 (visité le 23/06/2019).

Libre accès (édition scientifique), fr, Page Version ID : 160593388, juil. 2019, URL : [https://fr.wikipedia.org/w/index.php?title=Libre_acc%C3%A8s_\(%C3%A9dition_scientifique\)&oldid=160593388](https://fr.wikipedia.org/w/index.php?title=Libre_acc%C3%A8s_(%C3%A9dition_scientifique)&oldid=160593388) (visité le 09/07/2019).

Nuage de points (géométrie), fr, Page Version ID : 155223201, déc. 2018, URL : [https://fr.wikipedia.org/w/index.php?title=Nuage_de_points_\(g%C3%A9om%C3%A9trie\)&oldid=155223201](https://fr.wikipedia.org/w/index.php?title=Nuage_de_points_(g%C3%A9om%C3%A9trie)&oldid=155223201) (visité le 23/06/2019).

Numérisation, URL : <https://fr.wikipedia.org/wiki/Num%C3%A9risation> (visité le 23/06/2019).

Open science, en, Page Version ID : 900178688, juin 2019, URL : https://en.wikipedia.org/w/index.php?title=Open_science&oldid=900178688 (visité le 23/06/2019).

Photogrammétrie, fr, Page Version ID : 159009276, mai 2019, URL : <https://fr.wikipedia.org/w/index.php?title=Photogramm%C3%A9trie&oldid=159009276> (visité le 23/06/2019).

Reconnaissance optique de caractères, fr, Page Version ID : 139202141, juil. 2017, URL : https://fr.wikipedia.org/w/index.php?title=Reconnaissance_optique_de_caract%C3%A8res&oldid=139202141 (visité le 23/06/2019).

The CDH's vision : POLY-perspective – EPFL, en-GB, URL : <https://www.epfl.ch/schools/cdh/cdhs-vision/> (visité le 23/06/2019).

Traitement automatique du langage naturel, fr, Page Version ID : 159047154, mai 2019, URL : https://fr.wikipedia.org/w/index.php?title=Traitement_automatique_du_langage_naturel&oldid=159047154 (visité le 23/06/2019).

Acronymes

API *Application Programming Interface.* 46, 47, 89, 95, 118, 125

ASCII *American Standard for Information Interchange.* 12

BNF *Bibliothèque Nationale de France.* 18, 19

CSA *Coordination and Support Actions.* 105–107, 115

DH *Digital Humanities*, humanités digitales, humanités numériques. 56, 57

DHLAB *Digital Humanities Laboratory* ou Laboratoire d’humanités digitales. 3, 55, 59, 60, 66, 67, 69, 105, 109, 111, 123, 157

DPLA *the Digital Public Library of America.* 5, 77, 89, 94–96, 99–101, 116, 118, 157

EPFL École Polytechnique fédérale de Lausanne. 4, 55, 58–61, 63, 64, 66, 105, 121, 157

GLAM *Galleries, Libraries, Archives, Museums* ou Galeries, Bibliothèques, Archives, Musées. 17, 35, 71, 123, 129, 135

HTTP HyperText Transfer Protocol. 45

IIIF International Image Interoperability Framework. 47, 89, 118

OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting. 46, 89, 118, 125

RDF Resource Description Framework. 45, 46, 125

RFC Request for Comments. 116

UE Union Européenne. 21–27, 37, 38, 41, 51, 61, 64, 87, 92, 105, 107, 108, 111, 112, 117

UNESCO *The United Nations Educational, Scientific and Cultural Organization* ou Organisation des Nations Unies pour l’éducation, la science et la culture. 21, 22

URI Uniform Resource Identifier. 45

W3C *World Wide Web Consortium.* 44

WP *Work Package*, lot de travail. 105–107, 111, 112

Introduction

Les projets de numérisation de masse semblent issus des conséquences de la révolution numérique, ils sont portés en Europe par le constat effrayant d'un retard face aux entreprises et à la croissance américaines.

Pour rappel, en 2006, le monde comptait 50 sociétés de taille mondiale, dont 17 étaient européennes. Dix ans plus tard, 31 sont américaines et 8 sont chinoises. En 2017 Nestlé, le géant de l'alimentaire demeure seul représentant de l'industrie européenne¹.

Bénéficiant de soutiens politiques, les projets de numérisation s'inscrivent également dans l'histoire du développement des industries culturelles et patrimoniales et semblent incarner l'évolution logique des bibliothèques numériques. Perçus par les uns comme une réponse à la faible vélocité du marché économique européen, et un moyen incontournable permettant de garantir l'accessibilité au savoir dans une société aux pratiques de plus en plus connectées pour les autres², que sont réellement ces projets de numérisation de masse et que nous apprennent-ils sur l'évolution de notre société ?

Symptôme de la révolution numérique, l'information, la connaissance et le savoir créent de la valeur économique et sont de plus en plus traités comme « un bien immatériel appropriable »³. Ce constat voit s'affronter logiques du droit d'auteur et des brevets, accusés de privatiser le monde, et tentatives de définition des « biens publics communs » visant à préserver l'accessibilité d'un certain savoir à tous.

Le concept de patrimoine culturel européen date seulement des années 1950 et n'a été ancré dans l'agenda européen qu'à partir des années 1990. Présenté par la Commission, à la fois comme vecteur de l'identité culturelle et européenne, le patrimoine culturel est un élément important de la mémoire de l'Europe⁴. Or dans un contexte de mondialisation, au-delà de la question de la préservation du savoir, se pose le besoin de préserver cette diversité culturelle. Comment protéger dans cette course à la grandeur numérique, les spécificités intrinsèques au bien-être humain ? L'idée d'un universalisme culturel est-elle seulement envisageable⁵ ?

Alors que nous sommes sur le point de disposer des technologies rendant possible la réalisation du rêve de la « Cité mondiale » de Paul Otlet⁶, les acteurs actuels du marché de la numérisation de masse suscitent craintes, critiques et admiration.

Le projet Time Machine, pour lequel nous avons effectué notre stage du 15 avril au 31 août 2019 au sein du *Digital Humanities Laboratory* ou Laboratoire d'humanités

1. Daniel Battu, *L'histoire et l'économie du monde accompagnées par les TIC*, OCLC : 1030367775, London, 2018.

2. Michelle Wu, “Building a Collaborative Digital Collection : A Necessary Evolution in Libraries”, *Georgetown Law Faculty Publications and Other Works*. 699 (, 2011).

3. Armand Mattelart, *Histoire de la société de l'information*, OCLC : 1082200705, 2018, p.105.

4. Mikkel Thelle et Nanna Bonde Thylstrup, “Persuasive territories in European cultural politics : critical and controlled knowledgescapes”, *Library Hi Tech*, 29–4 (nov. 2011), p. 573-585, DOI : 10.1108/0737883111189705.

5. Anne-Marie Laulan, “Diversité culturelle et mondialisation”, *Hermès, La Revue*, 80–1 (2018), p. 168-174, URL : <https://www.cairn.info/revue-hermes-la-revue-2018-1-page-168.htm>.

6. Nanna Bonde Thylstrup, *The politics of mass digitization*, Cambridge, MA, 2018.

digitales (DHLAB) de l’École Polytechnique fédérale de Lausanne (EPFL), s’inscrit dans la continuité de ces débats idéologiques, ses objectifs trouvant une troublante incarnation dans l’idée récemment décrite du *mirrorworld*⁷ :

[Traduction] Nous sommes à l’aube de la création d’une nouvelle plate-forme, qui numérisera le monde. Sur cette plateforme, lieux et choses seront lisibles par les machines, sujets au pouvoir des algorithmes. Quiconque dominera cette grande plateforme s’inscrira parmi les plus riches et puissantes puissances de l’histoire. [...] L’histoire deviendra un verbe. D’un geste de la main, vous pourrez voyager dans le temps [...] ou dans le futur. Ces différents scénarios auront le goût de la réalité, car ils seront dérivés d’une reproduction à l’échelle de notre monde actuel. En ce sens, peut-être faut-il plus parler d’un monde en quatre dimensions que d’un miroir.⁸.

Sans prétendre apporter une réponse à tous les enjeux techniques, culturels, politiques, légaux que soulèvent les entreprises de numérisation de masse, nous proposons dans ce mémoire quelques clés pour mieux en comprendre les origines et la complexité afin d’interroger leur positionnement face aux autres acteurs de l’information. Car si la vision du *mirrorworld* semble résumer les motivations d’un projet tel que Time Machine, de nombreux débats et évolutions doivent encore être menés, qui influenceront fortement la future forme de ce double numérique et viendront bouleverser l’organisation de notre monde réel.

Notre mémoire est structuré en trois parties, visant à la fois à rendre compte du travail réalisé durant notre stage et à apporter un cadre théorique et réflexif pour comprendre le contexte dans lequel nous avons été amenée à élaborer certaines solutions. Dans la première partie, nous vous proposons de découvrir l’histoire des projets de numérisation de masse, ou comment les premières initiatives ont évolué vers des projets de grandes envergures. Ces derniers sont issus de révolutions technologiques, mais découlent également de l’évolution des pratiques bibliothéconomiques induites par ces mêmes révolutions. Leurs grandes portées les inscrit de plus dans une histoire économique et par conséquent politique, dont nous vous résumerons les étapes. Dans un deuxième temps, nous présenterons les caractéristiques de ces initiatives et les différents enjeux (*amener différents acteurs à*

7. Si le terme bénéficie d’une nouvelle popularité, il a été utilisé pour la première fois par l’informaticien de Yale, David Gelernter en 1991, dans son livre « *Mirror Worlds : Or the Day Software Puts the Universe in a Shoebox...How It Will Happen and What It Will Mean* » (Oxford University Press, 1991)

8. « We are now at the dawn of the third platform, which will digitize the rest of the world. On this platform, all things and places will be machine-readable, subject to the power of algorithms. Whoever dominates this grand third platform will become among the wealthiest and most powerful people and companies in history [...]. History will be a verb. With a swipe of your hand, you will be able to go back in time, at any location, and see what came before. [...] Or you’ll scroll in the other direction : forward. [...] These scroll-forward scenario will have the heft of reality because they will be derived from a full-scale present world. In this way, the mirrorworld may be best referred to as a 4D world. » Kevin Kelly, “AR Will Spark the Next Big Tech Platform—Call It Mirrorworld”, *Wired* (, févr. 2019), URL : <https://www.wired.com/story/mirrorworld-ar-next-big-tech-platform/> (visité le 03/07/2019)

collaborer, financement et partenariats public-privé, droit d'auteur, sortir des silos ou la quête de l'interopérabilité, stockage sur le long-terme et préservation) qui en découlent. Ceci afin de mieux appréhender les questions soulevées par le *comment* de la numérisation.

Notre projet de recherche étant motivé par notre expérience de stagiaire, nous poserons le contexte et la description du projet Time Machine dans la dernière section de cette première partie, abordant le contexte institutionnel, l'histoire de la recherche en humanités numériques et le lien entre ce secteur académique et les projets de numérisation. Les précédents développements menés par le laboratoire en vue de l'élaboration de Time Machine seront également rappelés (dont notamment une présentation de *Venice Time Machine*). Time Machine étant à l'heure actuelle un consortium réunissant quelque centaines d'institutions culturelles et patrimoniales sous la coordination du laboratoire lausannois, nous présenterons plus en détails l'organisation et les objectifs de cette initiative.

Le deuxième temps de notre mémoire, offre un regard plus précis sur quatre initiatives de numérisation de masse emblématiques du 21^e siècle, *Google Books*, Europeana, *HathiTrust* et *the Digital Public Library of America (DPLA)*. Nous proposerons une brève analyse des différentes réponses apportées par ces initiatives aux enjeux de la numérisation préalablement identifiés et conclurons sur la question des divergences et similitudes entre ces différents projets.

Nous consacrerons la troisième partie à notre expérience de stagiaire, dévouée à l'élaboration d'une feuille de route pour les opérations et l'infrastructure de Time Machine sur une échelle de dix ans, et sur les propositions contenues dans ladite feuille de route, qui nous ont permis de nous confronter à chacun des enjeux présentés. Nous présenterons notre travail et les réponses apportées aux questions soulevées par ces enjeux dans un premier temps. Puis nous analyserons les différentes innovations apportées par Time Machine par rapport aux précédentes initiatives, et les risques et opportunités auxquels il devra faire face. Nous conclurons par un élargissement sur les motivations et impacts soulevés par l'envergure d'un tel projet : le *pourquoi* des entreprises de numérisation de masse, et nous tenterons de définir si Time Machine s'inscrit dès lors dans la continuité des entreprises de numérisation portées par les acteurs culturels et patrimoniaux, ou s'établit en tant que nouvel acteur de l'information.

Première partie

De la numérisation à la numérisation de masse

Chapitre 1

Historique de la numérisation

Les projets de numérisation de masse ne sont pas soudainement apparus avec le tournant des années 2000, ils sont le fruit d'une multitude de développements technologiques et des transformations sociétales qui en découlent, initiés dès le 19^e siècle. Ce premier chapitre vise à replacer ces projets au sein d'un contexte historique, entre l'émergence du web et la formation des premières bibliothèques numériques, sans oublier les débats politiques associés à ces projets culturels.

Au même titre que les projets actuels combinent différents enjeux et acteurs, le développement de ces entreprises de numérisation sont au croisement des sciences de l'information et de l'informatique, et puisent leurs origines dans les deux domaines. Ce sont les progrès menés dans l'un et l'autre qui permettront la naissance des bibliothèques numériques dans les années 90, puis le passage à l'échelle des projets de numérisation. Nous verrons que l'idée des premières collections numériques existait bien avant, mais ces collections n'étaient pas identifiées en tant que tel¹.

Time Machine n'est pas le premier projet qui vise à mettre les données du passé à disposition du plus grand nombre, d'autres projets se sont construits autour d'objectifs similaires, avant même l'avènement du numérique.

Afin de ne pas alourdir plus que nécessaire notre mémoire, nous avons choisi de présenter une sélection des personnalités et des projets qui nous ont semblé les plus emblématiques et les plus à même d'illustrer les différentes évolutions de ces précurseurs des initiatives de numérisation.

1. Iris Xie et Krystyna K. Matusiak, *Discover digital libraries : theory and practice*, OCLC : 907120360, Amsterdam Boston Heidelberg, 2016, p. 8

1.1 1800-1990 : Prémisses des projets de numérisation

Préalablement à l'émergence des bibliothèques numériques et aux projets de numérisation de masse, et bien que leurs origines soient quelque peu incertaines à retracer, des scientifiques et penseurs ont contribué par leurs idées et outils, à poser les jalons de ce que deviendront les grandes entreprises de numérisation. Au-delà de la naissance des éléments théoriques propres aux bibliothèques numériques, les développements du matériel informatique, de l'hypertexte, de l'internet et du web ont permis la naissance des infrastructures techniques nécessaires à la concrétisation de ces idées. Sans vouloir en dresser une liste exhaustive, il est intéressant de montrer que l'histoire de la numérisation à grande échelle est basée sur les mêmes fondements historiques qui ont vu la naissance du web que nous connaissons aujourd'hui. Une frise temporelle, reprenant les éléments propres à l'histoire de la numérisation et élargissant sur l'histoire du web est détaillée en annexe A.

L'invention du microfilm, breveté par le français René Dagron en 1859, ouvre pour la première fois la perspective d'un nouveau support - substitut au livre papier pour l'accumulation et la diffusion du savoir. Dans son livre *Sur une forme nouvelle du livre : le livre microphotographique* paru en 1906, le belge Paul Otlet² suggère que les plus importantes transformations ne prendront pas place dans le livre lui-même mais dans son substitut³.

Conjointement avec Henri La Fontaine⁴, Paul Otlet initie dès 1895 le projet du *Mundaneum*. Décrit par ce dernier comme « [...] an Idea, an Institution, a Method, a Body of work materials and collections, a Building, a Network »⁵, ce projet de centre de documentation universel ne verra jamais la réalisation de son objectif, mais permettra la constitution d'une archive de quelques 12 millions de documents divers, et la naissance d'un musée. Reconnu par Google comme son ancêtre papier, l'entreprise américaine a même signé en 2013 un partenariat avec le musée subsistant⁶.

2. Paul Otlet (1868-1944), est un bibliographe, créateur du système de classification décimale universelle (CDU) et considéré comme un pionnier du web et des moteurs de recherche.

3. N. B. Thylstrup, *The politics of mass digitization...*, p. 7

4. Henri La Fontaine (1854-1943), homme politique et pacifiste belge.

5. *Ibid.*, p. 8

6. *Ibid.*



FIGURE 1.1 – Capture-d'écran du site internet du *Mundaneum*.

Diverses bibliothèques procèdent parallèlement au *microfilming* ou numérisation de leurs collections. En 1927, la Librairie du Congrès (États-Unis) procède au microfilmage de quelques trois millions de pages de livres et manuscrits de la *British Library* (Angleterre). Les microfilms générés durant cette grande période de réhabilitation des documents sont encore utilisés aujourd’hui⁷.

Vannevar Bush⁸ décrit son invention le *Memex*, convaincu que les méthodes traditionnelles d’indexation ne suffisent pas à répondre aux besoins des scientifiques modernes⁹, dans un fameux article publié en 1945, « *As We May Think* ». Il définit sans le savoir ce qui constituera nos futurs ordinateurs personnels.

[Traduction] Il s’agit d’un appareil dans lequel un individu peut stocker tous ses livres, fichiers et communications. Cet appareil est mécanisé de façon à pouvoir être consulté avec flexibilité et rapidité. Il est semblable à un élargissement, dans un format réduit, de la mémoire de l’individu.¹⁰

L’américain Richard Feynman¹¹ conçoit dès 1959 les perspectives offertes par les nouvelles technologies et leurs potentiels impacts sur les pratiques des bibliothécaires et autres spécialistes de l’information. Sa conférence *There's plenty of room at the bottom* donnée le 29 décembre 1959, lors de la réunion annuelle de l'*American Physical Society (California Institute of Technology)* propose de comprimer l’intégralité de l’encyclopédie britannique afin de la réduire à la taille d’une tête d’épingle. Il ira même plus loin en concluant ses propos par la proposition de réduire l’intégralité de la connaissance humaine

7. *Ibid.*, p. 9

8. Vannevar Bush (1890-1974), ingénieur et inventeur américain

9. I. Xie et K. K. Matusiak, *Discover digital libraries...*, p. 11

10. « [It is] a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory » Andrew Weiss, *Using massive digital libraries : a LITA guide*, First edition, Chicago, 2014 (LITA guides), p. 6.

11. Richard Feynman (1918-1988), physicien et théoricien américain, précurseur en nanotechnologies.

suivant la même méthode. Il imagine pour cela un système de conversion des caractères textuels en traits et points. Nous sommes un an avant le début des travaux sur l'*American Standard for Information Interchange* (ASCII)^{12 13}.

Dans son ouvrage *Libraries of the Future*, paru en 1965, J.C.R. Licklider¹⁴ propose d'étendre le monde des bibliothèques à celui des ordinateurs. Au-delà de l'objet livre, présenter faits et idées en fonctions, classes d'informations et domaines de connaissances, « *[engineers] need to substitute for the book a device that will make it easy to transmit information without transporting material, and that will not only present information to people but also process it for them*¹⁵. » Il faudra encore plus d'une génération pour que le numérique soit pleinement appliqué à la gestion des livres et du contenu informationnel qu'ils contiennent.

Nous concluons cette section par le projet Gutenberg, lancé par Michel Hart¹⁶ en 1971. Convaincu que la valeur ajoutée des ordinateurs ne se trouve pas uniquement dans leur puissance de calcul, mais également dans leur capacité à stocker et à retrouver des informations, il parvient à rassembler une équipe de volontaires pour saisir au clavier des textes libres de droit dans un format accessible et compréhensible par les ordinateurs : une version basique de l'ASCII. Ces textes analogues ainsi convertis dans une forme numérique sont conservés sur un serveur en mode texte, et diffusés auprès des membres du réseau ARPAnet^{17 18}. L'objectif n'est alors pas la précision, mais une plus ou moins haute correspondance entre texte écrit et rendu numérique¹⁹. Ce projet peut être perçu comme le premier projet de bibliothèque numérique, alors que l'internet et le web que nous connaissons aujourd'hui ne seront mis en place que dans les années 90.

12. Norme étasunienne qui sert comme premier système d'encodage informatique des caractères en anglais et apportera un premier niveau de standardisation. Les caractères utilisés en anglais sont encodés suivant une combinaison de 0 et de 1.

13. Association pour le patrimoine naturel et culturel du canton de Vaud, *Patrimoine numérique, numérisation du patrimoine*, Lausanne, 2012, p. 25

14. Joseph Carl Robnett Licklider (1915-1990), psychologue et informaticien américain.

15. A. Weiss, *Using massive digital libraries...*, p. 6

16. Michael Stern Hart (1947-2011), auteur américain et inventeur du livre électronique ou *e-book*.

17. *Advanced Research Projects Agency Network*, réseau qui a servi de modèle à notre actuel internet.

18. Association pour le patrimoine naturel et culturel du canton de Vaud, *Patrimoine numérique, numérisation du patrimoine...*, p. 10

19. N. B. Thylstrup, *The politics of mass digitization...*, p. 10



FIGURE 1.2 – Capture-d'écran du site internet du *Projet Gutenberg*.

Les principaux défis techniques permettant le passage d'un texte physique à une version numérique sont levés entre 1970 et 1990 (le coût des scanners devient abordable pour les bibliothèques, les formats des images digitales se normalisent, l'industrie des logiciels s'intéresse aux programmes de reconnaissance optique de caractères) ²⁰.

Désormais les écrans et les claviers servent d'interface entre hommes et ordinateurs, la puissance informatique a été augmentée et les premiers systèmes de recherche documentaires voient le jour ²¹. Pour plus de détails sur ces développements du matériel informatique, liés à l'avènement d'internet ²², référez-vous à l'annexe A.

20. Association pour le patrimoine naturel et culturel du canton de Vaud, *Patrimoine numérique, numérisation du patrimoine...*

21. A. Weiss, *Using massive digital libraries...*, p. 10

22. L'ensemble des supports de cours élaborés par Mr. Philippe Bootz, dans le cadre du cours "Sémantique du numérique", suivis durant ma première année de Master en *Humanités numériques : enjeux et technologies*, Université Paris 8, ont servi à l'élaboration de la synthèse des éléments techniques.

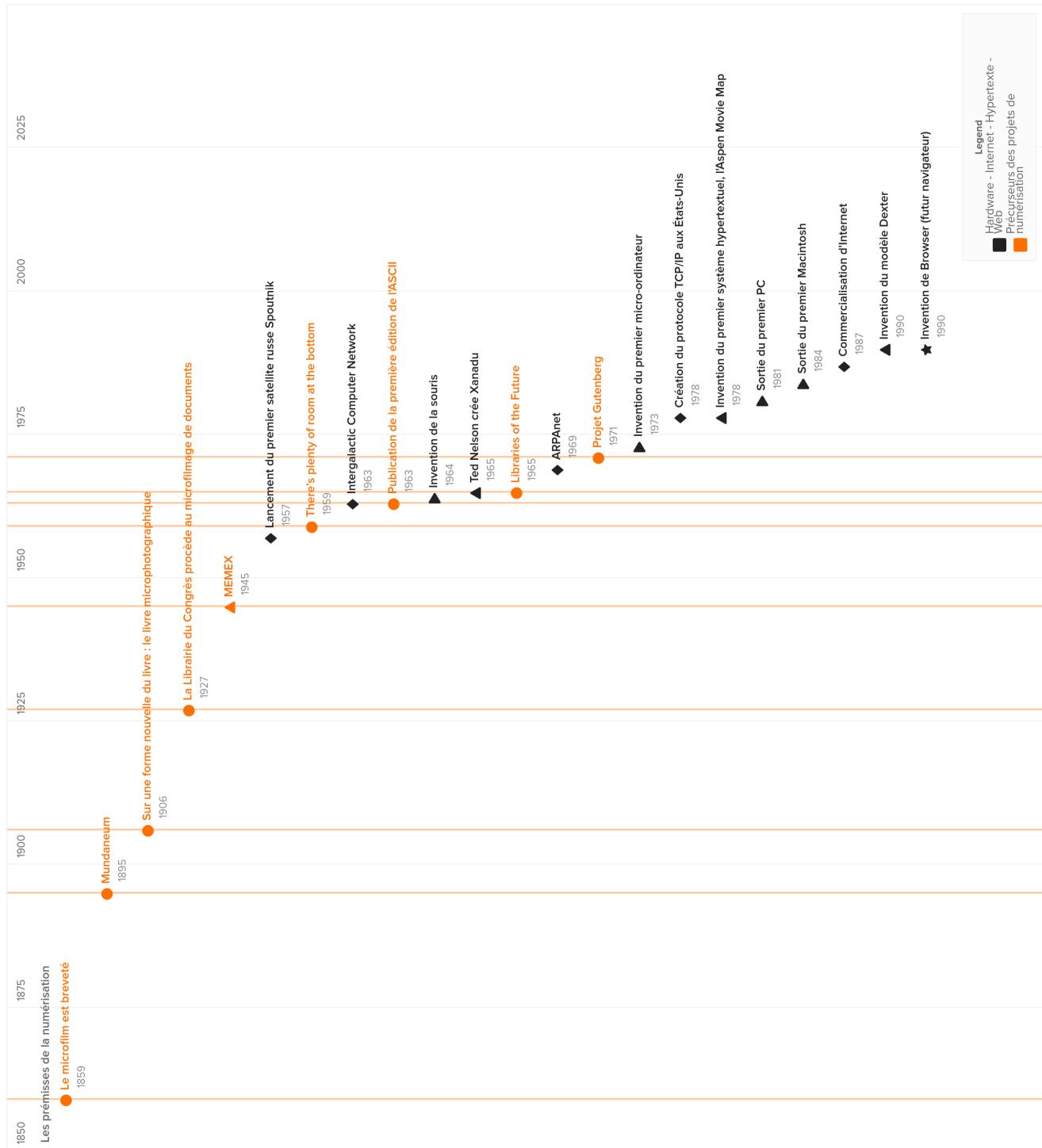


FIGURE 1.3 – Frise temporelle, *Les prémisses de la numérisation*.

1.2 1990-2000 : Passage à l'échelle

Bien que les bibliothèques publiques apparaissent dès le milieu du 19^e siècle²³, avec pour vocation de rendre la connaissance accessible à tous, ce n'est qu'au début des années 1990 et l'avènement des ordinateurs personnels, que le mouvement des bibliothèques numériques voit le jour. Au-delà des développements purement techniques, les professionnels réalisent que proposer des collections numériques signifie bien plus que l'acte de numérisation, et que le processus est étroitement lié à des enjeux et des controverses culturelles, politiques et légales. Mais comment est-on passé de la bibliothèque numérique, aux projets de numérisation de masse ?

Nous tenterons dans cette section de proposer une chronologie résumée des faits les plus marquants de cette révolution. Il est important de relever qu'il n'existe pas encore de définition exacte ni de catégorisation pour des projets d'une telle ampleur, et nous avons pu croiser, durant notre état de l'art, les expressions *very large digital libraries* ou *massive digital libraries*²⁴ au même titre que *large-scale digitization projects*^{25 26} ou *projets de numérisation de masse*^{27 28 29} pour désigner ces entreprises. Par souci de cohérence et puisque le cadre théorique fait encore débat, nous avons choisi de regrouper toutes les initiatives sous l'appellation de *projets de numérisation de masse*.

Avec la démocratisation de l'ordinateur personnel et la naissance du web, les premières bibliothèques numériques voient le jour. Différents termes servent à les désigner, *electronic library*, *virtual library*, *network-accessible libraries*, *libraries without walls*, reflétant l'évolution du sens de cette expression³⁰. D'abord perçues d'un point de vue technique et non comme institution ou objets porteurs d'influences sociales, « *In this sense, they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (text, image, sound; static or dynamic images) and exist in distributed networks*³¹ », les bibliothèques numériques garderont trace de ce premier focus technologique³².

Karen Calhoun les définit comme un système de services et de gestion des collections

23. Elisabeth Jones, “The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative : Assumptions, Intentions, and the Role of the Public”, *Information & Culture*, 52-2 (mai 2017), p. 229-263, DOI : 10.7560/IC52205, p.230

24. A. Weiss, *Using massive digital libraries...*

25. Cory Lampert, “Ramping up : Evaluating large-scale digitization potential with small-scale resources”, *Digital Library Perspectives*, 34-1 (févr. 2018), p. 45-59, DOI : 10.1108/DLP-06-2017-0020, p.47

26. Robin Yeates et Damon Guy, “Collaborative working for large digitisation projects”, *Program*, 40-2 (avr. 2006), p. 137-156, DOI : 10.1108/00330330610669262

27. N. B. Thylstrup, *The politics of mass digitization...*

28. C. Lampert, “Ramping up...”

29. I. Xie et K. K. Matusiak, *Discover digital libraries...*

30. *Ibid.*, p.3

31. E. Jones, “The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative...”, p.243

32. *Ibid.*, p.245

numériques, basé sur une architecture centrée sur les données³³.

Prenant conscience des potentialités en termes de services aux utilisateurs, promis par les interfaces numériques, les bibliothèques se mettent à réfléchir au développement de leurs activités, laissant les informaticiens développer les systèmes requis³⁴. Les premières entreprises de numérisation sont pensées dans la complémentarité de la mission d'accessibilité de la connaissance humaine, et motivées par les perspectives de préservation de documents rares ou trop fragiles, le rayonnement supplémentaire apporté aux collections et aux institutions par cet accès en ligne, et la possibilité de numériser différents matériaux (photographies, textes, manuscrits etc.)³⁵. Les projets de numérisation héritent toutefois des mêmes biais que les premières collections des bibliothèques publiques, dont les politiques documentaires visaient à éléver l'esprit des masses³⁶.

[Traduction] Les programmes de numérisation des bibliothèques ont contribué à la création d'un répertoire de collections spécialisées et de services dédiés à une audience similaire à celle de la bibliothèque physique. Puisque la plupart des bibliothèques engagées dans ces programmes de numérisation sont académiques, les collections sont destinées à un usage d'abord scientifique.³⁷

[Traduction] Les bibliothèques numériques et les bibliothèques publiques se sont aliénées de potentiels utilisateurs en se focalisant trop sur leurs idéaux. Les bibliothèques publiques en traitant de haut les travailleurs et ouvriers, les faisant se sentir indésirables et les bibliothèques numériques en développant des interfaces demandant une connaissance technique, culturelle ou thématique préalable à leur utilisation.³⁸

The Perseus Digital Library, est l'un des premiers projets de bibliothèque numérique. Initié en 1985 au sein de l'université Tufts (États-Unis), la version initiale regroupe des textes grecs et leurs traductions anglaises sur un CD-ROM. Le projet évolue vers une première plateforme en ligne en 1995, ouvrant ses collections au monde greco-romain et passant d'un outil d'enseignement à un moteur de recherche. Il rassemble des ressources

33. Karen Calhoun, *Exploring digital libraries. Foundations, practice, prospects*. OCLC : 894201348, Chicago, 2014

34. E. Jones, "The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative..."

35. Laurie Lopatin, "Library digitization projects, issues and guidelines", *Library Hi Tech* (, avr. 2006), DOI : 10.1108/07378830610669637, p.274

36. E. Jones, "The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative..."

37. « [...] library digitization programs have largely aimed to establish a repertoire of fairly specialized collections and services targeted at an audience that resembles the host libraries patron base - which, since most digitizing libraries are academics libraries, means designing for scholarly use » *Ibid.*, p.247.

38. « [...] both have alienated potential users by holding too strictly to this ideal, public libraries by making factory workers and other laborers feel patronized and unwelcome, and digital libraries by presenting interfaces that require a particular sort of technological, cultural, or domain-specific knowledge in order to use them effectively » *Ibid.*

liées aux études classiques (civilisations grecques et romaines) jusqu'à la renaissance anglaise³⁹.



FIGURE 1.4 – Capture-d'écran du visualisateur *Scaife* de la *Perseus Digital Library*.

Si ces premières initiatives de numérisation sont ciblées sur une certaine catégorie de collection et vers une audience particulière, c'est pour des raisons de financement. Les institutions culturelles et patrimoniales y voient une opportunité pour augmenter l'accessibilité de leurs collections et bénéficient dès les premiers temps de fonds publics⁴⁰. Les entreprises de numérisation articulent différents champs, nécessitant la création de nouveaux standards et bonnes pratiques, qui mèneront à l'élargissement de la portée de ces projets.

[Traduction] Il est apparu clairement que c'était un domaine multi-facettes, demandant une uniformisation des processus liés au contenu, aux technologies, à l'infrastructure, à la propriété intellectuelle et à la préservation. Les *Galleries, Libraries, Archives, Museums* ou Galeries, Bibliothèques, Archives, Musées (GLAM) ont très vite fait partie des participants les plus enthousiastes. En terme de contenu, les projets à petites échelles sont emblématiques des premières entreprises, se focalisant souvent sur ceux à haute valeur intellectuelle et culturelle. Cependant les activités de cette période visaient en parallèle à la mise en place de standards technologiques, au développement des infrastructures et au déploiement des processus de gestions, de préservation et de respect du droit d'auteur.⁴¹

39. Mark Dubis, "Web Resources for the Study of New Testament Backgrounds", *Journal of Religious & Theological Information*, 6-1 (janv. 2003), p. 3-9, DOI : 10.1300/J112v06n01_02

40. Margaret Coutts, *Stepping away from the silos : strategic collaboration in digitisation*, OCLC : ocn952385918, Cambridge, MA, 2017 (Chandos advances in information series), p.1

41. « It also became clear that it was a multifaceted field, requiring that same process to be applied

Les premières initiatives de numérisation de masse semblent également issues d'une forme de croyance répandue chez les informaticiens, déclarant que l'information contenue sur le papier deviendrait un jour obsolète, « [...] there was an attitude in computer science that putting things on dead trees was obsolete and getting it all into a searchable, digital format was a quest that had to be accomplished someday⁴². » En proposant des livres sous forme numérique, ces pionniers argumentent également qu'ils mettent leurs informations à disposition de tous, utilisant le même motif que pour la création des bibliothèques publiques et reproduisant à nouveau leurs biais⁴³.

[Traduction] Tout comme au 19^e siècle, il n'était pas suffisant de laisser le marché proposer des fictions de moindre qualité au peuple sans leur donner accès à des contenus jugés plus pertinents, au 21^e siècle, il n'est pas suffisant de laisser les chercheurs errer sur internet face à un océan de sites, où le meilleur de la connaissance - naturellement contenu dans les livres - n'est pas accessible.⁴⁴

Le premier projet de numérisation massive répertorié en France est un véritable précurseur et apparaît au tout début des années 90, il découle de la création de la *Bibliothèque Nationale de France* (BNF) en 1989. À la demande du Président François Mitterrand, Alain Giffard⁴⁵ est prié d'utiliser les dernières innovations technologiques pour rendre accessible les ouvrages du catalogue de la BNF et créer ainsi une très grande bibliothèque. Le projet conduira à la numérisation de 70'000 à 80'000 titres. Pour choisir quels ouvrages numériser, il n'eut pas recours à des bibliothécaires mais à des scientifiques et écrivains, ce qui fut possible car le projet était directement lié au Président. Alain Giffard explique d'ailleurs qu'il a personnellement acheté un grand nombre des livres à numériser, puisqu'ils n'étaient pas dans les collections de la BNF, et que celle-ci avait du mal à collaborer avec ce projet. Cet exemple illustre bien le fait que les projets de numérisation massive,

to issues of content, technology, infrastructure, intellectual property and sustainability. Universities, museums, galleries and national libraries were amongst the enthusiastic participants. In term of contents, small-scale projects typified the early work, often showcasing items of major intellectual and cultural value. The activity in this period, however, was as much concentrated on developing experience in and standards for the use of the technology and the provision of infrastructure, legal management and preservation » Ibid., p.12.

42. Michael Cook, *Google's Moon Shot : The quest for the universal library*, en-US, URL : <http://www.gutenbergnews.org/20070131/googles-moon-shot-the-quest-for-the-universal-library/> (visité le 04/07/2019)

43. N. B. Thylstrup, *The politics of mass digitization...*, p.8

44. « Just as in the nineteenth century, it was not seen as sufficient to let the market provide cheap paperback fiction for the masses without giving them access to higher quality materials, in the twenty-first century it is not enough to leave Internet searchers to their own devices in a sea of websites where the best knowledge - that contained in books - is nowhere to be found » E. Jones, “The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative...”, p.254.

45. Alain Giffard est directeur du groupement d'intérêt scientifique « Culture-médias & numérique. » Il a été directeur informatique de la Bibliothèque de France, directeur adjoint de l'Institut Mémoires de l'édition contemporaine (Imec), conseiller technique de la ministre de la Culture et de la Communication pour la société de l'information et président de la mission interministérielle pour l'accès public à l'internet.

au-delà des enjeux techniques, sont aussi des questions politiques qui soulèvent des débats de territorialité (institutionnelle ou nationale), matérialité et culture⁴⁶.

Ces documents numérisés seront intégrés en 1997 au projet plus connu de numérisation massive, *Gallica*. Ce projet, également mandaté par le Président François Mitterrand, vise à numériser les livres libres de droit issus des collections de la BNF, depuis le Moyen-Âge jusque vers 1930, avec une priorité pour les documents illustrant la culture francophone. Environ quatre millions de documents sont disponibles sur la plateforme du projet⁴⁷.

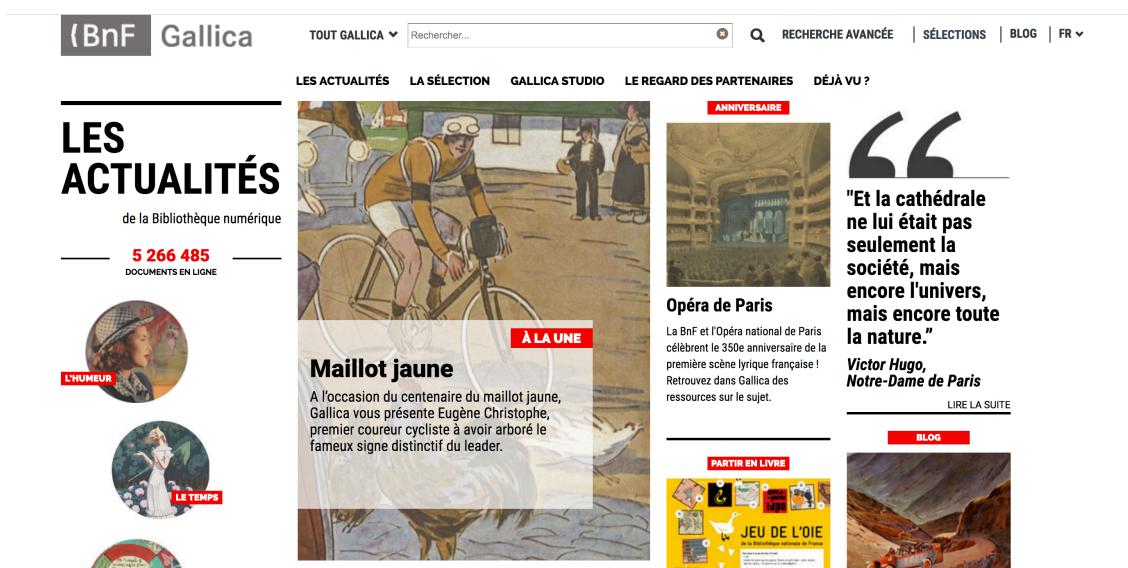


FIGURE 1.5 – Capture-d'écran de la plateforme du projet *Gallica*.

L'*Universal Digital Library* est initiée en 1995 par différents scientifiques en partenariat avec la Carnegie Mellon Foundation (États-Unis). En 1998, ce projet est à l'origine du *Thousand Book Project* qui sera finalement augmenté jusqu'à devenir le *Million Book Project*, successivement terminé en 2007. Il se distingue des autres projets de numérisation de masse, en incluant dès le départ des institutions chinoises, indiennes ou égyptiennes. L'entreprise de numérisation s'est toutefois terminée en 2008⁴⁸, et la plateforme du projet n'est pas parfaitement maintenue. Cet exemple, pourtant ambitieux, démontre que l'avenir réservé à de nombreux projets de numérisation est toujours bien incertain⁴⁹. Andrew Weiss appelle à une meilleure caractérisation des projets de numérisation de masse, afin d'en préserver l'accessibilité sur le long-terme et à sortir de tels projets des contraintes du marché économique : « *However, in dealing with consortia of public and nonprofit educational institutions, market forces should not be the sole factor determining their overall*

46. N. B. Thylstrup, *The politics of mass digitization...*, p.11

47. *Ibid.*, p.140

48. *Ibid.*, p.12

49. A. Weiss, *Using massive digital libraries...*, p.12

sustainability, especially when the content is of significant cultural and social value⁵⁰. »

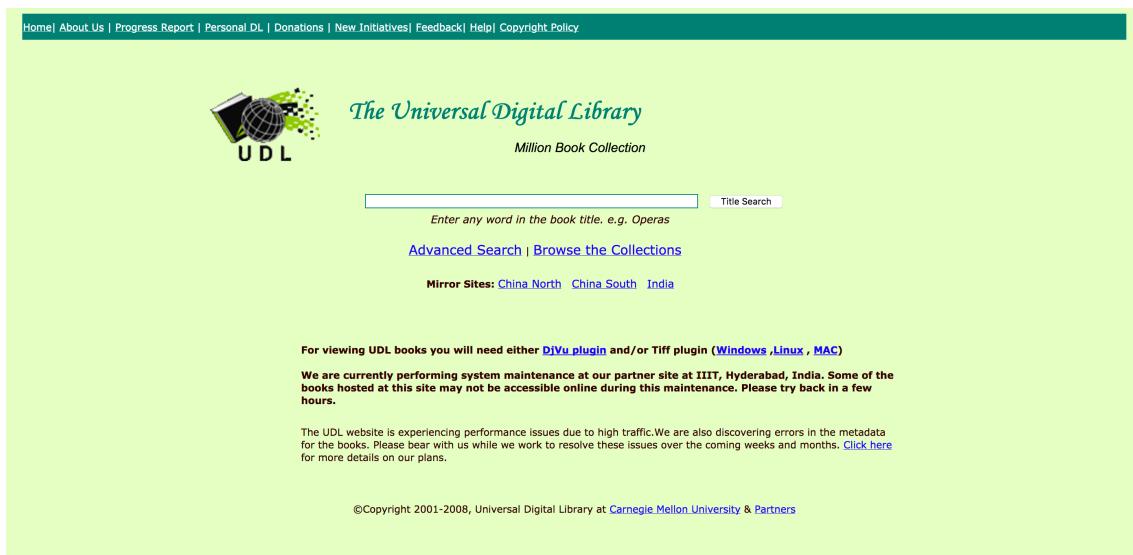


FIGURE 1.6 – Capture-d'écran de la plateforme du projet *Universal Digital Library*.

1996, marque également le lancement d'un projet se poursuivant aujourd'hui : *The Internet Archive*. Fondé par l'activiste du libre accès ou Open Access Brewster Kahle, avec pour objectif initial de préserver le matériel né numériquement (à l'instar des pages et des sites web). Le projet finira par numériser des livres dès 2005, avec l'aide d'une infrastructure regroupant de grandes institutions patrimoniales et acteurs privés, *The Open Content Alliance*⁵¹.

Dans leur désir d'accessibilité, les projets de numérisation de masse ne tiennent d'abord pas compte des aspects légaux propres aux ouvrages numériques⁵², et se voient imposer un certain nombre de restrictions à ce désir d'ouverture. Toutefois l'ambition de construire une bibliothèque numérique universelle commence à se répandre, notamment au sein de l'entreprise Google⁵³. Il est important de spécifier qu'au début des années 2000 de nombreux standards et bonnes pratiques ont été définis et ouvrent la voie aux futurs projets⁵⁴ : « *The digital lifecycle is now well defined, moving from creation through curation, preservation, discovery and use to the creation of new knowledge and content*⁵⁵. » Le mouvement des bibliothèques numériques va continuer de s'amplifier avec de nouveaux projets de numérisation de masse et l'apparition des dépôts de publication en libre accès

50. *Ibid.*, p.30

51. N. B. Thylstrup, *The politics of mass digitization...*, p.12

52. E. Jones, “The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative...”, p.251

53. M. Cook, *Google's Moon Shot...*

54. Alastair Dunning, “Digitising the past : Next steps for public-sector digitisation”, *JISC (Joint Information Systems Committee)* (, juil. 2009), URL : <https://core.ac.uk/download/pdf/11890177.pdf> (visité le 01/07/2019)

55. E. Jones, “The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative...”, p.245

ou Open Access^{56 57}.

Les années 2000 marquent l'histoire des projets de numérisation par le séisme de *Google Books* (2004) qui poussera les bibliothèques et leurs tutelles à revoir leurs stratégies de numérisation⁵⁸, et sa réponse européenne *Europeana* (2008). Le passage à l'échelle de ces initiatives est crucial dans l'industrialisation et la popularité du phénomène des *big data*^{59 60}.

Ces projets emblématiques des initiatives de numérisation de masse seront présentés en deuxième partie de ce mémoire, aux côtés d'autres initiatives similaires issues de ces dernières années.

L'apparition de projets de numérisation conduits par des entreprises à but lucratif semble marquer une séparation entre les initiatives issues du monde privé qui seraient par définition des *massive digitisation projects* et celles du monde public des bibliothèques numériques, soit des *massive digital libraries*⁶¹. Nous nous attacherons à déceler ces pré-supposées différences à travers l'analyse des réponses apportées par ces projets aux enjeux de la numérisation, qui sont souvent bien plus complexes.

