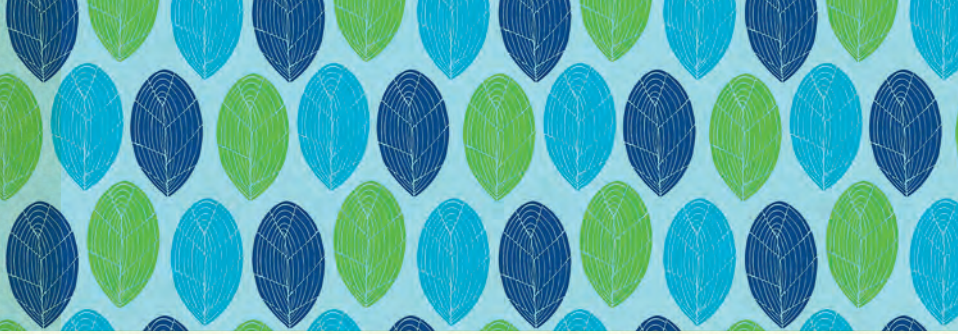

Pratiques de l'édition numérique

Sous la direction de
Michaël E. Sinatra et Marcello Vitali-Rosati

PRATIQUES DE L'ÉDITION NUMÉRIQUE

La collection **Parcours numériques** est accessible gratuitement en édition augmentée sur **parcoursnumeriques-pum.ca**

Pratiques de l'édition numérique

Sous la direction de
Michaël E. Sinatra
et Marcello Vitali-Rosati

Les Presses de l'Université de Montréal

Couverture : © innkey /123rf.com
Mise en pages : Yolande Martel

Catalogage avant publication de Bibliothèque et Archives nationales du Québec et Bibliothèque et Archives Canada

Vedette principale au titre :

Les pratiques de l'édition numérique

(Parcours numériques)
Comprend des références bibliographiques.

ISBN 978-2-7606-3202-8

1. Édition électronique. I. Vitali Rosati, Marcello, 1979- .
II. Eberle-Sinatra, Michaël, 1971- .

Z286.E43P72 2014 070.5'797 C2014-940172-8

Dépôt légal : 1er trimestre 2014
Bibliothèque et Archives nationales du Québec

ISBN (papier) 978-2-7606-3202-8
ISBN (epub) 978-2-7606-3204-2
ISBN (pdf) 978-2-7606-3203-5

Les Presses de l'Université de Montréal reconnaissent l'aide financière du gouvernement du Canada par l'entremise du Fonds du livre du Canada pour leurs activités d'édition et remercient de leur soutien financier le Conseil des arts du Canada et la Société de développement des entreprises culturelles du Québec (SODEC).

IMPRIMÉ AU CANADA

Introduction

Michaël E. Sinatra et Marcello Vitali-Rosati

Ce que nous sommes, en tant qu'êtres humains et en tant que sociétés, est profondément façonné par les formes de production et de circulation du savoir : comprendre ces formes, être capable de les analyser et d'en repérer les enjeux, n'est pas qu'une question de compétences techniques ou disciplinaires, c'est en fait la clé pour avoir une prise sur notre monde. Le numérique a engendré une transformation profonde des modèles de production et de circulation des contenus que nous connaissons depuis le XVIII^e^ siècle. Le web, en particulier, a déterminé un bouleversement majeur du sens même des contenus : nous étions dans une économie de la rareté, nous sommes aujourd'hui dans une surabondance d'informations. Les instances de choix, d'évaluation et de distribution des contenus étaient centralisées dans les mains de certaines institutions privées ou publiques qui en étaient les garants ; aujourd'hui, les systèmes de légitimation semblent absents ou déstructurés. Finalement, un modèle économique assez clair et universel régissait la production et la distribution des connaissances,

et désormais nous ne pouvons que constater la crise de ce modèle et la difficulté d'en proposer un nouveau.

Comprendre les caractéristiques spécifiques du numérique et analyser les pratiques éditoriales à l'époque du web signifient donc se poser la question du rôle du savoir et des connaissances dans notre société. En somme, parler d'édition numérique ne signifie pas seulement débattre de questions techniques ou de choix d'outils : on pourrait affirmer que les pratiques de production et de diffusion du savoir dans leur ensemble sont en jeu. Il est donc important de comprendre les différents aspects de l'édition numérique, au plan tant théorique que pratique, afin de mieux saisir l'importance du numérique dans notre société.

En premier lieu, il faut souligner que, traditionnellement, l'édition est caractérisée par trois instances qui correspondent à ses trois tâches principales : le choix des contenus, la légitimation de ces contenus et leur diffusion. L'éditeur commence par choisir ce qu'il est important de publier ; très souvent, il commande des contenus en choisissant les auteurs et les thématiques. Ensuite, l'éditeur se porte garant de la validité et de la qualité de ces contenus : il met en place des processus d'évaluation et confère de la crédibilité aux ouvrages qu'il publie. Pour finir, l'éditeur diffuse les contenus et s'occupe de les rendre visibles et accessibles. Bien évidemment, à ces trois instances s'en ajoute une quatrième, pragmatique, soit la mise en forme et la structuration. Le rapport complexe entre éditeur et auteur a toujours rendu difficile à saisir le rôle exact de l'éditeur en ce qui concerne cette dernière tâche. La mise en forme, pour la plupart des contenus, ne devait pas toucher au sens, qui était géré et assuré par l'auteur. C'est peut-être

là une des différences les plus évidentes entre l'édition papier et l'édition numérique. La mise en forme et la structuration des contenus ont à l'époque du numérique un rôle encore plus fondamental, et l'affaiblissement de la fonction auctoriale détermine que cette tâche soit aujourd'hui purement éditoriale.

En ce sens, quand on parle d'édition numérique, on peut utiliser le mot « éditorialisation », qui met l'accent sur les dispositifs technologiques qui déterminent le contexte d'un contenu et son accessibilité. Éditer un contenu ne signifie pas seulement le choisir, le légitimer et le diffuser, mais aussi lui donner son sens propre en l'insérant dans un contexte technique précis, en le reliant à d'autres contenus, en le rendant visible grâce à son indexation, à son référencement, etc.

C'est avec ce regard que ce livre a été pensé. Nous entendons par « édition numérique » un ensemble complexe de pratiques qui vont bien au-delà du rôle que l'éditeur a eu dans le modèle de l'édition imprimée à partir du XVIII^e siècle. L'édition numérique, en tant qu'éditorialisation, regroupe toutes les actions destinées à structurer, rendre accessible et visible un contenu sur le web. Même la recommandation d'un document dans un réseau social est une pratique d'éditorialisation. De ce point de vue, les questions relatives à l'histoire du web, son impact sur la circulation des connaissances dans notre société, le passage du web 1.0 au 2.0, puis au 3.0, ainsi qu'une réflexion plus philosophique sur la signification du mot « numérique » seront des sujets centraux de notre tractation. L'objectif de ce livre est d'expliquer ces enjeux et ces défis complexes à un public qui souhaite découvrir le monde de l'édition numérique. Il ne s'agit pas

d'un texte de recherche, mais d'un manuel qui propose une introduction à destination de tous ceux qui n'auraient aucune compétence particulière dans le domaine, qu'ils soient étudiants ou issus du monde de l'édition traditionnelle. Pour atteindre ce but, l'approche choisie mélange les points de vue historique, théorique et technique. Ce manuel est donc divisé en trois axes et treize chapitres.

Le premier axe illustre d'abord le rôle de l'éditeur et l'importance de la fonction éditoriale pour tracer ensuite l'histoire d'Internet et du web ainsi que celle des humanités numériques. Il s'agit de comprendre de quelle manière le numérique a acquis progressivement une importance fondamentale dans les pratiques éditoriales.

Le deuxième axe se concentre sur des questions plus théoriques afin de proposer des concepts clés pour faire comprendre au lecteur quelles sont les caractéristiques spécifiques de l'édition numérique, en donnant une description de ses retombées culturelles et en insistant sur l'importance et la fonction des métadonnées. Les deux derniers chapitres de cet axe discutent des modèles économiques possibles pour l'édition numérique, ainsi que des enjeux théoriques et politiques de l'accès libre aux différents types de contenus déjà produits et à venir.

Le dernier axe, plus long que les précédents pour répondre à un éventail de questions plus pointues et généralement moins connues, offre une présentation de questions techniques liées aux pratiques de l'édition numérique. Cet axe s'ouvre sur une description des protocoles d'Internet et du web, protocoles qui déterminent techniquement la manière dont les contenus circulent. Il offre ensuite un panorama des principaux formats de documents numé-

riques et de leurs spécificités, puis fait la lumière sur la gestion des métadonnées. Enfin, il propose une approche plus concrète de la production d'un objet éditorial numérique : une description des compétences basiques indispensables pour produire un livre au format ePub, les potentialités offertes par le numérique pour la recherche et la fouille de textes, et un aperçu des bonnes pratiques sous-jacentes à la conception d'un projet éditorial pour le web.

Ce manuel s'accompagne d'une version en ligne offrant le texte dans son intégralité sur le site de la collection « Parcours numériques », avec pour certains chapitres des versions plus longues, enrichies de contenus complémentaires (liens, images, etc.), afin de profiter au mieux des possibilités offertes par le numérique.

AXE HISTORIQUE

CHAPITRE 1

La fonction éditoriale et ses défis

Patrick Poirier et Pascal Genêt

Quel est, traditionnellement, le rôle de l'éditeur ? Ce rôle est-il, à l'époque d'Internet, encore indispensable ? La facilité de publication et d'auto-publication a-t-elle progressivement rendu inutile cet organe central ? Ce chapitre propose de répondre à ces questions en décrivant la fonction d'intermédiaire caractérisant l'éditeur. Il est le pont entre l'auteur et le lecteur, celui qui rend un contenu lisible. Cette fonction devient encore plus importante à l'ère du web, alors que, dans la surabondance des contenus, le lecteur ne peut que se perdre s'il n'est pas guidé par une fonction éditoriale.

Une fonction indispensable

Il n'est sans doute pas nécessaire de remonter aux premiers ateliers d'imprimerie de Gutenberg pour saisir tous les enjeux et les défis auxquels fait face aujourd'hui le milieu de l'édition, mais il n'est pas inutile, en revanche, de rappeler que, de Gutenberg à Diderot, puis de la Révolution française à la révolution industrielle, l'histoire de l'édition est peut-être avant tout celle, riche et tumultueuse, des idées et de la pensée, mais aussi celle de leur diffusion.

Aujourd'hui confrontée aux défis et aux promesses de la révolution numérique, on comprendra sans doute aisément que l'édition ou, plus précisément, que l'histoire de l'édition est moins celle de son évolution que des nombreuses « révolutions » – politiques et techniques – qui en ont profondément marqué le développement depuis ses origines. On ne s'étonnera pas, par exemple, que la fameuse *Lettre sur le commerce de la librairie* (1763) de Diderot, lettre dans laquelle il défendait une conception moderne de l'édition, paraisse dans le tumulte politique précédant les révolutions américaine et française. Étroitement liée à l'histoire des idées, l'évolution de l'édition, voire le métier même de l'éditeur, a historiquement accompagné un certain idéal démocratique qui est encore aujourd'hui le sien, mais dont la préservation s'annonce pourtant comme l'un des enjeux les plus importants pour les années à venir.

Au moment où, pour certains grands groupes éditoriaux, s'avère pratiquement impossible la publication d'un livre qui n'irait pas dans le sens immédiat du profit, il faut en effet s'inquiéter de ce qui relève de plus en plus d'une forme de « contrôle de la diffusion de la pensée dans les sociétés

démocratiques [...]. Le débat public, la discussion ouverte, qui font partie intégrante de l'idéal démocratique, entrent en conflit avec la nécessité impérieuse et croissante de profit[1] », comme le soulignait déjà l'éditeur André Schiffrin il y a quelques années. Ambassadeur culturel et intellectuel, intermédiaire essentiel entre l'auteur et le lecteur, l'éditeur a donc tout lieu, aujourd'hui plus que jamais, d'être vigilant, au risque de céder la place à une « édition sans éditeurs ». Le spectacle inquiétant que nous offre aujourd'hui le monde de l'édition, dont la surproduction n'est hélas qu'un des tableaux, nous rappelle à tout le moins que le capital financier investi par un individu ou un groupe de communication ne saurait à lui seul suffire de légitimation, de caution, à un texte, à un manuscrit ou à un document.

L'homme de lettres et l'entrepreneur

Ce n'est pas dire, pourtant, que les responsabilités qui incombent à l'homme de lettres proscrivent toute recherche de profit. La polysémie du mot « éditeur » traduit la réalité d'une profession qui, dès l'origine, est partagée entre deux rôles distincts : la fonction éditoriale et la fonction entrepreneuriale. La fonction éditoriale (*editor*) est propre à celui (ou celle) qui découvre, qui consacre et qui dirige la publication d'ouvrages et, plus largement, qui acquiert par le fait même un statut professionnel et une valeur symbolique spécifique dans le champ littéraire. Elle comprend le travail de développement du manuscrit et d'accompagnement de l'auteur, « la mise au point du texte et le choix des documents

1. André Schiffrin, *L'édition sans éditeurs*, La Fabrique éditions, 1999, p. 93-94.

éventuels qui l'accompagnent, la conception d'une maquette et le choix des éléments strictement techniques (format, papier, couverture, mode d'impression)[2] ». La fonction entrepreneuriale (*publisher*) est définie par des rôles et des responsabilités de gestionnaire et d'administrateur propres aux conditions de production et de diffusion des ouvrages[3].

Un commerçant ou un homme d'affaires ne devient éditeur qu'à partir du moment où il prend sur lui la (double) responsabilité matérielle et morale d'une œuvre. C'est là son rôle, sa fonction. C'est cette « contre-signature » à la fois financière (économique) et idéologique qui, dans une certaine mesure, fait d'un manuscrit un livre et d'un écrivain un auteur. En d'autres mots, les éditeurs sont ceux qui parviennent à concilier l'homme de lettres et l'entrepreneur, le « serviteur de la pensée française », pour reprendre ici une expression d'André Grasset, et le marchand. Comme le fait remarquer Pierre Bourdieu, « ces personnages doubles, par qui la logique de l'économie pénètre jusqu'au cœur du sous-champ de la production pour producteurs, doivent réunir des dispositions tout à fait contradictoires : des dispositions économiques qui, dans certains secteurs du champ, sont totalement étrangères aux producteurs, et des dispositions intellectuelles proches de celles des producteurs dont ils ne peuvent exploiter le travail que pour autant qu'ils savent l'apprécier ou le faire valoir[4] ».

---

2. Bertrand Legendre, *L'édition*, Le Cavalier Bleu éditions, 2008, p. 5.

3. Yves Winkin, « L'agent double. Éléments pour une définition du producteur culturel », *Les conditions économiques de la production culturelle*, Bruxelles, Cahiers JEB, n° 4, 1978, p. 43-50.

4. Pierre Bourdieu, « Le champ littéraire », *Actes de la Recherche en Sciences Sociales*, 1991, vol. 89, n° 1, p. 5.

Constituer un catalogue : la sélection des textes

Cette double dimension, économique et symbolique, de l'éditeur – celui qui a bâti sa légitimité culturelle en découvrant et en consacrant des auteurs – est intimement liée à sa personnalité, en plus de nécessiter des compétences particulières. Vu de l'extérieur, le monde du livre revêt parfois une « aura quasi mystique[5] » permettant à ses artisans d'entretenir une figure « mythologique » de l'éditeur, pour lequel « la découverte est à la conquête ce que l'invention est à la production : la manifestation de son autorité, ce par quoi lui sont reconnus le mérite de la révélation et le privilège de la propriété[6] ». Pourtant, être éditeur signifie tout simplement être un « professionnel de la chose éditoriale, celui qui possède un savoir et des compétences spécifiques, le savoir éditer[7] ». En ce sens, le terme « éditeur » renvoie à des rôles et à des fonctions qui contribuent à la dimension symbolique de la fonction éditoriale. C'est donc dire que, dans le cas de maisons d'édition de moyenne ou de grande envergure, ces rôles et ces fonctions peuvent être attribués ou répartis entre divers intervenants. La « sélection » d'un texte, qui constitue en quelque sorte la fonction éditoriale première, n'échappe pas à cette règle.

En choisissant un auteur, un texte, en publiant un titre, un éditeur contribue à consacrer un auteur tout en se faisant un nom auprès de ses pairs et du milieu littéraire. Il se constitue par le fait même une image de marque et une

5. Bertrand Legendre, *L'édition*, *op. cit.*, p. 9-11.
6. Hubert Nyssen, *La sagesse de l'éditeur*, L'Œil neuf éditions, 2006, p. 23.
7. Bertrand Legendre et Christian Robin, *Figures de l'éditeur*, *Représentations*, *savoirs*, *compétences*, *territoires*, Actes du colloque des 13 et 14 mai 2005 (Université Paris 13), Nouveau Monde éditions, 2005, p. 11.

identité singulières, tout en s'imposant comme le médiateur indispensable entre une œuvre et un marché, dans l'espoir de répondre aux désirs, aux attentes et aux goûts du public.

En ce sens, le catalogue d'un éditeur, c'est-à-dire l'ensemble des auteurs et des titres que chapeaute la maison d'édition, se veut en quelque sorte le reflet de cette coexistence antagoniste entre des valeurs économiques et culturelles, en plus de témoigner du parcours intellectuel, de la personnalité et, bien sûr, des intérêts de l'éditeur, qu'ils soient artistiques, techniques, économiques ou littéraires. Que l'éditeur soit seul à effectuer le « tri » parmi les textes et documents reçus ou que cette tâche soit confiée, en tout ou en partie, à des lecteurs extérieurs ou à un comité de lecture (dans le cas de maisons d'édition de plus grande taille), ce sont donc ces intérêts, ces choix esthétiques et/ou idéologiques, une certaine ligne ou politique éditoriale, qui influenceront de manière déterminante le processus de sélection des textes retenus. La décision finale, bien entendu, incombe à l'éditeur, ou à un comité éditorial, voire à un directeur de collection qui, de même, doit tenir compte d'un ensemble de critères et de caractéristiques présidant à la sélection des textes pour la collection dont il a la charge.

L'élaboration et la mise en forme du texte

Une fois le texte retenu (qu'il s'agisse d'une œuvre soumise ou d'un ouvrage commandé), c'est-à-dire une fois qu'est prise la décision de publier un texte, l'éditeur (mais il peut aussi s'agir du directeur de collection ou d'un directeur littéraire) doit ensuite accompagner l'auteur dans l'élaboration et la rédaction finale de son ouvrage, lui offrir son

soutien, ses conseils et ses encouragements, bref, travailler de pair avec l'auteur ou l'écrivain afin de mener le texte à son terme. Il incombera alors à l'éditeur, de concert avec les responsables financiers dans le cas de maisons d'édition de plus grande taille, d'établir le tirage envisagé et de mettre sous contrat l'auteur, entente qui définira les droits et les devoirs des deux parties.

S'ensuit une série d'étapes qui relève moins des fonctions éditoriales de l'éditeur que d'« artisans » du livre. C'est le cas de la révision et de la correction des textes, tâches des plus importantes qui sont presque toujours confiées à des correcteurs externes, ou de la traduction, également confiée à des professionnels lorsqu'elle s'impose, ou encore du travail de coordination entre ces divers intervenants, l'éditeur et l'auteur, dont est souvent chargé le secrétaire d'édition.

La fabrique du livre

Parfois seul (dans le cas de maisons d'édition artisanales), parfois chef d'orchestre, il revient aussi à l'éditeur de veiller à la fabrication du livre, c'est-à-dire d'en concevoir, mettre en place et suivre le processus de fabrication. Cette fonction éditoriale, à la fois technique et artistique, concerne la conception matérielle du livre qui, une fois le texte établi et corrigé, peut être mis en branle. Compte tenu de la complexité du processus et des nombreuses tâches qui sont impliquées, il est rare, aujourd'hui, même dans des maisons de type artisanal, que l'éditeur puisse seul prendre en charge chaque étape de la fabrication du livre. À ce stade, il faut en effet souvent faire appel à un infographiste qui veillera à la mise en pages et à la production des épreuves de l'ouvrage. De concert avec l'éditeur ou un directeur artistique, ou suivant

les caractéristiques graphiques et matérielles propres à une collection donnée, l'infographiste doit déterminer la taille et le type de caractères retenus, superviser la reproduction et l'insertion des illustrations s'il y a lieu (en couverture, notamment), établir et calibrer la maquette du livre, en estimer le nombre de pages, etc. La première épreuve produite sera alors envoyée en correction et il n'est pas rare qu'à ce stade l'éditeur souhaite également en remettre une copie à l'auteur pour fins de révision. Simultanément, l'éditeur ou le directeur de production soumettra l'épreuve à un imprimeur afin d'en estimer les coûts d'impression, une fois établi le choix du papier, le mode d'impression, ainsi que le type de couverture et de reliure du livre.

Au même titre que la sélection des textes, les conditions de production matérielle du livre (sa fabrication) tiennent également compte de critères et de choix esthétiques. Et si le catalogue des titres et des auteurs d'une maison d'édition est garant d'une ligne éditoriale, de la fiabilité de son contenu, de ses orientations idéologiques, voire politiques, le nom d'un éditeur ou d'une maison d'édition est aussi garant de la qualité matérielle et graphique d'un ouvrage. Sans chercher à surestimer l'importance de la valeur esthétique d'un livre, sa présentation matérielle (qualité de l'impression, choix des papiers et cartons), sa signature graphique (élégance, recherche du design, esthétique générale) sont pourtant eux aussi responsables de la signature et de l'image (de marque) de la maison d'édition. Loin d'être négligeables, ces facteurs peuvent donc s'avérer un atout indéniable. Et s'il est certain que la dimension économique (financière) revêt, en ce cas, une importance réelle, plusieurs maisons d'édition de petite taille parviennent néanmoins à produire

des livres de qualité, témoins en cela d'une attention portée à l'ouvrage bien fait.

La diffusion et la distribution du livre

En apposant son nom ou celui de sa maison d'édition au bas d'un livre, un éditeur s'engage *de facto* dans l'espace public : il participe activement à un acte de communication qui déborde largement la fabrication du livre. Non seulement l'éditeur – ou le responsable de la promotion et de la commercialisation – doit se faire le médiateur de l'œuvre auprès des libraires et du public, titre dont il aura travaillé à la mise en forme et au devenir final, mais il participe, grâce à la diffusion du livre, à la promotion de valeurs littéraires, à la légitimation d'une vision esthétique, d'un mouvement artistique ou d'un courant de pensée. C'est dire, somme toute, qu'il remplit par là un rôle social en assurant le développement et la pérennité de la vie intellectuelle, littéraire ou, plus largement, culturelle de la société. Plus encore, cette « médiation éditoriale inscrit le texte dans un projet d'entreprise et l'insère dans un processus de communication sociale qui lui donne un sens[8] ».

Si nombre de petites maisons d'édition assument plusieurs des tâches liées à la diffusion du livre (activités de promotion, communiqués de presse, rencontres avec les libraires, relations avec le milieu de la critique, etc.), la plupart des éditeurs préfèrent au contraire confier aux soins d'entreprises spécialisées, en tout ou en partie, ce qui relève de la distribution (entreposage des exemplaires, gestion des commandes, facturation et comptabilité, etc.),

8. *Ibid.*, p. 17.

plutôt que de devoir se contenter de circuits plus marginaux. Pour essentielle que soit une distribution optimale du livre publié (la révolution numérique offrant à ce titre de nouvelles possibilités), il n'en demeure pas moins que c'est en participant et en veillant au plus près aux activités de diffusion qu'un éditeur s'assure de la visibilité d'un livre dans l'espace public. À la participation traditionnelle de l'éditeur aux divers salons du livre et foires internationales, à la présence espérée d'un auteur dans les médias traditionnels (critiques littéraires dans la presse écrite, les revues et les magazines culturels, émissions littéraires à la radio et à la télévision) s'ajoute aujourd'hui la nécessité de plus en plus incontournable d'assurer une diffusion efficace du livre sur Internet et dans l'ensemble des réseaux sociaux.

De tout temps, l'éditeur aura été un intermédiaire intéressé entre le lecteur et l'auteur, le public et l'écrivain (ou créateur). Si cette activité (l'édition au sens large) doit aujourd'hui trouver de nouvelles formes, de nouvelles façons de faire dans la sphère numérique – et il s'agit là d'un enjeu majeur pour tous les éditeurs –, elle n'en demeure pas moins essentiellement une médiation qui, comme par le passé, requiert de l'éditeur qu'il aménage un « espace », dessine les contours d'un « lieu » où puissent se « rencontrer » auteurs et lecteurs, et où puissent circuler discours et idées auxquels il aura donné son *imprimatur*, sa caution.

Défis et enjeux

Le monde de l'édition – et plus encore le milieu de l'édition littéraire –, au Québec comme ailleurs, connaît aujourd'hui un des bouleversements les plus importants de son histoire.

Non seulement l'idéologie de marché et certaines politiques culturelles des dernières années ont profondément transformé le paysage de l'édition, mais les maisons d'édition et les librairies (ces dernières étant un maillon jadis essentiel dans la chaîne de diffusion) doivent aujourd'hui faire face aux défis que soulève la « révolution » numérique.

Pour le milieu de l'édition, cette révolution s'avère pourtant une chance, une ouverture inattendue, voire inespérée, à de nouvelles possibilités. Plus encore, elle conforte l'éditeur dans son rôle traditionnel, dans sa fonction première de médiateur indispensable à la création et à la diffusion du livre, quelle qu'en soit la forme. D'hier à aujourd'hui, du livre papier à son pendant dématérialisé, s'impose encore et toujours la nécessité d'un intermédiaire entre l'auteur et le lecteur; l'avènement du numérique en souligne d'ailleurs plus que jamais l'importance. « Au milieu de tant d'incertitude, et face à la marée montante des textes en tous genres qui saturent le net, il est certain que la marque des éditeurs sera demain plus que jamais un repère. Happés par les blogs de toutes natures, envahis par les textes qui circulent par millions sur la toile, les lecteurs de demain auront besoin de certifications, de labels, de garanties de qualité. Les éditeurs leur apporteront cette caution[9]. »

Historiquement, il faut en effet rappeler que l'éditeur fait son « apparition » dans la société « au moment où se crée un espace public pour la littérature[10] ». Depuis le moment où se dessine de manière encore limitée une telle agora jusqu'à

9. Olivier Bessard-Banquy, *L'industrie des lettres*, Pocket, 2012, p. 526.

10. Jacques Michon (dir.), *Histoire de l'édition littéraire au Québec au XX^e siècle. 1- La naissance de l'éditeur, 1900-1939*, Fides, 1999, p. 25.

l'émergence d'un espace public aux dimensions planétaires (le *World Wide Web*), la nécessité d'un truchement entre l'auteur et le lecteur demeure et, s'il faut en croire le portrait que trace Olivier Bessard-Banquy, elle s'imposera de manière plus essentielle encore dans les années à venir. Car à moins de s'en tenir au tintamarre de l'« opinion » et de se contenter de la rumeur uniforme qui lui tient lieu de « discours » sur la toile, le lecteur aura tout intérêt à se tourner vers quelque intermédiaire qui saura lui faire entendre autre chose, et mieux, et plus clairement. Devant la pléthore de titres qui encombrent les tables des libraires et qui déferlent dans l'espace virtuel, lequel choisir ?

On a longtemps dit de l'éditeur qu'il était le « banquier symbolique » de l'écrivain, mais cette métaphore financière dit aussi tout le poids d'un certain capital symbolique qui relève d'abord et avant tout de la responsabilité de l'éditeur, gage précieux dans l'univers virtuel. C'est en effet en engageant davantage que les conditions de production matérielle d'un livre (aussi importantes demeurent-elles pour le livre imprimé, et aussi dématérialisées soient-elles pour le livre numérique) que l'on assume le rôle et les fonctions d'éditeur. Ce ne sont pas tant les promesses de la révolution numérique qui, en ce sens, représentent un enjeu inquiétant, mais bien davantage le spectre d'une « édition sans éditeurs ». Comme le fait hélas remarquer Gustavo Sorá : « dans le passé, l'éditeur devait être (re)connu dans le milieu intellectuel ; aujourd'hui, il doit l'être dans le marché[11] ».

11. Gustavo Sorá, « La maison et l'entreprise », *Actes de la recherche en sciences sociales*, vol. 126, n° 1, 1999, p. 99.

Ainsi, souligne Bertrand Legendre, « pratiques artisanales et stratégies industrielles se rencontrent dans ce secteur qui connaît fortement les logiques de concentration et le phénomène de financiarisation[12] ». L'inflation industrielle des titres et la multiplication des fusions-acquisitions de maisons d'édition ou de groupes éditoriaux témoignent éloquemment, chaque jour, du triomphe d'une « édition sans éditeurs », selon l'expression d'André Schiffrin, où « le quantitatif prime sur le qualitatif[13] ». Devant les risques (réels) de concentration du milieu éditorial au sein de conglomérats dont les activités débordent largement le domaine du livre, devant la subordination de plus en plus importante des politiques éditoriales au profit d'impératifs commerciaux, face à l'abandon des politiques de fonds qui cèdent la place à la recherche du profit vite fait, rapide, insensé, on mesure mieux combien une large part de l'industrie du livre semble avoir relégué aux oubliettes le rôle premier de l'éditeur, cette caution dont le livre, tout particulièrement dans l'espace ouvert par la révolution numérique, a peut-être plus que jamais besoin. Car s'il est vrai que la révolution numérique entraîne aujourd'hui de réels changements dans la chaîne traditionnelle du livre (notamment en ce qui concerne sa production et sa distribution), ce sont davantage les concentrations économiques dans le milieu éditorial et la systématisation d'un « marketing forcené » qui remettent plus sérieusement en question « le rôle public

12. Bertrand Legendre, *L'édition*, *op. cit.*, p. 5.

13. Luc Pinhas, *Le livre et l'exception culturelle : le cas des politiques publiques du livre au Québec*, Paris, 2005, p. 10.

de l'édition et ses métiers éditoriaux comme espace et lieux de communication du savoir et de la connaissance[14] ».

Ainsi, précise Hubert Nyssen, « il y a de quoi faire hésiter puis vaciller des éditeurs, bons passeurs, dans le catalogue desquels des prédateurs se disent qu'il peut y avoir des reprises à faire, des compétences à s'approprier et une respectabilité à acquérir. Et ainsi nombre de maisons d'édition, détournées de leur vocation par promesses et mirages, cèdent-elles et consentent à s'adosser à des groupes financiers, voire à troquer leurs accomplissements pour une rente qui mettra leurs fondateurs à l'abri des besoins. Et comme cette dévoration qui a sa logique – oh oui, elle en a une, la logique de l'argent ! – s'accompagne d'une mainmise sur les grands moyens du monde médiatique et sur les réseaux de distribution, les réfractaires sont condamnés à la solitude et au silence. Et avec eux des auteurs devant lesquels les portes de l'édition se fermeront[15] ».

Le monde du livre est aujourd'hui plongé au cœur de nouveaux défis, voire de tensions, imposés par l'émergence des nouvelles technologies et, en particulier, par l'arrivée du numérique qui redéfinit les modes de production, de diffusion et, plus largement, les pratiques culturelles. Les transformations sont multiples : dématérialisation des supports, disponibilité illimitée des contenus dans le temps et l'espace, gestion et mise à jour de l'information pour faciliter le référencement, émergence de nouveaux modèles d'affaires et de nouveaux réseaux de diffusion et de distri-

14. Patrick Tillard, « Critique de l'admirable », *Spirale*, n° 243, « Nouveaux enjeux de l'édition », hiver 2013, p. 37.

15. Hubert Nyssen, *La sagesse de l'éditeur*, *op. cit.*, p. 101-102.

bution, etc. L'univers numérique laisse même déjà présager un avenir où, tout en aménageant un espace virtuel (pages, sites, plateformes numériques, etc.) préservant l'ensemble des fonctions éditoriales, la nécessité de ce que l'on appelle aujourd'hui une « maison d'édition » sera sans doute remise en question, voire appelée à disparaître. Autant de changements auxquels les métiers traditionnels de l'édition – et d'abord celui de l'éditeur lui-même – sont appelés à faire face dans l'urgence. Entrer en édition, affirme Nyssen, « c'est entrer dans la crise. Et c'est fort bien ainsi. La crise attire la détermination[16] ».

16. *Ibid*, p. 21.

CHAPITRE 2

D'Internet au web

Alain Mille

La naissance d'Internet, puis du web, a sans doute été le phénomène ayant déterminé le plus fort changement dans les modèles de production et de circulation des contenus. Ce ne sont pas tant les « nouvelles technologies » en général, mais le réseau en lui-même qui a bouleversé notre rapport à la connaissance. À ce propos, il est bon de rappeler qu'Internet et web ne sont pas synonymes – et ne sont par ailleurs pas nés au même moment. Internet est le réseau physique qui permet l'échange des données entre plusieurs ordinateurs et le web, l'ensemble des documents formatés en HTML accessibles avec un navigateur via le protocole HTTP. Ce chapitre raconte l'histoire allant des premières idées de réseau dans les années 1950 à la création d'Internet dans les années 1970, jusqu'à l'apparition du web et à son impressionnant développement à partir de la moitié des années 1990. Connaître cette histoire et ces étapes est fondamental pour comprendre les raisons des changements majeurs qui touchent aujourd'hui le monde de l'édition.

1950-1965 : faux départ

Les militaires américains réalisent dans les années 1950-1960, au plus fort de la guerre froide, que le système de communication de leurs forces armées, très hiérarchique par nature, est particulièrement fragile dans le cas d'un conflit nucléaire. La notion de réseau maillé avec des nœuds d'interconnexion naît à cette époque pour fournir la robustesse nécessaire au système de communication qui serait utilisé en cas de déflagration nucléaire. Paul Baran de RAND Corporation, sorte de *think tank* de l'époque, pose les principes d'un réseau centrifuge parsemé de nœuds d'interconnexion, avec des messages transitant sans avoir de route définitivement attribuée, mais seulement des indications du *from* et du *to* dans les champs du message. La route peut alors être redéfinie dynamiquement à chaque nœud si la route optimale précédemment calculée n'est plus disponible, de manière à chercher par tous les moyens à router le message vers un autre nœud à même de le faire progresser vers sa destination : routage dit de la « patate chaude ». Ce système, même avec des composantes bon marché à chaque nœud, est particulièrement simple et robuste, ce qui va être démontré formellement par un doctorant du Massachusetts Institute of Technology (MIT), Leonard Kleinrock. Donald Davies le met en œuvre en 1965 au National Physics Laboratory.

Malgré l'intérêt porté par les militaires, et les démonstrations faites, la multinationale américaine de télécommunications AT&T refuse de développer le réseau tel qu'il est décrit. En effet, ses dirigeants le considèrent comme une sorte d'auto-concurrence fatale pour leur *business*. Le projet

de Baran est évalué à 60 millions de dollars de l'époque, alors que le système qu'il peut remplacer coûte 2 milliards de dollars par an. En 1966, le projet est remisé sur les étagères.

1962-1968 : ARPAnet

Une autre structure va être au centre du développement d'Internet : l'ARPA (*Advanced Research Projects Agency*). En 1962, J.C.R. Licklider y est embauché pour s'intéresser d'une part au contrôle-commande et d'autre part aux sciences du comportement[1]. À partir de cette impulsion, l'orientation vers des systèmes informatiques interactifs, collaboratifs, ne cesse de se développer au sein de l'ARPA : Ivan Sutherland, après le développement du *SketchPad* au MIT, Robert Taylor issu de la NASA (*time sharing*), puis Lawrence Roberts qui, pour mettre en relation les chercheurs et partager non seulement les processeurs mais aussi tous leurs résultats, développe le premier réseau de communication numérique. En octobre 1965, l'ordinateur *Lincoln Labs TX-2/ANS/Q-32* communique avec le SDC's Q32 de Thomas Marill qui a développé le réseau. C'est ce réseau qui, plus tard, en 1968, prendra le nom d'ARPAnet. Le réseau développé à l'ARPA est dans sa technologie et ses principes très proche de ce qui a été proposé par Baran à la RAND Corporation, et mis au placard par AT&T. Lawrence Roberts aurait toutefois indiqué qu'il n'était pas au courant des travaux de Baran jusqu'à octobre 1967.

C'est une petite société, BBN (Bolt, Beranek and Newman), qui gagne le concours lancé par l'ARPA pour réaliser un

1. J.C.R. Licklider, « Man-Computer Symbiosis », *IRE Transactions on Human Factors in Electronics*, vol. HFE-1, mars 1960, p. 4-11.

réseau informatique de partage de ressources *(resource sharing computer network*). Le chef de projet, Frank Heart, a choisi son équipe et sa méthode de travail : « *Get the very, very best people and in small numbers, so they can all know what they're all doing.* » IBM, dans le même temps, déclare qu'un tel système ne peut pas être développé sans un très gros budget. En 1969, le réseau fonctionne et deux machines communiquent ; en 1971, il y a 15 nœuds. En 1972, le réseau est lancé à la conférence ICCC (*International Computer Communication Conference*) à Washington, et les 60 terminaux connectés permettent aux conférenciers de réaliser que ce type de réseau fonctionne réellement.

