2 | 2020

ABD BVD

ASSOCIATION BELGE DE DOCUMENTATION
BELGISCHE VERENIGING VOOR DOCUMENTATIE

Bladen voor **DOCUMENTATIE**
Cahiers de la **DOCUMENTATION**

Trimestriel | Driemaandelijks Juin | Juni

Dossier
L'archive photographique
Het fotografisch archief

Dossier
Saving the web

ISSN 0007-9804
Éditeur responsable | Verantwoordelijke uitgever: Samuel Piret, Rue de la Concorde 56 Eendrachtstraat - 1050 Ixelles - Elsene

2 | 2020

ABD BVD
ASSOCIATION BELGE DE DOCUMENTATION
BELGISCHE VERENIGING VOOR DOCUMENTATIE

Bladen voor DOCUMENTATIE
Cahiers de la DOCUMENTATION

Trimestriel | Driemaandelijks Juin | Juni

Pour tout renseignement sur les *Cahiers de la documentation*
ou pour soumettre un article :
Voor alle inlichtingen over de *Bladen voor documentatie*
of om een artikel voor te stellen:

cahiers-bladen@abd-bvd.net

Sommaire
Inhoudstafel

74ème année - 2020 - n° 2 74de jaargang - 2020 - nr 2

Les articles des numéros 1995/1 à 2019/2
sont disponibles à l'adresse :
http://www.abd-bvd.be/fr/publications/cahiers-de-la-documentation

De artikels van de nummers 1995/1 tot 2019/2
zijn beschikbaar op :
http://www.abd-bvd.be/nl/publicaties/bladen-voor-documentatie

ÉDITORIAL

WOORD VOORAF

par / door
Vincent DUTERME
Rédacteur en chef / Hoofdredacteur

Voici le premier numéro des *Cahiers* post confinement. Nous avons vécu ce printemps 2020, une épreuve inattendue avec l'apparition du Covid-19. Quelque chose qui ressemble à une hibernation printanière. Une assignation à résidence qui nous fait réfléchir sur nous-même et nous pousse à nous adapter et à nous réinventer d'autres formes de vie en société et de communication. Un moment aussi où la culture est mise à rude épreuve : la fermeture des musées des salles de spectacles, de théâtre et de cinéma, mais aussi la fermeture des bibliothèques : tout ce qui permet à l'expression artistique mais aussi à la connaissance de s'exprimer. *Mens sana in corpore sano*, nous apprend que l'entretien de l'esprit est tout aussi important que notre santé. Il nous faut voir au-delà et pouvoir en sortir meilleurs.

Réinventer d'autres formes de communication et d'expression marque ces deux derniers siècles, de la photographie au début du XIXème siècle à l'internet développé à la fin du XXème siècle. Ce sont précisément ces deux formes d'expression de communication que nous abordons dans ce numéro.

Après la peinture, l'écriture et la musique, la photographie apparue au début du XIXème siècle est le premier moyen d'expression sui generis : un instrument de communication qui révolutionnera notre époque et la fera rentrer dans l'ère de l'instantané. Chaque moyen d'expression nécessite son archivage et fera l'objet de la première partie des Cahiers.

L'autre tournant de notre époque est ce média sans lequel il ne nous est plus possible de communiquer : le Web. De son apparition à la fin du XXème siècle jusque maintenant, son développement a été exponentiel. Avec l'Internet, des milliards de gens sont connectés et le monde devient une gigantesque bibliothèque.

Le dossier sur l'Archive photographique, reprend le Doc'Moment idoine de novembre 2019 : Nico Gastmans a évoqué le fonds important de l'Agence de Presse Van Parys ; Erik Buelinkx et Julie Mauro ont présenté leur travail au sein de l'Institut National du Patrimoine artistique (IRPA). Il se termine par un article de Guy Delsaut sur le fonds

Dit is het eerste nummer van de *Bladen* na de lockdown. Het afgelopen voorjaar is door de komst van Covid-19 een onverwachte beproeving gebleken. Het leek wel een winterslaap in de lente. Het verplicht thuisblijven heeft geresulteerd in veel tijd voor bezinning en heeft ons gedwongen ons aan te passen en te zoeken naar nieuwe vormen van samenleven en communicatie. Het is een tijd waarin ook de cultuur zwaar op de proef is gesteld: sluiting van musea, toneelzalen, theaters en bioscopen, maar ook van bibliotheken, kortom van alles wat het tonen van kunst, maar ook het kennisnemen ervan, mogelijk maakt. De uitdrukking *Mens sana in corpore sano* leert ons dat onze geestelijke gezondheid even belangrijk is als de lichamelijke. We moeten vooruit kijken en proberen hier beter uit te komen.

Het zoeken naar nieuwe vormen van communicatie en expressie is kenmerkend voor de afgelopen twee eeuwen, van de fotografie begin 19e eeuw tot de ontwikkeling van internet eind 20e eeuw. Precies die twee expressie- en communicatievormen komen in dit nummer aan de orde.

Na de schilderkunst, het schrijven en de muziek, is de fotografie aan het begin van de 19e eeuw het eerste expressiemiddel van een geheel andere orde: een communicatie-instrument dat radicale veranderingen teweegbrengt met het inluiden van het instanttijdperk. Elk expressiemiddel verlangt zijn eigen archivering. Hierover gaat het eerste deel van de Bladen.

Het andere medium dat in ons tijdperk in een kentering heeft geresulteerd en dat inmiddels onontbeerlijk is voor onze communicatie, is internet. Vanaf de verschijning aan het einde van de 20e eeuw tot nu toe heeft internet een exponentiële ontwikkeling doorgemaakt. Dankzij internet worden miljarden mensen onderling verbonden en wordt de wereld één enorme bibliotheek.

In het dossier over het fotografisch archief wordt het Doc'Moment van november 2019 aangehaald: Nico Gastmans noemt de belangrijke collectie van het Persagentschap Van Parys. Erik Buelinkx en Julie Mauro presenteren vervolgens

d'archives d'images développé par Wikimedia Commons.

Nico Gastmans présente le fonds important constitué d'un siècle d'archives photographiques de l'agence de Presse Van Parys et constitué des supports successifs depuis l'invention de la photographie (plaque de verre, négatif noir et blanc et négatif couleur). Le fonds représente non seulement toute l'histoire du XXème siècle, mais aussi une révolution dans l'information. Germaine Van Parys, la fondatrice de la première agence de presse photographique a fait figure de précurseuse en inaugurant un type d'agence de presse entièrement nouveau et surtout en étant la première femme fondatrice d'une agence de presse. L'Agence de presse photographique Van Parys servit de modèle à l'ensemble des agences de presse photo actuelles.

Erik Buelinkx et Julie Mauro nous font connaitre l'IRPA, Institut royal du Patrimoine Artistique. Riche d'un fonds d'un million de négatifs, celui-ci est numérisé, identifié et conservé en interne par la cellule DIGIT créée en 2013 dans le cadre du projet lancé par Belspo. Le projet DIGIT a pour objectif également la numérisation du fonds des dossiers d'intervention de l'IRPA.

Enfin dans le dernier article du Dossier Archive photographique, Guy Delsaut présente le fonds Wikimedia Commons : un fonds de plus de 60 millions de médias dont plus de 57 millions d'images décrites et catégorisées, où chacun des contributeurs peut enrichir le fonds de nouvelles photos ou enrichir les descriptifs de ces photos. On montre également comment utiliser ces photos légalement.

Cet article fait la transition avec le second dossier : les archives du Web.

Le deuxième dossier est quant à lui consacré à l'archivage du Web. Le Web constitue l'autre tournant dans l'information opéré à la fin du XXIème siècle. L'archivage de ce média est développé dans cinq différents articles constituant le dossier.

Niels Brügger décrit l'archivage du Web et son intérêt pour la recherche. L'article indique les différences d'archivage entre les Etats Unis et l'Europe. Mais aussi la différence entre les pays procédant à un archivage du Web et d'autres pas.

Dans une interview coordonnée par Nadège Isbergue, Friedel Geeraert et Sally Chambers, Niels Brügger et Valérie Schafer évoquent le projet PROMISE qui a été développé lors du colloque Saving the Web: the Promise of a Belgian Web Archive. L'accent y est mis sur l'importance de l'archivage Web comme source d'information et de recherche.

hun werk binnen het Koninklijk Instituut voor het Kunstpatrimonium (KIK). En tot slot een artikel van Guy Delsaut over de door Wikimedia Commons ontwikkelde collectie beeldarchieven.

Nico Gastmans presenteert de grote collectie van het Persagentschap Van Parys, bestaande uit een eeuw aan fotografische archieven en de achtereenvolgende dragers sinds de uitvinding van de fotografie (glasplaat, zwart-witnegatief en kleurnegatief). De collectie toont niet alleen de hele historie van de 20e eeuw, maar ook een revolutie in de informatiewereld. Germaine Van Parys, oprichtster van het eerste fotopersagentschap, was baanbrekend, niet alleen doordat ze een geheel nieuw soort persagentschap heeft opgezet maar vooral doordat ze de eerste vrouw was die een persagentschap stichtte. Het Persagentschap Van Parys staat model voor alle huidige fotopersagentschappen.

Erik Buelinkx en Julie Mauro laten ons kennismaken met het KIK, het Koninklijk Instituut voor het Kunstpatrimonium. Dit instituut beschikt over een collectie van een miljoen negatieven die wordt gedigitaliseerd, geïdentificeerd en intern opgeslagen door de Cel Digitalisering, die in 2013 is opgericht in het kader van het project van Belspo. Het DIGIT-project heeft ook tot doel de collectie interventiedossiers van het KIK te digitaliseren.

Ten slotte presenteert Guy Delsaut in het laatste artikel van het Dossier fotografisch archief de collectie Wikimedia Commons: een collectie van ruim 60 miljoen mediabestanden, waaronder ruim 57 miljoen afbeeldingen, die zijn beschreven en gerubriceerd. Iedereen kan er een bijdrage aan leveren door nieuwe foto's aan te dragen of de beschrijvingen van de foto's aan te vullen. Ook is te zien hoe deze foto's legaal kunnen worden gebruikt.

Dit artikel vormt tevens de brug naar het tweede dossier: internetarchieven.

Het tweede dossier is gewijd aan de archivering van internet. Eind 20e eeuw heeft internet tot de andere kentering in de informatiewereld geleid. Aan de archivering van dit medium wordt aandacht besteed in de vijf verschillende artikelen van dit dossier.

Niels Brügger beschrijft hoe de archivering van internet plaatsvindt en wat het belang ervan is voor het onderzoek. In zijn artikel komen de verschillen in archivering tussen de Verenigde Staten en Europa aan de orde. Maar ook het verschil tussen de landen waar internetarchivering plaatsvindt en de landen waar dat niet het geval is.

In een interview dat is gecoördineerd door Nadège Isbergue, Friedel Geeraert en Sally Chambers,

Alejandra Michel nous fait partager l'archivage du Web sous l'angle juridique. L'importance est donnée à la préservation de ces archives et la mise à disposition pour les chercheurs et le public en général

Patricia Blanco évoque les différentes étapes du processus d'archivage du Web toujours dans le cadre du projet PROMISE. L'objectif étant de connaitre les besoins et les attentes des chercheurs mais aussi l'accès technique pour l'accessibilité de ces archives.

Avec le déconfinement, l'ABD reprendra progressivement ses activités phares : Inforum, Doc'Moment et événement commercial. Guy Delsaut nous a présenté le premier Doc'Moment sur Youtube[1] concernant Wikimedia Commons dans la continuité de son article dans ce numéro. En attendant je vous souhaite une bonne lecture de ce numéro. Et prenez soin de vous.

noemen Niels Brügger en Valérie Schafer het project PROMISE, dat is ontwikkeld tijdens de conferentie Saving the Web: the Promise of a Belgian Web Archive. Het accent wordt gelegd op het belang van de archivering van het web als informatie- en onderzoeksbron.

Alejandra Michel laat ons naar internetarchivering kijken vanuit een juridisch oogpunt. Hierbij staat het behoud van de archieven en de beschikbaarheid voor onderzoekers en het algemene publiek voorop.

Patricia Blanco noemt de verschillende stadia van het archiveringsproces, ook in het kader van het project PROMISE. Het doel is de behoeften en verwachtingen van onderzoekers in kaart te brengen, maar ook de technische toegang in verband met de toegankelijkheid van de archieven.

Nu de lockdown wordt afgebouwd, zal de BVD langzaamaan haar kernactiviteiten weer oppakken: Inforum, Doc'Moment en commercieel evenement. Guy Delsaut presenteerde het eerste Doc'Moment op Youtube[1] over Wikimedia Commons als vervolg op zijn artikel in dit nummer. Ondertussen wens ik u veel leesplezier met dit nummer. En zorg goed voor uzelf.

1. <https://youtu.be/5et83wMyHZg>

DOSSIER

L’archive photographique
Het fotografisch archief

Le "Plan Marshall" a 70 ans

Le 3 avril 1948, le président américain Truman signe le "Plan Marshall", gigantesque programme de prêts aux États européens, en l’échange de l’achat de biens américains. Truman qui doutait d’être réélu en novembre, proposa à Eisenhower, héros de la seconde guerre mondiale, d’être candidat à sa place, ce que ce dernier refusa. Truman sera réélu de justesse.

Septembre 1945. Le général Eisenhower, commandant des forces alliées US en Europe et qui organisa le débarquement de Normandie, salue la foule depuis le balcon de l’hôtel de ville, Grand-Place à Bruxelles.

Cette photo montre les photographes présents lors de l’événement.

Het "Marshallplan" is 70 jaar oud

Op 3 april 1948 ondertekende de Amerikaanse president Truman het "Marshallplan", een gigantisch programma van leningen aan Europese staten in ruil voor de aankoop van Amerikaanse goederen. Truman, die twijfelde of hij in november zou worden herkozen, stelde Eisenhower, een held uit de Tweede Wereldoorlog, voor in zijn plaats te komen, wat deze laatste weigerde. Truman zal ternauwernood herkozen worden.

September 1945. Generaal Eisenhower, commandant van de Amerikaanse geallieerde strijdkrachten in Europa en die de landing in Normandië heeft georganiseerd, begroet de menigte vanaf het balkon van het stadhuis, Grote Markt in Brussel.

Deze foto toont de aanwezige fotografen op het evenement.

L'ARCHIVE PHOTOGRAPHIQUE AU SEIN D'UNE AGENCE DE PRESSE PHOTO

Nico GASTMANS
Business Development Manager, GermaineImage

Article rédigé suite à la conférence donnée par l'auteurs dans le cadre du Doc'Moment "L'archive photographique", le 20 novembre 2019 à Bruxelles.

Het artikel is opgesteld naar aanleiding van een conferentie gegeven door de auteur in het kader van een Doc'Moment "Het fotografisch archief" op 20 november 2019 te Brussel.

■ Nous allons développer ce sujet en nous basant sur l'histoire de deux femmes qui vont s'imposer dans le monde de la presse en Belgique et leur descendance : d'une part Germaine Van Parys (1893–1983) et Odette Derèze (1932), les deux premières femmes photographes de presse en Belgique, et d'autre part Tom Gastmans (1964) et Nico Gastmans (1968) qui continuent à faire vire cette archive photographique.

■ We behandelen dit onderwerp op basis van de geschiedenis van twee vrouwen die zich hebben doen gelden in de Belgische perswereld en hun opvolgers: enerzijds Germaine Van Parys (1893–1983) en Odette Derèze (1932), de eerste twee vrouwelijke persfotografen in België, en anderzijds Tom Gastmans (1964) en Nico Gastmans (1968), die het voortbestaan van dit fotografische archief verzekeren.

Germaine Van Parys débute sa carrière à la toute fin de la Première Guerre mondiale. Très déterminée et d'une rare persévérance, elle parvient à se faire engager rapidement après ses débuts dans le monde professionnel de la presse belge et à travailler pour des quotidiens tels La Meuse ou Le Soir. Travailler pour ces quotidiens lui ont permis d'obtenir la carte de presse, un sésame très important à l'époque et qui ouvrait quasiment toutes les portes.

Vers la fin des années 1920, enfin reconnue par le milieu des photographes de presse et sa carte de presse officielle en poche, elle se met à son compte.

Femme aussi très clairvoyante, elle a investi pendant les premières années de sa carrière dans l'achat des supports photographiques de l'époque... les clichés en verre. De ce fait, et très légalement toutes les photos réalisées pour les différents quotidiens lui appartenaient définitivement. Une manière ingénieuse pour constituer son archive de base pour ses activités d'indépendante. Cette base, ainsi que son opiniâtreté, lui procure donc rapidement des rentrées financières confortables. Une femme photographe peu commune et une femme d'affaires sont sur le marché !

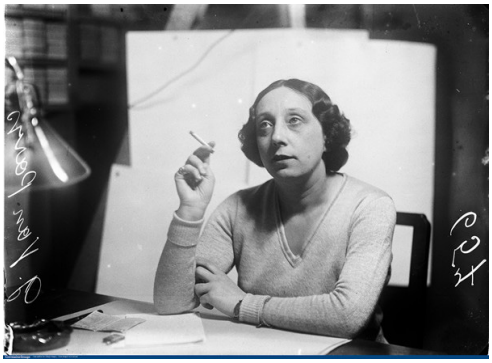

Fig. 1 : Photo de Germaine Van Parys

Début des années 1930, elle devient cofondatrice avec Victor Hennebert de l'Association belge des journalistes photographes. Une double première importante pour la profession ! D'une part, la naissance d'une Association reconnue qui va fédérer le métier et d'autre part la confirmation de Germaine comme égale de ses collègues masculins. Pour l'époque, un joli tour de force.

À l'époque, le métier de photographe de presse était très largement réservé à la gent masculine, et ce notamment à cause du poids et de l'encombrement du matériel de prise de vue. N'oublions pas non plus qu'à l'époque, les négatifs sont des clichés en verre qui pèsent un certain poids. Cela en plus évidemment des a priori classiques hommes/femmes dans le milieu du travail.

Germaine Van Parys ne passe pas inaperçue et elle se fait rapidement remarquer par les personnalités diverses et les clients, tous intrigués de voir une femme parmi "la meute" des photographes.

En effet, les photos et les reportages qu'elle réalise sont empreints d'une touche féminine, une marque de fabrique propre qui fera la différence avec ses concurrents masculins. Cette "touche féminine" va même convaincre un certain nombre de clients de

lui confier la réalisation de reportages ou de photos. Ainsi, elle tissera des liens privilégiés avec :

- le Théâtre Royal de la Monnaie,
- le Palais des Sports de Bruxelles et le Sportpaleis d'Anvers,
- les grands magasins Innovation,
- ViewMaster, la célèbre société éditant les premiers disques photo à visualiser pour le grand public

Ces commandes de reportages pour les clients précités lui assurent une bonne base de revenus réguliers et ceux-ci lui permettent alors de financer des reportages dits "de presse" où là, il n'y a pas l'assurance de vendre des photos aux clients (quotidiens, magazines ...). Sur ce marché-là, c'est le plus rapide et la meilleure photo qui l'emportent sur tout le reste.

Au cours de ces reportages, elle attire l'attention de la famille royale avec qui elle nouera des contacts fort étroits qui lui permettront de toujours avoir une bonne place lors des événements royaux. Événements que Germaine Van Parys couvrira tout au long de sa carrière et feront de ses archives une référence en matière d'archives royales. Ce travail sera d'ailleurs poursuivi par Odette Derèze qui reprendra ses affaires.

Germaine Van Parys est aussi une femme extravagante et définitivement extravertie. Sa maison devient au fil du temps un lieu de rencontre pour artistes de tout genre. Son attrait pour l'art, la culture et l'originalité la met aussi en contact avec Jeanne Walschot (femme collectionneuse d'art africain) et par ce biais, Germaine Van Parys fera son premier voyage vers le Congo en 1938.

La seconde guerre mondiale ne va pas entamer sa motivation, au contraire, elle continuera ses activités en tant que photographe, mais refusera de travailler pour l'occupant. Elle assurera les reportages photo du Secours d'hiver de manière officielle et travaillera pour la Résistance en même temps. Elle sera promue à la fin de la guerre commandant de l'armée de la Résistance par le Général Piron.

Fig. 2 : Photo d'Odette Derèze

Au fil des ans, une agence de presse se crée, se développe et prend racine dans le paysage de la photo de presse en Belgique.

En 1952, après deux ans de stages afin d'obtenir sa propre carte de presse, Odette Derèze vient rejoindre l'Agence Germaine Van Parys et c'est elle qui reprendra le flambeau dans les années 1960, lorsque Germaine Van Parys sera obligée d'arrêter ses activités pour raisons de santé.

Avec l'arrivée d'Odette Derèze, l'agence continue à se développer et devient l'Agence Van Parys. Puis un nouveau virage est pris lorsque l'Agence belge s'inscrit dans un réseau d'agences de presse internationales en s'associant avec la grande agence française: Reporters Associés (la première agence de presse photo française et qui s'étend au niveau mondial).

Du fait de cette alliance, l'Agence Van Parys va croître d'une part sur le marché national en pouvant offrir aux clients des photos ou des reportages venant de l'étranger (actualités, vedettes, reportages de fond, ...) et d'autre part sur le marché international en distribuant ses productions propres dans de nombreux pays, et ce au travers du réseau des Reporters Associés. Cette relation entre agences, basée sur l'échange de matériel photographique, va se poursuivre pendant de nombreuses années, jusqu'aux années 2000 et l'avènement de la digitalisation qui va bousculer les habitudes de distribution.