1.3 Politiques et numérisation

L'histoire de la numérisation constitue également une histoire politique. L'Union Européenne (UE) et *The United Nations Educational, Scientific and Cultural Organization* ou Organisation des Nations Unies pour l'éducation, la science et la culture (UNESCO) témoignent toutes deux d'une certaine ambition.

En 2003, l'UNESCO adopte la *Charte sur le patrimoine numérique*, argumentant que le temps des solutions individuelles est révolu et que des solutions internationales doivent être mises en place pour faire face à la croissance des coûts liés à ces collections numérisées. L'article 11 appelle à une coopération de tous les acteurs :

Vu la fracture numérique actuelle, il est nécessaire de renforcer la coopération et la solidarité internationales pour permettre à tous les pays d'assurer la création, la diffusion et la conservation de leur patrimoine numérique ainsi que la possibilité d'y accéder en permanence.

56. I. Xie et K. K. Matusiak, *Discover digital libraries...*

57. Comme notre mémoire porte sur les projets de numérisation de masse, nous n'aborderons la question du libre accès ou Open Access qu'à travers certains développements du projet Time Machine.

58. Bernadette Dufrêne, *et al.* (éd.), *Numérisation du patrimoine : quelles médiations ? quels accès ? quelles cultures ?*, OCLC : 859441811, Paris, 2013

59. N. B. Thylstrup, *The politics of mass digitization...*

60. Nous ne parlerons pas en détail de l'histoire des big data dans le présent mémoire, mais pour l'anecdote, Time Machine n'est pas le premier projet à vouloir créer un big data du passé, le projet *Collaborative for Historical Information and Analysis*, initié en 2011 par l'université de Pittsburgh poursuivait le même objectif. Patrick Manning, *Big Data in History*, London, 2013, DOI : 10.1057/9781137378972

61. E. Jones, "The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative...", p.245

En 2009, le constat dressé par l'UNESCO est amer, les recommandations non appliquées mettent l'humanité face au risque de perdre des pans entiers de son histoire⁶².

2009 est également l'année du lancement de la plateforme de la Bibliothèque numérique mondiale lancée par la Bibliothèque du Congrès sous l'égide de l'UNESCO et visant à proposer une interface à large échelle aux collections patrimoniales numériques. Se sachant incapable de rivaliser avec d'autres grandes bibliothèques numériques, elle fait du multilinguisme et du choix sélectif des documents, ses principaux atouts⁶³.

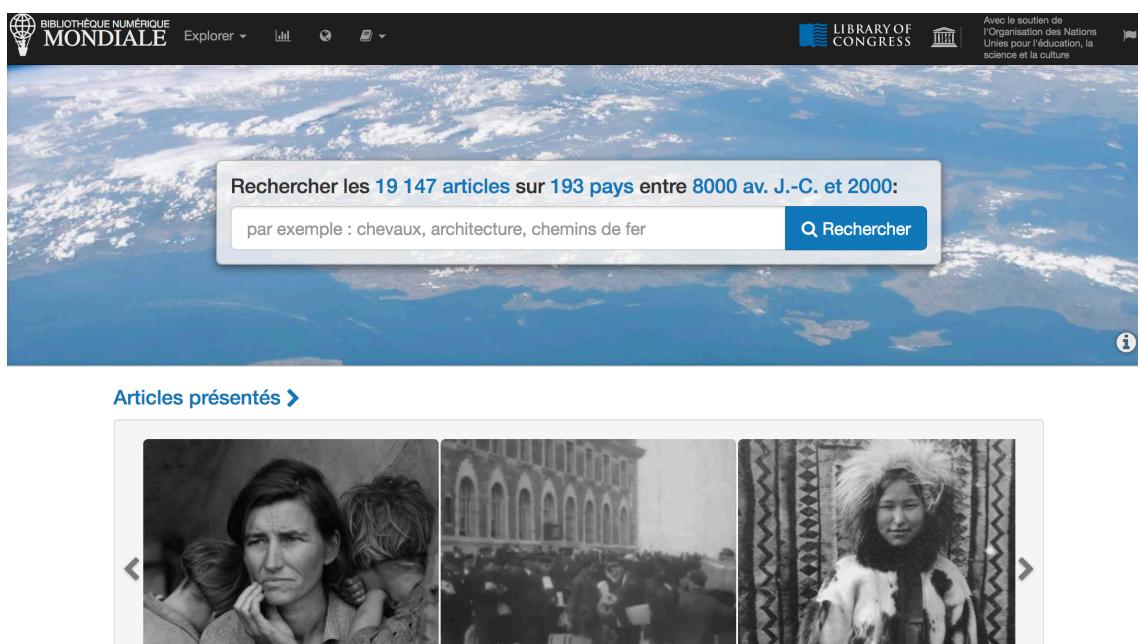


FIGURE 1.7 – Capture-d'écran de la plateforme du projet de Bibliothèque numérique mondiale.

La révolution numérique a fortement ébranlé l'UE qui, par crainte de manquer ce tournant économique, énonce en 2010 diverses mesures dans son agenda Europe 2020, dont une spécifiquement consacrée à la numérisation du patrimoine culturel européen à travers Europeana⁶⁴. Cet objectif, loin d'être nouveau, incarne l'évolution d'une volonté politique qui remonte à 2001. Retour sur l'histoire de ces principales actions politiques soutenant les entreprises de numérisation :

Les principes de Lund, définis en 2001, invitent les états membres à collaborer pour proposer un alignement de leurs programmes de numérisation, afin de favoriser la valorisation du patrimoine numérique et l'émergence d'une nouvelle forme d'économie⁶⁵.

62. Association pour le patrimoine naturel et culturel du canton de Vaud, *Patrimoine numérique, numérisation du patrimoine...*

63. *Numérisation du patrimoine...*

64. M. Coutts, *Stepping away from the silos...*

65. Muriel Foulonneau, "Recherche et numérisation du patrimoine en Europe", *Document numérique*, Vol. 7–3 (2003), p. 179-189, URL : <https://www.cairn.info/revue-document-numerique-2003-3-page-179.htm> (visité le 16/07/2019)

Consciente des nombreuses barrières technologiques à la mise en place d'une numérisation à l'échelle européenne et au développement d'un marché économique, l'UE s'intéresse également à l'interopérabilité et publie en 2002 un rapport sur les bonnes pratiques, *DigiCULT : paysages technologiques pour l'économie culturelle de demain, Faire connaître la valeur de l'héritage culturel*^{66 67}.

Ces principes s'incarneront sous plusieurs formes : le programme MINERVA⁶⁸ en 2003, *The European Library*⁶⁹ en 2005 et Europeana en 2008⁷⁰, et serviront de base pour définir le cadre politique et l'édition d'un certains nombres de recommandations⁷¹. Certains autres projets furent initiés par l'UE durant ces mêmes années à l'instar de MICHAEL *Multicultural Inventory of Cultural Heritage in Europe*, poursuivant des buts similaires et contribuant à rendre parfois trop touffu l'organigramme des initiatives⁷². Afin de clarifier les choses, l'UE a choisi de concentrer ses forces dans la plateforme d'Europeana.

La Commission européenne publie en 2011 des *COMMISSION RECOMMENDATION on the digitisation and online accessibility of cultural material and digital preservation*, qui précisent notamment que :

- La numérisation du patrimoine culturel européen comprend les objets imprimés, les photographies, collections des musées, documents d'archives, sons, documents audiovisuels, monuments et sites archéologiques.
- Pour couvrir les coûts élevés de la numérisation, des partenariats public-privé doivent être mis en place.
- Numériser est une manne économique, puisque ces données permettront aux industries créatives et culturelles de proposer de nouvelles formes de services et les aideront dans cette phase de transition liée à la révolution numérique. Il est urgent d'agir sous peine de manquer cette transformation numérique des industries : « *If Member States do not step up their investments in this area, there is a risk that the cultural and economic benefits of the digital shift will materialise in other continents*

66. Salzburg Research Forschungsgesellschaft et Europäische Kommission (éd.), *The DigiCULT Report : technological landscapes for tomorrow's cultural economy ; unlocking the value of cultural heritage : executive summary*, OCLC : 51714305, Luxembourg, 2002.

67. N. B. Thylstrup, *The politics of mass digitization...*

68. MINERVA vise à mettre en place un réseau de soutien politique pour harmoniser les entreprises de numérisation visant à proposer notamment, des recommandations pour les métadonnées et l'interopérabilité des plateformes. Thierry Claerr et Isabelle Westeel (éd.), *Manuel de la numérisation*, Paris, 2011 (Bibliothèques).

69. Bibliothèque numérique servant de portail à différentes bibliothèques nationales et qui servira d'agrégateur de contenus pour la future Europeana, avant que le projet ne soit gelé en 2016.

70. M. Coutts, *Stepping away from the silos...*

71. *Commission Recommendation of 27 October 2011 on the digitisation and online accessibility of cultural material and digital preservation*, en, rapp. tech. 32011H0711, 2011, URL : <http://data.europa.eu/eli/reco/2011/711/oj/eng> (visité le 22/07/2019)

72. Association pour le patrimoine naturel et culturel du canton de Vaud, *Patrimoine numérique, numérisation du patrimoine...*

and not in Europe.⁷³ »

- Pour diminuer les coûts, il est important d'améliorer les infrastructures techniques existantes et de collaborer à l'échelle européenne pour la mise en place des meilleures solutions.
- Certaines législations doivent être adaptées pour répondre aux besoins d'échanges et d'interopérabilité de ces initiatives. Lever ces barrières doit être fait sur le plan national et européen pour garantir un degré d'uniformisation.
- La conservation des données numérisées doit être faite sur le long-terme.
- Mettre en place des bases de données favorisant l'identification des œuvres sous droit d'auteur et faciliter leurs numérisations.

Ces recommandations spécifient également que le projet Europeana doit parvenir à numériser 30 millions d'objets d'ici 2015, afin de parvenir à la numérisation de l'intégralité de l'héritage culturel d'ici à 2025⁷⁴.

Faisant suite aux recommandations de la Commission, la directive sur certains usages des œuvres dites orphelines, datée d'octobre 2012⁷⁵, vise à permettre la numérisation des œuvres sous droit, si certains conditions sont remplies. Une fois qu'une recherche sérieuse des détenteurs des droits a été menée, une base de donnée permet de les enregistrer avant de procéder à leurs numérisations⁷⁶.

En 2015, l'UE, consciente que la révolution numérique a rétabli des barrières que les politiques européennes s'étaient efforcées de réduire dans le monde physique, notamment concernant la libre circulation des données, inhérente à la réussite de tout projet de numérisation de masse, propose la mise en place d'un marché unique numérique⁷⁷. L'idée est de pouvoir ainsi concurrencer les leaders mondiaux qui bénéficient eux-mêmes de vastes marchés économiques⁷⁸ en se dotant des moyens pour moderniser et décloisonner cette présence numérique⁷⁹. Pour pouvoir proposer un meilleur accès, sans frontières, aux commerces en ligne, il faut notamment lever le principe de géolocalisation et moderniser les lois sur le droit d'auteur⁸⁰. Le marché unique numérique est pensé pour aider l'UE à rattraper son retard économique.

73. *Commission Recommendation of 27 October 2011 on the digitisation and online accessibility of cultural material and digital preservation...*

74. M. Coutts, *Stepping away from the silos...*

75. Les œuvres orphelines sont des œuvres protégées par le droit d'auteur, mais dont le détenteur des droits ne peut être identifié.

76. *Orphan Works DB*, URL : <https://euipo.europa.eu/orphanworks/> (visité le 16/07/2019)

77. Johanna Hoekstra et Aysem Diker-Vanberg, “The proposed directive for the supply of digital content : is it fit for purpose?”, *International Review of Law, Computers & Technology*, 33-1 (janv. 2019), p. 100-117, DOI : 10.1080/13600869.2019.1562638

78. Roberto Viola et Olivier Bringer, “Vers un marché unique numérique : faire de la révolution numérique une opportunité pour l'Europe”, *Revue d'économie financière*, 125-1 (2017), p. 239, DOI : 10.3917/ecofi.125.0239, p.3

79. *Ibid.*

80. “EU Adopts Single-Market Digital Strategy”, *Information Management Journal* (, juin 2015)

[Traduction] L'UE est très en retard par rapport à d'autres pays et régions concernant les échanges commerciaux en ligne, les compétences numériques, la mise en place de règlements appropriés et l'investissement dans les infrastructures numériques. Cela est unanimement reconnu comme une conséquence d'un marché unique très fragmenté qui entrave les échanges commerciaux numériques entre pays européens et gêne le développement de jeunes plateformes européennes et start-ups.⁸¹

Les résultats de la mise en place de ce marché tardent à se faire sentir et certains auteurs argumentent que plutôt que de libéraliser, les propositions sont venues ajouter une couche de régulation administrative supplémentaire aux initiatives numériques⁸². Certains s'accordent à dire que l'UE s'est trompée de combat et plutôt que d'accroître la compétitivité par ce marché unique numérique, elle devrait s'attaquer aux véritables barrières qui se trouvent dans les lois sur le droit d'auteur, trop souvent envisagées uniquement du point de vue des créateurs de l'œuvre, voir plutôt de celui des détenteurs des licences⁸³.

Le nouvel agenda pour la culture, publié en mai 2018, est construit sur l'idée du marché numérique unique, et s'engage à mettre en place un réseau de centres de compétence à travers l'UE afin de préserver le patrimoine bâti par le biais de la numérisation de masse, et renforcer la collaboration entre acteurs culturels, industries créatives, autorités locales, partenaires sociaux, instituts de recherche et d'éducation, autour des initiatives de numérisation⁸⁴.

L'UE s'intéresse de près à l'évolution de ces entreprises de numérisation. La Commission européenne a publié en 2018 un rapport sur cette progression, *European Commission report on Cultural Heritage : Digitisation, Online Accessibility and Digital Preservation*⁸⁵.

81. « *The EU is still lagging far behind other countries and regions when it comes to digital cross-border trade, digital skills, innovative regulation and investment in digital infrastructure. It is widely acknowledged that this is to a large extent the result of a fragmented Single Market, which hinders digital trade between EU-countries and hampers the scaling of young European digital platforms and start-ups* » Paul-Jasper Dittrich, “Balancing ambition and pragmatism for the Digital Single Market”, Jacques Delors Institut (, juil. 2017), p. 14, URL : <https://www.delorsinstitut.de/2015/wp-content/uploads/2017/09/BalancingAmbitionandPragmatismfortheDigitalSingleMarket-Dittrich-JDIB-Sept2017.pdf> (visité le 01/07/2019).

82. *The Next Steps for the Digital Single Market : From Where do We Start ?*, URL : <https://ecipe.org/publications/the-next-steps-for-the-digital-single-market-from-where-do-we-start/> (visité le 04/07/2019)

83. Simone Schroff et John Street, “The politics of the Digital Single Market : culture vs. competition vs. copyright”, *Information, Communication & Society*, 21–10 (oct. 2018), p. 1305-1321, DOI : 10.1080/1369118X.2017.1309445

84. James Drennan, *European Framework for Action on Cultural Heritage*, en, Text, déc. 2018, URL : https://ec.europa.eu/culture/content/european-framework-action-cultural-heritage_en (visité le 12/07/2019)

85. *European Commission report on Cultural Heritage : Digitisation, Online Accessibility and Digital Preservation*, en, Text, juin 2019, URL : <https://ec.europa.eu/digital-single-market/en/news/european-commission-report-cultural-heritage-digitisation-online-accessibility-and-digital> (visité le 04/07/2019)

Avec pour objectif d'évaluer les progrès réalisés concernant l'application des recommandations de 2011, il propose un état des lieux des entreprises de numérisation au sein des pays membres et établit certaines bonnes pratiques⁸⁶. Parmi les faits à relever : un tiers des états font usage de la numérisation et de la 3D pour la préservation du patrimoine, bien qu'il y ait encore un manque de connaissances pratiques constaté dans les milieux professionnels ; un accroissement des partenariats public-privé est observé, pour mieux faire face aux coûts élevés de ces entreprises ; il est complexe d'identifier les œuvres étant du domaine public ; les œuvres sous droit d'auteur sont très peu représentées au sein des collections et la directive concernant les œuvres orphelines est peu utilisée car trop chère à déployer^{87 88}.

Signée durant les *Digital Days 2019*, une nouvelle déclaration européenne *the Declaration of cooperation on advancing digitisation of cultural heritage* s'engage à favoriser la coopération pour faire avancer les entreprises de numérisation patrimoniale. Elle s'articule autour de trois axes⁸⁹.

1. La mise en place d'une initiative européenne pour avancer la numérisation 3D des monuments, sites et artefacts culturels et patrimoniaux.
2. L'usage de ces ressources numériques pour développer l'engagement citoyen et favoriser l'émergence d'entreprises innovantes dans tous les secteurs.
3. Encourager les initiatives entre partenaires de secteurs et pays différents pour augmenter la capacité de développement de cette numérisation de masse de l'héritage culturel.

Il est intéressant de spécifier que le projet Time Machine est cité par la déclaration comme projet répondant à ces objectifs.

Consciente que les enjeux sont du côté de la gestion des droits des œuvres numériques, l'UE a adopté la *Directive on open data and the re-use of public sector information* en avril 2019. Cette dernière spécifie que les œuvres libres de droit numérisées doivent demeurer libres de droit. Si un droit exclusif pour numériser ou reproduire une œuvre est imposé, il ne doit pas excéder une période de dix ans et être au maximum non contraignant. « *Any licences for the re-use of public sector information should in any event place as few restrictions on re-use as possible, for example limiting them to an indication of source.* »

86. Council adopts Copyright Directive, en, URL : <http://era.gv.at/object/news/4678> (visité le 04/07/2019)

87. European Commission report on Cultural Heritage...

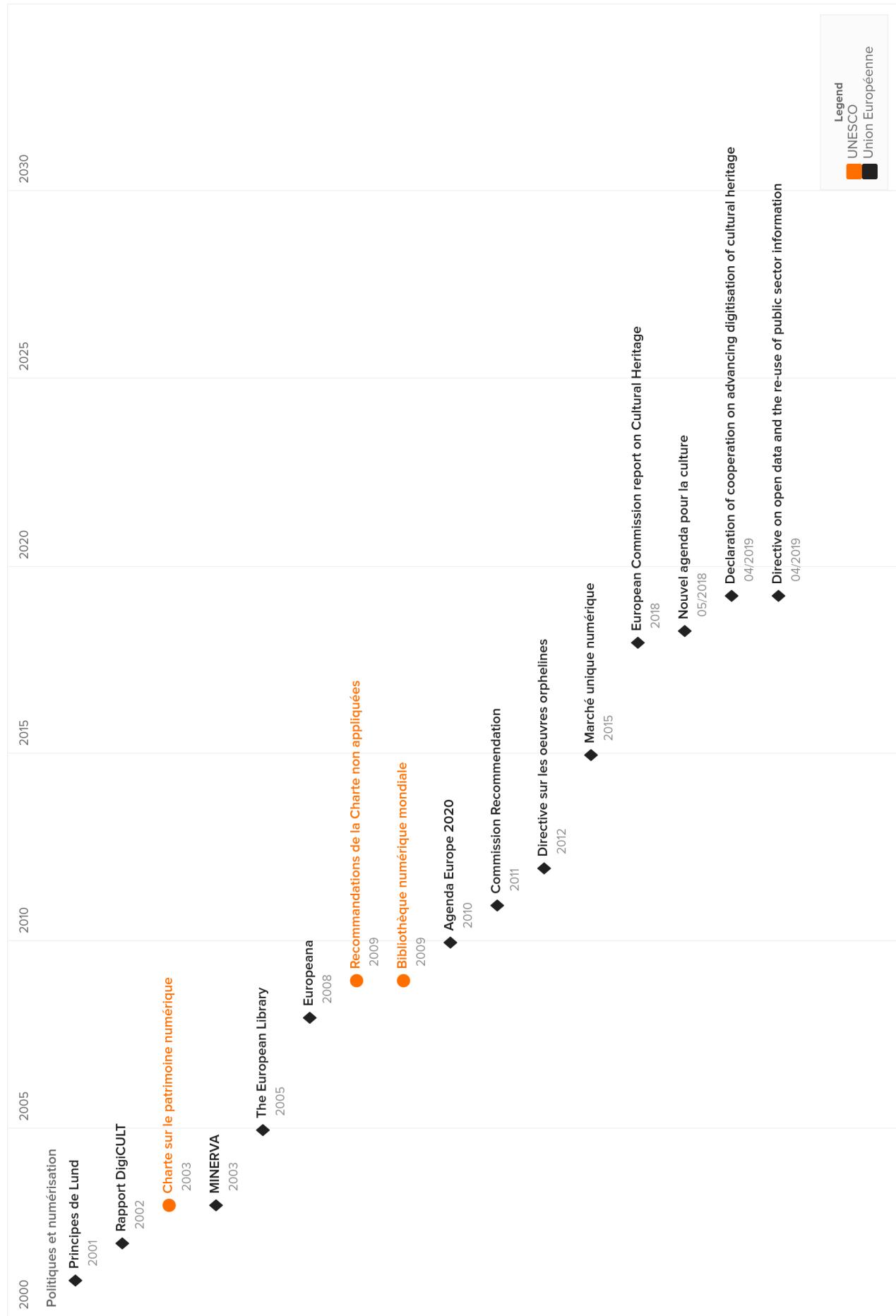
88. Maarten Zeinstra, Research : *Orphan Works Directive does not work for mass digitisation*, en-US, févr. 2016, URL : <https://www.communia-association.org/2016/02/16/orphan-works-directive-does-not-work/> (visité le 16/07/2019)

89. EU Member States sign up to cooperate on digitising cultural heritage, en, Text, avr. 2019, URL : <https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-digitising-cultural-heritage> (visité le 12/07/2019)

L'accélération du rythme de publication des recommandations et des directives, semble indiquer que l'UE a de hautes attentes, concernant la numérisation de son patrimoine. L'incendie récent survenu à Notre-Dame de Paris est venu encore renforcer un débat déjà existant⁹⁰. Europeana ayant dépassé ses objectifs de quantité, l'UE semble désormais choisir la voie du développement de la qualité (multilinguisme, plateforme, amélioration de la médiation du contenu etc.)⁹¹.

90. European Commission report on Cultural Heritage...

91. *Ibid.*

FIGURE 1.8 – Frise temporelle, *Politiques et numérisation*.

1.4 Leçons d'histoire pour Time Machine

De cette revue du contexte historique, nous avons déduit certaines « leçons ».

Contexte historique	Points d'attention pour Time Machine
Prémisses de la numérisation	Le projet de bibliothèque universelle n'est pas nouveau.
	Les développements technologiques impactent la portée et le cadre des projets.
	La version numérique n'est pas parfaitement identique à la version physique du document.
Passage à l'échelle	Pas de définition figée pour les projets de numérisation de masse.
	Les collections numériques ne sont pas le reflet de la connaissance universelle et portent des biais hérités des collections physiques et motivés par leurs sources de financement.
	Différents acteurs impliquent différents intérêts qui se mêlent au sein d'un projet. Point de vue technologique, point de vue patrimonial, point de vue du public ; partenaires publics, partenaires privés.
	La construction d'une collection numérique peut être définie par des choix politiques.
	La taille du projet induit un besoin de standardisation et la mise en place de bonnes pratiques.
	Ouverture et protection du droit d'auteur ne sont pas synonymes, et les enjeux légaux réels.
	Le projet se construit pour être au service de l'utilisateur final et peut être influencé par ce dernier.
Politiques et numérisation	La numérisation est au programme des politiques européennes.
	Il y a une volonté d'aligner les initiatives.
	Des partenariats public-privé servent à financer les projets.
	Les données numérisées permettent l'émergence d'un nouveau secteur économique.
	Les législations concernant le droit d'auteur doivent être adaptées et uniformisées.
	Les données numérisées doivent être conservées sur le long-terme.
	Renforcer la coopération autour des entreprises numériques est dans l'agenda européen.
	Numériser inclut le patrimoine bâti, avec l'aide de la 3D.
	Les œuvres libres de droit doivent demeurer, une fois numérisées.

TABLE 1.1 – Leçons d'histoire pour Time Machine

Chapitre 2

Comment numériser en masse

L'histoire des projets de numérisation a permis de mettre en lumière la complexité et la diversité de ces entreprises, s'inscrivant à la fois dans une forme de rupture et de continuité avec ce qui a précédemment été. Si de nombreux scientifiques s'intéressent au *comment* (quelles réponses apporter aux défis techniques, légaux etc.?) des projets de cette envergure, peu osent s'attaquer au *pourquoi* (quelles formes de connaissance sont créées par ces projets, quels sont les impacts sur l'organisation des entreprises culturelles, de quel contexte politique sont-ils issus, etc.?)¹. Les projets ne se laissent pourtant pas facilement restreindre à des aspects opérationnels et techniques, alors même que ces derniers sont déjà très nombreux et impliquent la définition de nombreuses typologies de standards :

[Traduction] Ces projets de numérisation de masse partagent de nombreuses caractéristiques. Ils impliquent le développement de standards pour une numérisation uniforme et pour les métadonnées techniques et descriptives ; la prise de décisions sur les droits d'utilisation ; le choix des logiciels de stockage, de l'environnement de mise à disposition des ressources et du matériel informatique.²

Nanna Bonde Thylstrup ose voir au-delà du *comment*, en argumentant que ce changement de contenant (passage d'informations cloisonnées dans un objet physique à la création de jeux de données) induit la création de nouvelles voies de transmission de la connaissance et la redéfinition des aspects politiques, légaux, financiers associés. Ce qui aboutit à d'inévitables controverses découlant de tant d'intérêts mélangés.

[Traduction] Les méthodes de numérisation de masse, créent de nouvelles formes de connaissances et de nouveaux moyens de la découvrir. Ce qui à

1. N. B. Thylstrup, *The politics of mass digitization...*

2. « *These mass digitization projects have many similar characteristics. They require standards for uniform digitization, standards for descriptive and technical metadata, decision on rights and permissions, decisions on software for storage and end-user access, and decisions on hardware and operating environments* » Stacy T. Kowalczyk, *Digital curation for libraries and archives*, Santa Barbara, California, 2018, p.11.

première vue semble se limiter à un simple acte de numérisation (la transformation des limites physiques d'un livre, en un jeu de données libre), révèle, si nous le regardons de plus près, un processus complexe, foisonnant de diverses controverses politiques, légales et culturelles.³

Pour mieux analyser ces projets et oser s'intéresser au *pourquoi* (dans la troisième partie de ce mémoire), il nous faut d'abord présenter leur matière même, ce *comment*. Nous vous présentons dans ce chapitre leurs principales caractéristiques puis certains enjeux qui en découlent⁴. Bien que les projets de numérisation de masse ne se laissent que mal résumer tant leurs complexités et spécificités sont grandes, notre revue de la littérature et notre expérience de stagiaire nous ont permis de définir les problématiques minimales induites par ces initiatives.

2.1 Caractéristiques des projets

Avant de tenter une caractérisation des projets de numérisation de masse, rappelons brièvement la signification du terme numérisation. Dans le monde du patrimoine, la numérisation implique traditionnellement⁵ :

- La conversion d'un objet physique ou analogique (en général des images, fichiers audio ou vidéo ou du texte) en une suite de données interprétables par des machines.
- La conversion des catalogues sur fiche des bibliothèques selon le même processus.

Le projet Time Machine y ajoute une troisième dimension puisqu'il désigne également :

- La transformation du patrimoine bâti ou phénomène géographique en modèles 3D.

Dans le cadre de ce mémoire, la définition de la numérisation se veut la plus large possible et peut se résumer par la création d'un contenu, copie ou enregistrement numérique, d'une information analogique contenue par un document, artefact, son, performance, élément géographique ou phénomène naturel.

Au-delà de l'acte lui-même, numériser désigne une série de processus d'analyse intellectuelle et d'indexation visant à satisfaire les besoins des utilisateurs avec, pour double objectif, d'assurer la préservation du document au-delà de la durée de vie de son support et de pouvoir valoriser le document numérisé et son modèle original.⁶. Tout projet de

3. « *The practice of mass digitization is forming new excuses of knowledge, and new ways of engaging with that knowledge. What at first glance appears to be a simple act of digitization (the transformation of singular books from boundary objects to open sets of data), reveals on closer examination, a complex process teeming with diverse political, legal, and cultural investments and controversies* » N. B. Thylstrup, *The politics of mass digitization...*, p.1

4. Ces derniers étant très nombreux et de nouveaux ne cessant d'apparaître, nous ne serions prétendre à une absolue exhaustivité.

5. *Numérisation du patrimoine...*, p.37

6. *Numérisation*, août 2013, URL : <https://www.essib.fr/le-dictionnaire/numerisation> (visité le 11/07/2019)

numérisation pose une série de difficultés, liées aux coûts souvent élevés, au temps nécessaire pour la réalisation, à la qualité de l'indexation, à la conservation à long-terme des données, au coût induit par la maintenance des équipements informatiques et à la qualification des équipes chargées de gérer la masse des documents numérisés⁷.

Les objets choisis pour la numérisation ne sont pas systématiquement le fait de politiques de sélection mais découlent des ressources financières, du temps, de la qualité et des objectifs que l'on veut accomplir : « *Therefore, looking to our traditional media, digitization is not a question of selection according to certain criteria, but according to financial resources, time line, priorities, quality and/or aims we want to reach.*⁸ »

Karen Coyle définit les projets de numérisation de masse comme la conversion de matériel à une échelle industrielle, sans véritable politique de sélection et utilisant les techniques de reconnaissance optique de caractères pour permettre les recherches plein-texte : « [...] mass digitization refers to converting materials on an industrial scale without curating specific materials for digitization. OCR is used to make the full text of digitized documents searchable »⁹.

Iris Xie met en avant leurs plateformes d'accès, en les désignant comme les nouvelles générations de systèmes de découverte, offrant un accès centralisé vers une grande variété du patrimoine culturel et scientifique et caractérisés par plusieurs millions de titres, des formats divers, une politique de gestion du droit d'auteur, une politique de sélection, et des degrés d'accessibilité dérivés en fonction, ainsi que par la mise en place de standards techniques pour assurer l'interopérabilité¹⁰.

Comme déjà mentionné¹¹, certains auteurs présentent¹² ou osent une distinction entre les projets menés par des partenaires commerciaux et les projets issus des institutions publiques. Justifié par un enjeu plus grand mis sur l'accessibilité des données dans le premier cas, et par la prise en compte de modération humaine dans la gestion des collections, avec un accent plus développé sur les aspects liés à la préservation des données et la réutilisations des métadonnées existantes dans le deuxième cas. Les objectifs de création d'un moteur de recherche universel et le développement de nouveaux services basés sur l'indexation plein-texte d'immenses corpus de données aux formats divers et issus du domaine public restent communs aux deux types d'initiatives¹³.

Il est toutefois certain que la complexité de tels projets ne saurait faire l'objet d'une seule définition tant les intérêts qu'ils mêlent soulèvent de multiples questions,

7. *Numérisation du patrimoine...*, p.37

8. Association pour le patrimoine naturel et culturel du canton de Vaud, *Patrimoine numérique, numérisation du patrimoine...*, p.19

9. Karen Coyle, “Mass Digitization of Books”, *The Journal of Academic Librarianship*, 32–6 (nov. 2006), p. 641-645, DOI : 10.1016/j.acalib.2006.08.002

10. I. Xie et K. K. Matusiak, *Discover digital libraries...*, p.23

11. Pour plus de détails, référez-vous à la section 1.2.

12. E. Jones, “The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative..”

13. C. Lampert, “Ramping up...”, p.47

« *Mass digitization brings together so many disparate interests and elements that any monotheoretical lens would fail to account for the numerous political issues arising within the framework of mass digitization*¹⁴ ».

La numérisation de masse semble surtout se distinguer des autres formes de numérisation par une contrainte moins importante sur la sélection des sources et par la vitesse des processus liée aux développements technologiques et à l'automatisation. Ces entreprises ne sont dès lors pas résumables en un tout ou par l'une de leurs parties, elles sont constituées de différents assemblages¹⁵. Si les projets sont difficiles à catégoriser, ils se regroupent derrière la mission de permettre un meilleur accès aux collections des institutions patrimoniales¹⁶.

Avec l'avènement et la construction de notre société numérique, le développement d'une économie liée et l'arrivée de nouvelles générations, il y a désormais un sentiment d'urgence à numériser le savoir qui motive la poursuite de projets de grande ampleur. Celui-ci fait écho aux premières craintes des informaticiens qui prédisaient que les pratiques des utilisateurs seraient à l'avenir uniquement orientées en ligne.

[Traduction] Dans un contexte d'accroissement de la dominance numérique et puisque la plupart des biens patrimoniaux et culturels existent incontestablement uniquement sous une forme physique, la numérisation demeure comme par le passé, un élément clé de tous futurs développements.¹⁷

[Traduction] Il existe aujourd’hui des risques considérables à ne pas faire progresser les projets de numérisation. Les nombreux utilisateurs exclusivement numériques pourraient ignorer, voire ne jamais être informés de l’existence de contenus pertinents pour leurs besoins et intérêts. Une connaissance négligée à travers le temps, devient une connaissance perdue, au détriment de tous.¹⁸

Cette intensification des projets commence à ouvrir les débats autour des questions du *pourquoi* de la numérisation, et certains auteurs s’attachent à proposer de nouveaux angles d’analyse, interrogeant la typologie des acteurs de la numérisation et les leçons à apprendre de ces initiatives^{19 20}.

14. N. B. Thylstrup, *The politics of mass digitization...*, p.5

15. *Ibid.*, p.26

16. C. Lampert, “Ramping up...”, p.58

17. « *In the light of increasing digital dominance, however, and the incontrovertible fact that so much of the world’s knowledge and cultural outputs still exist only in physical form, digitisation is as much a key element in future digital developments as it has been in the past* » M. Coutts, *Stepping away from the silos...*, p.4

18. « *There are considerable risks in not advancing digitisation at this time. The numerous digital-only users may ignore or never know of key content which is relevant to their interests and needs. Neglected knowledge, over time, becomes lost knowledge, to the detriment of all* ». *Ibid.*

19. C. Lampert, “Ramping up...”, p.47.

20. N. B. Thylstrup, *The politics of mass digitization...*

2.2 Enjeux des projets

Nous avons choisi de présenter les enjeux qui nous ont semblés incontournables pour toute entreprise de numérisation, cependant le développement rapide de ces initiatives et des technologies numériques, nous interdit de prétendre être exhaustive. Ces derniers se construisent autour des questions liées à la mise en place d'une collaboration effective entre différents partenaires, au coût, au droit d'auteur, à l'interopérabilité découlant de choix techniques ou de la construction des corpus, ainsi qu'à la préservation et au stockage de ces données. Ces différents axes interagissent les uns avec les autres au sein d'un même projet, et les décisions relatives peuvent influencer sa réussite finale. L'arrivée de *Google Books* a fortement impacté les stratégies des projets de numérisation et semble être l'origine de la nouvelle complexité de ces enjeux²¹.

Les réponses apportées par certaines initiatives et par Time Machine à ces enjeux du *comment* seront examinées respectivement dans la deuxième et la troisième partie du mémoire.

2.2.1 Amener différents acteurs à collaborer

Les projets de numérisation de masse ont poussé les bibliothèques à sortir de leur isolement. Pour arriver à une meilleure efficacité et une meilleure « compétitivité », il a fallu travailler à inscrire la numérisation du patrimoine dans les objectifs de politique publique, définir plus précisément la notion d'Europe culturelle déjà existante dans l'agenda européen²², agrandir l'offre à d'autres objets que les livres et partager les infrastructures existantes. Ce qui a conduit les GLAMs à vouloir davantage coopérer au sein d'un même projet²³.

[Traduction] Les bibliothèques et archives sont très conscientes des opportunités offertes par la collaboration en regard de ce qui pourrait être fait en travaillant seul. Nous constatons une grande volonté de leur part à développer de telles initiatives. Les bibliothèques publiques et les musées sont les premiers à viser l'amélioration de leurs services au public par le biais de la collaboration et de la mise en commun de collections et d'infrastructures.²⁴

S'inscrivant dans le contexte multilatéral et international de la mondialisation, les projets doivent orchestrer le double retour du local et du national²⁵.

21. *Numérisation du patrimoine...*

22. Pour plus de détail référez-vous à la section 1.3.

23. *Ibid.*

24. « *There is also a strong emphasis on the opportunities to achieve more by working with others than can be achieved by working alone, and this is specified by the library and archives sectors in particular. The public libraries and museums are prominent among those targeting improved public services through collaboration, and in sharing collections and infrastructure* » M. Coutts, *Stepping away from the silos...*, p.23.

25. *Numérisation du patrimoine...*, p.175.

La collaboration est également nécessaire pour définir les choix techniques inhérents à la mise en commun des collections de diverses institutions, que les frontières nationales soient dépassées ou non par le cadre du projet²⁶.

L'accélération des processus induits par l'envergure du projet implique souvent l'externalisation à un prestataire de certains actes de numérisation et le transport des documents concernés. L'organisation du flux des documents entre l'institution détentrice et le prestataire doit être soigneusement planifiée²⁷.

De la qualité de la collaboration découlent la possibilité de coordonner différentes initiatives de numérisation et de proposer une vision globale. Cet aspect semble souvent faire défaut aux initiatives, alors même que la collaboration est susceptible d'impacter fortement tous les autres enjeux²⁸. Parvenir à faire collaborer un nombre croissants de partenaires induit la mise en place d'outils de gestion, la définition d'un cadre, et la prise en compte du multilinguisme, et représente de fait un investissement nécessaire pour la réussite du projet²⁹.

[Traduction] Bien entendu, chaque organisation sera confrontée à ses propres limites, et certaines initiatives ne pourront être poursuivies faute d'une collaboration effective mise en place avec soin. Collaborer implique un environnement délicat et finement équilibré [...]. Collaborer signifie travailler avec un ensemble cohérent et complexe de partenaires régionaux, nationaux et internationaux.³⁰

2.2.2 Financement et partenariats public-privé

Les coûts élevés de la numérisation de masse et les limitations posées par le système de numérisation à la demande^{31 32}, ont poussé les initiatives à chercher des solutions du

26. Alexandre Moatti, “Bibliothèque numérique européenne : de l’utopie aux réalités”, *Réalités Industrielles* (, nov. 2012), p. 43-46.

27. Ministère de la Culture et de la communication : Direction des Archives de France, *Ecrire un cahier des charges de numérisation du patrimoine*, rapp. tech., URL : https://francearchives.fr/file/bf50d8fa5f554586dbf18fdc862d25970a1da0a7/static_4132.pdf.

28. M. Coutts, *Stepping away from the silos...*

29. Kristin R. Eschenfelder, Kalpana Shankar, Rachel D. Williams, Dorothea Salo, Mei Zhang et Allison Langham, “A nine dimensional framework for digital cultural heritage organizational sustainability : A content analysis of the LIS literature (2000–2015)”, *Online Information Review*, 43–2 (avr. 2019), p. 182-196, DOI : 10.1108/OIR-11-2017-0318.

30. « Ultimately, however, there will be limits to every organisation’s options, and there will be initiatives that cannot go forward because they lack the collaborative commitment that they require. Collaboration is a finely balanced and delicate environment [...]. Collaboration means working with a coherent and complex set of regional, national, and international partnerships » M. Coutts, *Stepping away from the silos...*, p.26.

31. Lionel Maurel, “Quel modèle économique pour une numérisation patrimoniale respectueuse du domaine public ?”, dans *Communs du savoir et bibliothèques*, Paris, 2017 (Bibliothèques), p. 73-84, DOI : 10.3917/elec.dujo.2017.01.0073.

32. En général, le premier utilisateur paie pour le processus, le fichier est ensuite publié si libre de droit, sur la plateforme de l’institution.

côté des partenaires privés (individus, banques, fondations, médias, industries du tourisme etc.)^{33 34}. En général, comme les coûts sont fortement imputés à l'infrastructure³⁵, ce sont vers des entreprises spécialisées dans ces technologies (à l'instar de Google, Proquest, IBM, Kodak)³⁶, que se tournent les institutions culturelles et patrimoniales. On trouve toutefois des engagements avec des fondations telles que Wikimedia et le recours au financement participatif ou *crowdfunding*^{37 38}.

L'arrivée des ces partenariats privés a reçu un accueil au départ très mitigé par crainte d'un monopole et de la réappropriation des ouvrages libres de droit publiés sous licence^{39 40}. Ces partenariats financiers entre institutions publiques et prestataires privés posent également des problèmes liés à la réutilisation des données, puisque certains prestataires imposent des conditions d'exclusivité^{41 42}. De récentes études montrent toutefois que l'exclusivité tend à ne plus porter directement sur l'utilisation des œuvres numérisées mais se déplace vers la commercialisation de services⁴³.

En dépit de la prolifération de ces partenariats, il manque un socle de règles communes visant à limiter les risques de monopole de certains acteurs privés et à fixer un cadre permettant d'harmoniser et légitimer ces pratiques⁴⁴, afin de trouver l'équilibre entre des financements gouvernementaux, locaux, privés et caritatifs⁴⁵. La mise en place d'une politique active de numérisation dans chaque état pourrait également permettre de rééquilibrer la situation^{46 47}.

Malgré la créativité déployée par les initiatives pour parvenir à couvrir leurs frais,

33. L. Lopatin, "Library digitization projects, issues and guidelines"...

34. R. Yeates et D. Guy, "Collaborative working for large digitisation projects"...

35. De grands progrès ont toutefois été faits sur les couches technologiques, la montée des coûts serait due aux recherches pour les droits d'auteur. Victoria Stobo, Kerry Patterson, Kristofer Erickson et Ronan Deazley, "I should like you to see them some time : An empirical study of copyright clearance costs in the digitisation of Edwin Morgan's scrapbooks", *Journal of Documentation* (, janv. 2018), DOI : 10.1108/JD-04-2017-0061.

36. *Numérisation du patrimoine...*

37. *European Commission report on Cultural Heritage...*

38. Cette piste demeure ambiguë, car pose des problèmes de réutilisation commerciale des œuvre ainsi numérisées. L. Maurel, "Quel modèle économique pour une numérisation patrimoniale respectueuse du domaine public ?"...

39. Voir aussi à ce propos dans la section 1.3 la récente directive de l'UE, sur les ouvrages libres de droit.

40. Association pour le patrimoine naturel et culturel du canton de Vaud, *Patrimoine numérique, numérisation du patrimoine...*

41. *Numérisation du patrimoine...*

42. Voir aussi à ce propos dans la section 1.3 la récente directive de l'UE, visant à limiter la durée des périodes d'exclusivité.

43. L. Maurel, "Quel modèle économique pour une numérisation patrimoniale respectueuse du domaine public ?"...

44. *Numérisation du patrimoine...*

45. R. Yeates et D. Guy, "Collaborative working for large digitisation projects"...

46. L. Maurel, "Quel modèle économique pour une numérisation patrimoniale respectueuse du domaine public ?"...

47. Les Pays-Bas ont par exemple mis en place une loterie nationale pour répondre aux besoins financiers des institutions culturelles.

un manque est constaté dans l'organisation des projets, qui ont tendance à négliger la perspective du long-terme. Les projets financés par des bourses locales, nationales ou européennes sont souvent les plus susceptibles de présenter un modèle financier lacunaire⁴⁸. Ceci conduit à une multiplicité de projets financés, mais plus accessibles car les contenus n'ont jamais été mis à jour et les plateformes datées⁴⁹.

Très peu d'études s'attachent à analyser le rapport entre l'audience générée par les plateformes et leurs coûts de production, ce qui permettrait aux initiatives de justifier ou d'adapter leurs investissements⁵⁰.

2.2.3 Droit d'auteur

Les projets de numérisation de masse suscitent une prise de conscience de l'urgence à ne pas limiter les collections aux ouvrages libres de droit⁵¹. La question du droit d'auteur devient incontournable, les institutions ayant la responsabilité de déterminer si le droit d'auteur s'applique ou non aux œuvres qu'elles souhaitent numériser⁵².

La problématique du droit d'auteur, dans tout projet de numérisation, intervient de différentes manières⁵³ :

- Certaines œuvres sont libres de droits, car dans le domaine public.
- Certaines œuvres sont couvertes par le droit d'auteur, mais dites orphelines⁵⁴.
- Certaines œuvres existent nativement sous forme numérique et sont couvertes par des licences numériques.

Les exceptions au droit d'auteur sont réglées par un certain nombre de régimes de droits nationaux et quelques directives européennes visant à faciliter leur gestion. Il n'existe pas, en Europe, de réel équivalent aux exceptions définies par le *fair use*⁵⁵ états-unien, qui a notamment servi à justifier la pratique de *Google Books*. L'UE est toutefois consciente des enjeux liés au droit d'auteur et a mis en place un certain nombre de recommandations et de directives⁵⁶.

La montée des coûts des entreprises de numérisation de masse se justifie de plus en plus par la recherche des auteurs des œuvres sous droit⁵⁷. En effet dans l'environnement

48. C. Lampert, “Ramping up...”.

49. K. R. Eschenfelder, K. Shankar, R. D. Williams, *et al.*, “A nine dimensional framework for digital cultural heritage organizational sustainability...”.

50. A. Moatti, “Bibliothèque numérique européenne : de l'utopie aux réalités”...

51. *Numérisation du patrimoine...*

52. L. Lopatin, “Library digitization projects, issues and guidelines”...

53. *Numérisation du patrimoine...*

54. Pour rappel, les œuvres orphelines sont des œuvres soumises au droit d'auteur mais dont le détendeur est inconnu.

55. Le *fair use* autorise cinq exceptions au droit d'auteur : le droit à opérer une copie privée, le droit de citation, le droit au pastiche ou à la caricature, l'usage pour l'enseignement, l'usage pour la recherche.