1969-1978 : Internet

Mais si le réseau ARPAnet est bien une réalité, il manque une pièce logicielle maîtresse pour que l'on puisse considérer qu'Internet est né. Il faut décrire les protocoles de communication, ce qui est fait en 1969 par un groupe d'étudiants relativement informel. Peu sûrs de leur légitimité et cherchant à discuter leurs idées pour édicter les protocoles, ils proposent des *Request For Comments* (RFC). Il s'agit pour Steve Crocker et ses amis de laisser ouvertes au maximum les possibilités de participation des pairs pour améliorer le protocole. Cette ouverture fait une grande partie du succès de la dynamique extraordinaire qui s'ensuit. Le premier résultat concret de ce travail est la création du NCP (*Network Control Protocol*), mais pas encore d'Internet. C'est par l'interconnexion d'un réseau filaire classique ARPAnet et d'un autre réseau hertzien de même nature (paquets) que la première communication par interconnexion de réseau

est réalisée par le Stanford Research Institute. La nécessité d'interconnecter ARPAnet avec les réseaux hertziens (PRNET) et les réseaux satellites (SATNET) pousse dans le sens d'Internet. Il est à noter que les travaux des chercheurs impliqués en France dans le projet Cyclades ont inspiré l'ensemble de ces travaux. En effet, Cyclades propose déjà à l'époque trois couches réseau : la couche transmission, la couche transport et la couche application. En mai 1974, Vint Cerf et Robert Kahn publient ce qui devient le protocole TCP (*Transport Control Protocol*). C'est après plusieurs révisions de TCP que le protocole IP (*Internetworking Protocol*) est proposé par Vint Cerf, Jon Postel et Danny Cohen en janvier 1978.

Fin des années 1970 et début des années 1980 : démocratisation des ordinateurs personnels

Il est donc démontré que ce type de réseau peut couvrir le monde, car il permet d'interconnecter les réseaux filaires, hertziens et satellites de manière robuste et économique. Cela peut s'appliquer au téléphone, naturellement, mais, encore une fois, l'industrie téléphonique se lance dans une version centripète (protocole X.25) de la transmission de paquets, issue des travaux du CCITT (*Consultative Committee on International Telegraphy and Telephony*) pour lesquels les choses doivent être contrôlées de bout en bout, sans reprendre à leur compte les notions de distribution et de diversité des réseaux. Le protocole X.25 garantit un fonctionnement de haute qualité quelle que soit la configuration, mais c'est une propriété finalement assez rarement nécessaire et le caractère centralisé et fermé de l'approche bloque le développement de cette technologie.

Ce choix des opérateurs de téléphonie n'est abandonné qu'en 2011 mais, malgré le faible soutien des professionnels des réseaux, Internet se développe de manière tout à fait impressionnante.

Qu'est-ce qui explique ce développement ? De la même façon que l'idée du réseau distribué s'est imposée à l'ARPA pour la mise en connexion des utilisateurs, c'est le surgissement soudain des ordinateurs personnels qui fait sortir l'usage d'Internet des universités et centres de recherche pour l'ouvrir à tous les secteurs d'activité de la société. Après quelques réalisations visionnaires, mais avec un succès limité, il faut attendre Ken Olsen et Harlan Anderson qui créent la Digital Equipment Corporation (DEC) avec le PDP1, sans doute le premier ordinateur interactif jamais conçu. Suivi du PDP8 et d'autres mini-ordinateurs, il scelle la fin du traitement par lot. Si les interfaces sont textuelles, on assiste très vite au développement d'interfaces graphiques, intuitives et, dès 1968, Engelbart peut faire sa célèbre démonstration sur l'usage de la souris. Si les idées et même certaines réalisations sont déjà là, il s'agit de systèmes confidentiels qui ne peuvent pas facilement se décliner en masse. L'apparition des microprocesseurs va changer la donne avec l'application de l'invention des circuits intégrés à la réalisation de calculateurs sur une surface très réduite. L'Intel 4004 est d'abord considéré comme un simple circuit logique peu évolué par les sociétés informatiques. Mais, en janvier 1975, la revue *Popular Electronics* annonce le Altair 8800, conçu autour du Intel 4004 : « *The era of personal computing in every home… has arrived.* » Le succès dépasse largement les attentes des concepteurs, même si l'usage de ces ordinateurs portables est difficile et laborieux. Les micro-

processeurs se succèdent (Motorola, Texas Instrument, Intel, etc.) avec des capacités de calcul et de mémoire qui augmentent significativement, créant les conditions nécessaires d'une offre massive d'ordinateurs personnels, mais pas encore les conditions suffisantes à l'émergence d'un phénomène mondial. Un club nommé le Homebrew Computer Club rassemble dans les années 1970 un mélange étonnant de militants de l'informatique de Berkeley et d'amateurs d'électronique. Les groupes hippies considèrent l'ordinateur comme un outil de réalisation personnelle et appellent dès 1972 à créer la Compagnie informatique du peuple (People's Computer Company). C'est dans cet état d'esprit que Steve Wozniak introduit le premier Apple I au Homebrew Computer Club, et c'est avec l'Apple II, en 1977, que la notion d'ordinateur personnel, installé par la personne à partir d'éléments très simples à assembler, s'impose. Ce qui explique ce succès, c'est qu'Apple met à disposition un manuel de référence qui détaille l'ensemble du code source et les schémas électriques et électroniques. Chacun peut alors développer librement de nouveaux logiciels et de nouveaux périphériques. Visicalc (1979) marque un tournant en fournissant à l'utilisateur non informaticien la possibilité de concevoir des traitements, de réaliser des calculs : la feuille de calcul est née. Cette fois, le marché s'impose vraiment et même IBM s'en aperçoit. En août 1981, IBM vend son premier PC, appelé le 5150. L'entrée d'IBM sur le marché se fait dans la précipitation, ce qui entraîne l'abandon du processus standard de développement au profit d'un développement par une petite équipe, et l'ensemble logiciel et matériel est ouvert, à l'image de ce qui a été fait par Apple. IBM conclut un accord avec une petite société,

Microsoft, qui est chargée de fournir un système d'exploitation au PC, avec la possibilité de vendre son système aux concurrents d'IBM. L'idée d'IBM est de permettre à ses utilisateurs d'installer facilement ce qui est développé par d'autres sur ses propres machines, mais c'est aussi la possibilité pour d'autres de cloner rapidement et efficacement sa propre machine avec des logiciels attractifs. Le phénomène des ordinateurs personnels, intégrant ensuite un nombre étonnant d'innovations, d'interactions, de design, de logiciels, etc. fournit les conditions alors suffisantes pour qu'Internet devienne le support d'interconnexion d'un nombre toujours plus grand d'utilisateurs d'ordinateurs personnels.

1975 : The Well, le premier réseau social

Les conditions de la massification de l'usage de l'informatique se renforcent, mais encore faut-il que les utilisateurs puissent accéder à Internet depuis leur domicile, par leur liaison téléphonique.

Les conditions d'utilisation du téléphone sont particulièrement réglementées, au point qu'il n'est pas autorisé de modifier une quelconque partie d'un récepteur téléphonique aussi bien aux États-Unis qu'en Europe. Ce n'est qu'à partir de 1975 qu'il est autorisé de mettre sur la ligne téléphonique d'autres terminaux que ceux livrés par l'opérateur. Ward Christensen (Bulletin Board Systems) invente la notion de modem et propose le xmodem pour les PC en 1977. Un *hacker* (terme qui faisait alors référence à des amateurs experts), Tom Jennings, développe Fidonet permettant avec un modem très simple de mettre en communication

des ordinateurs personnels entre eux (courriel et listes de discussions).

The Well utilise alors ces outils pour rassembler une communauté très dynamique autour d'une liste de diffusion. Il s'agit là du tout premier réseau social.

ARPAnet et la DARPA (Defense Advanced Research Projects Agency) connaissent ces développements, mais restent concentrés sur l'interconnexion d'ordinateurs puissants de leurs réseaux. D'autres réseaux de spécialistes se développent d'ailleurs sur ce protocole (MILNET, DECNET, MEFENET, CSNET, BITNET, etc.), alors que la téléphonie tolère les modems, mais ne s'intéresse toujours pas à la commutation de paquets.

1984 : le réseau des réseaux

Quand la NSF (National Science Foundation) décide d'interconnecter ses principaux réseaux régionaux par une dorsale (*backbone*) commune, en 1984, c'est véritablement Internet qui est élevé au niveau des grands réseaux scientifiques, diffusant TCP/IP sur les grosses machines comme sur les ordinateurs personnels connectés à ces réseaux. NFSNET se développe très rapidement avec plus de 160 000 machines connectées en 1989. En février 1990, ARPAnet est délié de ses liens avec les militaires et les connexions d'ordinateurs à NFSNET se mettent à croître de manière exponentielle : 500 000 en 1991, 1 million en octobre 1992, 2 millions en octobre 1993, etc. Dans le même temps, les nouvelles applications de recherche d'information comme Gopher, Veronica et Archie sont déployées et génèrent un trafic plus important que le courriel, ce qui préfigure le développement massif du web.

1979-2009 : une histoire des structures de gestion de l'Internet

Le développement de ce réseau de réseaux s'accompagne d'une mise en place originale de méthodes de gestion d'une structure toujours marquée par ses racines ouvertes et, d'un certain point de vue, libertaires. Vint Cerf avait créé l'Internet Configuration Control Board (ICCB) en 1979, transformé par Barry Leiner en Internet Activities Board (IAB) en 1983, réorganisé en 1989 entre l'Internet Engineering Task Force (IETF) et l'Internet Research Task Force (IRTF). En 1992, l'Internet Society (IS) est créée pour chapeauter l'IAC et l'IETF. Malgré la complexité des structures, le principe initial des *Requests For Comments* est gardé, avec cette idée du chantier sans fin. La question de l'attribution des noms de domaine (*Domain Name System* - DNS) a une histoire à rebondissements : dès 1983, une Internet Assigned Numbers Authority (IANA) est créée *ad hoc*, alors que le principe d'un système de nommage de domaine est imaginé la même année. En 1992, anticipant l'ouverture d'Internet, la NSF sélectionne la compagnie Network Solutions pour mettre à disposition les services d'assignation des noms de domaine et des adresses IP associées, en relation étroite avec l'IANA. Cette société a le monopole durant toute la durée de son contrat avec la NSF, jusqu'en 1998. Dès 1996, toutefois, l'International Ad Hoc Committee (IAHC) propose une nouvelle approche pour d'une part réserver des domaines de haut niveau aux gouvernements et d'autre part mettre en place des règles d'arbitrage pour éviter les conflits de nom. L'IAHC est agréée par les comités publics compétents, et ses conclusions reprises pour

créer une société à but non lucratif, l'Internet Corporation for Assigned Names and Numbers (ICANN) : « *it shall operate for the benefit of the Internet community as a whole* ». En 2009, après plusieurs controverses mondiales, l'ICANN cesse d'être uniquement contrôlé par les États-Unis, mais est piloté par une gestion intergouvernementale (Governmental Advisory Committee).

1990 : le web

C'est au CERN (Organisation européenne pour la recherche nucléaire) que se conçoit une sorte de « clé de voûte » parachevant l'environnement numérique qui est en train de s'étendre sur l'ensemble du globe par l'interconnexion des réseaux et par la mise en relation d'ordinateurs personnels de plus en plus nombreux. Les utilisateurs avides de pouvoir les utiliser comme dispositifs d'écriture-lecture collectifs et collaboratifs avec l'information circulante comme principe restent des amateurs surpassant les difficultés techniques pour y parvenir.

Dans les années 1980-1990, le CERN possède de très nombreux ordinateurs incompatibles entre eux, exploités par des chercheurs du monde entier. Tim Berners-Lee y développe d'abord, dans les années 1980 un logiciel, Enquire, ressemblant à une sorte de wiki (documents reliés par des hyperliens bidirectionnels) permettant d'expliciter les relations entre les personnes, les programmes et les systèmes qui se croisent au CERN. Cette idée rappelle naturellement celle de Vannevar Bush (1945) qui avait imaginé un mécanisme d'hypercartes (*Memex*), et Tim Berners-Lee s'appuie explicitement sur les contributions de Ted Nelson et Douglas Engelbart. Ted Nelson est en effet l'inventeur de la notion

d'hypertexte[2]. Toutefois, Enquire ne permet de mettre en relation des documents par liens bidirectionnels qu'entre des documents appartenant à un même espace de gestion de fichiers, et souvent en pratique sur le même ordinateur. Pour aller au-delà, et mettre en relation des composantes (pas seulement documentaires) de différents environnements hétérogènes, il faut disposer d'un système de nommage universel des composantes à mettre en relation, à tisser ensemble sur la même toile. Tim Berners-Lee propose le principe d'une identification universelle : URI (*Universal Resource Identifier*). L'idée du *World Wide Web* est née. Dès octobre 1990, d'après cette idée, Tim Berners-Lee crée le langage HTML (*HyperText Markup Language*) basé sur le langage documentaire SGML (*Standard Generalized Markup Language*) existant et, en décembre 1990, il propose un premier navigateur pour voir les documents produits sur le web. Ce travail ne reçoit pas tout de suite de validation formelle par la hiérarchie du CERN et se développe lentement jusqu'à ce que Robert Cailliau rejoigne l'équipe de Tim Berners-Lee pour développer un navigateur rudimentaire qui peut fonctionner sur n'importe quel système (et pas seulement sur NEXT, la machine de développement de Tim Berners-Lee), y compris sur les vieux télétypes. Le web devient visible sur toutes les machines connectées du CERN. Tim Berners-Lee et Robert Cailliau persuadent les autorités du CERN de mettre le répertoire téléphonique du

2. Theodor H. Nelson, *Complex Information Processing : A File Structure for the Complex, the Changing and the Indeterminate*, Proceeding ACM '65, Proceedings of the 20th National Conference, ACM New York, 1965, p. 84-100.

CERN sur le web, attirant ainsi les 10 000 personnes associées au CERN.

Les années 1990 : l'emballement

Rapidement, et parce que tout est ouvert et documenté, d'autres navigateurs sont développés avec, cette fois, des possibilités d'interaction de plus en plus sophistiquées et en particulier la possibilité du clic de souris : Perry Pei Wei (1992) développe le navigateur www pour Unix (déterminant pour la communauté scientifique), puis Erwise, Midas (Unix), Samba (Mac), etc. apparaissent, sans jamais vraiment encore s'imposer. Marc Andreessen et Eric Bina du National Centre for Supercomputing Applications (NCSA) développent un nouveau navigateur en adoptant un mode coopératif de développement, bien dans l'esprit du chantier sans fin : Mosaic est disponible en février 1993. En 1994, Andreessen quitte le NCSA pour créer Netscape avec James Clark. En août 1995, Netscape entre en Bourse avec un gros succès. Cette entrée en Bourse sonne le début d'un mouvement d'intérêt très important par le secteur économique. Le web devient le support d'innombrables initiatives pour proposer des services, en particulier dans le domaine du e-commerce. Les « .com » deviennent vite la coqueluche des investisseurs. Le Nasdaq voit passer son index de 1 291 en 1996 à 4 000 en décembre 1999, avec des valeurs parfois multipliées par 100 ou 200 en une seule journée après leur introduction. En 2000, la bulle Internet éclate, entraînant des faillites en chaîne, mais aussi de sévères restructurations dans les grands groupes qui ont surinvesti dans le domaine et, pour l'essentiel, les entreprises du secteur des télécommunications.

Depuis 1994 : l'histoire en construction

De la même façon que les structures de gestion d'Internet se sont constituées de manière informelle et en dehors des structures de normalisation internationales, Tim Berners-Lee propose de monter un consortium ouvert pour gérer l'évolution du chantier sans fin du web : le World Wide Web Consortium (W3C) naît en 1994 et assure toujours la promotion de l'interopérabilité et la standardisation des technologies web.

Chaque année, depuis 1994, la conférence mondiale du web rassemble plus de 1 000 acteurs internationaux pour partager et faire avancer ce chantier.

L'histoire de l'*open source* se tisse dans la même trame que celle de l'histoire du web. Ce principe d'ouverture est à l'évidence la clé de la réussite et du formidable développement d'Internet puis du web qui en est le service principal et dominant depuis 1995.

Les terminaux du web se multiplient et se diversifient avec des opérateurs téléphoniques qui cette fois investissent massivement le domaine avec une offre systématique de l'accès web, le téléphone devenant parfois une fonction accessoire du terminal. L'effet est immédiat sur le taux de connexion des usagers du monde entier, y compris dans les pays très peu pourvus technologiquement et pour lesquels le téléphone mobile est un équipement essentiel dans leurs activités.

Les chiffres clés d'Internet sont disponibles en permanence sur le web et, en 2013, ce sont 32 % des habitants du monde qui sont aussi des internautes. Internet et le web, son principal service, deviennent des biens communs à l'échelle mondiale.

Si le web 1.0, celui des documents, a été et est toujours un immense succès, il ne correspond toutefois pas à la vision qu'était celle de Tim Berners-Lee à la création du web. Tim Berners-Lee a d'emblée tenté de persuader les développeurs de navigateurs qu'il était important que ces navigateurs ne soient pas seulement des lecteurs d'information, mais aussi des éditeurs d'information. Les notions de lecture-écriture sont indissociables pour assurer la circulation de l'information bilatérale, mettant en évidence l'importance de l'interaction dans le développement de l'environnement numérique partagé que constitue le web.

Ce n'est qu'à partir de 2003 que ces notions deviennent majeures et que Dale Dougherty propose de nommer web 2.0 cette nouvelle phase. L'information et l'activité tendent à se fondre dans le chantier sans fin du web, ce qui pose naturellement la question du statut du document numérique à l'ère du web.

Les utilisateurs s'emparent en effet du web pour y écrire autant que pour y lire. Les réseaux sociaux se développent à l'échelle mondiale (YouTube et Facebook dépassent le milliard d'utilisateurs actifs) et les écrits des utilisateurs deviennent des données pour les logiciels du web et rentrent dans la liste des *Big Data* convoitées par un marché qui y voit une activité économique potentielle croissante.

Les écritures collaboratives émergent et démontrent des propriétés de confiance et de qualité qui confirment bien les principes d'ouverture qui ont toujours présidé au succès du développement de l'Internet et du web. Par exemple, Wikipédia en français compte 1 571 551 contributeurs, dont 5 000 très actifs. Plus de 2,4 millions d'articles Wikipédia en français sont consultés chaque jour.

Progressivement, les espaces d'activités collectives se développent sur le web, s'ouvrant à des publics de plus en plus variés et offrant des fonctionnements privés, semi-privés ou publics. Ils rassemblent dans le navigateur des services complémentaires pour la gestion de l'activité collective privée ou professionnelle. Si ces services ne relèvent pas du web, ils s'y expriment et leurs usages y constituent autant d'écritures collectives et individuelles. Parmi ces espaces d'activités collectives, il faut noter l'importance des jeux comme *World of Warcraft*, qui compte 8 millions de joueurs, avec une tendance à intégrer de manière de plus en plus importante les joueurs dans la conception et le renouvellement des jeux. Toutes les activités exploitent de manière plus ou moins directe les possibilités ouvertes par un accès web sur Internet, depuis les activités culturelles (avec un énorme développement dans le domaine de la musique, de la vidéo en particulier), de services, industrielles, touristiques, agricoles, éducatives, sociales, artistiques, militantes, politiques, sécuritaires, etc.

C'est toutefois la possibilité de *trouver* l'information qui déclenche l'immense succès du web comme conteneur d'information mondial. Dès 1995, Sergueï Brin et Larry Page proposent une autre méthode que l'indexation traditionnelle (Altavista, Yahoo!, etc.) thématisée et hiérarchisée, mais s'intéressent à l'indexation inverse, en partant des références aux pages et des termes des contenus comme autant d'index potentiels. Google était né. C'est donc une spirale de réussite qui se met en place entre les fournisseurs de contenus accessibles désormais sans effort et les fournisseurs de moteurs de recherche d'information. Les métiers des uns et des autres se rejoignent.

Au-delà de l'écriture explicite (texte) par les utilisateurs, ce sont donc toutes leurs interactions qui deviennent autant d'écritures rejoignant les données utilisables par les logiciels du web. Ces écritures sont le plus souvent nommées traces[3] dans la littérature, car elles posent la question de leur observation (enregistrement des actions des utilisateurs) et celle de leur interprétation pour des actions diverses (recommandation, personnalisation, sécurité, identification, etc.). La question du statut de ces traces est posée du point de vue de la vie privée, de leur propriété, de l'éthique de leur utilisation, de leur diffusion, de leur pérennité, de leur gestion, etc.

Le web reste par essence un chantier sans fin dont les dynamiques actuelles s'articulent autour de l'introduction d'écritures possédant une sémantique formelle, c'est-à-dire permettant des calculs interprétatifs explicites plutôt que des calculs implicites encapsulés dans les logiciels traitant les données issues des écritures des utilisateurs. Cette idée d'un web sémantique – auquel seront dédiés les chapitres 4 et 8 de ce manuel – est activement développée par une communauté de chercheurs et d'entreprises innovantes.

Dans le même esprit, sur une initiative du W3C et avec l'esprit initial du web, le web des données, ou données liées (*linked data*), se propose d'écrire dans le web l'information issue de données produites par ailleurs, avec un processus éditorial semi-automatisé. Les données des institutions, les données du monde deviennent, à leur tour, des lieux de lecture-écriture en s'inscrivant dans le web.

---

3. Alain Mille, *De la trace à la connaissance à l'ère du web*, Numéro thématique Intellectica, n° 59, 2013.

Documents, données, connaissances, activités sont tissés pour créer un environnement numérique évoluant maintenant de manière contributive et ouverte. Ce caractère contributif et ouvert a été et reste manifestement la condition *sine qua non* du développement de ce chantier sans fin.

CHAPITRE 3

Histoire des humanités numériques

Michaël E. Sinatra et Marcello Vitali-Rosati

Parallèlement à l'histoire du développement d'Internet et du web, une autre histoire est fondamentale pour comprendre les enjeux de l'édition numérique : celle des humanités numériques. Il y a encore quelques décennies, on pouvait penser que les ordinateurs et les technologies numériques étaient destinés uniquement aux sciences dures, les sciences exactes dont le calcul et les mathématiques sont les principaux outils. Cette idée est manifestement fausse aujourd'hui : le numérique habite l'ensemble de nos vies et touche aussi, et surtout, à nos activités purement « humanistes », ou même « humaines ». Ce chapitre a pour ambition de retracer l'histoire du rapport complexe entre les sciences humaines et l'informatique qui a mené des premières expériences de recherche assistée par ordinateur, dans le domaine des sciences humaines (*humanities computing*), aux actuelles *digital humanities*, ou à un possible humanisme numérique.

Qu'est-ce que les humanités numériques ?

L'expression « humanités numériques » est une traduction de l'anglais *digital humanities* (DH), un domaine de recherche très vaste, caractérisé par une forte interdisciplinarité. Dans le débat actuel, on essaie de ne pas penser les humanités numériques comme une discipline et de plutôt les envisager comme une approche globale, transdisciplinaire, adoptant une attitude et un point de vue sur la recherche qui devraient impliquer l'ensemble des chercheurs en sciences humaines et sociales.

Loin d'être un simple développement technologique ayant un impact sur le processus de recherche et de visualisation des données en sciences humaines et sociales, les humanités numériques nous amènent à repenser le sens même de la recherche et, par conséquent, l'ensemble du modèle de production et de circulation du savoir à l'époque de l'édition numérique.

D'une part, les humanités numériques pourraient être définies comme l'application d'une méthode d'analyse informatique aux sciences humaines. En d'autres mots, l'approche des DH consiste à prendre en compte le fait que la puissance ne doit pas être limitée aux sciences dures, mais peut et doit aussi être employée pour des recherches en sciences humaines. D'autre part, les humanités numériques transcendent cet aspect technique et peuvent être pensées comme un regard global posé sur les changements culturels déterminés par le numérique ; en ce sens, les humanités numériques pourraient conduire à une sorte d'« humanisme numérique ». Dans ces pages, nous traiterons de l'histoire de ce développement complexe ayant porté à concevoir l'articulation de ces deux niveaux.

Un changement de paradigme

Les textes qui ont orienté le développement d'outils destinés à aider la recherche en sciences humaines et sociales proposent en même temps une réinterprétation des structures conceptuelles à travers lesquelles l'homme se rapporte au monde et, surtout, structure et organise sa connaissance. En d'autres mots, nous sommes face à un changement de paradigme dans la façon d'agencer les contenus et, par ce fait même, à une nouvelle conception du savoir et de sa circulation dans la société.

Le texte de Vannevar Bush de 1945, « *As We May Think*[1] », en est un exemple particulièrement significatif. Bush essaie de définir une nouvelle manière d'organiser notre accès aux documents, qui serait, à son avis, rendue possible par l'invention des microfilms. L'idée proposée par Bush est de construire un bureau mécanique (*Memex*) qui puisse stocker une grande quantité de documents en microfilms et les relier entre eux grâce à des dispositifs mécaniques. Il s'agit, en d'autres termes, de créer des liens entre des textes et d'autres types de documents pour pouvoir organiser de façon dynamique nos parcours de lecture et de recherche et donc notre accès aux contenus. Vingt ans plus tard, Ted Nelson[2] reprend l'idée de Bush en l'adaptant aux technologies numériques. Le *Memex* se transforme en *Complex*, un dispositif électronique ayant les fonctionnalités du bureau mécanique de Bush. C'est dans cet article que Ted Nelson utilise pour la première fois le mot et le concept

1. *The Atlantic*, 1945.
2. *Complex Information Processing: A File Structure for the Complex, the Changing and the Indeterminate*, Proceedings of the ACM 20th National Conference, ACM New York, 1965, p. 84-100.

d'hypertexte, structure qui sera une des idées utilisées par Tim Berners-Lee pour concevoir le web. On comprend que, plus qu'un pur changement technologique, c'est un véritable changement de paradigme de la structuration du savoir qui commence à se développer à partir de la fin des années 1940. Les humanités numériques sont donc une discipline chargée de réfléchir aux outils technologiques dont les sciences humaines et sociales devraient se pourvoir, mais aussi de produire l'appareil théorique pour interpréter les structures conceptuelles fondamentales qui caractérisent notre culture actuelle.

Un bref historique

Essayons de parcourir les étapes historiques fondamentales de cette approche.

Les experts s'accordent à faire remonter aux travaux du père Roberto Busa l'origine de l'approche des humanités numériques. Entre la fin des années 1940 et le début des années 1950, ce jésuite met en place avec IBM un projet pour informatiser l'index de l'œuvre de Thomas d'Aquin (l'*Index Thomisticus*). Ce projet utilise pour la première fois l'informatique et démontre la puissance de cet outil pour faire de la recherche en sciences humaines. Pour le père Busa, l'informatique ne change en rien le sens des pratiques de recherche : elle ne fait que les simplifier, les automatiser, les rendre plus rapides. Ainsi, on parle à l'époque de *literary and linguistic computing*, c'est-à-dire d'une discipline qui met les outils informatiques à disposition des sciences humaines pour augmenter la capacité d'analyser des textes grâce à la puissance de calcul des premiers ordinateurs. Il s'agit, bien

évidemment, d'outils très coûteux, réservés à la seule communauté des chercheurs.

Dans les années 1960 et 1970, des travaux similaires à ceux du père Busa apparaissent en Amérique du Nord et en Europe avec, entre autres, la création de concordances pour d'autres corpus (l'index de textes allemands médiévaux de Roy Wisbey, et la concordance des poèmes de Matthew Arnold et W.B. Yeats par Stephen Parrish), l'usage de statistiques sur des corpus numérisés et le début de l'utilisation des outils informatiques dans les questions d'identification des auteurs[3]. La majorité de ces projets sont dédiés à une amélioration du processus mécanique de recherche et de la quantification de données, mettant à profit la capacité de calcul offerte par les serveurs informatiques et les programmes d'analyse de textes. Le brassage de textes à grande échelle (une quantité loin du milliard de livres de Google Books, mais qui aurait pris à l'époque plusieurs années pour une recherche manuelle) commence aussi à transformer de manière significative l'approche de certains chercheurs, encore toutefois assez isolés, et détermine la création de bases de données et d'outils dans plusieurs pays. La popularité de ces travaux et l'intérêt grandissant pour cette approche interdisciplinaire amènent à la fondation de l'Association for Literary and Linguistic Computing en Angleterre en 1972 et à la création, en 1976, de la revue *Literary & Linguistic Computing* qui y est associée, une revue qui existe encore aujourd'hui même si les articles

3. Susan Hockey, « The History of Humanities Computing », dans *A Companion to Digital Humanities*, Susan Schreibman, Ray Siemens et John Unsworth (dir.), Éditions Blackwell, 2004, p. 3-19.

publiés vont désormais au-delà des questions initiales, plutôt circonscrites à l'analyse des fréquences de mots ou à l'assignation d'autorité textuelle. Le mouvement est bien sûr international, comme l'atteste la création de l'Association for Computers and the Humanities, en 1978, aux États-Unis, et le Consortium pour ordinateurs en sciences humaines / Consortium for Computers in the Humanities, en 1986, au Canada. Lou Burnard indique que cette composante institutionnelle, qui voit se multiplier les centres et les programmes de formation universitaire au cours des années 1970 et 1980, détermine le passage du *literary and linguistic computing* aux *humanities computing*[4]. Les pratiques de recherche informatisées dans le domaine des sciences humaines et sociales ne sont plus l'activité de chercheurs isolés mais deviennent une véritable approche interdisciplinaire, partagée par des communautés de recherche qui commencent à se structurer et à s'organiser en tant que telles.

Le changement de nom, de *humanities computing* à *digital humanities*, se profile dès la deuxième moitié des années 1990 et se concrétise en 2004 avec la publication du *Companion to Digital Humanities*, sous la direction de Susan Schreibman, Ray Siemens et John Unsworth[5]. Cet ouvrage démontre la vitalité et le caractère interdisciplinaire des humanités numériques et en retrace l'histoire. La maturité des travaux, découlant de plusieurs décennies de recherche,

4. Lou Burnard, « Du *literary and linguistic computing* aux *digital humanities* : retour sur 40 ans de relations entre sciences humaines et informatique », dans Pierre Mounier (dir.), *Read/Write 2*, OpenEdition Press, 2012, p. 45-58.

5. Blackwell Éditions, 2004.

atteste d'une intégration suffisamment poussée de l'aspect scientifique (évoqué par le terme « *Computing* ») aux sciences humaines. Cette intégration est désormais claire et il est possible de s'émanciper de la référence directe à l'outil informatique : le nom de l'approche peut lui-même changer pour indiquer cette évolution. C'est le début d'une prise en compte de la révolution que ces outils ont engendrée en changeant profondément le concept même de *humanities*, c'est-à-dire l'ensemble des disciplines humanistes. La naissance d'Internet et sa diffusion rapide au milieu des années 1990 accroissent l'impact de cette évolution. Le changement de nom marque ainsi une étape importante, car il confirme la transformation d'une approche méthodologique en une discipline à proprement parler.

En effet, l'arrivée d'Internet est un moment clé dans le développement des humanités numériques, car il n'est pas un simple outil supplémentaire pour la recherche : il devient aussi un objet de recherche et, finalement, une technique qui modifie l'ensemble de nos pratiques au-delà de la communauté savante et, plus généralement, notre façon de voir le monde. Internet a aussi un effet direct sur les questions d'édition, permettant aux créateurs et aux utilisateurs de dépasser le processus de production traditionnel (auteur-éditeur-maison d'édition), comme l'évoquent plusieurs chapitres dans ce manuel. En effet, il ne s'agit plus seulement d'une question technique, mais d'une question de structuration du savoir en général. Cela signifie que, désormais, même un livre papier ne peut plus être conçu de la même manière, tant pour l'auteur que pour le lecteur.

De la même façon, avec le changement des supports, des modalités de publication et des mécanismes de visibilité,

d'accessibilité à l'information et de circulation des contenus, c'est l'ensemble de notre rapport au savoir qui est remis en question. D'une part, il est donc nécessaire de s'interroger sur l'implémentation de nouveaux outils – de recherche, d'édition, de diffusion, d'encodage, de forage (*data mining*), de curation (*data curation*) ou encore de visualisation et de représentation des données (textuelles, sonores, visuelles, etc.) – conçus par et pour les humanités, et de mesurer l'impact de ces outils sur la transformation de la recherche. D'autre part, il est tout aussi nécessaire de mettre en place une recherche qui puisse structurer le développement d'une théorie et d'une pensée du numérique. Pour le dire autrement, les humanités numériques doivent développer une réflexion sur la façon dont les outils numériques changent la recherche en sciences humaines, mais aussi mettre en place une recherche théorique sur ce qu'est le numérique lui-même.

Vers un humanisme numérique

Si la réflexion sur les outils numériques que l'on peut mettre au service de la recherche est très développée, plus rares sont les études sur l'impact qu'a le numérique sur les catégories conceptuelles qui structurent notre système culturel. À ces études ont contribué significativement les travaux de Milad Doueihi, qui conçoit le numérique comme un événement culturel. Selon Doueihi, on peut parler d'une véritable « conversion numérique[6] », un changement culturel qui peut être comparé à une conversion religieuse, car le numé-

6. Milad Doueihi, *La grande conversion numérique*, Seuil, 2008.

rique, comme une religion, touche à l'ensemble de notre vision du monde.

L'existence de ces deux approches des humanités numériques détermine une certaine ambiguïté caractérisant ce domaine de recherche : s'agit-il de se demander comment la technologie peut aider le développement des sciences humaines et sociales ou de porter un regard théorique sur la technologie ? Le numérique est-il le sujet ou l'objet des humanités numériques ? Cette ambivalence n'est pas toujours prise en compte. Les chercheurs en humanités numériques se sentent souvent obligés de choisir entre les deux approches : d'une part une recherche qui met au centre les outils, de l'autre un regard exclusivement théorique.

Cela nous amène à considérer l'histoire plus récente des humanités numériques, soit le changement de paradigme que la technologie a apporté aux sciences humaines et sociales, au-delà de son impact pragmatique, et à réfléchir au concept d'humanisme numérique proposé par Milad Doueihi. Avec cette notion, il ne s'agit pas, selon Doueihi, d'adapter l'ancien concept d'humanisme à l'époque du numérique, ni de régler le monde des nouvelles technologies sur les valeurs de l'humanisme. L'humanisme numérique est plutôt une situation de fait : il est « le résultat d'une convergence entre notre héritage culturel complexe et une technique devenue un lieu de sociabilité sans précédent[7] ». Cette approche permet d'éviter de penser la technique comme quelque chose qui s'oppose à l'humain, allant au-delà du cliché d'un conflit entre l'homme et la machine, pour penser, au contraire, une convergence entre technique

7. Milad Doueihi, *Pour un humanisme numérique*, Seuil, 2011, p. 9.

et culture. Une convergence qui est donc un fait : le numérique est d'ores et déjà une culture, une civilisation. Cette révolution copernicienne renverse le statut du numérique et le transforme d'objet en sujet : si le numérique est une nouvelle culture, il faut le penser comme une dimension de l'humain ou, encore mieux, il faut voir de quelle manière il change le sens même de l'humain.

Dans ce sens, une thèse de Tim Bray est particulièrement éclairante : « nous ne sommes plus de simples utilisateurs, mais tout simplement des humains[8] ». En d'autres termes, nous ne sommes pas des humains qui, entre autres, utilisent les nouvelles technologies ; l'humain est aussi constitué par la présence de la technique numérique. Nous sommes donc des humains numériques.

Une fois dépassée l'opposition apparente entre humanité et technologie, il n'est plus nécessaire de choisir entre une approche axée sur les outils et une approche théorique, car l'une ne peut pas exister sans l'autre : penser le numérique signifie développer des pratiques ; concevoir des outils signifie théoriser sur le numérique.

Le défi de l'édition numérique

Cette courte histoire nous fait comprendre le rapport complexe que le développement d'outils techniques entretient avec une réflexion théorique sur notre façon d'être au monde et, en particulier, de produire et d'organiser notre savoir. Les humanités numériques nous enseignent que les choix techniques qui sont à la base de la structuration des contenus ne sont pas neutres et qu'ils témoignent et

8. *No More Users*, 2010.

promeuvent des idées et des valeurs particulières. Pratiquer l'édition numérique signifie prendre en compte ce lien étroit entre la technique et la culture. Le risque du numérique est ce qu'on pourrait appeler un « déterminisme technologique » : nos pratiques et notre façon de penser pourraient finir par être déterminées par les outils. C'est le danger entraperçu par Nicholas Carr[9], par exemple, quand il déclare que la facilité d'accès aux contenus et leur multiplicité engendrent une fragmentation de l'attention et une incapacité à suivre des argumentations complexes. La réflexion théorique devrait aider à éviter ces écueils et ce déterminisme.