La croissance des différents marchés (national et international) va entraîner une croissance de la demande pour l'Agence Van Parys et par conséquence, ceci va augmenter au fur et à mesure le niveau de production de "photos belges", y compris des voyages à l'étranger pour couvrir des sujets typiquement belges.

Ainsi, différents photographes salariés et indépendants vont commencer à travailler pour et avec l'Agence.

En 1983, l'Agence change de nom et devient Van Parys Media et se transforme en même temps en une société anonyme.

À ce moment, cette petite entreprise occupe une quinzaine de salariés.

C'est aussi à ce moment que l'entreprise commence à s'informatiser... facturation, relevés des ventes vers les agents étrangers et photographes, archivages des photos déjà un petit pas vers le "tout digital".

Ayant été à bonne école, Odette Derèze développe deux axes dans son Agence.

D'une part, la production de photos relatives à des événements belges (en Belgique et/ou à l'étranger) et qui présente de l'intérêt pour le marché belge et international:

- CEE[1] (collaboration étroite avec le service de presse de la CEE et réalisation de commandes spéciales),
- OTAN (visite des chefs d'État ...)
- famille royale (y compris la couverture de tous les voyages à l'étranger de la famille royale)
- acteurs et actrices belges
- toute personnalité connue étrangère venant en Belgique

Et d'autre part la distribution de photos venant de fournisseurs étrangers :

- Grande Bretagne + international => Camera Press, Nunn Syndication, Retna UK,
- France : Imapress (production locale et internationale), Sygma (production locale et internationale), Presse-Sports / L'Équipe (production locale et internationale),
- États-Unis d'Amérique: Retna USA, Shooting Stars (vedettes, séries télévisées, films), Visages (portraits studio d'acteur ou de personnalités connues),

Pour ne nommer que les plus grandes ou les plus symboliques.

Cette activité de distribution deviendra d'ailleurs une "spécialité maison" qui fera la réputation de la société.

La distribution d'archives photo de tiers requiert d'une part de nombreux contacts internationaux et un solide carnet d'adresses, des déplacements et d'autre part l'accès aux clients sur le marché où ces représentations sont proposées.

Mais cela correspond aussi à la mise en place d'un réseau de distribution de la production propre vers l'étranger et une clientèle internationale bien souvent très intéressante au niveau pécuniaire.

En effet, le prix d'une photo à cette époque était déterminé par les éléments suivants :

- la rareté ou l'exclusivité de la photo,
- le cadre de la prise de vue, shooting studio ou prise de vue sur le vif d'un événement
- le tirage du support dans lequel est imprimée la photo,

De ce fait, les prix obtenus dans de grands marchés sont plus intéressants que ceux obtenus en Belgique francophone ou néerlandophone.
Ainsi, dans les années 1980, Van Parys Media dispose d'une distribution dans 45 pays.

Les archives de Van Parys Media (les propres et celles des différentes représentations étrangères) vont grandir très rapidement et même de manière quasi exponentielle au fil du temps. Et étant donné que nous sommes encore du temps de l'analogique, cela représente un volume très conséquent qui dépasse de loin les 100 m^3.

Ajoutez les espaces de bureau, les tables lumineuses, les zones de sélection, les bureaux des commerciaux, de la comptabilité et de la direction et bien entendu un laboratoire complet pouvant traiter le noir et blanc et la couleur, un studio de prise de vue et vous aurez une idée de la surface nécessaire à cette activité un peu particulière...

Ces archives sont constituées des clichés de verres de l'époque et de certains tirages réalisés, des négatifs noirs et blancs, des tirages qui vont avec et de diapositives.

Juste avant l'avènement du digital, l'analogique fera un dernier soubresaut avec le négatif couleur et les tirages couleurs.

Voici quelques chiffres indicatifs qui permettent de se faire une idée du volume de l'archive analogique brute produite depuis 1918 :

- Clichés de verre – environ 35.000 unités
- Sheet film N/B – environ 7.000 unités
- Sheet film couleur – environ 5.000 unités
- Négatifs 24x36 mm N/B – environ 750.000 unités
- Positifs 24x36 mm couleur – environ 750.000 unités
- Négatifs 24x36 mm couleur environ – 30.000 unités

Soit un total d'approximativement 1.577.000 originaux. Selon nos différentes expériences, environ 10% de ce volume est en train de rejoindre l'archive digitale.

Odette Derèze mettra un terme à sa carrière de photographe professionnelle en 1993, année du décès du Roi Baudouin. Néanmoins, elle restera encore active quelques années afin d'aider et de transmettre les rênes de la société à Tom Gastmans (son fils aîné).

Ce dernier poursuivra les efforts de développement de la société auprès de la clientèle de presse, des maisons d'édition mais développera en plus un nouvel axe commercial en introduisant la représentation de photos dites de "stock" dans la distribution existante et en approchant un nouveau marché qui est celui de la communication et de la publicité.

Les photos de "stock" sont principalement destinées aux usages commerciaux purs et publicitaires, en comparaison avec le marché de la presse ou éditorial.

Cela correspond aussi à une demande du monde publicitaire de l'époque faire des campagnes en utilisant des personnalités connues. Ce sera aussi un tournant dans la communication produit.

Ces usages "commerciaux" vont ouvrir des perspectives énormes pour Van Parys Media, mais cela entraînera aussi une modification de la structure de vente.

En effet, pour les usages de types éditoriaux, il n'y a pas besoin d'obtenir les droits à l'image des personnes représentées. Par contre, dans les usages commerciaux il est nécessaire, sous peine de très lourdes sanctions financières, d'obtenir les autorisations de la personne photographiée (bien souvent au travers de son agent) et du photographe (en direct ou au travers de son agent, en fonction de sa réputation). Ceci demande une bonne organisation administrative et un suivi très détaillé des dossiers.

C'est aussi Tom Gastmans qui assurera la transition de la photo analogique vers la photo digitale, une période inconfortable ou ces deux formats concurrents vont coexister pendant un certain nombre d'années. Chaque format trouvant un nombre équivalent de détracteurs, les débuts de la photo digitale furent périlleux et demandèrent de très gros investissements.

Nouveaux appareils photo, ordinateurs capables de gérer des photos, applications spécifiques, modes d'envois des fichiers digitaux et ces investissements ne sont pas uniquement financiers, il y a aussi l'apprentissage des nouvelles techniques de prise de vue, la manipulation des différentes applications bref, un très gros investissement humain aussi.

Petite anecdote à propos de l'évolution de la prise de vue.

Avec un appareil analogique, le cadre du viseur couvre l'entièreté de ce qui est photographié, mais ce n'était pas pareil avec les appareils digitaux de l'époque. Les capteurs ne se saisissent que d'une partie du cadre à l'époque beaucoup de photographes, habitué aux anciens appareils ont raté de nombreuses photos, ou pire, un reportage entier.

La photo va devenir petit à petit un produit moins artistique, moins artisanal ! De grands groupes apparaissent sur le marché (Getty Images, Corbis, Jupiter media, ...) et rachètent les petites structures à tour de bras. C'est le début de la globalisation du "one stop selling point".

Mais au bout du compte et de nombreuses années plus tard, cela ne s'avère pas aussi efficace que cela.

La photo reste avant tout un acte de création, heureusement d'ailleurs.

La distribution des photos va aussi se transformer pas uniquement au niveau de l'envoi de fichiers digitaux, mais aussi les accords de distributions avec l'introduction de la distribution non exclusive (plusieurs distributeurs peuvent présenter le même contenu sur le marché). Cette particularité vient de l'apparition sur le marché d'un nouveau type de produit : les photos dites "libres de droits". Appellation trompeuse, car le client/utilisateur paye bien un droit d'auteur, mais celui-ci est unique, peu importe le nombre d'utilisations ou le support de l'usage. De ce fait, ce type de photo a engendré une vente de type unique et permet la multiplication des distributeurs au sein d'une même zone géographique.

Par analogie et appétit du gain, certains acteurs du métier vont appliquer ce type de distribution à des photos en droits gérés. Une erreur qui aura coûté de l'argent au photographe et une colossale perte de temps au niveau de l'administration des ventes. En effet, les clients ont vite compris que pour ne pas payer un usage, il suffisait de dire que la photo provenait du distributeur concurrent.

Actuellement, la majorité des collections en droits gérés ont retrouvé une distribution exclusive. Et c'est ce que nous avons toujours fait pour la collection GermaineImage.

Depuis 2017, Nico Gastmans a intégré la structure avec différentes missions dont celles liées à la digitalisation des archives et leurs commercialisations en Belgique et à l'étranger.

Depuis, nous nous efforçons de faire vivre cette archive photographique avec les moyens digitaux actuels, mais les difficultés structurelles liées au marché de la photo ainsi que les investissements importants qu'exige la digitalisation d'une telle archive nous obligent à proposer d'autres services complémentaires.

Ainsi, parmi ces services, bien évidemment tous liés à la photo, nous proposons :

- des solutions IT (site internet, back office, ...) liées à la commercialisation d'une base de données photo) ;
- le marketing et le développement d'un réseau de distribution ;
- la gestion de contenu photo (édition – digitalisation – retouches – stockage ...) ;
- la gestion de thésaurus ;

- et surtout la vérification de la conformité des licences pour les usages en ligne.

Nous venons de lancer RightsControl en France et bientôt la Belgique suivra.

Pour réaliser cette numérisation, nous avons mis sur pied un protocole relativement précis :
- editing des documents ou le choix dans le reportage des images qui seront digitalisées
- digitalisation et création d'un fichier au format TIF
- indexation des images selon les protocoles IPTC et XMP
- retouches iconographiques et colorimétriques
- mise au format commercial, à savoir un JPG de 5000 pixels dans sa plus grande longueur
- organisation des copies de sauvegardes (sur disques et serveurs). 3 copies sont organisées.

Par rapport aux documents analogiques, le format digital présente un énorme avantage. Celui de pouvoir intégrer les métadonnées relatives à la photo (date, lieu, légende, mots-clés...) dans la photo elle-même et sous le standard IPTC – International Press and Telecommunication Council. Avant, le classement des documents était complexe :
- les photos N/B ou couleur
- les dias positives
- les clichés en verre
- les sheet-films

Cela demandait la mise en place d'un énorme fichier papier afin de retrouver ce qu'il y avait.

La photo digitale a résolu cela en un tour de main toutes les photos sont indexées et les moteurs de recherches internes des bases de données retrouvent les documents très facilement et rapidement.

Voici donc brièvement et au travers de la vie de 3 générations de personnes comment s'organise la vie dans une agence de presse. Il y aurait encore d'autres éléments à mettre en lumière, celui de la conservation des documents originaux dans des espaces adéquats pour chaque type de format de support, la conservation des documents digitaux et l'apparition très crainte du phénomène de "bit rot"...

Nico Gastmans
GermaineImage SA
Place des Acacias 8
1040 Etterbeek
media@ipsisworks.com

mai 2020

Notes

1. NDLR : à cette époque c'est la CCE (Commission des Communautés Européennes)

DE L'ANALOGUE AU NUMÉRIQUE, GESTION DES FONDS DOCUMENTAIRE DE L'IRPA

Julie MAURO
Documentaliste
Erik BUELINCKX
Responsable infothèque (photothèque, bibliothèque, dossiers d'intervention, numérisation)

Het artikel is opgesteld naar aanleiding van een conferentie gegeven door de auteurs in het kader van een Doc'Moment "Het fotografisch archief" op 20 november 2019 te Brussel.

Article rédigé suite à la conférence donnée par les auteurs dans le cadre du Doc'Moment "L'archive photographique", le 20 novembre 2019 à Bruxelles.

■ Riche d'un peu plus d'1 million de négatifs, le fonds de la photothèque de l'Institut Royal du Patrimoine Artistique (IRPA) est numérisé, identifié et conservé en interne par la cellule DIGIT créé en 2013 dans le cadre du projet DIGIT lancé par Belspo. Il est accessible via le site *BALaT (Belgian Art Links and Tools)*[1]. Le projet DIGIT a pour objectif également la numérisation du fonds des dossiers d'intervention de l'IRPA. Depuis 2018, l'IRPA développe en partenariat avec d'autres institutions nationales et internationales un projet de "research data management": *HESCIDA* (HEritage SCience Data Archive) pour rassembler et partager la multitude de données scientifiques hétérogènes créées par ses chercheurs.

■ De collectie van iets meer dan een miljoen negatieven van het fotoarchief van het Koninklijk Instituut voor het Kunstpatrimonium (KIK), wordt gedigitaliseerd, geïdentificeerd en intern opgeslagen door de Cel Digitalisering, die in 2013 is opgericht in het kader van het DIGIT-project van Belspo. De collectie is toegankelijk via de site BALaT (Belgian Art Links and Tools)[1]. Het DIGIT-project heeft ook tot doel de collectie interventiedossiers van het KIK te digitaliseren. Sinds 2018 werkt het KIK samen met andere nationale en internationale instellingen aan een 'research data management'-project: HESCIDA (HEritage SCience Data Archive), om de grote hoeveelheid door zijn onderzoekers voortgebrachte heterogene wetenschappelijke gegevens te verzamelen en te delen.

Établissement scientifique relevant des compétences du ministère de la Politique scientifique (BELSPO), l'Institut Royal du Patrimoine Artistique (IRPA) se consacre à la documentation, l'étude, la conservation et la valorisation des biens artistiques et culturels du pays.

L'une de ses missions principales est l'établissement d'un inventaire photographique du patrimoine belge et de le rendre accessible au grand public. Le fonds de la photothèque de l'IRPA a évolué avec les nouvelles technologies pour passer progressivement du négatif au numérique. C'est un instrument précieux et extraordinaire pour l'étude et la connaissance du patrimoine belge.

Le fonds des dossiers d'intervention, archives scientifiques de l'histoire de l'étude et de la conservation-restauration du patrimoine belge est une autre des richesses documentaires de l'IRPA.

Ces fonds s'inscrivent aujourd'hui dans une gestion globale de l'information digitale au sein de l'institution avec la mise en place depuis 2016 d'un projet de "research data management" en collaboration avec plusieurs institutions nationales et internationales. Le projet HESCIDA (HEritage SCience Data Archive)[2] a pour but de rassembler et partager la multitude de données scientifiques hétérogènes créées par les chercheurs.

Genèse de l'IRPA et de sa photothèque

L'histoire de l'IRPA et de sa photothèque commence en 1900 au sein des Musées Royaux d'Art et d'Histoire (MRAH) par la création d'un atelier de photographie. En 1920, le service de la documentation belge, l'ancêtre de la photothèque, est né. En 1934 Jean Capart, alors conservateur en chef des MRAH, nomme un jeune chimiste, Paul Coremans, chef du service de la Documentation belge avec pour mission de créer un Laboratoire de Recherches physico-chimiques. Les innombrables dangers que les œuvres et les monuments historiques subissent pendant la deuxième guerre mondiale, nourriront sa réflexion et seront à la base de son concept d'une approche interdisciplinaire pour la sauvegarde du patrimoine. Il fut d'ailleurs membre des *Monuments Men*, unité d'élite chargé de la sauvegarde des œuvres d'art pendant la guerre.[3]

Après la fin de la guerre, un arrêté du Régent daté du 24 juin 1948, fonde les Archives Centrales iconographiques d'Art national et le Laboratoire

central des Musées de Belgique (ACL). Cette nouvelle institution, indépendante des MRAH, se consacre officiellement à l'inventaire, l'étude scientifique et la conservation des œuvres d'art, au bénéfice de tout le pays.

Fig. 1: Bureau des ACL au sein des Musées royaux d'art et d'histoire. © KIK-IRPA, Bruxelles, 1947, B124866

L'arrêté royal du 17 août 1957 fait des ACL l'une des dix institutions scientifiques fédérales et devient l'Institut royal du Patrimoine artistique (IRPA). Le projet de travail pluridisciplinaire, défendu dès le début par Paul Coremans est enfin reconnu. En 1962, l'IRPA s'installe dans un nouveau bâtiment. Il est le premier au monde spécialement conçu pour rassembler toutes les disciplines œuvrant à la conservation du patrimoine artistique. Aujourd'hui, certaines parties de l'édifice sont classées.

Il est composé de trois départements : conservation-restauration, laboratoires et documentation et d'une cellule de valorisation-communication.

La documentation au sein de l'Institut

Les fonds documentaires de l'IRPA sont gérés, préservés et valorisés au sein du département documentation. La tâche principale de ce département est la constitution d'un inventaire photographique du patrimoine culturel belge. Les prises de vues sont réalisées par les ateliers photographiques, en collaboration étroite avec les historiens de l'art du département, les collègues d'autres départements ou d'autres institutions.

Le département comprend la cellule recherches en histoire de l'art et inventaire, le centre d'étude des Primitifs Flamands, la cellule imagerie (ateliers photographiques et l'imagerie scientifique) et l'infothèque (bibliothèque, dossiers d'intervention, photothèque et la cellule DIGIT).

La photothèque documentaire d'art, un outil au service de la recherche.

La collection photographique de l'IRPA est un fonds de photographie documentaire d'art qui compte aujourd'hui un peu plus d'1 million de clichés. Ce fonds s'est constitué au fil du temps grâce aux campagnes photographiques menées à travers le pays depuis 1900 et suite à divers achats et dons de fonds photographiques privés. Aujourd'hui, les photographes de l'Institut continuent les missions d'inventaire à travers le pays et à l'étrangers pour enrichir d'avantage cette collection qui couvre tous les aspects du patrimoine belge : beaux-arts, architecture, archéologie, paysages, etc.

La constitution du fonds de la photothèque s'est réalisée principalement en de trois grandes phases :[4]

- 1914-1918 : durant la Première Guerre mondiale, une équipe allemande d'environ trente historiens de l'art, photographes et architectes ont sillonné tout le pays pour photographier les monuments belges les plus importants. Ils ont réalisé plus de 10 000 prises de vue. Dix ans après la fin de la guerre, les négatifs originaux – tous sur plaques de verre – ont pu être achetés par l'État belge. Depuis lors, les "clichés allemands" sont gérés, conservés et valorisés par l'IRPA à Bruxelles.
- 1940-1945 : durant la deuxième guerre mondiale, les Musées royaux d'Art et d'Histoire entament d'urgence un inventaire photographique du patrimoine artistique. Ainsi, entre 1941 et 1945, plus de 165 000 clichés sont réalisés. Ces photos seront particulièrement utiles après la guerre pour reconstituer les œuvres endommagées. Elles resteront, dans certains cas, les seuls témoins d'œuvres d'art anéanties
- 1967-1984 : la disparition de plus en plus fréquente de biens mobiliers dans les églises inquiètes les ministres de la culture Pierre Wigny et Renaat van Elslande. Ils chargent l'IRPA de réaliser un répertoire photographique du mobilier des sanctuaires de Belgique, afin d'inventorier les œuvres présentant un intérêt pour le patrimoine. 250 000 prises de vues sont réalisées à travers le pays.

Les prises de vues en noir et blanc ont été réalisées jusque fin des années 80, ensuite la couleur est apparue pour faire place à la photo numérique début 2000.

Outre les photos d'inventaire, les ateliers photographiques ont également pour mission de documenter le travail qu'effectuent les restaurateurs et les chercheurs sur les œuvres étudiées et/ou traitées à l'IRPA ou in situ. L'imagerie scientifique vient également compléter ce travail documentaire des œuvres

par la réfléctographie infrarouge et la radiographie (RX). Ces photos techniques qui sont appelées en interne "photos Labos" sont ajoutées aux dossiers d'intervention des œuvres.

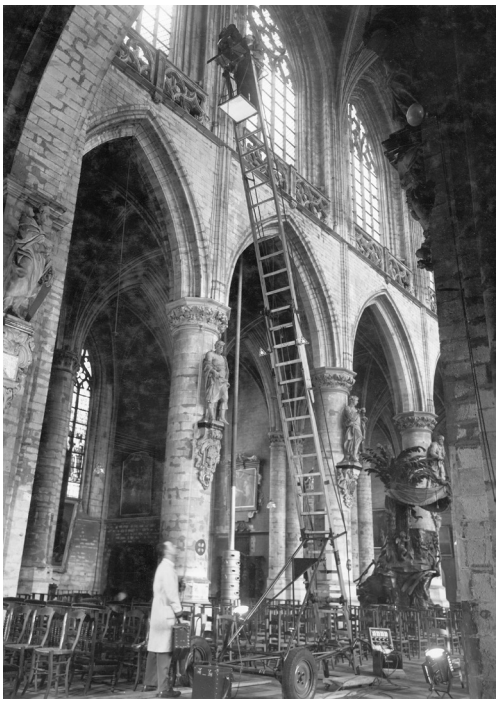

Fig. 2 : Mission d'inventaire photographique à Bruxelles, Eglise Notre-Dame de la Chapelle.
© KIK-IRPA, Bruxelles, 1954, B148813.

La numérisation, enjeu essentielle à la préservation des collections[5]

Jusqu'en 1989, la consultation des images de la photothèque se faisaient à travers les tirages positifs collés sur carton et muni d'une simple description. La numérisation de la collection photographique a débuté dans les années 90. A l'époque l'IRPA a choisi de faire numériser les tirages positifs par une firme externe, le Village n° 1 Reine Fabiola. L'objectif de cette première campagne de numérisation était de mettre rapidement à disposition du public une reproduction suffisante pour la consultation sur la base de données. Ce qui permet aujourd'hui d'accéder à près de 790.000 images téléchargeables gratuitement en ligne.

En 2004, le plan de numérisation DIGIT pour les collections des dix établissements scientifiques fédéraux et la cinémathèque royale de Belgique est lancé par BELSPO. Ce défi ambitieux qui a pour objectif une meilleure conservation et valorisation des collections fédérales, a amené à la création en 2013 de la cellule DIGIT au sein de l'Institut. Vu l'enjeu essentiel de la numérisation pour les Établissements scientifiques fédéraux (ESF), le gouvernement fédéral a approuvé fin 2018 une nouvelle prolongation du projet DIGIT qui s'achèvera en 2025.