56. Pour plus de détail, référez-vous à la section, 1.3

57. V. Stobo, K. Patterson, K. Erickson, *et al.*, ““I should like you to see them some time”...”.

numérique, les détenteurs des droits ont un contrôle presque absolu sur l'utilisation de l'œuvre originale⁵⁸. Les détenteurs de ces droits refusent souvent de libérer les œuvres et ont peu d'intérêt à réduire les coûts pour les initiatives de numérisation, car cela représente une perte financière⁵⁹. Les acteurs cherchent surtout à maintenir un équilibre économique, chèrement acquis par les industries créatives⁶⁰. « *The commercial publishers representing the greatest copyright holders were the most reluctant to grant the permission, fearing loss of their profits*⁶¹. »

Cette situation conduit à une sous-représentation des œuvres à partir du 20^e siècle au profit des œuvres plus anciennes généralement libres de droit, phénomène connu sous le nom de « Trou noir » du 20^e siècle⁶², *20th century black hole*⁶³ ou encore *digital black hole*⁶⁴.

Dans le cas des œuvres dites orphelines, la directive européenne de 2012⁶⁵ permet aux institutions de les numériser, mais impose la poursuite de recherches diligentes et le respect de nombreuses règles (interdiction de les utiliser ensuite pour un usage commercial, etc.) avant le lancement des processus⁶⁶. Ce qui s'avère trop coûteux pour la plupart des projets⁶⁷.

Pour les œuvres encore soumises au droit d'auteur, mais non trouvables dans le commerce, il n'existe aucune directive visant à faciliter la recherche des détenteurs de droit et à lever les coûts imposés par ces derniers⁶⁸.

Il y a un grand besoin de rétablir l'équilibre entre la protection des droits d'auteur et du profit qu'il génère pour certains, et la protection des intérêts du grand public et de la société à utiliser ces œuvres⁶⁹. Le droit d'auteur doit être adapté aux nouveaux modes de production de la connaissance⁷⁰.

[Traduction] Ces dernières années, le droit d'auteur semble avoir dérivé vers une protection des détenteurs du droit au détriment des bénéfices pour

58. Aleksandra Pavlovic, “The Serpent in the Garden of Eden : Intellectual property in the Digital Millennium”, *Academia.edu* (, déc. 2011), URL : https://www.academia.edu/34127494/_The_Serpent_in_the_Garden_of_Eden_Intellectual_property_in_the_Digital_Millennium (visité le 04/07/2019).

59. V. Stobo, K. Patterson, K. Erickson, *et al.*, ““I should like you to see them some time”...”.

60. *Ibid.*

61. Carole A. George, “Testing the barriers to digital libraries : A study seeking copyright permission to digitize published works”, *New Library World*, 106–7/8 (juil. 2005), p. 332-342, DOI : 10.1108/03074800510608648, p.54.

62. *Numérisation du patrimoine...*

63. *European Commission report on Cultural Heritage...*

64. Association pour le patrimoine naturel et culturel du canton de Vaud, *Patrimoine numérique, numérisation du patrimoine...*

65. Pour plus de détail, référez-vous à la section, 1.3

66. V. Stobo, K. Patterson, K. Erickson, *et al.*, ““I should like you to see them some time”...”.

67. *European Commission report on Cultural Heritage...*

68. *Ibid.*

69. A. Pavlovic, “The Serpent in the Garden of Eden : Intellectual property in the Digital Millennium”...

70. *Numérisation du patrimoine...*

la société. Les délégués semblent incapables ou peu disposés à changer cet équilibre par de nouvelles législations. Puisque les règlements existants sont mal équipés pour faire face aux nouvelles technologiques, les riches détenteurs de droits ont prestement utilisé ces nouvelles technologies pour accroître la protection de leurs œuvres ou intimider les usagers.⁷¹

Solution possible à cette situation paradoxale, le libre accès ou Open Access et l'Open Science méritent d'être davantage développés. Les exceptions au droit d'auteur doivent également être davantage explorées par les projets de numérisation de masse dont les institutions craignent trop souvent les retombées légales, ce qui ajouterait la gestion des risques à la liste des enjeux identifiés⁷². « Cependant ni la capitulation devant le droit d'auteur ni la transgression provocatrice des règles fondamentales ne peuvent résoudre le problème à la longue⁷³. »

[Traduction] Compte tenu des coûts trop élevés associés à la recherche des auteurs des œuvres orphelines dans le cas de collections de grandes tailles, et du manque d'alternatives législatives, nous proposons que les institutions culturelles et patrimoniales apprennent à vivre en acceptant une certaine forme d'incertitude. Elles doivent explorer les perspectives offertes par une gestion consciente des risques dans leurs stratégies de gestion et mettre à profit le large champ des exceptions prévues dans les législations sur le droit d'auteur.⁷⁴

2.2.4 Sortir des silos - la quête de l'interopérabilité

Tout projet de numérisation de masse implique à un certain moment, en plus de la numérisation, la réunion de collections déjà numériques et la mise à disposition de l'ensemble de ces données à travers une plateforme commune en ligne⁷⁵. Rassembler ces différentes données, puis les rendre trouvables, est complexe. De trop nombreuses bases de données ont été déployées par les institutions, épargnant les données à travers des

71. « *In recent years, the balance of copyright appears to have tipped more toward the rights of copyright owners over the benefits to society, with legislators unable or unwilling to change that balance through new legislation. Because existing statutory language is ill-equipped to handle new technologies, wealthy and powerful copyright holders have been quick to use technology to expand protection of their works or to intimidate users* » I. Xie et K. K. Matusiak, *Discover digital libraries...*, p.54.

72. V. Stobo, K. Patterson, K. Erickson, *et al.*, ““I should like you to see them some time”...”, p.661.

73. Association pour le patrimoine naturel et culturel du canton de Vaud, *Patrimoine numérique, numérisation du patrimoine...*

74. « *Given the unmanageable transaction costs associated with these orphan works schemes when dealing with collections of any notable size, as well as the lack of other useful cognate legislative form, we argue that Cultural Heritage Institutions must learn to live with the uncertainty inherent in copyright law : that is, they must explore risk management strategies in more depth, and utilise the full scope of the exceptions already available within the copyright regime* » V. Stobo, K. Patterson, K. Erickson, *et al.*, ““I should like you to see them some time”...”, p.661.

75. C. Lampert, “Ramping up...”.

systèmes différents et contribuant à les enfermer et les isoler les unes des autres, dans des « silos »⁷⁶.

L'interopérabilité est un argument pour beaucoup, pour l'UE qui souhaite mettre en place des procédures communes favorisant l'économie de la connaissance, pour les professionnels des bibliothèques et archives souhaitant sortir de leurs logiques de « silos » et pour le mouvement de *Open Science*. L'objectif de l'interopérabilité est de faciliter les échanges de données entre les différents systèmes des institutions culturelles et patrimoniales⁷⁷ et de proposer des services permettant d'accroître les interactions avec ces ressources au sein d'un même réseau⁷⁸.

[...] l'interopérabilité est devenue un enjeu à un autre niveau que par le passé. Il ne s'agit plus seulement d'échanger des notices pour éviter la redondance dans l'effort de catalogage, mais de mutualiser les outils à travers des « portails » pour tenter d'accrocher un usager devenu insaisissable⁷⁹.

Alors que les standards et les bonnes pratiques apportent de nouvelles réponses aux enjeux techniques posés par la numérisation, ces derniers ne semblent pas encore avoir trouvé leur forme finale.

De plus, un autre écueil contribue à isoler les collections des institutions culturelles et patrimoniales : celui de la sélection du contenu intellectuel. Un manque de coordination et de vision globale sur la création de ces collections les rend fortement disparates et complique davantage leur réunification au sein d'une même plateforme⁸⁰, augmentant également la redondance des données numérisées⁸¹.

De l'attention apportée aux enjeux techniques et aux enjeux sur le contenu, découle d'une part la qualité et la richesse de l'expérience proposée ensuite pour utilisateur, et d'autre part la qualité de la conservation sur le long-terme⁸².

Nous n'avons pas pour objectif de définir quelle sont les meilleures pratiques à adopter, puisque ces dernières sont propres au contexte de chaque projet et liées aux objectifs utilisateurs, mais nous vous proposons une introduction des différents éléments auxquels toute initiative sera amenée à se confronter.

Il n'existe pas encore de modèle de donnée européen ou universel propre à assurer l'interopérabilité. La coopération entre institutions culturelles, professionnels de l'informa-

76. Institut national d'histoire de l'art, #LundisNum / 11 février 2019 - Gautier Poupeau : rassembler les métadonnées des collections de l'INA, URL : https://www.youtube.com/watch?list=PLs18NWzVv6T2CQFtB0fnIA_EKLFeCFSUG&time_continue=52&v=KY0zoRPks8Q&fbclid=IwAR3QgiLH5rEPiDGFX-PPdIMVsH_83H3mJuHnn8jeJBVi2_01-jVWJa4hZW8 (visité le 18/05/2019).

77. L. Lopatin, "Library digitization projects, issues and guidelines"...

78. I. Xie et K. K. Matusiak, *Discover digital libraries*...

79. Emmanuelle Bermès, Antoine Isaac et Gautier Poupeau, *Le Web sémantique en bibliothèque*, OCLC : 866574281, Paris, 2013, p.21.

80. M. Coutts, *Stepping away from the silos...*

81. Institut national d'histoire de l'art, #LundisNum / 11 février 2019 - Gautier Poupeau...

82. I. Xie et K. K. Matusiak, *Discover digital libraries*...

tion et utilisateurs est essentielle pour la définition et l'implémentation d'un tel modèle⁸³ qui devra être à la fois souple, cohérent et garant de la provenance des données⁸⁴. Une étude européenne datée de 2004 définit trois critères à la réussite de l'interopérabilité : organisationnel, technique et sémantique⁸⁵.

Enjeux techniques

Les enjeux techniques se découpent en deux grandes familles. La première est liée aux processus de numérisation, ses choix portent sur la qualité des productions numériques et les formats⁸⁶ des fichiers ensuite produits. De cette qualité dépendent surtout la découverte proposée aux utilisateurs⁸⁷ et la préservation sur le long-terme des données⁸⁸. Les bonnes pratiques privilégient le choix de formats ouverts (par exemple .tiff ou .png pour les images)⁸⁹.

Pour s'assurer de la conformité aux critères de qualité, un contrôle doit être mis en place, faisant partie intégrante du projet. Il ne peut se réduire à une simple étape du processus. « Suivant les moyens disponibles et le niveau de qualité souhaité, on procédera au contrôle à différents moments de la production »⁹⁰.

Si la recherche de la qualité des produits de la numérisation a motivé les premières initiatives, il semble que les projets de numérisation de masse ont appelé à une redéfinition des priorités pour produire plus en allégeant les processus, ce qui ne se fait pas sans difficulté. « ANSI la non-qualité et la sur-qualité sont des extrêmes à éviter »⁹¹. La qualité minimale se fixant sur l'usage et non sur les conditions de préservation. Les entreprises de numérisation de masse sont ainsi appelées à faire preuve de flexibilité, en appliquant des critères de qualité différents en fonction des documents⁹².

83. Diane Rasmussen Pennington et Laura Cagnazzo, “Connecting the silos : Implementations and perceptions of linked data across European libraries”, *Journal of Documentation*, 75–3 (mai 2019), p. 643-666, DOI : 10.1108/JD-07-2018-0117.

84. Institut national d'histoire de l'art, #LundisNum / 11 février 2019 - Gautier Poupeau...

85. Anna Zharova, “Influence of the principle of interoperability on legal regulation”, *International Journal of Law and Management*, 57–6 (nov. 2015), p. 562-572, DOI : 10.1108/IJLMA-07-2014-0044, p.565.

86. Le format désigne la nature d'un document informatique et permet d'identifier le logiciel nécessaire à sa lecture. Chaque fichier porte une extension en trois lettres indiquant le format (par ex. pdf pour les documents)

87. Par exemple, une image sera accessible en plus ou moins haute résolution, ce qui ralentira ou non le navigateur de l'usager

88. Pour plus de détails, référez-vous à la section 2.2.5

89. Un format ouvert par opposition aux formats propriétaires est un fichier dont le code est accessible et transparent et peut être lu par différents logiciels.

90. *Manuel de la numérisation...*, p.271.

91. *Ibid.*

92. I. Xie et K. K. Matusiak, *Discover digital libraries...*



FIGURE 2.1 – Image de mauvaise qualité, numérisée par Google ©Google Books

La deuxième famille des enjeux techniques est liée aux données servant à décrire les produits numérisés, soit les métadonnées. La quête de l'interopérabilité permettant de faire communiquer ces corpus de métadonnées décrivant les collections, indépendamment de l'environnement technique de leurs bases de données, est l'une des solutions pour la sortie des silos, ou c'est, en tous les cas, la présomption qui en est faite.

« *It's a metadata's world, and you're just living in it*⁹³. » Les métadonnées sont pleinement intégrées dans notre société et passent inaperçues. Nous n'avons pas toujours conscience que lorsque nous achetons un livre en librairie, c'est le titre, l'auteur ou la couverture qui auront motivés notre choix. Les informations servant à décrire le livre ont toujours été utilisées dans les bibliothèques, pour permettre de retrouver et classer les documents et sont à la base des catalogues. Les métadonnées nous servent à simplifier la réalité. Par exemple, « Camille » est mon prénom, mais ce n'est pas moi en tant que telle.

Il existe plusieurs catégories de métadonnées, aux frontières parfois un peu floues : les métadonnées descriptives⁹⁴, les métadonnées administratives⁹⁵, les métadonnées struc-

93. Jeffrey Pomerantz, *Metadata*, Cambridge, Massachusetts ; London, England, 2015 (The MIT Press essential knowledge series), p.4.

94. Servant à simplifier l'objet ou l'information par le biais de mots-clés.

95. Informations sur l'origine d'un document, par exemple une photographie numérisée avec tel scanner, telle résolution, sous telle licence

turelles⁹⁶, les métadonnées de préservation⁹⁷, les métadonnées d'usage^{98 99}. Si les métadonnées descriptives ont été développées avec pour objectif d'améliorer la recherche sur internet, elles n'ont pas longtemps permis de le réaliser. Très vite, des entreprises ont tenté de tromper les moteurs en ajoutant de nouveaux mots-clés, et de fait, ce champ est souvent ignoré. Leur intérêt est néanmoins en train de se réveiller, mais plus pour améliorer l'affichage que pour influencer les résultats d'une recherche¹⁰⁰.

La multiplicité des catégories de métadonnées implique la création d'un fichier capable de toutes les contenir : le registre de métadonnées.

L'écriture de ces métadonnées est définie par un certain nombre de règles spécifiant la description de l'élément, on parle alors de schéma de métadonnées. Les schémas sont structurés en triplets : objet (par ex. Camille Besse), prédicat (par ex. est l'auteure), sujet (par ex. du présent mémoire), les règles du schéma servent à spécifier quels prédicats peuvent être utilisés et quelles sortes de déclaration sujet-prédicat-objet sont possibles¹⁰¹. La valeur prise par les différentes données du triplet est également normalisée, avec l'aide notamment de vocabulaires contrôlés¹⁰² et d'ontologies¹⁰³. « *If a metadata schema expert control over the kinds of statements that may be made, a controlled vocabulary expert control over the words and phrases that may be used in those statements.*¹⁰⁴ »

Bien que les métadonnées soient cruciales pour le succès des entreprises de numérisation de masse¹⁰⁵, elles ont souvent été développées sans souci d'uniformisation, rendant la situation actuelle très hétérogène¹⁰⁶. Les institutions culturelles ne traitent pas leurs métadonnées selon les mêmes schémas (différences notables entre les collections des bibliothèques, des archives, des musées)¹⁰⁷. Nous verrons, dans la deuxième partie de ce mémoire, les exemples d'Europeana et de *The Digital Public Library of America* qui, dans leurs objectifs de rassembler différentes collections, ont créé leurs propres schémas de métadonnées¹⁰⁸.

Pour pallier à cette hétérogénéité des schémas de métadonnées, le *Semantic Web Education and Outreach Interest Group*, créé en 2006 par le *World Wide Web Consortium* (W3C)¹⁰⁹ a proposé une nouvelle solution visant à faciliter l'échange de ces métadonnées,

96. Informations sur l'organisation de l'objet, par exemple un livre se compose de chapitres.

97. Informations spécifiques à la préservation sur le long-terme.

98. Informations sur l'usage de l'objet, par exemple combien de fois un livre a-t-il été lu.

99. *Ibid.*

100. *Ibid.*

101. *Ibid.*

102. Lexique servant à organiser les connaissances pour favoriser la recherche d'information

103. « Une ontologie est une façon de modéliser un domaine en identifiant les concepts y afférent et en les organisant les uns par rapport aux autres » Nicolas Delestre, Nicolas Malandin et Michel Bussi, *Du Web des documents au Web sémantique*, OCLC : 974379992, 2017.

104. J. Pomerantz, *Metadata...*, p.33.

105. *Ibid.*

106. D. Rasmussen Pennington et L. Cagnazzo, “Connecting the silos...”.

107. E. Bermès, A. Isaac et G. Poupeau, *Le Web sémantique en bibliothèque...*

108. J. Pomerantz, *Metadata...*

109. Communauté internationale chargée de développer les standards du web afin de garantir sa

pour permettre aux machines de les interpréter et construire de nouvelles applications et de nouveaux services¹¹⁰. S'exprimant à travers les technologies du web sémantique, le web de données est défini comme :

[...]un réseau où les données structurées qui se trouvent actuellement isolées dans les bases de données pourraient être exprimées sous une forme permettant aux machines de les interpréter et de construire de nouvelles applications et de nouveaux services. Pour cela, les données doivent être partagées dans un espace commun et reliées en utilisant des identifiants fiables et uniques¹¹¹.

Le web sémantique représente un ensemble de pratiques évolutives plus qu'une technologie spécifique. Il définit quatre principes¹¹² : les ressources doivent être identifiées par des Uniform Resource Identifier (URI)s¹¹³ ; ces URIs doivent être formulés suivant le protocole HyperText Transfer Protocol (HTTP)¹¹⁴ ; lorsqu'on accède à une ressource via son URI, il doit renvoyer des informations utiles et pertinentes standardisées ; les résultats doivent également contenir des liens vers d'autres URIs pour favoriser la découverte de nouvelles collections¹¹⁵.

Il est intéressant de relever que parmi les critères de qualité évaluant la mise en place du web de données, il est également recommandé de privilégier des formats ouverts pour les documents¹¹⁶.

Le Resource Description Framework (RDF) constitue la grammaire ou la colonne vertébrale du web sémantique, c'est le standard qui permet de formater la réponse à une requête via l'URI du document. RDF est un langage d'encodage des métadonnées et permet la description de leurs relations¹¹⁷. Suivant les mêmes bonnes pratiques que les schémas de données, il permet en plus la création de liens. Les données sont structurées en triplets, reliés les uns aux autres pour former un graphe. Une entité peut être à la fois le sujet ou l'objet d'autres triplets, ce qui permet de créer des liens entre plusieurs collections. La normalisation des valeurs du triplet est d'autant plus nécessaire à la mise en place du web de données¹¹⁸. « La navigation de lien en lien dans le web de données doit rendre possible l'exploitation conjointe de ressources décrites différemment, pourvu qu'elles aient un point de contact [...]¹¹⁹. »

Europeana et *The Digital Public Library of America*, ont créé en plus de leurs longévité.

110. E. Bermès, A. Isaac et G. Poupeau, *Le Web sémantique en bibliothèque...*

111. *Ibid.*

112. D. Rasmussen Pennington et L. Cagnazzo, “Connecting the silos...”.

113. Chaîne de caractère servant à désigner une ressource de manière non ambiguë.

114. Protocole qui permet la transmission d'information sur le web.

115. Dydimus Zengenene, “Global interoperability and linked data in libraries”, *New Library World*, 114–1/2 (janv. 2013), p. 84-87, DOI : 10.1108/03074801311291992.

116. D. Rasmussen Pennington et L. Cagnazzo, “Connecting the silos...”.

117. D. Zengenene, “Global interoperability and linked data in libraries”...

118. J. Pomerantz, *Metadata...*

119. E. Bermès, A. Isaac et G. Poupeau, *Le Web sémantique en bibliothèque...*, p.45.

schémas de métadonnées un modèle, permettant de définir la structure sémantique selon laquelle les métadonnées sont exprimées en classes et propriétés RDF¹²⁰.

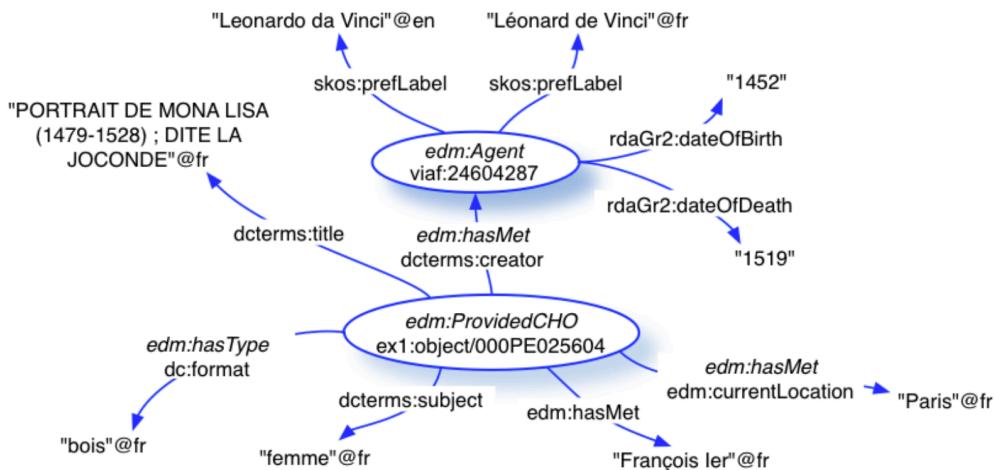


FIGURE 2.2 – Extrait de *Europeana Data Model Primer* ©Europeana.

Le web de données propose de nouvelles solutions, mais demeure une technologie avec certains défauts : il est très compliqué de faire face à un grand nombre de requêtes sur un corpus structuré selon le modèle RDF, et la création de liens entre les différents jeux de données est une pratique demandant beaucoup de temps pour son implémentation¹²¹. Il faut plus que des données structurées en RDF pour permettre un bon fonctionnement du web sémantique, la création de schémas de données de référence demeure un prérequis.

[Traduction] Si chaque service web développe son propre schéma, cela équivaut à chaque ville et village développant leur propre borne d'incendie : il ne pourrait y avoir aucune collaboration entre les départements de pompiers, puisque les lances incendies ne fonctionneraient qu'avec leur borne locale. C'est seulement lorsque tous utiliseront les mêmes standards qu'une collaboration à grande échelle pourra être mise en place.¹²²

D'autres technologies travaillent à rendre interopérables les métadonnées des collections, à l'instar du protocole Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) ou des *Application Programming Interface* (API), qui permettent à des applications de communiquer entre elles¹²³. Les APIs sont des mécanismes très populaires pour accéder aux métadonnées sur les documents et aux documents eux-mêmes, elles sont

120. *Ibid.*

121. D. Rasmussen Pennington et L. Cagnazzo, “Connecting the silos...”.

122. « If every web service developed its own schema for structuring its data, that would be the equivalent of every town and city developing its own type of fire hydrants : there could be no collaboration between fire departments because no department's hoses would fit any other department's hydrants. Only when everyone - or at least a significant portion of everyone - is using the same standards, is widespread collaboration possible » J. Pomerantz, *Metadata...*, p.186.

123. E. Bermès, A. Isaac et G. Poupeau, *Le Web sémantique en bibliothèque...*

dans la pratique presque davantage utilisées que les pages d'accueil des institutions¹²⁴. Les bibliothèques ont développé un standard très populaire spécifiquement dédié à l'interopérabilité des images, International Image Interoperability Framework (IIIF). Cette méthode de publication des données se base sur l'implémentation de différentes APIs et implique une structuration des métadonnées. Europeana a développé une extension de son modèle de métadonnées pour favoriser l'utilisation de IIIF par ses partenaires¹²⁵.

L'acquisition des compétences spécifiques pour les gestionnaires des collections numériques ne peut se limiter à la maîtrise d'une tâche, puisque la réussite du processus implique une vision d'ensemble et la prise de nombreuses décisions stratégiques (mise en place de standards de qualité pour la numérisation, définition des schémas de métadonnées, proposition d'un modèle, choix des outils pour favoriser l'interopérabilité etc.). « C'est tout une culture professionnelle du numérique qu'il faut élaborer et transmettre, de façon à permettre aux acteurs de s'inscrire dans un environnement global, de comprendre les implications de leurs choix et de leurs actions.¹²⁶ ».

Enjeux sur le contenu

Les collections numérisées sont disparates, proposant aux utilisateurs des ressources lacunaires ou contenant des doublons. Cela s'explique par des questions de financement, ou des questions de lois, de structures, de techniques^{127 128}, mais également par le processus de sélection, du point de vue intellectuel, des ressources à numériser, et qui contribue à la problématique des silos. Les différentes stratégies de numérisation ont d'abord cohabité, développant pour chaque projet des critères propres, avant l'émergence des premières initiatives visant à les réunir. Bien que les critères de sélection intellectuels sont essentiels dans les entreprises de numérisation, ils n'ont jamais fait partie des contraintes de standardisation.¹²⁹.

Si les projets de numérisation de masse se sont construits dans la volonté de tout numériser et semblent à priori se passer de critères de sélection¹³⁰, la réalité est différente. Même si la sélection semble plus large, une prédominance des livres et des journaux dans les collections numérisées est à relever. De plus pour les agrégateurs que sont Europeana ou *The Digital Public Library of America*, la sélection est fortement influencée par les collections et décisions des institutions partenaires, et par leurs positions géographiques.¹³¹

124. J. Pomerantz, *Metadata...*

125. IIIF Consortium, *Home — IIIF / International Image Interoperability Framework*, URL : <https://iiif.io/> (visité le 18/05/2019).

126. *Manuel de la numérisation...*, p.303.

127. I. Xie et K. K. Matusiak, *Discover digital libraries...*

128. L. Lopatin, “Library digitization projects, issues and guidelines”...

129. M. Coutts, *Stepping away from the silos...*

130. C. Lampert, “Ramping up...”.

131. La notion de « patrimoine européen » n'a pas attendu Europeana pour être diffusée à travers le monde. Le patrimoine des grandes bibliothèques américaines, numérisé par Google s'en est déjà chargé. C'est le patrimoine allemand, anglais, espagnol, français, italien de l'émigration européenne qui fonde les

La vision de la bibliothèque universelle étant à priori issue du monde occidental et de la recherche, les plateformes d'accès aux produits de ces initiatives empêchent souvent leurs usages par des communautés ne partageant pas les mêmes normes¹³². Ce biais influence également la représentation des collections non occidentales dans les collections, qui suivent des formats de documents souvent mal pris en charge par les acteurs du marché¹³³.

Les projets de numérisation à plus petite échelle ne vont pas s'arrêter au profit des plus grands et l'intégration de ces collections signifie l'acceptation de leurs critères de sélection. « *Digitisation work short of mass digitisation will continue, and selection criteria will be applied, whether in a structured or unstructured fashion*¹³⁴. »

Le manque de normalisation pour la sélection peut sans doute être expliqué par le contexte et le développement rapide des projets de numérisation de masse, qui ont souvent « palié au plus pressé » et se sont d'abord confrontés aux aspects techniques et légaux.

[Traduction] L'avancée de la numérisation a requis le développement et le renforcement de nombreux nouveaux éléments liés aux aspects techniques, structurels, légaux, financiers, au développement des interfaces de recherche. Ces aspects complexes évoluent de plus en plus rapidement, ce qui rend compréhensible le focus initial pour la création de standard et la coordination des pratiques associées. En conséquence, la mise en place de critères et d'aide à la sélection des ressources du point de vue de leurs contenus intellectuels a été négligé, alors même que ce champ d'étude était familier des acteurs de la culture et de l'information, habitués à les mettre en pratique dans la création de collections physiques.¹³⁵

Bien que les critères de sélection des objets à numériser se sont construits sans uniformisation, un groupe de critères communs émerge peu à peu. Margaret Coutts en propose huit, qui s'articulent en fonction du contexte des projets¹³⁶.

1. L'accessibilité

Les projets de numérisation se construisent autour de deux objectifs différents, celui

bases de celui des États-Unis. A. Moatti, “Bibliothèque numérique européenne : de l'utopie aux réalités”...

132. E. Jones, “The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative...”.

133. A. Weiss, “Examining Massive Digital Libraries (MDLs) and Their Impact on Reference Services”, *The Reference Librarian*, 57–4 (oct. 2016), p. 286-306, doi : 10.1080/02763877.2016.1145614.

134. M. Coutts, *Stepping away from the silos...*, p.78.

135. « [...] the advent of digitisation required that many entirely new elements in the technical, infrastructural, discovery, preservation, legal and financial areas had to be developed, consolidated and matured. These were complex fields that changed rapidly. It is understandable that there was much emphasis on the creation of robust standards and coordination of such practice ; they were imperatives to apply the new technology effectively. As a consequence, guidance and criteria for the selection of content in terms of its intellectual value, a field that was already very familiar to information and heritage practitioners in relation to analogue materials, was somewhat overshadowedIbid., p.80. »

136. *Ibid.*

de la préservation¹³⁷, et celui de l'accessibilité qui promet à la fois d'être élargie et enrichie par une publication en ligne. Ces deux notions font débat et dépendent fortement du contexte du projet.

2. La notion de valeur

Guidée par des critères intellectuels, historiques, physiques (format). Sélectionner en tenant compte de la valeur d'un document peut aussi être motivé par la notion de « valeur ajoutée », découlant des nouveaux usages apportés par la version numérisée. La valeur est un choix subjectif qui dépend des intérêts des utilisateurs.

3. La notion de rareté ou d'unicité

Proposer pour la première fois la version numérisée.

4. La sélection thématique

Une masse critique des documents numérisés est souvent atteinte par la mise en commun de plus petites entreprises, souvent articulées autour d'une thématique, d'un sujet particulier.

5. Le format et le support

Certains objets ont fait l'objet d'un numérisation plus rapide, à l'instar des livres, journaux, cartes et plans.

6. La cohérence

Critère souvent utilisé lorsqu'il s'agit de mettre en commun différentes collections.

7. La réunification virtuelle

Permettant la réunification de ressources liées mais disséminées physiquement à travers différentes institutions.

8. Le regroupement

Propose un développement du critère précédent. Visant à réunir et enrichir des ressources déjà numérisées afin de développer de nouveaux services.

Bernadette Dufrêne justifie la disparité des collections par leur développement initial autour de différents modèles, celui du musée où les pièces rares et trésors sont numérisés ; celui des archives où les documents sont numérisés selon une granularité d'ensembles thématiques liée à un fonds ; celui des bibliothèques, articulé souvent autour d'une charte documentaire détaillée visant à proposer la bibliothèque numérique idéale du chercheur et non élargir l'accès au patrimoine ; celui de Google, sans politique documentaire¹³⁸. Toutefois, l'absence de politique documentaire n'est pas forcément synonyme d'absence de critères de sélection comme nous l'avons vu plus haut.

Certains autres biais sont également listés, à l'instar des motivations des partenaires financiers qui peuvent venir imposer ou diriger certains choix, ou de la surreprésentation

137. Pour plus de détail, référez-vous à la section 2.2.5.

138. *Numérisation du patrimoine...*

des collections des bibliothèques au détriment de celles des musées ou des archives, qui s'explique par un plus rapide développement de la numérisation dans ce secteur¹³⁹.

La croissance de la masse des données implique la mise en place de véritables directives dans la poursuite de la numérisation, afin de favoriser la réunification de ces données, d'éviter la redondance des collections et promouvoir la naissance de nouveaux usages¹⁴⁰.

2.2.5 Stockage sur le long-terme - préservation

Si la numérisation peut servir à préserver les objets physiques¹⁴¹, la préservation signifie la maintenance sur le long-terme de l'accès aux objets numérisés ou à l'information contenue¹⁴². Assurer cette longévité dépend de moyens techniques, économiques, légaux, organisationnels et structurels^{143 144}. La prise en compte de cet enjeu est parfois laissée de côté dans les projets de numérisation de masse¹⁴⁵ alors que l'incertitude de pouvoir accéder sur la durée aux données a des impacts négatifs sur l'usage des plateformes et la conduite de projets de recherches dédiés¹⁴⁶.

Préserver implique souvent la copie de l'objet numérisé original, visant à la fois à servir de substitut et à proposer une version plus allégée pour les interfaces numériques¹⁴⁷. Les métadonnées de contexte sont importantes, puisque l'objet numérisé est souvent séparé de son format original dans le processus de numérisation¹⁴⁸.

Deux notions sont intrinsèques à la préservation, l'« intégrité » qui signifie que l'objet est le même que l'objet initialement numérisé, et l'« authenticité » qui consiste à s'assurer que l'objet est bien ce qu'il prétend être. La prise en compte de ces deux éléments doit assurer le long-terme de la compréhension et de l'usage de l'objet, alors même que les technologies seront modifiées¹⁴⁹.

En dehors de la question du stockage, on distingue trois technologies différentes pour la préservation des données¹⁵⁰.

139. M. Coutts, *Stepping away from the silos...*

140. En Suède, un système de coordination a été mis en place par le biais de la Bibliothèque Nationale. *Numérisation du patrimoine...*

141. I. Xie et K. K. Matusiak, *Discover digital libraries...*

142. M. Coutts, *Stepping away from the silos...*

143. I. Xie et K. K. Matusiak, *Discover digital libraries...*

144. Georgetown University, David Ribes, Thomas Finholt et University of Michigan, “The Long Now of Technology Infrastructure : Articulating Tensions in Development”, *Journal of the Association for Information Systems*, 10–5 (mai 2009), p. 375-398, DOI : 10.17705/1jais.00199.

145. L. Lopatin, “Library digitization projects, issues and guidelines”...

146. K. Shankar et K. R. Eschenfelder, “Sustaining Data Archives over Time : Lessons from the Organizational Studies Literature”, *New Review of Information Networking*, 20–1–2 (juil. 2015), p. 248-254, DOI : 10.1080/13614576.2015.1111699.

147. M. Coutts, *Stepping away from the silos...*

148. Joan E. Beaudoin, “Context and Its Role in the Digital Preservation of Cultural Objects”, *D-Lib Magazine*, 18–11/12 (nov. 2012), DOI : 10.1045/november2012-beaudoin1.

149. Regina Varnienė-Janssen et Jūratė Kupriénė, “Authenticity and Provenance in Long-Term Digital Preservation : Analysis of the Scope of Content”, *Informacijos mokslai*, 82 (déc. 2018), p. 131-160, DOI : 10.15388/Im.2018.82.9.

150. S. T. Kowalczyk, *Digital curation for libraries and archives...*

1. La préservation des technologies de lecture

Impliquant la conservation des anciens systèmes et des ordinateurs capables de les faire fonctionner.

2. L'émulation

Consiste à recréer un environnement technique capable d'exécuter les anciens programmes tout en fonctionnant avec une système moderne.

3. La migration

Passage d'un format de données vers un autre format. Souvent la méthode la moins coûteuse, mais implique certaines modifications de l'objet original, voir des pertes de données.

Les technologies numériques présentent un paradoxe en terme de préservation. Ils multiplient les risques et les difficultés liés à la préservation sur le long-terme (dus à l'obsolescence des formats et programmes notamment) et proposent de nouvelles solutions et outils spécifiquement développés pour y pallier¹⁵¹.

Face aux difficultés posées par la préservation, de nouvelles pratiques voient le jour, visant à prendre en compte tous les aspects du cycle de vie des données (ajout de métadonnées standardisées, mise en place d'un plan de préservation, prise en compte des besoins des utilisateurs, valorisation et accessibilité) : la curation de contenus numériques. Des infrastructures cross-institutions sont développées afin de partager les savoirs et les coûts¹⁵². Les bonnes pratiques développées pour le stockage des données ont conduits les dépôts institutionnels à devenir plus robustes, permettant à la fois l'ingestion de différents formats de données, leur organisation intuitive et proposant des fonctionnalités de recherche avancées¹⁵³.

L'UE a également financé des recherches visant à proposer des solutions de stockage sur le long-terme et des standards de métadonnées adaptés aux projets de numérisation de masse et leurs collections de données hétérogènes^{154 155}.

151. I. Xie et K. K. Matusiak, *Discover digital libraries...*

152. K. Shankar et K. R. Eschenfelder, "Sustaining Data Archives over Time..." .

153. Le mouvement de l'Open Science a poussé les institutions à proposer des solutions efficaces pour les données de la recherche, qui peuvent être utilisées pour les projets de numérisation de masse. S. T. Kowalczyk, *Digital curation for libraries and archives...*

154. Par exemple, le *Scalable Preservation Environments* ou SCAPE projet, conclu en 2014.

155. Bolette Ammitzbøll Jurik, Asger Askov Blekinge, Rune Bruun Ferneke-Nielsen et Per Møldrup-Dalum, "Bridging the gap between real world repositories and scalable preservation environments", *International Journal on Digital Libraries*, 16–3-4 (sept. 2015), p. 267-282, DOI : 10.1007/s00799-015-0152-4.

2.2.6 Résumé des enjeux de la numérisation

Enjeux	Résumé
Amener différents acteurs à collaborer	Susceptible d'impacter tous les autres enjeux, une collaboration de qualité doit être mise en place entre les différents partenaires.
Financement et partenariats public-privé	Les coûts élevés impliquent des fonds privés. Ces partenariats suscitent des craintes et manquent d'encadrement. Peu d'argent est dédié à la préservation sur le long-terme.
Droit d'auteur	Projets souvent limités aux œuvres libres de droit. Cela coûte cher de rechercher les détenteurs de droits. Les œuvres du 20 ^e siècle sont sous-représentées. Pas de cadre légal européen, mais de nombreux régimes nationaux. Réflexions autour de la directive sur les œuvres orphelines et la prise de risques.
Sortir des silos : enjeux techniques	L'objectif est de rassembler des collections de provenances différentes et permettre leurs interrogations en-dehors du silo de leurs institutions. Nécessité de fixer des standards pour la numérisation, et de définir des critères de qualité. Nécessité de fixer des standards pour les métadonnées, présentation des différentes typologies, du rôle du schéma et de la place des vocabulaires contrôlés. Présentation des quatre principes du web sémantique et usages du RDF, introduction d'autres outils favorisant l'interopérabilité (OAI-PMH, APIs, IIIF). Face à ces enjeux techniques en constante évolution, il est nécessaire de développer la formation des gestionnaires de collections.
Sortir des silos : enjeux sur le contenu	Un processus de sélection intellectuel non coordonné, qui contribue à la création de collections disparates et redondantes et ne favorise pas la mise en commun des collections. Nécessité de mettre en place des directives et de prendre conscience des typologies derrière ces critères de sélection (valeur, format, unicité, accessibilité, thématique, cohérence, réunification virtuelle, regroupement).
Stockage sur le long-terme - préservation	Préserver l'accessibilité sur le long-terme des données numérisées afin d'en favoriser l'usage. Différentes méthodes (émulation, préservation des technologies, migration) et systèmes permettent la mise en place de cette préservation. Le coût reste élevé, et les projets sont souvent construits sans plan de préservation.

TABLE 2.1 – Résumé des enjeux de la numérisation

Chapitre 3

Contexte du projet Time Machine

Pour comprendre l'articulation du projet Time Machine au sein de l'histoire des projets de numérisation de masse et les réponses apportées aux enjeux liés à l'envergure de ces initiatives, il nous faut premièrement introduire le contexte du projet. Nous tenterons de définir dans le troisième temps de ce mémoire, si ses objectifs et solutions l'inscrivent dans une démarche de continuité ou contribuent à positionner les entreprises de numérisation de masse en tant que nouvel acteur de l'information.

Cette présentation du cadre décrit d'abord le contexte académique de l'EPFL et de la recherche en humanités numériques qui l'ont vu naître, puis le projet *Venice Time Machine* et les diverses activités du DHLAB, racines des développements futurs.

Nous terminons ce chapitre par une présentation de Time Machine, ses objectifs et son organisation au moment de commencer notre stage.

Nous décrirons plus précisément notre mission de stagiaire et la réponse apportée par Time Machine aux enjeux de la numérisation dans la troisième partie de ce mémoire.

3.1 La recherche en humanités numériques

« Nouveau » domaine d'étude, dont l'appellation est apparue au tournant des années 2000, les *digital humanities (DH)*, humanités digitales ou humanités numériques pour les chercheurs francophones¹, proposent des programmes de recherche en cooccurrence avec les technologies informatiques et les différentes disciplines des sciences humaines.² Si les tentatives de définition abondent et demeurent volontairement larges, c'est aussi par refus d'établir « une définition qui préexisterait à l'usage »³.

Notre stage se déroulant au sein d'un laboratoire d'humanités numériques, il nous

1. Nous privilégions l'emploi du terme humanités numériques dans ce mémoire, qui est l'appellation utilisée en France, lieu de notre soutenance.

2. Jean-Guy Meunier, “Le paradoxe des humanités numériques”, *Quaderni*–98 (févr. 2019), p. 19-31, DOI : 10.4000/quaderni.1407, p. 2

3. Benjamin Caraco, “Les Digital humanities et les bibliothèques”, *Bulletin des bibliothèques de France (BBF)*–Bulletin des bibliothèques de France (BBF) (2012), p. 69-73, p. 2

semble pertinent d'introduire quelques fondamentaux du domaine, pour mieux comprendre les origines académiques du projet Time Machine. Les débats étant à ce jour vifs parmi les communautés de chercheurs, dont les historiens, « les réunions DH créent de nouveaux temps d'échange qui dépassent les thématiques et les périodes historiques, pour mettre en dialogue antiquisants, médiévistes, modernistes et contemporanéistes »⁴, nous ne prétendrons pas être exhaustif.

Les humanités numériques constituent des zones d'échange, de projets à la frontière entre plusieurs disciplines, créant dans leurs pratiques de nouveaux espaces d'expression et leurs organisations associées, à l'instar d'*Humanistica : Association francophone des humanités numériques*,⁵ des conférences *Digital Humanities*, humanités digitales, humanités numériques (DH) Benelux « *an initiative that aims to further the collaboration between Digital Humanities activities in Belgium, The Netherlands, and Luxembourg* »⁶, des THATCamp « *The Humanities and Technology Camp* »⁷, de Dariah en Europe « *a pan-european infrastructure for arts and humanities scholars working with computational methods* »⁸, ou encore des conférence DH organisées par l'« *Alliance of Digital Humanities Organizations* »⁹.

La variété des approches est grande et diffère en fonction des rôles spécifiques attribués à l'expert des humanités ou à l'expert de l'informatique. Dans le premier type d'approche, les humanités numériques sont la suite des études menées en humanités, le numérique étant alors un outil aux services de « *what is important today is not that we are doing work with computers, but rather that we are doing the work of the humanities, in digital form.* »¹⁰. Dans le deuxième type d'approche, l'informaticien est expert, et les données une fois générées sont offertes aux interprétations des chercheurs en humanités¹¹. Cette situation a été identifiée par Jean-Guy Meunier comme le paradoxe des humanités numériques, lorsque les experts techniques estiment que « les traitement interprétatifs manquent de rigueur et de scientificité »¹² et que les experts en humanités considèrent toute formalisation comme trop réductrice et peu capable de « rendre compte adéquate-

4. Frédéric Clavert et Valérie Schafer, “Les humanités numériques, un enjeu historique”, *Quaderni*-98 (févr. 2019), p. 33-49, DOI : 10.4000/quaderni.1417, p. 5

5. Association francophone des humanités numériques, *Humanistica*, fr-FR, URL : <http://www.humanisti.ca/> (visité le 15/06/2019)

6. DHBenelux 2019, *DHBenelux 2019*, en-US, URL : <http://2019.dhbenelux.org/> (visité le 15/06/2019)

7. THATCamp, *The Humanities and Technology Camp*, en, URL : <http://thatcamp.org/> (visité le 15/06/2019)

8. DARIOH, *Digital Research Infrastructure for the Arts and Humanities*, en-GB, URL : <https://www.dariah.eu/> (visité le 17/06/2019)

9. Alliance of Digital Humanities Organizations, *About*, URL : <http://adho.org/about> (visité le 15/06/2019)

10. Susan Schreibman, Raymond George Siemens et John Unsworth, *A new companion to digital humanities*, OCLC : 953120034, 2016, p.17

11. J.G. Meunier, “Le paradoxe des humanités numériques”...

12. *Ibid.*, p.6

ment de la complexité des objets traités »¹³. La recherche en humanités numériques est dès lors une démarche complexifiée par les questions de compréhension et d'interprétation posées par ses objets d'étude, « d'autant qu'il n'y a pas une, mais des communautés DH, qui distinguent notamment des approches anglo-saxonnes et européennes »¹⁴. Pour le Professeur Willard McCarthy¹⁵, ce n'est pas uniquement le champ qui doit être interdisciplinaire, mais les personnes qui le pratiquent qui devraient tendre à l'être, se tenir informées de ce qui se fait ailleurs pour enrichir leurs expériences de recherche¹⁶. Certains experts proposent une cartographie des champs de recherche :

La variété des approches en DH peut répondre à des interrogations en termes à la fois de recueils et gestion des sources de données, d'analyse et de représentation via des outils numériques - comme l'usage ou la constitution de bases de données, de cartographie de liens, d'étude sémantique ou lexicale etc. - ou encore de valorisation, notamment dans le cadre de l'histoire numérique publique.¹⁷

Si les démarches en humanités numériques sont multiformes et se distinguent de « tout un champ d'expérimentation croisant SHS et informatique (*literary / linguistic / humanities computing*) »¹⁸, les ancêtres de la discipline semblent communément acceptés, à l'instar du père Roberto Busa, jésuite pionnier de la numérisation et de l'analyse de corpus et de Franco Moretti, auteur de *Graphes, cartes et arbres. Modèles abstraits pour une autre histoire de la littérature, 2008*, qui reste une référence pour son concept de lecture distante.¹⁹ La *Text Encoding Initiative (TEI)*, créée en 1987, figure aussi en bonne position puisqu'elle place les données au centre de son étude et s'adapte aux évolutions des outils informatiques.²⁰ D'autres précurseurs communs aux projets de numérisation de masse pourraient figurer dans ce palmarès, notamment Paul Otlet, et Ted Nelson l'inventeur de la notion d'hypertexte.²¹ Les projets en humanités numériques s'offrent à la fois comme outil et objet d'étude, tant leurs histoires s'écrivent au présent.