Il ne s'agit donc pas seulement de comprendre ce que le numérique implique puisque nous ne pouvons pas limiter la pensée à une série de constats passifs : la pensée n'a de sens que si elle est normative, d'autant plus que l'évolution du monde numérique doit manifestement beaucoup à des événements extérieurs, impensables. L'évolution technologique est faite, surtout ces dernières années, d'une série de petites révolutions déterminées par des idées et par des pratiques qui émergent dans la communauté des usagers ; c'est pourquoi il est si difficile de prévoir le développement des nouvelles technologies et de parier sur une évolution plutôt qu'une autre. La plupart des innovations récentes ont produit des bouleversements majeurs dans le monde numérique, bouleversements imprévisibles et souvent issus d'une série de hasards. C'est le cas de Google ou de Facebook, dont le succès et l'influence sur la culture numérique ont largement dépassé les intentions premières de leurs concepteurs.

9. « Is Google Making Us Stupid ? », *The Atlantic*, 2008.

Dans ce contexte, les pratiques éditoriales assument un rôle fondamental. La structuration des contenus, leur organisation, la mise en place de dispositifs permettant leur validation et assurant leur visibilité ainsi que leur accessibilité sont les pratiques qui feront le web de demain et, par le fait même, le monde de demain. L'édition se transforme en éditorialisation : l'ensemble des pratiques d'organisation et de structuration de contenus sur le web. La différence principale entre le concept d'édition et celui d'éditorialisation est que ce dernier met l'accent sur les dispositifs technologiques qui déterminent le contexte et l'accessibilité d'un contenu, ainsi que sur la réflexion autour de ces dispositifs.

En d'autres termes, il ne s'agit pas seulement de choisir, de légitimer, de mettre en forme et de diffuser un contenu, mais il s'agit aussi de réfléchir à l'ensemble des techniques que l'on va utiliser ou créer pour le faire, ainsi qu'aux contextes de circulation produits par l'espace numérique. Si les humanités numériques s'occupent de produire des outils et de réfléchir à leur impact sur la production et la circulation du savoir, alors l'éditorialisation devient l'objet central de leur travail.

Le prochain chapitre de l'histoire des humanités numériques reste impossible à prévoir car nous en serons les acteurs : mettre en place de bonnes pratiques qui influenceront le développement culturel est sans doute le plus grand défi de l'édition numérique.

AXE THÉORIQUE

CHAPITRE 4

Pour une définition du « numérique »

Marcello Vitali-Rosati

On ne peut pas parler d'édition numérique sans approfondir le sens du mot « numérique » lui-même. L'édition numérique fait partie d'une série complexe de pratiques qui jalonnent désormais notre quotidien. Une réflexion théorique à propos de ce mot est indispensable pour pouvoir comprendre les caractéristiques structurales des nouvelles pratiques éditoriales et leur rapport avec la dimension de plus en plus numérique de l'ensemble de notre culture. Ce chapitre a l'ambition de clarifier la signification d'un mot omniprésent dans notre langage dans le but de développer un esprit critique par rapport aux caractéristiques spécifiquement « numériques » des modèles actuels de production et de circulation des contenus.

Questions de mots

Le mot « numérique » est de plus en plus présent dans notre vocabulaire. Il est en train de devenir un mot passe-partout qui sert à définir un ensemble de pratiques qui caractérisent notre quotidien et dont nous avons peut-être encore du mal à saisir la spécificité. Mais qu'est-ce que le numérique précisément ? Que dit ce mot à propos de nos usages ? De nos vies ? Au fil des années, plusieurs expressions différentes ont été utilisées pour parler de l'ensemble des pratiques et des possibilités qui ont émergé grâce au développement des technologies. On a souvent parlé de « nouvelles technologies » – parfois en précisant : « nouvelles technologies de l'information et de la communication » – ou de « nouveaux médias », ou encore d'« environnements virtuels » ou plus simplement d'informatique ou d'électronique.

Chacune de ces expressions privilégie et met en valeur un aspect particulier de ces expériences et pratiques et l'on peut dire que plusieurs de ces formulations, après avoir eu leur moment de gloire, sont devenues plutôt désuètes. C'est le cas de « nouvelles technologies de l'information et de la communication » ou de « nouveaux médias » : l'adjectif « nouveau » commence à être abandonné, car ces technologies ne sont plus si nouvelles que cela. Par ailleurs, cette expression renvoie à une approche particulière : celle des sciences de la communication, justement, qui a tendance à analyser les pratiques numériques essentiellement comme des pratiques de communication et d'information. Or, s'il est vrai que nous communiquons et que nous nous informons aujourd'hui surtout avec l'ordinateur, il serait réducteur de dire que le numérique n'est que cela.

L'adjectif « virtuel » aussi a fait son temps. Si l'on regarde les graphiques de *Ngram Viewer*, on peut facilement le constater : le mot commence à avoir beaucoup de succès dans la deuxième moitié des années 1980, arrive à son apogée en 2003, et son emploi commence ensuite à diminuer. En anglais, cette tendance est encore plus évidente. Le mot met l'accent sur le fait que les technologies informatiques donnent la possibilité de développer un véritable monde parallèle. Mais ce monde est apparemment – et dans l'acception la plus banale du mot « virtuel » – opposé au monde réel. Nos usages d'aujourd'hui nous obligent de constater qu'il n'y a rien d'irréel dans l'environnement numérique. C'est probablement pourquoi l'on délaisse le mot « virtuel ».

Pourquoi alors parler de numérique ? Quelle est la signification exacte de ce mot ?

Numérique et analogique

Le mot « numérique » est initialement utilisé pour caractériser le mode d'enregistrement de sons, d'images ou de vidéos en opposition à l'analogique. L'exemple des sons peut nous aider à mieux comprendre le sens de cette notion – en particulier en relation avec l'apparition du *compact disc* (CD) dans les années 1980 et avec la progressive substitution du vinyle. Le cas du CD est également significatif à cause de l'impact qu'il a eu dans l'imaginaire collectif par rapport au discours sur la qualité des enregistrements numériques et analogiques.

L'enregistrement analogique se base sur une reproduction du son de façon analogue à la réalité, à savoir en reproduisant sur un support – par exemple un vinyle – la continuité de

l'onde sonore. Concrètement, la pointe du dispositif d'enregistrement est ébranlée par le son et reproduit un mouvement analogue à celui du son. La courbe qui en ressort est continue et représente fidèlement le mouvement du son dans les moindres détails et dans la continuité du temps. Entre chaque point de cette courbe, il y a des points à l'infini – comme dans le cas d'une ligne droite continue : la courbe est donc « dense », dans le sens mathématique du mot, c'est-à-dire qu'elle ne comporte aucun saut.

L'enregistrement analogique garantit une fidélité parfaite au son d'origine, justement grâce à cette analogie et à cette continuité : en principe, l'enregistrement analogique devrait donc donner lieu à la meilleure qualité possible. Mais l'analogique pose un problème fondamental : celui de la reproduction. Chaque reproduction – à cause de la complexité de l'enregistrement, de sa densité, de sa continuité – comporte une perte de qualité. Le moindre grain de poussière sur un vinyle implique une déformation du son. Et, bien évidemment, chaque copie de l'enregistrement donne lieu à une perte de qualité : on copie analogiquement, ce qui implique qu'à chaque copie on est un peu plus loin du son original. La copie de la copie est de qualité inférieure et ainsi de suite. On pouvait le constater lorsque l'on copiait sur des cassettes des chansons depuis un vinyle. À force de copier, l'enregistrement devenait inécoutable. En somme, la complexité de l'analogique détermine une difficulté dans sa transmission et dans sa reproduction.

Le principe du numérique est de discrétiser le continu du son – ou de l'image ou de n'importe quelle autre information. Cette discrétisation est ce que l'on appelle « échantillonnage ». Concrètement, on prend le continu de l'onde

sonore et on choisit des échantillons, à savoir on ne considère pas l'ensemble du son, mais seulement les changements qui se produisent à des intervalles déterminés. Plus court est l'intervalle choisi, plus précis sera l'échantillonnage, et plus haute sera la qualité du son numérisé. Le son que l'on obtient de cette manière est essentiellement de qualité inférieure à l'analogique, car il ne rend pas compte de la continuité du son d'origine, mais seulement d'un nombre restreint – bien qu'élevé – d'échantillons. Mais le processus de discrétisation permet une simplification de l'enregistrement qui est réduit à une série de chiffres entiers et plus précisément de 0 et de 1. Cette simplification permet une meilleure gestion des reproductions. Pratiquement, il n'y a aucune différence entre différentes reproductions d'un enregistrement numérique : une copie n'a rien de différent par rapport au premier enregistrement, car aucune information ne sera perdue. En fait, on peut affirmer qu'il n'y a pas de copies, car il n'y a absolument aucune différence entre le premier enregistrement et sa reproduction. Dans chaque processus de copie analogique, même s'il est réalisé de façon mécanique, il y a une perte de données, donc chaque copie est un objet séparé. On pourra toujours identifier et distinguer la première de la deuxième ou de la troisième copie et ainsi de suite. Dans le cas d'un enregistrement numérique, cette distinction n'est pas possible. Une cassette est une copie d'un vinyle ; un CD n'est pas une copie du CD original, car il est absolument indistinguable de celui-ci.

Cela implique que, même si, lors de l'échantillonnage, il y a une perte de qualité obligatoire par rapport à l'original – car on transforme sa continuité en une série discrète

d'échantillons –, lors de la reproduction, la qualité du numérique restera la même, alors que celle de l'analogique diminuera. On comprend ainsi le discours commercial mis en avant comme argument de vente du CD : on parlait de sa meilleure qualité. Or il ne s'agissait pas d'une meilleure qualité – car le vinyle, étant analogique, était plus fidèle au son d'origine –, mais d'une meilleure gestion de la reproduction, qui permettait la transmission du son sans perte d'informations.

Le même discours vaut pour n'importe quel type d'information numérisée, qu'il s'agisse d'images, de vidéos ou de textes.

Internet et le web

Voilà expliqué le sens strict et l'emploi premier du mot « numérique ». Ce processus d'échantillonnage et de discrétisation est à la base de toutes les technologies électroniques qui fonctionnent à partir de chiffres discrets en base 2, à savoir, à partir d'une série de 0 et de 1. Concrètement, ces deux chiffres sont représentés par un circuit électrique où passe le courant (le 1) ou bien où le courant ne passe pas (le 0).

Or, si cette explication peut rendre compte du sens original du mot « numérique », elle n'est pas suffisante pour comprendre la généralisation de son emploi et surtout sa grande fortune dans les dernières années, où l'on commence à parler d'« environnements numériques », de « natifs numériques », d'« humanités numériques » et même de « culture numérique ».

On peut raisonnablement affirmer que ce développement de l'emploi du mot et sa valeur sociale et culturelle ont été déterminés avant tout par la naissance et la diffusion

d'Internet et, encore plus précisément, du web, soit à partir des années 1990. Le web, plus que la simple présence des ordinateurs, a déterminé un changement majeur de nos pratiques et de notre rapport au monde, car il a engendré de nouveaux modèles de production, de diffusion et de réception du savoir en général.

À la suite de l'omniprésence du web dans nos vies, le numérique est partout. Il y a encore quelques décennies, on pouvait considérer les technologies informatiques comme des outils puissants et aux fonctions multiples capables d'aider les hommes dans plusieurs champs de la production industrielle et culturelle. Aujourd'hui, cette définition serait du moins réductrice, sinon complètement fausse : le numérique est l'espace dans lequel nous vivons. Il ne s'agit plus d'outils au service des pratiques anciennes, mais d'un environnement dans lequel nous sommes plongés, qui détermine et façonne notre monde et notre culture.

Nous sommes obligés de prendre en compte le fait que l'on ne communique pas seulement sur le web : on organise sa journée, on achète des produits, on gère ses comptes en banque, on met en place des manifestations contre le gouvernement, on s'informe, on joue, on éprouve des émotions.

Voilà pourquoi le numérique n'est pas seulement une technique de reproduction qui s'oppose à l'analogique, mais il devient une véritable culture, avec des enjeux sociaux, politiques et éthiques fondamentaux et qu'il est urgent d'analyser et de prendre en compte.

Une culture numérique ?

Essayons d'approfondir le sens de cette expression : « culture numérique ». Que veut-on dire exactement par là ? Quel est

le sens de revendiquer un aspect proprement « culturel » du numérique ?

Comme nous venons de le suggérer, il s'agit de trouver une expression capable d'exprimer le fait que le numérique n'est pas qu'un ensemble d'outils : il n'est pas seulement un ensemble de dispositifs techniques qui permettent de mieux faire ce que nous faisions avant. Il ne peut pas être considéré comme une voiture qui nous permet de faire plus rapidement la même route que nous étions habitués à faire à pied.

Le numérique modifie nos pratiques et leur sens. Essayons de mieux comprendre cette affirmation avec un exemple simple : l'emploi de Twitter lors d'une conférence, d'un séminaire ou d'un cours universitaire. Cet outil change profondément la façon de participer à l'événement en question. Non seulement parce qu'il permet à des personnes qui ne se trouvent pas dans la salle d'être informées de ce qui y est dit et, éventuellement, de s'exprimer à ce sujet, mais surtout parce que cela produit une économie de l'attention différente et une façon différente de comprendre et de réfléchir aux contenus de la conférence, du séminaire ou du cours. Pendant que l'orateur parle, quelqu'un du public réagit à ce qu'il dit. Cette réaction est lue par d'autres personnes – présentes ou non –, ce qui crée souvent plusieurs couches de débat avec des niveaux différents d'approfondissement. Quelqu'un suit ce que dit l'orateur, quelqu'un est en train d'approfondir ce qu'il vient de dire – par exemple en cherchant des références sur Internet ou en demandant des précisions à un autre participant qui en sait davantage. Twitter change, en somme, la forme et le contenu du débat, mais aussi la forme de l'intelligence elle-même. On ne com-

prend plus les mêmes choses de la même manière ; notre rapport au monde change profondément. L'outil produit les pratiques et produit aussi le sens de ces pratiques, il modifie notre façon d'être au monde mais aussi notre « nature », car il change notre façon de comprendre, notre façon de gérer l'attention, notre façon de penser, notre perception du temps, de l'ennui et ainsi de suite.

Cela vaut pour Twitter, mais évidemment aussi pour Facebook – et l'emploi qu'en font, par exemple, les étudiants lors d'un cours –, Wikipédia ou le courriel… Penser et créer un outil signifie donc déterminer des pratiques et, par le fait même, changer notre façon d'habiter le monde. Ce constat nous oblige à mettre en question une conception naïve de la nature de l'homme. L'homme n'a pas une nature indépendante des outils dont il se sert. La nature de l'homme est – comme dans le mythe de Prométhée – dans ses outils et se transforme avec ceux-ci. Un homme numérique n'est pas simplement un homme qui se sert d'outils numériques, mais un homme différent, qui fonctionne différemment, qui a un rapport différent avec ce qui l'entoure : l'espace, le temps, la mémoire, la connaissance…

Mais il faut aller encore plus loin si l'on veut comprendre la valeur culturelle du numérique. On peut constater que ce n'est pas qu'en présence des dispositifs techniques ou technologiques que le rapport au monde change. Essayons d'expliquer cette affirmation avec un autre exemple lié à un dispositif numérique d'emploi courant : le GPS. Le fait d'avoir un GPS modifie notre rapport à l'espace. Nous percevons l'espace différemment – par exemple, il nous semble beaucoup plus rassurant, car nous savons toujours où nous sommes et ne pouvons pas nous perdre. C'est l'outil

qui façonne et agence notre rapport à l'espace et nos pratiques, ainsi que notre vision de l'espace, notre façon de le concevoir.

Or, faisons une expérience mentale (ou réelle) : éteignons le GPS lors d'un voyage. Même sans GPS, nous continuons à percevoir l'espace de la même manière. Il y a dix ans, nous aurions prêté une attention différente à la route, car la possibilité de nous égarer était toujours présente, comme une peur, une angoisse. Mais dès qu'il y a un GPS, même s'il est éteint, ce rapport change. L'espace a changé, même quand l'outil n'est plus là. Et nos valeurs ont changé, nos priorités, toutes nos structures mentales. La transformation nous a investis de façon totale.

Quelques caractéristiques du numérique

Nous avons établi la valeur culturelle du numérique et sa façon de modifier notre perception du monde. Maintenant, essayons d'identifier quelles sont ses caractéristiques et, plus précisément, quels aspects particuliers distinguent un objet numérique d'un objet non numérique.

On a souvent pu associer le numérique à l'immatérialité. L'environnement numérique – et il est clair que l'on pense ici en particulier au web – serait caractérisé par un espace immatériel qui s'opposerait à l'espace matériel non numérique. Or il est de plus en plus évident que cette affirmation est fausse. L'espace du web est, comme tout espace, un ensemble structuré de relations entre des objets. Les pages du web, par exemple, sont structurées et hiérarchisées à partir des relations qu'elles entretiennent entre elles. Ces relations sont bien définies et tout à fait concrètes. Une page

sera plus ou moins proche ou loin par rapport à une autre – selon les liens qu'il faut parcourir pour arriver de l'une à l'autre ou encore selon la place que les deux occupent dans l'indexation d'un moteur de recherche.

L'ensemble de ces relations structure l'espace numérique et ces relations sont elles-mêmes écrites et enregistrées dans les disques durs des différents acteurs du web : les fournisseurs d'accès, les moteurs de recherche, les différentes plateformes de services, etc. Rien de plus matériel.

Par ailleurs, ces objets existent dans une infrastructure très coûteuse et très matérielle – au sens qu'elle demande une grosse quantité de matériel, justement –, faite de serveurs, de câbles et même de pompes à eau pour refroidir les circuits. Récemment, la publication de photos des *data centers* de Google a fait prendre conscience aux usagers de cette réalité. On ne peut pas dire que l'immatérialité soit une caractéristique du numérique.

Cette sensation est peut-être produite par la facilité de copier les objets numériques dont on parlait plus tôt. Cette facilité peut nous induire en erreur en nous faisant croire que le numérique est immatériel. Mais elle est en réalité déterminée par une caractéristique qui est, en effet, la caractéristique principale du numérique : sa multiplicité.

Je m'explique : un objet numérique n'a pas besoin d'être copié. Quand nous envoyons un fichier par courriel, nous n'avons pas besoin de le copier, nous l'envoyons et le gardons en même temps. Et, entre les deux versions du fichier – qui sont en effet deux enregistrements –, il n'y a aucune différence. Le fichier n'est pas copié, il est nativement multiple.

Cette multiplicité se manifeste aussi dans la convertibilité des objets numériques : un texte peut être, par exemple,

converti automatiquement en son – avec un lecteur automatique – ou en image. Le même texte peut être visualisé de mille façons différentes – différentes polices, tailles, mises en page.

La multiplicité qui caractérise les objets numériques est déterminée par deux causes que l'on pourrait appeler la « discrétisation » et la « médiation ».

La discrétisation est le processus d'échantillonnage qui permet de transformer le continu du réel en une série de chiffres. Cette caractéristique du numérique est à la base de la facilité de gestion des objets numériques et de leur transformabilité.

La médiation est le processus d'interprétation nécessaire pour tout objet numérique. Il s'agit d'interpréter la série de chiffres en base 2 pour la comprendre en tant que code et d'interpréter ensuite ce code pour le rendre accessible et compréhensible pour l'usager. Je m'explique avec l'exemple d'une page web. À l'origine, elle est tout simplement une série de 0 et de 1. Cette série est interprétée par l'ordinateur et traduite – à travers un standard d'encodage – en un texte HTML. Ce texte est ensuite interprété par le navigateur qui le transforme en une page, avec ses caractéristiques graphiques, ses images, ses couleurs, ses polices, etc.

Ce processus de médiation permet évidemment des interprétations différentes. La même série de 0 et de 1 peut être interprétée de multiples manières et le même code HTML peut être affiché de multiples façons.

Voilà schématiquement expliquée la raison de la multiplicité du numérique. Or cette caractéristique fondamentale détermine ses traits distinctifs : sa facilité de circulation,

son ouverture, le fait qu'il soit facilement modifiable, réutilisable, qu'il permette des objets multimédias, etc.

À partir de cette réflexion sur le sens du numérique, il faudra reconsidérer l'ensemble de nos pratiques, en particulier dans le domaine de l'édition. La multiplicité caractéristique des objets numériques bouleverse notre rapport aux contenus et aux documents, des dynamiques de leur circulation selon les pays à la possibilité de modification et de copie, en passant par les lois sur les droits d'auteur. Cette multiplicité fait que les contenus ne sont plus assujettis à des temps de transmission ou à des coûts de copie mais deviennent présents partout en même temps. L'ubiquité des objets numériques s'associe à une grande facilité de gestion de ces derniers. Tout est facilement modifiable, réexploitable, transformable. Cela met évidemment profondément en crise les modèles traditionnels de gestion de contenus qui ne sont plus applicables au domaine du numérique. L'ensemble des pratiques liées à la production et à la diffusion du savoir doit être remis en question.

CHAPITRE 5

Les enjeux du web sémantique

Yannick Maignien

Le web change et, avec lui, changent les enjeux de l'éditorialisation. Ainsi, nous sommes passés du web statique des origines (1.0) au web participatif (2.0), puis au web sémantique (3.0) qui permet aux machines de comprendre la signification des données et de mieux les exploiter. Nous considérons que le passage au web sémantique est un enjeu majeur. Dans ce domaine, ce sont nos choix qui détermineront la structuration des connaissances dans le futur. Ce passage n'est pas neutre et comporte une série de questionnements politiques, philosophiques, économiques, sociaux et techniques. C'est pourquoi plusieurs chapitres de ce livre sont dédiés à cette question. L'objectif de celui-ci est d'établir une présentation théorique de cette problématique.

Le www, une toile de documents

Dès 1989, Tim Berners-Lee souhaite régler un problème de travail collaboratif sur des documents distribués au sein du CERN (Organisation européenne pour la recherche nucléaire). Comment accéder à l'information de milliers de documents hétérogènes, dans des formats divers, sur des systèmes d'exploitation différents, et faisant diversement référence les uns aux autres ? Autrement dit, comment appliquer à une communauté de travail (les membres du CERN) dispersée en de nombreux lieux du monde une technique d'hypertexte déjà existante ? La notion d'hypertexte était en effet déjà mise en œuvre, suite aux travaux de Ted Nelson (l'idée d'un univers de documents liés, le *Docuverse*), par exemple avec des logiciels comme *Hypercard* sur les Macintosh. Mais manquait alors l'espace commun où naviguer entre ces documents distants, organisé selon une architecture « client/serveurs ». Pour cela il fallait mettre en place un protocole commun, HTML (*HyperText Markup Language*) protocole issu lui-même des langages de structuration de documents (SGML). Par ailleurs, il fallait un protocole standard de transport de l'information sur Internet : le HTTP *(Hypertext Transfert Protocol*). Telles sont, avec l'adresse URL, les conditions d'interopérabilité.

Le modèle documentaire est alors clairement celui d'une bibliothèque distribuée, sans murs, autrement appelée bibliothèque virtuelle ou numérique, mais où l'information n'est pas classée de façon hiérarchisée, dans un arbre, mais distribuée en graphe (il existe autant de relations non hiérarchisées que de liaisons possibles entre les documents ou des parties de ces documents). Mais, dès cette époque, il

était clair pour Tim Berners-Lee que si l'on savait trouver une solution à l'échelle du CERN, on trouverait du même coup une solution à l'échelle d'Internet tout entier, pour n'importe quel document.

Il faut noter que, ensuite, HTML ne cessera d'évoluer en fonction des besoins des utilisateurs et des éditeurs de contenus, avec des logiciels de gestion de contenus (*Content Management System* - CMS) de plus en plus performants. Issu de SGML, HTML sera perfectionné à partir de 1995 en parallèle aux possibilités de XML (*eXtensible Markup Language*), un langage de balisage adaptant et simplifiant SGML pour le web, et assurant donc une plus grande interopérabilité.

Le web des documents sera complété dès 1993 par les outils de navigation (*browser*), afin d'exploiter l'ensemble de ces possibilités éditoriales, ainsi que des moteurs de recherche, afin d'indexer l'ensemble des contenus du web pour identifier et localiser les résultats des requêtes documentaires.

Dans cette architecture informatique, il importe de garder à l'esprit que ce qui est exploité au final par l'utilisateur humain est l'ensemble des possibilités (virtuelles) que d'autres humains ont intégrées au moment de l'édition numérique des documents (sites web, identifiés par des URL). Entre les deux, le système technique d'Internet et du web, universel et neutre, est indifférent aux contenus. Les « machines » informatiques ne connaissent en effet que les langages ou balises de description des structures physiques et logiques des documents, dont *href*, le lien hypertexte.

Sur cette base technique (la toile universelle), l'essor de l'échange et du partage de documents à l'échelle mondiale

représente une fantastique possibilité de développement des connaissances, démultipliant à la lecture les capacités d'écriture. En effet, toute information portée par ces documents s'affranchit des contraintes d'espace et de temps, de communication et de mémoire. On conçoit qu'une telle expansion et dynamique de liberté de l'information bouleverse l'ensemble des pratiques humaines localisées ou restreintes où le document écrit (mais aussi tout enregistrement sonore ou visuel, multimédia) était cantonné. Par exemple, l'affranchissement de l'information économique numérique devient le vecteur de la mondialisation des échanges, dérégulant de nombreux domaines protégés ou contrôlés.

Vers le web sémantique

Dès le début, mais plus concrètement vers la fin des années 1990, Tim Berners-Lee précise ce que pourrait être une évolution « sémantique » du web.

En premier lieu, une évolution du web des documents, des sites web, donnait accès, via des formulaires (et des applications très spécifiques), à des bases de données (dont les structures sont presque chaque fois différentes). Ces résultats de recherches, lisibles pour des humains, sont « illisibles » pour les robots de requête. On parle alors de « web profond », dont les contenus restent opaques pour les machines interprétant HTML.

D'autre part, si vous recherchez par exemple sur le web (de documents) l'indication d'un cardiologue ouvert le dimanche dans le département du Rhône, vous risquez de n'obtenir aucune réponse – ce qu'on appelle « silence » – à cette question précise ou, à l'inverse, beaucoup trop de

réponses – « bruit » : toutes les réponses indexant cardiologue / jours ouvrables / Rhône. Bien sûr, l'utilisateur humain peut successivement chercher dans l'annuaire des cardiologues, l'agenda de ces spécialistes et la liste dans la carte du Rhône, autant de bases de données différentes (du moins s'il a accès à ces différentes bases de données).

L'idée du web sémantique consiste, selon Bruno Bachimont[1], à « pouvoir déléguer à la machine une partie de l'interprétation des ressources du web ». Il s'agit de créer les conditions pour que ces informations contenues dans trois « silos » de données différents (annuaire de cardiologues, agenda, données géographiques du Rhône) puissent être lues automatiquement par les machines (serveurs et clients-navigateurs). La « machine » doit pouvoir inférer logiquement la ou les réponses possibles à partir de ces trois types de ressources pourtant hétérogènes.

L'importance des métadonnées

Par « sémantique », il ne s'agit donc pas d'envisager que la machine « comprenne » au sens humain le contenu de l'information de chacune de ces bases. Par contre, ces informations (données) peuvent faire l'objet d'un langage structuré décrivant ces données, et suffisamment standardisé pour être partageable par des machines. Ce langage est appelé « métadonnées » (des données décrivant des données, ou *metadata*). Dans l'univers des documents, de telles « métadonnées » existent depuis longtemps. Ainsi, dans les bibliothèques, les notices bibliographiques de documents contiennent des

1. « Enjeux et technologies : des données au sens », *Documentaliste-Sciences de l'Information*, vol. 48, n° 4, 2011, p. 24-41.

informations structurées décrivant par exemple un livre – Auteur = Victor Hugo, Titre = *Les Misérables*, etc. Auteur, Titre, Éditeur, Date, etc. sont autant de métadonnées standardisées dans un format d'échange permettant à tous les systèmes informatiques des bibliothèques de partager et de traiter cette information (dédoublonner, ou distinguer des éditions par exemple, gérer des prêts, des acquisitions, etc.). Les conditions requises supposent donc une standardisation des métadonnées, par exemple dans le format MARC (*MAchine-Readable Cataloging*, le standard de description bibliographique développé par la Library of Congress dans les années 1960).

On notera que les métadonnées sont désormais souvent produites en même temps que les données, par exemple pour les photos (format, date, géolocalisation, reconnaissance de formes, de couleurs, etc.), avec les appareils photos numériques, pour lesquelles l'on peut parler de métadonnées « embarquées ».

Peut-il s'agir d'étendre à tout le web un système unique de format de métadonnées, non seulement pour les livres, mais pour toutes les informations circulant dans le monde entier, de site en site ? Bien évidemment, non. La toile n'est pas une immense bibliothèque, dont chaque information singulière pourrait être structurée uniformément. Chaque secteur d'activité, de service, de négoce, chaque base de données développe des systèmes hétérogènes de métadonnées. Par exemple pour les formats d'image, les fichiers son, les données géographiques, les tableaux économiques, etc.

RDF, un modèle de données

Par contre, cette question, indépendamment d'Internet, s'était déjà posée en intelligence artificielle. Comment exprimer de façon informatique des données hétérogènes afin que les machines puissent procéder à des inférences logiques, comme pour les systèmes experts ? L'idée de Tim Berners-Lee, avec le web sémantique, est d'introduire un langage de format de métadonnées suffisamment simple et générique pour lier toutes ressources présentes sur le réseau, mais également toutes données de bases de données relationnelles (sous réserve de l'ouverture légale d'accès). Cette unité fondamentale du système se nomme RDF (*Resource Description Framework*), format d'expression des données qui sera standardisé par l'instance du W3C dès 1994.

RDF repose sur la structure logique de prédicat ‹sujet, prédicat, objet›, ou triplet, une sorte de phrase de grammaire simple : sujet, verbe, complément. Par exemple : Victor Hugo est l'auteur des *Misérables*, où le Sujet = « Victor Hugo » ; le Prédicat = « est auteur de » et l'Objet = « *Les Misérables* ». Ces triplets sont l'unité nécessaire et suffisante pour lier cette information dans des graphes de données de dimension énorme. La liste de la totalité des œuvres de Victor Hugo peut être simplement écrite. Si par exemple un autre triplet est du type « *Les Misérables* est une pièce de théâtre », ou « *Les Misérables* est jouée à Broadway », ou encore « Gavroche est un personnage des Misérables », « Jean Gabin a joué Jean Valjean », « Rodin a sculpté Victor Hugo », etc., toutes ces informations, dispersées sur la toile, peuvent être liées à « Victor Hugo », résultat d'une inférence logique automatique. Ces données liées dans un domaine particulier se nomment « jeu de données » (*dataset*).

La condition première pour que le web des données fonctionne est de doter chaque donnée d'un identifiant unique, ou URI (*Uniform Resource Identifier*), sur le modèle des URL (dans notre exemple : « Victor Hugo », « *Les Misérables* », « être auteur de », etc. doivent chacun avoir leur propre URI). Ces « adresses uniques des données » transmises par HTTP permettent de constituer les graphes de données sans ambiguïtés, chaque nœud étant unique.

Des domaines spécifiques d'information peuvent bien sûr standardiser ces systèmes de métadonnées et expliciter la sémantique associée dans des schémas (*RDF Schema*). *RDF Schema*, ou RDFS, est un langage extensible de représentation des connaissances permettant de déclarer une ressource comme classe d'autres ressources, par exemple les catégories documentaires bibliographiques citées plus haut, les notions de collection, d'œuvre, etc.

Dans le domaine documentaire en général, un format minimal de quinze métadonnées, le Dublin Core, permet d'exprimer de façon universelle ces entités documentaires sur la toile. Anticipant le mouvement du web sémantique, le Dublin Core avait déjà permis de « moissonner » les documents à partir de l'identité de format des métadonnées (OAI PMH). Une démarche du même type est mise en œuvre pour les archives (EAD), pour les œuvres de musées (CIDOC CRM), pour les événements culturels, etc., mais très vite, l'hétérogénéité des formats cloisonne les possibilités d'accès ou de service web.

Ne serait-ce que pour ces secteurs culturels, le web sémantique va permettre une convergence des données issues des archives, des musées, des bibliothèques (voir le Centre Pompidou virtuel, Europeana, Canadiana, data.bnf.fr, HdA

Lab, Érudit, etc.), mais aussi se croiser avec des données sur les besoins ou comportements des visiteurs, l'agenda, l'origine, l'organisation des expositions, le tourisme, etc.

Ce modèle de données, RDF, est lui-même exprimable dans une syntaxe XML, faisant penser à une continuité dans la description des documents. Cette homogénéité du web en ce qui concerne la syntaxe est importante, mais c'est bien avec RDF, le format des données et sa structure en graphe, qu'est assurée l'interopérabilité. RDF peut d'ailleurs se traduire par une syntaxe XML, mais aussi par d'autres syntaxes : Turtle, N-triples ou N3.

À ce stade, il importe de prendre conscience que le web des données a une dimension (quantitative et qualitative) qui répond à une rupture par rapport au seul domaine documentaire ou culturel (et à ses catalogues ou systèmes de métadonnées bien structurées). Il assure une interopérabilité inégalée jusqu'à maintenant. Le web des données correspond au besoin de traiter toute donnée liée (*linked data*), tout tableau à double entrée, toute corrélation ou fonction numérique. Cela correspond également au besoin étendu de recourir à des traitements automatiques de données « lisibles par les machines » pour tout système (dynamique) produisant des données, capteurs, enregistreurs de flux, systèmes d'objets, marchandises, processus de production, statistiques financières, etc.

Les secteurs scientifiques, et en particulier celui de la biologie avec ses immenses bases de génomique, sont des domaines privilégiés de développement du web sémantique. « Dans ce secteur, il existe un besoin urgent de croiser un très grand nombre d'informations pour trouver de

nouveaux médicaments », indique Tim Berners-Lee[2], soulignant l'importance d'une « intelligence collective » en croisant les données sur l'ensemble de la toile.

Un des exemples de développement du web des données à partir du web des documents est Wikipédia et son expression en RDF DBPedia. On peut aussi citer la base de données géographiques Geonames. La production distribuée et collaborative de connaissances par un Wiki (Wikipédia) est particulièrement propice au passage de ces informations encyclopédiques vers RDF, le Wiki étant un système éditorial structuré permettant la modification des données par des utilisateurs identifiés. L'ensemble des données structurées de Wikipédia, par exemple les données géographiques ou temporelles (dates) ou les entités nommées, sont automatiquement extraites et exprimées en un gigantesque silo de données liées (*triple store*) appelé DBPedia, et ce, pour chaque communauté linguistique de Wikipédia. Aussi, toute application du web des données peut se relier à ce silo et utiliser tout ou partie des données DBPedia. L'ensemble de ces silos correspond à des intérêts thématiques particuliers, mais peuvent interopérer.

Moteurs de recherche et requêtes SPARQL

Les moteurs de recherche, comme Google, véritables entrées sur le web par leur privilège de hiérarchiser les résultats de requêtes en fonction d'un algorithme de « réputation » (le nombre de liens), *page ranking*, fonctionnent essentiellement en indexant sans cesse le contenu « linguistique » de la toile. Mais cette logique du document tend de plus en

2. « Le web va changer de dimension », *La Recherche*, n° 413, 2007, p. 34.

plus à faire place à une logique de pertinence des données pour satisfaire des requêtes. Google lui-même bâtit une intense politique de développement de microformats, intégrant des données RDF au sein de pages HTML (RDFa), ainsi que des outils de liens de connaissance, comme *Knowledge Graph*.

Mais le web des données a son propre langage de requête, SPARQL (pour *SPARQL Protocol and RDF Query Language*). C'est une sorte d'équivalent de SQL pour les bases de données (comme MySQL), permettant de rechercher les données dans différents *triple stores*, participant au mouvement de décloisonnement des données et des applications.

Si nous reprenons notre exemple, avec le web des données, la requête SPARQL ira interroger les différents silos de données pour inférer les seules possibilités des agendas des cardiologues du Rhône effectivement ouverts le dimanche. Il faut bien sûr que ces données existent et qu'elles soient des ressources exprimées en RDF.