La numérisation a pour but de fournir des reproductions en haute résolution de la collection afin d'apporter aux chercheurs un outil précieux pour leurs recherches.

Vu l'importance de la collection, chaque année un plan d'action est établit pour déterminer quelle partie de la collection va être numérisé. Ce plan se base essentiellement sur 3 critères de sélections : le risque de dégradation, les projets de recherche (valorisation du fonds) et la demande du public.

La préservation de la collection analogue

La source du projet de numérisation est le négatif photographique, la prise de vue originelle réalisée par le photographe. Une grande attention lui est donc apportée pour sa conservation à long terme, en effet l'instabilité des supports photographiques traditionnels a déjà conduit à la perte de collections entières partout dans le monde et malheureusement aussi à l'IRPA.

La cellule DIGIT en collaboration avec la cellule de conservation-préventive est responsable de la gestion et de la préservation de la collection. L'Institut conserve environs 900.000 photos analogues sur différents supports photographiques comme les plaques de verre, les films Ektachrome, les dias, etc. Le reste est conservé sur des supports numériques.

Les négatifs sont conservés dans un dépôt au sein de l'Institut, où une température d'environ 13°C et un taux d'humidité relative d'environ 33 % sont maintenus. Ces indices sont contrôlés quotidiennement. Ce contrôle du climat est important notamment pour ralentir la dégradation chimique, physique ou biologique des photographies analogiques. Des mesures de protections strictes sont appliqués lors de la manipulation des supports originaux afin de

Fig. 3 : Stéphane Bazzo, photographe au travail dans l'atelier de conservation-restauration des sculptures en bois. Il photographie le retable Saint-Georges (MRAH).
© KIK-IRPA, Bruxelles, 2019, X136460.

prévenir les dégâts mécaniques (griffes, empreintes digitales, cassures, fissures, lacunes). Ces dégâts ont déjà été constatés par le passé suite à de mauvaises manipulations.

Fig. 4 : Dépôt de la collection des négatifs.
© KIK-IRPA, Bruxelles, 2014.

La numérisation : more than pressing a button

A l'IRPA, le processus de numérisation n'est pas simplement de placer un négatif sous un appareil photo et de presser le bouton. Il comprend également les activités de pré et post traitement comme le constat d'état du support original, la conservation des supports (comme vu plus haut), la numérisation en tant que tel, le post-traitement et l'archivage.

Conservation du négatif

Phase importante et décisive qui va influencer tant la qualité de la reproduction numérique que la sauvegarde du support analogique. Lors de cette phase, un constat d'état minutieux du négatif est établi, c'est-à-dire que les principales dégradations visibles sont relevées dans un formulaire type. Ce constat d'état des négatifs va permettre de dresser un bilan de santé de la collection, mais également de repérer plus facilement les nouveaux dégâts. Le négatif subit ensuite un nettoyage et est débarrassé de tous les éléments étrangers qui pourraient avoir une influence néfaste sur sa préservation (emballage acide, chemise, ruban adhésif, ...). Enfin, le négatif est reconditionné dans une pochette en papier antiacide à 2 rabats découpée sur mesure et ensuite dans une boîte en carton neutre adapté à son format pour le protéger des influences néfastes extérieures.[6]

Numérisation du négatif

La méthode utilisée est la rephotographie, c'est à dire que l'on photographie directement le négatif. Cette méthode a été choisie car elle donne les meilleurs résultats en terme de qualité de reproduction d'image ; afin de sauvegarder un maximum d'informations. Elle est également plus sûre et rapide pour les négatifs que les scanners. L'objet est numérisé en une seule prise et permet d'être manipulé avec plus de précautions. Une fois numérisé le négatif original étant fragile, il ne devra plus être consulté. Lors de la prise de vue un fichier RAW[7] est réalisé à l'aide d'un appareil photo HD (Nikon D810 ou Canon EOS EDsR) déclenché depuis un ordinateur à l'aide du programme Lightroom utilisé en mode "capture connecté". Le fichier RAW, renferme l'ensemble des informations provenant du capteur de l'appareil photo. Si la prise de vue était réalisée directement en TIFF ou JPEG, l'appareil appliquerait automatiquement une série de réglages non contrôlés sur les contrastes ou la netteté de l'image. Il permet donc de conserver tout le potentiel de l'image. Avant chaque session de numérisation, une charte de référence graphique est photographiée afin de définir la balance des blancs et d'enregistrer les conditions d'éclairages de la numérisation. Le négatif est numérisé dans son ensemble mais aussi dans l'état dans lequel il se trouve. Le fichier RAW sera développé grâce au programme Lightroom, sorte de chambre noire numérique à l'instar de la chambre noire du photographe pour l'analogique. Le résultat est sauvegardé au format TIFF[8] sans compression dans l'espace colorimétrique Adobe RGB[9] avec une profondeur de 48 bits (16 bits par canal de couleur).

Contrôle qualité et post-traitement

Après son développement, l'image est soumise à un contrôle qualité effectué par un autre opérateur de numérisation de la cellule DIGIT, afin d'appréhender l'image avec une autre œil. Les différents aspects de l'image comme le cadrage, les contrastes, la netteté de la photographie sont passés au crible. Ensuite vient le post-traitement qui a pour objectif de réaliser les opérations de retouches et d'ajustement éventuel de l'image (cadrage, rotation éventuelle de l'image, ajustement des contrastes, ...). A part pour certains négatifs couleurs, où l'on essaye de restaurer les couleurs d'origines, peu de restaurations numériques sont effectuées sur les images de la collection. Il a été décidé à l'Institut de montrer la collection comme elle est, car cela fait partie de son histoire matérielle. Si besoin, une restauration numérique peut être réalisée exceptionnellement pour une publication si les dégâts empêchent une lecture correcte de l'image. La restauration sera donc signalée.

Le fichier photographique, au format TIFF avec une résolution de 350 DPI, est prêt à être versé dans les archives numériques de l'IRPA en vue d'une conservation long terme. A partir de ces archives, des fichiers au format JPEG seront créés automatiquement pour la consultation en ligne.

Identification des images : encodage des métadonnées

Les images et toutes ses métadonnées liées sont cataloguées dans une base de données. L'application utilisée est *Adlib Xplus*[10]. La base de données permet de rassembler l'ensemble des informations sur le sujet immortalisé (œuvre d'art, événement, paysage, etc.) via une fiche "objet" reliée à celles de toutes les images de l'objet disponibles au sein de l'IRPA.

Le portail BALaT, outil de valorisation

La finalité de la numérisation est la valorisation de la collection auprès du grand public et des chercheurs via le portail web de l'IRPA, *BALaT* (Belgian Art Links and Tools)[11]. Cet outil, fruit d'une étroite collaboration entre le service informatique et le département documentation, permet d'accéder plus intuitivement au contenu de la base de données *ADLIB* et de télécharger gratuitement les photographies au format JPEG. Les utilisateurs qui souhaitent recevoir une image en haute définition peut la commander auprès de la cellule DIGIT[12]. Tout un chacun peut donc découvrir environs 750.000 photographies. Toutes reproductions issues de la collection doivent mentionner le copyright de l'IRPA.

Les dossiers d'intervention : archives scientifiques de l'IRPA

Collection riche d'environs 20.000 dossiers et qui continue de croître de jour en jour. Le fonds des dossiers d'intervention, mémoire des activités scientifiques de l'IRPA, s'est constitué depuis la création du laboratoire de recherches physico-chimiques dans les années 30. Un dossier d'intervention est ouvert et reprend l'ensemble de la documentation produite par les restaurateurs, les historiens d'art et les chimistes durant l'étude, le traitement et les analyses de laboratoires effectués sur une œuvre d'art. Il comprend également les photos techniques réalisées par les photographes et l'imagerie scientifique (RX, IR, photos prises pendant le traitement, ...). Les chercheurs ajoutent souvent dans le dossier leur propre documentation photographique.

C'est un fonds qui au fil du temps a pris une valeur historique, car il contient des documents uniques sur la conservation et la restauration des œuvres du patrimoine belge. Il est en effet important avant de restaurer une œuvre de connaître son histoire matérielle, en d'autres termes, les différentes interventions qu'elle a subies au cours du temps. Sa gestion est effectuée par le service "dossiers-archives" au sein de l'infothèque.

C'est un fonds accessible à tous mais il est consulté principalement par les membres du personnel et par un public de professionnels comme des chercheurs, restaurateurs, des étudiants en histoire de l'art et en conservation-restauration, des architectes, etc.[13]

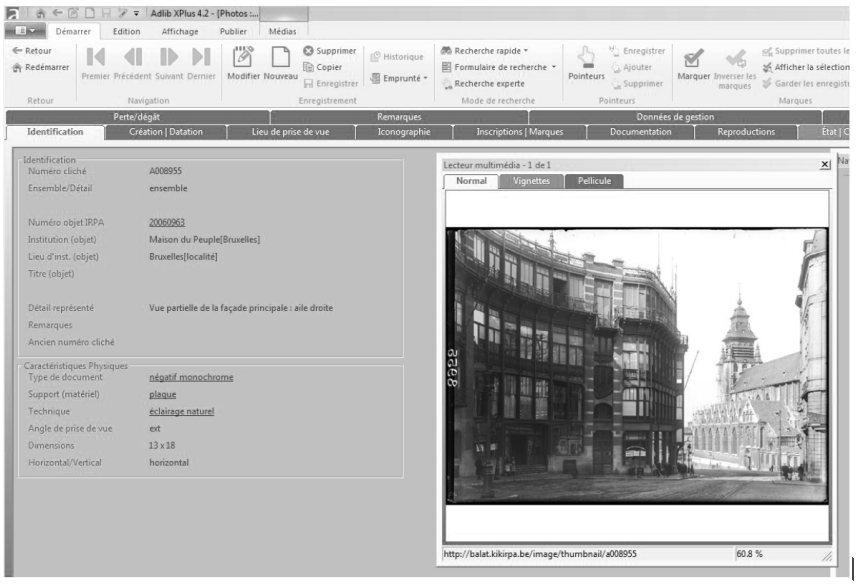

Fig. 6 : Fiche « photo » du logiciel Adlib.
© KIK-IRPA, Bruxelles, capture d'écran 2020.

Vers un archivage numérique des dossiers

Les dossiers sont dans la liste des priorités des documents à numériser pour l'Institut. Étant régulièrement consulté, leur numérisation est essentielle pour leur préservation à long terme. En effet avec le temps et les nombreuses manipulations, le risque d'endommager les documents ou de les perdre est grand. Vu l'ampleur de la collection, une sélection des dossiers à dématérialiser est indispensable. Cette sélection est réalisée d'une part en collaboration avec les différents ateliers et cellules avec qui une liste des dossiers prioritaires ayant une valeur historique et un intérêt majeur pour la recherche est établie, d'autre part en fonction des projets de recherche et des demandes interne ou externe des chercheurs.

La numérisation est effectuée en collaboration avec la cellule DIGIT avec qui une procédure spécifique a été mise en place. En effet, à l'instar des négatifs, leur numérisation est "more than pressing a button". En pré-traitement, un inventaire complet du contenu du dossier est réalisé dans un tableur Excel spécifique. La numérisation des documents est réalisée par la technique de la photographie suivant le même procédé que la rephotographie appliqué au négatif. Les fichiers sont sauvegardés en fichier TIFF et retravaillés si besoin en post-traitement par un opérateur de numérisation.

Ils sont archivés numériquement en PDF/A, en résolution 300 DPI en vue d'une préservation long terme sur des systèmes de stockages internes et externes.

Depuis l'informatisation de l'Institut, une partie des documents de la collection ont été produit dès le début sous format numérique par les chercheurs. Le service "dossiers-archives" essayent de capturer ces documents digital born pour que tout comme les dossiers numérisés, ils soient conservés dans un format pérenne et à long terme. La capture et la sauvegarde des documents à leur source numérique permettent de conserver une meilleure qualité de l'information qui peut se perdre lors de la numérisation.

Fig. 7 : Le fonds des dossiers d'intervention.
© KIK-IRPA, Bruxelles, 2015.

Encodage des métadonnées

L'encodage des dossiers est également effectué dans l'application *Adlib Xplus*, mais seul le numéro de dossier est mentionné dans la base de données en ligne *BALaT*. Les métadonnées et les dossiers numérisés sont accessibles pour le moment seulement au personnel interne via une plateforme *KIK-IRPA Tools* disponible sur le site Intranet de l'Institut.

Le dossier est le point d'accès à l'ensemble des informations disponible à l'IRPA sur une œuvre. Dans un esprit d'open access, le but est donc de les rendre accessible au plus grand nombre, ce qui est en train d'être développé par le Projet HESCIDA.

Projet HESCIDA & Research Data Management[14]

Le projet HESCIDA (HEritage SCience Data Archive) s'inscrit dans une infrastructure de recherche distribuée dans le domaine de la science du patrimoine, E-RIHS (European Research Infrastructure for Heritage Science)[15]. Une infrastructure de recherche distribuée est une organisation qui permet à la communauté scientifique d'utiliser des installations, des ressources et des services spécifiques qui sont géographiquement dispersés. E-RIHS donne accès aux installations, ressources et services dans le domaine de la science du patrimoine à travers l'Europe par le biais de quatre plateformes intégrées :

- E-RIHS ARCHLAB (archives) : Accès à des connaissances spécialisées et à des informations scientifiques structurées, y compris des images techniques, des données analytiques et des documents de conservation.
- E-RIHS DIGILAB (installations virtuelles) : Accès virtuel aux données scientifiques concernant le patrimoine matériel.
- E-RIHS FIXLAB (installations fixes) : Accès à des installations à grande et moyenne échelle (accélérateurs de particules et synchrotrons, sources de neutrons ; instruments d'analyse non transportables) offrant une expertise unique aux utilisateurs dans le domaine du patrimoine.
- E-RIHS MOLAB (installations mobiles) : Accès à une gamme d'instruments d'analyse mobiles avancés pour des mesures non invasives sur les œuvres d'art, sur des sites archéologiques et des monuments historiques.

Grâce au financement de BELSPO pour soutenir les institutions scientifiques fédérales qui jouent un rôle actif dans E-RIHS, le projet HESCIDA a été lancé en 2019 et servira de base au pilier E-RIHS DIGILAB. Il a pour objectif la gestion et l'accès aux données scientifiques du patrimoine.

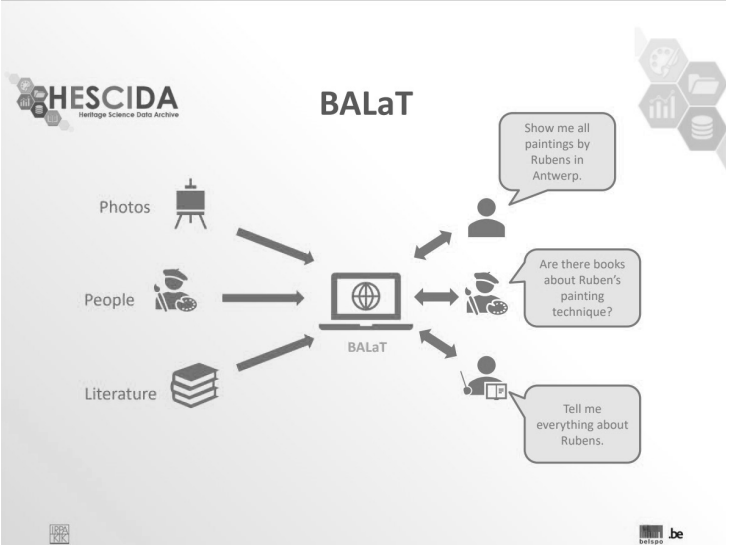

Fig. 8 : Slide issue de la présentation du Doc'Moment de novembre 2019 représentant les objectifs de BALaT+.

DIGILAB s'appuie sur un réseau de dépôts locaux fédérés où les chercheurs, les gestionnaires et autres professionnels du patrimoine déposent les résultats numériques de leurs travaux. En principe, DIGILAB ne conserve pas ces données en interne : il donne plutôt accès aux dépôts originaux où les données sont stockées. DIGILAB s'inspire du principe des données FAIR (Findable, Accessible, Interoperable, Reusable)

- Findable: permet de trouver les données grâce à un système de recherche avancée fonctionnant sur un registre contenant des métadonnées qui décrivent chaque ensemble de données individuel.
- Accessible : permet d'accéder aux données grâce à un système d'identité fédéré, tandis que les autorisations d'accès restent locales à chaque dépôt.
- Interoperable : garantit l'interopérabilité des données en adoptant un modèle de données standard.
- Reusable : favorise la réutilisation des données brutes et des métadonnées en mettant des services à la disposition des utilisateurs, pour traiter les données selon leurs propres questions de recherche ou leurs besoins d'utilisation.

BALaT sera mis à niveau en *BALaT+* et visera à devenir le premier dépôt local. La collection des négatifs, le catalogue de la bibliothèque, les PDF d'articles du Bulletin de l'IRPA et d'autres publications, une liste exhaustive des personnes et des institutions qui font autorité, plusieurs sites web thématiques spécialisés, les dossiers d'intervention et les données brutes des analyses seront tous reliés entre eux. L'utilisateur accédera donc plus facilement à aux informations en open access sur le patrimoine culturel belge.

Grâce à DIGILAB, ces données seront portées à un niveau supérieur, permettant aux chercheurs d'interroger et de comparer des ensembles de données multidisciplinaires provenant des plus prestigieuses institutions scientifiques du patrimoine européen, en utilisant une seule interface.

En ce qui concerne la conservation et la préservation à long terme des données, on se conformera aux exigences du dépôt de données CoreTrustSeal (Core Trustworthy Data Repositories), afin d'obtenir la certification de qualité pour nos dépôts.

Dans le projet HESCIDA, l'accent sera désormais mis sur deux éléments essentiels : le standard IIIF[16](International Image Interoperability Framework) et ElasticSearch (outil d'indexation open source)[17]. L'idée générale est de permettre aux chercheurs de faciliter la recherche sur de multiples aspects d'un objet du patrimoine culturel. Il pourra ainsi trouver plus facilement non seulement une des images accompagnées des métadonnées descriptives sur l'œuvre, mais aussi des données analytiques et des liens avec d'autres œuvres d'art.

L'ensemble de la collection de la photothèque a déjà été converti en TIFF pyramidal afin de pouvoir être proposée via la couche d'image du standard IIIF. Les autres images scientifiques et liées au travail des chercheurs seront également reprises, comme les photographies personnelles des restaurateurs, les photos indiquant les lieux de prélèvement, les photos microscopiques des échantillons exposés, etc.

ElasticSearch est également utilisé pour stocker des annotations concernant les indications de dégâts, les traitements de conservation, les analyses de

laboratoire, les lieux de prélèvement et l'iconographie. Un autre avantage d'ElasticSearch est la capacité de recherche à l'échelle du système pour trouver des informations aussi bien dans le CMS (Content Management System) que dans les PDF stockés sur le serveur des publications officielles de l'IRPA, comme les rapports d'intervention. Le système sera conçu de telle sorte que le principe " aussi ouvert que possible et aussi fermé que nécessaire " soit maintenu. Bien qu'une ouverture totale soit prônée, des considérations liées à la protection de la vie privée imposer certaines restrictions d'accès.

Conclusion

Avec environ 150 personnes, dont plus de 65 chercheurs, l'IRPA n'est pas une si petite institution. Etant donné que tant de tâches différentes sont effectuées dans les locaux de l'Institut et en mission, nous comptons principalement sur l'ingéniosité et la bonne volonté de nos collègues pour faire face aux questions de suivi des évolutions dans l'utilisation des données numériques (patrimoine) de la recherche scientifique. L'initiative ponctuelle de BELSPO nous a donné un coup de pouce pour créer un "research data repository" local qui fera partie à terme du domaine plus large de la science du patrimoine à l'échelle européenne (DIGILAB dans E-RIHS) et mondiale. En même temps, nous avons profité de cette occasion pour transformer le *BALaT* existant en *BALaT+*, pour qu'il soit prêt à être utilisé par les spécialistes et le grand public. Et cela en utilisant les dernières technologies disponibles pour une recherche approfondie et une visualisation optimale des données scientifiques patrimoniales.

Merci à nos collègues de l'infothèque, de la cellule DIGIT et du projet HESCIDA.