[...] les humanités numériques ont permis l'émergence d'un vaste et riche patrimoine numérique, sources primaires de leur propre histoire que les historiens pourront très clairement exploiter, par l'hybridation de leurs outils

13. *Ibid.*

14. F. Clavert et V. Schafer, “Les humanités numériques, un enjeu historique”..., p.41

15. Professeur au sein du département des humanités numériques du King's College de Londres.

16. S. Schreibman, R. G. Siemens et J. Unsworth, *A new companion to digital humanities...*

17. F. Clavert et V. Schafer, “Les humanités numériques, un enjeu historique”..., p.35-36

18. Pierre Mounier, *Lou Burnard : Du Literary and linguistic computing aux Digital Humanities : retour sur 40 ans de relations entre sciences humaines et informatique*, fr-FR, Billet, URL : <https://leo.hypotheses.org/3764> (visité le 20/06/2019)

19. Alexandre Gefen, “Des humanités numériques en 2017”, *Mélanges de la Casa de Velázquez. Nouvelle série*-47-2 (nov. 2017), p. 315-318, DOI : 10.4000/mcv.7957

20. F. Clavert et V. Schafer, “Les humanités numériques, un enjeu historique”...

21. *Ibid.*

maîtrisés de longue date (la lecture proche) avec ceux des *digital humanities*, au premier rang desquels la lecture distante jouera un grand rôle.²²

Time Machine fait de la question de la numérisation de masse l'un de ces enjeux. Cœur de nombreux projets en humanités numériques, l'acte de numérisation articule plusieurs moments (choix de la collection, conversion des documents papiers en versions électroniques, traduction du format électronique en format binaire, reconstruction de la structure du document, identification des caractères etc.)²³ et dans une même entreprise voient se confronter les différents rôles attribués à l'expert des humanités ou à l'expert de l'informatique.

Ainsi, pour un expert des humanités, la constitution d'un corpus apparaîtra avant tout comme une activité de type herméneutique et l'ordinateur ne sera qu'un assistant, à l'inverse pour un expert de l'informatique, la numérisation apparaîtra avant tout comme l'application de technologies computationnelles d'une grande complexité.²⁴

Tout en s'inscrivant dans les démarches historiques des humanités numériques, Time Machine semble incarner les différents paradoxes de cette pensée académique au sein d'un même projet et promet de devenir à son tour objet d'étude de cette histoire.

Ainsi, au contact des humanités numériques, les historiens sont amenés à revisiter plusieurs aspects de la fabrique d'une histoire qui aujourd'hui articule productions individuelles et collectives, sous l'effet certes du numérique, mais aussi d'une culture du projet, qui innervent toutes les disciplines [...].²⁵

3.2 Le Laboratoire d'humanités numériques de l'EPFL

L'EPFL ouvre ses portes en 1853 sous le nom de l'École Spéciale de Lausanne. Au fil des années, l'école grandit, voyant de nouvelles spécialisations s'intégrer à son cursus (électricité, physique, architecture etc.), et d'une académie, devient université. Elle prend son nom actuel en 1969, année qui coïncide avec le début de son établissement sur le campus d'Écublens-Dorigny (Suisse). Réputés sélectifs, les cursus ne désemplissent pas et l'offre continue de croître. Dès 1991, la construction d'un parc scientifique hébergeant une dizaine de compagnies favorise le développement des activités de recherche. En 2003, année de son 150^{ème} anniversaire, l'école compte plus de 6000 étudiants. 2014 marque la création de deux nouveaux campus en Suisse avec l'ouverture de l'EPFL Valais²⁶ (dix

22. *Ibid.*, p.43

23. J.G. Meunier, "Le paradoxe des humanités numériques"...

24. *Ibid.*, p. 29

25. F. Clavert et V. Schafer, "Les humanités numériques, un enjeu historique"..., p. 42-43

26. EPFL, *EPFL Valais Wallis*, fr-FR, URL : <https://www.epfl.ch/about/campus/fr/valais-fr/> (visité le 13/06/2019)

laboratoires actifs dans les domaines de l'énergie, de l'environnement et de la santé) et de l'EPFL Fribourg²⁷ (« *smart living lab* » dédié aux recherches en technologies de la construction, bien-être et comportements, interaction et processus de conception, et systèmes énergétiques). En 2015, le nombre d'étudiants franchit la barre des 10'000²⁸.

En 2019, l'EPFL compte cinq facultés, trois collèges, plus de 11'000 étudiants, et quelques 210 compagnies constituent son parc scientifique²⁹. Son rayonnement s'étend à l'international, l'EPFL est en bonne position dans les classements universitaires européens et mondiaux.

Fondé en 2002, le Collège des Humanités (CDH) est un « lieu de convergence des sciences humaines et sociales au cœur de l'EPFL, basé sur le concept de POLY-perspective, reflétant la nécessité pour les futurs ingénieurs et scientifiques d'adopter une perspective pluraliste face aux enjeux auxquels ils sont confrontés »³⁰. Les perspectives s'organisent autour de quatre grandes thématiques : l'interdisciplinarité, la conscience globale (comprendre l'histoire des technologies pour mieux les faire évoluer), la citoyenneté et la créativité (engagement à travers l'art et la production artistique). La recherche au sein du CDH se compose de deux instituts, le *Digital Humanities Institute (DHI)* et l'*Institute for Area and Global Studies (IAGS)*, et d'un pôle soutenant des projets transdisciplinaires entre l'Université de Lausanne (UNIL) et l'EPFL : le *Collaborative Research on Science and Society (CROSS)*³¹. Le CDH coordonne également le programme d'enseignement en sciences humaines et sociales pour tous les étudiants de l'EPFL et le DHI propose un master en humanités digitales et est lié au programme doctoral en humanités digitales³².

Le DHLAB dans lequel nous avons effectué notre stage, constitue l'un des trois laboratoires du DHI, aux côtés du *Digital and Cognitive Musicology Laboratory (DCML)* et du *Laboratory for Experimental Museology (eM+)* et de plusieurs groupes thématiques de recherche (*Social Computing Group, Social Media*).³³

Le DHLAB est fondé en 2012 par le Professeur Frédéric Kaplan avec pour objectif de développer de nouvelles approches informatiques pour l'étude du passé et la découverte du futur. Le Professeur Kaplan est également le créateur du Master en humanités digitales du DHI et co-organisateur de la *Digital Humanities 2014*, conférence internationale de

27. Id., *EPFL Fribourg*, fr-FR, URL : <https://www.epfl.ch/about/campus/fr/epfl-fribourg/> (visité le 13/06/2019)

28. Id., *History of EPFL*, en-GB, URL : <https://www.epfl.ch/about/overview/overview/history-of-epfl/> (visité le 13/06/2019)

29. Id., *Innovation Park*, en-US, URL : <https://www.epfl-innovationpark.ch/community/companies/> (visité le 13/06/2019)

30. Id., *La vision du CDH : POLY-perspective*, fr-FR, URL : <https://www.epfl.ch/schools/cdh/fr/la-vision-du-cdh-poly-perspective/> (visité le 13/06/2019)

31. Id., *CROSS – Collaborative Research on Science and Society*, fr-FR, URL : <https://www.epfl.ch/schools/cdh/fr/recherche/cross/> (visité le 13/06/2019)

32. *Formations offertes au CDH*, fr-FR, URL : <https://www.epfl.ch/schools/cdh/fr/formations-offertes/> (visité le 15/06/2019)

33. EPFL.DHLAB, *DH Research*, URL : <https://www.epfl.ch/schools/cdh/research-2/dhi/dh-research/> (visité le 13/06/2019)

l'*Alliance of Digital Humanities Organizations (ADHO)*³⁴ à Lausanne, qui demeure l'un des plus grands rassemblements scientifiques du domaine.

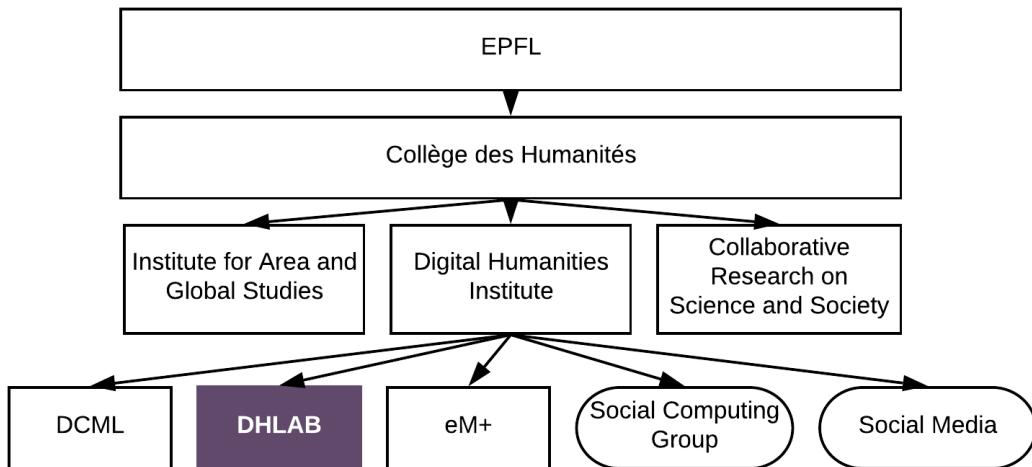


FIGURE 3.1 – Le DHLAB au sein de l'EPFL

3.3 Venice Time Machine

Né de la crainte de voir les informations non digitales tomber dans « les oubliettes de l'histoire »³⁵, initié en 2012 par l'EPFL et l'université Ca'Foscari de Venise, le projet vise à construire un modèle multidimensionnel de Venise et de ses évolutions, sur une période de 1000 ans : une machine à remonter le temps. Conscients dès le départ que les enjeux du projets nécessitent bien plus que de la puissance de calcul informatique, les initiateurs s'allient le secours d'historiens et d'archivistes pour mettre en place les processus de numérisation³⁶. La transformation des 80 kilomètres d'archives vénitiennes en un système d'informations nécessite en plus un grand travail d'indexation et de cartographie pour aboutir à ce qui se veut une version augmentée de *Google Maps*, offrant en plus un ancrage temporel³⁷ : un croisement entre histoire et *big data*³⁸.

34. Id., *DH Seminar Lecture 2019 - Prof. Frédéric Kaplan*, EPFL, mars 2019, URL : <https://tube.switch.ch/videos/9410934a> (visité le 18/05/2019)

35. Damien Dubuc, “Venice Time Machine, un canal à remonter le temps” (, déc. 2017), URL : https://www.lemonde.fr/tant-de-temps/article/2017/12/13/venice-time-machine-un-canal-a-remonter-le-temps_5229068_4598196.html (visité le 18/05/2019)

36. Alison Abbott, “The ‘time machine’ reconstructing ancient Venice’s social networks”, *Nature News*, 546–7658 (juin 2017), p. 341, DOI : 10.1038/546341a

37. WORLD.MINDS, *Frédéric Kaplan : The Venice Time Machine (2017 WORLD.MINDS Annual Symposium)*, URL : https://www.youtube.com/watch?time_continue=666&v=6brInBZ-jLk (visité le 18/05/2019)

38. RTS, *Remonter le temps à Venise - Vidéo*, fr, URL : <https://www.rts.ch/play/tv/mise-au-point/video/remonter-le-temps-a-venise?id=9020497> (visité le 18/05/2019)

Les archives numérisées par le projet proviennent de deux sources différentes³⁹ :

- **Les Archives d'Etat de Venise** : en 2018, 190'000 documents avaient été numérisés
- **La Fondation Cini** : la Fondation possède un million de photos d'art, en 2018 quelques 720'000 documents photographiques avaient été numérisés

Face à d'immenses corpus souvent manuscrits, les outils permettant la lecture automatique sont cruciaux. L'EPFL rejoint le projet européen *Recognition and Enrichment of Archival Documents (READ)*⁴⁰ poursuivant des objectifs similaires, et emploie une partie de ses équipes à développer des solutions d'apprentissage automatique ou machine learning⁴¹.

Avec le projet vénitien, l'EPFL souhaite développer, tester et démontrer l'efficacité d'une méthode appelée à se déployer avec Time Machine à travers l'UE⁴². Sans le succès du Venise Time Machine, il n'y a pas de Time Machine⁴³.

3.3.1 Développements technologiques

Le projet se poursuivant actuellement, de nombreux outils, techniques, méthodologies ont été spécialement dédiés à son avènement. En plus de la numérisation, il a fallu apprendre aux ordinateurs à automatiquement extraire certains types d'information, ce qui a impliqué un grand travail d'indexation manuelle. Tous ces éléments viendront enrichir la future interface de Time Machine et contribueront à la création de ce *big data* du passé. L'envergure de ce projet étant trop vaste pour en lister tous les composants, voici un aperçu de leurs variétés :

- **The Replica Scanner** : numériser massivement a induit à la création de divers scanners. Celui-ci, développé avec Adam Lowe de Factum Arte⁴⁴ est capable de numériser un document photographique recto verso toutes les quatre secondes.
- **DhCanvas** : moteur de recherche associé aux documents numérisés, permettant la recherche dans le document original et son contenu, fonctionnant comme un outil

39. Sandy Evangelista et Anne-Muriel Brouet, "Time Machine dans la course au FET Flagship européen" (, févr. 2018), URL : <https://actu.epfl.ch/news/time-machine-dans-la-course-au-fet-flagship-europe/> (visité le 20/06/2019)

40. *Recognition and Enrichment of Archival Documents / Projects / H2020 / CORDIS*, URL : <https://cordis.europa.eu/project/rcn/198756/factsheet/en> (visité le 17/06/2019)

41. nature video, *A virtual time machine for Venice*, URL : https://www.youtube.com/watch?time_continue=147&v=uQQGgYPRWfs (visité le 18/05/2019)

42. ARTE, *Venice Time Machine - History and Big Data (1/8)*, en, URL : <https://www.arte.tv/en/videos/075631-001-A/venice-time-machine-history-and-big-data-1-8/> (visité le 18/05/2019)

43. Frédéric Kaplan, *La cartographie en quatre dimensions Propos recueillis par Marc Frochaux*, fr, mai 2018, URL : <https://www.espazium.ch/fr/actualites/la-cartographie-en-quatre-dimensions> (visité le 18/05/2019)

44. Factum Arte S.L, *Factum Arte*, en, URL : <http://www.factum-arte.com/aboutus> (visité le 20/06/2019)

participatif de transcription. Les utilisateurs autorisés peuvent corriger les erreurs d'extraction⁴⁵.

- **Replica** : une interface de recherche spécialement dédiée aux fonds des images numérisées de la Fondation Cini, proposant une recherche visuelle (par motifs, position des personnages etc.) et textuelle des documents et permettant d'établir des connections entre des images⁴⁶.
- **Venice Scholar** : travail consacré aux ouvrages et journaux scientifiques vénitiens numérisés et dont les citations ont été automatiquement extraites, pour pouvoir étudier l'historiographie vénitienne⁴⁷.
- **DhSegment** : confronté à l'enjeu de traiter l'immense masse des images numérisées, ce système se base sur les réseaux de neurones artificiels pour en extraire automatiquement certaines parties (lignes, encadrés, zone de texte etc.)⁴⁸.



FIGURE 3.2 – Le scanner Replica à la Fondation Cini, © Copyright 2019 Factum Foundation

45. S. Evangelista et A.M. Brouet, “Time Machine dans la course au FET Flagship européen”...

46. *Ibid.*

47. *Ibid.*

48. EPFL.DHLAB, *Introduction — dhSegment documentation*, URL : <https://dhsegment.readthedocs.io/en/latest/intro/intro.html#use-cases> (visité le 20/06/2019)

3.4 Le projet Time Machine

Le 14 décembre 2016, Frédéric Kaplan, Professeur et occupant de la chaire d’humanités digitales de l’EPFL, publie dans le journal suisse *Le Temps*, le manifeste du projet Time Machine. L’intérêt du public et des institutions patrimoniales pour le projet Time Machine s’éveille et l’aventure est officiellement lancée ; Time Machine peut se lancer à la conquête de l’Europe et créer sa « machine à remonter le temps ».

L’Europe a inventé le web. Le web est devenu la matrice d’un monde nouveau. Les acteurs qui, les premiers, en compriront la logique, dominent aujourd’hui notre monde. Une trentaine d’années plus tard, le virage spatio-temporel de l’Internet, en plongeant l’information numérique dans un espace bien plus large, redéfinit les règles du jeu.

Le projet Time Machine peut donner à l’Europe la technologie de son renouveau : une occasion unique pour construire notre futur à partir de notre patrimoine commun, une occasion unique pour nous retrouver⁴⁹.



FIGURE 3.3 – Time Machine, logo © Copyright 2019 Time Machine

3.4.1 FET Flagships ou la recherche de financement

A travers l’initiative des *FET Flagships* ou « Initiatives-phare des technologies Futures et Emergentes », la Commission européenne mettait au concours un soutien financier d’un milliard d’euros réparti sur dix ans. Les projets éligibles se devaient d’être visionnaires, de proposer des solutions scientifiques et technologiques innovantes et de réunir

49. “L’Europe doit construire la première Time Machine” (, déc. 2016), URL : <https://www.letemps.ch/opinions/leurope-construire-premiere-time-machine> (visité le 18/05/2019)

sur le long-terme autour d'un objectif commun et d'une feuille de route aux solutions ambitieuses, des équipes de recherche interdisciplinaires^{50 51}.

L'ambition de Time Machine s'accordant avec les objectifs fixés par la Commission, ses initiateurs (*l'EPFL s'est alliée le concours de 32 partenaires pour développer et financer la proposition du projet, dont l'École Nationale des Chartes*) se sont lancés dans la course. L'histoire ne s'est pas faite sans embûches, Time Machine s'inscrivant sous le sigle de la culture, il fallut d'abord faire modifier les critères de sélection définis par l'UE qui s'opposaient aux candidatures issues du milieu⁵². En 2016, Time Machine est officiellement retenu comme candidat et en février 2019, le projet est classé premier parmi les six finalistes en lice pour la dernière ligne droite, avec cette fois une enveloppe d'un million d'euros et une année (à compter du 1er mars 2019) pour démontrer de l'organisation et de la faisabilité du projet.

Cette première victoire est nuancée, puisque coïncidant avec une année de votation, le programme de recherche européen *Horizon Europe*, doté d'une enveloppe de quelque cent milliards d'euros pour la période 2021-2027, peine à se stabiliser⁵³. Le projet des FET Flagships est abandonné par la Commission européenne sans que pour l'heure une solution concrète ait été proposée aux six projets concurrents⁵⁴.

Nous avons commencé notre stage en cette période d'incertitude financière, étant à la fois contraints par l'UE de planifier la mise en route du projet puisque le million a été perçu, mais sans certitude quant à l'ampleur finale du budget alloué.

3.4.2 Time Machine, quels objectifs ?

Dès 2014, le Professeur Frédéric Kaplan propose un modèle en forme de « champignon » pour expliciter l'ampleur des informations dont l'humain dispose pour évaluer une période donnée⁵⁵. L'image est porteuse, plus l'on se rapproche du présent, plus le cône s'élargit. Face à une société désormais digitale, le risque est grand de perdre pour toujours l'accès aux informations contenues uniquement sous formes physiques, les enjeux liés à la conservation de ces documents sont réels, mais comment parvenir à rétablir un équilibre

50. L'EPFL pilote déjà un projet doté d'un tel financement, le « Human Brain Project ».

51. Moriscl, *FET Flagships*, en, Text, déc. 2013, URL : <https://ec.europa.eu/programmes/horizon2020/en/h2020-section/fet-flagships> (visité le 22/06/2019)

52. RTS, *L'invité de La Matinale - Frédéric Kaplan, professeur en humanités digitales à l'EPFL - Radio*, fr, URL : <https://www.rts.ch/play/radio/linvite-e-de-la-matinale/audio/linvite-de-la-matinale-frederic-kaplan-professeur-en-humanites-digitales-a-lepfl?id=10247166> (visité le 18/05/2019)

53. *La crise révélée par la Time Machine*, fr, URL : <https://www.bilan.ch/opinions/fabrice-delaye/la-crise-revelee-par-la-time-machine> (visité le 22/06/2019)

54. Id., *Coup de frein de l'UE au projet de recherche "Time Machine" de l'EPFL*, fr, infoSport, mai 2019, URL : <https://www.rts.ch/info/sciences-tech/technologies/10435599-coup-de-frein-de-l-ue-au-projet-de-recherche-time-machine-de-l-epfl.html> (visité le 22/06/2019)

55. TED, *Frederic Kaplan : How I built an information time machine*, URL : https://www.youtube.com/watch?time_continue=296&v=2-Ev4rU27HY (visité le 18/05/2019)

entre l'information du passé et l'information du présent⁵⁶ ?

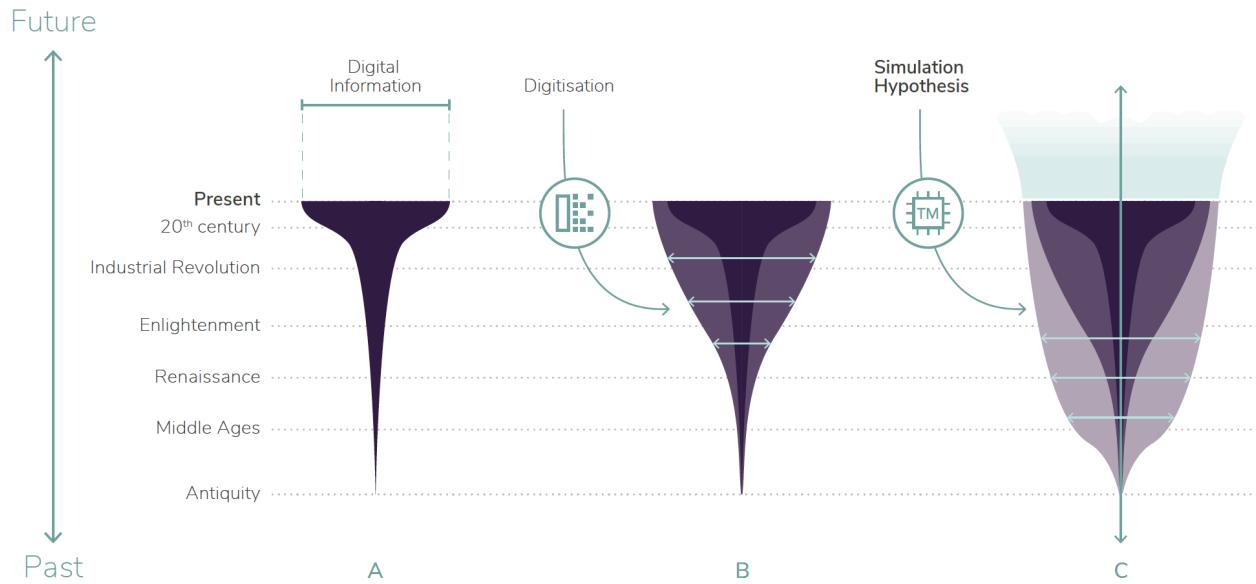


FIGURE 3.4 – Time Machine, Information's Mushroom © Copyright 2019 Time Machine

En visant comme résultat la création d'un graphe de données du passé, le *big data of the past*, impliquant le développement de processus, technologies et innovations nécessaires à sa réalisation, Time Machine poursuit différents objectifs⁵⁷ :

- Révolutionner les connaissances en intelligence artificielle et dans le domaine des sciences de l'information et de la communication, positionnant l'Europe comme meneur sur le marché de l'extraction et de l'analyse d'énormes jeux de données hétérogènes et complexes, induits par les activités et réalisations passées et présentes du genre humain.
- Offrir de nouvelles perspectives aux chercheurs en humanités et sciences sociales, leur permettant de positionner leurs objets d'étude à l'échelle plus large de l'histoire et de la culture européenne.
- Devenir moteur et acteur de l'*Open Science* et mettre à disposition les données du projet dans des dépôts ouverts et publics.
- Offrir un flux de connaissances à l'enseignement, encourageant les réflexions sur le long-terme et l'apprentissage de la pensée critique.
- Devenir un nouvel acteur économique, par la création d'emplois, services et produits appelés à influencer les secteurs clés de l'économie européenne.

56. La *factsheet* du projet, donnée en annexe B, offre une synthèse plus visuelle.

57. Time Machine Manifesto, 2019 URL : <https://documents.icar-us.eu/documents/2019/05/time-machine-manifesto.pdf>

3.4.3 Quelle(s) méthode(s) pour Time Machine ?

Créer un système d'information d'ampleur européenne basé sur un graphe de données du passé implique le déroulement de plusieurs étapes⁵⁸⁵⁹ et le déploiement de technologies nécessaires à leurs succès. Se basant sur les innovations et méthodologies du projet vénitien, des projets menés en parallèle par l'équipe du DHLAB, et de ceux conduits par les partenaires du futur réseau, Time Machine entend mener ses révolutions technologiques à différents niveaux⁶⁰ :

1. Numérisation massive des documents - archives - patrimoine bâti

Poursuivre le développement des outils de numérisation pour les documents (2D) et bâtiments (3D).

Depuis 2018, l'EPFL conjointement avec la fondation Cini et Factum Arte⁶¹ a créé un centre d'expérimentation et de formation aux techniques de numérisation à Venise : *ARCHiVe, Analysis and Recording of Cultural Heritage in Venice*⁶², diverses études sont en cours pour créer des scanners de nouvelles générations, mais également des programmes facilitant le suivi et l'automatisation des processus (traitement automatique de la post-production, ajout de métadonnées, indexation du texte). Les technologies d'apprentissage automatique ou machine learning sont utilisées par souci d'automatisation et de gain de temps.

Les innovations apportées par ce « modèle » vénitien seront amenées par Time Machine, avec l'aide des acteurs locaux du marché de la numérisation, à se répandre et se développer à travers l'Europe, créant un réseau Time Machine de centres de numérisation européen. Cette uniformisation des processus permettra l'harmonisation des coûts et la garantie de la qualité des données.

Une fois les données numérisées, leur stockage doit être assuré en vue de leur valorisation. Si un grand nombre d'institutions patrimoniales disposent de leurs propres serveurs, ce n'est pas forcément le cas des plus petites et ce n'était pas celui des archives vénitiennes. Une *Time Machine Box*⁶³ a été développée, visant à accompagner les partenaires dans leurs démarches de numérisation et leur offrir une alternative de stockage en local.

La numérisation du patrimoine bâti implique l'utilisation des techniques de la photogrammétrie et l'alignement de différents nuages de points. L'ambition de Time

58. *Ibid.*

59. F. Kaplan, *La cartographie en quatre dimensions Propos recueillis par Marc Frochaux...*

60. Les partenaires étant actuellement plus de 300, il est impossible de rendre justice dans le cadre de ce travail à la multitude des projets en cours appelés à contribuer à la mise au point de Time Machine.

61. F. A. S.L, *Factum Arte...*

62. *ARCHiVe Analysis and Recording of Cultural Heritage in Venice*, it-IT, URL : <https://www.cini.it/istituti-e-centri/archive-analysis-and-recording-of-cultural-heritage-in-venice> (visité le 23/06/2019)

63. EPFL.DHLAB, *Home / TimeMachineBox*, URL : <https://www.timemachinebox.eu/> (visité le 18/05/2019)

Machine étant à l'échelle de villes, les technologies actuelles devront être améliorées. Le projet *ScanVan (2016-2020)*, conduit par le DHLAB, a pour objectif le développement d'un véhicule capable de numériser les villes et de créer automatiquement des modèles 3D de ces territoires⁶⁴.

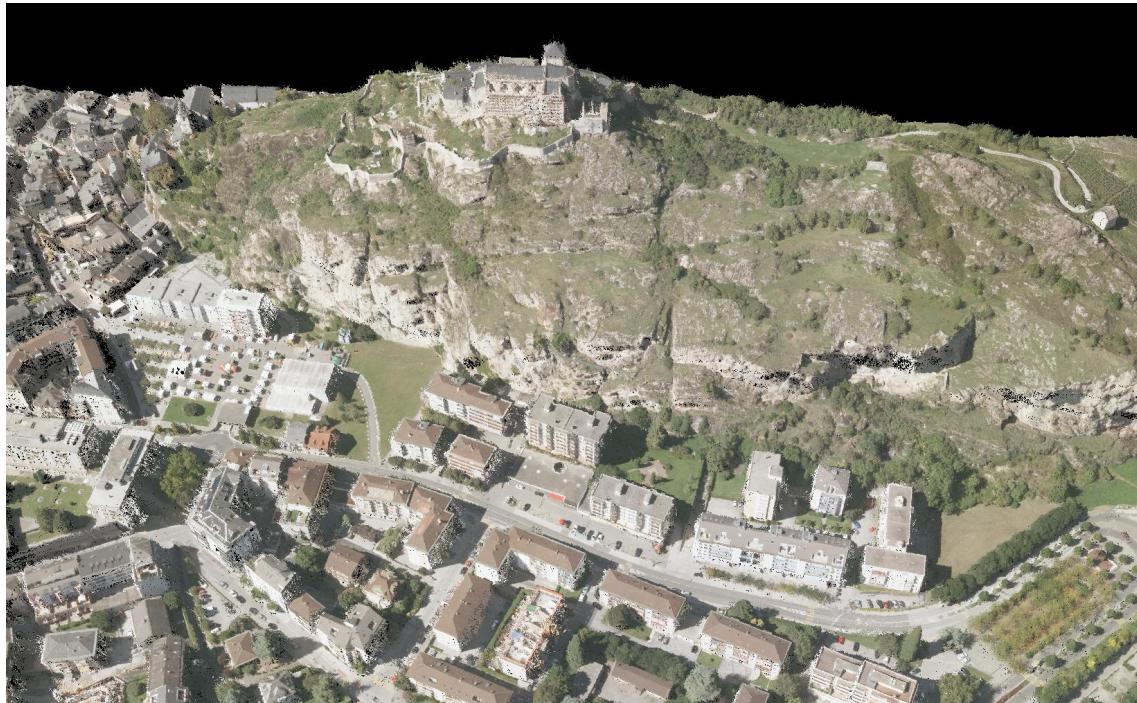


FIGURE 3.5 – Ville de Sion, compilée par ScanVan © Copyright EPFL DHLAB

Iconem⁶⁵, start-up partenaire du projet, s'est spécialisée dans la numérisation de sites patrimoniaux en danger, grâce à des drones et à des processus de photogrammétrie. Par la mise en commun de ces développements technologiques, Time Machine vise à la création et à l'implémentation d'un système des plus efficents.

2. Extraction et indexation des informations contenues

On peut déterminer différentes étapes dans les traitements d'un document après sa numérisation et la liste tend à s'agrandir. On parle notamment : de l'automatisation des processus de repérages du texte ou des motifs au sein d'une image, de la compréhension de l'architecture du document par le biais des technologies du traitement automatique du langage naturel, de la reconnaissance d'entités nommées, ou encore des technologies liées à la reconnaissance optique de caractères.

La plupart des projets du DHLAB conduisent des recherches dans ces domaines, c'est le cas notamment de *DHSegment*⁶⁶ cité pour le projet vénitien ou du projet *Impresso*

64. Id., 2016 – 2020 *ScanVan (FNS)*, en-GB, URL : <https://dhlab.epfl.ch/page-96350-en-html/page-152926-en-html/> (visité le 23/06/2019)

65. Iconem, *Iconem*, URL : <http://iconem.com/en/> (visité le 23/06/2019)

66. EPFL.DHLAB, *Introduction — dhSegment documentation...*

- *Media monitoring of the past (2017-2020)*⁶⁷, visant à automatiser l'extraction des informations de 200 ans d'archives de journaux afin d'en faciliter la recherche et la découverte.

Les différentes technologies déployées dans le cadre de ces deux projets seront amenées à enrichir les futurs composants de la Time Machine. Il n'existe actuellement qu'un nombre limité d'applications d'apprentissage profond ou deep learning pour le traitement des données culturelles et patrimoniales. Time Machine entend développer un réseau de chercheurs et de projets visant à améliorer cet existant⁶⁸.



FIGURE 3.6 – Impresso, visuel © Copyright 2019 EPFL, DHLAB

3. Extrapolation - simulation des données depuis d'autres sources

L'une des dimensions caractéristique des *big data*, est celle dite paradigmatic. Le nombre de données permettent elles-mêmes de déduire de nouvelles formes de sciences et de faire des découvertes⁶⁹. Ce processus s'apparente à celui mis en place par les moines copistes ou historiens confrontés à devoir « combler les trous » de l'histoire et faire communiquer entre elles des sources différentes. Ce travail suscite depuis longtemps de nombreux débats sans qu'il existe de normes sur la manière de

67. Id., 2017-2020 *Impresso* (FNS), en-GB, URL : <https://dhlab.epfl.ch/page-96350-en-html/page-150782-en-html/> (visité le 23/06/2019)

68. Dans le cadre du stage, nous avons produit différents états de l'art sur les technologies déployées dans Time Machine, ces informations sont contenues dans la feuille de route, figurant sur la clé USB accompagnant ce mémoire.

69. F. Kaplan et Isabella di Lenardo, “Big Data of the Past”, *Frontiers in Digital Humanities*, 4 (mai 2017), DOI : 10.3389/fdigh.2017.00012.

procéder. Time Machine soutient à priori qu'il faut veiller à précisément documenter ces techniques et les moyens déployés pour parvenir à cette « redocumentation » du passé⁷⁰. Face aux immenses jeux de données que sont les *big data*, les machines ont un rôle important à jouer dans la détermination du plus petit dénominateur commun capable de réunir différentes interprétations de l'histoire⁷¹.

Basé sur les méthodologies de l'intelligence artificielle, trois moteurs de simulation, dont un moteur d'inférence permettront à Time Machine de déduire de nouvelles informations des données numérisées.

Les moteurs d'inférence ou *Inference Engine* étaient un champ d'étude en intelligence artificielle très populaire ; les technologies liées à l'apprentissage profond ou deep learning ont quelque peu réduit l'intérêt des chercheurs en la matière⁷². Time Machine entend relancer cette recherche et utiliser les données produites pour accroître son graphe de données du passé.

4. Valorisation des données afin de transformer le patrimoine en un atout pour les industries et l'économie européenne

Afin de faciliter la découverte des données assemblées par Time Machine, une seule interface sera utilisée pour leurs valorisations, rassemblant en back-office tous les processus développés pour le projet⁷³.

Une première version de la future interface existe et est en cours de développement au sein du DHLAB. Cette dernière sera amenée à changer énormément au fil des avancées du projet afin de devenir un moteur de recherche diachronique : Diamond⁷⁴. Inclure la temporalité au sein du futur moteur de recherche implique pour Time Machine de favoriser les recherches menées sur la création de modèle 4D, actuellement dévouées à de petites échelles sans faire l'objet de standardisation⁷⁵.

70. *Ibid.*

71. *Ibid.*

72. Dans le cadre du stage, nous avons produit différents états de l'art sur les technologies déployées dans Time Machine, ces informations sont contenues dans la feuille de route, donnée en annexe sur la clé USB accompagnant le présent mémoire

73. EPFL.DHLAB, *DH Seminar Lecture 2019 - Prof. Frédéric Kaplan...*

74. Time Machine, *Diamond*, URL : <https://diamond.timemachine.eu/> (visité le 18/05/2019)

75. Dans le cadre du stage, nous avons produit différents états de l'art sur les technologies déployées dans Time Machine, ces informations sont contenues dans la feuille de route, figurant sur la clé USB accompagnant ce mémoire.

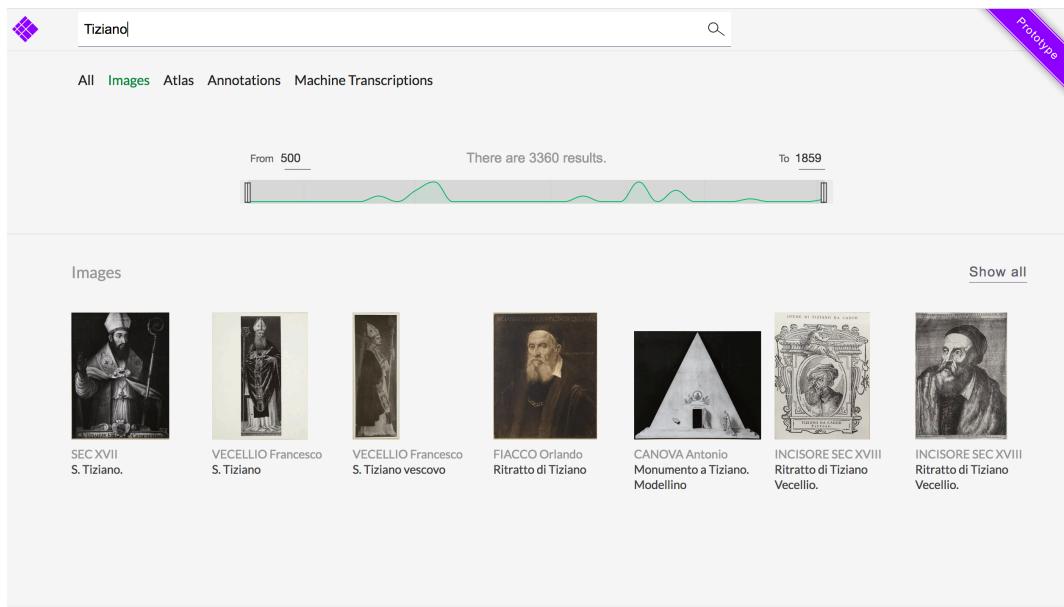


FIGURE 3.7 – Interface de Diamond, capture d'écran © Copyright 2019 Time Machine

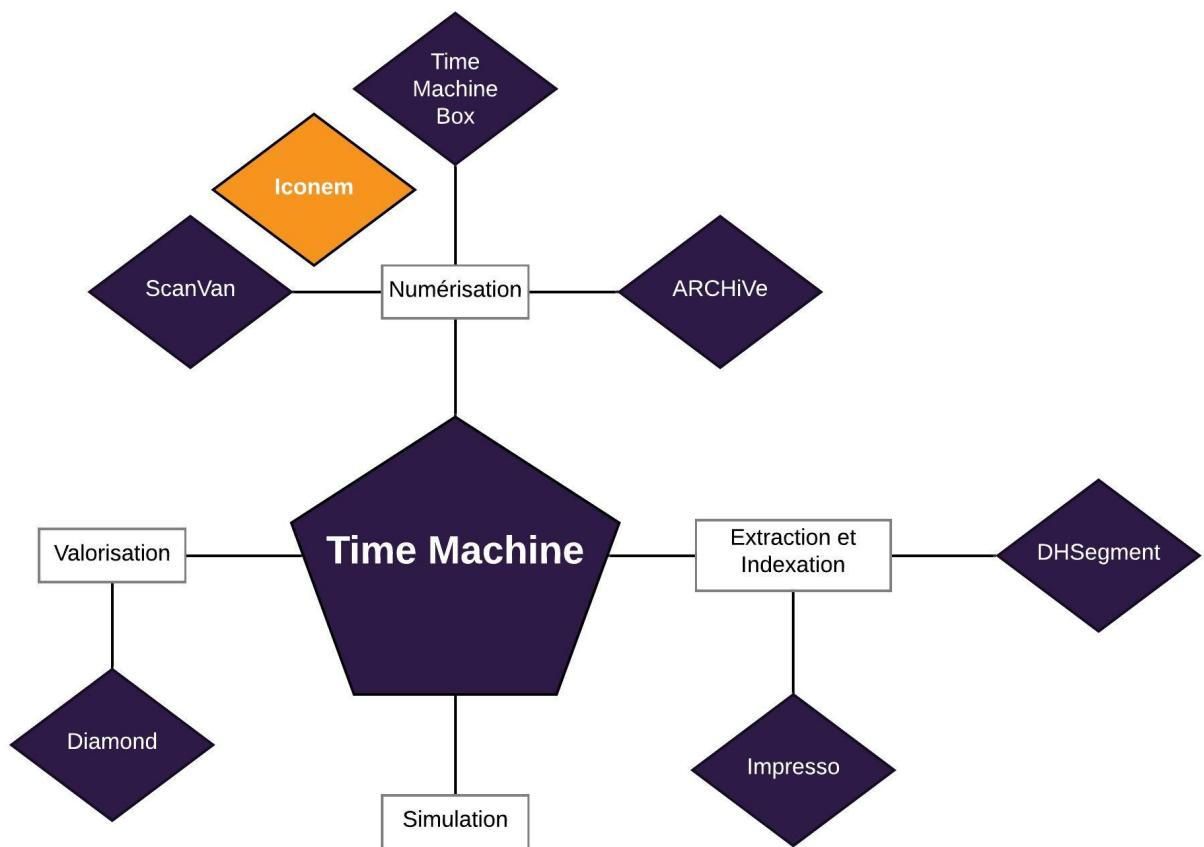


FIGURE 3.8 – Résumé visuel des projets et technologies évoqués

3.4.4 Collaboration public-privé de partenaires européens

Des 32 partenaires initialement impliqués dans Time Machine, le consortium s'est désormais agrandi à plus de 300 institutions réparties dans 34 pays et les chiffres continuent de croître. Notre stage nous a amené à en côtoyer un grand nombre⁷⁶ et offert la possibilité de questionner leurs motivations de manière informelle. De nos observations sur le terrain et de part nos activités de stagiaire, nous proposons d'établir certaines typologies, afin de mieux appréhender la multiplicité des profils et des attentes à l'encontre du projet. N'ayant pu mener de réelle étude quantitative, nous ne prétendrons pas être exhaustif :⁷⁷.

— Les institutions culturelles et patrimoniales

Composées de représentants des GLAM, ces dernières sont à la fois intéressées par les traitements de numérisation appliqués à leurs données et par l'exploitation des résultats pour leurs propres besoins de valorisation auprès de leurs publics-cibles.

— Les réseaux ou consortiums

Composés d'associations d'envergure actives dans le milieu du patrimoine, à l'instar d'Europeana⁷⁸ ou d'Icarus⁷⁹, ces dernières sont intéressées à intégrer un réseau qui leur permet à la fois de transcender leurs domaines d'activités, de rendre visibles leurs actions, d'accroître leurs membres et de participer au déploiement d'innovations technologiques applicables au sein de leurs communautés.

— Les laboratoires de recherche ou universités

Composés de représentants des sciences humaines et sociales aussi bien que des sciences dures, ces centres de recherches sont motivés à l'idée de prendre une part active dans les processus d'élaboration des différents composants techniques et de traitement de ce *big data* du passé. Ils sont également intéressés à déployer les résultats dans leurs missions d'enseignement.

— Les start-ups et entreprises privées

Actives dans les domaines innovants impliqués dans Time Machine, ces dernières sont intéressées à développer de nouveaux services et modèles d'entraînement grâce aux données du graphe du passé et à valoriser ou accroître leurs compétences au sein d'un réseau de grande envergure.

— Les futurs exploitants

76. Notamment lors de la conférence à Amsterdam, d'un séjour à Venise pour découvrir les coulisses du projet vénitien et des contacts nécessaires à la création de la feuille de route.

77. Id., *Introduction*, en, URL : <https://timemachine.eu/> (visité le 18/05/2019)

78. *Europeana Bot (@EuropeanaBot) / Twitter*, en, URL : <https://twitter.com/europeanabot> (visité le 24/07/2019)

79. ICARUS, *ICARUS / International Centre for Archival Research*, en-GB, URL : <https://icarus.eu/en/> (visité le 17/07/2019)

Le graphe une fois constitué laissera la possibilité à un grand nombre d'acteurs de créer leurs propres plateformes ou services basés sur ce *big data* du passé. Nous retrouvons ici, aux côtés des villes et des représentants des régions (le *big data* faisant écho au mouvement des *smart cities*⁸⁰), les industries culturelles et créatives, à l'instar du fabricant de jeu vidéo Ubisoft⁸¹, les industries du tourisme, de la recherche et de l'éducation, tous intrigués par la perspective de construire de nouveaux lendemains grâce à ces données du passé.

3.4.5 Un réseau de Time Machines locales

A l'instar de Venise, d'autres villes n'ont pas attendu la fin du processus de financement pour se lancer dans la construction de leurs propres Time Machines locales. Poursuivant des objectifs divers en fonction de leurs particularités régionales et de la temporalité de leurs archives, ces initiatives sont amenées à constituer le centre du futur réseau Time Machine, qui se veut un réseau de Time Machines locales (appelées à jouer un rôle similaire à celui des agrégateurs), composées de divers projets aux spécificités régionales, sous la gouvernance de la future *Time Machine Organisation*. Dans un contexte au financement incertain, elles montrent également que des solutions financières locales et régionales peuvent être trouvées⁸², dont le futur réseau Time Machine^{83 84} n'a pas encore fini d'explorer toutes les possibilités. Nous citerons ici quelques exemples parmi d'autres :

1. Antwerp Time Machine (1500 - 2000)⁸⁵
2. Amsterdam Time Machine (1550 - 2000)⁸⁶
3. Budapest Time Machine (1680-1990)⁸⁷
4. Dresden Time Machine (1200 - 2000)⁸⁸
5. Paris Time Machine (1000 - 2000)⁸⁹

80. Une ville intelligente utilise les données collectées sur internet pour mieux gérer et planifier l'utilisation de ses ressources.