Ontologies et inférences

Si le web des données est dans la suite du web des documents, il se différencie fortement par sa capacité de raisonnement ou d'inférence. Le modèle de données RDF, unité logique, peut être appelé unité minimale de connaissance dans la mesure où, selon Tim Berners-Lee[3], « cette description caractérise la donnée en la reliant à une catégorie. Par exemple, la donnée "pêche" sera reliée soit à la catégorie "fruit" soit à la catégorie "poisson", selon l'objet sur lequel elle porte. Ou une date de naissance sera reliée à la catégorie

3. *Ibid.*

"date". [...] RDF *Resource Data Framework* est aux données ce que HTML est aux documents. RDF permet de relier une donnée à une catégorie. »

Cette capacité logique de connaissance est elle-même formalisée pour le web sémantique. Les taxonomies ou thesaurus sont exprimés en SKOS ou OWL, appelés langages d'ontologies, standardisés par le W3C.

Ainsi, un modèle d'ontologie pour la description des personnes, le format FOAF (*Friend of a Friend*) ou ontologie des personnes, est utilisé dans l'organisation notamment des réseaux sociaux. À cet égard, on voit que le web 2.0, ou web collaboratif, n'est nullement disjoint, mais au contraire un élément dynamique de la structuration des données par le web sémantique.

Cette couche logique, elle-même basée sur des développements des logiques du premier ordre (*description logic*), permet de raisonner sur les données en liaison avec des références « métier » ou des catégories déjà organisées, propres à un domaine de connaissances, par exemple les relations logiques entre les entités biologiques (notion d'espèce, de classe, de genre, etc.).

L'aspect heuristique (de capacité de découverte) du web des données est sans doute la justification essentielle de cette délégation de pouvoir aux machines, pour le traitement de masses énormes de données, disjoignant l'origine éditoriale de ces données de leur réutilisation : « Beaucoup de grandes découvertes sont nées ainsi de la réutilisation d'informations qui avaient été collectées dans un tout autre but », précise ainsi Tim Berners-Lee[4].

4. *Ibid.*

Pour simplifier, on peut dire que l'on passe d'un web des documents à un web des données (ou *linked data*) avec RDF, mais, grâce aux raisonnements et inférences possibles sur ces données, on passe du *linked data* au web sémantique, ou web des connaissances.

Automatisation accrue ou nouvelles dimensions humaines du web ?

Il est important d'insister sur le double aspect technique et humain du web sémantique. Tim Berners-Lee pose comme principe que si, par le passé, on partageait des documents, dans l'avenir, on partagera des données.

Dans une première approche, nous l'avons souligné fortement, cette avancée technologique se marque par une délégation accrue des traitements des données aux machines, remplaçant ainsi de fastidieuses recherches humaines d'informations, de plus en plus imprécises, du fait de l'hétérogénéité des formats et du cloisonnement des applications.

Cette montée en puissance des automates de traitement des données est un fait. Mais, paradoxalement, l'avancée du web sémantique fait naître de façon croissante des préoccupations humaines collatérales.

La question de la qualité des données s'avère stratégique. Certes, plus il y a de données brutes disponibles, plus le web des données peut fonctionner. Mais toute erreur ou approximation introduite en amont se retrouve en aval et peut vicier définitivement des raisonnements apparemment bien construits. Le datajournalisme, par exemple, utilisant de grandes masses de données pour faire naître des informations nouvelles –, comme des statistiques de population carcérale, corrélées à des distributions géographiques

d'événements ou de risques – peut conduire à des conclusions erronées si le contexte de validité des informations n'est pas mûrement évalué.

Plus globalement, c'est toute la problématique de la transparence contre le secret qui se trouve profondément modifiée. Le journalisme, notamment, se trouve confronté au paradoxe de protection accrue des sources à mesure que se développe l'exigence de transparence des données.

À cet égard, rappelons que le document avait une fonction de preuve (origine, auteur, contexte, références, etc.). Les données sont au contraire « décontextualisées », dispersées, discrètes. Aussi peut-on raisonner logiquement sur des données sans se soucier de leurs références. Le web des données ouvre à la fois plus de possibilités, mais limite aussi les capacités intrinsèques de vérification. La question du contexte de validation de l'information demandera au contraire de nombreux travaux en fonction de nouvelles confrontations pour éclairer précisément les garanties que l'on est en droit d'attendre.

De nouvelles régulations entre transparence et confidentialité

De même, la nécessité de l'ouverture et la transparence des données publiques, afin que les organismes acceptent d'« exposer » ce qui était souvent au cœur de leurs missions, peut se confronter à des problèmes de limites de confidentialité. Il faut insister sur le fait que nous sommes les héritiers plusieurs fois millénaires de la relation au document, avec ce que cela implique en matière de possibilité de communication (de publication) ou de confidentialité (définissant la sphère privée, la signature). Passer au web des données, c'est

aussi affronter un changement de paradigme en matière de transparence et de confidentialité.

De fait, sous l'impulsion de l'administration américaine (reprise ensuite au Royaume-Uni et dans le reste de l'Europe), le mouvement des *data.gov* est en pleine expansion.

La question de la réutilisation des données (une fois réglée leur mise au format RDF) fait naître de nombreuses difficultés sociales de mise en œuvre. Ces situations nécessitent l'élaboration de consensus accrus, au terme de processus de discussion et de démocratie, afin que de nouvelles règles juridiques autorisent et encadrent cette nécessaire transparence. C'est typiquement le cas des données de santé, produites et « contrôlées » par des organismes de sécurité sociale, d'assurance-maladie ou par des institutions médicales. Le croisement de ces données sociales serait de la première importance pour le suivi d'usages de médicaments par exemple. Mais les conditions d'anonymisation sont souvent jugées insuffisantes par tel ou tel acteur des systèmes de santé, au point de refuser cette ouverture des données.

En tant qu'utilisateur en ligne, je suis identifié, d'autant plus que je suis par ailleurs présent sur des réseaux sociaux. Qui me garantit que mes requêtes (par exemple celle citée plus tôt sur l'identité d'un cardiologue ouvert le dimanche dans le Rhône) ne seront pas utilisées en les croisant avec d'autres données (d'ordre marketing ou d'opinion) ? La question de l'ouverture des données de santé est de fait une des principales problématiques sensibles, avec là encore la même ambivalence. Refuser l'ouverture des données de santé (y compris celles rendues anonymes), c'est s'opposer à des progrès scientifiques ou de services ; exiger une

transparence complète, c'est risquer des abus d'effraction dans la vie privée.

Au sein de secteurs économiques qui gagneraient à une ouverture et à un partage à grande échelle des données, par exemple pour les transports aériens, des résistances socio-économiques à la transparence venant des compagnies aériennes ou des agences de voyage bloquent des avancées possibles.

Plus généralement encore, avec le développement des réseaux sociaux et des formes contributives du net, la production « humaine » de données est sans limite. Tim Berners-Lee, en 2007, proposait même de laisser le terme « *World Wide Web* » pour celui de « *Giant Global Graph* », indiquant cette croissance des données relationnelles. Corrélativement, les « ressources », au sens du *linked data*, sont de plus en plus des éléments du monde physico-chimique lui-même, *data* produites automatiquement par toutes sortes de capteurs, de détecteurs ou de procédures de l'« Internet des objets », ou d'« objets communicants ». Comment s'orientera l'architecture, et plus encore l'« urbanisme », qui aura en charge la gestion équilibrée des données et la maîtrise harmonieuse du web sémantique ?

Enfin, ce n'est pas le lieu ici de traiter du paradoxe qu'il y aurait à voir d'une part se généraliser un système d'intelligence collective et distribuée avec le web sémantique, mais d'autre part observer que cette montée en puissance se réalise et s'organise avec de grands monopoles informatiques et économiques (comme Google, Facebook, Amazon ou Apple) particulièrement jaloux de leurs secrets d'entreprise. Pour le moins, des traditions divergentes de décentralisa-

tion et d'hégémonie sont à l'œuvre, sans que l'on sache de quoi sera fait le futur numérique.

Globalement, avec le web sémantique, c'est donc souvent la frontière entre données publiques et usages privés, et même entre signification du « bien public » et « intérêt privé », qui demande à être redéfinie à nouveaux frais, accentuant le caractère socioculturel du système technologique d'Internet.

CHAPITRE 6

Les modèles économiques de l'édition numérique

Gérard Wormser

L'édition papier, telle que nous la connaissons, s'est développée à partir du début du XVIIIe siècle. Plusieurs notions clés de cette forme de production du savoir sont nées et se sont affirmées pour rendre possible un modèle économique particulier : celui du copyright tel qu'il est défini à partir du Statut d'Anne, promulgué en Grande Bretagne en 1710. À partir de ce moment, les auteurs deviennent des professionnels, rémunérés par les maisons d'édition chargées de mettre en forme et de faire circuler leurs textes. Avec la naissance et le développement du web, à partir des années 1990, de nouveaux dispositifs de circulation des contenus se mettent en place et se présentent comme des modèles alternatifs. Cette concurrence nouvelle menace de mettre en crise le modèle traditionnel de l'édition papier. Quels modèles économiques pour la circulation du savoir sur le web ? Est-il possible d'éviter la désorganisation totale et la perte conséquente de qualité qui effraie certains usagers ? Quels sont les enjeux liés au développement de nouveaux modèles économiques ? Ce sont les questions auxquelles ce chapitre tente de répondre.

Un modèle économique en changement

Tant les travaux de Philippe Aigrain en Europe que ceux de Lawrence Lessig aux États-Unis l'affirment : dans nombre de domaines de nos activités intellectuelles, nous ne faisons pas principalement d'actes marchands, alors même que nous accédons à des biens de valeur auxquels nous consacrons une part importante de notre temps. Les biens non marchands se sont multipliés avec le développement des médias. La photographie « amateur » en est l'exemple paradigmatique, mais c'est le propre de la « société du spectacle » que de fondre en un geste unique un acte marchand et un projet de « développement personnel ». S'agissant de l'édition, cela n'est guère nouveau. Peu d'auteurs ont pu vivre des ventes de leurs ouvrages. Pendant environ 150 ans, dans les pays développés d'Europe et d'Amérique, on a pu croire qu'un réseau de libraires séduirait un public cultivé assez nombreux pour rétribuer toute la « chaîne du livre ». Cette économie s'est rompue en plusieurs temps. La montée en puissance de la diffusion de la musique et de la vidéo sur des supports compacts et bon marché s'est accompagnée d'investissements considérables dans des réseaux de promotion et de vente à distance : les programmes télévisés diffusant musique, films et interviews, les multiples opérations commerciales de la grande distribution et la généralisation des appareils mobiles de diffusion de ces sons et de ces images ont réduit l'emprise des éditeurs dans le monde de la culture et les ont contraints aujourd'hui à diffuser commercialement des fichiers numériques au lieu d'imprimer des livres. Cette transformation s'accompagne d'une révolution dans la forme des ouvrages. De plus en plus de possibi-

lités d'édition augmentées, multimédia, sont explorées. Le livre classique n'est plus la seule forme de transmission du savoir, ni peut-être la principale.

Cependant, l'édition traditionnelle a développé certains des aspects centraux de la diffusion de biens culturels. Leur utilisateur ne paie qu'un droit d'accès puisqu'il ne détruit pas le bien en le consommant, et que, une fois payés les frais d'impression et de distribution, ce dernier reste disponible pour tout autre. Le marché d'occasion contribue à pérenniser la notoriété des éditeurs et des auteurs et fait participer les lecteurs à l'établissement d'une cote des ouvrages. L'économie de l'édition repose aussi sur la mutualisation auprès d'un vaste public qui ne paie pas chaque ouvrage en particulier : les bibliothèques assurent cette circulation à moindre coût qui accroît la circulation des livres et des idées tout en assurant des revenus prévisibles aux éditeurs. Ces traits sont-ils conservés dans le cas de la transformation des ouvrages en fichiers numériques ? L'édition numérique recouvre des domaines suffisamment variés pour ressortir d'une multiplicité de modèles. Nous nous contenterons dans les pages suivantes de présenter ceux-ci sans pouvoir anticiper sur leur devenir.

De l'édition papier à l'édition numérique

La pérennité de l'édition papier venait d'un ensemble de facteurs déterminant sa capacité d'action. Le catalogue en était la marque éminente. Indiquant une ligne éditoriale, il confirmait une réputation et faisait bénéficier l'éditeur de la « longue traîne » : sur une longue durée, un ouvrage pouvait amortir ses coûts sans avoir jamais été un succès, par de petites ventes. L'édition est rarement un métier spéculatif,

mais ses activités passées ont des effets durables. Cet espoir se réalise d'autant mieux si l'éditeur maîtrise des circuits de diffusion et de distribution. Anciennement, nombre d'éditeurs étaient aussi libraires, et si certains éditeurs distribuaient leur production et celle d'autres maisons, il en résultait un chiffre d'affaires accru et toujours à l'équilibre. Bien évidemment, ces métiers exigent de forts capitaux et rapprochent l'édition du commerce de masse. La prescription et la recommandation existent également de longue date : les ouvrages scolaires ont pris le relais des manuels de religion, certains titres sont vendus par abonnement ou souscription, un ensemble de récompenses sous forme de « prix littéraires » vient stimuler les ventes sur quelques titres dont la profitabilité devient considérable, et la médiatisation intensifie encore cet effet. Ici, les maisons d'édition sont étroitement imbriquées avec le secteur des médias : certains auteurs relèvent du show business le plus outrancier, mais ils rapportent. De la célébrité naissent alors des droits dérivés (traduction, adaptation) qui augmentent les marges bénéficiaires de l'éditeur initial. Enfin, pour nombre d'ouvrages difficiles ou spécialisés, les subventions, les achats publics et divers dégrèvements fiscaux favorisent la durabilité des activités.

Ces aspects centraux du système éditorial se configurent de tout autre manière avec le numérique. La profusion nouvelle de documents en tout genre et des nouveautés fragilise le catalogue tant le meilleur des publications antérieures est de fait intégré aux publications nouvelles. Cela est moins vrai pour la fiction que pour les ouvrages documentaires, bien entendu. La maîtrise de la diffusion semble un lointain souvenir et les éditeurs composent avec des

plateformes nouvelles (notamment Amazon) qui les renvoient certes à leur métier fondamental, la création, mais les privent de la possibilité d'investir dans des circuits commerciaux. Il devient de plus en plus délicat de contrôler la recommandation là où le bouche à oreille a été remplacé par les réseaux sociaux. Le secteur de l'édition fait donc face à une adaptation d'ensemble, sans compter qu'il n'existe que peu d'éditeurs qui puissent dépenser en continu les sommes requises pour populariser leur production dans les circuits médiatiques. Les précédents de la musique ou de la presse n'augurent rien de réjouissant pour les acteurs du secteur marchand, confrontés tout à la fois à la puissance des entreprises nativement numériques et aux développements proprement ahurissants de l'économie des réseaux sociaux et de la recommandation. Dans le développement de l'édition numérique, quelle sera la place des éditeurs ? La réponse à cette question tient pour une large part aux types de licences et de modèles de diffusion qui seront plébiscités par le marché.

Le passage au numérique des procédés éditoriaux se produit à la fin du siècle dernier. L'édition sur des supports numériques prend alors acte de transformations radicales et sans retour. L'objectif le plus visible consiste à prendre place dans de nouveaux circuits commerciaux de biens « dématérialisés », dont la pénétration est considérable. Le public cultivé déserte les librairies et s'informe en réseau. Dès lors, les propositions des éditeurs, capables jusqu'à récemment de tenir le marché des centres urbains et des campus, sont concurrencées par les requêtes de lecteurs au « budget-temps » très contingenté. Le cycle ancien reposait sur une subtile asymétrie : les profits s'accumulent sur une

partie du catalogue et des nouveautés, alors que le niveau maximal des pertes est couvert par les frais d'impression et de stockage. Quelques choix judicieux et de bonnes ventes sur quelques titres suffisent, et le catalogue masque en réalité une part importante d'ouvrages à la rentabilité problématique. La règle économique de l'édition consiste à abaisser le taux marginal permettant de publier tout en concentrant les dépenses sur la petite partie des publications les plus rentables. Cela reste vrai, mais avec des chiffres d'affaires en péril. Nombre des métiers de l'édition disparaissent, et certains relais se mettent en place sans véritable organisation programmée, qu'il s'agisse de la numérisation de titres non réimprimés ou de nouvelles demandes aux auteurs pour céder aux éditeurs des « droits numériques » qui ne faisaient pas partie des contrats initiaux. Pour autant, de nouveaux éditeurs exclusivement numériques n'ont pas la partie facile. Faute de notoriété ou de catalogue, ils doivent se spécialiser et agir au cœur des tendances les plus actuelles, au risque de tout perdre s'ils surfent sur une mauvaise vague.

Nouveaux circuits commerciaux

Le défi à relever pour l'édition tient donc au changement des pratiques d'information, d'enseignement et de loisir. La diffusion de fichiers numériques contraint en réalité l'édition à repenser toutes ses pratiques, tant le « contenu » des livres tenait compte de leur matérialité, qui déterminait à son tour leur modèle économique. Une partie substantielle des profits récurrents disparaissent avec l'essor des sites portails, des moteurs de recherche et des communautés en ligne. Comment vendre de l'information générale, des

guides pratiques ou des débats d'idées quand leurs éléments sont abondamment présents sur divers blogs et que les actualisations se font au sein des communautés d'usagers ? L'*Encyclopedia Britannica* a ainsi cessé d'imprimer des ouvrages, et les dictionnaires et les cartes géographiques sont à présent intégrés aux services en ligne gratuits auxquels chacun peut accéder. Google fidélise ses usagers et récolte une infinité d'informations enrichissant ses pages en pourvoyant chacun d'une boîte à outils personnelle quasiment universelle. Et Facebook rivalise avec l'ensemble des services en ligne en diffusant *Home*, son logiciel pour intégrer les fonctions des téléphones intelligents sous Android à l'expérience Facebook. Comment les créateurs d'information peuvent-ils rivaliser avec de telles entreprises sans lesquelles ils disparaissent purement et simplement des écrans ?

Regardons du coté de l'information générale. Wikipédia restitue aux lecteurs les connaissances dès qu'elles ont été diffusées. Autant dire que cette encyclopédie et les divers sites gratuitement accessibles sur tous les sujets posent un sérieux problème d'obsolescence pour une grande partie des ouvrages généraux imprimés. La chaîne commerciale est fragilisée, car les chiffres d'affaires des réseaux de librairies générales étaient largement faits de ce genre d'ouvrages. Plusieurs grands distributeurs (par exemple Fnac, Barnes & Noble et Archambault) ont réagi en créant des librairies virtuelles de grande ampleur et, pour certains, en développant leurs propres tablettes de lecture (Kobo pour Fnac, Nook pour Barnes & Noble). Ces projets tendent vers la possibilité de concurrencer la politique d'Amazon et son Kindle.

Dans l'enseignement et la recherche, les institutions produisant les connaissances sont fortement incitées à diffuser gratuitement l'essentiel de leurs savoirs. En effet, la concurrence devenue internationale entre les grandes universités incite ces dernières à rivaliser auprès des étudiants de valeur en les attirant par des systèmes de diffusion numérique de leur enseignement. Et les centres de recherche ou de documentation publics sont invités par leurs financeurs à diffuser leurs données et leurs contenus. Cette diffusion institutionnelle est une manière de justifier les fonds obtenus par ces institutions et d'abaisser les coûts de production des pays développés en accroissant la mise à disposition d'informations. Il demeure un secteur privilégié, celui des ouvrages accompagnant la formation de cadres techniques de haut niveau : qu'il s'agisse de juristes ou de médecins, d'ingénieurs ou de managers, les synthèses de qualité, les ouvrages de pointe assurant la mise à disposition fiable de la meilleure information relèvent d'investissements pérennes dans les « ressources humaines », qui sont l'essentiel de l'avantage concurrentiel recherché par les entreprises et les institutions. Cette documentation professionnelle demeure un marché lucratif sous quelque forme qu'elle puisse être : il est encore possible de l'imprimer, et ses formats numériques peuvent faire l'objet d'une prescription lucrative, surtout si elle atteint un public international.

Enfin, pour l'édition de loisir, le temps disponible des lecteurs est allé aux médias de masse avant même que l'édition ne commence son virage numérique (télévision, musique, magazines, conversations en ligne, jeux vidéo). En conséquence, nombre de publications seraient aujourd'hui de moindre rapport si elles n'étaient pas diffusées par voie

numérique : des best-sellers aux bandes dessinées, la part des circuits commerciaux contrôlée par les éditeurs se réduit, avec pour effet de limiter le champ de la prescription (les libraires et les bibliothèques publiques cessent d'être les principaux indicateurs pour les éditeurs). La voie est libre pour les plus grands acteurs du numérique.

Nouveaux modèles

Dans un contexte où il lutte pour sa survie, le secteur éditorial s'efforce de sécuriser ses revenus pour concentrer ses efforts de créativité au service de contenus nouveaux. Ces derniers doivent naturellement comporter divers éléments que l'on ne trouvera pas aisément sous forme de téléchargement libre. Dès lors, les modèles-types qui s'imposent partagent l'édition en trois secteurs.

L'un se réfère au modèle de la presse. Celle-ci se débat actuellement entre la baisse inexorable du lectorat et des revenus publicitaires sur le papier, et la faible rentabilité des supports numériques. Le dilemme est entre la gratuité financée par la publicité ou la fermeture de l'essentiel des pages, réservées aux abonnés. Dans les deux cas, la fidélisation passe par une obligation de renouvellement permanent des contenus et leur rotation instantanée, en contrepartie d'une dématérialisation du support qui réduit considérablement les coûts initiaux. En France, le journal *Médiapart* est parvenu à l'équilibre sans publicité en faisant payer ses contenus. Mais la contrepartie en est la spécialisation de son lectorat sur la politique intérieure et un modèle axé sur les lecteurs « professionnels » qui ont un usage précis des informations qu'ils recherchent. Ce modèle *B to B* (*Business to Business*) rappelle le modèle des anciens cabinets

de lecture, face à un modèle *B to C* (*business to consumer*) où la gratuité est devenue une norme si elle s'accompagne de publicité. Les éditeurs de livres tentent bien de faire payer les contenus qu'ils éditent sous forme numérique. Mais les études montrent que la clientèle attend une baisse de tarifs d'environ 1/3 par rapport aux livres imprimés. Les conditions d'émergence de ce modèle semblent se réaliser avec la montée en puissance des libraires en ligne, du type Amazon, Fnac ou Archambault : ceux-ci peuvent à tout moment substituer des fichiers numériques aux exemplaires papier et jouer le rôle d'intermédiaires universels. À cette fin, ils se gardent bien d'empiéter sur les activités purement éditoriales (hormis l'autoédition sur la plateforme Amazon) pour ne pas tarir la source de leurs revenus alors qu'ils luttent pour imposer leurs services en ligne.

Dans ces conditions, les éditeurs peuvent accepter de vendre des livres numériques protégés par des DRM (*Digital Rights Management*). Les fichiers ne doivent pas pouvoir circuler aisément d'un ordinateur à l'autre. Les livres numériques achetés chez Barnes & Noble ne sont utilisables que par un lecteur dont la carte d'acheteur est valide. Un nombre limité de copies privées peut être autorisé, ou bien une licence d'usage collectif peut être envisagée dans un cadre institutionnel, une bibliothèque ou une école par exemple. Dans d'autres cas, et notamment les revues culturelles qui n'ont jamais constitué un secteur commercial à proprement parler, ces dispositifs de filtrage sont lourdement pénalisants : jusqu'à présent, ces revues n'ont pas trouvé le moyen de changer leur modèle fondé sur les abonnements institutionnels. Indépendantes ou liées à des groupes de presse et d'édition auxquels elles ne rapportent rien, elles ne peuvent

prendre d'initiatives par elles-mêmes et sont très fragilisées. Le débat des idées y perd une part considérable de son audience.

Le second modèle repose sur la prescription, la subvention et les achats publics. Les revues scientifiques en sont le meilleur exemple, pour un lectorat limité et professionnel. La constitution d'un marché de niche oligopolistique a concentré les savoir-faire et l'accès au marché sur un nombre limité d'acteurs : il est d'usage de mentionner ReedElsevier ou Springer. Ces éditeurs font autorité notamment en raison de la légitimité qu'ils confèrent aux articles qu'ils diffusent. En contrepartie, ils demandent une participation financière importante aux utilisateurs. Pour limiter cette contribution souvent perçue comme exagérée, et qui s'accompagne de fortes restrictions à la circulation de ces contenus en dehors des centres de recherche, plusieurs alternatives ont vu le jour : l'une consiste à diffuser gratuitement les contenus pour peu que ce soient les producteurs de l'information qui financent les lieux d'édition. Tel est le cas de la Public Library of Science (PLOS) qui développe ce modèle avec succès depuis une dizaine d'années. Cette inversion du modèle est bien évidemment tentante pour les pouvoirs publics qui financent des recherches et doivent ensuite commanditer les abonnements aux revues. On comprend donc bien l'intérêt pour les financeurs de demander une mise en circulation gratuite des contenus qu'ils ont contribué à subventionner. C'est le sens de la recommandation de juillet 2012 émanant de la Commission européenne et destinée aux États membres. Il suffit en effet de réserver un pourcentage des crédits de recherche en vue de la publication pour faire peser une menace de baisse de financement sur les éditeurs

scientifiques : soit ils changent de modèle et adoptent le principe d'un financement en amont et une ouverture de la circulation, soit leur valeur ajoutée doit être mesurable en termes d'expertise, de recommandation et de sélectivité des contenus.

Le troisième modèle s'inscrit dans une mondialisation de fait. La course aux best-sellers et aux séries déclinables sur divers supports, appuyés sur des circuits de diffusion « grand public », a bien évidemment favorisé les livres d'art et de voyage, dont la conception a depuis longtemps supposé des coéditions internationales pour amortir une qualité graphique et documentaire irréprochable. Mais ce modèle est à son tour fragilisé. De fait, l'abondance iconographique et multimédia accessible par YouTube, Dailymotion, Flickr ou Fotolia accroît le risque d'échec du lancement d'ouvrages de prestige que l'on ne sait actuellement pas commercialiser sous forme numérique. En effet, la qualité du papier, le statut d'objets de plaisir et de distinction de tels livres s'accommodent mal d'une diffusion par le biais de fichiers, d'autant que ces derniers devraient être de grande qualité et que leur circulation incontrôlée aurait des répercussions sur les divers ayants droit (photographes, musées, etc.) qui en attendent des revenus. La recommandation européenne porte bien sur une diffusion gratuite de contenus scientifiques subventionnés, mais non sur les documents issus du patrimoine culturel et audiovisuel que la France entend soustraire aux négociations d'un accord de libre-échange souhaité par les États-Unis.

L'édition de recherche, expérimentale, artisanale, innovante à petite échelle relève alors du bénévolat et de la constitution de solidarités locales, mais son économie reste

aléatoire ; elle ne peut guère conquérir de marchés ni transformer le prestige en profits. Une partie symboliquement significative de la création intellectuelle est progressivement poussée hors du jeu, alors qu'elle demeurait, dans le circuit de l'édition traditionnelle, au carrefour des trois secteurs que nous avons distingués. Le « grand lecteur » était prioritairement un « professionnel » capable d'amortir ses investissements dans un cadre de légitimité sociale : il donnait le ton. L'édition de création fidélisait ainsi ce premier public, lequel transformait cette production en lui conférant un accès à la prescription (scolaire et universitaire) et, pour une part de ces titres, à la mondialisation : traductions, adaptations, nouveaux tirages en format de poche étaient autant de garanties pour ce secteur. D'abord reconnu par les initiés, un auteur pouvait ensuite s'adresser à un public élargi. Ces trois moteurs sont aujourd'hui grippés, car les circuits de diffusion prépondérants se situent bien davantage du côté de l'expérience du lecteur que des enjeux pour la pensée. Le modèle de la longue traîne peut bien nous rassurer, mais un certain style de débat et d'écriture a vécu. Faute d'un modèle de diffusion numérique adapté au renouvellement du lectorat qui avait fait du livre le vecteur du renouvellement des idées, mais aussi en raison d'évolutions proprement culturelles, les activités culturelles sont toujours plus ramifiées et variées, et la norme de leur appropriation n'est plus tant l'approfondissement d'une question que le survol rapide et le croisement des références. Robert Darnton avait fort bien exposé cette évolution dans des articles déjà anciens.

Nouveaux formats hybrides

Il reste donc aux responsables de maisons d'édition à se demander comment adapter leurs réalisations aux paramètres actuels. Parmi les questions spécifiques posées à l'édition numérique, celle des compétences d'écritures multimédias et hypertextuelles est centrale. Elle ouvre à la question des droits, des citations et des emprunts, car ces écritures sont en grande partie référentielles : elles supposent d'envisager divers parcours de lecture et d'en formaliser quelques-uns. L'édition numérique grand public a toutes les chances de créer des formats de compilations et d'anthologies, et d'investir dans des navigateurs qui permettront aux lecteurs de circuler au sein d'ensembles composites et potentiellement infinis. Voici longtemps qu'Umberto Eco avait thématisé l'idée de l'œuvre ouverte, quand bien même l'auteur du *Nom de la Rose* est devenu particulièrement pessimiste relativement à la transmission culturelle. Plus que jamais, à l'époque des réseaux sociaux, la constitution de chaînes de prescription numérique (par des actions de marchandisation « virale », mais aussi par capillarité au sein des réseaux personnels où s'établissent la réputation et la notoriété) devient la règle. L'hybridation en cours consistera à ajouter des services à valeur ajoutée aux documents librement mis en circulation : des modules payants pourront s'ajouter à des modules gratuits ou à prix cassés. C'est une des façons qu'ont les éditeurs de pouvoir espérer intégrer l'obsolescence rapide des productions et de leur cycle de renouvellement. Mais en contrepartie, ils devront aussi s'efforcer de repérer les auteurs potentiels et les capter dans un monde de diffusion ouvert.

En conclusion, l'édition a quitté l'ère du patrimoine et de la transformation du capital symbolique en capital économique pour entrer dans l'ère des « services transactionnels » à usage immédiat. Qu'il s'agisse de plaisir ou de formation, de loisir ou de compétences spéciales, il ne s'agit plus de « transformation » mais bien d'« application ». Selon l'adage de McLuhan, il s'agit bien du médium comme message. L'édition numérique devient avant tout une « boîte à outils » produite par des « ingénieurs » et des « techniciens » pour des publics différenciés et identifiés, segmentés.

CHAPITRE 7

Le libre accès et la « Grande Conversation » scientifique

Jean-Claude Guédon

La facilité de circulation des contenus garantie par le web et la baisse prodigieuse des prix de production donnent lieu à de nouvelles possibilités et à de nouveaux modèles de publication. Nous considérons que le libre accès est bien plus qu'un des modèles possibles pour la publication des contenus scientifiques. Nous pensons qu'il est la caractéristique principale des documents numériques. C'est ce qui sera expliqué dans ce chapitre.

Libre accès et *sustainability*

La problématique du libre accès repose sur deux conditions préalables : en premier lieu, elle ne touche la sphère de la publication que dans le contexte précis de la production de connaissances ; de plus, le libre accès n'est viable qu'en présence de la numérisation en réseau. Dans le premier cas, cette affirmation revient à dire que le caractère marchand – et ses conséquences –, attaché aux documents scientifiques depuis l'imprimerie, est remis fondamentalement en question. Le libre accès, en effet, ne relève nullement d'une perspective mollement idéologique sur les relations qui pourraient exister ou non entre le libre, le gratuit et l'Internet ; il s'appuie plutôt sur une analyse serrée de la nature même de la communication scientifique et/ou savante, bref de la « Grande Conversation » qui, à travers le temps et l'espace, noue et structure le territoire mondial de la recherche. Dans ce contexte particulier, qui n'a rien à voir avec les livres de jardinage ou les romans, les documents offrant des résultats validés de recherche sont mis à la disposition des autres chercheurs pour leur permettre d'avancer dans leurs propres travaux. Cette mise à disposition comporte évidemment un coût qui doit être pris en charge d'une manière ou d'une autre, mais cette prise en charge financière n'est nullement synonyme de commerce. Que le commerce se soit immiscé dans la communication scientifique et/ou savante ne peut être nié, mais partir de cette constatation pour en déduire que le format « marchandise » est nécessaire aux publications savantes constitue un saut dont l'illogisme est évident.

La recherche, sans exposition à l'examen critique des pairs et même du public, ne veut rien dire. En d'autres

mots, la recherche n'existe pas sans publication (c'est-à-dire mise à la disposition du public) et, symétriquement, la publication correspond à une phase essentielle de la recherche, au même titre que l'heuristique, l'expérimentation, l'observation, l'interprétation, etc. Banale en apparence, cette remarque conduit pourtant à des conséquences importantes ; en effet, la recherche scientifique (distinguée ici du « développement » qui lui est souvent associé dans le raccourci classique de Recherche et Développement) est largement soutenue par de l'argent public ou provenant de fondations sans but lucratif. Cela revient à dire que la recherche scientifique n'est pas *sustainable* au sens anglais du terme : en effet, la *sustainability*, souvent traduite par « durabilité », incorpore une dimension de pérennité fondée sur des ressources engendrées en interne. La subvention, de ce fait, n'entre pas dans le cadre du *sustainable* ; en revanche, le modèle commercial apparaît acceptable parce qu'il engendre en principe des profits. Émerge alors un petit paradoxe : depuis la Révolution scientifique (au moins), la recherche scientifique se déploie et croît sans jamais avoir été *sustainable*… Il en va de même pour les publications scientifiques qui l'accompagnent.

Pourquoi les gouvernements subventionnent-ils la recherche scientifique ? La réponse la plus simple revient à dire qu'il s'agit en fait d'une infrastructure économique et militaire, au même titre que les routes, par exemple. Les routes, en effet, ne sont pas non plus *sustainable*, mais leur pérennité et leur durabilité ne soulèvent aucun doute.

Le libre accès ne peut réellement exister sans numérisation ni réticulation. Bien sûr, un désir de libre accès se manifeste dès l'imprimé, et l'importance croissante des

bibliothèques en constitue l'un des signes ; de la même manière, les tirés à part que les revues offraient de façon routinière à leurs auteurs jusque dans les années 1980 alimentaient un commerce intellectuel sans préoccupation commerciale. Mais la numérisation permet d'obtenir deux résultats essentiels pour le libre accès : on peut produire des copies parfaites à un coût marginal proche de zéro, et la présence des réseaux ouvre la possibilité d'une dissémination à un coût marginal à peu près nul. De plus, les outils de publication permettent d'abaisser considérablement les coûts associés à la production de documents de bonne facture visuelle. Tout cela explique les explorations diverses menées par divers chercheurs dans le monde à partir de la fin des années 1980. Par exemple, au Québec, la revue *Surfaces* (dont j'ai été l'un des fondateurs) a commencé à publier à la fin de 1991, se situant alors parmi la première douzaine de revues savantes électroniques – d'ailleurs en accès libre – dans le monde.

Un peu d'histoire

Ce serait succomber à une forme simpliste de déterminisme technologique que de présenter le libre accès comme la conséquence de la numérisation et d'Internet. La « Grande Conversation » scientifique n'avait pas attendu ces innovations, pourtant majeures, pour subir des transformations au total peu comprises de l'extérieur, mais qui ont abouti à rien de moins qu'une contre-révolution silencieuse dans le domaine. De marginales dans le système de communication scientifique, les grandes maisons d'édition ont commencé à devenir dominantes après la Deuxième Guerre mondiale. Jouant sur le succès de la troncature de l'édition

scientifique mondiale qu'a réussi à imposer, dès les années 1970, le *Science Citation Index*, les éditeurs ont réussi à créer un marché inélastique de revues scientifiques. Une fois définie comme membre des revues centrales (*core journals*), une revue devient incontournable, en particulier aux yeux des bibliothécaires, et peut alors voir son prix augmenter fortement sans que la demande puisse diminuer. Les chercheurs sont en fait des lecteurs subventionnés ; mais, ne payant pas la note, ils ignorent souvent la structure économique soutenant une importante section des publications savantes.

Les bibliothécaires, en revanche, financent les achats de livres et d'abonnements ; il n'est donc pas surprenant, dès les années 1970, de les voir sonner l'alarme en documentant ce qu'il est convenu d'appeler la *Serial Pricing Crisis*. Ensuite, avec la numérisation des revues qui débute au milieu des années 1990, les bibliothèques ont dû réagir à une situation transactionnelle entièrement nouvelle : l'acquisition et la possession effective d'objets imprimés, en particulier de livres et de numéros de revues, a été remplacée par des licences d'accès à des bases de données d'articles.