Julie Mauro
Erik Buelinckx
Institut royal du patrimoine artistique
Département Documentation
1 parc du Cinquantenaire – 1 Jubelpark
1000 Bruxelles
julie.mauro@kikirpa.be
erik.buelinckx@kikirpa.be
www.kikirpa.be

Mai 2020

Notes

1. Voir <http://balat.kikirpa.be>
2. Voir <http://hescida.kikirpa.be/>
3. Voir Publication : Dominique Deneffe, Dominique Vanwijnsberghe. *A Man of Vision. Paul Coremans and the Preservation of Cultural Heritage Worldwide*. Proceedings of the International Symposium Paul Coremans Held in Brussels, 15-17 June 2015. Scientia Artis 15. Institut royal du Patrimoine artistique, 2019. ISBN 978-2-930054-34-6
4. Voir article : Claes, Marie-Christine. Héritage bénéfique des guerres mondiales en Belgique. Le concept er les collections de l'IRPA. *Bruxelles patrimoines*. 2014, vol. 11-12, p. 60-73.
5. Voir article : De Groof, Stijn, De Vocht, Tim, De Zutter, Élodie, Raes, Sander , Reyniers, Jeroen. La numérisation de la collection photographique de l'IRPA: le cas des Clichés allemands. *In Bulletin de l'Institut royal du Patrimoine artistique / Bulletin van het Koninklijk Instituut voor het Kunstpatrimonium*, 35, 2019, p. 157-175.
6. Tous les matériaux utilisés pour la conservation répondent aux exigences du Photographic Activity Test (PAT) (norme ISO 18916 :2007).
7. RAW ("brut" en anglais)= Un fichier Raw contient les données brutes du capteur et les paramètres nécessaires à la transformation en fichier image visible sur écran.(Wikipédia, 13/05/2020)
8. TIFF = Tag(ged) Image File Format : Format non compressé recommandé au niveau international pour sa stabilité et pour sa possibilité de s'ouvrir sur différents logiciels.
9. RGB= Red-green-blue : ces trois "couleurs primaires lumière" forment le blanc par synthèse additive.
10. Application développée par le groupe international Axiell.
11. Voir <http://balat.kikirpa.be>
12. Les commandes peuvent être envoyées à l'adresse suivante : shop@kikirpa.be.
13. Les dossiers d'intervention sont consultables sur rendez-vous :dossiers@kikirpa.be

14. Voir article : Stephanie Buyle ; Wim Fremout ; Edwin De Roock ; Erik Buelinckx en Emmanuel di Pretoro. Een nieuwe toekomst voor erfgoedwetenschappelijke data: een overzicht van research data management-initiatieven in het Koninklijk Instituut voor het Kunstpatrimonium. *META*, 2020, n°4 (consulté le 13/05/2020). <https://www.vvbad.be/meta/meta-nummer-20204/een-nieuwe-toekomst-voor-erfgoedwetenschappelijke-data-een-overzicht-van>
15. Voir <http://www.e-rihs.eu/>
16. Voir <https://iiif.io/>
17. Voir <https://www.elastic.co/fr/>

WIKIMEDIA COMMONS
DES MILLIONS D'IMAGES EN LIBERTÉ

Guy DELSAUT
Professionnel de l'information

■ Plus de 60 millions de médias sont consultables, modifiables et réutilisables dans *Wikimedia Commons*, l'un des projets de la Wikimedia Foundation. Les images représentent la part la plus importante de ces médias : plus de 57 millions de photos, de schémas, de peintures, de cartes... Chaque image est décrite et catégorisée. Les contributeurs peuvent téléverser de nouvelles photos mais aussi enrichir les descriptifs ou la classification de ces œuvres. Dans cet article, nous survolerons les droits relatifs à ces médias, le contenu du site, le travail collaboratif qui peut y être apporté. Nous verrons également comment rechercher une image et comment la réutiliser légalement.

■ Er kunnen ruim 60 miljoen mediabestanden worden geraadpleegd, gewijzigd en hergebruikt in Wikimedia Commons, een van de projecten van de Wikimedia Foundation. De afbeeldingen vormen het grootste deel van de bestanden: ruim 57 miljoen foto's, schema's, schilderijen, kaarten, enz. Elk beeld is beschreven en gerubriceerd. Deelnemers kunnen nieuwe foto's uploaden, maar ook beschrijvingen of rubriceringen van de werken aanvullen. In dit artikel kijken we in vogelvlucht naar de rechten die aan deze bestanden verbonden zijn, de inhoud van de site en het opensource-werk dat eraan kan worden besteed. Ook zien we hoe we een afbeelding kunnen zoeken en legaal kunnen gebruiken.

Imaginez. Vous cherchez la photo d'une loupe pour embellir une présentation sur la recherche documentaire, d'une veilleuse pour illustrer un article sur la veille ou d'une infirmière pour remercier le personnel soignant sur votre site. De plus, vous souhaitez respecter le droit d'auteur mais vous n'avez pas vos propres photos de loupe, de veilleuse ou d'infirmière. Aujourd'hui, il est inimaginable de communiquer sans illustration. Même un tweet a plus de chance d'être lu s'il comporte une image. Et si certains oublient que le droit d'auteur existe, d'autres souhaitent le respecter scrupuleusement. C'est là qu'interviennent les licences qui permettent à des créateurs (photographes amateurs ou professionnels, graphistes, dessinateurs...) d'abandonner certains droits pour que leurs œuvres puissent être diffusées sans qu'aucune demande formelle ne soit nécessaire.

Wikimedia Commons[1] rassemblent ainsi plus de 60 millions de fichiers médias[2] pouvant être réutilisés, à condition de respecter la licence qui y est attachée. Le nom de *Wikimedia Commons* ne vous est peut-être pas familier. Pour résumer, c'est une banque de médias libres créée en 2004 par la Wikimedia Foundation. Ces médias servent à illustrer les différents projets de la fondation, dont *Wikipédia*, *Wikivoyage* ou *Wikinews*, mais peuvent aussi être utilisés ailleurs, y compris pour un usage commercial.

Des médias libres de tout droit

Avant de parler du contenu, il faut parler des droits. Le respect strict du droit d'auteur et du droit à l'image influence nettement le contenu du site.

Licence des médias importés

Tous les médias importés dans *Wikimedia Commons* doivent être publiés sous une licence libre permettant au minimum la rediffusion du média original, sa transformation et la diffusion du média dérivé, et son utilisation commerciale. Ces autorisations sont données pour une durée illimitée. De ce fait, celui qui importe un média en est soit l'auteur (ou l'ayant droit), soit un tiers qui peut prouver que le média remplit ces caractéristiques.

Au moment de l'importation dans *Wikipédia*, l'auteur d'un fichier doit indiquer "Je, [nom d'utilisateur], détenteur du droit d'auteur sur ces œuvres, accorde de façon irrévocable à quiconque le droit d'utiliser ces œuvres sous la licence...". Par défaut la licence proposée actuellement est la licence Creative Commons[3] "Attribution - Partage dans les mêmes conditions 4.0 International (CC BY-SA 4.0)" mais d'autres licences sont proposées comme les licences CC BY 3.0 et 4.0, qui n'obligent pas la redistribution sous la même licence, et la licence CC0, c'est-à-dire le versement de l'œuvre dans le domaine public. Les licences interdisant l'utilisation commerciale (NC) ou la rediffusion d'œuvres dérivées (ND) ne sont donc pas proposées (voir fig. 1). Avant 2009, les projets Wikimedia utilisaient les licences GNU Free Documentation License avant de passer aux Creative Commons, ce qui explique la double licence sur certains médias.

On ne peut donc pas importer une image trouvée sur Internet ou une carte postale, même ancienne, sans s'être assuré que le créateur est bien mort depuis

Quelle est la licence indiquée ?	Licence	Acceptée ou non ?
© Tous droits réservés	Soumise au droit d'auteur	Non acceptée
Certains droits réservés	CC-BY-NC-ND	Non acceptée
Certains droits réservés	CC-BY-NC-SA	Non acceptée
Certains droits réservés	CC-BY-NC	Non acceptée
Certains droits réservés	CC-BY-ND	Non acceptée
Certains droits réservés	CC-BY	Acceptée
Certains droits réservés	CC-BY-SA	Acceptée
Aucune restriction au droit d'auteur connue	Domaine public	Acceptée

Fig. 1 : Tableau des licences autorisées.

plus de 70 ans ou que la photo a été publiée sous une licence compatible avec *Wikimedia Commons*.

Droit sur l'œuvre représentée

Si l'image représente une œuvre, il faudra également s'assurer qu'il est légal d'en diffuser une reproduction, même à des fins commerciales[4]. Il faut donc que l'œuvre soit dans le domaine public (l'auteur est mort depuis plus de 70 ans), soit qu'elle bénéficie d'une exception au droit d'auteur dans le pays concerné.

La Liberté de panorama[5] fait partie de ces exceptions mais elle diffère d'un pays à l'autre. Il sera donc interdit d'y placer une photo de la Petite sirène de Copenhague dont l'auteur est mort en 1959 puisque la loi danoise a réservé la Liberté de panorama aux seuls bâtiments. Pas de souci, par contre, pour les immeubles, statues ou fresques murales situées en Belgique depuis l'adoption de la Liberté de panorama en 2016[6].

Autre exception marquante, les œuvres produites par l'administration des États-Unis sont libres de droit. Pour cette raison, on peut retrouver des enregistrements audio de Présidents des États-Unis ou la photo des pièces et des billets américains, à quelques exceptions près. D'autres pays, comme l'Inde ou la Russie, appliquent également une politique ouverte en matière de droit d'auteur pour les documents diffusés par l'État. Cette politique permet d'inclure dans la banque d'images des photos de visites officielles, par exemple. En Europe, le Parlement européen publie des photos sous licence Creative Commons et la Banque centrale européenne autorise la reproduction des dessins des billets de banque.

La communauté wikimédienne veille à une stricte application du droit d'auteur. Ainsi, même la photographie d'un dessin d'enfant prise lors d'une manifestation a été supprimée car le contributeur ne pouvait pas prouver l'autorisation des parents pour la reproduction de l'œuvre.

Autre exemple personnel et plus complexe. Il y a quelques années, j'avais obtenu de la Monnaie royale de Belgique des photos de qualité de toutes les pièces belges. Le responsable avait signé une déclaration autorisant la diffusion de ces photos sur *Wikimedia Commons* mais un contributeur pointilleux a fait supprimer les photos. Pourquoi ? Il remettait en doute que la Monnaie royale de Belgique puisse donner cette autorisation à la place du créateur, autrement dit du graveur. Circonstance aggravante : l'une des pièces représentait Tintin. Il s'agissait donc d'une œuvre dérivée. Seules sont restées les pièces dont le graveur était mort depuis plus de 70 ans après des recherches sur chaque artiste.

Droit à l'image

Qu'en est-il du droit à l'image ? Une photo d'un chanteur sur scène, d'une actrice défilant sur le tapis rouge à Cannes ou une femme politique en plein meeting électoral peut-elle être importée dans *Wikimedia Commons* ? La loi diffère d'un pays à l'autre. Généralement, on considère qu'une photo prise dans un lieu public, qui plus est quand il s'agit

Fig. 2 : Les pièces américaines ne sont généralement pas soumises au droit d'auteur. (US Mint, domaine public)

d'une personne publique exerçant son métier, est autorisée. Si la photo est prise dans un lieu privé mais qu'on voit clairement que la personne a posé, elle sera également acceptée.

Cependant, il existe un devoir moral. La photo ne peut pas ridiculiser l'individu, le diffamer ou l'humilier. Une photo d'une personne en train de se moucher ou grimaçante, même prise dans l'espace public, n'aura pas vraiment sa place. Pas plus qu'une photo de style "paparazzi" montrant une célébrité à la plage ou occupée à faire du shopping en famille.

Il arrive que certaines personnalités demandent le retrait d'une photo qui ne leur plaît pas. Ces demandes n'aboutissent que si elles sont valablement motivées. Prenons l'exemple du comédien Stéphane Freiss. Son assistante demandait en 2019, la suppression d'une photo de 2008 parce que l'acteur ne la trouve "ni bonne ni représentative de ce qu'il est aujourd'hui". L'argument n'a pas été jugé satisfaisant, puisqu'il pose clairement pour la photo et qu'elle a été prise dans un lieu public.

Fig. 3 : Un artiste comme Raphael peut être photographié sur scène.
(Guy Delsaut, CC BY-SA 4.0)

Images d'œuvres protégées par le droit d'auteur dans *Wikipédia*

Il est à noter que certaines images présentes sur *Wikipédia* ne sont pas gérées dans *Wikimedia Commons*. Ce sont des images d'œuvres protégées par le droit d'auteur, ne bénéficiant pas d'exception, mais dont les différentes communautés wikipédiennes ont estimé qu'elles étaient utiles et qu'elles ne posaient pas de problèmes majeurs. Elles ne peuvent cependant être utilisées que dans des conditions précises sur le principe du "fair use" américain ou de "l'utilisation équitable" canadienne. La communauté wikipédienne francophone accepte ainsi les logos de marques déposées, les photos de bâtiments récents (même en l'absence de Liberté de panorama) et les photos de pièces de monnaie. La version anglophone de *Wikipedia* accepte notamment, à certaines conditions, les timbres, les couvertures de livre, les pochettes de disques, les captures d'écran de logiciel ou encore les images historiquement importantes. *Wikipedia* en néerlandais ne tolère, par contre, aucune exception.

Contenu

Quittons les aspects juridiques pour parler du contenu. Tout d'abord, précisons que *Wikimedia Commons* ne contient pas uniquement des images. Comme son nom l'indique, le site est ouvert à toute sorte de médias libres. Cependant, certains médias sont plus difficilement libres de droit. Une vidéo d'un concert, même téléchargée par le caméraman, pose un problème de droit pour la musique, les paroles, l'éclairage, etc. Les images représentent donc plus de 95 % des médias contenus sur le site. Nous nous attarderons davantage sur ce média mais parlons quand-même brièvement des autres[7].

Audio, vidéo, 3D

L'audio ne représente qu'une petite part des médias présents dans *Wikimedia Commons* mais on compte quand-même près d'un million et demi de fichiers. Bien sûr, vous ne trouverez pas le dernier album d'Angèle ou un enregistrement inédit d'un concert des Rolling Stones. Par contre, il est possible d'écouter des poèmes de Guillaume Apollinaire récités par le poète en personne. Certains discours politiques sont également présents. Citons Barack Obama ou Joseph Goebbels. Côté musique, vous pourrez écouter quelques hymnes nationaux ou quelques extraits de musique classique. Enfin, de nombreux bruitages et la prononciation de certains noms propres constituent une part non négligeable de ces sons.

Wikimedia Commons n'est pas *YouTube*. On y trouvera cependant plus de 150 000 vidéos. Des vidéos à caractère scientifique ou historique, comme, par exemple, des images d'une chenille qui se déplace sur une feuille, des manifestations françaises contre la réforme des retraites ou encore du relais de la torche pour les Jeux olympiques de Londres.

Enfin, ajoutons l'ouverture à des fichiers 3D, avec actuellement environ 1200 fichiers.

Images

Passé les aspects légaux, toutes les photos sont quasi autorisées à condition d'avoir une dimension éducative. On trouvera donc plus de 57 millions d'images réutilisables par tout un chacun. On peut donc trouver de tout : du coucher de soleil sur le Lac Léman à la famille de tigres blancs en passant par le schéma du cœur, Adeline Dieudonné lors d'une séance de dédicaces ou encore une carte des États-Unis montrant les résultats de la primaire républicaine de 2016.

Fig. 4 : Photo lauréate du concours de la photo de l'année 2019. (Rodney Ee, CC BY 2.0)

Limitations

Wikimedia Commons n'est évidemment pas un site pour partager des photos de vacances. Des photos comme Mamy prenant un bain de soleil, fiston couché sur son crocodile gonflable ou les parents posant à la terrasse d'un café ont quand-même un intérêt assez limité, en dehors des proches. Si le contributeur a téléchargé la photo dans le but précis d'illustrer un article de *Wikipédia*, pourquoi pas mais s'il téléverse 350 photos de sa famille sur la plage ou à la piscine, ça posera un problème.

Toutes ces photos sont hébergées sur un serveur aux États-Unis et ont pour vocation d'être utilisées partout dans le monde. Contrairement à *Facebook*, qui limite la publication de certains types de photos jugées déplacées, Wikimedia n'interdit pas d'images sous prétexte qu'elles peuvent heurter la sensibilité d'une partie du public, tant que ces images restent légales dans la plupart des pays, et sensiblement aux États-Unis, et qu'elles peuvent avoir une dimension éducative.

On y trouvera, par exemple, des photos de nus et même d'organes génitaux. Cependant, *Wikimedia Commons* n'est pas un site pornographique et, dans ce cas précis, n'acceptera pas de photos qui seraient trop similaires à d'autres déjà dans la banque d'images. À noter d'ailleurs que ces images peuvent être incluses dans des pages *Wikipédia* et qu'elles ne sont pas cachées.

On y trouvera aussi des images violentes : des images de guerre, de corps mutilés, de corps sans vie, de manifestations violentes, etc. L'idée n'est évidemment pas de montrer gratuitement de la violence. Ces photos ont généralement une valeur historique ou didactique, même si on peut en débattre.

De même, on trouvera des caricatures antisémites, des représentations anciennes du Prophète Mahomet ou des dessins anticléricaux. À nouveau, ces images ne servent pas de propagande mais ont une valeur historique.

Importations et dons d'images

Si beaucoup d'images ont été téléchargées par des particuliers, des accords ont également été conclus avec des institutions qui ont offert de nombreuses photos. Ces dons favorisent la mise à disposition de photos historiques ou d'images de qualité pour des sujets scientifiques. Ainsi les articles de *Wikipédia* peuvent être illustrés mais aussi des présentations ou des articles.

Parmi ces institutions, citons des instituts d'archives nationaux (Bundesarchiv en Allemagne, Nationaal Archief aux Pays-Bas, Archives fédérales suisses, Istituto Centrale per gli Archivi en Italie, National Archives aux États-Unis...), des bibliothèques (Bibliothèque interuniversitaire de Santé de Paris, Biblioteca europea di informazione e cultura de Milan, ETH-Bibliothek de Zurich,...), des musées (Museo Nazionale della Scienza e della Tecnica " Leonardo da Vinci " de Milan, Nordiska museet de Stockholm, Muséum de Toulouse...), des universités (Université de Neuchâtel, Universidad de la Comunicación à Mexico,...) ou d'autres institutions (Naturalis Biodiversity Center de Leyde, Château de Versailles, Konrad-Adenauer-Stiftung...)[8].

Des photographes professionnels ou des agences photographiques ont également apporté de nombreuses images plus anciennes : des portraits de personnalités pour les studios Harcourt, des clichés pris lors de nombreux voyages depuis les années 1950 pour les photographes Robert Brumter ou Françoise Foliot, ...

"Et la Belgique ?" me direz-vous. Si aucune institution belge ne figure bizarrement dans la liste des partenariats, on peut cependant trouver plusieurs institutions qui ont téléversé des médias. KBR, le KMSKA ou le Museum Plantin Moretus proposent principalement des reproductions d'œuvres d'art issues de leurs collections : peintures, estampes, dessins. Le Musée Horta quant à elle a mis à disposition des photos anciennes d'immeuble de l'architecte. On y trouvera par exemple une photo des magasins Waucquez, avant que s'y installe le Musée de la Bande dessinée (voir fig. 5). En plus de certains documents numérisés, le Mundaneum a mis en ligne des photos historiques notamment du Palais mondial ou de Paul Otlet.

D'autre part, les licences libres étant de plus en plus répandues, il n'est pas rare qu'on retrouve sur *Wikimedia Commons* des images qui ont été initialement mises en ligne sur d'autres sites. L'exemple le plus flagrant est *Flickr*. Ce site de photos propose en effet différentes licences Creative Commons.

Toutes ne sont cependant pas compatibles avec *Wikimedia Commons* mais, quand elles le sont, les images peuvent être importées légalement dans la banque d'images.

Les points faibles

Tous les dons d'institutions permettent de réduire le manque de photos anciennes. Par "anciennes", on peut pratiquement dire d'avant la popularisation de la photo numérique. Celle-ci a, non seulement, fait exploser le nombre de photos prises dans le monde mais a permis aux banques de photos sur Internet de se développer. Si les photos prises avec un appareil photo argentique ou un procédé plus ancien sont les bienvenues sur *Wikimedia Commons*, elles doivent d'abord être numérisées, et si possible, en bonne qualité. Cela demande donc du matériel et plus de temps.

Fig. 5 : Les magasins Waucquez, don de la Maison Horta. (Domaine public)

Il est impossible de savoir combien de photos sont originellement numériques dans *Wikimedia Commons* mais on constate clairement un décalage entre les époques. Ainsi, si on regarde le nombre de photos de la construction de la Sagrada Família à Barcelone, on voit que les photos récentes sont bien plus nombreuses : 36 photos prises au 20e siècle (dont aucune dans les années 1960 et 1970), 308 prises dans les années 2000, 954 dans les années 2010.

Les photos de personnalités sont aussi l'un des points faibles. Un petit tour sur *Wikipédia* en français nous permet de voir que plus de 12 000 personnalités liées au cinéma présentes dans l'encyclopédie ne disposent pas de photos. Ce chiffre monte à près de 14 000 pour les politiques[9].

Métadonnées des médias

Chaque média doit pouvoir être retrouvé et est donc accompagné de métadonnées importées automatiquement ou données par l'utilisateur qui a téléchargé le média.

Légende

Depuis quelques temps, il est demandé aux utilisateurs d'indiquer une "légende". Ce champ est un peu redondant avec la description qui est également requise. L'idéal serait que cette légende soit celle par défaut quand on souhaite incorporer une image dans un article de *Wikipédia* mais il n'en est rien. Il est possible d'indiquer une légende dans différentes langues.

Données structurées

Depuis peu également, il est possible de joindre à la photo des éléments *Wikidata*, la base de données libres que la Wikimedia Foundation essaie d'imposer ces dernières années. L'avantage par rapport au système de catégories que je décrirai plus tard et aux autres champs complétés par l'utilisateur est le multilinguisme. Les éléments *Wikidata* sont déjà traduits dans différentes langues. Donc, si je télécharge une photo de singe, j'indique "singe" comme donnée structurée, un utilisateur germanophone verra "Affe", un lusophone verra "macaco", etc. Cependant, *Wikidata* reste le produit Wikimedia le plus compliqué et on ne voit pas bien l'utilité, si la donnée structurée ne peut pas directement servir à la recherche.

Description

La description décrit le média. Plus une description sera détaillée, mieux on pourra retrouver le média. Elle peut être introduite dans différentes langues.