81. Ubisoft, *Welcome to the official Ubisoft website*, URL : <https://www.ubisoft.com/en-us/> (visité le 17/07/2019)

82. RTS, *Coup de frein de l'UE au projet de recherche "Time Machine" de l'EPFL...*

83. Time Machine, *Local Time Machines / Time Machine Europe*, URL : <https://timemachine.eu/time-machines/> (visité le 23/06/2019)

84. La *factsheet* du projet, donnée en annexe B, présente une liste des Times Machines locales

85. University of Antwerp, *Antwerp Time Machine*, URL : <https://www.uantwerpen.be/en/projects/antwerp-time-machine/> (visité le 23/06/2019)

86. Amsterdam Time Machine, *Amsterdam Time Machine*, en-US, URL : <https://amsterdamtimemachine.nl/> (visité le 23/06/2019)

87. Hungaricana, *Budapest Time Machine*, URL : <https://hungaricana.hu/en/budapest-idogep/> (visité le 23/06/2019)

88. *Dresden Time Machine - Google Search*, URL : <https://www.google.com/search?client=firefox-b-d&q=Dresden+Time+Machine> (visité le 23/06/2019)

89. L'École Nationale des Chartes est membre du projet.



FIGURE 3.9 – Time Machine, visualisation du réseau © Copyright 2019 Time Machine

Deuxième partie

Exemples de projets de numérisation de masse

Chapitre 4

Différentes typologies de projets

Afin de mieux appréhender les réponses des projets de numérisation de masse aux enjeux de la numérisation et de mieux comprendre leurs similitudes et divergences, nous allons en présenter quatre encore actifs aujourd’hui. Pour rappel, nous exposerons dans la troisième partie de ce mémoire, les solutions proposées pour Time Machine. Suivant les recommandations de Margaret Coutts, nous distinguons deux typologies parmi ces exemples¹.

Typologie 1 : Les premières initiatives présentées (*Google Books* et *HathiTrust*), ont pour objectif de rassembler en ligne des corpus massifs de données numérisées. Aux côtés d’autres initiatives que nous ne détaillerons pas dans ce mémoire, telles que *Microsoft Live Book Search*² ou *the Open Content Alliance*³, ils ont fortement contribué à l’accroissement des données numérisées et au développement de l’expertise technique nécessaire à la gestion de telles collections⁴.

Typologie 2 : Les deuxièmes, bien que soutenant des entreprises de numérisations au sein de leur réseau, visent surtout à rassembler les données numérisées par de multiples institutions et les mettre à disposition via une interface commune : ce sont les agrégateurs (Europeana, *DPLA*). Ces agrégateurs se caractérisent par une architecture distribuée, terme désignant un réseau informatique dont l’ensemble des ressources ne se trouvent pas au même endroit ou sur la même machine, mais sont stockées sur les serveurs de leurs institutions⁵.

Nous ne parlerons pas d’une autre forme d’initiatives visant à rassembler en ligne

1. M. Coutts, *Stepping away from the silos...*

2. Le projet s’étant arrêté en 2008, le contenu est désormais accessible à travers la plateforme d’*Internet Archive* : <https://archive.org/index.php>

3. Initié en 2005, ce groupe a rassemblé des institutions privées (Yahoo!), et publiques (*Perseus Digital Library*) pour proposer une alternative aux objectifs commerciaux de Google. Ils étaient administrés par *Internet Archive*, dont le fondateur Brewster Kahle est un militant pour l'accès universel à la connaissance via la numérisation et un opposant au projet de Google : <https://archive.org/details/opencontentalliance>. Le consortium ne semble plus actif depuis 2009.

4. *Ibid.*, p.78.

5. *Ibid.*, p.79.

des corpus de données numérisées, les *shadow libraries*. Ces bibliothèques « de l'ombre », sont devenues des infrastructures de numérisation à grande échelle et proposent un accès gratuit et souvent illégal à des livres et articles académiques⁶. Caractérisées par une construction fragile voire éphémère, elles constituent un phénomène digne d'intérêt, mais que nous avons choisi de ne pas aborder afin de ne pas trop alourdir notre mémoire.

Si les typologies proposées par Margaret Coutts s'articulent autour de la construction des projets, elles font toutefois écho à celles mentionnées dans les sections 1.2 et 2.1, qui évoquaient des critères différenciant initiatives publiques et privées.

Pour pouvoir mieux comprendre les débats autour de ces efforts de catégorisation ou de classement des projets de numérisation de masse⁷, nous nous intéresserons également à leurs origines et motivations. Nous proposons un résumé comparatif des différentes réponses apportées aux enjeux de la numérisation et une brève conclusion à ce débat dans le chapitre 5.

4.1 Rassembler une masse de données numérisées

4.1.1 Google Books

Le projet initié en 2004 sous le nom de *Google Print*, qui deviendra *Google Partner Program*, a pour objectif de rendre accessible tout ce qui a été écrit par l'homme sans distinctions⁸. Les fondateurs de l'entreprise Larry Page et Sergey Brin annoncent à la foire du livre de Francfort la création d'une nouvelle plateforme d'édition pour permettre aux auteurs et éditeurs de commercialiser en ligne leurs ouvrages.

En complément de la création de cette plateforme d'édition, le *Google Library Project*, qui vise à numériser sur dix ans, quelque 15 millions d'ouvrages à la fois libres et sous droit est lancé la même année. Cinq bibliothèques de langue anglaise sont partenaires de cette première étape du projet : celle de l'Université du Michigan, Stanford, Harvard, Oxford (Bodleian) et la bibliothèque publique de New York^{9 10}.

Afin d'uniformiser les deux programmes qui travaillent à la réalisation du même objectif, le *Google Partner Program* et le *Google Library Project*, changent de nom en 2005 pour se rassembler derrière la bannière de *Google Books*.

Google retire de nombreux avantages de son projet de création de cette future bibliothèque universelle : des données pour contribuer à la réalisation d'une intelligence

6. N. B. Thylstrup, *The politics of mass digitization...*

7. E. Jones, “The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative...”, p.245

8. M. Thelle et N. Bonde Thylstrup, “Persuasive territories in European cultural politics...”.

9. N. B. Thylstrup, *The politics of mass digitization...*

10. A. Weiss et Ryan James, “Assessing the coverage of Hawaiian and Pacific books in the Google Books digitization project”, *OCLC Systems & Services : International digital library perspectives*, 29–1 (févr. 2013), p. 13-21, DOI : 10.1108/10650751311294519.

artificielle capable de comprendre toute la symbolique humaine, de nouveaux espaces publicitaires, une offre commerciale de livres numérisés¹¹.

La mise en œuvre du projet va très vite se heurter à des barrières d'ordre légal (comment numériser en restant dans la légalité du droit d'auteur américain), technique (quels moyens déployer pour numériser à grande échelle et comment valoriser et organiser ces données) et politique (comment dépasser les limites de fonctionnement territoriales et régionales).

Les éditeurs initialement du côté de Google vont prendre peur en constatant que ce dernier numérise également les ouvrages des bibliothèques. S'ensuivront différents procès et le lancement des débats liant droit d'auteur et numérique. Toutefois Google gagnera successivement ces derniers, grâce au principe du *fair use*¹² états-unien, contribuant à créer les bases légales légiférant la circulation des biens culturels et patrimoniaux en ligne^{13 14}. Google a compris avant tout le monde que « les barrières techniques classiques à un contournement du droit d'auteur tombaient avec le numérique¹⁵ ».

Il semble que la firme ne dispose pas d'équipe chargée de vérifier les droits d'auteur et qu'en cas de doute, elle préfère directement limiter l'accessibilité. L'utilisateur est informé par des pages introducives multilingues sur les droits d'usage de chaque ouvrage¹⁶, ces derniers figurent également dans les métadonnées.

Google Books propose une recherche plein-texte gratuite par mots-clés¹⁷, limitée par quatre différentes vues en fonction de l'accessibilité des documents. Un accès au texte intégral (le livre peut être téléchargé), un accès à quelques pages donnant un aperçu de l'ouvrage (en accord avec l'éditeur, avec un lien permettant de l'acheter), un accès aux métadonnées et à quelques phrases permettant de comprendre le contexte du mot-clé de la recherche (lorsque la license n'est pas claire) ou un accès uniquement aux métadonnées (lorsque l'ouvrage n'a pas été numérisé)^{18 19}.

11. *Numérisation du patrimoine...*

12. Pour rappel, le *fair use* autorise cinq exceptions au droit d'auteur : le droit à opérer une copie privée, le droit de citation, le droit au pastiche ou à la caricature, l'usage pour l'enseignement, l'usage pour la recherche.

13. I. Xie et K. K. Matusiak, *Discover digital libraries...*

14. V. Stobo, K. Patterson, K. Erickson, *et al.*, ““I should like you to see them some time”...”.

15. *Numérisation du patrimoine...*, p.37.

16. Kalev Leetaru, “Mass book digitization : The deeper story of Google Books and the Open Content Alliance”, *First Monday*, 13 (oct. 2008), DOI : 10.5210/fm.v13i10.2101.

17. *Ibid.*

18. Anna Lauren Hoffmann, “Google Books, Libraries, and Self-Respect : Information Justice beyond Distributions”, *The Library Quarterly*, 86–1 (janv. 2016), p. 76-92, DOI : 10.1086/684141.

19. A. Weiss, *Using massive digital libraries...*

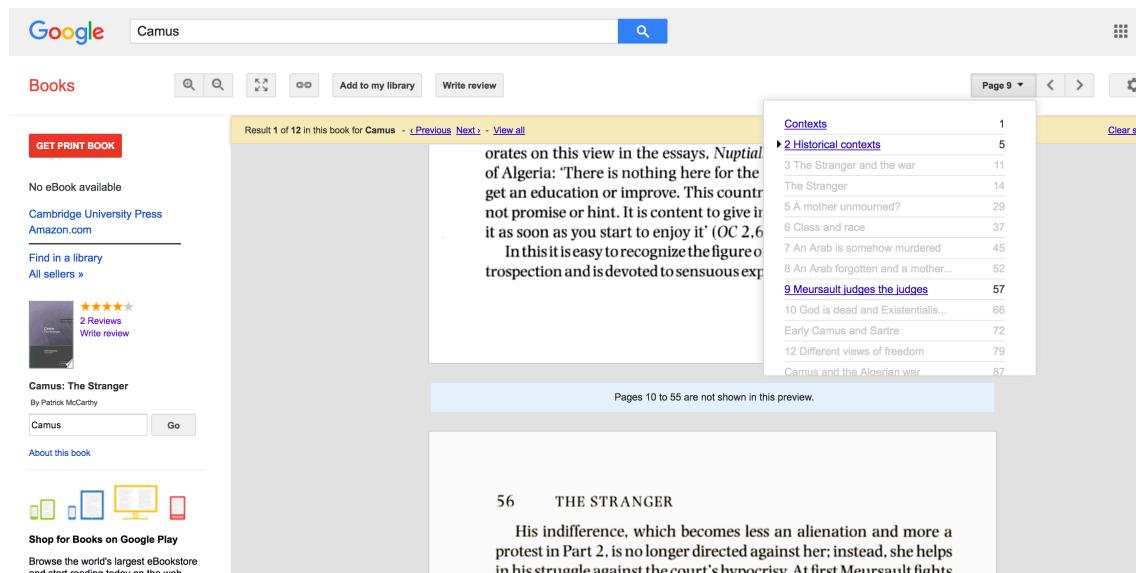


FIGURE 4.1 – Affichage d'un résultat *Google Books*, capture d'écran

Passés les débuts rapides et presque « agressifs » de l'initiative, le droit d'auteur est néanmoins avancé pour justifier le récent ralentissement de l'entreprise et une nouvelle préférence pour la numérisation d'ouvrages libres de droit^{20 21 22}.

Les débuts sont caractérisés par de nombreux problèmes relatifs à la qualité des métadonnées et des images²³. Les langues non romanes sont sous-représentées et les spécificités propres à d'autres histoires littéraires (telles que celles japonaises où les livres sont construits sur d'autres logiques que celles européennes) sont difficilement prises en compte dans les processus de numérisation²⁴. Les biais dans la construction des collections sont hérités des bibliothèques partenaires²⁵. De nombreux problèmes sont également constatés concernant la gestion des ouvrages en plusieurs volumes. En vérité, par rapport aux standards élevés des bibliothèques, *Google Books* a du mal à faire concurrence. Cela ne signifie pas qu'il réussit moins bien son objectif de desservir le plus grand nombre, puisqu'il permet d'accéder aux recherches au sein de l'environnement Google.

Google ne partage pas publiquement les processus et les algorithmes de traitement de ces données²⁶. La firme se charge du transport des documents vers le lieu de numérisation, puis restitue les données enrichies une fois numérisées, en garantissant leur intégrité.

20. N. B. Thylstrup, *The politics of mass digitization...*

21. A. Weiss, "Examining Massive Digital Libraries (MDLs) and Their Impact on Reference Services" ...

22. *Viewpoint - What's Next for Google Books*, URL : <http://www.infotoday.com/IT/sep16/Quint--Whats-Next-for-Google-Books.shtml> (visité le 03/07/2019).

23. A. L. Hoffmann, "Google Books, Libraries, and Self-Respect..." .

24. A. Weiss, "Examining Massive Digital Libraries (MDLs) and Their Impact on Reference Services" ...

25. A. Weiss et R. James, "Assessing the coverage of Hawaiian and Pacific books in the Google Books digitization project" ...

26. A. L. Hoffmann, "Google Books, Libraries, and Self-Respect..." .

Certaines bibliothèques témoignent d'un rythme de numérisation de quelques 5000 livres par mois²⁷.

La préservation sur le long-terme n'étant pas sa priorité, la firme remet tout de même une copie des fichiers aux institutions afin qu'elles puissent y veiller²⁸, pour autant que ces fichiers demeurent stockés sur les serveurs de l'institution²⁹.

Contrairement aux initiatives publiques, Google n'a pas le souci d'inclure l'usager dans le déploiement de *Google Books*. Il a démontré qu'il était capable de réagir et d'intégrer certaines critiques (notamment celles levées dans les premières années du projet qui reprochaient la mauvaise qualité de la numérisation), mais n'est pas prêt à ouvrir son projet plus largement aux besoins des individus³⁰, qui sont limités au rôle de critiques. Les développements technologiques semblent être prioritaires³¹.

Sa quête de profit le pousse également à limiter les usages commerciaux faits par la suite et prévient le référencement des titres numérisés par d'autres moteurs de recherche^{32 33}. La firme semble cependant largement autoriser les usages pour la recherche³⁴.

L'initiative de *Google Books* ne laisse personne indifférent, organisant dès le départ le monde en deux camps, celui des partenaires et celui des opposants. Google est entre-temps devenu un géant de l'information, dont le monopole et le non-respect de la protection des données des utilisateurs suscitent autant la crainte que l'admiration³⁵. De ce clivage va naître la plupart des projets de numérisation de masse du 21^e siècle³⁶.

En 2018, quelques **30 millions** de livres ont été numérisés par cette initiative³⁷ et le fonds continue de croître.

27. Association pour le patrimoine naturel et culturel du canton de Vaud, *Patrimoine numérique, numérisation du patrimoine...*

28. K. Leetaru, "Mass book digitization...".

29. Association pour le patrimoine naturel et culturel du canton de Vaud, *Patrimoine numérique, numérisation du patrimoine...*

30. E. Jones, "The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative...".

31. A. L. Hoffmann, "Google Books, Libraries, and Self-Respect...".

32. F. Clavert et Serge Noiret, *L'histoire contemporaine à l'ère numérique Contemporary history in the digital age*, OCLC : 894315404, 2013, URL : <http://site.ebrary.com/id/10819452> (visité le 25/06/2019).

33. A. Moatti, "Bibliothèque numérique européenne : de l'utopie aux réalités" ...

34. K. Leetaru, "Mass book digitization...".

35. F. Clavert et S. Noiret, *L'histoire contemporaine à l'ère numérique Contemporary history in the digital age...*

36. N. B. Thylstrup, *The politics of mass digitization...*

37. *Ibid.*

Enjeux	Réponses de Google
Amener différents acteurs à collaborer	Chaque contrat est négocié au cas par cas, mais Google impose certaines règles. Un contrat avec Google n'empêche pas la poursuite d'autres projets de numérisation mais prévient certains usages commerciaux.
Financement et partenariats public-privé	Les profits générés par Google suffisent à couvrir les frais des projets de numérisation. Nul besoin de trouver d'autres partenariats financiers.
Droit d'auteur	Google a gagné plusieurs procès contre la guilde des auteurs aux États-Unis. Toutefois, ces attaques incessantes sont souvent évoquées comme raison aux choix de numériser des ouvrages libres de droit, caractérisant les plus récentes initiatives.
Sortir des silos : enjeux techniques	Déploiement de centres de numérisation et dépôt de divers brevets visant à accélérer les processus (scanners, mélodie aidant à la concentration, automatisation etc.). Si Google est plutôt transparent sur les contrats de numérisation, il protège farouchement les technologies associées, les centres de numérisation ne sont pas visitables. Comme le projet est inclus dans l'environnement de recherche le plus utilisé, les ouvrages numérisés bénéficient d'une grande visibilité et sont accessibles dans plusieurs langues.
Sortir des silos : enjeux sur le contenu	Alors même que le projet a pour objectif de tout numériser, sans sélection préalable, certains choix ont été faits, qui le sortent de cette neutralité. En choisissant les plus grands et prestigieux établissements culturels, Google a reproduit leurs biais, privilégiant les ouvrages scientifiques et classiques au détriment de la littérature populaire. Le projet se consacre également à la seule numérisation de livres et privilégie les partenariats avec des institutions dont les ouvrages sont en langue romane.
Stockage sur le long-terme - préservation	La démarche de Google est une mise à disposition. La qualité des données est trop variable pour prétendre travailler à leur préservation, cette responsabilité est laissée aux institutions partenaires.

TABLE 4.1 – Les réponses de *Google Books* aux enjeux de la numérisation

4.1.2 HathiTrust

Initié en 2006 sous l'impulsion de l'université du Michigan, qui propose aux autres bibliothèques partenaires de *Google Books*³⁸ de développer un dépôt commun pour le stockage sur le long-terme des ouvrages numérisés³⁹. Les missions d'*HathiTrust* sont de contribuer à la recherche, l'éducation et le bien commun en rassemblant, organisant, préservant et rendant accessible, de manière collective et collaborative, les données de la connaissance humaine⁴⁰. Ainsi, au-delà de l'objectif de préservation, l'initiative se soucie de leur accessibilité dans le respect du cadre légal, afin de répondre aux besoins de la communauté scientifique : « *It represents an example of a « light archive », meaning, that the repository also functions as a digital library and provides access to some of their collections.* »⁴¹ Le nom de la plateforme fait écho aux ambitions de ce projet, puisque *Hathi* signifie éléphant en indien, animal connu pour sa mémoire, force et sagesse⁴².

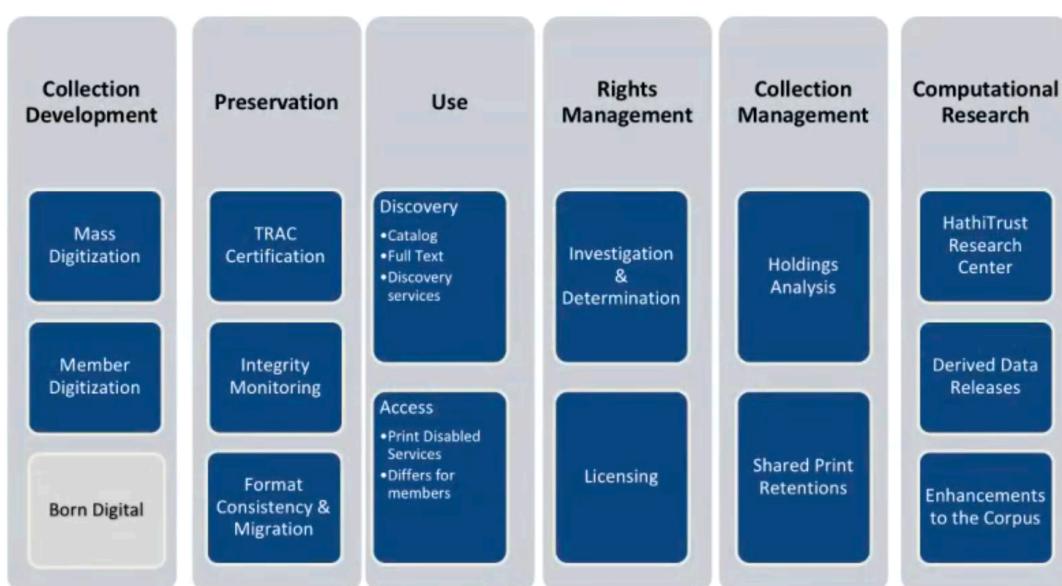


FIGURE 4.2 – Résumé des différentes activités entreprises par *HathiTrust*, capture d'écran

La plateforme, inaugurée en 2008, propose des livres, articles de journaux sous licences ou libres de droit. Voulue indépendante du projet de Google, elle intègre également de nombreuses autres collections découlant de divers projets de numérisation de masse, publics (*Internet Archive*) ou privés (*Microsoft*). *HathiTrust* regroupe plus d'une centaine

38. Comme nous l'avons mentionné plus haut, Google s'inscrit dans une logique d'accès et ne se focalise pas sur le stockage sur le long-terme.

39. *Viewpoint - What's Next for Google Books...*

40. HathiTrust Digital Library, *Charting HathiTrust's Strategic Directions* / www.hathitrust.org/charting-hathitrusts-strategic-directions (visité le 23/07/2019).

41. I. Xie et K. K. Matusiak, *Discover digital libraries...*, p.275.

42. *Ibid.*

de bibliothèques partenaires (140) à travers le monde⁴³ et offre un accès à quelques **seize millions** de documents^{44 45}. Active dans différents domaines, l'organisation contribue notamment à la poursuite des projets de numérisation au sein des institutions partenaires en permettant le partage des coûts, et à accroître l'accessibilité et l'usage pour la recherche des données préservées. À travers le développement de différents outils et la mise en place du centre pour les recherches analytiques de *the HathiTrust Research Center Analytics*, l'initiative soutient les travaux de recherches d'envergure conduits dans un objectif scientifique et par des organisations à but non lucratifs⁴⁶.

Pour regrouper les collections des différents partenaires et répondre aux besoins d'interopérabilité, un groupe collaboratif a développé un modèle de gestion des métadonnées visant à regrouper les différentes données des institutions et à en permettre l'exportation dans le format du projet : *Zephir*⁴⁷. Portée par l'expertise des bibliothèques partenaires, la recherche offre des fonctionnalités similaires à celles que l'on peut trouver via les interfaces de ces institutions, permettant de rechercher dans les index (titre, auteur etc.), en plus de la recherche avancée et d'une recherche plein-texte. De la qualité de ses métadonnées, découle la robustesse et les performances de son moteur de recherche⁴⁸. L'affichage des résultats est toutefois limité en fonction du droit d'auteur⁴⁹, puisque *HathiTrust* vise avant tout à préserver les collections et s'inscrit moins dans une démarche d'accessibilité universelle. Certains ouvrages soumis au droit d'auteur ne peuvent être recherchés ou certaines fonctionnalités de la plateforme sont réservées aux membres des institutions partenaires (comme le téléchargement pdf des textes libres de droit), ce qui a valu au projet quelques critiques^{50 51 52}.

L'anglais représente la langue majoritaire dans les collections⁵³, même si l'initiative poursuit l'objectif de permettre la recherche dans tous les systèmes d'écriture existants (actuellement celle-ci est disponible en caractères cyrilliques, hébreux, grecs, chinois, japonais et coréens)⁵⁴.

43. Kevin O'Brien, "Large-Scale Book and Journal Digitization Projects and Interlibrary Service : Opening the Discussion", *Journal of Interlibrary Loan, Document Delivery & Electronic Reserve*, 25-1-2 (mars 2015), p. 39-42, DOI : 10.1080/1072303X.2016.1150380.

44. I. Xie et K. K. Matusiak, *Discover digital libraries...*

45. HathiTrust Digital Library, *Charting HathiTrust's Strategic Directions / www.hathitrust.....*

46. Id., *HTRC Analytics*, URL : <https://analytics.hathitrust.org/> (visité le 23/07/2019).

47. Id., *Zephir / www.hathitrust.org*, URL : <https://www.hathitrust.org/zephir> (visité le 23/07/2019).

48. A. Weiss, *Using massive digital libraries...*

49. Soline Lau-Suchet, *HathiTrust : une bibliothèque numérique éléphantesque... fr-FR*, Billet, févr. 2014, URL : <https://bulac.hypotheses.org/967> (visité le 23/07/2019).

50. M. Wu, "Building a Collaborative Digital Collection : A Necessary Evolution in Libraries"...

51. K. O'Brien, "Large-Scale Book and Journal Digitization Projects and Interlibrary Service...".

52. I. Xie et K. K. Matusiak, *Discover digital libraries...*

53. A. Weiss, *Using massive digital libraries...*

54. S. Lau-Suchet, *HathiTrust...*

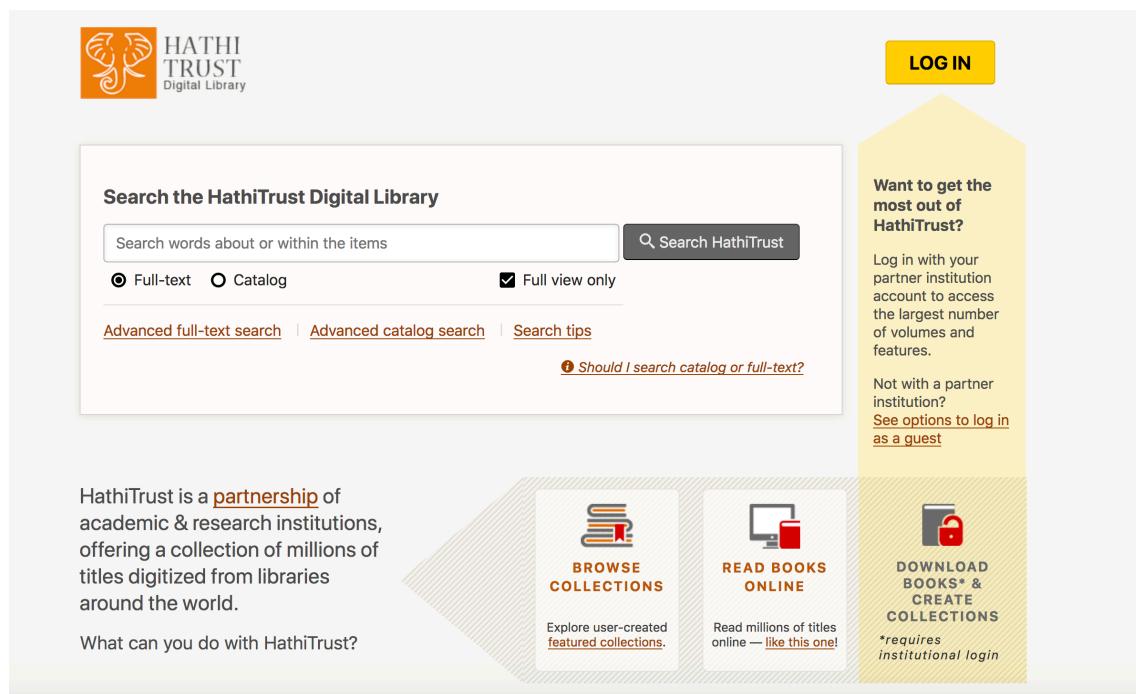


FIGURE 4.3 – Page d'accueil de la plateforme d'*HathiTrust*, capture d'écran

À l'inverse de Google, *HathiTrust* sait gérer les ouvrages en plusieurs volumes, ce qui contribue à leurs découvertes⁵⁵ et s'attèle à vérifier le statut d'un document lorsque des doutes liés au droit d'auteur persistent⁵⁶. La collaboration entre institutions partenaires est réalisée dans la plus grande transparence, et le site internet, mis à jour régulièrement, offre les derniers détails sur les décisions stratégiques soumises à l'approbation générale (dont les directions stratégiques envisagées pour la période 2019-2023)⁵⁷. *HathiTrust* développe des interfaces spécialement destinées aux personnes en situation de handicap⁵⁸, ce qui contribue à accroître l'accessibilité de ses collections.

55. A. Weiss, “Examining Massive Digital Libraries (MDLs) and Their Impact on Reference Services”...

56. *Ibid.*

57. HathiTrust Digital Library, *Charting HathiTrust's Strategic Directions* / www.hathitrust.org....

58. *Ibid.*

Enjeux	Réponses de HathiTrust
Amener différents acteurs à collaborer	Projet collaboratif, porté par quelques 140 bibliothèques partenaires. Une complète transparence semble être garante d'une cohésion d'ensemble.
Financement et partenariats public-privé	Les bibliothèques financent les infrastructures, selon une somme calculée sur différents facteurs à la fois liés aux collections de l'institution mais également à la taille du réseau et au pourcentage d'œuvres sous droits. Le projet ne semble pas recourir à des financements privés, mais bénéficie de leurs entreprises de numérisation.
Droit d'auteur	HathiTrust a gagné son procès contre la guilde des auteurs aux États-Unis. Il limite l'accès plein-texte des ouvrages sous droits et entreprend de véritables recherches afin de déterminer si des licences s'appliquent.
Sortir des silos : enjeux techniques	La qualité des métadonnées caractérisent le projet, favorisant l'efficacité du moteur de recherche. Un système de gestion dédié a été développé : <i>Zephir</i> . Ce dernier attribue un score à chaque fichier de métadonnées en fonction des informations trouvées dans les différents champs de métadonnées. Si ce score est trop bas, l'institution sera priée de corriger les éléments manquants ou une notice développée par un autre partenaire sera utilisée. Différentes APIs favorisent l'accès aux métadonnées ou aux documents.
Sortir des silos : enjeux sur le contenu	Le projet est motivé par des besoins académiques et ses collections portent de fait le même biais scientifique que les institutions partenaires. On observe une plus forte représentation des ouvrages en langue anglaise. Très peu d'institutions sont partenaires en dehors de ce milieu linguistique. Les formats des collections sont limités aux livres et articles de presse. Le projet rassemblant des données préalablement numérisées, ses collections offrent une forte redondance avec celles d'autres initiatives.
Stockage sur le long-terme - préservation	Le point fort du projet. L'infrastructure a été conçue pour garantir la longévité des données et suit les normes récentes. Les métadonnées de provenance et de contexte, et l'intégrité et l'authenticité des documents sont préservés.

TABLE 4.2 – Les réponses de *HathiTrust* aux enjeux de la numérisation

4.2 Agréger pour mieux valoriser

4.2.1 Europeana

Issue de diverses réactions politiques à l'arrivée de *Google Books*⁵⁹, Europeana est lancée par l'UE en 2008, motivée par la perspective de créer une nouvelle économie de la connaissance. Le projet vise à favoriser le partage de contenus de qualités vers une audience européenne, à agrandir le nombre des partenaires afin de toucher toutes les parties concernées et à susciter l'intérêt du public quel qu'il soit, pour les contenus des institutions culturelles et patrimoniales⁶⁰. La plateforme donne accès à quelques **58 millions** d'objets numérisés et élargit le spectre de leurs typologies. Les collections multilingues se composent d'images (33.5 millions), de textes (22.8 millions), de fichiers audio (700'000) ou vidéo (1.2 millions) et d'objets 3D (28'000) issus de plus de 3500 institutions^{61 62 63}.

En temps qu'aggrégateur, Europeana bénéficie des entreprises de numérisations menées à l'échelle d'une nation ou d'une région, faisant converger les attentes de chacun au sein d'une infrastructure commune : « *Europeana produces a new form of cultural memory politics that converge national and supranational imaginaries with global information infrastructures*⁶⁴ ».

59. En 2004, le Président français Jacques Chirac demande au directeur de la BNF Jean-Noël Jeanneney de lancer un projet européen similaire. Le directeur étant effrayé à la perspective de laisser les lois anglo-saxonnes s'imposer en Europe, il appelle à un effort national de numérisation et à la création d'une infrastructure européenne à même de préserver les différences culturelles des nations. Toutefois Europeana ne s'est pas construit comme une réponse à « l'invasion américaine », d'autres voix plus pragmatiques, à l'instar de celle du président de la Commission européenne, José Manuel Barroso, ont argumenté du besoin de créer une économie de la connaissance N. B. Thylstrup, *The politics of mass digitization...*

60. Europeana, *Europeana Collections*, URL : <https://www.europeana.eu/portal/en> (visité le 24/07/2019).

61. *Ibid.*

62. I. Xie et K. K. Matusiak, *Discover digital libraries...*

63. *European Commission report on Cultural Heritage...*

64. N. B. Thylstrup, *The politics of mass digitization...*, p.57.

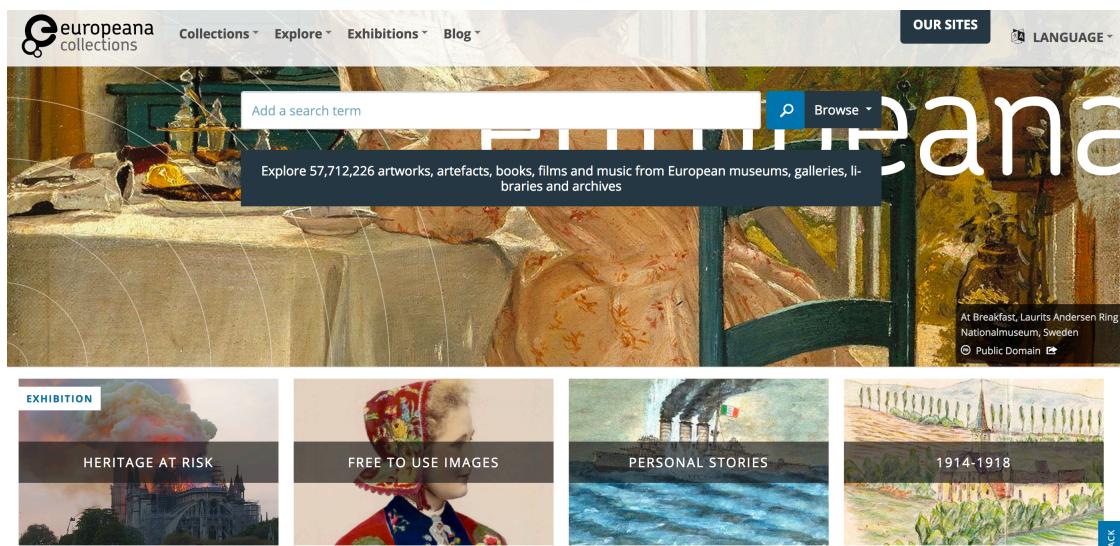


FIGURE 4.4 – Page d'accueil de l'interface découverte d'Europeana, capture d'écran

Bien que se réclamant d'une souveraineté nationale, Europeana travaille avec les grands réseaux du web (Google, Microsoft etc.) pour financer la numérisation des données qui seront rassemblées sur sa plateforme. Présenté comme une initiative européenne, le projet réagit à des règles de privatisation induites par ces partenariats⁶⁵. Les frontières entre public-privé tendant à se moduler à travers l'infrastructure, et celles entre institutions et environnement suivent la même logique⁶⁶. Ce paradoxe sera d'ailleurs souvent critiqué^{67 68}.

[Traduction] Si Europeana est un projet de numérisation de masse construit autour d'anciennes logiques souveraines, dont le discours officiel valorise l'État nation, en même temps qu'il abolit les frontières en déployant des infrastructures interopérables, la question demeure : quel est le résultat de cet assemblage d'un point de vue culturel ?⁶⁹

Europeana étant un projet sous financement européen, il respecte le droit d'auteur⁷⁰, alors même que ce système est très complexe à déployer. Comme l'Europe ne possède pas de législation commune, il navigue entre les différentes particularités nationales, l'émergence d'un marché global et la mouvance du libre accès ou Open Access^{71 72}. Europeana

65. M. Thelle et N. Bonde Thylstrup, "Persuasive territories in European cultural politics...".

66. N. B. Thylstrup, *The politics of mass digitization...*, p.63.

67. M. Thelle et N. Bonde Thylstrup, "Persuasive territories in European cultural politics...".

68. *Ibid.*

69. « *If Europeana is a late-sovereign mass digitization project that maintains discursive ties to the national imaginary at the same time that it undercuts this imaginary by means of networked infrastructure through increased interoperability, the final question is : what does this late-sovereign assemblage produce in cultural terms ?* » N. B. Thylstrup, *The politics of mass digitization...*, p.73

70. par opposition à Google avec son application du *fair use*

71. *Ibid.*

72. A. Weiss, *Using massive digital libraries...*

favorise les contenus en accès libres sous licence *Creative Commons CC0*⁷³, ce qui pose des problèmes d'application pour certains États (dont la France), plus restrictifs⁷⁴. On retrouve également différentes formes d'affichage en fonction de ces restrictions (accès aux contenus interactifs et aux métadonnées, accès aux contenus statiques et métadonnées, extraits et métadonnées, métadonnées seulement)⁷⁵. Les œuvres à partir du 20^e siècle sont sous-représentées au profit des œuvres libres de droit^{76 77}.

Pour tenter de répondre à ces différentes contraintes tout en effectuant sa mission, Europeana investit le principe d'interopérabilité : proposant notamment un cadre de publication à destination des institutions directement ou des agrégateurs nationaux et thématiques de collections numérisées⁷⁸ ; une stratégie de contenu ; un modèle de métadonnées suivant les recommandations du web sémantique⁷⁹ et proposant un enrichissement à l'aide de vocabulaires contrôlés⁸⁰ ; des modèles de licences basés sur les *Creative Commons*. Plusieurs APIs (dont celles développées par IIIF) contribuent à l'accessibilité des collections et la naissance de nouveaux produits^{81 82}. Les données sont également rendues accessibles via le protocole OAI-PMH⁸³. Soucieuse de garantir une meilleure utilisation des données, Europeana conjointement avec DPLA et d'autres institutions, fait partie du *RightsStatements.org* Consortium proposant des « déclarations de droits standardisées pour le patrimoine culturel disponible en ligne »⁸⁴. Les institutions peuvent choisir entre 12 déclarations (en cours de traductions dans tous les idiomes européens), favorisant l'usage des fichiers par le public⁸⁵.

De nombreuses communautés sont impliquées dans Europeana. Le projet s'est construit

73. Licence la moins restrictive, n'obligant pas à citer la provenance du contenu.

74. *European Commission report on Cultural Heritage...*

75. Marieke Willems et Rossitza Atanassova, “Europeana Newspapers : searching digitized historical newspapers from 23 European countries”, *Insights the UKSG journal*, 28–1 (mars 2015), p. 51-56, DOI : 10.1629/uksg.218.

76. *Numérisation du patrimoine...*

77. V. Stobo, K. Patterson, K. Erickson, *et al.*, ““I should like you to see them some time”...”.

78. Afin de mieux interagir avec les plateformes et institutions existantes, Europeana a créé en 2012 le *Europeana Aggregator forum*. Ce lieu permet la mise en place d'échanges professionnels et assure l'implication de ces instances stratégiques au sein du projet. Les agrégateurs offrent à leurs institutions partenaires conseils et support dans leurs projets de numérisation concernant le choix les licences ; formats ; traitement du multilinguisme etc... Europeana, *Breathing new life into the Europeana Aggregator Forum*, en-GB, URL : <https://pro.europeana.eu/post/breathing-new-life-into-the-europeana-aggregators-forum> (visité le 24/07/2019). Plus de 2/3 des États membres ont des agrégateurs nationaux *European Commission report on Cultural Heritage...*

79. Id., *Linked Open Data*, en-GB, URL : <https://pro.europeana.eu/page/linked-open-data> (visité le 24/07/2019).

80. Lexique servant à organiser les connaissances pour favoriser la recherche d'information

81. Henriette Roued-Cunliffe et Andrea Copeland (éd.), *Participatory heritage*, OCLC : ocn971483437, London, 2017.

82. à l'instar de l'*EuropeanaBot*, garant d'une certaine forme de sérendipité, puisque publient aléatoirement des images issues des collections. *Europeana Bot (@EuropeanaBot) / Twitter...*

83. Id., *Everything you need to publish in Europeana*, en-GB, URL : <https://pro.europeana.eu/services/data-publication-services/existing-provider> (visité le 24/07/2019).

84. *RightsStatements.org*, URL : <https://rightsstatements.org/fr/> (visité le 24/07/2019).

85. *European Commission report on Cultural Heritage...*

dans la mouvance des sciences citoyennes ou *citizen science*, dès lors ces communautés participent entre autres aux corrections, transcriptions, classification, contextualisation des données numérisées. L'expérience promue par le projet est de ne pas simplement demander aux utilisateurs de cliquer, mais de leur offrir l'opportunité de prendre une part active dans un projet de recherche⁸⁶. Ce souci d'impliquer l'utilisateur se retrouve dans le développement de la plateforme et des services, héritage probable des pratiques mises en place au sein des institutions partenaires⁸⁷.

Pour améliorer l'expérience utilisateur et lui permettre de trouver son chemin face aux collections foisonnantes, des expositions virtuelles ont été créées⁸⁸, la recherche peut se faire dans toutes les langues européennes mais les métadonnées, proposées en résultat ne sont pas traduites⁸⁹. Europeana propose deux interfaces différentes, s'adressant d'un côté à l'utilisateur « grand public » et mettant en avant la découverte, ou interagissant en toute transparence sur sa plateforme professionnelle avec les partenaires du réseau et les entités ou individus motivés à profiter pleinement des opportunités offertes par ses collections massives (éducation, recherche, industries etc.)⁹⁰. L'interface professionnelle décrit précisément l'organisation, les spécificités techniques et propose des accompagnements ciblés selon si l'on se positionne en tant que partenaire du réseau, chercheur, ou futur exploitant⁹¹.

Europeana propose depuis les débuts des interfaces mobiles, ce qui est important pour le projet, car la plupart de son trafic provient des portables⁹².

86. N. B. Thylstrup, *The politics of mass digitization...*

87. E. Jones, “The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative...”.

88. Europeana, *Europeana Collections...*

89. *European Commission report on Cultural Heritage...*

90. Ioana Roiu, Cristina, “From Family Memories to Digital Archives – Europeana and the Educational Value of its Digital Collections” (), p. 137-143.

91. Europeana, *Europeana Collections...*

92. I. Xie et K. K. Matusiak, *Discover digital libraries...*

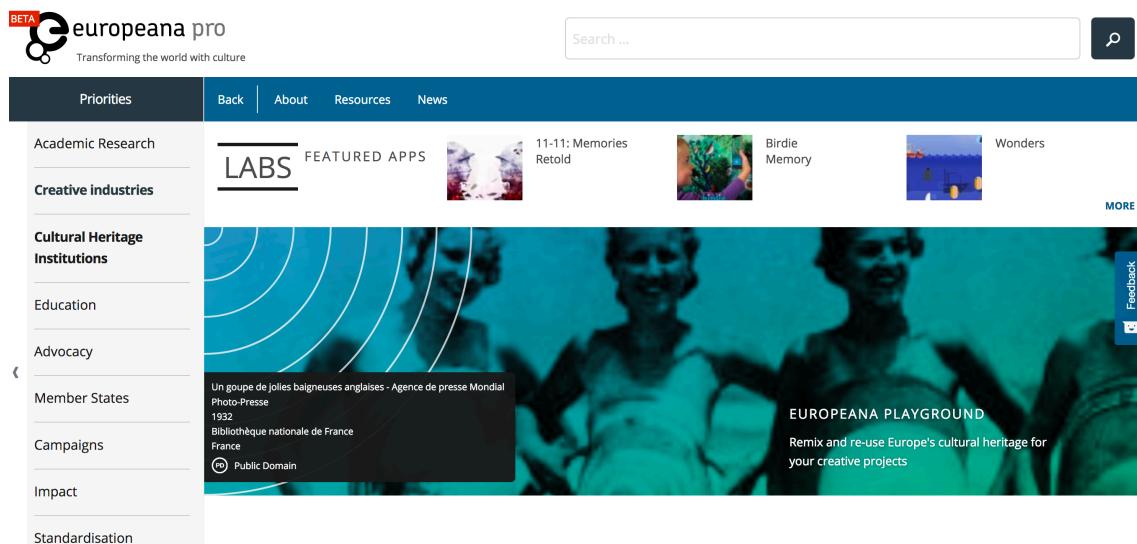


FIGURE 4.5 – Page d'accueil de l'interface professionnelle d'Europeana, capture d'écran

Se voulant reflet du patrimoine culturel européen, ses collections n'en sont toutefois pas pleinement représentatives. Certains pays ont moins développé la numérisation (biais géographique), certaines typologies d'objets sont sous-représentées, les collections héritent des biais des institutions partenaires⁹³. Ce qui amènera un responsable d'Europeana à caractériser, en 2010, la différence entre Google et Europeana, par le fait que d'un côté l'usager accède à tout, alors que de l'autre les collections ont été soigneusement sélectionnées par les institutions en vue de leur numérisation⁹⁴. Toutefois Europeana tente de proposer de nouvelles formes de numérisation, impliquant le citoyen dans le processus de sélection⁹⁵, et a mis en place une stratégie de contenu visant à pallier aux manques des collections⁹⁶.

Europeana n'est pas responsable du stockage sur le long-terme des collections numérisées. Une forte disparité est constatée au niveau européen concernant l'existence de législations relatives à la préservation des données. Cette dernière est souvent dépendante du bon vouloir des institutions⁹⁷.