Les protestations des bibliothèques à l'égard des coûts d'abonnement, puis des licences d'accès, furent finalement entendues par des chercheurs dans les domaines biomédicaux : au début de 2001, une pétition fut lancée sous le nom de *Public Library of Science* (PLOS). Cette pétition, en apparence, fut un échec simplement parce que, naïvement, elle avait incorporé des menaces inapplicables. Pour autant, elle contribua à rendre visible la situation de plus en plus insupportable dans laquelle la communication scientifique en général, et les bibliothèques en particulier, se débattaient.

Diverses fondations, alertées par la pétition PLOS, commencèrent à s'intéresser à cette question. En particulier, l'*Information Program* des *Open Society Foundations* (OSF) de George Soros a organisé une rencontre à Budapest à la fin de 2001. Cette réunion conduisit à la publication de la *Budapest Open Access Initiative* (BOAI), le 14 février 2002, qui catalysa un mouvement mondial en faveur du libre accès. Une dotation de plusieurs millions de dollars des OSF, annoncée en 2002, permit de mettre en route un certain nombre d'activités, de réunions, d'organisations qui ont joué un rôle essentiel dans l'émergence, puis le déploiement du libre accès.

Sans surprise, les bibliothèques commencèrent presque immédiatement à s'organiser pour appuyer le mouvement naissant. Elles le firent en se dotant de dépôts institutionnels dont l'objectif était de récupérer les publications locales de leur institution, de leur donner une exposition nouvelle et de mieux saisir les résultats de la recherche locale. Ce faisant, elles contribuaient à appliquer concrètement l'une des deux solutions proposées par la BOAI dès 2002, en l'occurrence la voie que Stevan Harnad, l'un des grands pionniers de l'accès libre, baptisera « voie verte », et qui consiste à reposer sur l'auto-archivage des publications des chercheurs. Cette solution a graduellement été renforcée par divers moyens, tels des répertoires mondiaux de ces dépôts. Des standards de métadonnées tel l'*Open Archives Initiative Protocol for Meta-Data Harvesting* (OAI-PMH), standards conçus par Carl Lagoze et Herbert Van de Sompel, ont également vu le jour. Il ne faut pas oublier non plus le modèle qu'offrait, en physique des hautes énergies, la base de données ArXiv de *pre-prints* que Paul Ginsparg animait, depuis 1991, au laboratoire national de Los Alamos aux États-Unis.

Ce type de dépôt disciplinaire s'est ensuite étendu à d'autres disciplines, par exemple en économie (RePEc), avec des variations dans le détail des opérations.

En parallèle, et de façon plus prévisible, l'effort pour créer des revues en accès libre s'est évidemment intensifié (« voie d'or », toujours selon la terminologie de Stevan Harnad). D'ailleurs, avant même la BOAI, Vitek Tracz avait créé en 2000 un ensemble de revues en libre accès, ensemble connu sous le nom de *Biomed Central*. Tracz et le premier directeur de *Biomed Central*, Jan Velterop, ont essentiellement inventé la notion d'auteur-payeur dans le domaine des publications libres. Selon cette formule, que l'on assimile trop souvent à la voie d'or, les frais de publication sont reportés en amont et payés par l'auteur ou une institution se substituant à celui-ci, par exemple l'institution d'appartenance, ou un organisme de subvention de la recherche. Le modèle financier de *Biomed Central*, tout à fait acceptable dans la perspective de la *sustainability*, fut repris par un PLOS nouveau genre, descendant de la pétition de 2001. Le nouveau PLOS est en fait un éditeur de publications scientifiques en accès libre qui fut initialement soutenu par une dotation de 9 millions de dollars de la Gordon and Betty Moore Foundation.

À côté des revues financées selon le modèle auteur-payeur, il existe un nombre encore plus grand de revues qui utilisent d'autres moyens, en particulier des subventions de diverses sources, pour alimenter des revues dans toutes les disciplines. Le *Directory of Open Access Journals*, au moment où ces lignes sont écrites en mai 2013, compte en effet 9 191 revues savantes, toutes en libre accès, dont 268 provenant du Canada. Notons pour l'Université de Montréal,

par exemple, la revue *Altérités* gérée par les doctorants du Département d'anthropologie et la revue *Bioéthique Online* qui a démarré en 2012, en collaboration avec l'Université McGill.

Avec près de 10 000 revues en libre accès recensées, au-delà de 2 000 dépôts en tous genres dans le monde, et des manifestations annuelles régulières, telle la *Open Access Week* chaque automne, le mouvement du libre accès ne peut plus être ignoré et, effectivement, il ne l'est plus. La prochaine section va le démontrer en esquissant quelques-unes des lignes de front qui se sont dessinées au cours des dernières années et qui révèlent clairement la nature des forces en faveur et contre le libre accès.

Quelques lignes de front autour de l'accès libre

En dix ans, le libre accès a vécu la trajectoire classique des mouvements contestataires que décrit fort bien ce petit adage de la sagesse populaire : d'abord, ils vous ignorent ; puis ils rient de vous ; et enfin ils vous combattent. Le libre accès, actuellement, et ce, depuis plusieurs années, n'est plus ignoré et les rires se sont tus. Restent les combats…

Ces combats prennent évidemment plusieurs formes. Prenons le cas des dépôts, qu'ils soient institutionnels, thématiques ou nationaux. Après avoir transféré ses droits à une maison d'édition, un auteur ne peut déposer ses propres travaux dans un dépôt public qu'avec l'assentiment de la maison d'édition. Pour ne pas aliéner frontalement les chercheurs, bon nombre de maisons d'édition consentent donc à ce dépôt, mais en l'assortissant de conditions chaque fois particulières. En résulte un paysage flou d'une lisibilité limitée. Par exemple, un chercheur peut se demander quelle

version de son texte il peut déposer, mais ces questions juridiques ne sont pas dans ses priorités. Aussi, le manque de réponse claire à la question précédente tend à inhiber le processus de dépôt. En effet, un chercheur publie en pensant surtout à son dossier personnel. Sa visibilité lui importe aussi, mais les avantages que confère l'accès libre à cet égard demeurent flous, en partie parce que les évaluations institutionnelles, primordiales pour la carrière, fonctionnent sur la base d'une forme de réputation dictée davantage par l'évaluation des revues que par celle des articles.

La variété fluctuante des politiques de dépôts effectivement concédées par les maisons d'édition conduit à démobiliser les chercheurs. L'entretien passif d'une zone de confusion par les maisons d'édition est clair : les auteurs autoarchivent peu, autour de 10-20 % du total possible.

De plus en plus d'institutions, conscientes des limites inhérentes au dépôt volontaire, exigent le dépôt des travaux de recherche de leurs chercheurs. La première université à s'engager dans cette voie fut celle de Minho, au Portugal, dès décembre 2004. Sous le leadership d'Eloy Rodrigues, le directeur des services de documentation de l'Université de Minho, une politique de dépôt obligatoire fut promulguée il y a près de dix ans, ce qui entraîna immédiatement une croissance forte du dépôt institutionnel local. Plus tard, l'Université de Liège instaura, sous la houlette de son recteur, Bernard Rentier, une procédure d'évaluation des professeurs qui repose exclusivement sur les documents confiés au dépôt institutionnel local. Là aussi, le taux de dépôt des articles publiés par les chercheurs de Liège a augmenté très vite.

Actuellement, des politiques diverses d'obligation de dépôt existent dans environ 170 institutions, plus un certain nombre de facultés et de départements. Dans certains cas, comme à la Faculté des arts et des sciences de Harvard, au Massachusetts Institute of Technology, l'obligation de dépôt a été votée par les chercheurs eux-mêmes, ce qui assure à la politique en force une plus grande légitimité.

L'intensité des combats autour du libre accès s'est accentuée lorsque le Wellcome Trust, puissante *charity* anglaise soutenant la recherche biomédicale, entreprit de placer en accès libre les publications résultant de ses programmes de soutien à la recherche. Sous l'impulsion de son directeur, Sir Mark Walport, la fondation britannique décida en 2005 d'énoncer une politique exigeant des chercheurs financés par le Wellcome Trust le dépôt en accès libre des articles publiés six mois (ou moins) auparavant. Cette décision se révéla extrêmement importante ; elle conduisit en 2008 à une loi analogue aux États-Unis[1], qui s'appliqua dès lors aux milliers de publications issues des 18 milliards de dollars de recherche provenant des National Institutes of Health (NIH). Au Canada, le Canadian Institutes of Health Research (CIHR) s'est aligné sur les normes des États-Unis, et de plus en plus d'agences de financement de la recherche, un peu partout dans le monde, suivent ces modèles.

Cette attitude nouvelle des organismes subventionnaires a évidemment inquiété les maisons d'édition. Leurs réactions ont pris une variété de formes, allant de la volonté d'être nécessairement impliquées dans le processus de

1. Division G, Title II, Section 218 de PL 110-161 (Consolidated Appropriations Act, 2008).

dépôt des articles contre rétribution, tactique perverse mais intelligente pour inventer de nouvelles sources de revenus, jusqu'à la tentative de renverser la loi américaine de 2008 sur le dépôt obligatoire.

Elsevier, pour sa part, a résolument opté pour cette dernière solution : cette compagnie a tenté de faire annuler la loi exigeant le dépôt en libre accès des articles publiés avec des fonds du NIH. Ce projet de loi, connu sous le nom de H.R.3699 ou *Research Works Act*, fut introduit en novembre 2011 à la Chambre des représentants du Congrès des États-Unis. Une réaction directe des chercheurs en forme de pétition, organisée sous le nom de *The Cost of Knowledge*, engendra un débat suffisamment intense dans la presse pour convaincre Elsevier d'abandonner son soutien pour le *Research Works Act* en février 2012. Le projet de loi fut alors retiré.

Autour du libre accès, une autre ligne de front est également en train d'apparaître dans divers pays émergents. Longtemps exclues des périmètres définissant les *core journals*, un très grand nombre de revues savantes, par exemple en Amérique latine, ont végété dans l'ombre des revues européennes et américaines, condamnées à la marginalité et à l'invisibilité pour des raisons qui ne sont pas toujours liées au manque de qualité. Pour contrer ce phénomène de « science perdue », pour reprendre l'expression de W.W. Gibbs, des chercheurs brésiliens ont décidé de construire une plateforme imposante de revues, intitulée *SciELO*, qui couvre maintenant plus d'une douzaine de pays, et environ 900 revues. L'effort de *SciELO* s'est porté sur plusieurs fronts : d'abord, professionnaliser la production de revues savantes dans la région, mais aussi construire des métriques

de citations incluant les citations contenues dans le réseau *SciELO*. Le résultat de ces efforts a été de produire un ensemble de revues de plus en plus visibles, même dans les pays de l'Organisation de coopération et de développement économiques (OCDE). Malheureusement, *SciELO* se heurte à deux types de problèmes : les organismes subventionnaires d'Amérique latine, de manière peut-être encore plus rigide et mécanique que les gestionnaires de la recherche dans les pays riches, se fient aux facteurs d'impact tirés du *Science Citation Index*. De fait, les métriques de *SciELO* ne se sont pas encore imposées en Amérique latine. Mais, ironie peut-être prévisible, ces mêmes métriques sont prises au sérieux par certaines maisons d'édition comme Elsevier. Des rumeurs circulent au sujet d'ambassadeurs de ces compagnies qui iraient rencontrer les équipes éditoriales des revues les plus prometteuses sur le plan financier pour tenter de les attirer dans les écuries de ces maisons d'édition. Si tel est le cas, l'idée, bien évidemment, est d'écrémer les revues prometteuses sur le plan lucratif. *SciELO* se trouverait alors pris dans le rôle peu enviable d'antichambre d'Elsevier et autres maisons d'édition, ce qui, bien évidemment, n'est pas le but de ce consortium.

Les quelques lignes de bataille esquissées ici sont loin d'épuiser le sujet. En fait, une guerre sourde se joue actuellement dans les organismes officiels, les parlements, les associations professionnelles. De temps en temps, un événement plus éclatant vient donner un coup de projecteur sur un paysage où dominent les guerres d'influence, les perfidies et les traîtrises propres au monde du lobbying. Cela dit, chaque fois qu'une de ces affaires atteint la presse et le

grand public, il y a défaite partielle des maisons d'édition : celles-ci, en effet, cherchent généralement à maintenir une apparence de sérénité, de coopération et de bonne volonté, preuve s'il en est que tout va pour le mieux dans le meilleur des mondes…

Les combats autour du libre accès, on peut le prédire avec assurance, vont se poursuivre et se multiplier pendant encore bien des années, en fait jusqu'à ce que la communication scientifique trouve enfin un nouveau point d'équilibre. Il ne s'agit pas moins, en effet, que de vivre la transition de l'imprimé au numérique. Actuellement, pour reprendre la belle expression de Gregory Crane, nous vivons simplement la phase des « incunables numériques » et, dans le cas des publications scientifiques, le libre accès constitue en fait une façon de dépasser cette phase. Le libre accès, il ne faut jamais l'oublier, constitue aussi un symptôme indéniable de la transition en cours vers le numérique.

Un peu de prospective

Pendant que les premiers combats autour du libre accès se déroulaient de façon plus ou moins visible, le monde numérique ainsi que le libre accès n'ont cessé d'évoluer. S'il fallait ne mentionner qu'un élément de ces transformations, il faudrait tout de suite souligner que, en quinze ans, le numérique a révélé que tout document bénéficie désormais de deux types de lecteurs : les lecteurs humains et les lecteurs-machines. Google illustre cette deuxième catégorie de manière éclatante, mais des expériences telles que l'*executable papers*, d'Elsevier montrent bien que le document numérique assume des rôles totalement hors de portée

de l'imprimé. Notons en particulier le lien entre articles de recherche et données sous-jacentes, avec les algorithmes permettant de traiter ceux-ci à travers ceux-là.

La mise en œuvre de dépôts en tous genres et de larges collections de revues a également révélé l'importance de la plateforme par rapport aux formes classiques de rassemblement d'articles que sont les revues. De manière presque prévisible, ces constats ont conduit à la création de méga-revues dont le premier modèle vient du monde du libre accès : PLOS One. Ces méga-revues qui publient plusieurs milliers d'articles par mois non seulement transforment le paysage de l'édition savante, mais en modifient aussi subtilement le sens : dans une méga-revue, la sélection se fait sur la base de la qualité de l'article soumis, sans référence aucune à quelque orientation éditoriale que ce soit. Cela veut dire que toute question, si exotique puisse-t-elle paraître, sera acceptée si l'évaluation par les pairs établit la qualité du travail. Ainsi, une recherche bien conduite sur une maladie négligée sera acceptée. En d'autres mots, les contraintes qui se sont souvent exercées sur le champ des questions intéressantes et possibles viennent de s'amoindrir de manière extrêmement significative.

Autour du libre accès, on commence aussi à discuter des questions d'évaluation. Ces questions, dominées depuis une génération par les métriques issues du *Science Citation Index*, sont de plus en plus contestées. Dans les pays émergents, on cherche à revaloriser les questions d'intérêt local (ce qui n'exclut d'ailleurs pas des solutions de valeur universelle) ; ailleurs, on s'inquiète des distorsions de valeurs qu'engendrent les formes exacerbées d'une concurrence gérée au moyen de ces métriques. Or l'accès libre se prête bien à des

formes d'évaluation plus justes, que les publications fermées ne permettent pas aussi facilement. Pensons, par exemple, aux enjeux associés à la volonté de fonder l'évaluation sur les articles, et non les revues.

Plus avant dans la réflexion, on commence aussi à entrevoir comment l'ouverture des vecteurs de la « Grande Conversation » permettront de donner des réponses innovantes à des questions de collaboration scientifique internationale ou de structuration d'espaces de recherche (comme c'est le cas en Europe avec la European Research Area). On commence aussi à comprendre qu'un régime généralisé de données ouvertes va transformer les pratiques de la recherche de fond en comble. La Research Data Alliance, actuellement en voie de constitution avec les États-Unis et l'Europe en son cœur, en constitue un signe évident. L'Australie y adhère déjà, mais *quid* du Canada, pourtant sollicité ?

Conclusion

Ce texte, on l'aura vu, positionne le libre accès comme une facette fondamentale, indispensable, incontournable en fait, d'un passage réussi au numérique, du moins dans les domaines associés à la recherche. En effet, en recherche, ce qui prime, ce ne sont ni les vecteurs, ni les dispositifs, et encore moins les institutions sur lesquelles s'appuyaient autrefois les centres d'où partaient les imprimés scientifiques ou savants, mais bien plutôt les processus au travers desquels se constitue cette « Grande Conversation » qui forme le pivot de l'argument de ce texte. Le libre accès découle directement et nécessairement des besoins de la « Grande Conversation », et l'évolution des dix dernières années démontre en fait que, petit à petit, sa logique se

renforce et s'impose. Parfois, cela se passe en bousculant des ordres établis : ainsi, les métiers regroupés dans une maison d'édition traditionnelle sont en train de se redistribuer différemment au sein d'entités beaucoup plus proches des milieux de la recherche. Les nouvelles alliances entre presses universitaires et bibliothèques, par exemple à l'Université du Michigan, en témoignent. Ainsi, des formes de contrôle sur les questions admissibles ou intéressantes en recherche se voient contournées ou transformées par l'invention de nouvelles formes de vecteurs, ou par la prise en charge de certaines formes de traitement de documents par des algorithmes. Cela dit, il faut toujours dépasser tout réflexe nous confinant aux « incunables numériques » pour garder les yeux fixés sur le bon compas, celui qui pointe vers les promesses d'une intelligence humaine réellement et universellement distribuée. Utopie ? Sans doute ! Mais la communauté des chercheurs s'est constituée sur la base d'une utopie sociale qu'a bien décrite le sociologue R.K. Merton. Et le fil allant de la « Nouvelle Atlantide » de F. Bacon aux structures de la Société royale de Londres demeure visible.

AXE TECHNIQUE

CHAPITRE 8

Les protocoles d'Internet et du web

Jean-Philippe Magué

Nous avons affirmé l'importance d'Internet et du web dans les changements de modèles de circulation des connaissances. Ce manuel propose une conception de l'édition numérique comme l'ensemble des pratiques qui structurent aujourd'hui la production et la circulation des contenus et, dans ce sens, nous ne cesserons de le répéter, c'est le développement du web – et non du numérique en général – qui doit être au centre de notre attention. Voilà pourquoi il est indispensable d'approfondir, d'un point de vue plus technique, les dispositifs concrets sur lesquels se base l'échange de données via Internet et la circulation des documents sur le web. Ce chapitre introduit le sujet des protocoles d'échange de données et d'informations, du protocole TCP/IP, sur lequel est basé Internet, au protocole HTTP, qui permet le partage des documents sur le web.

Des conventions aux protocoles

Envoyer une lettre à un ami est une action relativement banale. Il convient tout d'abord d'écrire la lettre, de l'insérer dans une enveloppe, de noter l'adresse de l'ami en question sur l'enveloppe, d'y coller un timbre et de la glisser dans une boîte aux lettres. L'enveloppe sera alors récupérée par la poste, qui l'acheminera à l'adresse qui y est indiquée, celle de l'ami destinataire.

À bien y regarder, aussi banale que soit cette action, elle repose sur la mobilisation d'un nombre non négligeable de « conventions » que chacun des acteurs impliqués doit respecter. L'expéditeur indique l'adresse sur l'enveloppe d'une manière très codifiée (que l'on inscrive « la troisième maison à gauche après l'école » et l'on est certain que la lettre se perdra), et colle le timbre en haut à droite de l'enveloppe et pas ailleurs. Les divers employés du service postal qui manipuleront la lettre doivent respecter l'ensemble des processus liés à leur fonction qui feront que la lettre passera correctement de centre de tri en centre de tri jusqu'à la boîte aux lettres du destinataire. Quant à ce dernier, il faudra qu'il sache interpréter tout un ensemble d'autres conventions utilisées par l'expéditeur pour écrire sa lettre : la date qui figure certainement en haut à droite est la date de rédaction (éventuellement accompagnée d'un lieu) et le nom indiqué tout en bas de la lettre est la signature, le nom de l'expéditeur. La langue même dans laquelle est rédigée la lettre est une convention entre l'expéditeur et le destinataire.

C'est parce que cet entremêlement de conventions est respecté par l'ensemble des acteurs qui constituent la chaîne allant de l'expéditeur au destinataire que la

communication entre ces derniers est possible. Il en va de même sur Internet lorsque, par exemple, on envoie un courrier électronique ou qu'un navigateur affiche une page d'un site web : la communication a alors lieu entre deux ordinateurs (celui de l'expéditeur et celui du destinataire dans le cas d'un courrier électronique, le serveur du site web et le poste de l'utilisateur dans le cas d'une page web), et cette communication est possible parce qu'un ensemble de conventions sont respectées. Pour des communications entre machines, on n'utilise pas le terme de convention mais celui de « protocole », et ce sont ces protocoles (ou du moins une partie d'entre eux) qui seront détaillés dans ce chapitre.

Les protocoles d'Internet

Pour permettre à deux ordinateurs d'échanger des informations entre eux, il faut un lien physique entre ces deux ordinateurs (éventuellement sans fil). Si on relie non pas deux, mais plusieurs ordinateurs qui vont pouvoir s'échanger des informations, on construit un réseau informatique.

Internet est un réseau de réseaux, c'est-à-dire un ensemble de technologies qui permettent à plusieurs réseaux de s'interconnecter de manière à permettre l'échange d'informations entre ordinateurs connectés non seulement au même réseau, mais aussi sur des réseaux différents. Par exemple, les ordinateurs dans une bibliothèque universitaire sont connectés, certainement par un câble Ethernet[1], au réseau

1. Ethernet est la technologie standard pour connecter un ordinateur à un réseau avec un câble. Mais nous n'entrerons pas ici dans le détail des normes et protocoles des matériels physiques qui constituent Internet.

interne de l'université. Chez un particulier, l'ordinateur familial sera plutôt connecté au réseau de son fournisseur d'accès à Internet, notamment par ADSL[2]. Le réseau de l'université comme le réseau du fournisseur d'accès à Internet sont connectés à d'autres réseaux, eux-mêmes connectés à d'autres réseaux. Pris ensemble, tous ces réseaux forment Internet.

Pour gérer la transmission de données sur ce réseau de réseaux, deux protocoles sont utilisés et constituent le fondement d'Internet : IP, pour *Internet Protocol*, et TCP, pour *Transfert Control Protocol*.

Le protocole IP

La métaphore postale permet de saisir le rôle du protocole IP : les deux éléments indispensables pour permettre l'envoi d'une lettre d'un expéditeur à un destinataire sont l'adresse du destinataire et un service postal qui acheminera la lettre jusqu'au destinataire. Il en va de même de la transmission de données sur Internet : le protocole IP permet d'attribuer une adresse unique à chaque ordinateur (nommée « adresse IP ») et fournit les mécanismes pour acheminer les données à bon port.

Certains ordinateurs des réseaux qui constituent Internet, les « routeurs », ont pour fonction d'aiguiller les données de l'expéditeur au destinataire. L'expéditeur encapsule les informations dans un « paquet IP » qui contient, outre les données à transférer, un certain nombre d'informations, dont

2. *Asymmetric Digital Subscriber Line*, une technologie permettant de se connecter à un réseau informatique via l'infrastructure des réseaux téléphoniques.

l'adresse de l'expéditeur. L'ordinateur expéditeur transmet ce paquet à un routeur de son propre réseau, qui à son tour le transmet à un autre routeur, et ainsi de suite jusqu'à ce que le paquet parvienne au destinataire. Une des spécificités d'Internet est d'être entièrement décentralisé : il n'y a aucune machine centrale ayant une vue globale de l'ensemble de sa structure. Les routeurs ne connaissent d'Internet que les autres routeurs auxquels ils sont connectés. Un routeur qui reçoit un paquet IP doit donc déterminer lequel des routeurs auxquels il est connecté est le plus approprié pour rapprocher le paquet IP de son destinataire. Ce choix est fait par un ensemble d'algorithmes associés à d'autres protocoles d'échange d'informations entre routeurs qui ne seront pas détaillés ici.

Cet aspect décentralisé d'Internet est une de ses forces. Il existe toujours plusieurs chemins possibles par lesquels faire transiter un paquet IP entre deux ordinateurs. Et si, à la suite d'une panne par exemple, un chemin venait à disparaître, un autre chemin serait disponible et pourrait être utilisé. Cela fait d'Internet un réseau extrêmement robuste.

Deux propriétés du protocole IP doivent être soulignées. La première est qu'il est non fiable (ce qui ne veut pas dire « pas fiable »). Une fois envoyé, chaque paquet IP suit son propre cheminement et, s'il se perd, par exemple à cause d'une panne d'un routeur, ni l'expéditeur ni le destinataire n'en sont informés. Rien ne garantit non plus que deux paquets envoyés successivement par le même expéditeur au même destinataire arriveront dans l'ordre dans lequel ils ont été envoyés. La seconde propriété est que la taille des paquets IP est limitée, au mieux à 1 280 octets, soit la taille d'un texte de 200 mots.

Le protocole TCP

La plupart des informations transitant sur Internet (courriers électroniques, pages web) dépassent largement les tailles maximales des paquets IP. Elles doivent donc être découpées en plusieurs paquets de taille appropriée par l'ordinateur expéditeur et reconstituées par l'ordinateur destinataire. Par ailleurs, les informations sont échangées dans le cadre d'interactions complexes nécessitant des transferts d'information dans les deux sens (voir par exemple le protocole HTTP décrit plus bas). Le rôle du protocole TCP est de composer des échanges de paquets IP pour proposer des services plus adaptés aux types d'échanges d'information se déroulant sur Internet. La métaphore appropriée pour décrire le protocole TCP serait celle du téléphone : une machine en contacte une autre pour établir une connexion et, une fois que cette dernière a accepté, elles disposent d'un canal, stable tant qu'aucune des deux n'interrompt la connexion, par lequel des informations de taille arbitraire peuvent transiter dans un sens comme dans l'autre.

Pour rendre cela possible, lorsqu'une machine souhaite transmettre des données, le protocole TCP se charge de les découper en un ensemble de paquets IP. Du côté de la machine réceptrice, il va se charger de réordonner les paquets IP reçus, d'en accuser la réception ou, au contraire, de redemander ceux qui se seraient perdus et de les réassembler pour reconstituer les données initiales.

TCP est une couche de communication construite pardessus la couche IP, encapsulant et masquant les détails du protocole IP (du moins les notions de paquets et de routage, car l'adresse IP est utilisée pour identifier les machines), de

la même manière qu'IP encapsule et masque les détails des protocoles liés aux connexions physiques (Ethernet, ADSL, etc.). Cette construction par couches successives, chaque couche utilisant les services de la couche inférieure pour en proposer de nouveaux, est un principe constituant d'Internet. TCP et IP sont deux protocoles permettant de transmettre des données, mais restent complètement neutres vis-à-vis des données transmises. Cela fait d'Internet une infrastructure très générique, sur laquelle différents services, ou applications, ont été conçus en construisant de nouvelles couches de protocoles par-dessus TCP. Ces protocoles spécifient la manière dont les machines doivent traiter les données que font transiter IP et TCP : par exemple, le courrier électronique, le web (basé sur le protocole HTTP décrit plus bas) ou le transfert de fichiers (ce que fait le protocole FTP, abordé plus bas également).

Lorsqu'une machine accepte une connexion TCP, il faut qu'elle puisse connaître le protocole qui structure les données qui transiteront dans cette connexion (par exemple HTTP ou FTP) afin de les interpréter et de les traiter correctement. Pour ce faire, elle dispose de différents « ports », identifiés par des numéros, et chaque demande de connexion TCP doit indiquer le port visé. À chacun de ses ports, la machine attendra des données conformes à un protocole donné. Ainsi, par exemple, HTTP sera typiquement associé au port TCP 80.

Noms de domaines

Si les adresses IP permettent d'identifier les ordinateurs sur Internet, ayant été conçues à l'usage des machines, elles sont difficiles à mémoriser pour les humains : 192.0.43.10

est un exemple d'adresse IP. Les applications construites sur Internet telles que le web ou le courrier électronique sont au contraire conçues pour être utilisées par les humains. Pour pallier ce problème de lisibilité des adresses, le *Domain Name System* (DNS), a été conçu. Il permet que les machines, en parallèle à leurs adresses IP, soient identifiées par un « nom de domaine », plus facilement mémorisable. Ce sont les noms qui sont couramment utilisés lors de l'accès à une page web ou de l'envoi d'un courriel. Il est plus simple de se rappeler que l'adresse d'un site web est « www.exemple.com » plutôt que « 192.0.43.10 » ou d'envoyer un mail à « nom@exemple.com » plutôt qu'à « nom@192.0.43.10 ». Il y a une correspondance entre adresse IP et nom de domaine, et des machines spécialisées, les « serveurs DNS », permettent d'effectuer la traduction du nom de domaine en adresse IP.

Le web et le protocole HTTP

Le web, ou plus précisément le *World Wide Web*, est un système documentaire construit sur Internet dans lequel les documents, nommés hypertextes[3] ou pages web, sont reliés les uns aux autres par des hyperliens. Ces documents sont affichés dans des navigateurs qui permettent, grâce aux hyperliens, de naviguer d'une page à une autre. Les données qui sont échangées sur le web le sont avec le protocole HTTP, *HyperText Transfert Protocol*[4].

Lorsque deux machines s'échangent des données sur le web avec le protocole HTTP, leur rôle est asymétrique. L'une

3. On trouve également d'autres types de documents sur le web : images, sons, vidéos, etc.

4. On trouve aussi HTTPS, pour HTTP *Secure*, qui fonctionne en tout point comme HTTP, si ce n'est que les données sont chiffrées.

d'entre elles, le « serveur », a pour fonction de fournir des ressources ou des services, tandis que l'autre, le « client », utilise les ressources et les services du serveur. Le protocole HTTP fonctionne selon un principe de requête/réponse : le client formule une requête auprès du serveur, lequel renvoie en retour une réponse au client (il y a une polysémie autour des termes « client » et « serveur » : « client » désigne à la fois l'ordinateur qui émet la requête et le logiciel exécuté par cet ordinateur, typiquement le navigateur web ; de même, « serveur », désigne à la fois l'ordinateur destinataire de la requête et le programme chargé de traiter la requête et formuler une réponse). Un exemple typique de requête est un navigateur réclamant à un serveur une certaine page web, lequel répond en envoyant au client le code HTML de la page désirée. Le navigateur traite alors ce code HTML pour afficher la page à l'utilisateur.

La première fois qu'un client envoie une requête HTTP à un serveur, il ouvre une connexion TCP avec le serveur (généralement sur le port 80). Il envoie alors sa requête à travers cette connexion, et le serveur envoie en retour sa réponse par la même connexion. Cette connexion TCP peut être maintenue ouverte pour des requêtes ultérieures. Ce que spécifie HTTP est la structure des messages que le client et le serveur vont échanger à travers la connexion TCP.

URL

Les requêtes HTTP sont, de manière générale, déclenchées lorsqu'un utilisateur clique sur un lien hypertexte ou saisit une adresse dans son navigateur. Dans un cas comme dans l'autre, la requête est construite à partir d'une URL, *Uniform Resource Locator*.

Par exemple :

http://fr.wikipedia.org:80/w/index.php?title=Uniform_Resource_Locator&oldid=93045880#URL_absolue

Cette adresse, cette URL, a une structure particulière :

- La première partie, « http:// », spécifie qu'il faut émettre une requête HTTP pour traiter cette URL. Des URL peuvent spécifier d'autres protocoles. On rencontre par exemple sur le web des URL débutant par « mailto: », suivi d'une adresse électronique. Ces URL doivent être traitées pour l'envoi d'un courrier électronique, et cliquer sur un hyperlien pointant vers une URL de ce type déclenche, en principe, l'ouverture du logiciel de messagerie électronique.
- La seconde partie, « fr.wikipedia.org », indique le nom de la machine et le nom du domaine où se trouve la ressource (ici, il s'agit de la machine « fr » dans le domaine « wikipedia.org »), c'est-à-dire la machine avec laquelle le client doit établir une connexion TCP. Avant d'établir cette connexion, le client doit faire appel à un serveur DNS qui lui transmettra l'adresse IP correspondante.
- La troisième, « :80 », spécifie le port vers lequel la connexion TCP doit être établie. L'indication du port est facultative. S'il n'est pas mentionné, c'est par défaut sur le port 80 que la connexion se fait. Mais lorsque le serveur est configuré pour accepter les connexions HTTP sur un autre port que le port 80 (on rencontre par exemple le port 8080 lorsque le serveur repose sur la technologie Java), cette information doit être précisée.
- La quatrième partie de l'URL, le « chemin d'accès », ici « /w/index.php », donne un chemin dans une hiérarchie similaire à celle des fichiers sur un disque dur. Si, aux débuts du web, ce chemin indiquait en effet la localisation d'un fichier HTML sur le disque dur d'un serveur, cela est beaucoup moins vrai aujourd'hui. Les ressources ne sont pas nécessairement des fichiers préexistants à la requête, mais sont bien souvent

au contraire calculées dynamiquement par le serveur. Le chemin est alors au mieux configurable, au pire imposé par une technologie chargée de traiter la requête et construire la réponse. Dans cet exemple, l'extension du fichier, « .php », indique que le fichier index.php est un programme écrit dans le langage de programmation PHP[5], qui permet précisément de construire dynamiquement des documents HTML.

- La construction de la ressource par un programme exécuté sur le serveur peut faire intervenir des paramètres. Ce sont de tels paramètres qui sont indiqués par la partie suivante de l'URL, la « chaîne d'interrogation » : « ?title=Uniform_Resource_Locator&oldid=93045880 ». Il y a ici deux paramètres, le premier ayant pour nom « title » et pour valeur « Uniform_Resource_Locator », le second ayant pour nom « oldid » et pour valeur « 93045880 ». Dans cet exemple, le programme « /w/index.php » construit la version passée d'une page Wikipédia, la version ayant pour identifiant « 93045880 » de la page ayant pour titre « Uniform Ressource Locator ».
- La dernière partie de l'URL, « #URL_absolue », désigne un « fragment » de la ressource. La réponse du serveur à une requête à cette URL sera la totalité de la ressource, mais le navigateur focalisera l'affichage sur le fragment ayant pour identifiant « URL_absolue ». Dans cet exemple, la section « URL absolue » de la page Wikipédia.

Structure des requêtes et des réponses HTTP

Les requêtes et réponses HTTP sont constituées de deux parties, un en-tête qui contient des informations propres au protocole HTTP, et un corps qui contient les données échangées (par exemple, le code HTML de la page renvoyée par le serveur).

---

5. PHP est un acronyme récursif signifiant « PHP: *Hypertext Preprocessor* ».

L'URL contient une grande partie des informations nécessaires à la construction d'une requête. Le nom de domaine et le port sont utilisés pour établir la connexion TCP, tandis que le chemin d'accès et la chaîne d'interrogation font partie des informations constituant l'en-tête de la requête.

Si, comme l'illustre l'exemple ci-dessus, HTTP permet à un client de réclamer une ressource à un serveur, d'autres types de requêtes sont possibles. Le type d'une requête est identifié par la « méthode HTTP», une autre information que l'on va trouver dans l'en-tête d'une requête HTTP. Ces méthodes sont au nombre de neuf, mais les navigateurs web n'utilisent typiquement que deux d'entre elles. La « méthode GET » est celle qui serait typiquement utilisée à partir de l'URL utilisée en exemple ci-dessus. Cette méthode indique au serveur que le client demande une ressource. À l'inverse, il arrive que le client souhaite transmettre des informations au serveur plutôt que d'accéder à une ressource particulière. C'est le cas, par exemple, lorsqu'un utilisateur remplit un formulaire pour se créer un compte sur un service en ligne. La « méthode POST » est alors utilisée par le client qui insère alors les informations fournies pas l'utilisateur (son nom, son mot de passe, son adresse de courrier électronique, etc.) dans le corps de la requête. Le serveur a alors pour tâche de traiter ces informations afin qu'elles puissent servir lors des futures interactions entre l'utilisateur et le service (par exemple, lorsqu'il s'identifiera).

Outre la méthode, le chemin d'accès et la chaîne d'interrogation, l'en-tête HTTP est également constitué d'un ensemble de paramètres. Ceux-ci permettent, entre autres choses, de transmettre les identifiants de l'utilisateur ou de

préciser la requête en spécifiant, par exemple, le format ou la langue de ressource attendue en réponse. Un de ces paramètres est le « cookie ». Un cookie, ou témoin de connexion, est une suite d'informations envoyée par un serveur au client dans une réponse HTTP, et que le client renvoie systématiquement lors des requêtes ultérieures à ce même serveur. Les cookies permettent aux serveurs d'enregistrer des informations sur l'utilisateur à l'origine d'une requête et de les réutiliser lors des requêtes suivantes. Ces informations peuvent être des préférences de l'utilisateur quant au fonctionnement ou à l'aspect du site, ou encore le fait qu'il s'est déjà identifié. Elles assurent une continuité dans le traitement d'un utilisateur.