Date

Le système repère automatiquement la date et l'heure de la photo. Elle peut cependant être modifiée.

Source

L'utilisateur doit indiquer s'il s'agit d'un travail personnel ou si le média provient d'une autre source. Dans certains cas, il est indiqué qu'il provient d'un autre fichier.

Description	**Français :** Place Dumon à Woluwe-Saint-Pierre, en Région bruxelloise, Belgique, en juillet 2018 **Nederlands:** Dumonplein te Sint-Pieters-Woluwe, Brusselse Gewest, België, in juli 2018
Date	31 juillet 2018, 17:05:12
Source	Travail personnel
Auteur	Guy Delsaut
Lieu de la prise de vue	50° 50′ 23,8″ N, 4° 27′ 53,9″ E Voir cet endroit et d'autres images sur : OpenStreetMap - Google Earth

Fig. 6 : Exemple de description.

Auteur

On parle ici de l'auteur de la photo. S'il s'agit d'un travail personnel, c'est forcément l'utilisateur mais il peut indiquer un autre nom que son nom d'utilisateur.

Lieu de la prise de vue

Le système récupère les coordonnées géographiques si elles figurent dans les métadonnées de la photo. Le contributeur peut également les ajouter lui-même, si cela a un sens. L'utilisateur, lui, pourra localiser directement l'endroit où a été prise la photo.

Conditions d'utilisation / exception au droit d'auteur

Sous la description, on retrouve toujours au moins la licence, avec le résumé des conditions d'utilisation. Il arrive aussi qu'on indique sur quelle base juridique, la photo est autorisée. Il existe des modèles pour faciliter la vie des contributeurs. Par exemple, en ajoutant {{FoP-Belgium}}, un utilisateur ajoute un bandeau reprenant les dispositions de la loi relatives à la Liberté de panorama en Belgique. Le texte apparaît dans la langue de l'utilisateur.

Catégories

En bas de page, vous trouverez les catégories auxquelles est rattaché le média.

Autres données

S'il s'agit d'une photo, des données techniques (marque et modèle de l'appareil photo, ouverture focale, orientation...) sont également disponibles. Vous trouverez aussi la liste des pages des projets Wikimedia qui utilisent l'image.

Organisation des médias

Les médias sont rattachés à une ou plusieurs catégories, elles-mêmes rattachées à une ou plusieurs catégories, qui forment donc une arborescence. Cette classification a été élaborée au cours du temps et en fonction des besoins. Les médias peuvent être rattachés à tous les niveaux.

Toutes les catégories sont libellées en anglais. Certains noms propres, intraduisibles dans la langue de Shakespeare, restent dans leur langue d'origine. Cependant, les règles ne sont pas clairement définies et, pour Bruxelles, ça pose un problème de cohérence. Ainsi les catégories liées aux communes bruxelloises portent le nom de la commune en français sauf Bruxelles (Brussels) et Forest (Vorst-Forest).

La finesse de la classification dépend beaucoup du nombre de photos et de la volonté des contributeurs. Prenons un exemple. Au départ, il y a quelques photos des rues et places d'Ixelles, placées dans la catégorie "Streets in Ixelles". Au fur et à mesure que le temps passe, le nombre de photos de la Place Eugène Flagey augmente. Quelqu'un crée la catégorie "Place Eugène Flagey/Flageyplein" reliée à la catégorie "Streets in Ixelles". Parmi les photos de la place, un grand nombre montre l'ancien siège de l'Institut national de Radiodiffusion (INR), on crée donc une sous-catégorie "Flagey Building", etc., etc.

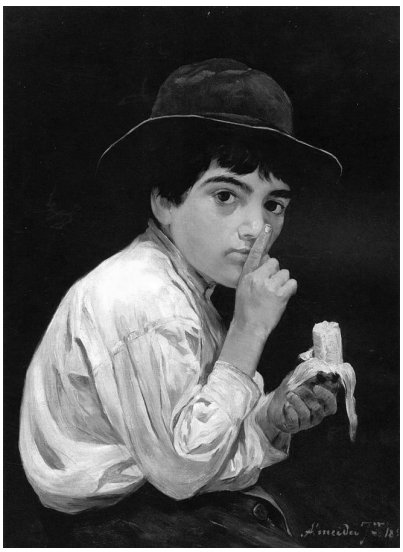

Fig. 7 : Tableau Garoto com banana de José Ferraz de Almeida Júnior. (Domaine public).

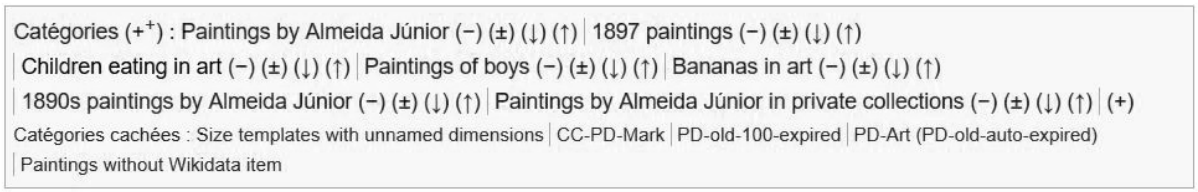

Fig. 8 : Catégories sur la photo du tableau Garoto com banana.

Chaque catégorie et chaque média peut être relié à une ou plusieurs catégories. Une œuvre d'art, par exemple, va être reliées à son auteur mais aussi au lieu où elle est exposée (musée, rue...), à ce qu'elle représente, à l'époque de l'œuvre, au courant artistique, etc. Ces catégories doivent permettre de retrouver tout aussi facilement un tableau exposé au Louvre (Category:Paintings in the Louvre), une peinture de Pierre Paul Rubens (Category:Peter Paul Rubens) ou une œuvre d'art représentant des enfants occupés à manger (Category:Children eating in art).

Comment chercher ?

Wikimedia Commons est doté d'un moteur de recherche. Il est très utile pour une recherche précise sur un nom propre, il l'est un peu moins pour une recherche plus vague. Les catégories vous seront alors d'un grand secours.

Outil de recherche interne

La recherche, en haut, à droite de la page, est basique : on tape un ou plusieurs mots et l'outil recherche en texte intégral. Comme nous sommes dans une base de données avec des descriptifs dans toutes les langues, ça peut poser des problèmes. Vous voulez une photo d'élan pour illustrer un site web sur la faune scandinave ? Si vous tapez "élan" dans l'outil et vous trouverez des photos de voitures (la Lotus Elan), de basket (l'équipe Elan Chalon ou le basketteur Elan Buller) et même des photos prises par l'utilisateur Elan5. Une recherche en anglais ("elk") permet déjà de trouver des photos de l'animal mais ne vous épargnera pas les homonymes comme une église polonaise ou des chars de l'OTAN.

En cliquant sur la loupe, on obtient la recherche avancée. Elle permet de limiter la recherche à certaines catégories (et leurs sous-catégories). Cependant, l'outil reste très limité. La recherche dans les sous-catégories est limitée à 5 niveaux et à 256 catégories. Vu l'arborescence, ces limitations sont vraiment contraignantes.

L'idéal est de trouver une image qui corresponde plus ou moins à ce que vous cherchez et de remonter les catégories. Ainsi, en tapant "elk", vous trouverez l'image d'un animal mais la photo ne vous convient pas. Grâce à la catégorie de la photo (en bas de page), vous pourrez retrouver d'autres images dans la même catégorie et même remonter dans l'arborescence et puis redescendre si la photo ne représente pas l'espèce recherchée mais un cousin. Dans mon exemple, la recherche sur "elk" m'a mené à une photo d'un Cervus cadanensis nanodes (remarquons au passage l'usage des noms scientifiques en latin). Je peux remonter à la catégorie "Species of Cervidae" et redescendre par exemple aux Alces alces.

Fig. 9 : Encart sur Wikipédia menant à Wikimedia Commons.

Via *Wikipédia*

Le moteur de recherche n'étant pas toujours très facile, le mieux est parfois de démarrer de *Wikipédia*. Je cherche une photo de la Grand-Place de Bruxelles, je trouve l'article qui s'y rapporte. Les photos qui illustrent la page ne sont qu'une sélection parmi toutes les images disponibles.

Souvent, il y a un encart, en bas de page (fig. 9) qui donne un lien direct vers la catégorie correspondante sur *Wikimedia Commons*.

S'il n'y a pas d'encart, cliquez sur une photo. Elle s'agrandira et un bouton "plus de détails" avec le logo de *Wikimedia Commons* vous redirigera sur la page de la photo. Comme précédemment, vous pourrez accéder à la catégorie et choisir la photo qui vous plaît.

Cette technique a aussi l'avantage d'éviter la question de la langue et du nom utilisés pour la catégorie. "Grand-Place" ? "Grote Markt" ? "Main Square of Brussels" ? Ce sera finalement "Grand Place (Brussels)".

Moteurs de recherche classique

Bien sûr l'utilisation de la recherche d'images dans un moteur de recherche classique reste également une option, soit en limitant la recherche au site

"commons.wikimedia.org", soit en filtrant les résultats pour n'obtenir que les images réutilisables. Selon les moteurs de recherche, on filtra par "Droits d'usage" (*Google*) ou "Licence" (*Bing*). Cela produira une recherche plus large, avec des résultats d'autres sites, mais les photos de *Wikimedia Commons* seront souvent dominantes. Notons que certains moteurs ne proposent pas cette option (*DuckDuckGo*) ou gèrent mal les licences indiquées sur *Wikimedia Commons* (*Qwant*).

Fig. 10 : Icône "Plus de détails".

Réutilisation des médias

Si vous souhaitez utiliser la superbe image que vous avez trouvée, vous devrez respecter la licence, qui vous demandera au minimum d'indiquer le nom de l'auteur, la licence (la forme abrégée du type "CC BY-SA 4.0" est autorisée) et si possible un lien vers la licence. Dans certains cas, il vous est demandé d'indiquer si vous avez modifié l'image.

Aspects collaboratifs

Comme tous les projets de la Wikimedia Foundation, *Wikimedia Commons* est un site collaboratif. Cet aspect collaboratif ne se limite pas au fait que chacun y télécharge ses photos et participe à la construction d'une immense banque d'images. Un utilisateur peut apporter également sa contribution à l'amélioration d'une photo, à la création de nouveaux médias sur base des images existantes ou à l'amélioration des métadonnées.

Retouche d'image

On le sait, la plupart de photos que vous trouvez dans la presse sont des photos retouchées à l'aide de *Photoshop* ou d'un logiciel similaire. Ces outils ne servent pas qu'à retirer les rides ou les cernes des personnalités mais aussi à améliorer les couleurs, les contrastes, à gommer un reflet... Tout le monde n'a pas les connaissances, ni les outils adéquats pour améliorer une photo. C'est là qu'une entraide peut exister. Certains utilisateurs retouchent les photos des autres.

Cependant, il vaut mieux éviter de modifier des photos de personne que vous ne connaissez pas. Éclaircir une sculpture à contre-jour, rendre transparent un fond inutile dans le cas de logos, armoiries ou objets comme une pièce de monnaie, gommer les yeux rouges,... sont autant de cas qui sont appréciés et même souhaités. Cependant modifier les couleurs d'un coucher de soleil immortalisé par un photographe professionnel pourrait vous attirer les foudres dudit photographe, même si c'est totalement autorisé par la licence qu'il a consentie.

Création d'autres images

Pour illustrer un article de *Wikipédia*, par exemple, une carte est parfois nécessaire. Il est évidemment inutile de réinventer la roue à chaque fois. Si on a besoin d'une carte d'Europe illustrant le revenu moyen par pays, le nombre de malades du covid-19 ou localiser les monarchies, on ne va pas redessiner la carte du continent à chaque fois. Autant réutiliser une carte et modifier les couleurs.

Autre scénario : on veut illustrer un article sur une personnalité. Il y a bien une photo d'elle sur

Fig. 11 : La même carte d'Europe a servi pour réaliser une carte montrant le statut des unions homosexuelles et celui de la peine de mort.
(Silje L. Bakke, CC BY-SA 2.5, et Lcmortensen, domaine public)

Fig. 12 : Le Louvre ? Vraiment ?
(Ice201:Játi Þór Mikjálsson, CC BY 2.5)

Wikimedia Commons mais elle n'est pas seule. On peut parfaitement recadrer la photo et isoler l'individu, si la photo le permet. L'inverse est vrai aussi, il peut être utile de créer une image à l'aide de plusieurs images. C'est souvent le cas pour l'image de l'infobox[10] des articles de *Wikipédia* portant sur des villes. Une seule photo étant réductrice, des utilisateurs ont créé des montages avec les principaux monuments de chaque ville.

Amélioration des métadonnées

Le principe d'une banque de photos est quand-même de retrouver les images qui correspondent à notre besoin. Donc, plus on est précis et plus elle a des chances d'être retrouvée. Cependant, les imprécisions et les erreurs sont fréquentes. Quelqu'un prend la photo d'une rue d'Oslo et la télécharge sur *Wikimedia Commons* avec comme seule information "Rue d'Oslo", en français dans le texte. Un Osloïte peut voir cette photo et préciser le nom de la rue, voire les bâtiments qu'on y voit. Il peut aussi traduire la légende.

Parmi les perles, on trouvera par exemple, cette photo (voir fig. 12) de la Place Rogier, à Bruxelles, du temps où elle était pourvue d'une pyramide de verre. Le contributeur islandais, observant une pyramide, y a vu une succursale du Louvre et a indiqué comme description "Le Louvre í Brussel, Belgíu". Raté ! Le Louvre n'a jamais eu de succursale à Bruxelles. Cette description n'avait jusqu'ici jamais été corrigée depuis son téléchargement en 2007.

Plus étonnant encore, cette photo de la statue de Pieter Bruegel l'Ancien par Tom Frantzen (voir fig. 13) que le contributeur avait nommé "Un géant parmi nous" avec comme descriptif en anglais "Gulliver is upon us"[11] et sans aucune catégorie. Aujourd'hui, elle a été correctement réintitulée, décrite et catégorisée par d'autres contributeurs que le photographe.

Organisation

Un apport collectif consiste aussi à réorganiser les médias. Une catégorie avec trop de photos ne permet pas de s'y retrouver facilement. On peut donc rassembler des photos dans des catégories plus petites et plus précises. Cela nécessite de créer de nouvelles catégories.

De même, on peut relier des catégories à d'autres catégories pour les trouver plus aisément. Par exemple, si une sculpture fait l'objet d'une catégorie reliée à l'artiste et à la ville où elle est située, on peut ajouter une catégorie reprenant le sujet de l'œuvre. Exemple la catégorie "Music in art" si elle représente un musicien.

Fig. 13 : Gulliver ou Bruegel ?
(Petit lait, CC BY-SA 4.0)

Conclusion

Si vous ne deviez retenir qu'une seule chose de cet article, c'est sans doute ceci : plus de 60 millions de médias libres de droit sont proposés dans *Wikimedia Commons*. C'est beaucoup et, en même temps, on peut encore faire mieux, notamment, pour les photos anciennes ou des sujets plus difficilement abordables. Alors, n'hésitez pas à convaincre vos institutions, si elles possèdent un fonds de photos, à les partager sur *Wikimedia Commons*. N'hésitez pas non plus à améliorer les descriptifs et les catégories des médias présents dans la banque de données.

Guy Delsaut
Val des Seigneurs 142 bte 50
1150 Bruxelles
delsautg@gmail.com

Avril 2020

Notes

1. *Wikimedia Commons* [en ligne]. <https://commons.wikimedia.org> (consulté le 23 avril 2020).
2. Exactement 61 875 265, à l'heure de la dernière relecture de cet article (4 juin 2020).
3. Pour l'explication des licences Creative Commons, je vous renvoie à : Thoumsin, Pierre-Yves. "Certains droits réservés" : L'utopie pragmatique de Creative Commons. *Cahiers de la Documentation = Bladen voor Documentatie* [en ligne], mars 2010 (consulté le 18 mars 2020), vol. 64, n° 1, p. 6-11. <https://www.abd-bvd.be/wp-content/uploads/2010-1_Thoumsin.pdf>.
4. La loi française, par exemple, autorise la diffusion d'une oeuvre exposée dans l'espace public, sans autorisation spécifique de l'auteur, à certaines conditions. L'une d'entre elles spécifie que la diffusion ne peut se faire à des fins commerciales. Cette condition exclut les photos d'oeuvres exposées dans l'espace public français et dont le droit d'auteur n'a pas encore expiré, puisque la photo ne pourra pas être exploitée commercialement.
5. À ce sujet, je vous invite à lire : Delsaut, Guy. Le statut de la liberté de panorama en Belgique et ailleurs : Entre droit d'auteur modernisé et flou artistique. *Cahiers de la Documentation = Bladen voor Documentatie* [en ligne], juin 2017 (consulté le 18 mars 2020), vol. 71, n° 2, p. 30-39. <https://www.abd-bvd.be/wp-content/uploads/2017_2_Delsaut.pdf>.
6. *Wikimedia Commons* adopte l'interprétation de la loi belge qui permet la diffusion commerciale de représentation d'œuvres situées dans l'espace public. Cette interprétation est cependant contestée, comme je l'expliquais dans notre mon article de juin 2017.
7. Les statistiques par format de fichier sur : Statistiques sur les médias. *Wikimedia Commons* [en ligne], 10 avril 2020 (consulté le 10 avril 2020). <https://commons.wikimedia.org/wiki/Special:MediaStatistics>.
8. Pour une liste des partenariats : Commons:Partenariats. *Wikimedia Commons* [en ligne], 26 octobre 2019 (consulté le 8 avril 2020). <https://commons.wikimedia.org/wiki/Commons:Partnerships/fr>
9. Ces chiffres sont donnés grâce aux catégories cachées de *Wikipédia* "Article à illustrer...".
10. On appelle "infobox" dans *Wikipédia*, le cadre à droite qui reprend l'essentiel de l'information sous forme de fiche et généralement d'une image.
11. "Gulliver est parmi nous"

DOSSIER

Saving the web

Logo du projet PROMISE

Logo van het PROMISE project

NATIONAL WEB ARCHIVES
THE LAND OF PROMISE FOR RESEARCHERS

Niels BRÜGGER
Professor, Department of Media and Journalism Studies, Aarhus University

This article is based on the keynote "National web archives — the land of promise for researchers", presented at the conference "Saving the Web: the Promise of a Belgian web archive", KBR, the Royal Library of Belgium, Brussels, 18 October 2019.

■ This article outlines some of the major topics of interest related to national web archives and their potential use by researchers. In the first main section the major differences between the US and Europe are briefly outlined, followed by an identification of some of the issues that are relevant when discussing web archives in Europe; this is done by focusing on which countries do have one, more or no web archives, which strategies can be used, and what characterises the access conditions. In the next main section the potential impact of the characteristics of web archives on transnational European history writing is debated, and to that end a case study of the Danish web domain is used. Finally, in the last main section, the possible future of the situation in Belgium is briefly discussed.

■ Cet article présente certains des principaux sujets d'intérêt liés aux archives nationales du web et à leur utilisation potentielle par les chercheurs. Dans la première grande section, les principales différences entre les États-Unis et l'Europe sont brièvement exposées, suivies d'une identification de certains des problèmes pertinents lors de la discussion sur les archives du web en Europe ; cela est fait en se concentrant sur les pays qui ont une, plusieurs ou pas d'archives du web, les stratégies qui peuvent être utilisées, et ce qui caractérise les conditions d'accès. Dans la grande section suivante, l'impact potentiel des caractéristiques des archives du web sur l'écriture de l'histoire européenne transnationale est débattu, et à cette fin, une étude de cas du domaine web danois est utilisée. Enfin, dans la dernière grande section, l'avenir possible de la situation en Belgique est brièvement abordé.

■ In dit artikel wordt dieper ingegaan op een aantal belangrijke aandachtspunten in verband met nationale internetarchieven en het potentiële gebruik ervan door onderzoekers. In de eerste hoofdparagraaf worden de voornaamste verschillen tussen de VS en Europa beknopt toegelicht. Daarna volgt een identificatie van een aantal problemen die relevant zijn als het over internetarchieven in Europa gaat. Dit gebeurt door te focussen op welke landen een, meerdere of geen internetarchieven hebben, welke strategieën er kunnen worden gebruikt, en wat de toegangsvoorwaarden inhouden. In de volgende hoofdparagraaf wordt de potentiële impact besproken van de kenmerken van internetarchieven op de transnationale Europese geschiedschrijving. Hiertoe wordt een case study van het Deense internetdomein gebruikt. In de laatste hoofdparagraaf wordt kort ingegaan op de mogelijke toekomstige situatie in België.

Why national web archives matter — a personal story

I am very fond of beers, Belgian beers in particular. My interest in Belgian beers started in the early 1990ies when together with a colleague I went to Belgium to sign the contract for the edited volume Lyotard, *les déplacements philosophiques* (co-edited with F. Frandsen and D. Pirotte) that was published in 1993 by De Boeck-Wesmael in Brussels. We had been driving for an entire day and were very thirsty when arriving, so we immediately went to a small bar, and here the exciting world of Belgian beers opened up. When back in Denmark I wanted to pursue this new interest, but since beers from Belgium were by no means available in Denmark at the time, I had to limit my interest to reading about them. Therefore, I joined the national Belgian beer consumers association, "De Objectieve Bierproevers — Les Taste-Bière objectifs" that was later renamed "Zythos" (in 2002). Although I did not understand all the words of the association's magazine Den Bierproever it still gave me the feeling of being part of this new and exciting world (years later I have had many opportunities to take revenge and not only read about Belgian beers).