Malgré son ancrage européen, Europeana est embranchée dans un écosystème Google, laissant le projet à la merci de changements de sa stratégie de référencement et par conséquent face au risque de perdre en visibilité⁹⁸. Afin de mieux traiter la richesse des langues européennes, le projet a d'ailleurs intégré une fonction Google, *Google Translate*⁹⁹. Depuis

93. N. B. Thylstrup, *The politics of mass digitization...*

94. M. Thelle et N. Bonde Thylstrup, "Persuasive territories in European cultural politics..." .

95. Dans le cadre du projet Europeana 1914-1918, les individus ont été sollicités afin de créer une archive de la Première Guerre Mondiale basée sur leurs images et écrits familiaux Ioana Roiu, Cristina, "From Family Memories to Digital Archives – Europeana and the Educational Value of its Digital Collections" ...

96. Europeana, *Everything you need to publish in Europeana...*

97. European Commission report on Cultural Heritage...

98. N. B. Thylstrup, *The politics of mass digitization...*

99. M. Thelle et N. Bonde Thylstrup, "Persuasive territories in European cultural politics..." .

2011, les bibliothèques européennes signataires d'un accord de numérisation avec Google voient leurs documents indexés par Europeana¹⁰⁰.

L'UE a récemment fixé les nouvelles priorités pour Europeana, qui visent à : développer le multilinguisme de la plateforme ; améliorer la valorisation des collections et proposer de nouveaux services de découverte ; mieux prendre en compte les besoins des petites institutions¹⁰¹.

100. A. Moatti, "Bibliothèque numérique européenne : de l'utopie aux réalités"...

101. European Commission report on Cultural Heritage...

Enjeux	Réponses d'Europeana
Amener différents acteurs à collaborer	Une grande transparence et l'élaboration du cadre technique en collaboration étroite avec les partenaires semblent être garants d'une bonne collaboration. De nombreux groupes de travail impliquant les institutions contribuent aux évolutions du projet. Les agrégateurs nationaux ou thématiques sont intégrés au sein d'un forum, favorisant les échanges. Le grand public devient partenaire des projets de recherche.
Financement et partenariats public-privé	Comme la plupart des projets de numérisation sont conduits de manière décentralisées par les institutions, ces dernières se tournent vers des institutions privées (Google, Microsoft etc.) pour les financer. L'Europe finance l'infrastructure et quelques projets de numérisation ciblés.
Droit d'auteur	Le projet agit dans la légalité et applique des licences et restrictions aux œuvres sous droit. Un cadre de publication règle les conditions d'échange des métadonnées et s'assure que la mention de la licence y figure. Les licences <i>Creative Commons</i> sont privilégiées. Membre du <i>RightsStatements</i> , le projet propose des déclarations de droit standardisées. Différentes formes d'affichage garantissent un accès légal aux collections.
Sortir des silos : enjeux techniques	Un cadre de publication contenant un modèle de métadonnées a été développé. Différentes APIs (dont IIIF) sont déployées, le protocole OAI-PMH est disponible, les métadonnées sont exposées suivant les principes du web sémantique et enrichies grâce à des vocabulaires contrôlés.
Sortir des silos : enjeux sur le contenu	Conscient des biais liés aux formats, à la géographie et à l'héritage des institutions partenaires, le projet a mis en place une stratégie de contenu afin de pallier aux manques constatés. De nombreux formats de documents sont disponibles, dont des fichiers audio, vidéo et 3D. Les documents sont multilingues mais le français est sur-représenté.
Stockage sur le long-terme - préservation	Europeana n'a pas vocation à stocker directement les données, (cette responsabilité est laissée aux institutions) toutefois les critères de qualité imposés aux métadonnées contribuent à la préservation de l'authenticité et de l'intégrité des documents.

TABLE 4.3 – Les réponses d'Europeana aux enjeux de la numérisation

4.2.2 Digital Public Library of America

La *Digital Public Library of America* ou DPLA, pensée dès 2010 et officiellement lancée en 2013 vise à rendre le contenu des institutions culturelles et patrimoniales états-unies accessible à tous. Deux programmes ont été mis en place pour répondre à cette mission : l'agrégation des données numériques des institutions de tous les États-Unis et un service d'ebooks^{102 103}. Ce projet est en grande partie dû aux travaux intellectuels du *Berkman Klein Center for Internet and Society*¹⁰⁴ de l'université d'Harvard et de grands théoriciens du numérique, et des bibliothèques, qui ont travaillé à le positionner comme une alternative publique à l'infrastructure privée de *Google Books*¹⁰⁵. Cependant, comme dans toutes les initiatives de numérisation de masse, la notion de « public » suscite de nombreuses controverses, puisque l'on reproche à ces projets de détourner l'argent autrement réservé aux infrastructures plus petites¹⁰⁶.

Le réseau rassemble quelques 41 États, plus de 4000 institutions et permet de découvrir à travers sa plateforme plus de **30 millions** d'objets de différentes typologies (images, cartes, fichiers audio, manuscrits, œuvres muséales etc.)¹⁰⁷. Au même titre qu'Europeana, le projet n'héberge pas ces données, mais offre un portail permettant d'y accéder^{108 109}.

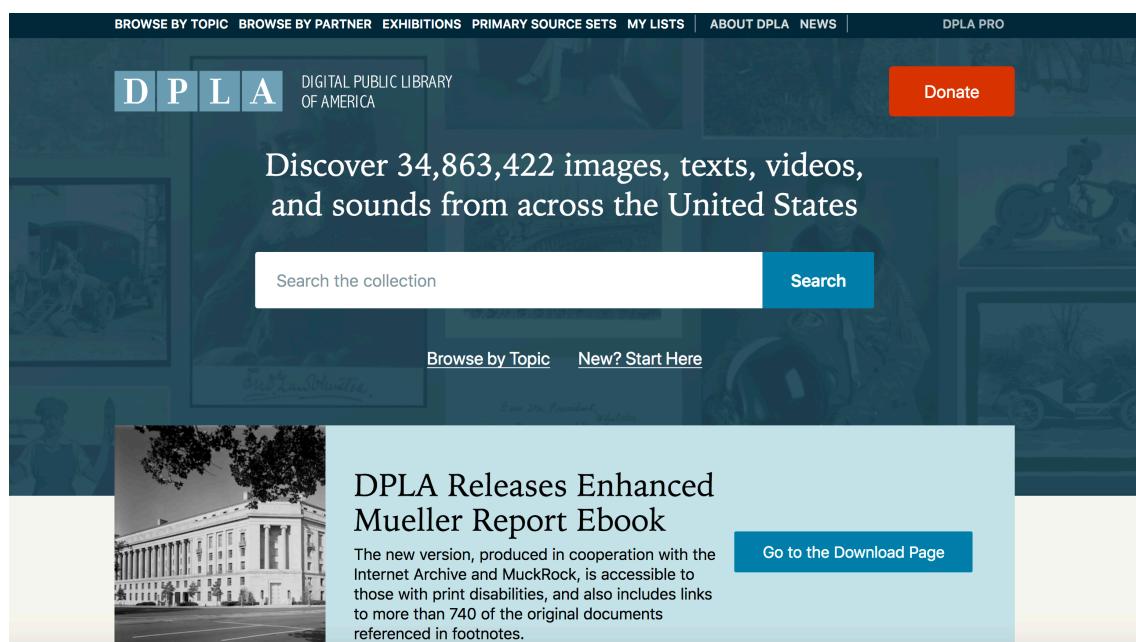


FIGURE 4.6 – Page d'accueil de l'interface découverte du projet de DPLA, capture d'écran

102. Digital Public Library of America, *Strategic Plan*, en, URL : <https://pro.dp.la/about-dpla-pro/strategic-plan> (visité le 25/07/2019).

103. Id., *History*, en, URL : <https://pro.dp.la/about-dpla-pro/history> (visité le 25/07/2019).

104. La mission de ce centre est notamment d'explorer le cyberspace et d'identifier le besoin de nouvelles législations ou des ajustements de celles existantes.

105. N. B. Thylstrup, *The politics of mass digitization...*

106. *Ibid.*

107. Digital Public Library of America, *Strategic Plan...*

108. J. Pomerantz, *Metadata...*

109. I. Xie et K. K. Matusiak, *Discover digital libraries...*

Bien que sa structure soit uniquement basée aux États-Unis, DPLA a été développé pour être interopérable avec Europeana (le même modèle de métadonnées est utilisé¹¹⁰, les deux institutions collaborent au sein du *RightsStatements.org* Consortium etc.). Les liens entre les deux initiatives contribuent ainsi à la création d'un réseau mondial¹¹¹ et les placent à l'avant-garde d'un mouvement grandissant, visant à proposer des modèles de métadonnées spécifiques à certains domaines¹¹². DPLA s'appuie également sur un réseau d'agrégateurs nationaux qui facilitent l'intégration des données de leurs membres, et proposent des solutions de stockage pour les petites institutions qui n'en disposeraient pas^{113 114}. Des APIs permettent des requêtes sur le contenu et les métadonnées, la plate-forme propose également un export total de ses bases de données¹¹⁵.

L'initiative soutient une gestion des risques et l'application du *fair use*, offerte par son ancrage géographique, pour justifier la numérisation des œuvres orphelines ou dont le libre accès suscite des doutes. Il incite toutefois les institutions à entreprendre des recherches sérieuses afin de déterminer si un contenu est libre de droit ou non, et offre un accompagnement dans cette démarche à ses partenaires¹¹⁶.

DPLA travaille au développement et à la mise en place de solutions techniques afin de soutenir l'engagement des communautés, la recherche et l'éducation. Ses missions actuelles s'articulent autour de différents projets visant à : favoriser l'usage des ebooks par le grand public et les bibliothèques ; accroître l'emploi des collections par le système éducatif ; clarifier l'usage des œuvres sous droit ; construire des solutions de dépôt numérique ; améliorer la présence de journaux issus de communautés marginalisées ; former les nouveaux partenaires aux technologies de la plateforme et les soutenir dans leurs entreprises de numérisation¹¹⁷.

Tout comme Europeana, l'initiative dispose de deux plateformes, l'une permettant la découverte des collections et la deuxième, professionnelle, s'adressant plus précisément : aux agrégateurs partenaires ; aux candidats à la création d'un nouvel agrégateur ; aux communautés de développeurs ; aux chercheurs ; aux utilisateurs d'ebooks ; aux communautés de volontaires souhaitant promouvoir le projet¹¹⁸. Soucieuse de faciliter l'accès à ses collections, la plateforme publique propose également des expositions virtuelles et a

110. Digital Public Library of America, *Metadata Application Profile*, en, URL : <https://pro.dp.la/hubs/metadata-application-profile> (visité le 25/07/2019).

111. I. Xie et K. K. Matusiak, *Discover digital libraries...*

112. J. Pomerantz, *Metadata...*

113. Digital Public Library of America, *Becoming a Service Hub*, en, URL : <https://pro.dp.la/prospective-hubs/becoming-a-service-hub> (visité le 25/07/2019).

114. I. Xie et K. K. Matusiak, *Discover digital libraries...*

115. *Ibid.*

116. Digital Public Library of America, *Understanding Copyright*, en, URL : <https://pro.dp.la/projects/understanding-copyright> (visité le 25/07/2019).

117. Id., *Projects / DPLA*, URL : <https://pro.dp.la/projects> (visité le 25/07/2019).

118. Id., *Our Values*, en, URL : <https://pro.dp.la/about-dpla-pro/our-values> (visité le 25/07/2019).

développé un *Bot*¹¹⁹ afin de ramener la notion de sérendipité dans la découverte¹²⁰.

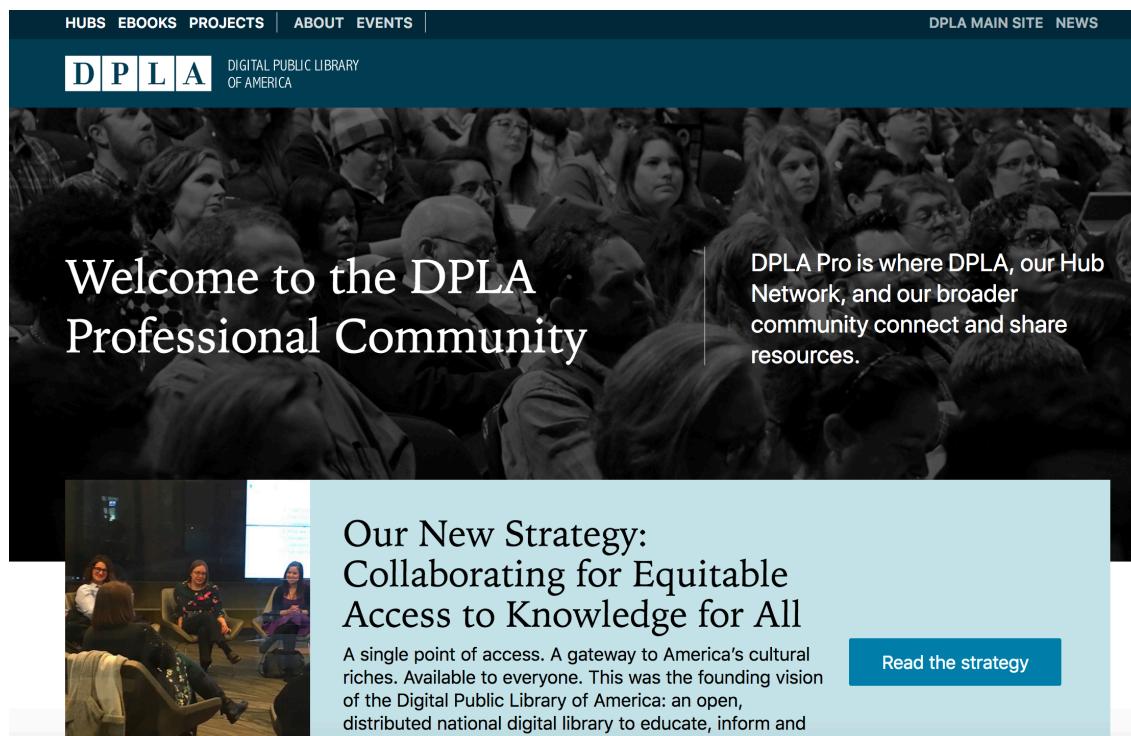


FIGURE 4.7 – Page d'accueil de l'interface professionnelle du projet de DPLA, capture d'écran

Ces projets et l'infrastructure sont financés par des fondations à but non lucratif, par des fonds du gouvernement américain et par une redevance prélevée auprès des institutions partenaires (qui varie entre 10'000 et 15'000 USD par an)¹²¹. Les institutions doivent elles-mêmes assurer les fonds nécessaires à la numérisation de leurs données mais peuvent être aidées dans cette recherche¹²². Le projet indexe également des données numérisées par Google et stockées par *HathiTrust*.

La collaboration fait partie intégrante de sa démarche, de nombreux groupes de travail sont mis en place afin de réfléchir aux développements de nouvelles solutions. Loin d'être partenaire du clivage Amérique - Europe, l'institution chercher à collaborer avec les réseaux existants¹²³.

119. Programme autonome capable d'interagir avec un système ou un utilisateur.

120. N. B. Thylstrup, *The politics of mass digitization...*

121. Digital Public Library of America, *Membership Program*, en, URL : <https://pro.dp.la/hubs/membership-program> (visité le 25/07/2019).

122. Id., *Our Supporters*, en, URL : <https://pro.dp.la/about-dpla-pro/our-supporters> (visité le 25/07/2019).

123. Id., *Our Values...*

Enjeux	Réponses de the Digital Public Library of America
Amener différents acteurs à collaborer	Collaborer fait partie des missions du projet. Une plateforme professionnelle garantit la transparence de la gouvernance. Différents groupes de travail permettent aux partenaires de participer aux développements futurs. Les communautés professionnelles, ou non, sont impliquées dans le projet.
Financement et partenariats public-privé	Le financement est assuré par le gouvernement américain, des fondations à but non lucratif et une redevance versée annuellement par les partenaires. Les projets de numérisation sont à la charge des partenaires qui peuvent collaborer avec des groupes privés (tels Google, Microsoft etc.).
Droit d'auteur	Favorable à l'usage des licences <i>Creative Commons</i> , encourage les partenaires à tirer parti du <i>fair use</i> et à mettre en place une gestion des risques. Membre du <i>RightsStatements</i> , le projet propose des déclarations de droit standardisées.
Sortir des silos : enjeux techniques	Adoption du modèle de métadonnées développé par Europeana et interopérabilité avec sa plateforme. APIs pour accéder aux métadonnées ou aux contenus.
Sortir des silos : enjeux sur le contenu	Hérite des biais des institutions partenaires et sur-représentation de certains formats de documents (textes et images). Un effort est fait pour identifier les manques des collections et les corriger par la mise en place de projets de numérisation ciblés.
Stockage sur le long-terme - préservation	Développe des solutions de dépôts institutionnels pour les petites institutions et favorise les agrégateurs qui en proposent à leurs communautés. Le stockage des données sur le long-terme demeure sous la responsabilité des institutions. La qualité des métadonnées favorisent le respect de l'intégrité et de l'authenticité des objets.

TABLE 4.4 – Les réponses de la *Digital Public Library of America* aux enjeux de la numérisation

Chapitre 5

Comment catégoriser ?

La comparaison plus détaillée de ces quatre initiatives de numérisation de masse permet de nuancer les propos de ceux qui distinguent initiatives financées par des institutions commerciales de celles financées par des institutions publiques ou à but non lucratif, puisque les activités de ces deux typologies de projet semblent étroitement imbriquées. Nous appuyons la réflexion contenue dans cette section sur les réponses apportées aux enjeux de la numérisation, bien que d'autres critères non étudiés dans le cadre de ce mémoire auraient pu servir à cette comparaison (tels que le rôle joué par l'homme au sein de ces projets, le nombre de publications officielles proposées etc...) ¹.

Les critères proposés par Cory Lampert argumentent que les premières (financées par des institutions commerciales) se distinguent par un enjeu plus grand sur l'accessibilité des données, lorsque les deuxièmes (financées par le secteur public) offrent une modération humaine dans la gestion des collections avec un accent plus développé sur les questions de préservation et la réutilisation des métadonnées². Bien que corroborées en un sens par nos exemples, l'analyse de Cory Lampert semble oser quelques raccourcis et ne pas refléter la complexité des réponses apportées aux enjeux de la numérisation. Si *Google Books* offre l'avantage de pouvoir rendre accessible ses données au sein de son propre environnement et bénéficie de fait d'une meilleure visibilité, les autres projets déploient de grands efforts pour améliorer l'accessibilité des leurs. Les différents développements des plateformes d'Europeana, *HathiTrust* et DPLA témoignent de leurs investissements. Ces trois initiatives se distinguent d'ailleurs par la prise en compte de plusieurs typologies de public et le déploiement de services spécialement adaptés. En ce sens, il est vrai que la modération humaine joue un rôle important pour les initiatives publiques. La préservation n'est certes pas considérée par Google, mais hormis *HathiTrust*, Europeana et DPLA semblent surtout déléguer cette responsabilité aux institutions détentrices des données. Leur position d'agrégateur justifie la réutilisation des métadonnées, afin de ne pas refaire à double le travail mené par leurs institutions partenaires, et constitue plus une conséquence

1. K. Leetaru, "Mass book digitization..."

2. C. Lampert, "Ramping up...", p.47

qu'une véritable opportunité.

S'il est vrai que les entreprises publiques que sont Europeana, DPLA et *HathiTrust* sont davantage transparentes concernant la gouvernance ou le détail de leurs activités, des développements techniques propres à leurs plateformes et des processus de numérisation, demeurent souvent réalisés par des institutions privées et ne sont dès lors que peu documentés puisqu'ils répondent à des enjeux de protections commerciales évidents : « [...]the important infrapolitical question in mass digitization, namely, how, why, and when to manage visibilities [...]»³. A ce titre, il est possible d'argumenter que Google offre la même ouverture sur son organisation que les initiatives publiques⁴.

La collaboration n'a sans doute pas la même valeur pour *Google Books* que pour les autres projets (puisque dirigé par une seule compagnie qui peut imposer les mêmes règles à ses partenaires⁵), qui doivent amener un grand nombre d'institutions publiques ou privées à collaborer. Les critères économiques semblent toutefois pousser les initiatives à se développer de manière complémentaire et à collaborer entre réseaux pour trouver des solutions communes (DPLA et Europeana dans le cadre du *RightsStatements.org*, Europeana et Google pour le partage de certaines fonctionnalités et l'indexation des données numérisées).

Les projets financés uniquement par des fonds publics peinent à construire sereinement leurs activités, Europeana a souvent été contrainte par les instances politiques à revoir ses objectifs et son positionnement⁶. Il semble que pour parvenir à un équilibre sur le long-terme, les institutions doivent déployer des partenariats public-privé ou prélever une redevance auprès de leurs partenaires (*HathiTrust*, DPLA). De plus l'acte de numérisation est finalement souvent financé par des institutions privées, qui imposent dès lors certaines restrictions d'usage⁷.

Concernant le droit d'auteur, Google, de par son emploi du *fair use*, propose un accès moins restrictif à ses collections que les autres exemples étudiés. De plus il a mis en place avant les autres initiatives différents affichages en fonction des droits liés au document⁸. DPLA défend d'ailleurs la mise en place d'une gestion des risques dans les projets de numérisation américains afin de favoriser un accès plus large aux collections potentiellement sous droit. Les différents projets peinent toutefois à proposer une solution satisfaisante, ce qui résulte en une sous-représentation des œuvres à partir du 20^e siècle.

La quête de l'interopérabilité est au centre des préoccupations des projets. Si l'environnement de Google permet à *Google Books* de s'affranchir des barrières, DPLA, Europeana et *HathiTrust* sont contraints de développer plusieurs solutions techniques pour

3. N. B. Thylstrup, *The politics of mass digitization...*

4. K. Leetaru, "Mass book digitization...".

5. *Ibid.*

6. N. B. Thylstrup, *The politics of mass digitization...*

7. M. Thelle et N. Bonde Thylstrup, "Persuasive territories in European cultural politics...".

8. K. Leetaru, "Mass book digitization...".

sortir des silos des institutions.

Bien qu'il y ait au sein de toutes les initiatives une volonté officialisée de donner accès à la connaissance qu'elle soit universelle ou européenne, sa définition ne semble pas faire l'unanimité. Google se limite aux livres et restreint sa politique de numérisation à ce format, tandis qu'Europeana et DPLA en élargissent le champ pour inclure d'autres formes de création. Les critères de sélection liés au développement des collections, mis en place par les institutions, sont reproduits au sein des projets de numérisation de masse, et si certains argumentent que *Google Books* n'a pas de politique documentaire⁹, il en a en vérité une très complexe, au vu du nombre de ses partenaires¹⁰. Il est cependant indéniable que certains projets travaillent plus activement à détecter et corriger leurs « manques », à l'instar de DPLA et Europeana.

HathiTrust est le projet le plus actif dans la préservation des données sur le long terme, puisqu'il s'est construit autour de cette mission et pour pallier au désintérêt montré par Google sur la question. Toutefois les deux initiatives sont les seules à stocker elles-mêmes leurs données numérisées. Europeana et DPLA, en temps qu'agrégateurs semblent surtout déléguer cette responsabilité aux institutions, bien que DPLA travaille au développement et à l'implémentation de dépôts pour celles qui n'en auraient pas.

Vouloir effectuer une catégorisation entre ces différentes initiatives, semble relever aujourd'hui d'un exercice difficile. La distinction public-privé s'efface au profit d'une collaboration. La proposition de Margaret Coutts que nous avons choisie d'appliquer, bien que cohérente, semble manquer de plus en plus d'initiatives actives au sein du premier groupe et encourt d'une part, le risque d'isoler la démarche d'*HathiTrust* et de restreindre ses activités de valorisation derrière la bannière de la préservation, et d'autre part, d'occulter le rôle joué par Google au sein du deuxième groupe. Les initiatives de numérisation de masse évoluent rapidement et leur croissance continue rend l'exercice de définition de ces géants de plus en plus complexe, tant ils s'articulent autour d'enjeux et d'objectifs multiples¹¹.

Peut-être qu'alors il serait plus raisonnable de considérer les projets de numérisation de masse comme un tout, chaque nouvelle initiative se construisant sur les précédentes et contribuant à renforcer les bonnes pratiques et à proposer de nouvelles orientations, plutôt que de vouloir séparer des projets pourtant motivés par les développements des autres. Une approche inclusive de ces initiatives permettrait l'ouverture de nouvelles perspectives de recherches, palliant au risque d'une catégorisation trop réductive incapable de refléter leur envergure.

9. Numérisation du patrimoine...

10. M. Coutts, *Stepping away from the silos...*

11. N. B. Thylstrup, *The politics of mass digitization...*

Troisième partie

**Une nouvelle voie à définir pour
Time Machine**

Chapitre 6

Stagiaire au sein du projet Time Machine

Nous avons été engagée en tant que stagiaire au sein du DHLAB de l'EPFL, du 15 avril au 31 août 2019. Nous avons pris une part active dans l'élaboration de la future feuille de route du projet Time Machine.

6.1 État des lieux avant le commencement du stage

Lorsque nous avons commencé notre stage, Time Machine venait tout juste de recevoir le million d'euros de l'UE pour démontrer la faisabilité du projet et tenter de remporter un milliard. Cette période appelée à se dérouler du 1er mars 2019 au 29 février 2020 est appelée *Coordination and Support Actions* (CSA) par l'UE, et implique la collaboration des 32 partenaires du projet dans l'élaboration de solutions. Dans la proposition initialement soumise à l'UE, quatre piliers, *PILLARs*, correspondant à des thématiques particulières et devant proposer chacun une feuille de route détaillée contenant des solutions concrètes ont été identifiés et les responsabilités de chacun d'eux réparties parmi autant de groupes de travail composés de membres du consortium.

En plus de ces différents piliers amenés à durer au-delà de la phase du CSA, différents *Work Package*, lot de travail (WP) visent à accompagner la mise en place du projet durant cette année charnière. La temporalité des actions à mener est également préalablement définie au même titre que les interactions entre les différentes tâches composant chaque pilier, et les interactions entre les différents piliers.

La liste des activités conduites durant la période du CSA peut se résumer ainsi :

1. *WP 1 : Project Management*
2. *PILLAR 1 : Science and Innovation*

Ayant pour objectif de proposer les différentes thématiques de recherches nécessaires au fonctionnement des futurs processus de l'infrastructure Time Machine.

3. *PILLAR 2 : Time Machine Operations*

Ayant pour objectif de mettre en place le projet, de définir son fonctionnement tant au niveau organisationnel que technologique.

4. *PILLAR 3 : Exploitation Avenues*

Ayant pour objectif de définir quels produits et services pourront être construits avec les données du projet et comment mettre en place la collaboration avec ses futurs utilisateurs.

5. *PILLAR 4 : Innovation and Outreach*

Ayant pour objectif de garantir la création d'une structure suffisamment flexible pour évoluer fortement en fonction des développements technologiques et demeurer suffisamment souple pour ne pas freiner l'innovation.

6. *WP 6 : Governance scheme*

Ayant pour objectif de mettre en place la gouvernance de Time Machine.

7. *WP7 : Dissemination and Promotion*

Ayant pour objectif la communication et valorisation du projet Time Machine.

8. *WP8 : Overall Time Machine Strategy and Implementation Plan*

Ayant pour objectif de garantir la cohérence globale et le développement futur de Time Machine.

Certains groupes de travail doivent proposer des solutions plus tôt, afin que d'autres puissent bâtir leurs propositions en tenant compte des objectifs énoncés, c'est notamment le cas pour les *PILLARS* 1, 2 et 3.

Notre maître de stage, le Professeur Frédéric Kaplan, est chargé de la gestion du projet durant cette période du CSA.

Un consultant externe, déjà engagé pour la proposition initiale, se charge de superviser les différentes activités et propositions et offre une vision plus stratégique, nécessaire au bon développement du projet.

Un coordinateur de projet a également été engagé pour la durée du CSA, chargé de mettre en place les outils collaboratifs nécessaires à des groupes de travail répartis entre plusieurs territoires et de professionnaliser les pratiques de Time Machine, en prenant en charge le management des risques et en proposant une supervision globale du projet. Les activités du chef de projet forment un WP spécifique au temps du CSA.

Un comité exécutif réunissant les responsables des différents *PILLARS*, le chef de projet, le consultant et le coordinateur se réunit virtuellement chaque semaine, afin de garantir un niveau d'échanges minimal entre les différents participants et pour faire face aux inévitables enjeux posés par un projet d'une telle envergure. Nous serons amenée à en faire partie durant toute la durée de notre stage.

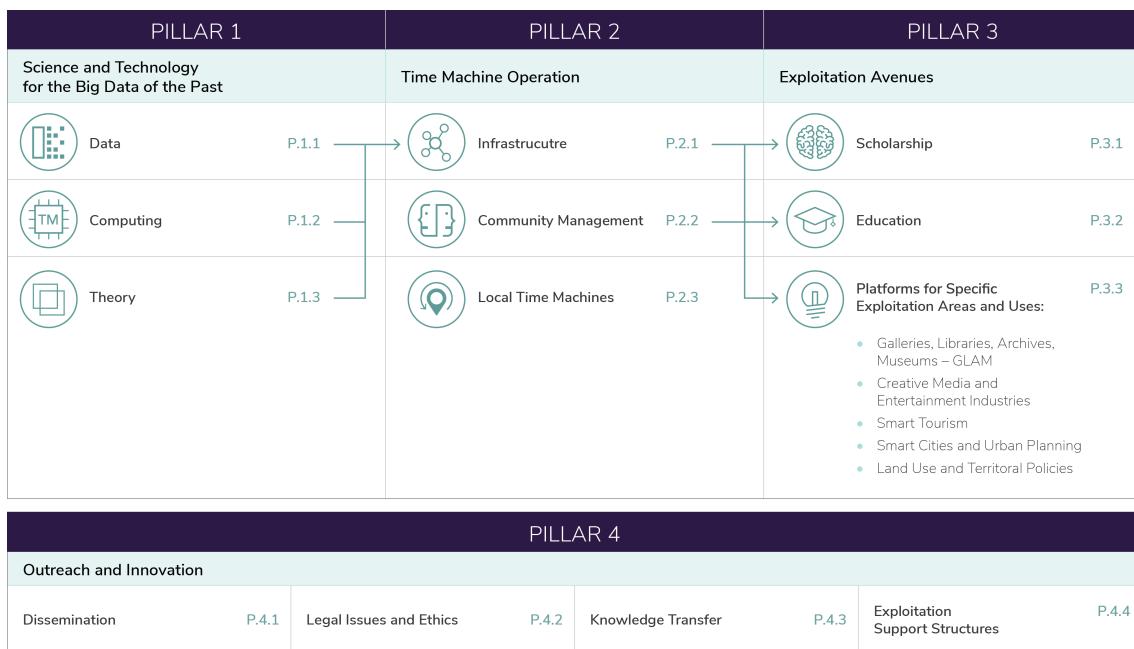


FIGURE 6.1 – Organisation et interactions des premiers *PILLARS* © Copyright 2019 Time Machine

Cette période du CSA est également rythmée par différentes conférences¹, visant à réunir les partenaires afin de faciliter la collaboration au sein des membres d'un même PILLAR ou WP, de faire connaître le projet et d'inviter les membres croissants à collaborer à ce processus décisionnel. En effet, le réseau Time Machine grandit en parallèle de la période du CSA et bien que les nouveaux membres ne soient pas liés contractuellement à l'élaboration du projet, ni rémunérés par l'UE, les échanges sont fortement encouragés. Lors de la conférence d'Amsterdam des 9 et 10 mai 2019, plus d'une centaine de participants ont fait le déplacement.

Bien que très vite, la perspective du milliard d'euros ait été retirée², les participants ont poursuivis le plan d'élaboration et des rendus fixés par l'UE, dans l'attente que les commissions concernées trouvent la manière appropriée de « faire atterrir » le projet, étant toujours liés contractuellement puisqu'ayant touché le million.

La période de notre stage coïncide avec la recherche de nouvelles formes de financement et de pressions auprès de l'UE conduites par le comité exécutif, en vue de demeurer un projet bénéficiant du soutien européen.

1. Nous participerons à la deuxième conférence, réunissant les groupes de travail des différents PILLAR, les 9-10 mai à Amsterdam

2. Détails dans le chapitre 3.4.1.

6.2 Déroulement du stage

Les missions du stage, initialement attachées à l'étude et à l'implémentation d'un des composants du réseau Time Machine précédemment mentionné, la *Time Machine Box*³ ont été modifiées dès notre premier jour. Notre maître de stage, ayant évalué qu'il serait plus intéressant de nous impliquer dans la conception du réseau de *Local Time Machines*, ce qui correspond au *PILLAR 2*.

Concrètement, nous avons eu pour mission de rendre début juin 2019 la première version de la feuille de route du projet à destination de l'UE, en charge de valider les différentes propositions. Une période de consultation des propositions contenues étant planifiée de juillet à septembre, avant le rendu final devant avoir lieu à la fin de l'année 2019.

Notre stage a consisté en la création de la première proposition de ladite feuille de route pour le *PILLAR 2*, en étroite collaboration avec notre maître de stage, des membres d'Icarus⁴ et le consultant du projet⁵. La version finalisée de la feuille de route devant être développée pour décembre, elle ne fait pas partie de nos missions de stage.

Le *PILLAR 2* se compose de trois grandes sections, la première traitant de l'infrastructure (sous la responsabilité du Professeur Frédéric Kaplan), la deuxième des communautés (sous la responsabilité des membres d'Icarus) et la troisième des enjeux organisationnels et opérationnels du réseau de Time Machines locales et par conséquent du projet Time Machine dans son ensemble. Nous fûmes responsable de la troisième section et de l'articulation des différentes parties les unes avec les autres, avec pour objectif de garantir la cohérence entre les propositions de recherche identifiées dans le *PILLAR 1*, et les impacts envisagés par le *PILLAR 3*. Mettre en place les opérations et l'infrastructure nécessaires à un réseau tel que Time Machine sous-tend de définir chacun de ses composants et leurs rôles respectifs et de précisément établir le cadre, ce que nous nous sommes efforcée de faire afin de pouvoir définir les directives associées.

La création de cette première version s'est déroulée suivant quatre étapes qui ont rythmé nos activités de stagiaire :

1. Définition

Clarification du contexte, des différentes instances et partenaires du réseau Time Machine et de leurs rôles respectifs.

2. Écriture

Transformation des idées définies en 1 en objectifs et réalisations concrètes. Rédaction de la feuille de route.

3. Intégration et validation

3. EPFL.DHLAB, *Home / TimeMachineBox...*

4. ICARUS, *ICARUS / International Centre for Archival Research...*

5. Vous trouverez ladite feuille de route sur la clé USB jointe au mémoire.

Consultation et évaluation des propositions par les partenaires et intégration des commentaires. Création de la cohérence entre les différentes feuilles de route proposées en *PILLAR 1* et *PILLAR 3*.

4. Consultation externe

Définition de questions ciblées pour la phase de consultation externe, ouverte au grand public.

Nous avons été aidée dans la réalisation de notre mission par les membres du DH-LAB et particulièrement par la coordinatrice de la *Venice Time Machine*, Dr. Isabella di Lenardo, membre également de l'unité de coordination du projet Time Machine⁶.

6. Nous participerons à un voyage d'étude à Venise (30 mai au 1^{er} juin 2019), pour découvrir les coulisses de la *Venice Time Machine*.

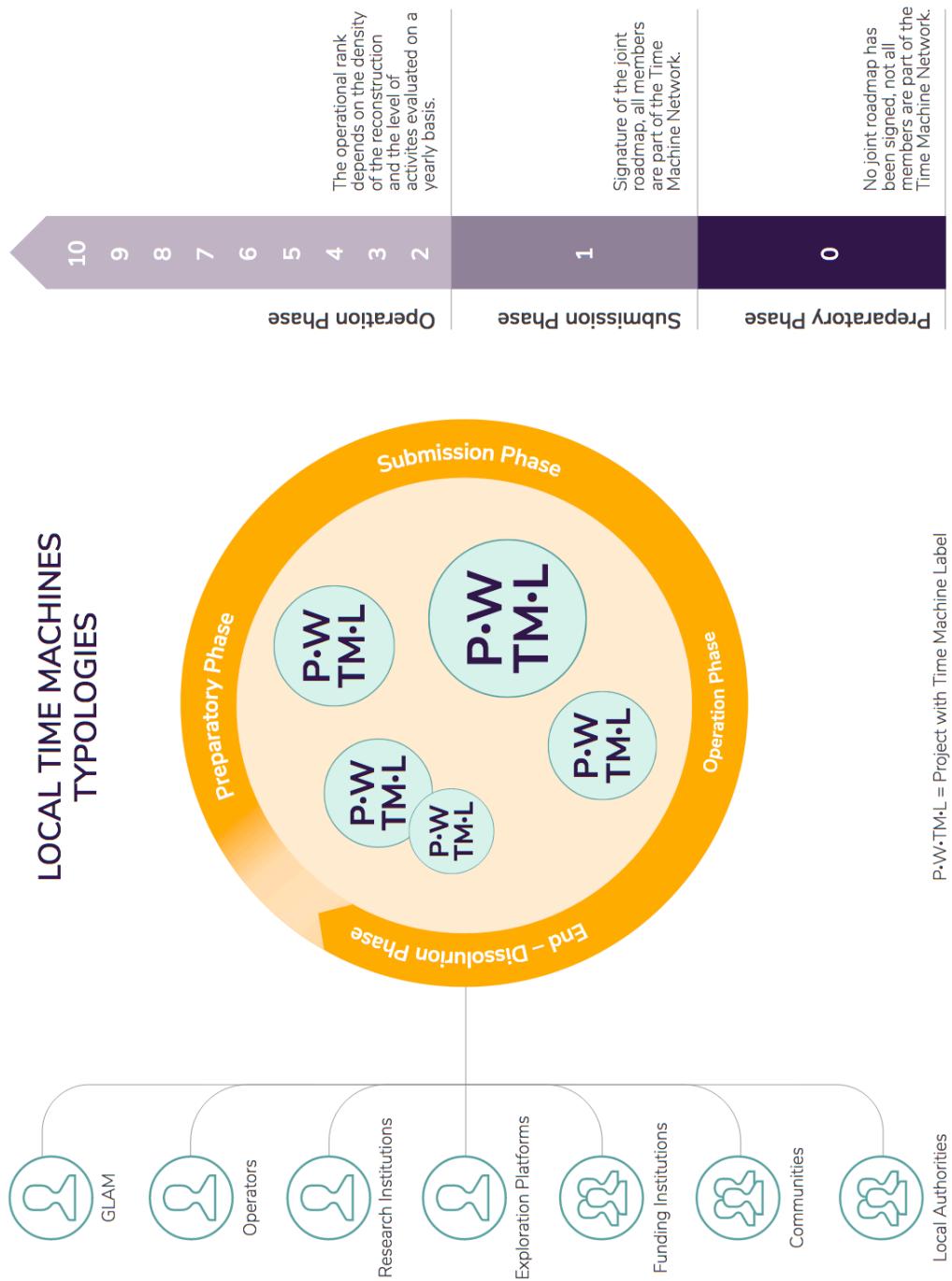


FIGURE 6.2 – Local Time Machines, document de travail © Copyright 2019 Time Machine

6.3 Rédaction de la feuille de route

L’élaboration de la feuille de route s’est très vite révélée une tâche ambitieuse, puisqu’elle implique à la fois l’intégration des idées motrices à l’initiative Time Machine, telles que pensées par le DHLAB, dans le contexte de la recherche en humanités numériques, et la prise en compte des envies et attentes des différents partenaires. Rédigée à destination d’un réviseur mandaté par l’UE et dans le but d’obtenir un financement, ce document doit à la fois faire rêver et concrétiser les idées de tous.

Nous nous sommes vite trouvée confrontée à la réalité que derrière le nom du projet et la quête commune visant à proposer un graphe du passé, se cachaient autant de définitions que de membres du réseau. Notre travail de stagiaire a consisté en un difficile exercice de transcription de tous ces intérêts parfois divergents, parfois audacieux, vers une feuille de route commune, servant de base au futur cadre des opérations, mais suffisamment flexible et vaste pour intégrer les inévitables changements induits par sa longue durée et son financement des plus incertains. Mener à bien cette mission nous a poussée à comprendre très précisément les différents composants du futur réseau Time Machine, qu’ils soient d’ordre technique ou opérationnel. Le projet étant prévu pour se déployer sur dix ans, la feuille de route reste à un niveau relativement macro, puisqu’il est difficile de prévoir certains choix techniques ou stratégiques avant que l’infrastructure, le financement et la gouvernance ne soit eux-mêmes clarifiés.

Au début de notre stage, le projet en est à ses prémisses, tout reste à définir et les nombreuses questions que nous nous posons ne font que s’ajouter à la longue liste de celles auxquelles nous devons proposer une réponse, et celles qui sont posées par les membres du comité exécutif. Le temps n’est d’ailleurs pas aux palabres et à la réflexion, puisque l’échéance du rendu intervient tôt, nous n’avons malheureusement pas eu l’opportunité de mener l’intégralité du travail réflexif présenté dans ce mémoire avant le rendu de la feuille de route. Ce travail de recherche nous a d’ailleurs permis de remarquer certains manques qui seront signalés.

Notre méthodologie de travail s’est construite autour des éléments et des enjeux figurant dans la première proposition du projet faite à l’UE, complétée par nos discussions avec les membres du consortium, du comité exécutif, les équipes du DHLAB et notre maître de stage.

Travaillant au sein d’un environnement ouvert où chaque membre peut suivre la construction des idées de manière transparente, nous avons également bénéficié de remarques ciblées émanant de certains spécialistes partenaires du consortium. Les premières semaines de notre stage ont consisté en un grand travail de définition, afin de proposer un socle commun construit autour d’une clarification du contexte des différents éléments du projet, sur lequel chaque WP ou *PILLAR* peut venir appuyer ses idées. Ce travail a conduit à l’élaboration de différents graphiques et documents, qui trouvent une version

plus synthétisée dans la deuxième partie de la feuille de route, *Pillar 2 Key concepts*.

Nous avons également été confrontée à la difficulté que chacun des trois premiers *PILLARs*, bien que planifiés pour être écrits en parallèle, était destiné à influencer les choix des autres. Veiller à articuler ces propositions pour garantir une harmonisation des processus (entre les pistes de recherche identifiées dans le *PILLAR 1* et les impacts envisagés par le *PILLAR 3*) a impliqué des ajustements jusqu'aux derniers jours avant le rendu. Le détail de ces interactions figure également dans la deuxième partie de la feuille de route, *Interaction of pillar 2 with pillars 1 and 3*.

De manière générale, la construction de la feuille de route a nécessité une grande flexibilité, et de nombreuses versions ont vu le jour avant d'aboutir à celle introduite dans ce mémoire et disponible en annexe⁷. Écrit à plusieurs mains, ce document, à la hauteur des projets de numérisation, est complexe, et chaque choix pourrait être analysé.

Pour ne pas alourdir inutilement ce rendu et rester dans le cadre de notre problématique, nous ciblons notre présentation sur les réponses apportées par Time Machine aux enjeux de la numérisation. Bien que ces derniers ne soient pas décrits ainsi dans la feuille de route, les différentes solutions développées dans la partie trois *Research and Innovation Plan*, complétés des éléments proposés par les autres *PILLARs* et WPs contribuent à y répondre et offrent un aperçu cohérent de notre expérience de stagiaire.

La feuille de route figure sur la clé USB accompagnant ce mémoire.

7. La version soumise à l'UE fait quelque 60 pages et est écrite en anglais. Vous pouvez la consulter sur la clé USB accompagnant ce mémoire.

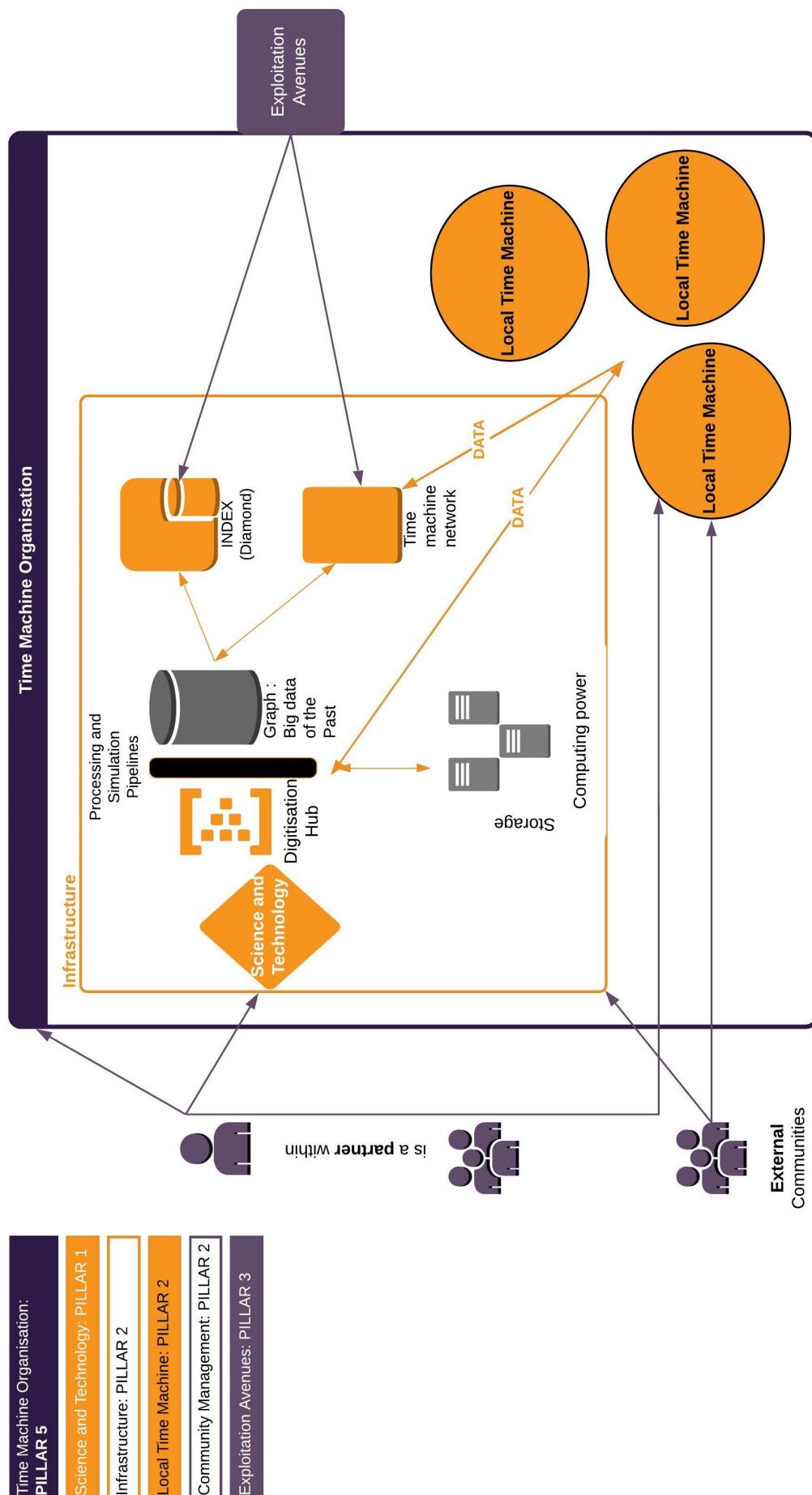


FIGURE 6.3 – Les différents composants de la *Time Machine Organisation*, document de travail

Chapitre 7

Quelles réponses aux enjeux de la numérisation ?