Le premier élément de l'en-tête d'une réponse HTTP, le « statut », indique la manière dont le serveur a traité la requête du client. Par exemple, le statut « 200 OK » indique que la requête a été correctement traitée : le corps de la réponse contiendra alors la ressource demandée par une requête GET ou un document attestant que les informations transmises par une requête POST ont bien été traitées. À l'inverse, le statut « 404 Not Found » indique que la ressource désignée par l'URL n'a pas pu être fournie par le serveur, soit parce qu'il ne l'a pas trouvée, soit parce qu'il n'a pas su la construire.

L'en-tête contient également des métadonnées sur la ressource renvoyée : son format, sa taille, sa langue. L'entête peut aussi comprendre un cookie, que le client enregistrera pour des requêtes futures et des informations de redirection, indiquant au navigateur qu'il doit se rendre sur une autre page, c'est-à-dire formuler une nouvelle requête.

D'autres protocoles basés sur Internet : SSH et FTP

Si le web et le courrier électronique sont les applications les plus populaires d'Internet, ce ne sont pas les seules. Deux autres sont présentées ici, l'échange de fichiers au moyen du protocole FTP (*File Transfert Protocol*) et la connexion à distance en ligne de commande avec le protocole SSH (*Secure Shell*).

L'échange de fichiers par le protocole FTP

Le principe du protocole FTP est de permettre à un client de parcourir l'arborescence de fichiers d'un serveur et de transférer des fichiers du serveur vers le client et inversement. FTP permet aussi de créer et de supprimer des répertoires sur le serveur et d'y supprimer des fichiers.

Pour se connecter à un serveur FTP, un utilisateur a besoin d'un client FTP (par exemple Filezilla) auquel il spécifie le nom du serveur (qui sera traduit en adresse IP par un serveur DNS), le port (par défaut 21), son identifiant et son mot de passe. L'identification de l'utilisateur permet au serveur de déterminer quels sont les droits de lecture et d'écriture de l'utilisateur pour chaque répertoire et fichier de l'arborescence. Certains serveurs acceptent des connections anonymes.

La connexion à distance en ligne de commande avec le protocole SSH

Les systèmes d'exploitation de la famille UNIX (tels que Linux ou Mac OS) proposent, en plus de leur interface gra-

phique, une interface en ligne de commande[6]. Bien que purement textuel, ce mode d'interaction avec l'ordinateur se révèle souvent plus puissant que l'interface graphique. Il est donc intéressant de pouvoir se connecter en ligne de commande à une machine distante. C'est ce que permet de faire le protocole SSH.

Comme avec le protocole FTP, l'utilisateur qui souhaite utiliser le protocole SSH doit avoir un compte sur le serveur auquel il souhaite se connecter. À la différence de HTTP et de FTP, pour lesquels il existe des logiciels clients avec une interface graphique (les navigateurs web en sont l'exemple le plus évident), une connexion SSH ne s'établit qu'à partir de la ligne de commande[7], à l'aide de la commande « ssh ». Pour Windows, il est nécessaire d'utiliser un client spécifique, tel que PuTTY.

Voici un panorama des protocoles d'échange de données sur Internet. Le respect de ces standards a rendu possible la construction d'un réseau complexe comme le web au sein duquel un nombre très élevé de machines peuvent communiquer. La facilité de la circulation de l'information déterminée par la mise en place de ces conventions est une des bases du modèle actuel de diffusion des contenus.

---

6. Les systèmes d'exploitation Windows aussi, mais il s'agit d'une interface beaucoup plus pauvre.

7. Des clients HTTP et FTP en ligne de commande existent aussi. Ainsi, la commande FTP permet d'initier une connexion avec un serveur FTP et de saisir des commandes directement avec le clavier. *Lynx* est un navigateur web entièrement en mode texte, avec lequel on interagit donc uniquement avec le clavier. La commande cURL est très générique et supporte un grand nombre de protocoles.

CHAPITRE 9

Les formats

Viviane Boulétreau et Benoît Habert

Les contenus numériques sont, par nature, encodés. Pour pouvoir être partagée, une information doit être structurée selon des standards : les formats. Le choix d'un format a des implications profondes : les informations que l'on peut transmettre changent, ainsi que leur lisibilité, leur universalité, leur agencement, leur transportabilité, leur transformabilité, etc. Qu'est-ce qu'un format ? Lequel choisir ? Pour quel usage et pour quelle pérennité ? Ce chapitre propose une présentation éclairante de cette notion.

Les formats : invisibles ou pénibles

Il en va des formats informatiques comme des dimensions des roues de nos voitures : nous ne nous y intéressons que lorsque contraints et forcés. Pour les roues, quand il nous faut par exemple leurs dimensions pour acheter des chaînes pour la neige. Pour les formats informatiques, quand « ça ne marche pas ». L'objectif de ce chapitre est de comprendre, essentiellement pour les documents textuels dans la version papier, comment ça marche, pour prévenir les problèmes ou être capable de les résoudre.

Les formats nous sont le plus souvent invisibles ou presque. Lorsque nous cliquons ou double-cliquons sur un nom de fichier dans un dossier, sur une icône de pièce jointe dans un courriel, sur un lien dans une page en ligne, quand nous saisissons une adresse (URL) dans un navigateur, « il se passe quelque chose », qui est le plus souvent approprié. Nous visionnons un film, nous entendons de la musique, nous visualisons une image, nous lisons un texte qui peut d'ailleurs « contenir » du son, de l'image, de la vidéo. Il y a donc un mécanisme qui associe au format de chaque document un outil adapté. Ce mécanisme apparaît d'abord quand l'application (le navigateur, le système d'exploitation) effectue un diagnostic – elle fait l'hypothèse que le document relève d'un format déterminé – et qu'elle nous propose éventuellement d'utiliser un certain outil pour ce document et pour tous les autres du même format. Ce mécanisme apparaît également quand l'application ne trouve pas d'outil adéquat. Elle nous demande alors de chercher sur l'ordinateur utilisé l'outil nécessaire ou de l'installer. Notre ordinateur utilise donc un mécanisme de détection de for-

mat et maintient un « dictionnaire » évolutif associant à un format l'outil à utiliser. On comprend en passant qu'un format peut parfois être utilisé par plusieurs outils.

L'outil qui peut traiter un format change au fil du temps. De nouvelles versions surviennent. Le décalage de version entre celle nécessaire et celle de l'outil peut gêner dans les deux directions. Une ancienne version d'un logiciel peut ne pas prendre en compte un nouveau format. Inversement, une version peut être incapable d'interagir avec un format trop ancien.

Dans la plupart des cas, tout se passe bien. Nous obtenons une interaction normale. Il y a adéquation entre l'outil utilisé et ce qui lui est fourni. Nous pouvons aussi être confrontés à des comportements plus ou moins bizarres, à des dérèglements. L'outil peut enfin se révéler incapable de traiter ce dont il a la charge. Il peut même se « bloquer », voire bloquer l'ordinateur.

Un format : une mise en (bonne) forme de données

Le mot « fromage » vient de « formage », de mettre en forme (une pâte grâce à un moule). On en garde la trace dans le mot « fourme ». De la même manière, un format, c'est quelque chose qui met en forme d'une manière conventionnelle des données destinées à représenter du texte, du son, de l'image, de la vidéo, ou une combinaison des quatre. C'est une sorte de « gabarit » qui met certaines données à des endroits déterminés. Les outils qui vont traiter ce format s'attendent à trouver tel élément à tel endroit, organisé de telle manière, et tel autre à un autre endroit, organisé d'une autre façon.

Prenons l'exemple des pages web. Elles relèvent de la famille de formats HTML. La première ligne d'une page web au format HTML (‹!DOCTYPE html›) indique le choix fait quant à la famille (exemple : HTML5). Le format HTML repose sur des couples de balises ouvrantes – comme ‹title› – et fermantes – comme ‹/title› – qui constituent des « boîtes » attendant chacune un certain type d'information. Ce qui est dans la boîte « title » est utilisé ainsi par le navigateur pour donner un titre à l'onglet dans lequel est affichée la page. Ce qui est dans la boîte « em » est rendu par des italiques. Une boîte peut contenir une ou plusieurs autres boîtes : la boîte « HTML » contient les boîtes « head » et « body ». Une boîte peut contenir également du texte, ou bien être vide : il n'y a alors qu'une seule balise, avec une oblique avant le chevron fermant. La balise ouvrante d'une boîte peut comporter une ou plusieurs associations, marquées par le signe =, entre un attribut et une valeur. Ces associations permettent de spécifier davantage l'utilisation à faire des informations de la boîte. Le format HTML autorise seulement un certain nombre de boîtes et certains types d'inclusion des boîtes entre elles : la boîte « head » ne peut pas contenir une boîte « body ». La conformité avec le format HTML passe par trois volets, de plus en plus exigeants. Si le document est bien fait de boîtes incluses les unes dans les autres, sans « débordements », avec une seule boîte globale, le document est dit « bien formé ». *A contrario*, si la boîte « title » n'est pas fermée, toutes les autres boîtes du document y sont incluses et le navigateur n'arrive pas à afficher quoi que ce soit. Si le document emploie les bons types de boîtes dans les bonnes relations d'inclusion, c'est

une « page HTML valide ». Si les contenus des boîtes correspondent à leur fonction, c'est une « page HTML cohérente ». Ce ne serait pas le cas si était mis dans la boîte « title » le contenu de la page.

L'outil qui traite un document censé relever d'un certain format commence par vérifier que le format est effectivement respecté. De très petites erreurs de mise en forme peuvent rendre le document inutilisable par l'outil qui le signale ou qui « abandonne la partie » plus ou moins doucement. Inversement, certains outils sont plus tolérants. C'est le cas des navigateurs qui doivent faire face à la maîtrise plus ou moins grande des formats HTML. Ainsi, la modification de la boîte englobante de HTML en HTM ne trouble pas certains navigateurs qui affichent le même résultat.

Les formats : des clauses cachées au contrat explicite – standards et normes

On distingue plusieurs catégories de formats :

- Les « formats propriétaires » : leurs spécifications techniques sont contrôlées par une entité privée et ont en général fait l'objet d'un brevet. Leur usage est donc limité.
  - Ces spécifications ne sont pas diffusées, on parle alors de « format opaque » ; les données ne peuvent donc être utilisées que par leur application d'origine, ce qui pose de nombreux problèmes de compatibilité et de portabilité.
  - Ces spécifications ont été publiées, mais sont associées à des autorisations d'utilisation liées aux brevets. Dans ce cas, les données sont la plupart du temps exploitables par d'autres applications, mais le risque de perte d'information existe et ne doit pas être négligé.
- Les « formats libres ou ouverts » : leurs spécifications techniques sont publiques et il n'y a pas de restriction d'accès

ou de mise en œuvre. Chaque éditeur de logiciel peut donc librement proposer les modules permettant de lire ou d'écrire des données selon ces formats.

L'ancien format .doc de Microsoft Word relevait des formats propriétaires opaques, l'actuel format .docx de Microsoft Word ou le format PDF (Adobe) sont des formats propriétaires publiés, tandis qu'HTML est un format ouvert, spécifié par le consortium qui gère le web, W3C.

Un format occupe en fait un des stades du processus de normalisation mis en œuvre au sein de chaque communauté. À partir des bonnes pratiques observées et des « standards de fait » que l'on voit régulièrement émerger, des instances collégiales spécialisées (comités techniques) s'organisent afin d'élaborer un ensemble de référentiels communs : des « normes explicites ». Dans le domaine du document numérique, trois instances principales assurent la gestion des processus de normalisation : l'International Organization for Standardization (ISO), organisme international composé des représentants des organisations nationales (Standards Council of Canada – Conseil canadien des normes ; AFNOR pour la France) ; l'Organization for the Advancement of Structured Information Standards (OASIS), centrée sur la normalisation des formats de fichiers ; le World Wide Web Consortium (W3C), centré sur le web. À l'inverse, certains types de documents ne disposent pas (encore) de formats partagés facilitant l'échange et la reprise. C'est le cas des blogs comme Drupal ou WordPress.

L'exemple le plus représentatif de format ayant fait l'objet d'une procédure de normalisation est PDF. Né en 1993, le *Portable Document Format* de la société Adobe Systems avait

l'avantage de préserver la mise en page des documents, quelle que soit la plateforme de lecture, son système d'exploitation, etc. Il s'agissait, alors, du seul format ayant cette propriété. Il offrait également des options de sécurisation qui, à une époque où la circulation et le partage de documents n'étaient pas aussi naturels qu'ils le sont devenus, l'ont rendu très attractif. La politique commerciale d'Adobe consistant à distribuer gratuitement l'outil de lecture, à commercialiser à des tarifs très raisonnables les applications permettant de générer des fichiers au format PDF et à autoriser des applications tierces à utiliser – gratuitement – le format, a fait le reste. Ce format est devenu le « standard de fait » pour l'échange de documents. À partir des années 2000, quatre sous-ensembles du format PDF ont fait l'objet d'une normalisation par l'ISO, dont PDF/A dans une perspective de pérennisation.

Dans ces processus de normalisation, XML occupe un rôle central. Ce n'est pas un format, mais un métalangage qui permet de définir pour un ensemble de documents donné la « forme » qu'ils doivent suivre : les types d'informations possibles (les boîtes) et les relations entre eux (les relations d'inclusion ou de succession). Cette forme se matérialise par des balises, comme pour HTML. XML permet d'associer à la définition d'un format des outils de validation qui permettent de certifier qu'un document suit bien les conventions du format. Recourir à XML, c'est donc expliciter le contrat que constitue un format et permettre de vérifier son respect. Les formats OpenDocument et Office Open XML des suites bureautiques d'Open Office et de Microsoft Office reposent ainsi sur XML.

Documents autonomes ou dépendants, documents composites

Un document réalisé sous un certain format peut comporter tous les éléments pour être « reconstruit » sur un autre poste de travail. Il peut également faire référence à des « ressources ». Si celles-ci ne sont pas transmises avec le document, la reconstruction de ce dernier va être plus ou moins dégradée. C'est par exemple le cas si le fichier HTML est transmis sans la feuille de style qui indique la mise en page à appliquer et qui est appelée dans une des lignes de code. Ou dans un document PDF, lorsque les polices de caractères n'ont pas été intégrées au moment de sa fabrication et que le poste de travail sur lequel est visualisé le document ne dispose pas de la police appropriée. Les caractères manquants sont alors remplacés par d'autres caractères. L'auteur, s'il n'est pas dûment averti de l'existence de ressources secondaires et de leur caractère nécessaire, risque fort de ne pas les sauvegarder, d'omettre de les transmettre et de rendre son document peu lisible.

Contrairement à l'approche initiale consistant à fusionner en un seul fichier, sous un seul format, des constituants de nature différente, les outils actuels privilégient un agrégat de formats élémentaires dédiés à chaque ressource composant un document. Le document lui-même se trouve alors transformé en *container*, permettant de regrouper et de manipuler un ensemble de ressources hétérogènes : textes, images, éléments de mise en forme, sons, objets mathématiques, etc. C'est le cas des formats OpenDocument, Office Open XML, ePub, PDF… Ces différents formats sont en général massivement basés sur XML et ils reposent sensi-

blement sur les mêmes formats standard pour le stockage des ressources élémentaires. La plupart utilisent la notion d'archive (format ZIP, RAR…) pour regrouper en un seul fichier l'ensemble des ressources qui constituent le document. Enfin, un certain nombre d'éléments communs se retrouvent quel que soit le format envisagé : un groupe de métadonnées, le plus souvent exprimées en XML, donnant des informations de type documentaire, des informations liées à l'application, mais également liées aux licences d'utilisation du document ; un catalogue listant les ressources composant le document en donnant leur typologie (éventuellement complétée d'informations applicatives) ; une structure décrivant l'organisation des ressources les unes par rapport aux autres en matière d'ordonnancement et de hiérarchie. Ces nouveaux formats préservent, pour chaque constituant, le moyen de stockage le plus riche possible. Le recours à XML et à la validation systématique par rapport à des modèles standard confère un caractère pérenne à chaque ressource prise indépendamment. La migration d'un format à l'autre se réduit à une somme de migrations élémentaires.

Jeux de caractères : les formats des données textuelles

Un fichier numérique est une suite de *bits*, c'est-à-dire de 0 et de 1. Le format spécifie comment interpréter cette suite de 0 et de 1. Un texte est d'abord une suite de caractères. Le flux de bits est découpé en *octets*, en séquences de 8 bits. Un octet correspond à 2^8 combinaisons possibles de 0 et de 1, c'est-à-dire à 256 nombres en notation binaire (de 0 – 00000000 – à 255 – 11111111). Chaque nombre est interprété

comme un numéro d'ordre dans un jeu de caractères. Le *a* minuscule a le numéro 95 (1100001). Les caractères consécutifs ont des numéros consécutifs (96 pour *b*, etc.). Certains caractères sont « invisibles », comme les deux qui peuvent servir au changement de ligne et qui sont empruntés au fonctionnement des machines à écrire : passage à la ligne (faire tourner le rouleau d'un cran – *line-feed* – caractère 10 – symbolisé par \l) et retour-chariot (ramener le rouleau à son point de départ – *carriage-return* – 13 – \r). Le premier jeu de caractères standardisé, en 1963, est l'ASCII (*American Standard Code for Information Interchange*). Il note (code) 128 caractères sur 7 bits (le 8e est à 0). Comme son nom l'indique, il permet de coder l'américain et d'échanger des documents en américain ou en anglais. Il est insuffisant pour noter les langues comprenant des caractères autres, comme les caractères accentués français. Une deuxième étape a été d'utiliser 256 positions (8 bits), en gardant l'ASCII pour les 128 premières, et en utilisant le reste pour d'autres caractères. Mais comme 256 positions ne suffisent pas à toutes les langues occidentales, les 128 positions autres que l'ASCII ont donné lieu à des jeux de caractères reliés mais partageant le noyau ASCII. C'est la famille ISO 8859, standardisée par l'ISO. La branche ISO 8859-1 ou Latin 1 permet de noter le français (sauf pour les caractères *Œ*, *œ*, *Ÿ*, et le signe €) et d'autres langues occidentales. La nécessité de pouvoir échanger en mêlant des textes de plusieurs langues et en intégrant des langues aux jeux de caractères très larges (comme le japonais ou le chinois) a conduit plus récemment à la mise au point du standard Unicode. Il utilise 1 114 112 positions possibles et représentait en janvier 2012 une centaine de scripts qui totalisaient 100 181 carac-

tères. Les scripts sont des collections cohérentes de caractères en usage dans un domaine particulier. Ils incluent les symboles monétaires, les opérateurs mathématiques, le braille, etc. À chaque caractère sont associées des propriétés : opposition majuscules/minuscules ou « casse » ; place dans le tri – le é en français doit être trié avec le e et non mis après le z ; direction d'écriture, de gauche à droite ou de droite à gauche. Unicode reprend comme noyau Latin 1 et donc aussi l'ASCII. Pour représenter le million de positions possibles en « économisant » les octets, on a souvent recours au format UTF, qui utilise de 1 à 4 octets pour fournir la position d'un caractère. Les caractères les plus fréquents, ceux de l'ASCII, sont codés sur un octet. Les caractères accentués du français sont codés sur 2 octets. On le comprend par exemple lorsqu'un fichier HTML obéit au format UTF-8 tandis que le navigateur attend de l'ISO Latin 1. Les lettres accentuées sont alors remplacées par deux caractères « bizarres ». À l'inverse, lorsque le fichier HTML est en ISO Latin 1 et que le navigateur attend de l'UTF-8, celui-ci remplace les lettres accentuées par une marque conventionnelle de « gêne ». On peut faire en sorte que le navigateur utilise un autre jeu de caractères que celui défini dans ses préférences ou changer ces dernières.

On parle parfois de format texte (seul), c'est-à-dire d'un ensemble de caractères sans indications de mise en forme. Pour pouvoir utiliser un fichier au format texte, il faut connaître le jeu de caractères qu'il utilise et aussi la manière dont il matérialise les changements de ligne. Historiquement, le monde Windows utilisait la suite retour-chariot/passage à la ligne, tandis que l'univers Mac se contentait du seul retour chariot, et le monde Linux/Unix du passage à la

ligne. Confrontés à un fichier texte seul, les traitements de texte actuels proposent souvent, si nécessaire, de convertir le fichier pour qu'il utilise les conventions qui sont les leurs. Un autre format textuel courant est CSV pour *Comma Separated Values*. C'est une manière au départ de représenter des tableaux de nombres en séparant les colonnes par des virgules (*commas*), puisqu'en anglais le point est le séparateur décimal, et les lignes par des changements de ligne. Pour le français, c'est alors souvent la tabulation (caractère 9 – \t) qui sépare les colonnes, la virgule étant le séparateur décimal. Les tableurs importent ou exportent des données au format CSV.

Quand on copie une portion de document dans une application (navigateur, par exemple) et qu'on la colle dans une autre (traitement de texte ou courriel, par exemple), via une zone d'échange appelée « tampon », le format de ce qui est copié peut être plus ou moins conservé. Si le maintien du format compte, on utilisera plutôt les fonctions de conversion explicite (en import ou en export) des logiciels en question.

Quels formats pour quels usages ?

Les formats sont le plus souvent invisibles : nos actions suffisent en général à mobiliser ceux qui nous servent. Ils sont en fait *trop* souvent invisibles : le choix est fait « à notre insu », sans qu'aient toujours été pesées les contraintes à respecter en fonction de l'usage à faire du document en question, maintenant et plus tard.

- « Consommation » ou modification : le document doit-il seulement être affiché ? Doit-il également être imprimé et, si oui,

la mise en page doit-elle impérativement être conservée ? Lors des utilisations futures, est-il supposé être modifié ? Seulement par son auteur ou par d'autres personnes ? Si oui, doit-on garder un historique de ces modifications ? PDF est préférable dans les deux premiers cas, OpenDocument ou .docx pour les suivants, avec les possibilités de suivi de révision qu'ils offrent.

- Public visé : à qui est destiné le document ? L'auteur sera-t-il son seul utilisateur, ou bien d'autres personnes seront-elles amenées à le manipuler ? Les futurs utilisateurs sont-ils *a priori* connus ou non ? Connaît-on la nature des dispositifs (matériel et applications) que le public visé utilisera, connaît-on son degré de maîtrise de ces outils ? Par exemple, il est tout à fait normal d'échanger des fichiers sous un format très spécifique destiné à une application particulière avec des collègues de travail dont on sait qu'ils sont équipés pour les exploiter. Inversement, un CV à un futur employeur devra impérativement être transmis sous le format le plus répandu et le plus simple d'utilisation possible.
- Mode de transmission : le vecteur utilisé pour transmettre (réseau : site web, FTP, courriel ; support physique mobile : disque, clé USB) peut amener à privilégier un format par rapport un autre. Un document PDF est ainsi souvent plus « léger » que le document Word ou PowerPoint source.
- Durée de vie du document : pendant combien de temps doit-il être utilisable ? Quelques jours ? Quelques semaines ? Doit-il être exploitable sur du plus long terme comme un document administratif pour lequel on considère qu'une dizaine d'années est la durée de vie « légale » ? Dans ce cas, on utilisera le format PDF/A conçu dans cette optique.

Nos objectifs, une fois correctement énoncés, nous permettent la plupart du temps de choisir une famille de formats. Les contraintes techniques, économiques, les choix politiques, les usages d'une communauté suffisent à affiner

ce premier tri. Aussi ne sommes-nous pas toujours aussi libres que nous pourrions le souhaiter. En nous plaçant délibérément dans un contexte moins contraint, deux critères complémentaires peuvent nous aider à choisir le format adéquat.

- Interopérabilité : elle est essentielle non seulement si l'objectif est de partager un document avec d'autres utilisateurs, mais également dans une perspective de préservation sur la durée, puisqu'elle permettra d'exploiter un document avec d'autres dispositifs si celui d'origine devenait obsolète. Deux approches sont possibles : l'une basée sur le dispositif consiste à vérifier que, même s'il fonctionne de préférence avec un format qui lui est propre, il est capable de lire et de produire des formats dits d'échange (comme CSV pour les tableaux, PDF, etc.), en garantissant une perte d'information minimale. Cette approche est généralement privilégiée lorsqu'il s'agit de documents spécifiques et que les outils mis en œuvre disposent de fonctionnalités très particulières (publication assistée par ordinateur – PAO –, dessin technique, outils de traitement du signal, etc.). La seconde consiste à choisir d'abord un format gage d'interopérabilité, puis le dispositif qui permettra de le manipuler. Un tel format devra être au moins un standard *de facto*, voire une norme. On privilégiera bien entendu les formats les plus ouverts possible afin de préserver un minimum de liberté dans le choix du dispositif.
- Pérennité : cette préoccupation doit être présente à l'esprit dès la création du document. L'utilisation de formats et de dispositifs disposant d'une solide communauté d'utilisateurs est également un facteur à prendre en compte pour la gestion à long terme des documents. En effet, au-delà du choix du format initial, la préservation des documents suppose une veille constante afin de pouvoir opérer les migrations de support, de format ou d'application chaque fois que l'obsolescence de l'un de ces éléments est signalée.

Les contrats que constituent les formats s'insèrent dans des relations de confiance plus ou moins grandes qui lient éditeurs de logiciels et utilisateurs. La connaissance fine des types de contrats possibles et de leurs enjeux nous permet de contribuer à bon escient à ces échanges.

CHAPITRE 10

L'organisation des métadonnées

Grégory Fabre et Sophie Marcotte

Nous avons déjà insisté sur l'importance des métadonnées. Nous pouvons affirmer que se joue là un des enjeux fondamentaux de la structuration des contenus. Selon notre façon de baliser les informations, nous les rendons accessibles, visibles et exploitables, mais nous construisons aussi une taxinomie du monde. Les métadonnées constituent un outil technique destiné à rendre compte de notre vision du monde et de la structuration de l'ensemble de nos connaissances. C'est parce qu'elle a produit une taxinomie précise et vaste que l'œuvre d'Aristote a autant marqué notre culture occidentale. L'enjeu est donc très important aujourd'hui, lorsque cette taxinomie est destinée à être reprise et généralisée dans tous les domaines de la connaissance – de la description des relations entre individus, décrite par les métadonnées d'un réseau social comme Facebook, à la hiérarchisation des informations bibliographiques du Dublin Core. Ce chapitre illustre, de manière très pratique, les principales caractéristiques des métadonnées et les fonctionnalités automatiques qu'elles permettent.

Mise en place d'une ontologie

Les technologies actuelles permettent de plus en plus aux chercheurs d'archiver, de diffuser et d'échanger aisément sur l'évolution de leurs recherches. Cela répond d'ailleurs à une nouvelle réalité dans le milieu de la recherche : celle de la création d'outils de diffusion mieux adaptés au transfert et à la circulation des connaissances. Pour ce faire, il importe de concevoir des outils et d'élaborer des pratiques et des protocoles communs.

La diffusion sur support numérique de textes ou de résultats de recherche fragmentaires ou achevés implique un processus dont la complexité devient parfois un obstacle difficile à surmonter pour les chercheurs habitués à publier leurs travaux en passant par les supports imprimés traditionnels (livres, revues, journaux, etc.). Le web n'est pas encore, en effet, le média de publication le plus employé pour la publication d'articles savants, d'édition critiques ou de textes de fiction, si bien que la plupart des protocoles restent à être définis et, surtout, à être apprivoisés par les chercheurs. Or il faut s'interroger sur la manière de faire évoluer les pratiques sans réduire la portée des contenus, sur la manière de changer les processus sans travestir les résultats, tout en insistant sur la valeur ajoutée des différentes pratiques inhérentes à l'édition et à la publication numériques.

Si l'un des objectifs du web 2.0 a été de simplifier un usage parfois complexe d'Internet, en se souciant notamment de l'ergonomie des interfaces et en plaçant l'utilisateur (non spécialiste) au cœur des enjeux – l'utilisateur partage, contribue, commente –, les visées du web 3.0 (ou

web sémantique) sont autres. Il s'agit désormais de penser le web en termes de structuration et de programmation des métadonnées. Pour le dire autrement, il y a passage d'une version du web horizontale, où la catégorisation est mise à plat (par exemple, le *hashtag*), à une version hiérarchisée et informée, où le relief de la structuration met en évidence auprès des autres systèmes sémantiques la nature du contenu diffusé.

Dans ce contexte, les modalités de recherche deviennent également un enjeu crucial. Le SEO (*Search Engine Optimization*, optimisation pour les moteurs de recherche) a longtemps été l'apanage des compagnies privées qui promettent à leurs clients un meilleur positionnement au sein des moteurs de recherche les plus populaires et performants. Face à la pléthore de sites indexés chaque jour par ces derniers, il convient d'adopter des stratégies visant à ne pas voir son site noyé dans un océan de références, surtout lorsqu'il s'agit d'initiatives de recherche ou à visée éducationnelle. Dès lors, la métacatégorisation du contenu – le fait, autrement dit, de rendre intelligibles les contenus publiés sur son site – est à la base des très nombreuses stratégies possibles pour ne pas pâtir des *crawlers* de Google, Yahoo ! et consorts programmés pour indexer (donc, classer).

Rendre une donnée intelligible, partageable et catégorisable nécessite une conception et un regard communs sur l'objet à qualifier. En d'autres termes, le systématisme imposé par la machine n'est pas toujours évident à faire cohabiter avec la réalité d'une recherche en cours. En effet, la catégorisation implique une compréhension globale et préalable du sujet étudié. Comment parvenir à un consensus

permettant l'élaboration d'un système dans un contexte aussi mouvant ?

Le choix des outils occupe à cet égard une place prépondérante. Ces derniers doivent permettre la flexibilité imposée par l'évolution constante des recherches. En effet, les catégories dans lesquelles les contenus sont « classés » doivent être redéfinies régulièrement. Ainsi, l'environnement de publication en ligne (tel Drupal, qui sera évoqué dans les pages qui suivent) aura nécessairement besoin d'être pourvu d'outils appropriés pour faire face à cette mouvance des contenus inhérente à la recherche. En ce sens, le module *Taxonomy Manager* est une interface très performante pour la gestion d'ontologies. Il permet entre autres l'ajout de nouveaux termes, mais également la fusion de plusieurs termes, ainsi que le déplacement d'un terme au sein d'une autre catégorie (« vocabulaire »).

Une ontologie informatique est un ensemble de termes et concepts réunis pour décrire un domaine ou un objet d'étude. Structurée sous une forme hiérarchisée (parent/enfant ou vocabulaire/terme), son utilisation permet d'informer, par des métadonnées, le contenu diffusé et de renseigner sur la nature des différents champs composant un site web. Une telle structuration s'explique à la fois par la volonté d'établir une interrelation féconde entre deux systèmes (par exemple, un site et un moteur de recherche de type Google), mais permet également d'élaborer une circulation de l'information efficace au sein d'un seul et même système. Des relations croisées entre différents vocabulaires ou termes d'une ontologie peuvent ainsi être envisagées afin d'approfondir une requête au sein d'un environnement de recherche propre à un système.

Les formats sémantiques

Différents formats et langages de description des données permettent d'organiser et de partager efficacement de l'information dans l'environnement du web. Parmi ceux-ci, on trouve Microformat et Microdata, qui sont des langages permettant la structuration des données à partir de balises HTML (*HyperText Markup Language*) existantes.

Microformats

Les microformats, aussi connus sous le terme « entités », sont des conventions essentiellement conçues pour la description d'informations précises comme le partage d'événements, de contacts ou de précisions géographiques. Les entités possèdent toutes leurs propres propriétés. Par exemple, un événement sera défini selon les propriétés « date », « lieu », « type d'événement », « heure », « contact », etc. La cohabitation des microformats avec les autres formats du web sémantique, notamment le RDF, duquel il sera question plus loin, n'est toutefois pas toujours efficace.

Exemple :

```
<address class="vcard">
<p>
<span class="fn">Charlie Temple</span><br/>
<span class="org">Laboratoire NT2</span><br/>
<span class="tel">514-xxx-xxxx</span><br/>
<a class="url" href="http://figura-concordia.nt2.ca/">Site
web</a>
</p>
</address>
```

Utilisé pour établir des profils d'individu contenant une somme d'informations limitée, les microformats sont communément employés dans le but de constituer des annuaires de personnes. Dans l'exemple ci-dessus, on peut distinguer la nature des différents champs composant une « vcard » (standard de carte d'affaire électronique). Ils viennent plus précisément offrir de l'information sur le profil de la personne au travers de son organisation ou de son numéro de téléphone, par exemple.

Microdata

Le format Microdata permet quant à lui de créer des liens sémantiques entre des contenus déjà présents sur le web en ajoutant des balises à la structure HTML. Des navigateurs web ou des moteurs de recherche comme Google, Bing et Yahoo !, entre autres, sont en mesure d'extraire les contenus en Microdata des sites web afin de mettre certaines informations en évidence sur leur propre site et ainsi fournir aux utilisateurs les résultats de recherche les plus pertinents.

Le site schema.org recense les différentes balises prévues à cet effet. Il propose en fait une série de marqueurs HTML qu'il est possible d'employer pour le balisage des sites de manière à ce que ceux-ci soient reconnus par les principaux moteurs de recherche, qui s'appuient sur ces conventions pour permettre aux internautes de bénéficier des meilleurs résultats : « *A shared markup vocabulary makes it easier for webmasters to decide on a markup schema and get the maximum benefit for their efforts. So, in the spirit of sitemaps.org, search engines have come together to provide a shared collection of*

schemas that webmasters can use[1].» Il reste en outre possible de programmer des balises spécifiques selon les besoins.

Extrait de la convention schema.org (portion de la hiérarchie d'un fichier type):

Thing additionalType, description, image, name, url
Class
Creative Work about, accountablePerson, aggregateRating, alternativeHeadline, associatedMedia, audience, audio, author, award, awards, comment, contentLocation, contentRating, contributor, copyrightHolder, copyrightYear, creator, dateCreated, dateModified, datePublished, discussionUrl, editor, educationalAlignment, interactionCount, interactivityType, isBasedOnUrl, isFamilyFriendly, keywords, learningResourceType, mentions, offers, provider, publisher, publishingPrinciples, review, reviews, sourceOrganization, text, thumbnailUrl, timeRequired, typicalAgeRange, version, video
Article articleBody, articleSection, wordCount
BlogPosting
NewsArticle dateline, printColumn, printEdition, printPage, printSection

Exemple:

```
<div itemscope itemtype="http://schema.org/Person">
<span itemprop="name">Charlie Temple</span>
<img src="smarcotte.jpg" itemprop="image" />
<span itemprop="jobTitle">Professor</span>
<div itemprop="address" itemscope itemtype="http://
schema.org/PostalAddress">
<span itemprop="streetAddress">
1400, boul. de Maisonneuve Ouest
</span>
```

1. Voir: <http://schema.org/>.

```
<span itemprop="addressLocality">Montreal</span>,
<span itemprop="addressRegion">QC</span>
<span itemprop="postalCode">H3G 1M8</span>
</div>
</div>
```

Les *microdatas* peuvent être envisagés comme une version évoluée des microformats, de par leurs fonctionnalités ainsi que la technologie au sein de laquelle ils s'inscrivent (HTML5). Un des grands intérêts de ce format est de pouvoir être interprété par les principaux moteurs de recherche (Google, Bing, Yahoo !) en affichant de manière distincte les informations émanant de ces derniers. Ainsi, on peut constater dans l'exemple ci-dessus que Google distingue par un affichage différent le nom de la personne de son organisation et de son affiliation professionnelle.

RDF

RDF (*Resource Description Framework*) est un modèle d'information parmi les plus reconnus du web sémantique. Il définit les règles qui relient les informations entre elles. Les documents RDF sont structurés grâce à des ensembles de triplets (sujet, prédicat, objet). Pour être compréhensible par les machines, ce format doit être interprété par les agents logiciels qui échangent de l'information entre eux. Par conséquent, les systèmes (sauf exception) doivent être équipés de telles fonctionnalités pour pouvoir utiliser/interpréter ce format.

Exemple :

```
‹rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
        ‹rdf:Description rdf:about="http://figura.uqam.ca/
membre/marcotte-sophie">
                ‹dc:title›Sophie Marcotte‹/dc:title›
            ‹dc:publisher›Laboratoire NT2‹/dc:publisher›
    ‹/rdf:Description›
‹/rdf:RDF›
```

Ici, on peut constater la manière dont l'information est structurée en RDF grâce à des balises permettant d'accroître la granularité de l'information première (description) et de préciser ainsi la nature du contenu diffusé. Devenus référants, d'autres formats (FOAF, OAI/PMH, Dublin Core, etc.) viennent compléter les termes de base du RDF grâce à un vocable plus précis.