When preparing the keynote on which this article is based this early encounter with a cornerstone in Belgian culture immediately came to my mind, and since I had now moved my academic interests from contemporary French philosophy to web history I asked myself, if it would be possible to study the history of the website of "De Objectieve Bierproevers" from the early web until the association became "Zythos", that is the period 1994-2002. But I got disappointed, because at the time of writing there exists no national Belgian web archive. I had to rely on the American Internet Archive where I found many copies of "De Objectieve Bierproevers'" website from 2000-01[1]. This small example of web historiography shows that if a researcher wants to study any part of the Belgian web that has been online in the past, this researcher has to use whatever has been archived of the Belgian web by a US-based institution. The first 25 years of the Belgian online cultural heritage is either lost,[2] or has to be found outside of Belgium, with an organisation where no Belgian cultural heritage institution has had any say about curatorial choices, where no quality control is ensured, and where there exist no assurance of the long term preservation of the Belgian web. Undoubtedly, the Internet Archive is doing a remarkable job, and has been doing this

for almost 25 years, but it is not within their remit to archive national web domains, and therefore one cannot rely on that what is present in the Internet Archive is complete. That national web domains are not archived in any comprehensive manner by the Internet Archive is clearly indicated in a study of the presence of the Danish web in the national Danish web archive Netarkivet compared to what can be found in the Internet Archive. Netarkivet's archiving of the Danish web domain is based on the authoritative list of domain names on the top-level domain .dk, and only a fraction of these were found in the Internet Archive in the years 2006, 2009, and 2012.[3]

Within recent years there has been a growing national and international interest in using the archived web as a historical source because it constitutes one of the main entry points to our contemporary societies. As explained but by the UK Professor of History Jane Winters: "*the web already constitutes an unprecedentedly rich primary source, combining information from personal blogs, to formal reportage, to the communications of local and national government. It is where we socialise, learn, campaign and shop.*".[4] Also, a number of methodological and theoretical texts have been published, and the literature of empirical studies based on the holdings of web archives continue to grow.[5] Thus, no doubt that all nations need one or more web archive, for the very same reasons that they have already established libraries and archives: the cultural heritage has to be collected and preserved to be made available to researchers and society at large, no matter in which form it presents itself.[6]

In the remainder of this article I shall outline some of the major topics of interest related to national web archives and their potential use by researchers. In the first main section the major differences between the US and Europe are briefly outlined, followed by an identification of some of the issues that are relevant when discussing web archives in Europe; this is done by focusing on which countries do have one, more or no web archives, which strategies can be used, and what characterises the access conditions. In the next main section the potential impact of the characteristics of web archives on transnational European history writing is debated, and to that end a case study of the Danish web domain is used. Finally, in the last main section, the possible future of the situation in Belgium is briefly discussed.

Mapping different web archiving forms

In this section I shall outline some of the main topics that are relevant to have in mind when debating national web archives and their differences and similarities.

The US vs. Europe

There exists a divide in the ways that web archiving has been approached in the US and in Europe, the major difference being that in the US web archives are not based on a legal deposit law which they are in many European countries (exceptions are Portugal, and the Netherlands, where web archives exist that are not based on legal deposit). This implies that in contrast to Europe the web archiving initiatives in the US are either bigger or smaller than the national territory. On the one hand, there exists one supranational player, the Internet Archive, that is not entitled to archive "the US web", but rather acts as an independent player that aims at archiving the web as such, without any focus on national borders. On the other hand, a large number of smaller thematic collections of archived web have been established at universities, museums, and the like.[7] In between these two levels one finds the Library of Congress which is the closest one gets to an institution with the aim of preserving "the US web", but since the collection is focusing only on topics of interest to US culture and politics it is not intended to be comprehensive in the sense that it covers the US web space.

Mapping trends in Europe

When looking at the differences within Europe it is striking that not all countries do have a national web archive, and that some countries have more than one. If we first look at the "have" and the "have-nots" some countries do not have a national web archive (e.g. Poland and Italy), whereas in a few countries national web archives are in the making, for instance pilot projects have been established in Hungary and in Belgium.[8] But in the vast majority of European countries national web archives have been established. It is also worth mentioning the specific challenge of transnational web domains such as the .eu. The .eu web domain was composed of almost four million websites as of 2016,[9] but this large portion of the web, including all EU-funded projects with a web presence on .eu, is not preserved by any national web archive.[10] If we then look at the "have" and the "have-too-many" some countries have more than one national web archive. Notable examples are the UK where one finds the UK Web Archive at the British Library as well as the UK Government Web Archive, and the UK Parliament Web Archive; and France where there exists two national web archives, one at the Bibliothèque nationale de France that aims at archiving the entire French web space, and one at the Institut National de l'Audiovisuel, with a

focus on preserving websites related to the audio and visual cultural heritage.

If we then look at the different archiving strategies that are used to preserve as much of the national web as the national web archive is entitled to there also exist a great variety. In some countries the national web archives are established to archive an entire national web domain based on an existing list with all the web domain names that are present on the national web domain, as is the case in the UK Web Archive where the .uk web domain has been collected since 2013; this archiving strategy where the archiving institution receives a list of domain names from the national registrar of country code top-level domain names is in some cases supplemented with manual tracking of material of national interest sitting on web domains outside the national web domain, such as material on .com, .org and similar, as is the case in the Danish national web archive Netarkivet. Another strategy that is often used aims at archiving a selection of the national web domain, and what is deemed relevant to be selected varies from archive to archive; it can be specific types of websites, certain time frames, valuable cultural heritage, etc. (as is the case with the national Dutch web archive). And differences may even exist within the same national web archive, either in the form of a combination of two or more strategies, as is seen in the Danish case where three different strategies have been used since the beginning of the archiving activities (the entire national domain as well as a selection of rapidly changing websites, and websites in relation to selected events are archived); or the combination of different strategies may be the result of changes over time, for instance between 2005 and 2013 the UK Web Archive used a selective strategy, focusing on websites of historical, social and cultural significance whereas from 2013 the entire .uk web domain is archived.

Finally, when looking at the access conditions the European web archives are also very different. First, legal frameworks are not the same which is why some national web archives have free online access for everyone, e.g. Archivo.pt in Portugal, or the Icelandic Web Archive, whereas others are open for researchers only, and even within this group differences apply. Some countries have onsite access only (e.g. the Netherlands), others have onsite access, but distributed, either to other national libraries like in the UK, or to regional libraries like in France, and yet others have online access (like Netarkivet in Denmark). National web archives with no access also exist, such as the Norwegian web archive, and the web collection at the National Library of Ireland. Second, different technical access forms are in place. Most national web archives offer something like the Wayback view that is familiar to users of the Internet Archive, that is a way of presenting the web page as close as possible to what it looked like when it was online in the past. But other technical forms of accessing the archived web content are also available, some countries offer an open API access (e.g. the Portuguese Archivo.pt), others make prepared datasets ready for download (e.g. the "JISC UK Web Domain Dataset (1996–2013)" in the UK), and extraction has even been made possible in Denmark where researchers can now have subsets of the national web archive extracted and shipped to a secure computer environment at their academic institution.

In summary, one can observe that national web archives are unevenly distributed in the different European countries, some have a web archive whereas other do not, and some even have more than one, a great variety of archiving strategies are used, in several cases even within the same web archive, and access conditions are very different. The question then is how to make transnational European historical studies based on web archives.

Impact on transnational history writing — the case of national webs

It is by no means unusual that national collections, be that in libraries or archives, have not been established in the same ways when comparing them across borders. Nevertheless the differences between different national web archives are in many ways more fundamental than in other cultural heritage institutions, and they may therefore have a great impact on future studies, in particular cross-national studies. In the following I shall discuss a case where an entire national web domain has been studied, namely the historical development of the Danish web, and the question is to what an extent it would be possible to replicate such a study in another national setting, thus enabling comparisons with other national web domains.

Studying a national web

In the Danish research project Probing a nation's web domain the overall aim was to study how an entire national web domain had developed. This was done by using selected material from the Danish web archive Netarkivet, as it had been archived from 2006 to 2015, and by asking broad and explorative research questions such as: What is the size of the Danish web? How much written text and how many images are there on the Danish web? Which are the most popular social media on the Danish web? Figures 1-4 show some of the first results, where focus is on the size of websites, the hyperlinks to

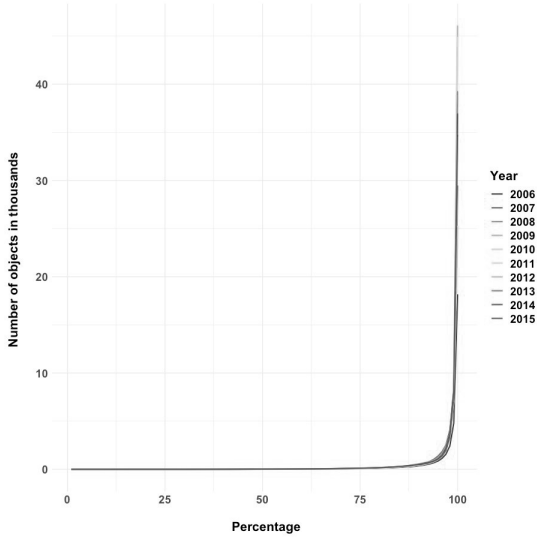

Fig. 1: Size of websites

other top-level domains (generic and countries), and to social media.[11]

Figure 1 shows the size of websites, and the figure should be read like this: the X-axis indicates how big a percentage of the websites on the Danish web has the size that is seen on the Y-axis (measured in number of files in thousands). As can be seen the vast majority of the websites on the Danish web are very small, and only a small fraction are very big, which gives the Danish web the shape of a long tail. And what is striking is that this shape has not changed during the ten years that were investigated.

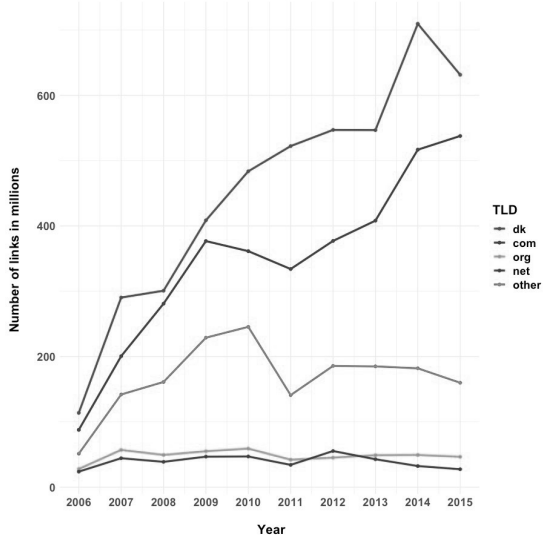

Fig. 2: Number of links from the Danish web to top-level domains

Figures 2-4 are all based on a study of all outgoing hyperlinks from all web pages on the Danish web, approximately 3 billion hyperlinks in 2006, increasing to app. 10 billion in 2010 at which level the number stays each of the following years. Figure 2 looks at the top-level web domains to which most links from the Danish web point, and as can be seen the Danish web links very much to itself (that is: to .dk), but there are also a large number of links pointing to .com. However, when looking more in detail at the numbers on which the figure is made a very large portion of the links to .com are either to web infrastructure websites, such as websites that help run websites on the Danish Web (like mysql.com, phpbb.com, blogspot.com, adobe.com, or addthis.com, in particular between 2006 and 2010), or to social media, where it is difficult to determine whether the links point to Danish pages on these social media sites or not. (in particular after 2010).

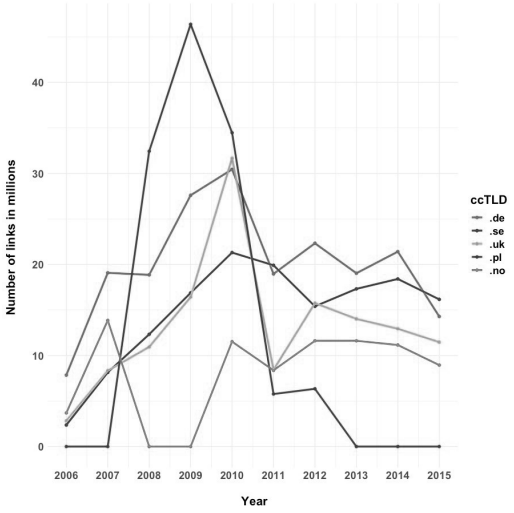

Fig. 3 : Number of links from the Danish web to other country-code top-level domains

Figure 3 depicts the country code top-level domain names to which the Danish web links, and not surprisingly the Danish web links most to its nearest neighbouring countries (Norway, Sweden, Germany, and the UK). However, in the years 2008-10 Poland is an outlier (as for now we have no good explanation to why these many hyperlinks to Poland are there in this interval).

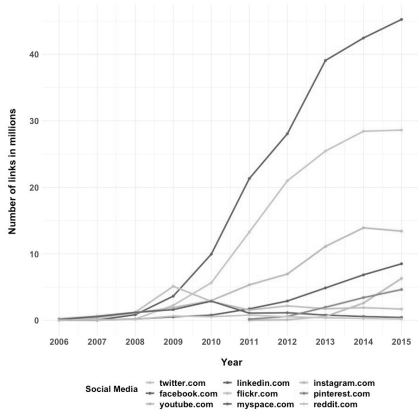

Fig. 4 : Number of links from the Danish web to social media.

Finally, Figure 4 shows the development of the links from the entire Danish web to social media. The overall trend is that in the beginning of the period there were a lot of links to myspace.com and flickr.com, but from 2010 *Facebook*, *Twitter*, and *YouTube* have been the most predominant, with pinterest.com and instagram.com as the two newcomers (the former dominates first, the latter later).

Transnational web studies of national web domains

Let us now have a closer look at the challenges of remaking this study in another European nation state, with a view to making transnational web studies of national web domains.

First, regarding the "have", the "have-nots", and the "have-too-many". If a country does not have a national web archive the study has to be based on what may be available in a transnational web archive such as the Internet Archive, with the limitations that this gives, because one is then delimited by this archive's collection strategy and coverage (cf. above about the completeness of the Internet Archive when it comes to national web domains). However, if a country has more than one national web archive this may give other challenges, because the study then has to combine the content from different web archives, potentially archived with different software, different strategies, etc. which is by no means not trivial.

Second, different archiving strategies may have been applied, including different technical settings. In cases where an entire national web domain has been archived we may have the needed data, but in cases where only a selection of the national web domain has been archived a lot will be missing, and comparisons will be made difficult — we may have to supplement with material from the Internet Archive or another national web archive with the challenges that this poses for our study. And in cases where an entire domain and a selection were archived by the same web archive we also have to combine two different collections.

Third, different access conditions may apply. One challenge is to get access to the web archive's holdings, which may or may not be possible if the researcher comes from another country. Maybe only onsite access is possible, and the researcher then has to travel to the web archive, but the web archive may even not provide the data in a form that is ready for analysis, which is very likely to be the case. For instance, the above mentioned project Probing a nation's web domain relied heavily on the availability of the web data in a form that the researchers could study. Having the data in the right form will often imply that the web data are extracted in exactly the form the researcher wants to, which is only the case in one national web archive (at least for now), at least if the data is not made available via an API.

As this case analysis clearly shows the great varieties among web archives in different national settings challenge transnational studies across web archives. Because of the size and complexity the study of the development of the Danish web domain is indeed an extreme case, but nevertheless all the identified challenges also apply even at a smaller scale. However, it is possible to mitigate or minimize these challenges. Most importantly researchers need to voice their demands and to do this in a coordinated, transnational manner. Since 2012 the research network *RESAW* (*A Research Infrastructure for the Study of Archived Web Materials*) has offered an open and welcoming forum where researchers from a great variety of academic disciplines who all study the archived web can formulate and discuss the limitations and the possibilities related to doing national as well as transnational web archive based studies.[12] And recently the international researcher network WARCnet, funded by the Independent Research Fund Denmark | Humanities, has continued this work.[13] On the web archive side the International Internet Preservation Consortium (IIPC) provides a home for discussions among web archives of any kind, national as well as more topic based. Fortunately, these three organisations have reached out to each other for years. *RESAW* and WARCnet invite web archives as members of the network, and the IIPC have opened up for presentations from researchers at their annual meeting. In summary, international organisational structures are in place to promote the fruitful use of web archives and for addressing the challenges related to studies of web archives. Also, on a national level, close collaborations exist in many countries between web archiving institutions and research communities, as is for instance seen in France, Luxembourg, the UK, Portugal, and Denmark.

Belgium — the land of PROMISE

Belgium is one of the few European countries that does not have a national web archive yet which has the consequences identified above, as also pointed out by Chambers, Mechant and Geeraert: "*the Belgian web is not systematically archived. Without a Belgian web archive there is a significant risk that essential born digital resources for contemporary and future historians will not be preserved and a significant portion of Belgian history will be lost forever.*".[14] But the project *PROMISE*, which ran for two years in 2017-19, has paved the way for the establishing of a national Belgian web archiving service.[15]

However, that Belgium does not yet have a national web archive does not mean that no web archiving initiatives have been launched. Belgium is a federated country, and therefore the existing initiatives have to be understood in this context. Chambers, Mechant and Geeraert identify the following initiatives: "Felix Archive (FelixArchief, n.d.) and the ADVN, archives and research centre on the Flemish movement (ADVN, n.d.) in Antwerp, KADOC, documentation and research centre on religion, culture, and society (KADOC, n.d.), as well as the AMSAB Institute of Social History (AMSAB, n.d.) in Ghent that have been archiving websites as part of their archival responsibilities. Furthermore, Ghent University Library has been archiving websites as part of its long-term digital preservation activities since 2007 (Archive-It, n.d.).".[16] And in 2019-20 AMSAB-ISG and Liberas/Liberal Archive have conducted the project *Catching the Digital Heritage* about the best way to archive websites within the framework of the two organisations. Also, studies of geographical Belgian entities on the web exist, most notably a study of the generic top-level domain .brussels by Waty et al.[17] In summary, despite the ongoing initiatives very large portions of the first 25 years of the national Belgian web have already been lost because of the scattered and fragmented nature of web archiving in Belgium.

It is interesting to observe that when it comes to the fragmented underwood of local and/or topic oriented web archives the situation in Belgium somewhat resembles the case of the US, but in contrast to the US neither a supra-national nor a national web archiving initiative exist. However, if — or hopefully when — Belgium establishes a national web archiving service it is important to keep intact the already existing smaller collection initiatives. The reason for this is that although an institution may be established with the remit of collecting and preserving the national web domain like .be it is not possible to do this in a fine grained manner, because none of the prevailing archiving strategies enables this. If an entire web domain is archived this takes approximately 2-3 months (as is the case in the Danish Netarkivet), and in this period a lot of the web material that was archived earlier in the archiving process is probably updated and therefore the new versions are not archived. If a more time sensitive strategy is used then the scale of what is archived cannot be very big, simply because it takes time to archive websites in-depth, and this is not possible to do in a time sensitive way with millions of websites. Thus, national web archives are trapped between the Scylla of big scale archiving and the Charybdis of archiving everything "now", between the constraints imposed by space and by time. Therefore, it can be considered a strength that a national web archive can act "on top" of more focused, but smaller web collections. However, it should not be neglected that smaller web archiving initiatives are challenged by providing the resources to sustain their web archiving activities over time, including keeping up with the most recent technological innovation on the online web as well as regarding web archiving software. One way of countering this challenge is to have the national web archive offer the technical solutions of archiving and preserving of the web, but leaving the curation to the smaller web archive collections and handing them a copy of the material that was curated by their institution so that they can make it available as they please.[18] In the Belgian case this could be termed a "federated web curation" in tandem with a national web archiving service. All in all, that Belgium has not yet established a national web archive opens up new opportunities of taking stock of all existing national web archive initiatives — take the best and make the best — and with its underwood of smaller web archiving initiatives and with a national web archive in the making Belgium can be considered the land of *PROMISE*.

Niels Brügger
Department of Media and Journalism Studies
Aarhus University
Helsingforsgade 14
8200 Aarhus N
Denmark
nb@cc.au.dk

March 2020

Bibliography

Brügger, N. *The archived web. Doing history in the digital age*. MIT Press, 2018.

Brügger, N. Web som lokalhistorisk kilde — hvad er udfordringerne? [The web as a source for local history — what are the challenges]. In Andersen, K.H.; Jansen, C.R. (eds.) *Lokalhistorie. Fortid, nutid og fremtid.* Forlaget Skippershoved, 2014, p. 279-295.

Brügger, N. (ed.). Web 25. *Histories from the first 25 years of the world wide web*. Peter Lang, 2017.

Brügger, N.; Milligan, I. (eds.). *The Sage handbook of web history*. SAGE Publishing, 2019.

Brügger, N.; Laursen, D. (eds.). *The historical web and Digital Humanities. The case of national web domains*. Routledge, 2019.

Brügger, N.; Schroeder, R. (eds.). *The web as history. Using web archives to understand the past and present*. UCL Press, 2017.

Brügger, N., Laursen, D., Nielsen, J. Exploring the domain names of the Danish web. In Brügger, N.; Schroeder, R. (eds.) *The web as history. Using web archives to understand the past and present*. UCL Press, 2017, p. 62–80.

Brügger, N.; Nielsen, J.; Laursen, D. Big data experiments with the archived Web: Methodological reflections on studying the development of a nation's Web. *FirstMonday*, 2020, vol. 25, no 3, 18 pages. <https://journals.uic.edu/ojs/index.php/fm/article/view/10384>.

Chambers, S.; Mechant, P.; Geeraert, F. Towards a national web archive in a federated country: A Belgian case study. In Brügger, N.; Laursen, D. (eds.) *The historical web and Digital Humanities. The case of national web domains*. Routledge, 2019, p. 29-44.