Le projet Time Machine étant encore dans un processus de définition, nous ne pouvons garantir que les propositions énoncées ci-après seront réellement implémentées après la période du CSA. Toutefois, elles reflètent les premières hypothèses et volontés organisationnelles, qui seront amenées à se concrétiser plus précisément au fil de l'avancement du projet.

7.1 Amener différents acteurs à collaborer

Time Machine est par essence un projet collaboratif qui vise à agréger les données de différentes institutions culturelles et patrimoniales.

Né d'une association de 33 institutions, le consortium compte désormais plus de 300 membres. Pour parvenir à co-construire ce projet, la collaboration est centrale à différents niveaux : entre les partenaires appelés à concevoir l'organisation du projet et à partager des ressources ; entre le projet et les instances extérieures qu'elles soient politiques, académiques, privées, publiques ; entre le projet et le grand public. Afin de préserver ces relations, différentes démarches ont été entreprises et le déploiement des solutions continues dans la feuille de route devront contribuer à préserver le futur de cette coopération.

Conscients de l'importante de la collaboration pour une initiative de cette envergure et par conviction du pouvoir de l'intelligence collective, la collaboration sera intégrée aux futures valeurs du projet. Différents moyens seront mis en place afin de promouvoir l'existence d'espaces d'échanges et encourager un décloisonnement des activités au profit d'une logique de groupe. Un système de label sera par exemple développé afin d'encourager les différents projets composant Time Machine à favoriser cette pratique (des moyens qualitatifs serviront à l'évaluer).

Afin de favoriser l'acceptation, le développement et la mise à jour des futurs stan-

dards du réseau, un processus inspiré de celui utilisé pour définir les standards d'internet sera adopté : Request for Comments (RFC)¹. Ce système consiste en la publication en ligne de propositions, dont les thématiques, voir les pistes de réponses, ont d'abord été définies par un comité d'édition en fonction des besoins du projet. Publiées de manière transparentes, elles sont soumises à la validation et aux critiques de la communauté, et ne sont implémentées que lorsque le consensus est atteint. La feuille de route contient de nombreuses propositions de RFCs visant à mettre en place l'infrastructure et les opérations.

Le projet s'inscrivant dans une démarche de transparence, les composants techniques, ainsi que toute la documentation du projet seront accessibles sous licences *Creative Commons*, favorisant leurs réutilisations et le développement de nouveaux services. Le partage des connaissances et les développements technologiques font partie des objectifs du projet, qui s'inscrit dans la démarche d'Open Science.

Partenaire de grands réseaux culturels et patrimoniaux (à l'instar d'Europeana), Time Machine s'attachera à mettre en place des moyens dédiés à favoriser synergies et interactions entre ces différents acteurs, favorisant les échanges de bonnes pratiques et contribuant à l'accroissement des activités de ces réseaux et le développement d'outils communs.

Les communautés de bénévoles spécialistes ou amatrices seront invitées à prendre une part active au sein des projets. Inscrivant Time Machine dans la mouvance des sciences citoyennes ou *citizen science*.

Au sein de chaque Time Machine locale, un espace *Smart Clusters*, servira à rassembler les potentiels futurs exploitants (industries culturelles et créatives, industries touristiques, instances politiques régionales etc.) et les partenaires de Time Machine, afin d'impliquer les premiers dans la conception de projets, et assurer un ancrage de l'initiative locale au sein de son environnement et le déploiement d'une partie de ses activités autour de particularités et besoins ciblés.

7.1.1 Complément aux propositions de la feuille de route :

Les données produites au sein des Time Machines locales étant destinées à alimenter le graphe du passé, ces dernières constituent un réseau de partenaires (agrégateurs) stratégiques, avec lesquels il faut particulièrement soigner la collaboration. La mise en place d'un forum similaire à celui d'Europeana ou DPLA favorisant les échanges devrait ainsi être considéré. La place des agrégateurs thématiques et régionaux déjà existants devra également être réfléchie.

1. Pour en savoir plus : https://en.wikipedia.org/wiki/Request_for_Comments

7.2 Financement et partenariats public-privé

Un modèle de franchise sera mis en place, permettant aux différents partenaires de bénéficier de l'infrastructure de Time Machine, de formation et d'aide ciblée tout en contribuant à l'autonomie financière du projet. Ce dernier tiendra compte de différents critères et visera à clarifier les relations entre les Time Machines locales et la future gouvernance.

Au-delà des partenariats financiers, le réseau mise sur la création de la *Time Machine Infrastructure Alliance*, regroupant des partenaires extérieurs au projet mais intéressés à contribuer au développement de nouvelles technologies, et des philanthropes, qui seront invités à participer par des moyens plus concrets (puissance de calcul informatique, espace de stockage) à la durabilité du projet.

Les différentes Time Machines locales seront à priori responsables du financement de leurs projets.

Le consortium étant déjà composé d'un mélange entre partenaires publics et privés, cette spécificité sera conservée au sein de la future *Time Machine Infrastructure Alliance*. Une attention particulière sera portée sur ces partenariats afin de garantir la mission de transparence et d'ouverture que s'est donné le projet.

Il est possible que des institutions privées souhaitant bénéficier des outils de Time Machine puissent le faire, sans que leurs données soient utilisées pour alimenter le futur graphe de données du passé, ces dernières devront payer pour ce service en fonction d'une tarification préétablie.

Le financement servant à développer l'infrastructure du projet, n'est pas encore connu, mais le projet est en quête du soutien de l'UE.

7.3 Droit d'auteur

Time Machine est amené à traiter deux typologies différentes de données : les privées et les publiques. Les données publiques (produites par des institutions publiques, ou données libres de droit) seront partagées sous licences *Creative Commons*, de préférence CC0. Les données privées (produites par des particuliers ou des institutions privées, données personnelles, ou données sous droit) seront protégées par des licences plus strictes et leur consultation pourra être sujette à certaines restrictions. Les données à caractère sensible (violentes ou illicites) publiées sous licence *Creative Commons*, seront également associées à des restrictions d'accès.

Les différentes licences seront explicitées dans les métadonnées de chaque document.

Les outils et processus permettant de faire face à ces typologies de données restent encore à déterminer.

7.3.1 Complément aux propositions de la feuille de route :

La feuille de route ne mentionne pas précisément le besoin de créer et rendre visible un système de déclarations des droits d’usage pour améliorer l’accès aux ressources. Le traitement des œuvres orphelines ou hors commerce n’est pas non plus abordé alors que, suite aux exemples étudiés, nous préconiserions l’étude d’une stratégie de gestion des risques pour ces œuvres au statut légal incertain.

7.4 Sortir des silos : enjeux techniques

Europeana étant partenaire du réseau Time Machine, à priori son modèle de métadonnées (suivant les recommandations du modèle RDF) sera utilisé au sein du projet (pour les données non déduites) et l’architecture de Time Machine sera interopérable avec celle d’Europeana. Un modèle pour les données déduites sera élaboré, s’assurant de faire figurer explicitement les processus de création de ces informations dans leurs métadonnées. Une charte technique, contenant ces différentes spécifications devra être signée par tous les partenaires du réseau Time Machine. Time Machine prendra part aux réflexion menées sur les vocabulaires contrôlés² et les ontologies³

En plus de la standardisation des métadonnées, Time Machine vise à normaliser les protocoles de numérisation, afin de garantir une qualité et une cohérence à travers tous les centres de son futur réseau européen de numérisation.

Time Machine rendra obligatoire l’utilisation de IIIF pour les images, veillera au déploiement de différentes APIs et utilisera le protocole OAI-PMH .

7.4.1 Complément aux propositions de la feuille de route :

La qualité de la numérisation devra être suffisamment élevée pour garantir un stockage sur le long-terme des données numérisées et les critères de cette qualité précisément étudiés.

En plus de l’usage de IIIF, OAI-PMH et autres APIs pour garantir l’interopérabilité, il serait judicieux d’offrir la possibilité aux futurs exploitants de disposer d’une exportation totale des bases de données (suivant l’exemple de DPLA).

2. Rappel : Lexique servant à organiser les connaissances pour favoriser la recherche d’information

3. « Rappel : Une ontologie est une façon de modéliser un domaine en identifiant les concepts y afférents et en les organisant les uns par rapport aux autres N. Delestre, N. Malandain et M. Bussi, *Du Web des documents au Web sémantique...* »

7.5 Sortir des silos : enjeux sur le contenu

Mise en place d'un système de sélection des données (politique documentaire) s'inspirant des recommandations élaborées par *The National Information Standards Organisation*⁴, qui sont basées sur des critères liés aux besoins de recherche des institutions, unicité des documents, valeur du document etc. Bien qu'ayant vocation à numériser toutes les données du passé, il est probable que certains documents servant de base au développement de la future plateforme soient numérisés en premier (tels que les cadastres, journaux, cartes et plans). Time Machine ne se focalise pas sur un seul format de document et veut développer particulièrement la numérisation 3D.

7.5.1 *Complément aux propositions de la feuille de route :*

Les manques hérités des politiques de numérisation des institutions partenaires, ou les différents biais qui peuvent être amenés à persister dans de nouvelles initiatives, devraient être étudiés et certains projets spécifiquement entrepris afin de rétablir l'équilibre.

Certaines mesures spécifiquement dédiées à faciliter la numérisation de données à partir du 20^e siècle devraient être examinées afin de contribuer à sortir l'Europe de ce « trou noir » de l'information.

7.6 Stockage sur le long-terme - préservation

La préservation sur le long-terme des données de Time Machine sera prise en compte par le projet, afin que l'énorme travail de numérisation n'encourt pas le risque d'être perdu.

Time Machine propose des solutions de stockage à destination des petites institutions qui ne disposent pas d'espace suffisant sur leurs propres serveurs : les *Time Machine Box*. Ces dernières existent à l'état de prototype et devront être améliorées (elles ne disposent pas encore de copie de sauvegarde).

Les données du projet seront stockées de manière partagée (à travers les ressources de la *Time Machine Infrastructure Alliance*) sur des serveurs spécialement conçus pour garantir leur préservation durable et prévenir tous risques d'accidents ou de dommages.

Le poids des jeux de données et l'impact environnemental des serveurs faisant partie des préoccupations du projet, des solutions de stockage innovantes sont aussi envisagées.

4. Pour en savoir plus : <https://www.niso.org/>.

7.7 Résumé des réponses de Time Machine aux enjeux de la numérisation

Enjeux	Réponses de Time Machine
Amener différents acteurs à collaborer	Inscription de la collaboration dans les valeurs du projet et création d'un système de labels. Utilisation des <i>Request for Comments</i> . Publication et partage des différents développements sous licence <i>Creative Commons</i> . Échange avec les réseaux partenaires. Intégration du grand public (sciences citoyennes). Création de <i>Smart Clusters</i> . Création d'un forum favorisant les échanges entre les Time Machines locales et la gouvernance.
Financement et partenariats public-privé	Moyens financier assurés par les partenaires (redevance). Soutien en nature possible à travers la <i>Time Machine Infrastructure Alliance</i> . Vente de services à l'extérieur du projet. Les Time Machines locales sont responsables du financement de leurs projets. Partenariats public-privé déjà existants et amenés à se développer. Recherche de financement pour l'infrastructure auprès des instances européennes.
Droit d'auteur	Favorable à l'usage des licences <i>Creative Commons</i> . Restriction d'accès dans le cadre de données privées ou sensibles. Métadonnées explicites. Déploiement d'un système de déclaration des droits d'usage.
Sortir des silos : enjeux techniques	Adoption du modèle de métadonnées développé par Europeana et interopérabilité avec sa plateforme. Recours à des vocabulaires contrôlés. Utilisation des différents outils tels que APIs, IIIFs, OAI-PMH. Création de normes de qualité pour la numérisation, amenées à se déployer à travers les futurs centres de numérisation du réseau Time Machine.
Sortir des silos : enjeux sur le contenu	Politique documentaire. Intégration des différentes typologies de données et développement spécifique de la 3D. Analyse des manques existants et actions ciblées pour y pallier.
Stockage sur le long-terme - préservation	Développement de solutions de stockage pour les petites institutions. Le stockage des données sur le long-terme est essentiel. Les données seront stockées pour être préservées sur les serveurs de la <i>Time Machine Infrastructure Alliance</i> .

TABLE 7.1 – Résumé des réponses de Time Machine aux enjeux de la numérisation

Chapitre 8

Les innovations du projet

En quoi Time Machine représente une solution innovante par rapport à d'autres initiatives de numérisation de masse ? Nous vous proposons un retour sur certaines ambitions du projet, qui, selon nous, contribuent à sa popularité toujours croissante parmi les institutions européennes.

8.1 Mélanger les acteurs

Une des forces de Time Machine, est d'avoir cherché à fédérer dès le départ en-dehors du cercle des institutions culturelles et patrimoniales. Représentatif des recherches conduites en humanités numériques, qui favorisent le mélange des pratiques, Time Machine réunit : le monde politique ; les industries créatives et touristiques ; les scientifiques de tous horizons ; les réseaux et infrastructures existants ; et le grand public. Tous ces acteurs sont invités à la création d'un projet commun, s'affranchissant des barrières et n'ayant pas crainte d'inviter les futurs exploitants à participer à la création de l'infrastructure.

S'offrant comme un objet d'étude malléable, le projet nourrit une ambition suffisamment grande (la numérisation du passé européen sous toutes ces formes), pour que chacun puisse envisager d'en bénéficier.

L'élargissement des différentes typologies de partenariats est pris en compte par tous les groupes de travail du projet, qui veillent à demeurer les plus inclusifs possibles.

8.2 Innovations technologiques

Porté par l'EPFL dont la renommée scientifique n'est plus à démontrer, Time Machine promet d'être à la pointe de la recherche en intelligence artificielle, mais également de contribuer à l'émergence et la démocratisation de nouvelles technologies, permettant à ses partenaires publics et privés d'en bénéficier. Rapide retour sur certaines de ces

ambitions technologiques :

1. Données enrichies et données déduites

L'automatisation des processus d'extraction des informations contenues dans les documents numérisés constitue le premier développement technologique, permettant la construction du graphe de données du passé : le *big data* du passé.

Les données numérisées sont ensuite complétées de données déduites à l'aide de trois moteurs de simulation : *4D Simulation Engine*¹, *Universal Representation Engine*², *Large-scale Inference Engine*³. Ainsi les technologies de l'intelligence artificielle permettront aux données du passé d'acquérir le même poids que les données de notre présent numérique.

L'introduction de ces données « déduites » et toutes les transformations technologiques qui en découlent constituent un des changements majeurs apporté par Time Machine.

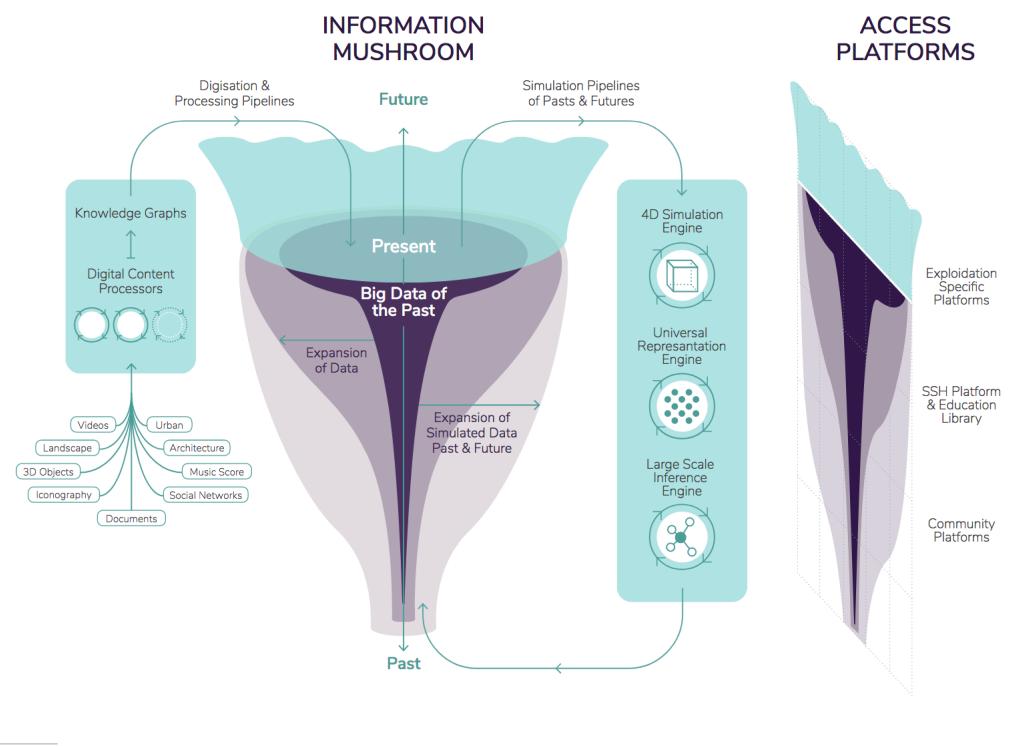


FIGURE 8.1 – Représentation schématique des processeurs de contenus et des moteurs d'inférence © Copyright 2019 Time Machine

1. Ce moteur permet de simuler tous les présents et passés possibles liés à une donnée numérisée, prenant compte du temps et de l'espace.
2. Permet la reconstitution d'espaces ou éléments physiques en fonction d'informations issues dans les données (textes, images, vidéos, 3D). Ce moteur est capable de reconstruire un bâtiment en se basant uniquement sur sa description, son origine géographique et son âge.
3. Permet de déduire des informations en reliant toutes les données de la base, ce moteur d'inférence permet de corroborer les informations déduites par le *4D Simulation Engine*

2. Le facteur temps

En ajoutant la temporalité aux outils de découverte traditionnellement déployés par les GLAMs, Time Machine enrichira l'expérience utilisateur. Son moteur de recherche diachronique ouvrant de nouvelles perspectives de recherches et de valorisation des collections.

3. 3D et photogrammétrie

Le DHLAB s'est spécialisé ces dernières années dans la reconstruction 3D du patrimoine. La photogrammétrie sera au centre des activités de la future initiative qui ne résume pas l'acte de numérisation du patrimoine aux objets 2D, mais promet de mettre l'accent sur la numérisation du patrimoine bâti et des collections muséales.

Au même titre que Time Machine vise à organiser des centres de numérisation à travers l'Europe afin d'harmoniser les coûts et assurer un certain niveau de qualité, la photogrammétrie bénéficiera de son réseau de spécialistes européens.

4. Archivage dans l'ADN

Les données 3D sont très lourdes et l'ampleur du projet est telle, que de nouvelles solutions de stockage doivent être trouvées afin de permettre une préservation et une consultation de qualité.

Le projet travaille avec différents partenaires actifs dans cette thématique, dont les propositions futuristes influenceront probablement notre monde de données de demain. L'archivage de données dans l'ADN fait partie de ces technologies prometteuses, permettant de stocker de manière stable et durable des informations, tout en minimisant l'impact écologique alloué à leur préservation.

5. Outils de numérisation

La marché de la numérisation est encore dans une phase de construction, de nouvelles technologies seront amenées à améliorer les performances obtenues et à s'adapter à la fragilité de certaines typologies de documents. La tomographie⁴ constitue un outil des plus prometteurs, permettant de numériser un objet sans l'ouvrir et d'y déceler des particularités (liées aux fibres du papier etc.) jusqu'alors non visibles par une numérisation plus traditionnelle. Similaire à une radiographie, l'outil est sensible au taux de fer présent dans l'encre des manuscrits et permet d'obtenir des images en coupe de l'intégralité d'un volume. Time Machine promet de contribuer à la recherche sur ces nouveaux outils, et à leurs déploiements.

4. ARTE, *Tomography - The Digitisation of the Future (3-8) - Venice Time Machine*, en, URL : <https://www.arte.tv/en/videos/075631-003-A/tomography-the-digitisation-of-the-future-3-8/> (visité le 18/05/2019).

Chapitre 9

Risques et opportunités

Au-delà des réponses aux enjeux de la numérisation, Time Machine doit faire face à certains risques et opportunités qui découlent de son contexte de création, mais sont aussi induits par les évolutions technologiques et les nouveaux services que le projet se propose de déployer.

9.1 Plateforme, usages et accessibilité

La masse des documents numérisés permet à la fois des recherches orientées, mais offre aussi de nouvelles méthodes de découvertes « [...]across genres and disciplines, as well as across institutional and national borders.¹ ».

Au-delà de l'objectif de création de ce corpus de données, se pose la question de son accessibilité. Comment s'assurer que le travail mis en place par Time Machine puisse convenir aux usages des futurs utilisateurs ? Les nouvelles générations semblent peu enclines à la lecture, mais sont entraînées à la communication interactive², seront-elles dès lors à même de s'approprier le contenu et les interfaces de Time Machine ? Comment ne pas se retrouver dépassé par cette masse de données, pour mieux les présenter et les trier³. Avec la numérisation de masse, la capacité humaine de curation de ces collections numériques croissantes, semble dépassée. Les capacités cognitives des citoyens ordinaires et la puissance informatique sont de plus en plus sollicitées pour trouver du sens à cette accumulation de données⁴. Les fournisseurs de contenu ne peuvent plus se permettre de choisir un outil parmi d'autres, mais se doivent de construire avec tous les standards existants (API, OAI-PMH, RDF etc.) pour s'assurer de délivrer leurs données aux plus larges audiences⁵.

1. N. B. Thylstrup, *The politics of mass digitization...*, p.123.

2. *Numérisation du patrimoine...*

3. I. Xie et K. K. Matusiak, *Discover digital libraries...*

4. N. B. Thylstrup, *The politics of mass digitization...*

5. A. Dunning, “Digitising the past : Next steps for public-sector digitisation”...

Le projet, à l'instar des précédentes initiatives, propose de réfléchir à de nouvelles formes d'outils de découvertes, promettant de nouvelles expériences aux utilisateurs. « *An important question for builders of mass digitization projects has therefore been how to build visual and semantic infrastructure that offer the user a sense of meaningful direction as well as a desire to keep browsing.*⁶ »

9.1.1 Franchir les barrières territoriales

S'adressant à une communauté européenne, Time Machine doit être capable de gérer plusieurs langues. Ses utilisateurs souhaiteront pouvoir filtrer les résultats en fonction du langage et effectuer des recherches dans leurs propres idiomes⁷. Le déploiement de solutions techniques devra tenir compte des particularités culturelles propres à chaque région et permettre une méthode de partage des données qui traduise automatiquement une requête dans toutes les langues et adresse des résultats provenant de chaque communauté linguistique⁸.

La future plateforme du projet devra veiller à ne pas cloisonner ces résultats par le déploiement de normes occidentales qui préviendraient l'usage des produits à l'extérieur de sa propre communauté⁹.

9.1.2 Être mobile

La majeure partie du trafic d'Europeana étant générée par des mobiles, et étant donné l'évolution des pratiques et usages, le projet devra développer au plus tôt une telle interface^{10 11}.

9.1.3 Favoriser la cocréation

Pour s'assurer de l'adéquation de la plateforme avec les pratiques des différentes communautés et garantir que ce projet, tourné vers différents publics, ne vienne pas à fonctionner comme la vitrine d'un projet de recherche mené par une communauté professionnelle¹², Time Machine se doit de questionner les méthodes de développement traditionnelles et impliquer les usagers dès le début du projet, afin de proposer des services à même de satisfaire leurs besoins interactionnels¹³.

6. N. B. Thylstrup, *The politics of mass digitization...*, p.110.

7. I. Xie et K. K. Matusiak, *Discover digital libraries...*

8. Anne R. Diekema, “Multilinguality in the digital library : A review”, *The Electronic Library*, 30-2 (avr. 2012), p. 165-181, DOI : 10.1108/02640471211221313.

9. E. Jones, “The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative...”.

10. I. Xie et K. K. Matusiak, *Discover digital libraries...*

11. L'interface mobile de *Diamond* est en cours de développement.

12. A. Moatti, “Bibliothèque numérique européenne : de l'utopie aux réalités”...

13. A. Dunning, “Digitising the past : Next steps for public-sector digitisation”..., p.7.

De nombreuses études (instaurées dès la fin des années 1990), ont cherché à comprendre pourquoi les ressources mises en place dans le contexte des humanités numériques n'étaient que peu utilisées par les chercheurs. Ces dernières ont conclu que leurs pratiques ne correspondaient que peu aux recherches empiriques, qui s'intègrent difficilement dans leurs routines de travail. De plus, les usagers semblent se désintéresser rapidement des outils numériques s'ils ne parviennent pas à accéder dès le départ à une information pertinente.^{14 15}.

Les outils mis en place alors ne s'intéressaient pas du tout aux besoins et usages de leurs utilisateurs, qui étaient considérés comme incapables de comprendre les enjeux du numérique et dès lors de proposer leurs avis

Pour pallier à ce manque d'intérêt et d'adoption des données numériques par les chercheurs en humanités, il faut penser la création des plateformes dans la continuité des pratiques des communautés (ce qui signifie les étudier au préalable et les impliquer dans le processus de création) tout en offrant de nouveaux avantages. « *If digital resources fit well with what they [users] want to do with them, users will adopt them.*¹⁶ ». « La plateforme fournit une infrastructure ouverte et participative pour des interactions créatrices de valeurs entre des producteurs et des consommateurs externes, dans le cadre des conditions de gouvernance définies par celles-ci.¹⁷ »

Les pratiques ne vont pas évoluer d'un coup, il faudra du temps pour que les communautés scientifiques adoptent véritablement ces nouveaux outils. Elles tendent toutefois à évoluer, poussant par exemple les historiens à s'adapter aux nouvelles méthodes d'analyse mises à leur disposition et à développer leurs outils de travail, les faisant prendre conscience de la nécessité de s'impliquer dans les processus de création de ces nouvelles ressources informationnelles. « If historians do not develop their tools themselves and embrace the goals of digital humanities, they are in danger of having method forced on them that are not compatible with their practice¹⁸. »

Les données doivent être mises à jour régulièrement, si les utilisateurs ont le sentiment que ces dernières sont dépassées ou datées, ils auront encore plus de difficultés à se les approprier. Leur préservation sur le long-terme doit aussi être garantie, car la recherche scientifique implique de pouvoir réutiliser des jeux de données pour comparer les résultats obtenus. La révolution numérique a instauré le doute concernant la durabilité

14. Claire Warwick, Melissa M Terras et Julianne Nyhan, *Digital humanities in practice*, OCLC : 836872277, London, 2012, URL : http://www.123library.org/book_details/?id=92814 (visité le 10/06/2019).

15. C. Warwick, “Studying users in digital humanities”, dans *Digital Humanities in Practice*, dir. Claire Warwick, Melissa Terras et Julianne Nyhan, 1^{re} éd., 2012, p. 1-22, DOI : 10.29085/9781856049054.002.

16. *Ibid.*, p.18.

17. D. Battu, *L'histoire et l'économie du monde accompagnées par les TIC...*, p.100.

18. F. Clavert et S. Noiret, *L'histoire contemporaine à l'ère numérique Contemporary history in the digital age...*, p.25.

et la reproductibilité de ces ressources¹⁹. Si ces critères de mise à jour et de préservation ne sont pas respectés, il ne sert probablement à rien d'investir dans la création du projet. « *This is a waste of the (probably) very large amount of money that was spent on its creation. Institutions have only recently begun to develop strategies to deal with this problem.*²⁰ » La pauvreté des informations constitue également une menace, si les jeux de données sont construits uniquement pour un contexte précis, ils ne sauront être utilisés pour d'autres usages. Il faut garantir leur accessibilité intellectuelle afin de prévenir toute mauvaise compréhension des informations²¹.

Au-delà du public scientifique, les voix des communautés et des individus issues du « grand public » peinent également à être entendues au sein des projets de numérisation de masse²². La plateforme visant à s'adresser à la plus vaste audience possible, les publics empêchés doivent également être invités à participer au développement des futures interfaces.

Certains auteurs appellent à la construction d'une cyberinfrastructure au service de l'histoire, qui sera coconstruite avec les chercheurs en humanités et s'appuiera sur l'héritage des précédentes initiatives de numérisation, afin d'éviter que le savoir universel demeure détenu, de façon monopolistique, par quelques compagnies privées et puisse bénéficier à tous. Cette cyberinfrastructure devra permettre la création de nouveaux standards capables d'inviter les utilisateurs à chercher de l'information et à la transformer en connaissance²³.

Il semble que Time Machine soit dès lors bien placé pour répondre à ce souhait, pour autant qu'il soit capable d'intégrer ses différents publics-cibles dans ses processus de création et d'inviter le grand-public à se positionner aux côtés de la communauté scientifique.

9.1.4 Moteur de recherche ou plateforme de découverte ?

Time Machine doit se positionner face aux géants de l'information déjà existants, qui ne travaillent pas uniquement à la création de corpus mais à l'indexation du web dans son ensemble. Entre site internet et moteur de recherche, le débat est encore vif et n'apporte que peu de réponse. Certains argumentent qu'un site ne saura rivaliser avec un moteur capable d'en indexer des millions²⁴ lorsque d'autres s'interrogent sur la valeur en terme de diffusion, de fonds numérisés ou de moteurs de recherche²⁵. Il semble qu'à ce

19. *Ibid.*

20. C. Warwick, “Studying users in digital humanities”..., p.14.

21. F. Clavert et S. Noiret, *L'histoire contemporaine à l'ère numérique Contemporary history in the digital age...*, p.67.

22. M. Thelle et N. Bonde Thystrup, “Persuasive territories in European cultural politics...”.

23. F. Clavert et S. Noiret, *L'histoire contemporaine à l'ère numérique Contemporary history in the digital age...*

24. A. Moatti, “Bibliothèque numérique européenne : de l'utopie aux réalités”...

25. *Numérisation du patrimoine...*

jour, la seule méthode consiste en une forme de collaboration avec ces autres géants, et au déploiement de méthodologie de référencement²⁶, que nous n'aborderons pas plus en détails dans ce présent mémoire.

9.2 Crowdsourcing et sciences citoyennes

Avec l'avènement du web social, de nouveaux moyens de création de ressources informationnelles ont vu le jour, sous la forme du *crowdsourcing* d'abord, reflétant le transfert de processus de travail vers une main d'œuvre composée d'internautes, puis évoluant vers l'approche plus participative des sciences citoyennes ou *citizen science*. Les entreprises de numérisation de masse et les GLAMs, utilisent traditionnellement les compétences de ces communautés ou individus pour transcrire, corriger ou indexer une sélection de matériel²⁷.

Le *crowdsourcing* n'est toutefois pas apparu avec l'avènement du numérique, puisque déjà en 1859 la *Philological Society* s'adressait aux citoyens afin de créer ce qui deviendra la première version du *Oxford English Dictionary*²⁸.

De nombreux projets ont recours au *crowdsourcing* pour la création des collections à numériser, invitant les individus à déposer leurs archives personnelles. Cette pratique tend à augmenter l'intérêt et la visibilité de ces initiatives, bien que suscitant des débats sur la qualité et la pertinence de ces nouveaux contenus²⁹. Ces communautés sont également appelées à collaborer dans le déroulement des activités de photogrammétrie liées à la numérisation 3D, qui offrent de nouvelles opportunités à forte valeur ajoutée « [...]which has particular intrinsic and instrumental value for themselves, and for their museums in their communities.³⁰ », dans un contexte où les institutions n'ont souvent pas les moyens pour engager des équipes professionnelles³¹. Certains auteurs appellent à travailler avec les communautés d'expertise déjà existantes afin d'alléger les coûts engendrés par les recherches des détendeurs de droit des œuvres orphelines, favorisant leur numérisation et garantissant ainsi une marge d'erreur moins grande dans les résultats, puisque ces pratiques intègrent souvent des processus d'auto-vérification³².

Si le *crowdsourcing* peut prendre une grande variété de formes, en fonction des

26. Terme rassemblant toutes les actions visant à améliorer la visibilité d'un site web dans les résultats naturels d'un moteur de recherche. Ces actions impliquent notamment le déploiement de schéma de métadonnées spécifiques, l'usage de mots-clés etc.

27. M. Coutts, *Stepping away from the silos...*

28. Dinesh K. Gupta et Veerbal Sharma, “Enriching and enhancing digital cultural heritage through crowd contribution”, *Journal of Cultural Heritage Management and Sustainable Development*, 7-1 (févr. 2017), p. 14-32, DOI : 10.1108/JCHMSD-12-2014-0043.

29. M. Coutts, *Stepping away from the silos...*

30. Eugene Ch'ng, Shengdan Cai, Tong Evelyn Zhang et Fui-Theng Leow, “Crowdsourcing 3D cultural heritage : best practice for mass photogrammetry”, *Journal of Cultural Heritage Management and Sustainable Development*, 9-1 (févr. 2019), p. 24-42, DOI : 10.1108/JCHMSD-03-2018-0018, p.22.

31. *Ibid.*

32. V. Stobo, K. Patterson, K. Erickson, *et al.*, ““I should like you to see them some time”...”.

organisations et du contexte des projets, certains facteurs favorisent leur réussite : accessibilité des plateformes, coût, environnement de travail, ressources humaines etc.³³. Les projets ayant recours au *crowdsourcing* doivent être pensés dans la durée, puisqu'il est contreproductif de créer une communauté puis de la laisser tomber³⁴.

S'il faudra encore attendre quelques années pour bien comprendre l'articulation de ces sciences citoyennes ou *citizen science* au sein des projets de numérisation, il est indiscutable qu'elles auront un rôle à jouer. Le modèle économique de Google semble basé sur une certaine idée du *crowdsourcing* faisant appel à l'inventivité et l'intelligence collective du plus grand nombre, « [...]ces plateformes gratuites rebattent largement les cartes des règles commerciales et des frontières entre le marchand, le public, le commun et leurs différents domaines.³⁵ ». Certains y voient aussi le dernier rempart pour garantir l'accessibilité de toute la connaissance humaine et échapper aux logiques sélectives, dictées par le plébiscite des audiences légitimant certains contenus³⁶ :

Seules des « sciences citoyennes » qui échappent à l'élitisme tout en se gardant de faire le jeu du populisme peuvent faire contrepoids au projet de société globale de l'information porté par les géants du numérique, leur culture du résultat et du retour sur investissement à court terme. C'est là une condition nécessaire à l'essor de nouveaux usages démocratiques du potentiel du réseau des réseaux.³⁷

Time Machine a déjà pris en compte ces communautés dans la création de la feuille de route, il s'agira pour le projet de définir quelle forme et quel rôle attribuer à ces sciences citoyennes ou *citizen science*, garantissant à la fois la préservation d'une continuité démocratique, et porteuses de changement.

9.3 Pour un savoir cohérent, égalitaire et éthique

Au-delà des critères de sélection hérités des politiques documentaires des institutions, Time Machine a l'opportunité de contribuer à créer une connaissance humaine plus juste et égalitaire, qui ne reproduit pas forcément les inégalités hiérarchiques héritées de notre société humaine. Ce défi n'est pas des moindres, puisque jusqu'à présent les plateformes informationnelles ont au contraire servi à renforcer ces mécanismes.

[Traduction] Ce que nous avons constaté dans la pratique, c'est que les plateformes numériques ont la capacité de produire davantage d'homogénéité

33. D. K. Gupta et V. Sharma, “Enriching and enhancing digital cultural heritage through crowd contribution”..., p.27.

34. C. Warwick, “Studying users in digital humanities”..., p.15.

35. *Numérisation du patrimoine*..., p.37.

36. A. Mattelart, *Histoire de la société de l'information*..., p.107.

37. *Ibid.*, p.110.

plutôt que l'inverse, à reproduire les logiques hiérarchiques, inégalités et biais qui structuraient l'existence humaine dans la société avant l'avènement d'internet³⁸.

Le monde occidental étant très représenté au sein des acteurs de la numérisation de masse, Time Machine devra porter une attention particulière à la création de liens entre des réseaux de normes et de traditions culturelles différents de son propre milieu, afin de permettre une représentation plus objective de la connaissance qui, elle, s'affranchit aisément des frontières (ce biais impacte également le déploiement d'outils spécifiques à la numérisation, et le développement des plateformes d'accès)^{39 40}.

En tant que futur agrégateur des collections d'institutions diverses, Time Machine devra veiller à proposer des collections satisfaisantes pour les besoins d'une immense variété de publics. Trouver un équilibre entre la construction de la cohérence et l'exhaustivité à laquelle tend le projet sera l'un des défis de sa future politique documentaire.

De nombreuses questions éthiques sont également soulevées par les initiatives visant à numériser la connaissance universelle. Que doivent-elles faire concernant les objets aux contenus racistes, sexistes, appelant à la haine ou à la suprématie de certains peuples ?⁴¹ Comment garantir que le projet est fidèle aux critères moraux européens ? Si Time Machine veut réussir là où les précédentes initiatives ont échoué, il s'agira d'apporter des réponses à ces questions incontournables et d'offrir davantage qu'une restriction d'accès, comme solution au caractère sensible de certaines données.

38. « *What we have seen in practice is that online platforms have the capacity to produce more homogeneity rather than less, to cultivate and reproduce the hierarchical logics, inequalities, and biases that structured human existence in the world before the internet.* » Dorothy Kim, Treaandrea Russworm, Corrigan Vaughan, Cassius Adair, Veronica Paredes et Cowan, “Race, Gender, and the Technological Turn : A Roundtable on Digitizing Revolution”, dir. Everett et Guisela Latorre, *University of Nebraska Press*, *Frontiers : A Journal of Women Studies* 39–1 (), p. 149-177, p.22

39. A. Weiss, “Examining Massive Digital Libraries (MDLs) and Their Impact on Reference Services”...

40. E. Jones, “The Public Library Movement, the Digital Library Movement, and the Large-Scale Digitization Initiative...”.

41. A. Weiss, “Examining Massive Digital Libraries (MDLs) and Their Impact on Reference Services”...

9.4 Résumé des risques et opportunités pour Time Machine

Risques et opportunités pour Time Machine	Résumé
Plateforme, usages et accessibilité	<p>S'affranchir des barrières territoriales : multilinguisme et prévention contre l'usage de pratiques "occidentales". Être mobile.</p> <p>Favoriser la cocréation : impliquer les futurs usagers dans le développement d'une plateforme ouverte et participative, favorisant des interactions créatrices de valeurs. Veiller aux mises à jour et à l'accessibilité sur le long-terme. Intégrer l'héritage des précédentes initiatives.</p> <p>Moteur de recherche ou plateforme de découverte ? Trouver un positionnement adéquat et déployer des techniques de référencement.</p>
Crowdsourcing et sciences citoyennes	<p>Utiliser les compétences des communautés pour transcrire, corriger ou indexer les collections, sélectionner les documents, participer à la numérisation 3D (photogrammétrie), rechercher les détendeurs des droits d'œuvres orphelines.</p> <p>Garantir sa réussite, par le développement de plateformes accessibles, une prise en charge des coûts sur le long-terme, un environnement de travail valorisant, des ressources humaines dédiées.</p> <p>Définir le rôle du volontaire entre gardien de la démocratie et de l'accessibilité et briseur des frontières entre marchand, public et commun.</p>
Pour un savoir cohérent, égalitaire et éthique	<p>Ne pas reproduire les inégalités hiérarchiques de notre société.</p> <p>Créer des passerelles au-delà du monde occidental pour permettre une représentation nouvelle du savoir.</p> <p>Trouver un équilibre entre construction de la cohérence et quête d'exhaustivité.</p> <p>Adresser les questions éthiques posées par la création des collections et les objectifs du projet.</p>

TABLE 9.1 – Résumé des risques et opportunités pour Time Machine

Chapitre 10

Pourquoi numériser en masse ?

Nous avons essayé à travers ce mémoire, de montrer la complexité des entreprises de numérisation de masse. Du passage à l'échelle des premières entreprises de numérisation jusqu'à l'avènement des big data, ces projets ont franchi les portes de leurs institutions pour se mêler à l'agenda politique et prendre une part active dans la définition des standards du web. Porteurs de nombreux intérêts mélangeant biens communs et usages commerciaux, ils semblent rassembler sous la même bannière, des idées à priori opposées et des acteurs issus du monde public et du monde privé. Leurs poids financiers et les débats légaux qu'ils suscitent les poussent à inventer de nouvelles formes de partenariats et confrontent la sphère politique aux usages du numérique. Time Machine, bien que né dans un centre de recherche en humanités digitales, fédère au-delà de son milieu académique, s'affranchissant des barrières nationales, et à travers ses ambitions de recherche et de valorisation, semble incarner les questionnements de tous sur notre société numérique. Le projet, comme ses précurseurs, ne se laisse pas résumer par une seule question. Dernière initiative de la lignée des projets de numérisation de masse, Time Machine entend préserver le passé tout en modifiant l'avenir.

Au-delà des aspects techniques et opérationnels liés à l'envergure de ces projets, il semble urgent de considérer les questions du *pourquoi*, afin de mieux comprendre l'impact d'initiatives de telle ampleur sur l'organisation de notre société et les industries qui la composent. Les questions soulevées sont nombreuses et nous ne prétendrons pas pouvoir leur apporter à toutes un début de réflexion, nous avons préféré introduire ce débat à travers deux questions qui nous semblent peu abordées par la communauté scientifique dans le cadre de tels projets¹ : la surveillance de masse et le positionnement de ces initiatives par rapport aux acteurs culturels et patrimoniaux plus traditionnels.

1. Le temps que nous avons eu à disposition pour la rédaction du mémoire étant cependant limité, il est probable que de nombreuses publications n'aient pas été prises en compte dans notre revue de la littérature. Notre impression est dès lors subjective.

10.1 Influence de masse

Bien que les projets de numérisation de masse se construisent autour de l'objectif de rendre accessible le savoir en ligne, et que les complexités liées à la construction des infrastructures et les questions financières, politiques, éthiques et légales qui en découlent tendent à accaparer toute l'attention, une autre forme de connaissance est acquise de manière plus discrète. Elle est sans doute motivée par des intérêts divergents en fonction des acteurs de chaque initiative et justifiée par le souci de proposer une meilleure expérience aux utilisateurs ou obtenir de nouvelles formes de financement. Nos actes et pensées d'usagers sont néanmoins précieusement récoltés, nous livrant à la merci de nouvelles logiques économiques et régimes politiques : « [...]as we collect and connect, we are also ourselves collected and connected.² ». Si les défenseurs de l'innovation argumentent qu'il ne sert à rien de s'effrayer de l'impact de ces collectes de données, qui auraient moins d'effet que les spots publicitaires traditionnels accompagnant les campagnes politiques : « We freak out over the unsavory influence of social media on our politics, while TV's partisan influence on elections is far, far greater than Facebook's.³ ». Il semble important de nous rappeler que nous avons un rôle à jouer dans la définition éthique des projets de cette envergure.

Les projets de numérisation de masse ont de plus recours aux technologies dites persuasives pour développer l'interface de leurs plateformes⁴, ces technologies basées sur les principes du « moindre effort » contribuent à la mise en place de mécanismes de pouvoir invisibles, susceptibles d'influencer notre liberté d'opinion et la démocratie⁵.

Le pouvoir de ces entreprises étant souvent le résultat d'un petit nombre de personnes qui décident pour toutes les autres⁶, ce sont néanmoins les utilisateurs finaux qui permettent à ce marché d'exister. Dès lors une prise de conscience de ces enjeux permettrait de rétablir un équilibre entre tous ces intérêts divergents, et contribuerait à garantir que l'emploi de ces technologies se fasse bien au service de la démocratie.

2. N. B. Thylstrup, *The politics of mass digitization...*, p.138.

3. K. Kelly, “AR Will Spark the Next Big Tech Platform—Call It Mirrorworld” ...

4. Les technologies persuasives sont conçues pour modifier les pratiques et comportements des usagers à travers la persuasion et l'influence sociale, sans utiliser la force. Elles s'incarnent souvent dans le graphisme des interfaces et sont motivées par la recherche du « moindre effort ». Par exemple un usager ayant appris à effectuer des recherches sur Google, sera d'autant plus enclin à utiliser la même méthode pour accéder aux ouvrages de *Google Books*. M. Thelle et N. Bonde Thylstrup, “Persuasive territories in European cultural politics...”