OWL

OWL (*Web Ontology Language*) est un prolongement de RDF. Il s'agit d'un vocabulaire qui permet la définition d'ontologies structurées selon le modèle d'organisation des données de RDF. Ce vocable XML permet de spécifier ce qui ne peut pas être compris d'emblée par la machine en fournissant un langage propice à l'élaboration d'une ontologie.

En ce sens, OWL fournit une aide précieuse pour la gestion et la compréhension des informations par les machines. Différentes versions de OWL (Lite, DL, Full) existent et doivent être utilisées en fonction de la granularité (ou niveau de détails) des hiérarchies souhaitées.

Même si le balisage de ces formats demeure relativement peu standardisé à ce jour, il existe des initiatives importantes qui permettent une certaine uniformisation dans l'organisation des données.

FOAF (Friend of a Friend) – FOAF a été créé au début des années 2000 par Libby Miller et Dan Brickley. Il s'agit d'une ontologie reposant sur le vocabulaire singulier du RDF et de l'OWL qui permet de définir, de manière descriptive, les spécifications des individus, de leurs activités et des relations qu'ils entretiennent avec d'autres personnes ou objets. Par exemple, elle spécifiera le prénom et le nom de l'individu, l'adresse de son site web personnel, la liste de ses réalisations et de ses projets en cours, ses activités professionnelles, etc.

Exemple :

```
‹foaf:Group›
‹foaf:name›Laboratoire NT2 Crew‹/foaf:name›
‹foaf:member›
‹foaf:Person›
‹foaf:name›Sophie Marcotte‹/foaf:name›
‹foaf:homepage rdf:resource="http://figura.uqam.ca/membre/
marcotte-sophie"/›
‹foaf:workplaceHomepage rdf:resource="http://figura.uqam.ca"/›
‹/foaf:Person›
‹/foaf:member›
‹/foaf:Group›
```

Malgré ses nombreux avantages, l'utilisation de ce format reste relativement marginale. FOAF a toutefois été adopté par des communautés en ligne importantes comme WordPress et Identi.ca. Afin de contrer de possibles utilisa-

tions fallacieuses des données diffusées, FOAF permet notamment le cryptage (SHA1 : *Secure Hash Algorithm*) des adresses courriels.

OAI-PMH – Open Archives Initiative Protocol for Metadata Harvesting est un mécanisme favorisant l'interopérabilité entre différentes sources de référence. La structuration des données doit être réalisée par les fournisseurs de données. OAI-PMH est un ensemble de six verbes ou services qui sont invoqués par le biais du protocole HTTP. OAI-PMH peut être considéré comme une couche venant ajouter de l'information supplémentaire aux principales balises RDF. Ce format sera probablement celui sur lequel le Laboratoire NT2 s'appuiera au cours des prochaines années, notamment pour son projet CELL, évoqué en conclusion. Il permettra aux différentes équipes se mêlant au projet de coordonner, et ainsi homogénéiser, leur vocabulaire taxonomique.

Dublin Core – *Dublin Core Metadata* correspond à une liste de métadonnées qui sont liées aux sites web. Celles-ci ont été rassemblées sous la convention DCMI (*Dublin Core Metadata Initiative*), qui comporte notamment une liste officielle de 15 propriétés normalisées[2].

Parrainé, en 1995, par le Online Computer Library Center (OCLC) et le National Center for Supercomputing Applications (NCSA), ce projet a alors réuni 52 chercheurs et experts désireux de faire évoluer la réflexion sur la structuration des métadonnées. L'initiative Dublin Core est notamment née de la volonté de normaliser les différentes balises nécessaires

2. ISO 15836-2003, RFC 5013.

à la métadescription des références bibliographiques, ainsi que d'établir des relations entre ces références et d'autres ressources.

Exemple :

```
<meta name="DC.title" lang="fr" content="De quoi t'ennuies-tu, Évelyne?" />
    <meta name="DC.date" scheme="DCTERMS.W3CDTF" content="1945" />
<meta name="DC.creator" lang="fr" content="Gabrielle Roy" />
<meta name="DC.language" scheme="DCTERMS.RFC4646" content="fr-FR" />
<meta name="DC.description" lang="fr"
        content="Bonheur d'occasion est un roman urbain écrit par Gabrielle Roy et publié en 1945. Il a valu à l'écrivaine le prix Fémina, le premier prix littéraire prestigieux français remporté par un écrivain canadien" />
    <link rel="DC.source" href="urn:ISBN:978-2890525757" />
```

DRUPAL

Tel qu'évoqué précédemment, le web 3.0 implique une structuration méthodique des données dans le but de les rendre potentiellement réutilisables par d'autres sites ou d'autres systèmes. La hiérarchisation de ces données, également appelée « schéma », est particulièrement facile à intégrer au sein d'un système de gestion de contenu (CMS) comme Drupal. En effet, Drupal évite aux programmeurs les processus de « capture » et de gestion des métadonnées. En plus d'être pourvu nativement d'un module RDF permettant le balisage des différents champs des sites web, la communauté propre à Drupal a développé plusieurs modules garantissant le respect des normes de publication.

Drupal est un logiciel *open source* publié sous licence générale (GNU) et maintenu par une très vaste communauté à travers le monde. Dès ses premières versions, le logiciel a été orienté vers la métacatégorisation du contenu hébergé par l'utilisation des outils taxonomiques. La création d'une ontologie taxonomique composée d'un vocabulaire et de termes qui lui sont reliés constitue l'une des bases pour produire, gérer et diffuser efficacement du contenu via cet environnement de développement. Car il ne s'agit plus uniquement, aujourd'hui, de bien structurer son site, mais de faire en sorte qu'il interagisse avec d'autres systèmes en rendant son contenu intelligible pour que les échanges puissent être efficaces, et ce, indépendamment de la langue.

Afin de rendre le dialogue possible entre les systèmes, permettre la recherche et la gestion des données RDF disponibles sur la toile, un autre langage du nom de SPARQL a été créé. La version 1.1 permet à ce langage d'enregistrer et de fusionner des données émanant de sources différentes.

On comprend alors la pertinence d'une telle structuration et d'un tel appareillage pour la constitution, par exemple, d'une communauté virtuelle autour d'un même objet d'étude ou pour les besoins d'une équipe de recherche disséminée géographiquement et ne partageant pas forcément les mêmes systèmes d'archivage et de diffusion.

Les limites du moteur de recherche propre à Drupal entraînent la nécessité de convoquer un autre moteur plus puissant, qui puisse s'intégrer de manière harmonieuse à la logique de ce CMS. Apache Solr est l'une des solutions possibles. Solr a été créé par la fondation Apache qui distribue l'ensemble des logiciels qu'elle développe sous licence libre.

Programmé en Java[3], un langage orienté vers l'objet, Apache Solr est doté d'une interface en HTML permettant de faciliter la gestion de ses fonctionnalités. La prise en compte des facettes – les critères pour le filtrage des données – générées par les taxonomies et la structure de Drupal (les types de contenus, notamment) font de cet outil un compagnon idéal pour la recherche avancée au sein d'une base de données complexe.

Dans une utilisation couplée de Drupal et d'Apache Solr, les facettes présentent une double utilité. Jouant avant tout le rôle de métadonnées, les facettes informent le contenu de la teneur du message dont il est porteur; elles vont également fournir au moteur de recherche des éléments servant à filtrer le contenu accumulé (mais « facetté ») dans la base de données. Dans le même sens, la prise en compte par Apache Solr du format RDF fait de cet outil un incontournable de l'organisation des métadonnées sous Drupal.

En outre, l'indexation partagée constitue un moyen idéal pour indexer l'ensemble des sites appartenant à un même écosystème. L'objectif alors poursuivi est de pouvoir effectuer une recherche de manière transversale sur l'ensemble des sites. Pour y parvenir, l'utilisation d'une ontologie commune est nécessaire. D'autres ontologies, plus spécifiques et propres à chacun des projets de recherche, peuvent venir s'y greffer et ainsi densifier les informations premières. Une réflexion ayant pour finalité l'établissement de passerelles sémantiques afin d'intégrer ces nouveaux fragments taxonomiques s'impose alors d'emblée.

---

3. James Gosling *et al.*, *The Java Language Specification*, Boston, Addison-Wesley, 2005.

Un cas particulier : le TEI

Le TEI, pour *Text Encoding Initiative*, a été créé en 1987 par un consortium formé de trois sociétés savantes (ACH, Association for Computers and the Humanities ; ACL, Association for Computational Linguistics ; et ALLC, Association for Literary and Linguistic Computing). L'objectif premier du TEI, qui était à l'origine fondé sur le langage SGML, était de « faciliter la création, l'échange et l'intégration de données textuelles informatisées[4] ». Désormais lié à l'emploi de balises XML, les conventions élaborées dans le cadre du TEI visent à permettre la description de la manière dont un document a été créé ainsi que la façon dont il a été structuré : pages, paragraphes, lignes, chapitres, dialogues, soulignements, ajouts marginaux, ratures, etc. C'est pourquoi les protocoles proposés par le TEI sont surtout adoptés dans le cas de projets d'édition de textes anciens, de manuscrits ou de dossiers génétiques, où il s'agit parfois de reconstituer le processus ayant mené à la version définitive d'un texte. Toutefois, puisque le TEI ne constitue pas un modèle entièrement adapté au contexte du web sémantique, ces protocoles doivent impérativement être accompagnés du RDF pour permettre la mise en relation des sources, afin d'éviter que les textes ne se trouvent insularisés.

En somme, comme ce bref parcours aura permis de le constater, s'intéresser à l'organisation des métadonnées dans le contexte du web sémantique implique de connaître les

4. Lou Burnard, « Le schéma XML TEI pour l'édition », Université d'été de l'édition électronique ouverte, 2009 : ‹http://leo.hypotheses.org/2630› (document audio).

potentialités des différents formats et langages tels Microformat, Microdata, RDF et OWL, qui font autorité dans le domaine. Certaines pratiques, on l'a vu, répondent davantage aux besoins de professions ou d'activités spécifiques, notamment dans les domaines de l'édition et de la bibliothéconomie. Les figures de l'individu et de l'événement sont par ailleurs particulièrement bien représentées.

Dans un contexte plus général, les travaux visant la constitution d'une ontologie pour le web propre aux différents types d'objets de recherche demeurent embryonnaires. Une initiative a récemment été lancée, en ce sens, par l'Electronic Literature Organization (ELO). Elle consiste à réunir, dans un souci d'interopérabilité, les informations des bases de données réparties au sein de laboratoires de recherche qui s'intéressent à la création et à l'étude des œuvres hypermédiatiques. Ce projet de grande envergure, intitulé CELL, rendu possible grâce à la mise en commun des ontologies propres à chacune des unités de recherche, va permettre, à terme, l'échange, la normalisation et la densification des données partagées.

CHAPITRE 11

Le livrel et le format ePub

Fabrice Marcoux

Il ne serait pas possible de tracer un panorama complet des enjeux de l'édition numérique sans parler des formats qui tentent de reproduire le mode typique de circulation des contenus papier – le livre – en l'adaptant au support numérique. C'est ce que l'on appelle « livre électronique » ou « livrel » (*eBook* en anglais). Il ne faut pas confondre le livre électronique avec la liseuse qui est le support de lecture. Le livre électronique est un fichier, formaté selon des standards déterminés. Bien évidemment, ces formats ne sont pas stables et changent très rapidement. Aujourd'hui, le standard ouvert de référence est l'ePub. Ce chapitre essaie d'en donner une description.

Introduction au livre numérique

Le *Grand dictionnaire terminologique* de l'Office québécois de la langue française définit ainsi le livre numérique, aussi appelé « livrel » : « Livre disponible en version numérique, sous forme de fichier, qui peut être téléchargé, stocké et lu sur tout appareil électronique qui en permet l'affichage et la lecture sur écran ».

Cette définition est assez générale au vu du fait qu'il existe plusieurs types de livrels : on peut notamment en identifier trois. Le premier, et le plus répandu, est le livrel « homothétique », qui est une transposition à l'identique d'un livre papier en version numérique. Le deuxième est le livrel enrichi, qui prend une place de plus en plus importante : il utilise les possibilités techniques du format numérique afin d'apporter un enrichissement, autant au contenu qu'à la mise en forme de l'ouvrage imprimé qu'il vient compléter. Enfin, il existe un troisième type de livrels : le livrel « originairement numérique », créé *par et pour* le numérique, et non en complément ou par imitation d'un « original papier ».

Le livrel a commencé à prendre de l'importance sur le marché à partir de 2003 et, depuis 2010 environ, les nouveautés publiées par la majorité des maisons d'édition sont immédiatement disponibles en version « homothétique ». Le prix des livrels avoisine généralement 75 % du coût de la version papier et peut s'élever à 100 % (dans le cas des livres présentant de nombreuses illustrations, par exemple). Les livrels sont souvent protégés contre la copie par filigrane et parfois par verrou numérique (DRM, *Digital Rights Management*).

Parmi les formats existants qui permettent de réaliser un livrel, le plus populaire est actuellement l'ePub, format

de fichiers non propriétaire maintenu par l'International Digital Publishing Forum (IDPF) qui a pour mandat d'en faire le standard pour l'édition de livre numérique. L'ePub est une norme ouverte qui permet de créer des livrels inspirés du web ou de livres papier, ou encore de faire des versions enrichies de livres papier pour les liseuses électroniques et pour le web.

Historique de l'ePub

L'ePub a pour ancêtre le format Open eBook, créé par SoftBook Press, société fondée en 1996. La norme *Open eBook Publication Structure* (OEBPS), à la base de la norme OPS (*Open Publication Structure*), fut élaborée à partir de la technologie développée pour la liseuse électronique SoftBook, lancée en 1998.

La prolifération des formats, survenue en 1998-1999, a créé un éclatement de l'offre. Chaque format de livrel n'est alors compatible qu'avec un seul modèle de liseuse : la nécessité de mettre en place une norme ouverte et commune s'impose.

Dès 1998, le National Institute of Standards and Technology (NIST) initie le processus de normalisation en mettant sur pied l'Open eBook Initiative. C'est ce groupe qui « élabore l'*Open eBook* (OEB), un format de livres numériques basé sur le langage XML (*eXtensible Markup Language*: langage de balisage extensible) et destiné à normaliser le contenu, la structure et la présentation des livres numériques[1] ». En septembre 1999, la version 1.0 de l'OEBPS, sur laquelle se fonde le format OEB, est déjà disponible.

1. Marie Lebert, *Une courte histoire de l'*eBook, NEF, Université de Toronto, 2009, p. 75.

En janvier 2000, l'Open eBook Initiative devient l'Open eBook Forum puis, en avril 2005, l'International Digital Publishing Forum (IDPF). Ce consortium a alors une double vocation de commercialisation et de normalisation. Sa mission centrale est « d'établir une norme globale, interopérable et accessible pour les livres électroniques et d'autres publications, afin de contribuer à la croissance de l'industrie de l'édition numérique[2] ». Quelques mois plus tard, en juillet 2005, le format ePub remplace l'OEB. Mais ce n'est qu'en 2007 que l'ePub2 devient une norme de l'IDPF.

L'IDPF lance, en octobre 2011, l'ePub3 : le format ePub s'appuie dès lors sur les règles du HTML5 (plutôt que sur le XHTML) pour la structuration du contenu et sur celles des feuilles de style CSS3 pour la mise en forme.

En mars 2013, l'IPA (Union internationale des éditeurs) déclare officiellement qu'elle reconnaît le format ePub3 comme norme internationale. Au moment d'écrire ces lignes, l'ePub3 suit le processus d'accréditation pour recevoir le statut de norme ISO/IEC (International Organization for Standardization/International Electrotechnical Commission).

Particularités techniques du format ePub

L'ePub est basé sur les mêmes langages de balisage que ceux employés pour la réalisation de sites web : il s'agit de fichiers HTML. C'est donc un format permettant de faire des livres numériques ayant à la fois les caractéristiques du livre papier et les caractéristiques d'un site web.

---

2. Bill McCoy, « Portable Documents for the Open Web – Part 1 : What Role does ePub Play in the Cloud-Centric World? », *O'Reilly-TOC*, 2012 (traduction libre). Bill McCoy est président de l'IDPF.

Le fichier de format ePub est organisé selon une arborescence qui comprend un dossier dans lequel figurent des fichiers HTML, des CSS et des documents de différents types (audio, image, vidéo, etc.) regroupés eux-mêmes par sous-dossiers. Ce qui procure à cet ensemble une unité « organique » est la rigueur avec laquelle tous les éléments sont répertoriés dans l'élément « *manifest* » du fichier OPF (document structuré fondé sur XML). Il faut également que tous les documents du contenu soient regroupés dans le dossier OPS. Le dossier initial est ensuite compressé suivant le protocole d'archivage ZIP (norme ouverte de compression de dossiers). L'extension « .epub » vient remplacer l'extension « .zip » du fichier qui en résulte.

Le principe du « bien formé » qui préside à la construction des documents structurés fondés sur XML suppose qu'il ne doit y avoir qu'un seul élément racine dans lequel tous les autres s'emboîtent. Un document XML est nécessairement du même type que son élément racine (‹html› dans le cas des pages web). Mais, pour éliminer toute ambiguïté quant à ce que signifie ce type, il peut être opportun d'introduire, au début du document, une déclaration XML qui renvoie à une DTD (définition de type de document). La DTD sert à prescrire ce que peut contenir chaque élément et sous-élément. Tout élément de contenu doit être encadré par une balise ouvrante (‹élément) et une balise fermante (élément/›). Les sous-éléments doivent être entièrement compris dans les éléments de niveau supérieur (pas de chevauchements). Les consignes de mise en forme, s'il y a lieu, sont complètement séparées du contenu. On peut ainsi changer la mise en forme sans affecter le contenu, ou

extraire des informations du contenu sans que le code des consignes de mise en forme fasse interférence.

Le détail des spécifications techniques constitutives des deux versions du format (ePub2 et ePub3) est disponible gratuitement sur le site de l'IDPF. Celles-ci sont formelles et indiquent ce sur quoi doivent se baser les développeurs qui veulent créer une application capable de lire correctement un fichier au format ePub. Il est possible de faire valider un fichier réalisé suivant l'une ou l'autre des versions de la norme recommandée par le consortium.

Caractéristiques générales

Plusieurs des caractéristiques générales de l'ePub sont reliées à l'adoption du principe des documents structurés.

La première caractéristique de l'ePub est de pouvoir contenir tous les éléments typiques d'un document sur le web. En d'autres mots, le format ePub est un cousin des sites web et leurs structures sont semblables. Les documents contenus dans un fichier ePub sont rangés dans les dossiers pour les chapitres (texte), les images, le son (audio) et les films (vidéo). Les deux versions supportent bien les hyperliens. La version 3.0, à l'instar du HTML5, permet les scripts (Javascript, une autre norme ouverte) dans les documents de contenus.

La deuxième caractéristique de l'ePub est de permettre d'avoir l'ensemble des éléments paratextuels propres au livre : un livrel peut avoir une table des matières, un index, une page de couverture. Un nom de fichier est réservé pour chacun de ces éléments. Enfin, ce qui rapproche beaucoup l'ePub du livre est le principe de répartition du contenu : un chapitre par fichier. Cependant, les fonctionnalités typiques

du numérique permettent d'aller au-delà de la structuration linéaire du livre papier. À travers des systèmes de signets, on permet de transformer une idée héritée du livre papier en une véritable forme d'hyperliens.

Les livrels ePub (2 ou 3) peuvent contenir divers éléments caractéristiques des deux environnements (web et livre papier). Ceux-ci incluent notamment les tableaux, les listes, les images (et autres figures) et les encadrés (‹div›). S'y ajoutent les divisions qui servent pour l'en-tête et le pied de page, et qui sont explicitement prévues dans le HTML5. Il en va de même pour les « notes de bas de page », qui deviennent des « notes marginales » (élément « *aside* »). Des graphiques de type SVG (*Scalable Vector Graphics* – Graphiques vectoriels adaptables), basés sur le balisage XML, peuvent également être inclus, garantissant ainsi l'interopérabilité. Il s'agit de dessins de type vectoriel, qui se construisent avec des balises et leurs attributs (du texte) et non par du code binaire.

À ces caractéristiques de base du fichier ePub peuvent se greffer des dispositifs optionnels tels que les dictionnaires et les DRM.

Certaines applications, comme Aldiko, offrent la possibilité de consulter des dictionnaires en cliquant sur des mots contenus dans le livrel. L'usager peut ainsi avoir la définition d'un terme en temps réel, et éventuellement sa traduction en une autre langue. Ces dictionnaires peuvent être inclus dans l'application de lecture, ou alors être disponibles sur le web et demander une connexion pour être consultés.

En ce qui concerne les DRM, dans le cadre d'une « édition homothétique », les éditeurs peuvent avoir tendance à essayer de transposer en numérique le modèle économique

de l'édition papier. Ce modèle est cependant mis en crise par la facilité de copier des livres numériques. De plus, les versions « homothétiques » livrent l'intégralité du contenu du livre papier. Les concepteurs du format ePub ont alors prévu la possibilité de configurer des verrous numériques (DRM), pensant qu'une majorité d'éditeurs n'adopteraient ce format qu'à cette seule condition. Mais ces restrictions au partage n'étant pas unanimement acceptées, l'IDPF a rendu les DRM optionnelles.

Fonctionnalités

Passons maintenant en revue les six principales fonctionnalités de l'ePub (et des autres livrels).

Premièrement, il est portatif puisqu'en tant que fichier informatique il ne pèse rien de plus physiquement par rapport à l'appareil de lecture (sur lequel on peut stocker des milliers de livrels). Et son contenu est « recomposable » (*reflowable*), le rendant consultable sur des plateformes variées. Le texte flottant, à savoir le fait que le texte se met en page automatiquement selon la taille de l'écran et les options choisies, est un des traits qui distinguent l'ePub du PDF.

Deuxièmement, l'ePub permet la recherche en plein texte. L'architecture de l'information rigoureuse qui le caractérise donne la possibilité de créer des applications pour aller chercher (extraire) des données. Dans le cas de l'ePub3, les CFI (*Canonical Fragment Identifiers*) améliorent cette capacité. Il s'agit d'une spécification qui définit des méthodes standards pour baliser puis référencer des fragments de contenus (un mot, une partie du texte, une image, etc.) d'un fichier ePub. Cela permet de créer des hyperliens

et de naviguer plus aisément, de façon non linéaire, dans les contenus d'un fichier.

Une troisième fonction fondamentale des documents ePub est la possibilité d'associer des annotations au contenu d'un livrel. Il faut préciser que cette fonctionnalité dépend pour le moment davantage de l'application que du format. C'est pourquoi l'ePub3 permettra d'associer des annotations plus riches grâce, justement, aux identifiants de fragments (CFI). Ces identifiants devraient aussi favoriser le partage des annotations et des marque-pages (ou signets) qui sont une forme d'annotation standard dans la plupart des logiciels de lecture.

L'ePub3 offre davantage de possibilités d'« échanges » avec le livrel que l'ePub2. Il prévoit ainsi la création de quiz, des animations dans les fenêtres surgissantes (*pop-up*), des couches de multimédias superposées, etc. Les effets de styles autorisés par l'ePub 3 (Javascript et CSS3) sont également plus riches, mais l'IDPF recommande d'utiliser animations et autres scripts « sophistiqués » avec « prudence ».

En quatrième lieu, les fichiers ePub sont configurables. Pour les options de présentation, ils permettent un ajustement des paramètres concernant le rendu du texte à l'écran (polices, tailles, marges). Il est aussi possible de déterminer une couleur de fond d'écran, pour atténuer au besoin le contraste, par exemple. D'autres aspects peuvent être configurés, comme la prise en charge des jeux de caractères particuliers, requérant parfois le chargement de bibliothèques, comme MathML. Ce langage pour l'affichage correct de formules mathématiques est supporté par l'ePub3, grâce à HTML5 et CSS3.

Il est aussi faisable – et même souhaitable parfois pour les livrels enrichis – de configurer le livrel avec une mise en page déterminée (*fixed layout*), même lorsqu'il s'agit d'un format ePub (version 3). Mais on perd alors l'avantage de la mise en page « flottante » (« recomposable »).

En cinquième lieu, les applications de lecture proposent généralement plusieurs options pour la présentation des pages et la manière de circuler à travers le contenu. Les flèches à gauche et à droite permettent de revenir en arrière ou d'avancer. Une barre de défilement permet de se situer dans l'ensemble du document et de se rendre directement à un autre point. Un champ présentant le numéro de page peut être visible et offrir la possibilité d'entrer un autre numéro pour accéder directement à la page choisie (quand les pages sont indexées). On peut compter sur la présence d'une table des matières avec hyperliens vers les parties en question (pourvu que les parties du contenu aient été incluses dans le fichier assigné à cet usage : le fichier TOC). L'index est un autre moyen de se repérer pour naviguer.

Sixième fonctionnalité fondamentale, les métadonnées globales sont incluses dans l'élément « *metadata* » du fichier dont l'extension est « .opf », et sont structurées en Dublin Core. Le Dublin Core est une norme pour les métadonnées du milieu documentaire. Pour référencer des fragments au moyen d'un identifiant numérique (dans le but de pouvoir y associer des métadonnées), l'ePub3 spécifie comment procéder grâce au « protocole epubcfi ». Les métadonnées permettent d'évaluer la pertinence d'un document et de le qualifier. Elles sont « encapsulées » dans l'enveloppe (l'élément « *package* ») du document.

Évaluation

Les possibilités énumérées ci-dessus découlent de la structure même du format ePub : il est nécessaire que les concepteurs de chaque livrel les implémentent de manière optimale pour en tirer tout le bénéfice. Il faut néanmoins tenir compte des limites des applications de lecture, dont l'ergonomie peut rendre parfois difficile l'exploitation des fonctionnalités offertes par ce format.

On peut proposer un bilan provisoire du développement du format ePub en essayant d'en montrer les principales qualités mais aussi d'en identifier les limites.

On constate que la plupart des forces principales de l'ePub proviennent de son organisation et de la cohérence que lui procure l'utilisation des principes de séparation de la forme et du contenu, caractéristiques des documents structurés.

La première chose qu'un auteur souhaite, quand il confie son livre à un éditeur, est que son œuvre soit lue. Or, si le fait de publier par et pour le web ouvre à tout un nouveau public, il faut relever le défi de rendre les contenus visibles à travers la mer d'informations qui envahit la toile.

Le fait que les fichiers électroniques permettent la recherche en plein texte est un avantage. Mais s'ils ne sont pas repérés eux-mêmes, c'est totalement inutile. Or l'ePub présente l'avantage, justement, de favoriser la repérabilité de ses contenus, puisqu'il en présente (s'il est bien fait) les informations pertinentes sous forme de métadonnées globales (Dublin Core) et spécifiques (CFI). La possibilité de bien structurer les métadonnées dans le fichier ePub est tout à fait alignée au développement progressif du web vers le web sémantique. En ce sens, il n'en tient qu'aux éditeurs

de profiter de l'occasion que l'ePub leur offre de rendre leurs contenus beaucoup plus repérables et récupérables en s'assurant d'indiquer clairement dans l'élément « *metadata* » toutes les informations utiles.

Avec l'ePub3, il est aussi possible d'entrer plus profondément dans la structure pour qualifier des fragments de documents (CFI) qui méritent d'être portés à l'attention des moteurs de recherche. L'attribution des métadonnées est un savoir-faire qui doit être appris et appliqué.

L'accessibilité est d'une importance capitale dans le contexte de la société de l'information. Cela passe par l'adoption de normes communes par les développeurs de navigateurs (W3C). À l'instar des pages web, les livrels deviennent des outils pour relier des contenus de types différents. Il est donc important que des normes assurant la conformité à certains standards soient établies et suivies. C'est ce qui donne la possibilité aux créateurs d'applications d'assistance technique de fournir des outils universellement applicables pour les personnes aux prises avec des obstacles à la lecture ou à l'écoute. Ces extensions des logiciels de lecture rendent le contenu disponible autrement pour pallier une difficulté d'accessibilité. L'ePub se positionne avantageusement à cet égard depuis longtemps.

La grande force du format ePub est certainement d'être ouvert. Comme nous l'avons indiqué, cela signifie que les spécifications constitutives du format sont disponibles gratuitement pour tout le monde. De cette façon, il est possible à tous les distributeurs, éditeurs ou développeurs de se baser sur ces indications pour élaborer des logiciels qui seront capables de lire les livrels au format ePub. Il n'y a ainsi pas de raison pour qu'une plateforme populaire

n'offre pas la possibilité de lire des ePub. C'est ce que l'on appelle l'interopérabilité. Et, surtout, cela encourage la collaboration.

On peut, en revanche, indiquer certaines limites du format ePub.

Premièrement, il faut reconnaître que les contraintes liées au langage de balisage peuvent rendre rigide la structure qui devra être donnée aux livrels pour qu'ils puissent être validés. Deuxièmement, même si les principes du XML qui sous-tendent l'organisation des livrels au format ePub sont d'une logique relativement simple à comprendre, ils supposent la connaissance de règles syntaxiques rigoureuses et la maîtrise de nouveaux langages qu'il peut être difficile – et coûteux pour les éditeurs – d'apprendre. Troisièmement, les deux versions de la norme (ePub2 et ePub3) peuvent, pour le moment, apporter de la confusion.

En dernier lieu, il pourrait y avoir des résistances à adopter le format ePub en raison de l'attrait supplémentaire qui offrent des formats concurrents. Les formats privatifs, en particulier, appartenant à des corporations comme Amazon (AZW et KF8) et Apple (iBooks, un format privatif dérivé de l'ePub), bénéficient de la visibilité que leur offrent les plateformes de distribution et les appareils de lecture de ces compagnies.

CHAPITRE 12

Les potentialités du texte numérique

Stéfan Sinclair et Geoffrey Rockwell

Pour avoir une idée précise des possibilités apportées par le numérique à l'édition, il est indispensable de pouvoir assumer le point de vue du lecteur : que peut-on faire avec un texte numérique ? Comment peut-on l'exploiter ? Que peut-on en apprendre en utilisant des outils d'analyse ? Se poser ces questions est indispensable lorsque l'on s'interroge sur les bonnes pratiques de la production des contenus numériques. Seul le fait d'être conscient des potentialités offertes par le texte numérique nous permettra de mettre en place de bonnes pratiques et d'en produire. Ce chapitre propose une introduction à ces thématiques.

Introduction

Pour quelqu'un de passionné par la littérature, la fouille informatisée de textes peut paraître bien exotique, voire subversive : pourquoi voudrait-on céder le moindre plaisir du texte à une calculatrice ? Tout le monde sait que l'ordinateur ne comprend rien aux relations humaines, au langage métaphorique, à l'ironie et à bien d'autres ingrédients encore qui donnent aux textes leur piquant. À quoi peut bien servir la machine pour l'étude et la critique littéraires ?

Ce chapitre tentera de répondre à cette question, sans pour autant vouloir convertir qui que ce soit. Nous reconnaissons que la réaction sceptique chez le littéraire est tout à fait naturelle et nous estimons d'ailleurs qu'une bonne dose de scepticisme est essentielle lorsque se conjuguent analyse informatisée et herméneutique. Cela dit, nous souhaitons remettre en cause l'image dominante que nous fournit la société de l'ordinateur comme générateur prodigieux de données infaillibles et de graphiques inéluctables. Ce que l'on ignore souvent, c'est que l'ordinateur, grâce à la nature même du numérique, peut s'avérer une aide très puissante pour faire proliférer le nombre et les types de représentations d'un texte. Loin d'en réduire la souplesse et la richesse, les outils informatiques peuvent servir à multiplier la matière brute qui mène à de nouveaux constats, de nouvelles associations, de nouveaux arguments. La machine est l'engin du heureux hasard, contrainte seulement par l'imagination de son utilisateur.

Ce chapitre est divisé en deux sections. Dans la première, nous rappellerons quelques caractéristiques clés des textes numériques ; il est capital de bien comprendre la nature du

texte numérique avant de pouvoir procéder à son analyse. Dans la deuxième section, nous introduirons certains concepts de base pour la lecture informatisée de textes et nous signalerons quelques outils repères qui sont adaptés au contexte des sciences humaines (compétences techniques, orientation épistémologique, etc.).

Le texte numérique

Le monde change vite : d'après l'Association of American Publishers, la part du marché des livres numériques (aux États-Unis) est passée d'environ 1 % en 2008 à presque 23 % quatre ans plus tard[1]. Cette croissance explosive semble se stabiliser et la situation n'est pas la même dans toutes les régions et pour toutes les langues (ni pour tous les genres : une proportion démesurée de lectrices et lecteurs semblent préférer la discrétion de l'édition numérique de *50 nuances de Grey*, par exemple). Toujours est-il que l'édition numérique s'installe, se normalise. La convivialité des liseuses (leur poids, leur écran, leur interface) et la disponibilité des titres font que l'attachement historique à la page imprimée se délie, même pour la lecture dans le bain.

Pour la lecture conventionnelle (séquentielle), la question du format est somme toute secondaire. Si on lit un texte du début à la fin, peu importe de tourner une page matérielle ou d'appuyer sur une flèche pour avancer. Certes, l'expérience de lire n'est pas identique dans les deux cas (on peut penser aux indices tactiles que représentent le nombre de pages que l'on a lues dans la main gauche par rapport aux

1. Andi Sporkin (2013), « Trade Publishers' Net Revenue Grows 6.2 % for Calendar Year 2012 » : ‹http://bit.ly/1ek9HN1›.

pages qui restent à lire dans la main droite), mais bien d'autres facteurs interviennent également dans l'expérience personnelle (lieu de lecture, heure de la journée, couverture souple ou cartonnée, etc.). On peut d'ailleurs s'étonner de voir l'étendue du *skeuomorphisme* dans les éditions numériques, c'est-à-dire à quel point on cherche à rassurer les lecteurs en reproduisant les caractéristiques familières du volume imprimé dans la version numérique. Et pourtant, il peut exister aussi des fonctions inédites dans la version numérique, tel que les notes publiques (les annotations ajoutées par d'autres lecteurs qui se font à une tout autre échelle que les quelques gribouillages clandestins que l'on peut trouver dans les marges d'un livre imprimé).

La question qui nous préoccupe ici n'est pas comment on « lit » un texte du début à la fin, mais plutôt comment on « étudie » et « analyse » un texte. Il y aura toujours des divergences de préférences pour la consommation conventionnelle des textes, que ce soit la page imprimée, la page web, l'écran de la liseuse, le livre audio ou d'autres supports encore. Les préférences personnelles peuvent jouer également dans le choix de format pour l'analyse de texte (informatisée ou non), mais force est de constater que tous les formats ne sont pas égaux. Rappelons d'emblée que l'étymologie du mot « analyse » évoque le déliement et la décomposition. Or telle est justement la nature du numérique : être coupé, être représenté par des unités discrètes d'informations, des *bits*. Alors que la page imprimée est un support analogue dans la mesure où elle représente une séquence continue de texte (l'unité des lettres individuelles n'étant pas à confondre avec la continuité de la page où elles

sont inscrites), le support numérique traite déjà chaque lettre comme une entité indépendante et mobile.

Déjà dans son article de 1985 intitulé « Quelques réflexions sur le statut épistémologique du texte électronique », Serge Lusignan décrivait les retombées du texte numérique :

> Le texte magnétique ou électronique possède des caractères de flexibilité et de malléabilité qu'ignore le texte imprimé. Les caractères et les mots incrustés dans le papier ne peuvent être ni déplacés, ni ré-ordonnés, ni modifiés, tandis que les caractères et les mots magnétisés sont complètement mobiles. Ce trait propre au texte électronique permet de lui appliquer, grâce à l'ordinateur, différentes procédures algorithmiques de manipulation[2].

Rien du format numérique n'oblige une réorganisation des lettres et des mots (comme en témoigne la grande majorité des éditions numériques). En revanche, il est possible de découper le texte imprimé comme l'ont fait de façon enjouée les Dadaïstes ou, de façon moins enjouée mais bien avant, les moines du XIII^e siècle qui ont inventé la concordance en réorganisant le texte de la Bible par chaque occurrence d'un mot clé avec un peu de contexte. Simplement, la nature même du texte numérique facilite le découpage et la réorganisation, elle se prête naturellement à l'analyse. En empruntant le jargon du domaine de l'interaction homme-machine, on pourrait parler de l'*affordance* du texte numérique : la structure de l'information textuelle en unités mobiles suggère d'elle-même sa propre utilisation pour l'analyse.

---

2. Serge Lusignan, « Quelques réflexions sur le statut épistémiologique du texte électronique », *Computers and the Humanities*, vol. 19, n° 4, A Special Double Issue on Activities in Canada Part I (oct.-déc. 1985), Springer, p. 209-212.

Nous reviendrons aux possibilités de l'analyse informatisée dans la prochaine section, mais il vaut la peine de s'attarder un moment sur la simplicité trompeuse du texte numérique car, sans une compréhension de la matière brute avec laquelle on travaille, il est difficile de véritablement comprendre ce que l'on construit.