Hockx-Yu, H.; Laursen, D.; Gomes, D. The curious case of archiving .eu. In Brügger, N.; Laursen, D. (eds.) *The historical web and Digital Humanities. The case of national web domains*. Routledge, 2019, p. 64-72.

Laursen, D.; Møldrup-Dalum, P. Looking back, looking forward. 10 years of web development to collect, preserve and access the Danish web. In Brügger, N. (ed.) *Web 25. Histories from the first 25 years of the world wide web*. Peter Lang, 2017, p. 207–228.

Milligan, I. *History in the Age of Abundance? How the Web Is Transforming Historical Research*. McGill-Queen's University Press, 2019.

Schroeder, R., Brügger, N. Introduction. The web as history. In Brügger, N.; Schroeder, R. (eds.) *The web as history. Using web archives to understand the past and present.* UCL Press, 2017, p. 1-19.

Vlassenroot, E.; Chambers, S.; Di Pretoro, E.; Geeraert, F.; Haesendonck, G.; Michel, A.; Mechant, P. Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 2019, vol. 1, p. 85-111 <10.1007/s42803-019-00007-7>.

Waty, M., Van Hooland, S., Hengchen, S., Coeckelbergs, M., De Wilde, M., & Decroly, J. M. How hot is .brussels? Analysis of the uptake of the .brussels top-level domain name extension. *Brussels Studies. La revue scientifique électronique pour les recherches sur Bruxelles/Het elektronisch wetenschappelijk tijdschrift voor onderzoek over Brussel/The e-journal for academic research on Brussels*, 2018, collection générale, n° 119 <https://journals.openedition.org/brussels/1609>

Webster, P. Users, technologies, organisations: Towards a cultural history of world web archiving. In Brügger, N. (ed.) Web 25. *Histories from the first 25 years of the world wide web*. Peter Lang, 2017, p. 175–190.

Winters, J. Breaking in to the mainstream. Demonstrating the value of internet (and web) histories. *Internet Histories,* 2017, vol. 1, n° 1-2, p. 173-179.

Notes

1. The Internet Archive is a non-profit web archive that aims at archiving the entire web, and has been doing so since 1996, cf. Brügger, 2018, p. 92-93. The Internet Archive's collections are freely available online at archive.org.
2. The lifetime of web content is probably around two months; about this issue, see Brügger, 2018, p. 75-77.
3. Brügger, Laursen, Nielsen, 2017, p. 72.76.
4. Winters, 2017, p. 173.
5. The following publications constitute relevant places to start when wanting to explore the literature related to web archives and researcher use of web archives: Brügger, 2018; Brügger, Milligan, 2018; Brügger, Schroeder, 2017; Brügger, Laursen, 2019; Brügger, 2017; Milligan, 2019.
6. For overviews of existing web archiving initiatives see Vlassenroot et al, 2019; Schroeder, Brügger, 2017; Laursen, Møldrup-Dalum, 2017; Webster, 2017.
7. See Brügger, 2018, p. 93-98.

8. In Belgium, the project *PROMISE* (*PReserving Online Multiple Information: towards a Belgian StratEgy*), read more at <https://promise.hypotheses.org>.
9. Hockx-Yu, Laursen, Gomes, 2019, p. 65.
10. A pilot project was run by Archivo.pt, the Portuguese Web Archive. About the challenges related to the archiving of the .eu web domain see Hockx-Yu, Laursen, Gomes, 2019.
11. The project as well as the results are described in more detail in Brügger, Nielsen, Laursen, 2020.
12. Read more about RESAW on <http://resaw.eu>.
13. Read more about WARCnet on <http://warcnet.eu>.
14. Chambers, Mechant, Geeraert, 2019, p. 30.
15. *PROMISE* (*PReserving Online Multiple Information: towards a Belgian StratEgy*) was funded by the Belgian Science Policy Office (BELSPO), and the aim was to investigate the feasibility of establishing a web archive in Belgium, cf. Chambers, Mechant, Geeraert, 2019.
16. Chambers, Mechant, Geeraert, 2019, p. 31.
17. Waty et al., 2018; see also Chambers, Mechant, Geeraert, 2019, p. 30.
18. These ideas of combining national web archiving initiatives with local web archives, with local curation, but national operation are expended in Brügger, 2014.

PROMISE : UN PROJET DE RECHERCHE POUR UN ARCHIVAGE DU WEB BELGE AU NIVEAU FÉDÉRAL

Rolande DEPOORTERE
Chef du service Archivage digital, AGR
Friedel GEERAERT
Assistante scientifique en archivage du web, KBR
Gerald HAESENDONCK
IDLab UGent
Sébastien SOYEZ
Chef de travaux Service Archivage digital, AGR
Sophie VANDEPONTSEELE
Directrice des Collections contemporaines, KBR

Het artikel is opgesteld naar aanleiding van een colloquim "Saving the web: the promise of a Belgian web archive"[1], georganiseerd in KBR op 18 oktober 2019 te Brussel.

Article rédigé suite au colloque "Saving the web: the promise of a Belgian web archive"[2], organisé à KBR le 18 octobre 2019, à Bruxelles.

■ De juillet 2017 à décembre 2019, la Bibliothèque royale de Belgique (KBR) a lancé, en partenariat avec les Archives de l'Etat (AGR) et plusieurs universités et hautes-écoles belges, le projet de recherche PROMISE en vue de définir une stratégie fédérale pour l'archivage du web belge. Le présent article retrace le parcours mené par l'équipe de recherche durant plus de deux ans : état de l'art, définition d'une stratégie, étude de plusieurs scénarios et de leurs coûts et infrastructure technique. Les résultats de ce projet de recherche, financé par la Politique scientifique belge (BELSPO) ont également été présentés en octobre 2019 lors du colloque "Saving the web".

■ Van juli 2017 tot december 2019 werkte de Koninklijke Bibliotheek van België (KBR) – in samenwerking met het Algemeen Rijksarchief (ARA) en meerdere Belgische universiteiten en hogescholen – aan het onderzoeksproject PROMISE met als doel een federale strategie uit te stippelen voor de archivering van het Belgische web. Dit artikel schetst het parcours dat het onderzoeksteam gedurende meer dan twee jaar heeft afgelegd: state of the art, een strategie bepalen, studie van meerdere scenario's evenals hun kosten en technische infrastructuur. De resultaten van dit onderzoeksproject, dat werd gefinancierd door het Belgische wetenschapsbeleid (BELSPO), werden in oktober 2019 voorgesteld tijdens het colloquium "Saving the web".

Introduction au projet PROMISE

Le web occupe une place centrale dans la société et contient donc de nombreuses traces de notre histoire contemporaine. De ce fait, le web devient une ressource très intéressante pour les générations futures, raison pour laquelle différentes bibliothèques et archives (nationales) du monde entier archivent et préservent depuis des années, voire dans certains cas depuis des décennies, leur web national ou des parties de celui-ci, et donnent accès à ces collections.

En 2017 fut lancé au sein de KBR (Bibliothèque royale de Belgique) et des Archives de l'État (AGR), le projet de recherche *PROMISE*. Celui-ci avait pour objectif d'élaborer une stratégie fédérale de préservation du contenu du web belge. Le projet a été financé par la Politique scientifique belge (BELSPO) dans le cadre de leur programme BRAIN.be. En raison d'aspects techniques, juridiques et opérationnels et pour étudier les besoins des utilisateurs en matière d'archivage du web, les Archives de l'État et KBR ont collaboré avec les universités de Gand (Research Group for Media, Innovation and Communication Technologies; Ghent Centre for Digital Humanities) et de Namur (Centre de Recherche Information, Droit et Société (CRIDS)) et avec la Haute-École Bruxelles-Brabant.

Le projet, qui s'est clôturé en décembre 2019, était construit autour de quatre objectifs : 1) analyser les bonnes pratiques en archivage du web, 2) élaborer une stratégie d'archivage du web belge, 3) tester l'archivage du web et donner accès aux collections et 4) formuler des recommandations pour l'implémentation d'un service d'archivage du web durable.

L'objectif de cet article est d'offrir un aperçu de la structure globale du projet et des principaux résultats de recherche. La première partie contient un aperçu des bonnes pratiques qui ont été identifiées. La deuxième partie présente la stratégie esquissée dans le cadre du projet et la troisième partie présente les différents scénarios d'archivage du web au sein des Archives

de l'État et de KBR ainsi que l'analyse des coûts y afférents. Dans la dernière partie, l'infrastructure technique utilisée est présentée succinctement.

État de l'art

Dans le cadre du projet *PROMISE*, un certain nombre de projets d'archivage du web en Belgique et à l'étranger ont été étudiés. Certaines institutions patrimoniales belges archivent depuis des années du contenu du web, notamment : Felixarchief Antwerpen, Universiteitsbibliotheek Gent, Liberaal Archief, AMSAB - Instituut voor Sociale Geschiedenis, ADVN (Archief voor Nationale Bewegingen), KADOC, het Letterenhuis, Archief Gent ou l'Université Catholique de Louvain. De nombreuses formes d'expertises sont donc présentes en Belgique. La principale différence entre les ambitions de KBR et les Archives de l'État d'une part, et ces initiatives d'archivage du web, concerne les critères de sélection. KBR et les Archives de l'État souhaitent archiver le web belge de la manière la plus large possible alors que les institutions susmentionnées se concentrent sur certains sites web, et parfois aussi sur des médias sociaux, ayant un lien direct avec les priorités de leurs collections. Ceci est logique, mais on est en présence d'une réelle différence d'échelle. Le résultat est que KBR et les Archives de l'État ne peuvent généralement pas s'appuyer sur la même infrastructure ou approche que ces institutions, leurs pratiques n'étant pas modulables, comme par exemple l'utilisation de fichiers .zip dans une structure de dossiers déterminée. Les institutions patrimoniales belges constituent toutefois d'importantes parties prenantes pour le futur archivage du web belge puisque l'échange d'expertise et de pratiques entre les différentes institutions peut contribuer à la création d'une communauté pour l'archivage du web en Belgique.

Les initiatives en matière d'archivage du web qui ont été étudiées à l'étranger ont été sélectionnées de manière spécifique afin d'obtenir un bon mélange d'initiatives qui ont beaucoup d'expérience en matière d'archivage du web, qui opèrent dans des pays de différente taille, avec plusieurs langues nationales ou dans lesquels aussi bien les archives nationales que la bibliothèque nationale s'occupent de l'archivage du web et d'initiatives qui gèrent tous les aspects de l'archivage du web en interne ou collaborent avec des prestataires de services externes. Les initiatives ont été examinées au moyen d'une étude bibliographique complétée d'interviews semi-structurées avec des représentants des institutions en question. Les institutions suivantes ont été étudiées : la Koninklijke Bibliotheek et le Nationaal Archief aux Pays-Bas, la Bibliothèque nationale de France et l'Institut national de l'audiovisuel en France, la Bibliothèque nationale de Luxembourg, la British Library et les UK National Archives en Grande-Bretagne, la Kongelige Bibliotek au Danemark, l'Arquivo.pt au Portugal, l'Irish National Library et les Bibliothèque et Archives Canada et la Bibliothèque et archives nationales de Québec.

La politique menée au sein de ces institutions a été comparée sur quatre niveaux : sélection, accès, contexte juridique et infrastructure technique. Voici la problématique de la sélection et l'accès.

En ce qui concerne la sélection, la plupart des bibliothèques nationales combinent, dans des pays qui disposent d'un dépôt légal, des "crawls"[3] larges avec des "crawls" sélectifs. Les "crawls" larges servent à archiver le web national de manière superficielle alors que les collections sélectives visent généralement certains thèmes, événements ou circonstances imprévues. Les sites web qui sont archivés dans des collections sélectives, le sont de manière plus approfondie que ce n'est le cas pour les "crawls" larges. Il est aussi important de noter qu'il n'existe pas de définition univoque de ce que couvre exactement un web national et que les pratiques diffèrent donc d'un pays à l'autre. Quant aux archives nationales, elles constituent des collections de sites web d'institutions dont les archives doivent être déposées chez elles selon la loi sur les archives.

La moitié des institutions étudiées constitue, outre les sites web, des collections de médias sociaux. *Twitter* est archivé le plus fréquemment, suivi de *YouTube*, *Facebook*, Instagram et *Flickr*. Les institutions en question précisent toutefois que chaque canal de médias sociaux requiert une approche spécifique et que les API[4] sous-jacentes et l'infrastructure technique de certaines plateformes changent tellement vite qu'il faut pratiquement un suivi journalier pour permettre l'archivage de ce contenu de manière permanente. En outre, certaines plateformes permettent à peine d'être archivées. Initialement, l'équipe du projet *PROMISE* avait souhaité tester l'archivage aussi bien des médias sociaux que des sites web. Sur base des résultats des interviews, il a toutefois été décidé d'exclure l'archivage des médias sociaux du projet pour des raisons de faisabilité.

La manière de donner accès aux collections diffère aussi très fortement. En raison de la législation sur les droits d'auteur, la plupart des archives web sont uniquement accessibles dans les salles de lectures des institutions sauf si l'institution a reçu l'autorisation explicite des ayant-droits de donner accès librement au contenu du web archivé. Cette contrainte décourage en toute logique l'utilisation des archives du web étant donné que le public est habitué au web "live" qui est généralement en accès libre. Faire le pas vers la salle de lecture d'une institution est pour beaucoup un pont trop loin dans ce contexte. L'accès

à certaines archives du web comme au Danemark est encore plus restreint, notamment à un certain groupe d'utilisateurs, à savoir les chercheurs.

Non seulement l'endroit où des archives du web peuvent être consultées est important, mais aussi la manière dont les collections seront mises à disposition. La recherche par URL est possible dans toutes les institutions étudiées, mais la recherche plein texte ne l'est pas toujours. Cela est principalement dû à la taille des collections et donc aussi aux index établis sur la base du texte présent. Il n'est pas possible de créer l'infrastructure nécessaire pour toutes les initiatives. Dans de rares cas, il est également possible de parcourir les collections alphabétiquement ou thématiquement, mais cela ne concerne que certaines petites collections d'archives du web[5].

L'étude d'autres initiatives d'archivage du web était une phase très intéressante du projet. En effet, l'équipe du projet *PROMISE* a permis de tirer des leçons importantes des expériences acquises par d'autres institutions.

Stratégie

Pour développer la stratégie, KBR et les Archives de l'État (AGR) ont tenu compte des résultats de l'état de l'art, de l'analyse juridique réalisée par l'Université de Namur ainsi que des résultats d'une enquête de l'Université de Gand sur les besoins des utilisateurs dans le contexte de l'archivage du web.

Cette stratégie a été construite sur base du modèle d'exigences fonctionnelles de l'archivage électronique, à savoir celui publié dans la norme ISO 14721.2012, plus connu sous le nom d'OAIS[6]. Outre les 7 fonctions classiques que prévoit ce modèle (versement, gestion de données, stockage, planification de la préservation, accès, gestion opérationnelle et gestion stratégique), nous lui avons adjoint trois fonctions complémentaires, en phase avec les processus d'archivage du web, à savoir : la sélection, la collecte et le contrôle-qualité. Ce découpage fonctionnel nous a permis de construire notre stratégie sur base d'une structure logique, de la collecte jusqu'à la diffusion, en passant par la mise en archive.

Au niveau de la description du corpus de données, notre choix s'est porté sur un schéma existant qui permettait d'utiliser, tant pour KBR que pour les AGR, une liste de métadonnées identique. Il s'agit du schéma de métadonnées établi par le *Web Archiving Metadata Working Group* de l'OCLC[7]. Ce schéma, construit pour établir une liste de 14 descripteurs pour toute archive web[8], permet en outre de faire converger deux préoccupations professionnelles distinctes, à savoir celle du monde des bibliothèques et celle des archives. En effet, ce schéma peut aisément renvoyer vers les descripteurs de chaque domaine, à savoir MARC21[9] pour les bibliothécaires, et la DTD-EAD[10] pour les archivistes.

Dès que ces deux premiers choix stratégiques - le schéma fonctionnel et le schéma de métadonnées - ont été posés, il était important de définir quelle serait la base sélective de la collecte des archives web en Belgique. En effet, plusieurs choix pouvaient s'opérer. Première possibilité, celle de sélectionner l'ensemble des sites ayant ".be" comme noms de domaine génériques[11], éventuellement en y ajoutant les sites régionaux ou locaux[12]. Sur base des derniers chiffres connus[13], on peut estimer ce volume total à environ 1 million. Seconde possibilité, on peut définir des listes sélectives sur base de choix précis, comme par exemple la pertinence de sites web pour nos deux institutions. Pour établir ces listes, KBR et les AGR ont, par exemple, sélectionné des sites - ou des parties de sites - qui étaient particulièrement intéressants compte tenu de leurs missions spécifiques, ou qui constituaient une obligation légale de conservation. À titre d'exemple, pour KBR, des sites en lien notamment avec la littérature, la musicologie ou la BD, et pour les AGR, des sites (para-)publics liés au fonctionnement de l'État fédéral, régional ou local. Après analyse, KBR a retenu un peu moins de 1.000 sites et environ 1.400 pages. Les AGR quant à elles ont établi leurs listes sélectives selon 3 niveaux : environ 650 sites fédéraux, 1.400 sites régionaux et locaux, et 300 d'origine privée. Lors de la mise en place de la stratégie globale, il restera à choisir ou à combiner l'une ou l'autre de ces listes, et d'envisager le cas échéant la sélection d'un échantillon aléatoire[14].

La stratégie de la collecte proactive est une des options retenues par le projet[15]. Une option alternative pourrait être celle d'attendre que des gestionnaires de sites web nous versent, soit volontairement soit par obligation légale, leurs contenus informationnels. Mais la pratique nous démontre que les résultats pourraient être relativement peu uniformes et parcellaires. Outre ce choix, nos deux institutions devront également s'accorder sur la fréquence (annuelle) et sur la profondeur (complète ou quelques niveaux) de cette collecte. Pour effectuer techniquement cette collecte proactive, notre choix s'est porté sur l'outil *Heritrix*[16]. Son fonctionnement est relativement simple : lors de son passage sur une page web, cet outil prend une copie complète du contenu d'un site et le sauvegarde dans le format de fichier WARC[17]. Ce sera à partir de ce fichier WARC qu'une nouvelle consultation du site sera possible lors de l'étape de l'accès. Après cette collecte, il est indispensable d'effectuer un contrôle-qualité, soit systématique pour un volume limité de sites, soit sur base d'échantillons. Il existe deux possibilités

pour effectuer ce test de qualité : manuellement ou semi-automatiquement. Manuellement tout d'abord, en prenant un nombre limité de sites collectés et en les comparant avec la version d'origine, de manière visuelle mais également en testant les hyperliens. Cette méthode, très gourmande en ressources, est la seule qui permet de faire un contrôle systématique du contenu archivé. Sur base de notre expérience, il faut compter en moyenne 10 jours de travail pour contrôler environ 600 sites. Par ailleurs, il est possible de mettre en place des contrôles semi-automatisés sur des sites, sur base de paramètres techniques prédéfinis[18]. Dans le cadre du projet *PROMISE*, plusieurs paramètres ont été testés, mais ne constituent pas à eux seuls la solution. Notre conclusion porterait davantage sur une solution mixte, de contrôle manuel d'une partie limitée de la collection, et de contrôle semi-automatique des collectes larges.

Dès que la collecte et le contrôle-qualité ont été réalisés, les collections sont ensuite transférées dans les dépôts de chaque institution, par le biais de leur procédure de transfert en place. Ensuite, il est indispensable de régler la question de la mise en archive proprement dite, à savoir de mettre en place une gestion des (méta)données, une infrastructure de stockage ainsi que garantir la pérennisation des objets numériques archivés. Pour ce qui est de la gestion de (méta)données, chaque institution (KBR et AGR) peut choisir son mode de gestion propre, pour autant que la description avec le schéma commun soit maintenue. Les données administratives de gestion seront intégrées dans chaque catalogue spécifique (*Syracuse* pour KBR et *SAM* pour les AGR), et il sera décidé de leur adjoindre les métadonnées de description complémentaires, à savoir MARC21 pour KBR, et la DTD EAD pour les AGR. Cette transposition sera facilitée par le choix initial du schéma de métadonnées de l'OCLC. La liaison entre ces métadonnées et les fichiers WARC, constituant l'archive web, sera effectuée à l'aide de fichiers METS. Pour ce qui est du stockage, deux stratégies sont envisageables : soit chaque institution gère sa propre collection sur ses propres espaces de stockage, soit elles mutualisent une de leurs *infra*structures communes prévue par le LTP[19], financée actuellement par la Politique scientifique fédérale (BELSPO). Ces choix devront principalement conduire à une rationalisation des coûts de stockage (cf. *infra*). Les deux institutions avertiront la plateforme LTP lorsqu'un format déterminé sera arrivé en fin de vie. Chaque année, tous les formats de fichiers se trouvant dans les collections de l'archive web seront identifiés sur base de quoi un tableau de contrôle sera établi. Les institutions effectueront un monitoring technologique continu pour rester au courant des derniers développements.

Enfin, plusieurs éléments concernant la stratégie de l'accès se sont dégagés au terme du projet *PROMISE*. Tout d'abord, d'un point de vue de l'accès physique, et en respectant l'actuelle législation sur le droit d'auteur, il ne sera possible d'accéder à certains sites web que par le biais de l'une des salles de lecture de nos institutions. Bien entendu, cette restriction ne s'appliquera qu'aux sites web soumis à ce droit d'auteur. Pour la majorité des sites web créés par des autorités publiques, ce problème ne se posera pas, et une consultation directement en ligne est envisagée. En ce qui concerne l'interface d'accès aux collections web de KBR et des AGR (cf. *infra*), construite autour du modèle commun de description (OCLC), il est envisagé de développer les fonctionnalités suivantes : recherche par URL, recherche plein texte et par mots-clés et recherche par organismes "responsables" du site. L'interface se basera sur un "replay" des sites[20], à partir des collections archivées. Il est également envisagé de créer une interface de récupération de sets de données plus larges, ce qui permettrait aux chercheurs/ses d'analyser avec d'autres outils ces contenus de données.