5. *Ibid.*

6. N. B. Thylstrup, *The politics of mass digitization...*

10.2 Vers la création d'un nouvel acteur de l'information

Nous nous sommes interrogées, en introduction de ce mémoire, sur la place occupée par ces projets au sein de la famille des institutions culturelles ayant pour mission de rendre la connaissance accessible à tous (GLAM), puisque les initiatives de numérisation de masse sont d'abord et premièrement motivées par une mission similaire : celle de restituer au monde la connaissance, en s'assurant de son accessibilité dans notre société numérique, à travers la numérisation.

L'étude des différents enjeux de la numérisation et l'analyse d'autres initiatives témoignent que le succès de ces projets n'est pas seulement corrélé aux seules capacités des institutions culturelles et patrimoniales, mais implique un mélange d'acteurs issus du monde privé et du monde public répondant à des communautés de pratiques et à des politiques diverses, s'organisant au sein d'un réseau. L'infrastructure commune aux membres de ce réseau doit être par essence souple et flexible, capable de s'adapter à la diversité de ses partenaires, tout en définissant un certain nombre de standards et de routines permettant une convergence des pratiques. La mise en place de ces standards aboutit à des formes de politiques internes au réseau et contribuent à renforcer son pouvoir vis-à-vis des politiques extérieures. Le pouvoir du réseau se construisant souvent en même temps que le réseau lui-même : « *through standardization, inventions become commonplace, novelties become mundane, and the local becomes universal.* ⁷ ».

Le déploiement d'activités de numérisation de masse offre la perspective de création de nouvelles sources de revenus, attirant industries et entreprises privées au sein de leurs réseaux. Les entreprises culturelles étant depuis toujours empreintes de capitalisme, la numérisation de masse a de facto introduit une nouvelle forme de capitalisme imposée à notre mémoire culturelle : le capitalisme numérique⁸.

Ces initiatives semblent, de plus, contribuer à inscrire les institutions patrimoniales et leurs usagers au sein d'une nouvelle constellation de pouvoir et de politiques : « [...] *these infrapolitical imaginaries in fact show the complexity of mass digitization projects in their reinscription of users and cultural memory institutions in new constellations of power and politics.* ⁹ ».

Plus que de simples plateformes de valorisation visant à rendre les données du passé accessibles, le déploiement de l'infrastructure des projets de numérisation de masse soulève de grandes questions concernant l'éthique, les politiques, le pouvoir et la protection des individus au sein de la sphère numérique. Issues de différents courants politiques, ces initiatives viennent bouleverser l'organisation existante du pouvoir.

7. *Ibid.*, p.31.

8. *Ibid.*

9. *Ibid.*, p.131.

[Traduction] La luxuriante dimension des initiatives de numérisation de masse semble émerger des décombres de processus politiques perturbateurs et tumultueux, qui viennent violemment disloquer les frontières établies et les dynamiques de pouvoir, donnant naissance à de nouvelles formes qui doivent encore être interprétées.¹⁰

Difficile en conséquent de considérer ces initiatives comme une simple évolution des activités conduites par les institutions culturelles et patrimoniales. Si ces projets tendent à bouleverser les règles établies et sont suffisamment puissants pour impacter les politiques et proposer une nouvelle forme d'économie, ils se doivent probablement d'être considérés en tant que nouvelle entité (ou acteur de l'information), dont les particularités et caractéristiques doivent encore être pleinement étudiées et comprises, afin de ne pas laisser ce pouvoir émergent se construire en-dehors des règles fondatrices de notre société.

10. « Rather, the productive dimensions of mass digitization emerge from the rubble of disruptive and turbulent political processes that violently dislocate established frontiers and power dynamics and give rise to new ones that are yet to be interpreted. *Ibid.*, p.137 »

Conclusion

Étudier la numérisation de masse, tout en participant à la construction d'une telle initiative a permis de nourrir un intéressant dialogue entre problématiques du terrain et théorie. Les enjeux de la numérisation de masse se dessinant au fil de notre état de l'art et de la création de la feuille de route, l'apport pratique a contribué à élargir le champ initial de la recherche, et la recherche a pu influencer la réalité des propositions. La réalisation de notre mission de stagiaire au sein du projet Time Machine, notre collaboration à l'élaboration des propositions de la feuille de route et la rédaction du présent mémoire, nous ont demandé un important travail de réflexion sur trois axes : la définition des projets de numérisation (contexte, histoire, caractéristiques et enjeux), l'analyse de précédentes initiatives (motivations et réponses apportées aux enjeux identifiés), l'étude du projet Time Machine (réponses apportées aux enjeux, innovations, risques et opportunités, potentiel impact). Nous proposons d'esquisser un bilan des propositions contenues dans ces trois parties de notre mémoire ainsi que des enseignements tirés de notre expérience de stagiaire.

La première partie, entièrement théorique, nous a permis d'abord de retracer l'origine des premières initiatives de numérisation, mêlant développements technologiques et élargissement des activités des institutions culturelles et patrimoniales. Les développements numériques tendant à s'accélérer, les années 1990 voient apparaître les précurseurs des projets de numérisation de masse et les années 2000 marquent la naissance des premières recommandations et directives européennes, appelées à favoriser la croissance de l'économie numérique et à contribuer à l'orientation de ces initiatives d'envergure, la photogrammétrie figurant désormais en bonne place dans les objectifs de numérisation européens. Time Machine s'inscrit ainsi dans la continuité des premiers projets de bibliothèque universelle qui se voyaient également limités par des développements technologiques, dont la taille nécessitait un besoin de standardisation et ne manquaient pas de voir des décisions politiques ou historiques influencer la création de leurs collections. L'histoire nous montre que l'avènement du numérique a très vite impliqué une grande mixité parmi les acteurs des projets, qui se sont confrontés aux questions du droit d'auteur et ont œuvré au déploiement de solutions politiques.

Dans un deuxième temps, nous nous sommes attelée au *comment* des initiatives de numérisation de masse, à la définition de leurs caractéristiques et aux enjeux reliés. Par nature complexes, ces initiatives ne se laissent pas facilement résumer et semblent gagner en difficultés à mesure que l'on cherche à les comprendre. Nous nous sommes basée sur notre analyse historique et expérience de stagiaire afin de proposer une sélection de cinq enjeux à étudier plus précisément (la collaboration, le financement public-privé, le droit d'auteur, l'interopérabilité, le stockage sur le long-terme).

Au vu de la taille des projets, appelés à articuler intérêts divers et pratiques multiples, la collaboration apparaît comme un facteur clé de réussite. Une numérisation massive, implique des coûts conséquents dont les institutions publiques ne peuvent, seules,

prendre la responsabilité. Des partenariats public-privé semblent inévitables, mais rendent plus difficiles l'inscription de ces projets dans une démarche d'ouverture et de transparence. Les directives actuelles du droit d'auteur ne facilitent pas la circulation des données numériques et préviennent la représentation de celles issues de notre passé le plus récent, Time Machine devra oser la prise de risque s'il ne veut pas être limité par ces barrières. Sortir des « silos » peut se faire à condition de veiller à une interopérabilité technique et des politiques documentaires à la hauteur des ambitions des projets de numérisation de masse, en arrière-fonds la collaboration semble encore détenir les clés du succès. Appréhender la préservation des données pour en favoriser les usages est essentiel à tout projet de numérisation de masse. Time Machine aura pour tâche de comprendre les nuances propres à chaque enjeu et d'intégrer la matérialité des solutions au sein de son infrastructure. La création de cette infrastructure ne se fera pas sans la mise en place d'un cadre opérationnel à même de réconcilier les pratiques du monde de la documentation, du web et de l'industrie, afin de planifier et exécuter la transformation des données du passé en données numériques.

Notre travail nous permet de conclure que si le projet veut réussir à établir de nouvelles normes et apporter des réponses satisfaisantes à ces différents enjeux, il lui faudra non seulement disposer de grands moyens financiers, mais être capable d'instaurer un dialogue privilégié avec chaque membre de son réseau, tout en composant avec le cadre externe du projet (légal et politique) qui, sans avoir apporté de réponses complètes aux problématiques du droit d'auteur, nourrit de hautes ambitions pour la numérisation du patrimoine européen. Alors que le financement du projet est des plus incertain, Time Machine doit réussir à se positionner comme réseau d'influence, à même d'accroître le soutien politique dont il bénéficie et espérer ainsi briser certaines barrières territoriales et légales qui rendent compliquée la réalisation du big data du passé.

La deuxième partie est consacrée à l'étude des motivations d'autres projets de numérisation de masse (*Google Books*, Europeana, *HathiTrust* et *the Digital Public Library of America*) et aux réponses apportées par ces initiatives aux enjeux de la numérisation. Cette comparaison nous a permis de déduire que les frontières séparant ces projets sont poreuses, chacun se basant sur les réalisations des autres pour construire de nouveaux développements. Les initiatives sont conscientes de cette complémentarité et travaillent ensemble à élaborer de nouvelles solutions aux différents enjeux. Nous concluons cette recherche par le constat que plutôt que de chercher à catégoriser ces initiatives, il semble plus intéressant de les considérer comme un tout, afin de pouvoir déplacer le spectre de la réflexion au niveau de leur ensemble et ne pas seulement s'arrêter sur les détails qui les composent. Time Machine devra lui aussi développer des ponts avec ces grands réseaux, et trouver un équilibre entre le phénomène de mondialisation culturelle qui les accompagne et la préservation des particularités locales et régionales de ses partenaires.

Enfin notre troisième partie, plus concrète, est d'abord pour nous l'occasion de

présenter les propositions contenues dans la feuille de route en analysant les réponses apportées aux différents enjeux par Time Machine. Puisque l'agenda, entre les recherches conduites pour le mémoire et les échéances de rendu européennes, n'a pas toujours coïncidé, nous proposons un certain nombre d'amendements. Ils visent à accroître la collaboration avec les agrégateurs que sont les Time Machines locales, à mieux prendre en compte les usages du public dans la gestion des droits d'auteur, à favoriser la préservation sur le long-terme par des directives techniques et à compléter les outils destinés à garantir l'interopérabilité afin de sortir de la logique des « silos ». Nous argumentons ensuite que ce qui distingue Time Machine des autres entreprises de numérisation de masse, est l'intégration de toutes les parties publiques ou privées ou sein d'un même réseau (le projet ne faisant pas la distinction entre collaborateurs internes et futurs exploitants), et le déploiement d'innovations technologiques, pour certaines basées sur l'intelligence artificielle, au service de la création du big data du passé. Fort de ces propositions novatrices, Time Machine accroît sa popularité auprès d'acteurs attirés par les nouvelles perspectives offertes par la démarche inclusive et la crédibilité technologique du projet.

Notre travail d'analyse historique et des enjeux, nous a permis de définir un certain nombre d'éléments constituants des risques ou opportunités. Time Machine devra, au-delà du déploiement d'infrastructures et routines cohérents, proposer une plateforme en accord avec les besoins et attentes des différentes communautés composant son public-cible en évitant les méthodologies limitées au monde occidental, pour ne pas biaiser le développement de ses collections et cloisonner les usages futurs. Favoriser la cocréation devrait permettre à Time Machine de proposer des solutions fédératrices au-delà des frontières territoriales et académiques. Résolument tourné vers le futur, le projet se devra également de refléter les nouvelles pratiques induites par le numérique et proposer au grand public un rôle actif et créateur de valeur, en promouvant les sciences citoyennes ou *citizen science*. Time Machine a l'opportunité de pouvoir donner à l'Europe une plateforme aux contenus égaux et éthiques, ne cherchant pas à refléter l'histoire telle que déjà écrite, mais à en offrir une nouvelle reproduction.

Nous concluons cette partie sur le constat que si de nombreuses études se consacrent aux enjeux techniques et organisationnels liés aux projets de numérisation de masse, trop peu de chercheurs s'intéressent aux questions du *pourquoi*. Pourtant la portée de ces projets dépasse les frontières du monde culturel et patrimonial et ces réseaux détiennent une forme de pouvoir susceptible d'avoir des répercussions sur notre société.

Réseau définitivement tourné vers les innovations numériques, Time Machine semble déjà incarner les attentes d'une classe dirigeante européenne inquiète de bénéficier des retombées économiques offertes par ce nouveau monde. S'inscrivant dans le suivi des politiques de mondialisation, et s'adressant à une audience plus large que celle réservée aux institutions culturelles et patrimoniales, le projet ouvre le dialogue au sein des plus hautes sphères du pouvoir européen et suscite avant même le déploiement de son infrastructure,

l'intérêt des foules. Ce nouvel acteur des initiatives de numérisation de masse semble déjà détenir une forme de pouvoir. Seules des recherches plus avancées permettront de définir avec assurance si ce dernier est limité aux partenaires de son réseau, où s'étend au-delà et touche la sphère politique. Nous pensons que pour permettre l'étude de ces nouveaux phénomènes et être en adéquation avec leur complexité, il est plus que temps de considérer les acteurs de la numérisations de masse comme de nouvelles entités informationnelles, dont les contours évoluent certes, mais entraînent des répercussions en dehors du milieu des institutions culturelles et patrimoniales, qui viennent ébranler un certain ordre établi.

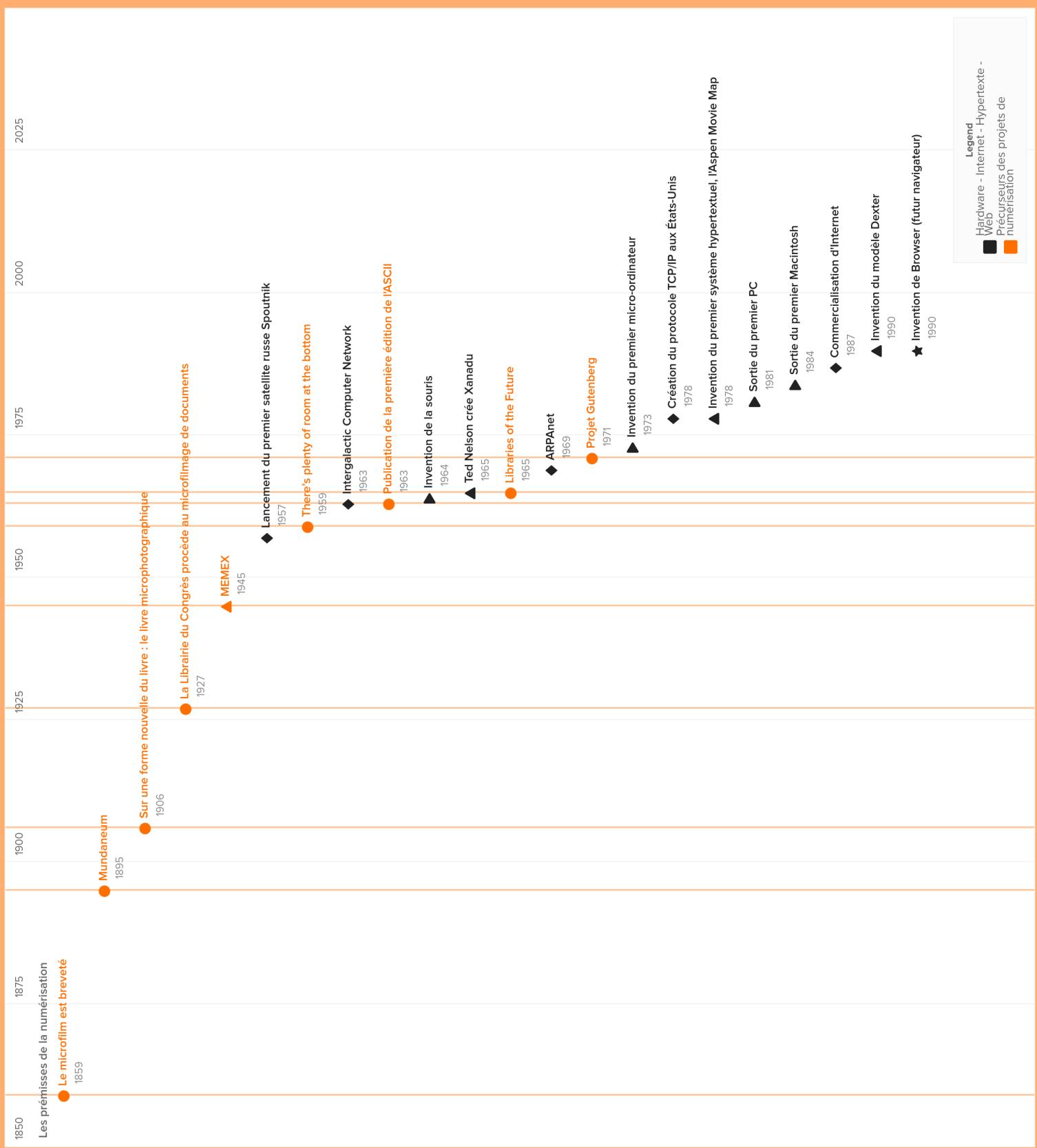
A l'instar des précédentes initiatives de numérisation de masse, le projet Time Machine semble apporter autant de nouvelles solutions que de questionnements. Si le projet aboutit et réussit une intégration équilibrée de tous ces éléments, il deviendra sans doute un acteur informationnel, politique, économique et culturel de poids. Espérons alors, que la prophétie du *mirrorworld* se trompe lorsqu'elle décrit les futurs services développés comme appelés à devenir des biens payants au même titre que l'eau et l'électricité¹¹, et que Time Machine demeurera un outil au service de l'égalité et de la démocratie, libre et gratuit.

11. K. Kelly, "AR Will Spark the Next Big Tech Platform—Call It Mirrorworld"...

Annexes

Annexe A

Prémisses de la numérisation



Les prémisses de la numérisation

- **Le microfilm est breveté**
1859

L'invention du microfilm est brevétée par le français René Dagron.
- **Mundaneum**
1895

Conjointement avec Henri La Fontaine, Paul Otlet initie le projet du Mundaneum, centre de documentation universel.
- **Sur une forme nouvelle du livre : le livre micropographique**
1906

Paul Otlet publie son livre et suggère que les plus importantes transformations ne prendront pas place dans le livre lui-même mais dans son substitut.
- **La Librairie du Congrès procède au microfilmage de documents**
1927

Près de trois millions de livres et manuscrits issus des collections de la British Library sont ainsi numérisés.
- ▲ **MEMEX**
1945

Vannevar Bush invente les bases de l'hypertexte, non pas comme un système pour relier des informations entre elles à travers des interfaces d'ordinateurs mais comme un système documentaire en soi. Son système s'inspire de la représentation du cerveau humain et place l'utilisateur au centre. Cette méthode servira à poser des bases au concept de lien et d'ancre utilisé sur internet.
- ◆ **Lancement du premier satellite russe Spoutnik**
1957

Les américains qui ne veulent pas rester sur le banc de touche se lancent à leur tour dans la conquête de l'espace. L'ARPA (Advanced Research Projects Agency) est créée une année plus tard.
- **There's plenty of room at the bottom**
1959

Lors d'une conférence, l'américain Richard Feynman propose de comprimer l'intégralité de l'encyclopédie britannique afin de la réduire à la taille d'une tête d'épingle.
- ◆ **Intergalactic Computer Network**
1963

Le premier réseau de machines voit le jour.
- **Publication de la première édition de l'ASCII**
1963

L'American Standard Code for Information Interchange (ASCII) est un standard d'encodage des caractères permettant de rendre compréhensible l'écriture latine en la transformant en une suite binaire (0 et 1), lisible par les systèmes des ordinateurs. La première version est limitée aux caractères utilisés dans la langue anglaise.
- **Invention de la souris**
1964

Douglas Engelbart, membre de Xerox entreprise au sein de la Silicone Valley invente la souris.
- ▲ **Ted Nelson crée Xanadu**
1965

Ce prototype se base sur les travaux de Vannevar Bush mais distingue deux points de vue, celui de l'utilisateur et celui du producteur. En effet si un lien est une association entre deux documents, celui qui crée l'information connaît le système des liens qu'il crée par opposition à celui qui découvre cette navigation. C'est à lui que l'on doit la création du terme hypertexte.
- **Libraries of the Future**
1965

Dans son ouvrage, J.C.R Licklider propose d'étendre le monde des bibliothèques à celui des ordinateurs.
- ◆ **ARPAnet**
1969

ARPAnet est le premier réseau de machines réellement implémenté, installé pour relier quatre universités de Californie.
- **Projet Gutenberg**
1971

Initié par Michel Hart, le projet incarne la première version de ce que deviendront les bibliothèques numériques. Une équipe de volontaires saisit au clavier des textes libres de droit dans un format accessible et compréhensible par les ordinateurs : une version basique de l'ASCII. Les textes analogues sont ainsi convertis sous une forme numérique et sont ensuite diffusés auprès des membres du réseau ARPAnet.
- **Invention du premier micro-ordinateur**
1973

Inventé par les français Micral, ce premier micro-ordinateur fonctionne avec des boutons.
- ◆ **Création du protocole TCP/IP aux États-Unis**
1978

C'est le protocole servant à l'échange entre les machines. C'est également à ce moment-là que le nom d'internet est utilisé à la place de celui d'Arpanet.
- ▲ **Invention du premier système hypertextuel, l'Aspen Movie Map**
1978

Inventé par le MIT, l'Aspen Movie Map se constitue de documents et de noeuds, c'est-à-dire de documents reliés entre eux par des liens.

▶ Sortie du premier PC	Créé par IBM, son coût est relativement faible.
1981	
▶ Sortie du premier Macintosh	Apple dévoile son premier Macintosh.
1984	
◆ Commercialisation d'Internet	Internet commence à se commercialiser. A noter qu'en France le réseau Minitel développé par France Telecom va retarder l'entrée du pays dans l'ère d'Internet.
1987	
▲ Invention du modèle Dexter	Ce modèle d'hypertexte définit les standards que l'on retrouve encore dans les modèles actuels. Il normalise la description de l'http.
1990	
★ Invention de Browser (futur navigateur)	Au CERN en Suisse, Tim Berners-Lee invente un logiciel pour retrouver ses documents, qu'il nomme Browsers (futur navigateur). Le CERN va contribuer au développement de cette idée et créer un intranet pour son usage interne. En 1993, Tim Berners-Lee crée le W3C (World Wide Web Consortium) afin de s'assurer de la bonne évolution du web. C'est à lui que l'on doit les différentes versions d'HTML (Hypertext Markup Language).
1990	

Annexe B

Factsheet

Creating a common future for Europe, based on our shared past



Why Europe should invest to preserve its cohesion and identity

- ◆ The cohesiveness of European cultural identity is being threatened by the resurgence of unresolved conflicts deep-seated in European memory.
- ◆ Democratic dialogue is endangered by the dominion of private platforms over historical and cultural data.
- ◆ Managed by proprietary algorithms, such platforms may prioritise popularity and personal agendas, opening the way to fake news.



Time Machine allows Europe to restore its engagement with its past and use it as a vital resource for a common future

- ◆ Time Machine (TM) is a large-scale research initiative aiming to develop the big data of the past, a huge distributed digital information system mapping the European social, cultural and geographical evolution across times.
- ◆ By designing and implementing advanced new digitisation and Artificial Intelligence (AI) technologies to mine Europe's vast cultural heritage, Time Machine will provide fair and free access to information that will support future scientific and technological developments.
- ◆ Open platforms for navigating the multicultural and multilingual perspectives of our common past will turn our long history into a pan-European cultural, economic and social asset.

Concrete outcomes and expected impacts for Society And Economy



Creating new disruptive business models in key economic sectors

- ◆ Time Machine will act as an economic motor for new professions, services and products, impacting key sectors of European economy (ICT, creative industries and tourism, the development of Smart Cities and land use).
- ◆ The European creative industries contribute 6.8% of GDP and 6.5% of employment in the EU. Europe is the most visited tourism region in the world, and in the EU, tourism contributes 10% to EU GDP and creates jobs for 26 million people. Cultural Heritage is a unique asset for European businesses.
- ◆ Time Machine develops a Franchise model for cities that wish to make a creative use of their historical past.



A transformational impact on Social Sciences and Humanities (SSH)

- ◆ Identifying larger patterns, correlations and connections will open new frontiers in our capacities for in-depth analysis and informed decision making.
- ◆ Sharp increase in the demand for digital and traditional humanists and social scientists at a time where these disciplines and corresponding university degrees do not guarantee jobs in these fields.



Making education more accessible, interactive and diversified

- ◆ Time Machine will offer more depth to educational curricula, sharpening the critical thinking of learners, and contributing to informed decision-making at all levels.
- ◆ The resulting online courses, materials, simulations and other experiences will promote active engagement with our combined cultural heritage and make continuous learning more accessible and inclusive.
- ◆ Time Machine will create a dynamic new industry for the production of educative digital material based on aligned massive cultural datasets.



A strong boost in EU competitiveness in AI and ICT

- ◆ An AI trained on Big Data of the Past will offer a strong competitive advantage for Europeans in the global AI race.
- ◆ TM will also introduce disruptive technologies in machine vision, linguistic and knowledge systems, multimodal (4D) simulation, HPC and long-term data storage, strengthening the competitive position of EU industry in these fields.



Building a cornerstone for international European excellence

- ◆ Time Machine comes at a time where culture occupies a central role in the UN 2030 Agenda for Sustainable Development.
- ◆ Europe has a leading role in the digitisation of culture and Artificial Intelligence for Cultural Heritage. TM will strengthen this role at a time where this field gains momentum in Asia and the USA.



Key facts

Timeline

2019

Europe invests in Time Machine

The European Commission chose Time Machine as one of the six proposals retained for preparing large-scale research initiatives.

2018

Time Machine develops algorithms that outperform humans in transcription of Venetian handwriting

AI methods open new way to search in ancient documents.

2016

Manifesto "L'Europe doit construire la première Time Machine" published in Le Temps, then translated in 9 languages

A call for action to invite Europe to invest in an infrastructure for mining "Big Data of the Past".

2014

Increasing interest

Frederic Kaplan's TED Talk "How to build a Time Machine" reaches more than 1 Million views.

2013

Venice Time Machine starts

EPFL and University Ca'Foscari launch a project that aims at building a multidimensional model of Venice and its evolution covering a period of more than 1000 years.

Consortium

- ◆ 225 institutions from 32 countries
- ◆ 33 founding institutions
- ◆ 7 national libraries
Austria, Belgium, France, Israel, Netherlands, Spain, Switzerland
- ◆ 19 state archives
Belgium, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, Germany, Hungary, Lithuania, Malta, Norway, Poland, Romania, Slovenia, Spain, Slovakia, Sweden, and Switzerland
- ◆ Key European Museums
Louvre, Rijksmuseum, Belvedere
- ◆ 95 academic and research institutions
- ◆ 30 European companies
- ◆ 18 governmental bodies

Current local Time Machines

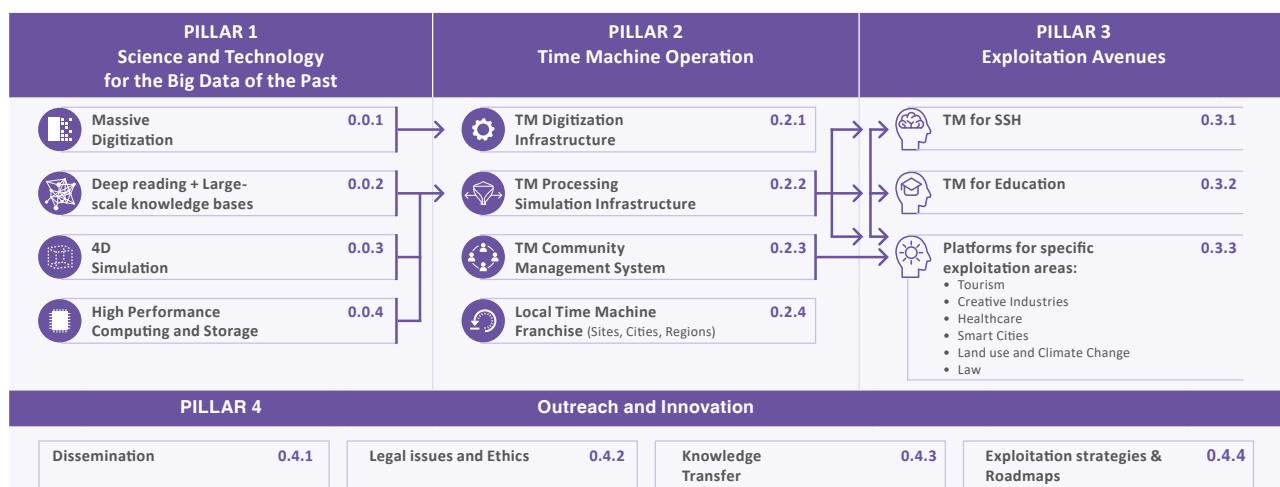
- ◆ Antwerp TM (1500-2000)
- ◆ Amsterdam TM (1550-2000)
- ◆ Budapest TM (1680-1990)
- ◆ Dresden TM (1200-2000)
- ◆ Ghent-Bruges TM (800-2000)
- ◆ Jerusalem TM (2000 BCE-2000)
- ◆ Limburg TM (1775-2000)
- ◆ Lower Austrian TM (800-2000)
- ◆ Naples TM (800-2000)
- ◆ Nuremberg TM (1000-2000)
- ◆ Paris TM (1000-2000)
- ◆ Regensburg TM (1200-2000)
- ◆ Utrecht TM (40-2000)
- ◆ Venice TM (1000-2000)

Further expanding an already vast network of partners

The Time Machine community is currently expanding to create a dense Time Machine ecosystem of leading scientists, innovators and other key players of the civil society, having as target to reach the number of 2000 supporting organisations in the beginning of 2020.



Time Machine is an integrated programme with clearly defined pillars and thematic areas



Glossaire

agrégateur Dans le cadre de ce mémoire, ce terme désigne une organisation offrant un service de collecte, validation, harmonisation, stockage et souvent enrichissement des données issues de projets de numérisation des collections d'institutions culturelles et patrimoniales.... 23, 47, 72, 77, 87, 89, 95, 99, 101, 116, 131, 141

apprentissage automatique ou machine learning Champ d'étude de l'intelligence artificielle qui se fonde sur des approches statistiques pour donner aux ordinateurs la capacité « d'apprendre » à partir de jeux de données. Cet apprentissage peut se faire de manière supervisée (le modèle mathématique utilise des informations sur les jeux de données entrant et sur les résultats attendus), semi-supervisée (certains jeux de données entrant ne sont pas explicités dans le modèle mathématique et l'ordinateur doit déduire leurs fonctions), ou non supervisée (le modèle mathématique contient uniquement des informations sur les données entrantes et aucune sur les transformations attendues en résultat)¹.... 61, 66

apprentissage profond ou deep learning L'apprentissage profond fait partie de la famille des méthodes d'apprentissage automatique ou machine learning. Basé sur les réseaux de neurones, cette méthode utilise un système de couches pour extraire progressivement des informations².... 68, 69, 155

bases de données Une base de données permet de stocker des données brutes ou informations en rapport avec un thème ou une activité. Ces dernières permettent ensuite de rechercher les informations ainsi stockées. Il existe différentes natures de base de données³.... 40, 95

big data Grands jeux de données, dont l'analyse par traitements informatiques permet de mettre en valeur d'insoupçonnés motifs et associations, contribuant à enrichir

1. *Apprentissage automatique*, fr, Page Version ID : 160226362, juin 2019, URL : https://fr.wikipedia.org/w/index.php?title=Apprentissage_automatique&oldid=160226362 (visité le 23/06/2019)

2. *Deep learning*, en, Page Version ID : 902979782, juin 2019, URL : https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=902979782 (visité le 23/06/2019)

3. *Base de données*, fr, Page Version ID : 160423450, juin 2019, URL : https://fr.wikipedia.org/w/index.php?title=Base_de_donn%C3%A9es&oldid=160423450 (visité le 18/07/2019)

notre connaissance sur le genre humain⁴.... 21, 60, 61, 65, 68, 69, 71, 72, 122, 133, 140, 141

Creative Commons Association à but non lucratif dont l'objectif est d'offrir une solution alternative légale (par le biais de plusieurs licences) aux personnes souhaitant libérer leurs œuvres des droits de propriété intellectuelle standards dans leurs pays, lorsque jugés trop restrictifs⁵.... 89, 116, 117

curation Néologisme servant à désigner la pratique visant à sélectionner, éditer et partager des ressources pertinentes du web en réponse à une requête donnée.⁶.... 51, 125

graphe Dans le contexte de ce mémoire, ce terme est utilisé pour désigner une grande base de connaissance, compilant les informations de plusieurs sources différentes⁷.... 65, 66, 69, 71, 72, 111, 116, 117, 122

intelligence artificielle Ensemble des théories et techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence⁸.... 65, 69, 78, 121, 122, 155, 156

libre accès ou Open Access Mise à disposition de manière pérenne, libre, gratuite et en ligne, des travaux de recherches financés par les pouvoirs publics⁹.... 20, 21, 40, 88

numérisation Dans le contexte du projet Time Machine, ce terme indique non seulement la conversion des informations d'un support (texte, image, audio, vidéo, artefact) ou signal électrique en données numériques¹⁰, mais également la transformation du patrimoine bâti, élément géographique, en modèles 3D.... 15, 40, 42, 58, 60, 61

Open Science Mouvement visant à rendre les données de la recherche au sens large (jeux de données, notes de laboratoire, processus, images etc.), accessibles à tous¹¹.... 40,

4. *Big data*, fr, Page Version ID : 159931315, juin 2019, URL : https://fr.wikipedia.org/w/index.php?title=Big_data&oldid=159931315 (visité le 23/06/2019)

5. *Creative Commons*, en, Page Version ID : 906058802, juil. 2019, URL : https://en.wikipedia.org/w/index.php?title=Creative_Commons&oldid=906058802 (visité le 24/07/2019)

6. *Curation de contenu*, fr, Page Version ID : 160878957, juil. 2019, URL : https://fr.wikipedia.org/w/index.php?title=Curation_de_contenu&oldid=160878957 (visité le 30/07/2019)

7. *Knowledge Graph*, fr, Page Version ID : 138894172, juil. 2017, URL : https://fr.wikipedia.org/w/index.php?title=Knowledge_Graph&oldid=138894172 (visité le 23/06/2019)

8. *Intelligence artificielle*, fr, Page Version ID : 160343358, juin 2019, URL : https://fr.wikipedia.org/w/index.php?title=Intelligence_artificielle&oldid=160343358 (visité le 23/06/2019)

9. *Libre accès (édition scientifique)*, fr, Page Version ID : 160593388, juil. 2019, URL : [https://fr.wikipedia.org/w/index.php?title=Libre_acc%C3%A8s_\(%C3%A9dition_scientifique\)&oldid=160593388](https://fr.wikipedia.org/w/index.php?title=Libre_acc%C3%A8s_(%C3%A9dition_scientifique)&oldid=160593388) (visité le 09/07/2019)

10. *Numérisation*, URL : <https://fr.wikipedia.org/wiki/Num%C3%A9risation> (visité le 23/06/2019)

11. *Open science*, en, Page Version ID : 900178688, juin 2019, URL : https://en.wikipedia.org/w/index.php?title=Open_science&oldid=900178688 (visité le 23/06/2019)

41, 51, 65, 116

photogrammétrie La photogrammétrie peut se résumer en une technique de reconstruction numérique en 3D d'un objet physique, permettant de déterminer les dimensions et les volumes des objets à partir de mesures effectuées sur des photographies montrant les perspectives de ces objets¹².... 66, 67, 123, 129

POLY-perspective Concept indiquant que les futurs chercheurs et scientifiques devraient adopter une approche plurielle pour mener à bien leurs activités, et ne pas hésiter à collaborer avec des spécialistes d'autres domaines pour proposer des solutions innovantes¹³.... 59

reconnaissance d'entités nommées La reconnaissance d'entités nommées est une sous-tâche de l'activité d'extraction d'informations dans des corpus documentaires, consistant à rechercher des objets textuels (mots ou groupes de mots) catégorisables dans des classes (noms de personnes, de lieux, quantités, distances, valeurs, dates etc.)¹⁴.... 67

reconnaissance optique de caractères La reconnaissance optique de caractères, ou océrisation, désigne les procédés informatiques pour la traduction d'images, de textes imprimés ou dactylographiés en fichiers de texte¹⁵.... 13, 33, 67

réseaux de neurones artificiels Système imitant le fonctionnement des neurones biologiques, fait partie des technologies d'apprentissage profond ou deep learning, utilisé également par l'intelligence artificielle. Ce réseau implique le traitement d'une information en couches successives, permettant d'affiner les résultats proposés par la dernière couche¹⁶.... 62

sciences citoyennes ou *citizen science* Désigne la recherche scientifique conduite entièrement ou en partie par des scientifiques amateurs ou non-professionnels. Ce mouvement est également décrit comme la participation du public dans la recherche scientifique¹⁷.... 90, 116, 129, 130, 141

12. *Photogrammétrie*, fr, Page Version ID : 159009276, mai 2019, URL : <https://fr.wikipedia.org/w/index.php?title=Photogramm%C3%A9trie&oldid=159009276> (visité le 23/06/2019)

13. *The CDH's vision : POLY-perspective – EPFL*, en-GB, URL : <https://www.epfl.ch/schools/cdh/cdhs-vision/> (visité le 23/06/2019)

14. *Reconnaissance d'entités nommées*, fr, Page Version ID : 144628341, janv. 2018, URL : https://fr.wikipedia.org/w/index.php?title=Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es&oldid=144628341 (visité le 23/06/2019)

15. *Reconnaissance optique de caractères*, fr, Page Version ID : 139202141, juil. 2017, URL : https://fr.wikipedia.org/w/index.php?title=Reconnaissance_optique_de_caract%C3%A8res&oldid=139202141 (visité le 23/06/2019)

16. *Artificial neural network*, en, Page Version ID : 901207463, juin 2019, URL : https://en.wikipedia.org/w/index.php?title=Artificial_neural_network&oldid=901207463 (visité le 23/06/2019)

17. *Citizen science*, en, Page Version ID : 906579059, juil. 2019, URL : https://en.wikipedia.org/w/index.php?title=Citizen_science&oldid=906579059 (visité le 24/07/2019)

traitement automatique du langage naturel Domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle visant à créer des outils de traitement de la langue naturelle pour diverses applications¹⁸.... 67

18. *Traitement automatique du langage naturel*, fr, Page Version ID : 159047154, mai 2019, URL : https://fr.wikipedia.org/w/index.php?title=Traitement_automatique_du_langage_naturel&oldid=159047154 (visité le 23/06/2019)

Table des figures

1.1	Capture-d'écran du site internet du <i>Mundaneum</i>	11
1.2	Capture-d'écran du site internet du <i>Projet Gutenberg</i>	13
1.3	Frise temporelle, <i>Les prémisses de la numérisation</i>	14
1.4	Capture-d'écran du visualisateur <i>Scaife</i> de la <i>Perseus Digital Library</i>	17
1.5	Capture-d'écran de la plateforme du projet <i>Gallica</i>	19
1.6	Capture-d'écran de la plateforme du projet <i>Universal Digital Library</i>	20
1.7	Capture-d'écran de la plateforme du projet de Bibliothèque numérique mondiale.	22
1.8	Frise temporelle, <i>Politiques et numérisation</i>	28
2.1	Image de mauvaise qualité, numérisée par Google © <i>Google Books</i>	43
2.2	Extrait de <i>Europeana Data Model Primer</i> ©Europeana.	46
3.1	Le DHLAB au sein de l'EPFL	60
3.2	Le scanner Replica à la Fondation Cini, © <i>Copyright 2019 Factum Foundation</i>	62
3.3	Time Machine, logo © <i>Copyright 2019 Time Machine</i>	63
3.4	Time Machine, Information's Mushroom © <i>Copyright 2019 Time Machine</i>	65
3.5	Ville de Sion, compilée par ScanVan © <i>Copyright EPFL DHLAB</i>	67
3.6	Impresso, visuel © <i>Copyright 2019 EPFL, DHLAB</i>	68
3.7	Interface de Diamond, capture d'écran © <i>Copyright 2019 Time Machine</i> .	70
3.8	Résumé visuel des projets et technologies évoqués	70
3.9	Time Machine, visualisation du réseau © <i>Copyright 2019 Time Machine</i> .	73
4.1	Affichage d'un résultat <i>Google Books</i> , capture d'écran	80
4.2	Résumé des différentes activités entreprises par <i>HathiTrust</i> , capture d'écran	83
4.3	Page d'accueil de la plateforme d' <i>HathiTrust</i> , capture d'écran	85
4.4	Page d'accueil de l'interface découverte d' <i>Europeana</i> , capture d'écran . . .	88
4.5	Page d'accueil de l'interface professionnelle d' <i>Europeana</i> , capture d'écran .	91
4.6	Page d'accueil de l'interface découverte du projet de DPLA, capture d'écran	94
4.7	Page d'accueil de l'interface professionnelle du projet de DPLA, capture d'écran	96

6.1	Organisation et interactions des premiers <i>PILLARs</i> © Copyright 2019 <i>Time Machine</i>	107
6.2	Local Time Machines, document de travail © Copyright 2019 <i>Time Machine</i>	110
6.3	Les différents composants de la <i>Time Machine Organisation</i> , document de travail	113
8.1	Représentation schématique des processeurs de contenus et des moteurs d'inférence © Copyright 2019 <i>Time Machine</i>	122

Liste des tableaux

1.1	Leçons d'histoire pour Time Machine	29
2.1	Résumé des enjeux de la numérisation	53
4.1	Les réponses de <i>Google Books</i> aux enjeux de la numérisation	82
4.2	Les réponses de <i>HathiTrust</i> aux enjeux de la numérisation	86
4.3	Les réponses d' <i>Europeana</i> aux enjeux de la numérisation	93
4.4	Les réponses de la <i>Digital Public Library of America</i> aux enjeux de la numérisation	97
7.1	Résumé des réponses de Time Machine aux enjeux de la numérisation . . .	120
9.1	Résumé des risques et opportunités pour Time Machine	132

Table des matières

Résumé	iii
Remerciements	v
Bibliographie	vii
Acronymes	xxxi
Introduction	3
I De la numérisation à la numérisation de masse	7
1 Historique de la numérisation	9
1.1 1800-1990 : Prémisses des projets de numérisation	10
1.2 1990-2000 : Passage à l'échelle	15
1.3 Politiques et numérisation	21
1.4 Leçons d'histoire pour Time Machine	29
2 Comment numériser en masse	31
2.1 Caractéristiques des projets	32
2.2 Enjeux des projets	35
2.2.1 Amener différents acteurs à collaborer	35
2.2.2 Financement et partenariats public-privé	36
2.2.3 Droit d'auteur	38
2.2.4 Sortir des silos - la quête de l'interopérabilité	40
2.2.5 Stockage sur le long-terme - préservation	50
2.2.6 Résumé des enjeux de la numérisation	53
3 Contexte du projet Time Machine	55
3.1 La recherche en humanités numériques	55
3.2 Le Laboratoire d'humanités numériques de l'EPFL	58

3.3	Venice Time Machine	60
3.3.1	Développements technologiques	61
3.4	Le projet Time Machine	63
3.4.1	FET Flagships ou la recherche de financement	63
3.4.2	Time Machine, quels objectifs ?	64
3.4.3	Quelle(s) méthode(s) pour Time Machine ?	66
3.4.4	Collaboration public-privé de partenaires européens	71
3.4.5	Un réseau de Time Machines locales	72
II	Exemples de projets de numérisation de masse	75
4	Différentes typologies de projets	77
4.1	Rassembler une masse de données numérisées	78
4.1.1	Google Books	78
4.1.2	HathiTrust	83
4.2	Agréger pour mieux valoriser	87
4.2.1	Europeana	87
4.2.2	Digital Public Library of America	94
5	Comment catégoriser ?	99
III	Une nouvelle voie à définir pour Time Machine	103
6	Stagiaire au sein du projet Time Machine	105
6.1	État des lieux avant le commencement du stage	105
6.2	Déroulement du stage	108
6.3	Rédaction de la feuille de route	111
7	Quelles réponses aux enjeux de la numérisation ?	115
7.1	Amener différents acteurs à collaborer	115
7.1.1	<i>Complément</i> aux propositions de la feuille de route :	116
7.2	Financement et partenariats public-privé	117
7.3	Droit d'auteur	117
7.3.1	<i>Complément</i> aux propositions de la feuille de route :	118
7.4	Sortir des silos : enjeux techniques	118
7.4.1	<i>Complément</i> aux propositions de la feuille de route :	118
7.5	Sortir des silos : enjeux sur le contenu	119
7.5.1	<i>Complément</i> aux propositions de la feuille de route :	119
7.6	Stockage sur le long-terme - préservation	119

7.7 Résumé des réponses de Time Machine aux enjeux de la numérisation	120
8 Les innovations du projet	121
8.1 Mélanger les acteurs	121
8.2 Innovations technologiques	121
9 Risques et opportunités	125
9.1 Plateforme, usages et accessibilité	125
9.1.1 Franchir les barrières territoriales	126
9.1.2 Être mobile	126
9.1.3 Favoriser la cocréation	126
9.1.4 Moteur de recherche ou plateforme de découverte ?	128
9.2 Crowdsourcing et sciences citoyennes	129
9.3 Pour un savoir cohérent, égalitaire et éthique	130
9.4 Résumé des risques et opportunités pour Time Machine	132
10 Pourquoi numériser en masse ?	133
10.1 Influence de masse	134
10.2 Vers la création d'un nouvel acteur de l'information	135
Conclusion	139
Annexes	145
A Prémisses de la numérisation	145
B Factsheet	149
Glossaire	153
Table des figures	157
Liste des tableaux	159
Table des matières	161