Comme nous l'avons déjà évoqué, l'ordinateur fonctionne comme un système binaire qui traite de l'information encodée à un niveau élémentaire en 0 et 1 (ce qui représente par la suite la présence ou l'absence de courant dans un transistor qui contrôle des circuits logiques). Il est remarquable que toute la magie informatique soit rendue possible par cette dichotomie élémentaire d'une grande simplicité : les mots dans un texte numérique, les requêtes quasi instantanées d'un moteur de recherche sur des centaines de milliards de pages indexées, le graphisme réaliste d'un jeu vidéo, le système qui gère les mouvements complexes d'un avion commercial, et ainsi de suite.

Si l'on prend les 26 lettres de l'alphabet romain, il suffit de 5 bits (5 colonnes de 0 et de 1) pour représenter toutes les possibilités[3]. Effectivement, à l'aube de l'ère informatique (dans les années 1940 et 1950), c'est justement avec 5 bits que les textes étaient représentés, mais cela ne servait que pour les lettres en majuscules. Il faudrait au moins 52 possibilités pour inclure aussi les lettres en minuscules, sans parler des caractères avec diacritiques, de la ponctuation et d'autres marques typographiques, et des variantes de l'espace blanc (espace simple, espace insécable, espace de

3. 5 bits = 2^5 = 2*2*2*2*2 = 32.

tabulation, fin de ligne, etc.). L'histoire de l'informatique trace d'ailleurs une progression des jeux de caractères de plus en plus grands et inclusifs : 7 bits (128 possibilités) pour ASCII en 1963, 8 bits (256 possibilités) pour l'ASCII étendu, et jusqu'à 16 bits pour Unicode (UTF-32 pouvant représenter jusqu'à 4 294 967 296 possibilités). On peut se demander pourquoi ne pas avoir créé un grand jeu de caractères dès le début, mais n'oublions pas que la mémoire était alors précieuse pour l'informatique. C'était le même principe pour la représentation des années avec deux chiffres qui a causé beaucoup de soucis à la fin du millénaire. Les 4Ko disponibles dans l'Apple II de 1977 sont 1 000 000 de fois inférieurs au 4 Go disponibles dans un modèle d'ordinateur portable de base aujourd'hui.

Tout aussi important que la taille des jeux de caractères est leur standardisation. Rien d'inhérent ne définit la lettre *A* comme le code décimal 65 (comme le font les normes ASCII, ISO-8859-1 et Unicode) ; c'est une convention. Simplement, pour que les systèmes puissent se parler, pour qu'il y ait interopérabilité des données, pour éviter une tour de Babel, les standards sont essentiels pour les caractères.

Il en va de même pour les formats de document. Le format le plus simple est un document en texte brut, mais ce format ne permet pas de préciser le jeu de caractères utilisé, ce qui cause des ennuis pour un texte en français qui s'affichera différemment selon qu'il s'agit de Latin-1, Mac OS Roman, Unicode ou autre. Un texte formaté peut exprimer l'encodage des caractères, mais toute instruction de formatage entraîne un coût important pour la complexité du format. Or plus un format est complexe, plus le logiciel de traitement devra être complexe et plus

la pérennité du format sera en cause, surtout pour les formats dits propriétaires (ou fermés). C'est justement cette logique qui a motivé Michael Hart, le fondateur du Projet Gutenberg, à privilégier le format texte brut pour sa collection de textes du domaine public qui comprend aujourd'hui quelques 40 000 titres. Le format ouvert ePub (qui utilise en partie le même balisage HTML que pour les pages web) s'est établi comme un format de prédilection pour la diffusion de textes numériques, surtout pour sa mise en page flexible qui facilite la représentation multiplateforme sur des écrans de tailles très différentes (téléphone intelligent, tablette, liseuse, ordinateur, etc.). La capacité du format ePub à gérer (de façon facultative) les droits d'accès a sans doute énormément contribué au succès commercial du format. L'ePub réussit à encoder les caractères et la structure de base des textes, mais il n'est pas conçu pour représenter de façon standardisée d'autres détails textuels et métatextuels, tel que les types de stances dans un poème, les variantes orthographiques entre différentes éditions ou les trous illisibles dans une page manuscrite, pour ne donner que trois exemples parmi un nombre presque infini. C'est pourquoi les chercheurs et archivistes soucieux de capter une gamme beaucoup plus large de détails préfèrent la *Text Encoding Initiative* (TEI), un langage XML qui permet de décrire les caractéristiques sémantiques d'un texte plutôt que sa présentation (qu'une séquence de mots exprime un titre de livre, par exemple, et non seulement un bloc quelconque à représenter en italique).

Quel que soit le format, la première étape pour l'analyse de texte informatisée est souvent l'extraction du texte en

format brut[4]. Il serait bien de pouvoir exploiter les balises sémantiques dans un fichier de format TEI, par exemple, mais très peu d'outils sont conçus pour le faire (et donc les balises deviennent superflues ou même nuisibles à l'analyse).

La lecture informatisée

Il nous est arrivé à tous d'avoir un passage d'un livre à l'esprit, et même le souvenir de quelques mots clés, mais d'éprouver du mal à le retrouver en feuilletant les pages. Quiconque a utilisé la fonction « Rechercher » dans un logiciel pour trouver un mot ou une phrase dans un fichier PDF, un document MS Word ou une page web connaît déjà l'utilité du format numérique ; c'est une opération de dépistage que la page imprimée n'offre pas. Il ne faut surtout pas sous-estimer la valeur des procédures simples avec les textes numériques. Par exemple, si l'on souhaite mieux comprendre la façon dont Molière parle de « ma pensée » dans une version PDF des *Œuvres complètes*, il suffit de faire une recherche de phrase (mots entre guillemets) dans le logiciel Aperçu (ou Adobe Acrobat) pour voir une concordance qui permet de naviguer facilement entre chaque occurrence (autrement dit, certaines fonctions d'analyse de texte sont à la portée de tous). La recherche de mots clés nous permet de trouver ce que l'on cherchait, mais elle permet aussi de découvrir ce que l'on ignorait. La découverte est d'ailleurs le premier des principes de la recherche savante (« *scholarly primitives* ») identifié par John Unsworth en parlant des méthodologies dans les sciences humaines

4. Un des meilleurs outils de conversion et d'extraction s'appelle Calibre (calibre-ebook.com).

(les autres étant l'annotation, la comparaison, le renvoi, l'échantillonnage, la démonstration et la représentation[5]).

Comme c'est généralement le cas dans l'analyse de texte, la recherche de mots clés paraît simple, mais elle peut vite devenir étonnamment compliquée. Revenons à l'exemple de « ma pensée » : une séquence de caractères que le logiciel devrait pouvoir trouver dans un texte source. Mais que se passe-t-il si la séquence se trouve en début de phrase et que la lettre initiale est en majuscule ? Ou si deux espaces ou une fin de ligne séparent les deux mots ? Et nous passons sous silence les variantes orthographiques, les coquilles, etc. Certains logiciels s'occupent de normaliser les différences (par exemple Aperçu), d'autres offrent des paramètres avancés de recherche (par exemple MS Word) et d'autres encore permettent ce que l'on appelle des expressions régulières (par exemple TextMate), une syntaxe très souple pour effectuer des recherches. Ainsi, si l'on voulait trouver « ma pensée » ou « ta pensée » de façon très flexible, on pourrait définir une expression régulière comme : /\b[mt]a\s+pensée\b/i (la barre oblique commence l'expression, \b indique la frontière d'un mot, [mt] correspond à l'une ou l'autre des lettres, \s+ trouve un ou plusieurs caractères d'espace blanc, la deuxième barre oblique termine l'expression et le *i* indique que l'expression n'est pas sensible à la casse).

La recherche de mots clés mène naturellement à une deuxième forme d'analyse informatisée qui est facilitée par

5. Ce sont nos propres traductions de « *discovering* », « *annotating* », « *comparing* », « *referring* », « *sampling* », « *illustrating* », « *representing* » ; voir John Unsworth « Scholarly Primitives : What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This ? » : ‹http://bit.ly/1bea3Xz›.

la nature numérique des textes : le comptage des fréquences. Les fréquences individuelles ne valent rien en soi, elles gagnent à être comparées, soit à d'autres termes dans le même texte, soit à la fréquence d'un terme dans un autre document. Dire par exemple qu'il y a 305 occurrences du mot « monsieur » dans notre édition du *Bourgeois gentilhomme* de Molière manque de contexte et de poids. Par contre, cela peut devenir plus intéressant quand l'on considère que le même texte ne contient que 180 occurrences de « madame », presque la moitié des occurrences de « monsieur ». On pourrait commencer à formuler une hypothèse au sujet de la présence relative des sexes dans ce texte, mais avec les opérations quantitatives il faut toujours rester conscient des particularités des données. *Le Bourgeois gentilhomme* est une pièce de théâtre, bien entendu, et notre édition en texte brut indique le nom de chaque personnage devant sa réplique – la fréquence de certains mots peut être déjouée par la présence d'éléments métatextuels (l'indication des répliques de Monsieur et Madame Jourdain). D'inclure dans le décompte le nom des personnages n'est pas nécessairement faux, mais il faut tout simplement en être conscient et pouvoir le justifier (la proportion relative des termes « monsieur » et « madame » reste à peu près pareille avec ou sans les indications de personnage).

On peut comparer des fréquences absolues dans un même document, mais aussi les fréquences absolues ou relatives par rapport à d'autres documents. Le concept de fréquence relative est important : si deux documents ont exactement 1 000 mots, les fréquences absolues et relatives sont identiques. Cependant, puisque les textes sont presque toujours de longueurs différentes, il est souvent utile de

relativiser les valeurs : 10 occurrences d'un terme dans un document de 1 000 mots ne valent pas 10 occurrences dans un document de 500 mots – le terme est deux fois plus fréquent dans le deuxième texte. On cherche donc une échelle relative ; on peut dire que le premier texte a 100 occurrences par 10 000 mots alors que le deuxième a 200 occurrences toujours par 10 000 mots (l'échelle importe peu, l'important est de trouver une façon de comparer les valeurs).

Certains cas de comparaison se prêtent très bien à une étude de l'agencement ordonné des valeurs. Une forme d'agencement possible (et commune) est l'ordre chronologique des documents. Si l'on prend les fréquences relatives de « monsieur » et « madame » dans l'ensemble de l'œuvre théâtrale de Molière disposée en ordre de parution, on observe une augmentation assez marquée vers la fin de sa carrière. L'outil *Google Books Ngram Viewer* permet de telles requêtes diachroniques à une échelle inouïe (plus de cinq millions de livres, toutes langues confondues), même si les détails pour chaque texte demeurent malheureusement inaccessibles. *Google Books Ngram Viewer* a été devancé par plusieurs projets de textes numériques à grande échelle en français, y compris *Le Trésor de la langue française* au début des années 1970, American and French Research on the Treasury of the French Language (ARTFL) au début des années 1980 et Gallica au début des années 1990.

Les graphiques linéaires peuvent être très puissants pour visualiser les données « comparatives » (comme les fréquences relatives), et il en va de même pour les données « associatives » en réseau. Si l'on veut examiner la collocation des termes – autrement dit la proximité des termes –, on peut utiliser les concepts précédents de recherche, comp-

tage et comparaison pour créer une constellation de termes à haute fréquence qui ont tendance à paraître ensemble. On pourrait vouloir examiner le réseau de mots liés (par proximité) à « monsieur » et « madame » : quels mots sont les plus associés à l'un ou l'autre de nos mots clés ? Quels mots sont associés aux deux ? Le graphique nous suggère que « mari », « dieu » et « fille » sont plus associés à « madame », alors que « maître », « comédie » et « philosophie » sont plus associés à « monsieur », et enfin que « comtesse », « vicomte » et « oui » sont associés aux deux (entre autres). Encore une fois, les associations ne sont pas concluantes en soi, mais le graphique peut nous mener à poser des questions qui ne nous seraient pas venues à l'esprit autrement.

Plusieurs outils sont disponibles pour la lecture et l'analyse informatisées mais, à quelques exceptions près, il s'agit de variantes des procédures élémentaires présentées ici : la recherche, le comptage, la comparaison et l'association. La préparation des données textuelles peut être très exigeante (et souvent quelques compétences de programmation s'avèrent indispensables, surtout les langages de script comme Python, Ruby et PHP). Mais des outils comme *Voyant Tools* démontrent qu'il est possible de prendre plusieurs documents dans des formats variés (texte brut, MS Word, PDF, etc.), de les ajouter à l'outil et de commencer à lire, à explorer et à analyser. Ces activités diverses représentent les atouts des textes numériques.

Conclusion

Les textes numériques nous entourent de plus en plus, mais nous ne les connaissons guère. Nous avons voulu ici revisiter la nature des textes numériques et surtout

voir comment la décomposition de l'information en unités discrètes nous permet d'élargir le sens de la lecture au-delà du texte séquentiel pour inclure la recherche, le comptage, la comparaison et l'association. Cela dit, il est essentiel de comprendre le rôle de l'informatique dans l'entreprise littéraire : elle peut nous aider à répondre à des questions quantitatives, mais celles-ci ont tendance à être banales et à manquer de contexte. Le vrai potentiel du texte numérique réside en sa capacité de faire proliférer les représentations des textes et de nous amener à poser de nouvelles questions interprétatives.

CHAPITRE 13

Concevoir un projet éditorial pour le web

Hélène Beauchef

Nombreux sont les contenus qui circulent désormais en ligne et plus nombreux encore sont ceux qui y circuleront demain. Le livre a longtemps été le principal vecteur d'information et de connaissances, mais il se fait progressivement remplacer par le web. L'édition numérique concerne de plus en plus les pratiques de production, de validation et de diffusion des contenus en ligne. Dans ce cadre, il est indispensable de considérer un site web (un blogue, une plateforme à la structure complexe) comme un projet éditorial et de le traiter avec les attentions et les compétences d'un éditeur. Comment projette-t-on un site web ? Comment agence-t-on ses contenus ? Comment choisit-on les solutions techniques en cohérence avec ces contenus ? De l'expression du besoin à la phase d'exploitation, en passant par les choix éditoriaux, techniques et graphiques, ce chapitre expose les rouages et explique les bonnes pratiques qui doivent être à la base de toute conception et réalisation d'un projet éditorial en ligne.

Préparation du projet : le cahier des charges fonctionnel

Le niveau de détails et d'implication dans la conception d'un projet éditorial pour le web varie selon la taille du projet, de la structure dont il émane, du budget, etc. Dans certains cas, une ou deux personnes aux compétences plurielles seront en charge de l'intégralité du projet, dans d'autres, le nombre d'intervenants pourra être bien plus important et les tâches compartimentées. Les missions éditoriale, technique et graphique peuvent être intégrées – avec des équipes au sein de la structure à l'origine du projet – ou externalisées – soit remplies par des prestataires missionnés ou par une agence.

La première étape du projet est l'expression et la contextualisation du besoin. Elle peut se faire sous la forme d'un cahier des charges fonctionnel. Après avoir récolté toutes les informations nécessaires, le responsable de projet (ou le maître d'ouvrage, MOA) doit y définir, de manière claire et complète, le projet, son environnement, ses objectifs et les moyens disponibles pour sa mise en place. Le cahier des charges fonctionnel définit les lignes conductrices et doit être une référence pour tous les acteurs du projet, tout au long de son développement.

On doit notamment y trouver les informations suivantes :

- Quelle est la structure à l'origine du projet ? Il s'agit de la carte d'identité du commanditaire : type d'organisme ou société, histoire, mission, nombre d'employés, présence sur le web, stratégie, etc.
- En quoi consiste le projet et quels en sont les objectifs ? Le projet doit être expliqué en quelques phrases et le besoin formulé

précisément. Est ici décrit ce que doit dire, faire et apporter le projet (fonctionnalités attendues) et non comment il doit le faire (fonctionnement technique).

- Quel est le public visé ? Quels sont ses usages et ses besoins ? Selon l'âge, les habitudes de lecture, les pratiques liées au web – mais aussi la langue, la situation géographique ou le champ d'activité (étudiants en médecine, professeurs en sciences humaines, cuisiniers, etc.) – le public du futur projet doit être identifié et défini. S'il est difficilement envisageable de proposer un site répondant précisément aux besoins de chacun des usagers, il est cependant important de tracer un portrait-type du ou des usagers moyens, prenant en compte leurs principales caractéristiques et leurs attentes. L'usager-type est-il à l'aise avec les nouvelles technologies ? De quel type de connexion dispose-t-il (haut débit, limité ou illimité) ? Sur quel support navigue-t-il le plus souvent (ordinateur, tablette, téléphone intelligent) ? Est-il un consommateur de vidéos et de *podcasts* ? Utilise-t-il les réseaux sociaux et lesquels ? Peut-il lire de longs contenus en ligne ou a-t-il une attention plus volatile ? Est-il prêt à participer à l'élaboration des contenus ou préfère-t-il être un simple spectateur ? C'est pour ce ou ces usagers-types que sera pensé et développé le projet web.
- Que font les autres ? Qu'existe-t-il dans ce domaine ? Aussi novateur et révolutionnaire soit un projet, il est fort probable que d'autres projets similaires aient déjà été développés et mis en ligne. Qu'il soit présenté comme un état de la concurrence ou une analyse comparative, l'inventaire de l'existant permet d'évaluer la faisabilité du projet, de mesurer son potentiel, et de lister les bonnes et les mauvaises idées glanées au fil des exemples choisis. La recherche doit être circonscrite (ciblée sur les sites qui semblent les plus intéressants) tout en restant ouverte (les exemples étrangers ou issus de domaines connexes ne sont pas à négliger) afin d'avoir un panorama pertinent et varié de ce qui a déjà été fait. Ce tour d'horizon est l'opportunité de se défaire des idées préconçues, de ne pas omettre les fonctionnalités indispensables et de se démarquer.

Le cahier des charges fonctionnel détaille également les moyens à disposition et les contraintes à prendre en compte pour la réalisation du projet :

- Les moyens humains pour constituer une équipe. Quelles sont les ressources humaines disponibles pour intervenir à chaque étape du projet (conception, développement, suivi, rédaction, exploitation, etc.) ? A-t-on besoin d'une aide complémentaire (éditoriale, technique, etc.) ou de formations ?
- Le budget. De combien d'argent dispose-t-on ?
- Les délais. Pour quand le projet doit-il être prêt ?

À ce stade, un nom de domaine peut déjà être avancé : c'est l'adresse qui permettra aux internautes d'accéder à la page d'accueil du site. Il doit être pertinent et facilement mémorisable. Il comprend aussi l'extension, à choisir en fonction de la localisation et de l'ampleur du projet ou du public ciblé. Par exemple : .com pour un positionnement à l'international (extension créée à l'origine pour les sites commerciaux, c'est aujourd'hui la plus utilisée), .ca pour le Canada, .fr pour la France, .org, .net, etc. Il est possible d'en choisir plusieurs. Chaque nom de domaine étant unique, sa disponibilité doit être au préalable vérifiée en ligne à l'aide d'une recherche *Whois*.

Choix et contenus éditoriaux : la charte éditoriale

Pour pouvoir choisir et modeler le type de site le plus approprié, un inventaire exhaustif des contenus qui devront y figurer doit être dressé.

En premier lieu, les contenus « principaux », liés directement au projet, sont ceux qui constitueront l'« essence » du site, sa raison d'être. Ils doivent être pensés spécifiquement

pour le web et peuvent être de tous types : texte, photo, illustration, vidéo, *podcast*, PDF, etc. Certains existent peut-être déjà et vont pouvoir être utilisés tels quels ou, dans la majorité des cas, adaptés pour le web ; d'autres vont devoir être créés et rédigés dans cette même optique.

En second lieu, les « métacontenus » fournissent des informations plus générales, contextualisent le site et ont trait à son fonctionnement :

- Mentions légales (obligatoires au Canada comme en France) : le nom du responsable de la publication, les coordonnées de l'éditeur et de l'hébergeur, le contact du webmestre (ou de l'agence qui a conçu le site), les droits d'auteur (sous quelle licence sont publiés les contenus ?), etc.
- Charte de la vie privée : informations quant à la politique de protection des données personnelles (par exemple, pour les adresses courriels collectées lors de l'inscription à l'infolettre).
- Crédits : nom et coordonnées de l'ensemble des personnes ayant contribué à la réalisation du site.
- Le cas échéant : partenaires, remerciements, FAQ (foire aux questions), conditions générales d'utilisation (CGU), conditions générales de vente (CGV), etc.

On peut citer également les textes préformatés, qui apparaissent lors de l'affichage des pages d'erreur (par exemple, la page erreur 404 identifiant une page qui n'existe pas ou plus) et dans les courriels automatiques (par exemple, les confirmations d'inscription ou de désinscription à l'infolettre), et qui peuvent être personnalisés pour améliorer l'expérience utilisateur.

À ces contenus s'ajoutent l'ensemble des fonctionnalités souhaitées : moteur de recherche, fil RSS (*Really Simple Syndication*), module de commentaires, options de notification

et de partage sur les réseaux sociaux, formulaire de contact, fil d'ariane, forum, comptes utilisateurs, etc.

Enfin, dans le cadre d'un projet web, le travail d'édition ne se limite pas (ou plus) au site en lui-même. La dissémination des contenus, l'éditorialisation élargie, fait elle-aussi partie intégrante de la mission éditoriale. La déclinaison du site sur d'autres supports (téléphone intelligent et tablette) et sa présence sur d'autres plateformes doivent ainsi être anticipées afin de maximiser les modes d'accès aux contenus et leur visibilité. Ces choix se font en pertinence avec l'objectif du projet, du public ciblé, des pratiques, des moyens disponibles, etc. Des profils peuvent être notamment créés au nom du projet sur les carrefours d'audience suivants :

- Réseaux sociaux (Facebook, Twitter, Google +, LinkedIn, Pinterest)
- Plateformes de partage de vidéo et photo (Youtube, Vimeo, Vine, Instagram, Flickr)
- Plateformes de *podcast* (iTunes U, Podcast addict)
- Blogs et micro-blogs (WordPress, Tumblr, Blogger)

Tous ces « objets » sont triés, catégorisés, organisés et liés pour créer des ensembles et sous-ensembles cohérents et interconnectés : le rubriquage. La hiérarchisation et le maillage de ces rubriques permettent alors de structurer le contenu et de faire le croquis d'une arborescence simple, logique et exhaustive pour le futur site.

La ligne éditoriale du projet est ensuite définie dans une charte éditoriale. Cette charte comprend l'ensemble des règles et des bonnes pratiques qui devront être suivies pour la production et la publication des contenus pour le web. Elle apporte une cohérence et une identité au projet, un

style, un ton. Tous les types de contenus (texte, vidéo, *podcast*, illustration, etc.) et de publications (sur le site, les réseaux sociaux, par courriel, etc.) doivent être spécifiés : le niveau de langage, la façon de s'adresser aux internautes, le style et la taille (nombre de signes ou de mots) des titres, des accroches et des textes, la longueur et le format des vidéos, les choix d'images, etc.

La charte détermine également la fréquence de publication, sur le site comme en dehors : envoi d'une infolettre, interventions sur les réseaux sociaux ou sur d'autres plateformes de partage. En fonction du public ciblé, le rythme ainsi que le choix du jour et de l'heure de ces publications sont déterminants pour gagner en visibilité et en efficacité. Une grille de programmation peut-être mise en place pour un meilleur suivi.

Par ailleurs, elle stipule les normes, caractéristiques du web, en matière d'ajout de liens internes et externes (renvoi vers d'autres pages du site ou d'autres sites) et de placement de mots clés dans le but d'optimiser la navigation mais aussi le référencement naturel, et donc le positionnement du site dans les résultats des moteurs de recherche.

Elle rappelle la législation en vigueur en matière de droits d'auteur et de droits des personnes ainsi que les mesures et précautions à prendre pour la respecter. Les membres du projet doivent être sensibilisés aux conditions liées à l'utilisation ou à la reproduction de tous types de contenus provenant d'autres sources : respect des droits, demande d'autorisation auprès des auteurs, citation de leurs noms, lien vers la publication d'origine, etc. Ils doivent aussi être informés des conditions de réutilisation par des tiers des contenus publiés sur le site : autorisée ou non,

modifiable ou non, pour usage commercial ou non commercial, citation de l'auteur, restriction par pays, etc. Le choix des licences devra ainsi être spécifié clairement : tous droits réservés (*copyright*), licence *creative commons*, domaine public. La responsabilité des contenus qui seront publiés revient à l'éditeur du site, tel qu'il est énoncé dans les mentions légales.

Enfin, elle mentionne les rôles de chaque acteur du site : qui rédige les contenus, qui les relit et les publie, qui en assure la diffusion, etc.

L'inventaire des contenus, leur organisation et leur maillage ainsi que l'ensemble de ces principes seront à l'origine des choix graphiques et techniques. Ils pourront ensuite être ajustés, au fur et à mesure, en fonction des retours, de l'expérience et des analyses statistiques.

Choix techniques et graphiques : le cahier des charges technique et la charte graphique

En fonction du cahier des charges fonctionnel (objectif, besoins exprimés, ressources disponibles, etc.), le responsable technique (ou le maître d'œuvre, MOE) définit, dans un cahier des charges technique, les choix qui vont structurer le site : l'architecture, les langages ainsi que les outils et technologies qui seront utilisés. Il y décrit le futur *front office*, soit le site tel qu'il apparaîtra en ligne à l'internaute (graphisme, arborescence, etc.), et le *back office*, l'outil d'administration qui sera utilisé par l'équipe éditoriale pour alimenter et mettre à jour le site.

Un premier choix doit être fait entre une solution logicielle propriétaire et une solution logicielle *open source*.

Avec des logiciels propriétaires peut être développée une plateforme sur mesure, entièrement adaptée à la demande. Cette option nécessite une solide équipe technique (agence et prestataires externes ou service informatique intégré) et une longue phase de programmation, puisque l'ensemble du site doit être construit de A à Z (*front office* comme *back office*). Elle aura l'avantage de proposer une structure unique, un site « haute-couture », mais l'inconvénient de créer une forte dépendance vis-à-vis de l'équipe ayant développé le projet. À moins d'un lourd transfert de compétences et de la cession des droits du code source, elle seule pourra effectuer la maintenance et les mises à jour du site.

La solution logicielle *open source* – comme les systèmes de gestion de contenus (CMS, *Content Management System*) Drupal, Spip, Joomla, WordPress, etc. – permet d'utiliser des interfaces déjà développées, dont le code et les sources sont en accès libre. Ces interfaces sont entièrement configurables et personnalisables (*front office* et *back office*) afin de s'adapter aux besoins et à l'identité du projet : habillage graphique, choix des modules, ajout de fonctionnalités, etc. Le recours à l'*open source* permet une plus grande indépendance face à l'équipe technique, un délai de développement plus court et des évolutions plus fréquentes. Qui dit solution *open source* dit également le soutien possible d'une large communauté en ligne, à même de partager son expérience et de s'entraider techniquement à travers des questions/réponses, tutoriels, etc. En revanche, il est nécessaire d'adhérer à la philosophie et au cadre de la solution choisie, qui doivent être adaptés au projet et à ses exigences. C'est également faire un pari sur sa stabilité et sa pérennité à long terme.

Le choix du type d'hébergement est également déterminant. Le site, et toutes les données qui le constituent, doit être stocké (hébergé) sur un serveur connecté en permanence sur Internet. L'hébergeur garantit l'accessibilité du site, sa sécurité et sa sauvegarde. Il peut fournir un serveur mutualisé, dont les ressources (espace disque, bande passante) sont partagées par plusieurs « clients ». Cette solution, plus économique (les coûts étant eux aussi partagés), est aussi la moins stable : un pic de trafic sur un site peut entraîner l'inaccessibilité des autres sites hébergés sur le serveur. Il crée également une dépendance vis-à-vis de l'hébergeur, qui prend en charge l'assistance mais garde le contrôle sur ses machines et sur les technologies utilisées. Un serveur dédié (à chaque « client » est attribué un serveur) aura quant à lui l'avantage, non négligeable pour les sites à forte audience, d'offrir un trafic illimité et de pouvoir être techniquement personnalisé. Il sera en revanche plus coûteux et nécessitera certaines compétences techniques. Il est également possible d'utiliser un serveur privé, à « domicile », qui ne sera lié à aucun hébergeur mais nécessitera une connexion Internet puissante et stable ainsi que des connaissances techniques approfondies.

Le nom de domaine préalablement choisi doit être inscrit au registre des noms de domaine. Il peut être réservé auprès d'un bureau d'enregistrement ou registraire (*registrar*) accrédité auprès d'un registre (*registry*) : ACEI (Autorité canadienne pour les enregistrements Internet) pour les noms de domaine se terminant par .ca, AFNIC (Association française pour le nommage Internet en coopération) pour les .fr, VeriSign pour les .com et .net, Public Interest Registry

pour les .org. Le registraire peut être un hébergeur, un fournisseur d'accès à Internet, une agence web, etc. C'est à lui que revient la gestion du nom de domaine (enregistrement, pérennité). Il s'agit d'une location payante, pour une durée déterminée, qui doit être renouvelée périodiquement.

Concernant le *front office*, plusieurs propositions d'habillage graphique sont présentées. Après validation d'une des propositions, une charte graphique est rédigée. Celle-ci liste l'ensemble des règles et codes visuels qui devront être respectés pour garantir la cohérence et l'homogénéité graphique du projet, son identité visuelle. Elle comprend notamment les règles d'utilisation des couleurs et des polices, le style graphique et iconographique (choix des pictogrammes et illustrations, formats), les principes de navigation, l'organisation visuelle des pages (*zoning*), etc. : autant de repères qui permettront à l'internaute d'évoluer avec aisance dans les pages du site et lui donneront envie de prolonger sa visite. S'il existe déjà une charte graphique au sein de la structure à l'origine du projet, la charte graphique du site web peut en être une déclinaison ou une prolongation.

L'élaboration de la charte graphique est, de plus en plus, intimement liée à un travail réalisé sur l'ergonomie dont la visée est d'optimiser l'expérience utilisateur en rendant l'interface intuitive et simple d'utilisation. Grâce à l'analyse des usagers-types (tels que définis plus tôt) et de leurs pratiques, l'architecture de l'information et la structuration des pages seront adaptées afin de gagner en lisibilité, en accessibilité et donc en efficacité.

Sur ces bases, et celles de l'arborescence simplifiée réalisée lors de l'inventaire des contenus, une arborescence détaillée

peut être élaborée. Elle liste l'ensemble des pages et présente schématiquement la manière dont l'information sera organisée et articulée.

À l'issue de ces choix, un calendrier est préétabli, planifiant les phases successives de la réalisation du projet : habillage graphique, développement et intégration technique, saisie des contenus, mise en ligne, sans oublier les tests et les étapes de validation. Chaque phase est chiffrée (nombre de jours travaillés par nombre de personnes sollicitées) pour dresser un budget prévisionnel.

Lorsque les différentes parties impliquées se sont entendues sur les propositions technique et graphique, le calendrier et le budget, le travail de développement peut commencer.

Développement et mise en ligne

L'équipe technique développe le site conformément aux décisions prises préalablement.

La structure du site (comme celle des contenus) doit être pensée stratégiquement pour garantir un bon référencement, c'est-à-dire un bon positionnement dans les résultats des moteurs de recherche, principales sources d'accès au site. Le SEO (*Search Engine Optimization*, optimisation pour les moteurs de recherche) consiste à mettre en place un ensemble d'éléments normés qui permettront aux robots d'indexation de lire et de comprendre de manière optimale le site et son contenu. Ces bonnes pratiques amélioreront le référencement naturel, offrant au site une plus grande visibilité et de meilleures chances d'atteindre le plus grand nombre d'internautes. On peut citer, entre autres choses, l'écriture et la réécriture des URL (utilisation de mots clés,

bannissement des caractères spéciaux), le *sitemap* et les métadonnées.

Pour pouvoir mesurer l'audience du site et l'évolution de son trafic, un outil d'analyse statistique doit être implémenté (du type *Google Analytics* ou Xiti). La majorité des CMS fournissent leurs propres outils.

Dès lors que la partie *back office* est opérationnelle, l'équipe éditoriale peut entamer la saisie des contenus, préalablement adaptés ou créés pour le web.

Une période de tests est ensuite planifiée afin d'éprouver la pertinence de l'arborescence et de la navigation, de détecter et réparer les éventuels *bugs*, de vérifier les liens internes et externes, de pallier les oublis et de corriger les coquilles.

Puis, la mise en ligne peut avoir lieu : le site devient accessible à tous. Idéalement, elle a lieu suffisamment en amont du lancement officiel pour, d'une part, prendre le temps de régler les derniers détails et faire les derniers tests et, d'autre part, permettre aux moteurs de recherche de scanner le site et ainsi initier son référencement naturel (ce qui prend habituellement quelques jours).

La diffusion de l'information sur tous les canaux peut alors commencer : annonce sur les réseaux sociaux, envoi de courriels, communication traditionnelle (relations de presse et publiques), etc.

Suivi du projet : la phase d'exploitation

La mise en ligne du site ne marque pas la fin du projet. Son succès dépend tout autant de l'activité de veille et du suivi qui lui est apporté en phase d'exploitation.

Une veille statistique, ciblée sur les principaux indicateurs (nombre de visites et de pages vues, temps passé sur le

site, rubriques les plus consultées, taux de rebond, sources des visites, mots clés utilisés, etc.) ainsi qu'un suivi du référencement donnent de précieuses informations sur le trafic, les usages liés au site et l'intérêt qu'il suscite. Ces informations sont primordiales pour ajuster le discours et améliorer l'expérience utilisateur.

L'enrichissement constant ainsi que l'animation des contenus créent une relation avec l'internaute l'incitant à revenir régulièrement, poussent les moteurs de recherche à scanner le site plus souvent, et accroissent le référencement comme le trafic : mise à jour des textes, ajout de nouveaux articles et de liens, mise en avant événementielle sur la page d'accueil, envoi d'une infolettre, prise en compte des suggestions des internautes, réponse aux commentaires, etc. Une veille éditoriale constante permettra de s'adapter aux pratiques et aux usages en temps réel.

En complément, d'autres techniques de référencement peuvent alors être mises en place, comme l'achat de mots clés sur les moteurs de recherche (liens sponsorisés) : c'est ce que l'on appelle le SEM (*Search Engine Marketing*).

La valorisation des contenus, grâce à une éditorialisation externe mesurée et continue sur les grands carrefours d'audience (notamment les réseaux sociaux), apportera au site une large exposition, assurant également sa visibilité et sa pérennité. Le relai des actions décrites ci-dessus, l'instauration d'un dialogue, le rebond sur l'actualité, l'appel à contribution ainsi qu'une veille active (suivre ce qui se dit sur le site) sont autant d'occasions de communiquer sur ces terrains.

D'un point de vue technique, maintenance et veille informatiques doivent être assurées (mises à jour, correction des

bugs), ainsi que, à plus long terme, la mise en place d'évolutions techniques ou l'ajout de nouvelles fonctionnalités.

Un projet éditorial conçu pour le web est un projet ouvert, nécessitant un travail d'éditorialisation constant. Un site « clos », aux contenus figés, est en effet destiné à plonger dans l'oubli : celui des internautes comme celui des robots d'indexation des moteurs de recherche, qui ne viendront plus le visiter. Les objets éditoriaux se transforment, les habitudes de lecture et de consultation aussi ; il en est de même, par conséquent, pour le métier de l'éditeur.

Table des matières

Autre titre de la collection
« Parcours numériques »

Maurizio Ferraris, *Âme et iPad*, 2014

La collection Parcours numériques est accessible gratuitement en édition augmentée sur parcoursnumeriques-pum.ca.

L'apparition du numérique a entraîné ces dernières années une transformation profonde des modèles de production et de circulation des livres, qui ont peu changé depuis le XVIII^e siècle. Le web, en particulier, a provoqué une remise en question du sens même du partage des connaissances : d'une économie de la rareté, nous sommes passés à la surabondance. Auparavant, une poignée d'institutions centralisatrices, privées et publiques, étaient garantes du choix, de l'évaluation et de la distribution des contenus ; aujourd'hui, il n'y a plus de systèmes de légitimation, ou alors ils sont déstructurés. Après avoir fait le constat de la crise de ces modèles et de la difficulté d'en proposer de nouveaux, ce livre présente les enjeux et les défis complexes du nouveau monde de l'édition numérique.

Michaël E. Sinatra est professeur au Département d'études anglaises de l'Université de Montréal.

Marcello Vitali-Rosati est professeur adjoint de littérature et culture numérique au Département des littératures de langue française de l'Université de Montréal.

ISBN 978-2-7606-3202-8

9 782760 632028

14,95 $ • 13 €