Scénarios et calcul des coûts

Il est difficile de présenter tous les scénarios envisagés tant les combinaisons possibles sont nombreuses. En effet, si l'on se base sur les différents types de collecte, dont les coûts sont mutualisés entre KBR et AGR, ou non, et ce pour chaque phase du modèle OAIS, il existe plus de cent combinaisons possibles. Il a été nécessaire de faire un choix et quatre scénarios combinant un intérêt tant sur l'approche de la sélection que sur la mutualisation des coûts ont été sélectionnés.

Dans l'objectif de prendre une décision quant à la définition d'une politique structurelle pour l'archivage du web belge, la KBR et les AGR ont travaillé sur différents scénarios d'archivage du web. Cette analyse a permis d'offrir une variété d'approches institutionnelles possibles de l'archivage web en fonction des ressources qui peuvent être mises à disposition. Ces scénarios sont basés sur différentes approches concernant la sélection et couvrent trois niveaux différents : complet, intermédiaire et basique.

Le scénario complet couvre la collecte de collections sélectives et un large éventail comprenant la collecte de 100% du web belge. En ce qui concerne les AGR, les collections sélectives comprendraient d'abord les sites web des institutions fédérales dont les archives doivent être légalement conservées, ensuite les sites web des villes, des communes et des organismes parapublics et enfin les sites web des archives privées. Pour KBR, les collections sélectives

comprendraient des sites web étroitement liés aux collections existantes et s'inscrivant pleinement dans la charte de développement des collections qui définit les grands principes d'acquisition des collections. Les sites web qui font partie des collections sélectives seraient collectés dans leur entièreté. Par contre, dans le cas de la collecte large, la sélection se limiterait à parcourir uniquement les couches supérieures des sites web, constituant ainsi un échantillon.

Dans le scénario intermédiaire, la collecte large serait limitée à un échantillon choisi au hasard à hauteur de 10 % du web belge. Les collectes sélectives pour les AGR seraient limitées aux sites web des institutions fédérales soumises à la loi sur les archives. Pour KBR, les collections sélectives couvriraient le même contenu que dans le scénario complet.

Le scénario dit basique comprendrait les mêmes collections sélectives que dans le scénario intermédiaire, mais en excluant la collecte large.

Un quatrième scénario a également été envisagé : il s'agit de celui de l' " outsourcing ". Cette piste pourrait constituer une alternative intéressante dans le cas où il ne serait pas possible d'effectuer l'archivage du web belge en interne. Il semble aussi intéressant de connaître le prix de ce type de prestation et de le comparer au scénario le plus complet. Cependant, cette piste présente quelques limites : si l'abonnement est stoppé, il sera alors nécessaire de développer une infrastructure propre pour consulter les fichiers WARC.

Afin d'estimer le coût annuel de l'archivage du web belge, nous avons listé toutes les tâches à exécuter à chaque étape du modèle OAIS, en distinguant les tâches à répéter chaque année, les coûts annuels de maintenance et les investissements périodiques pour garantir la pérennité de l'infrastructure. Le calcul a porté sur les ressources humaines à mobiliser et sur l'infrastructure douce et dure, incluant le coût de la maintenance et des futures mises à jour. La projection porte sur une période de cinq ans, dont a été tirée une moyenne annuelle. Les ressources humaines ont été estimées en nombre d'heures pour effectuer les traitements, sur base de tests effectués sur un échantillon de sites web capturés. Le calcul du coût humain intègre le salaire horaire moyen des différents profils de compétences nécessaires, en fonction de leur niveau de qualification, pour trois familles de compétences identifiées: "archiviste numérique", "bibliothécaire numérique" et "informaticien". Par scénario envisagé, le total général est exprimé en équivalent temps plein (ETP). Il été tenu compte du salaire élevé des métiers en pénurie.

Dans le scénario le plus complet, la collecte sélective couvre l'intégralité de 2.350 sites pour les AGR et de 920 sites pour la KBR, plus 1.400 pages d'autres sites pour la KBR. S'ajoute à cette sélection l'échantillon du million de sites web belges moissonnés chacun partiellement. Ce scénario exige 5,76 ETP en personnel, soit un montant de 275.000 euros. La mise à jour de la liste des sites à moissonner intégralement, le processus de collecte (incluant le contrôle qualité de ces sites) et la gestion opérationnelle sont les tâches les plus gourmandes en personnel, la collecte large du million de sites étant entièrement automatisée sans contrôle qualité humain. Le volume de la collecte est estimé à 114 Tb pour la première année, dont 40 Tb pour la collecte large du million de sites et 70 Tb pour la collecte sélective des sites intéressant la KBR. Ce volume considérable s'explique par la nature et la fréquence de la collecte : certains sites de presse se composent de beaucoup de fichiers lourds (vidéos, sons, images fixes) et ils seront moissonnés plusieurs fois par jour en raison de leur caractère ultra-dynamique. En comparaison, la collecte sélective des 2.350 sites intéressant les Archives de l'État n'atteint que 4 Tb parce que ces sites sont moins souvent modifiés et qu'il a été jugé suffisant de les moissonner une fois par an, en complément à la collecte d'autres archives numériques des administrations concernées. Un accroissement annuel de 10% a été intégré dans le calcul. L'infrastructure coûterait environ 205.000 euros. Au total, le scénario le plus complet revient à 480.000 euros annuels.

Dans le scénario intermédiaire, le nombre de sites à moissonner intégralement est fort réduit pour les Archives de l'État mais reste identique pour KBR, tandis que la collecte large est divisée par dix. Le volume total annuel de stockage descend à 75 Tb, le coût global de l'infrastructure baisse à 136.500 euros, soit une réduction due uniquement à l'impact sur la capacité de stockage, car le reste de l'infrastructure douce et dure est le même que dans le scénario complet. Les ressources humaines ne diminuent que d'1 ETP car la collecte large se fait sans intervention humaine. Le coût total de ce scénario est estimé à 360.000 euros.

Dans un troisième scénario dit basique, la collecte sélective est identique à celle du scénario intermédiaire mais aucun site supplémentaire n'est moissonné. Or ceci n'entraîne pas de réduction notable des coûts. Les besoins en personnel restent identiques puisque, la collecte large n'impliquant pas d'intervention humaine, son abandon n'impacte pas les ressources à prévoir. La capacité de stockage est réduite de 4 Tb. Au total, l'économie entre le scénario intermédiaire et le basique se limite à 6.700 euros.

Infrastructure technique

Un des objectifs du projet *PROMISE* était de tester l'archivage du web et de donner accès aux collections. Pour ce faire, il fallait bien sûr également prévoir l'infrastructure technique nécessaire, surtout en ce qui concerne la sélection, la collecte et l'accès.

Un prototype du module de sélection a été développé. Ce module permet d'introduire un URL et de créer manuellement les métadonnées descriptives nécessaires basées sur le modèle OCLC pour métadonnées descriptives dans les archives du web[21]. Ce module, développé par la Haute-École Bruxelles-Brabant, est basé sur Python, Django et PostgreSQL. Le logiciel pourrait être développé davantage dans le futur afin de pouvoir commander l'exploration et le contrôle qualité.

Afin de capturer les sites web, le logiciel *Heritrix* a été utilisé dans le cadre du projet *PROMISE*. Ce "software", appelé également "web crawler", part d'une liste d'URLs sélectionnés[22]. En suivant les liens internes entre les différentes pages web, le "crawler" stocke une copie de tout le contenu des pages web et de leurs métadonnées techniques dans un fichier WARC[23]. Le format de fichier WARC est comparable à un container pour tous les contenus web déterminés et informations contextuelles afférentes. Il s'agit du format de fichier le plus courant pour le contenu du web archivé au niveau international. Le principal avantage de *Heritrix* est la rapidité de l'exploration et le fait que le logiciel, utilisé depuis longtemps, a déjà fait ses preuves. En revanche, ce logiciel a du mal à moissonner des sites web complexes, comme ceux avec de nombreux médias sociaux ou Javascript. Un certain nombre de tests ont dès lors été effectués avec des outils tels que *Browsertrix* et *Brozzler*, spécialement développés pour capturer des sites web au contenu dynamique[24]. Ces outils donnent des résultats de grande qualité, mais utilisent nettement plus de puissance de calcul que *Heritrix*. En outre, ces outils sont récents et donc pas encore stables.

Le prototype du module d'accès est basé sur *WARCLight*, une application spécialement axée sur la découverte d'éléments dans une archive, et sur *pyWB*[25]. *PyWB* est le "replay software" qui permet aux utilisateurs d'interagir avec un site web archivé comme sur le "live web". *PyWB* permet d'afficher une version spécifique avec un certain horodatage d'un site web. Actuellement, le prototype permet uniquement des recherches sur base d'URL, mais l'objectif est de développer, dans le futur, davantage de possibilités de recherche et de filtres.

Conclusion

Le projet de recherche *PROMISE* constitue une étape fondamentale pour l'implémentation d'une politique structurelle d'archivage du web au niveau fédéral. L'étude des meilleures pratiques en Belgique et à l'étranger a permis de tirer des leçons des expériences des institutions qui s'occupent de l'archive du web depuis un certain temps. Ces leçons se trouvaient, entre autres, à la base de la stratégie commune qui a été développée dans le cadre du projet *PROMISE*. Actuellement, l'archive du web belge n'existe que comme prototype, mais le but est de développer une archive du web belge fonctionnelle dans les prochaines années. Les scénarios concrets et les coûts afférents peuvent servir de base à KBR et aux Archives de l'État pour prendre des décisions stratégiques ultérieures. Il faudra évidemment tenir compte des moyens disponibles, tant au niveau financier qu'au niveau des ressources humaines. Les conclusions de ce projet constituent, par ailleurs, une opportunité pour défendre le besoin de disposer d'un budget structurel dédié à la gestion, tant pour les questions de préservation que pour améliorer l'accès pour le citoyen, des archives numériques qu'elles proviennent de la numérisation ou qu'elles soient nativement numériques.

Rolande Depoortere
Sébastien Soyez
Archives Générales du Royaume
Rue du Ruysbroeck 2
1000 Bruxelles
rolande.depoortere@arch.be
sebastien.soyez@arch.be
http://www.arch.be
Friedel Geeraert
Sophie Vandepontseele
KBR
Boulevard de l'Empereur, 4
1000 Bruxelles
friedel.geeraert@kbr.be
sophie.vandepontseele@kbr.be
http://www.kbr.be
Gerald Haesendonck
IDLab UGent
Technologiepark-Zwijnaarde 126
9052 Gent
gerald.haesendonck@UGent.be
https://www.ugent.be/ea/idlab

Mai 2020

Références

Archives Unleashed. *Warclight* [en ligne]. <https://github.com/archivesunleashed/warclight> (consulté le 23 janvier 2020).

International Organisation for Standardisation. *ISO 28500:2017. Information and documentation — WARC file format* [en ligne]. 2017 (consulté le 23 janvier 2020) <https://www.iso.org/standard/68004.html>.

Dooley, Jackie & Bowers, Kate. *Descriptive metadata for web archiving. Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group.* Online Computer Library Center, 2018 (consulté le 23 janvier). Technical report. <https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html>.

Internet Archive. *Heritrix* [en ligne]. <https://github.com/internetarchive/heritrix3/wiki> (consulté le 23 janvier 2020).

Internet Archive. *Brozzler* [en ligne]. <https://github.com/internetarchive/brozzler> (consulté le 23 janvier 2020).

Vlassenroot, Eveline, Chambers, Sally, Di Pretoro, Emmanuel, Geeraert, Friedel, Haesendonck, Gerald, Michel, Alejandra, Mechant, Peter. Web archives as a data resource for digital scholars. *International Journal of Digital Humanities* [en ligne], mars 2019 (consulté le 23 janvier 2020), vol. 1, n°1. <https://link.springer.com/article/10.1007/s42803-019-00007-7>.

Webrecorder. *Browsertrix* [en ligne]. <https://github.com/webrecorder/browsertrix> (consulté le 23 janvier 2020).

Webrecorder. *PyWB* [en ligne]. <https://github.com/webrecorder/pywb> (consulté le 23 janvier 2020).

Notes

1. <https://www.kbr.be/nl/colloquium-saving-the-web-the-promise-of-a-belgian-web-archive/>
2. <https://www.kbr.be/fr/colloque-saving-the-web/>
3. To crawl = collecter, moissonner.
4. Application Programme Interface.
5. Vlassenroot, Eveline et al. Web archives as a data resource for digital scholars. *International Journal of Digital Humanities* [en ligne], mars 2019 (consulté le 23 janvier 2020), vol. 1, n°1. <https://link.springer.com/article/10.1007/s42803-019-00007-7>.
6. OAIS : pour *Open Archive Information System*, un Système Ouvert d'Archivage d'information <https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf> (consulté le 30 mars 2020).
7. OCLC, pour *Online Computer Library Center* <https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html>
8. La description selon 14 métadonnées prend du temps et n'est concrètement possible que pour un nombre limité de sites web. En fonction des choix stratégiques de collecte sélective et/ou large, il est envisagé de limiter ces descriptions à 2 ou 3 métadonnées, obtenues de manière automatisée.
9. MARC 21, cf. <https://www.loc.gov/marc/bibliographic, consulté le 31/03/2020>.
10. DTD-EAD, cf. <https://www.loc.gov/ead/ead2002a.html, consulté le 31/03/2020>.
11. En anglais, on parle de "ccTLD" pour "Country Code Top Level Domain", <https://www.dnsbelgium.be/fr/nouvelles/les-extensions-de-noms-de-domaine-et-le-monde> (consulté le 31 mars 2020)
12. Comme par exemple pour les régionaux .brussels, ou .vlaanderen, et pour des locaux .gent.
13. Cf. les rapports annuels de DNS, <https://www.dnsbelgium.be/fr>,consulté le 31 mars 2020.
14. Cet échantillon pourrait par exemple comprendre de collecter une partie seulement de tous les sites ".be" (3-4 niveaux de profondeur), ou 10% de tous les sites ".be" de manière intégrale.
15. Cette stratégie pourrait évoluer suivant l'adaptation du cadre légal en Belgique, notamment lorsque le web sera intégré comme élément du dépôt légal numérique, ou quand le droit d'auteur tiendra compte de la particularité patrimoniale du web.
16. Cf. page web de l'outil : <http://crawler.archive.org/index.html>,consulté le 31 mars 2020.
17. WARC : Web ARChive format, est un format d'archivage de sites web reconnu par l'ISO : <https://www.iso.org/fr/standard/68004.html>, consulté le 31 mars 2020.
18. Ces paramètres peuvent être la correspondance visuelle, la correspondance interactive, la complétude, la pertinence de la taille et du contenu. Référence : Ayala Brenda Reyes, *A Grounded Theory of Information Quality in Web Archives* (PhD, 2018), cf. <https://digital.library.unt.edu/ark:/67531/metadc1248497>, consulté le 26 mai 2020.

19. LTP : *Long Term Preservation Platform*, cf. <https://www.belspo.be/belspo/organisation/doc/Org/Contrat%20d%27administration%202016-2018%20SPP%20PS>.pdf, consulté le 31 mars 2020.

20. Ce "replay" des sites web sera basé sur la technologie développée par *Internet Archive*, à savoir la *Wayback Machine*. Il peut être mis en œuvre techniquement en installant par ex. *PyWB* (Python Wayback Machine), cf. <https://github.com/webrecorder/pywb, consulté le 31/03/2020>.

21. Dooley, Jackie & Bowers, Kate. *Descriptive metadata for web archiving. Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. Online Computer Library Center, 2018 (consulté le 23 janvier). Technical report. <https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html>.

22. Internet Archive. *Heritrix* [en ligne]. <https://github.com/internetarchive/heritrix3/wiki> (consulté le 23 janvier 2020).

23. International Organisation for Standardisation. *ISO 28500:2017. Information and documentation — WARC file format* [en ligne]. 2017 (consulté le 23 janvier 2020) <https://www.iso.org/standard/68004.html>.

24. Webrecorder. *Browsertrix* [en ligne]. <https://github.com/webrecorder/browsertrix> (consulté le 23 janvier 2020). Internet Archive. *Brozzler* [en ligne]. <https://github.com/internetarchive/brozzler> (consulté le 23 janvier 2020).

25. Archives Unleashed. *Warclight* [en ligne]. <https://github.com/archivesunleashed/warclight> (consulté le 23 janvier 2020). Webrecorder. *PyWB* [en ligne]. <https://github.com/webrecorder/pywb> (consulté le 23 janvier 2020).

TOUR D'HORIZON SUR LES ASPECTS LÉGAUX DE L'ARCHIVAGE DU WEB[1]

Alejandra MICHEL
Chercheuse au CRIDS et Maître de conférences en Droit des médias à l'Université de Namur

■ De nos jours, il est indéniable que le web regorge de contenus susceptibles d'animer le débat public ou d'alimenter le patrimoine culturel. Leur préservation pour les générations futures et leur mise à disposition pour la communauté des chercheurs et le public en général représentent un enjeu sociétal majeur. Les initiatives de préservation et d'archivage de la mémoire du web facilitent en effet l'exercice du droit à l'information des citoyens en leur offrant un outil précieux leur permettant de rechercher et d'accéder à des contenus d'une importance considérable. Toutefois, ces initiatives soulèvent leur lot de questionnements juridiques qu'il convient de résoudre pour développer des stratégies d'archivage du web et de mise à disposition du public des archives web durables et respectueuses des intérêts en présence. La présente contribution propose une vue d'ensemble des principaux aspects légaux impliqués.

■ Het valt niet te ontkennen dat internet tegenwoordig bol staat van inhoud die kan aanzetten tot het publieke debat of het cultureel erfgoed kan voeden. Het behoud daarvan voor toekomstige generaties en de beschikbaarheid voor onderzoekers en het algemene publiek, vormen een enorme maatschappelijke uitdaging. De initiatieven die worden genomen om het geheugen van internet te behouden en te archiveren, maken het voor burgers gemakkelijker hun recht op informatie uit te oefenen, doordat zij daarmee een waardevol instrument in handen krijgen om deze omvangrijke inhoud te doorzoeken. Deze initiatieven werpen echter juridische kwesties op die moeten worden opgelost om strategieën voor de archivering van internet te ontwikkelen en aan het publiek duurzame internetarchieven ter beschikking te kunnen stellen die de bestaande belangen respecteren. Deze bijdrage geeft een overzicht van de belangrijkste juridische aspecten.

À l'ère numérique, le web est sans conteste un instrument propice à l'échange d'idées et à la communication d'information. Ce dernier fournit un outil primordial pour la liberté d'expression et le droit à l'information de milliards d'internautes[2]. Il est en effet aujourd'hui indéniable que le web permet de se connecter à une base de données sans précédent alimentée par une multitude d'informations, d'articles et de publications que l'on retrouve sur les sites Internet, les blogs, les pages personnelles ou encore sur les diverses plateformes en ligne. Sur ces derniers, se côtoient bien évidemment des matériaux des plus légers au plus sérieux. Par ailleurs, d'innombrables contenus d'intérêt général ou pouvant alimenter le patrimoine culturel local, régional, national, européen ou mondial sont uniquement disponibles sur Internet. L'archivage du web permet ainsi leur préservation pour les générations futures ainsi que la création d'une "mémoire numérique" considérable.

Du point de vue légal, l'archivage du web suscite une kyrielle d'interrogations auxquelles il importe de prêter attention pour privilégier la mise en place d'une stratégie durable et respectueuse des intérêts en présence, qu'il s'agisse de ceux des auteurs, des personnes impliquées, des institutions nationales de préservation du patrimoine culturel voire des citoyens en général. Ces enjeux légaux apparaissent et coexistent tout au long du processus de préservation du patrimoine en ligne : de la sélection à la collecte des sites web, en passant par leur archivage jusqu'aux possibilités d'accès aux archives ainsi constituées pour la communauté scientifique et le public en général. L'on peut ainsi lister la répartition des compétences, rôles et responsabilités entre les institutions nationales de préservation du patrimoine culturel qui retrouvent dans leurs missions d'intérêt public la préservation de la mémoire du web, la définition du "web national" pour la sélection des sites web, la liberté d'expression, le droit au respect de la vie privée, la protection des données à caractère personnel, le sort à réserver aux contenus illégaux ou dommageables en ligne, le droit d'auteur, le droit *sui generis* des bases de données ainsi que la valeur probante des archives web. Ce cadre légal exerce une influence considérable sur les politiques et les stratégies d'archivage du web qui sont développées par les institutions nationales de préservation du patrimoine culturel.

Dans la présente contribution, les aspects légaux de deux situations distinctes mais bel et bien liées sont développés : d'une part, ceux de la préservation du web qui consiste à collecter l'information pertinente et à l'archiver et, d'autre part, ceux relatifs à la constitution de collections d'archives d'Internet dans l'optique de les rendre accessibles au citoyen. Après avoir souligné l'enjeu sociétal majeur qu'elles représentent et leurs liens avec le droit à la liberté d'expression consacré par l'article 10 de la Convention européenne des Droits de l'Homme (ci-après "la Convention"), la présente contribution analyse les principaux aspects légaux liés à la préservation de la mémoire du web et à la mise à disposition du public des archives web.

Archiver le web : un enjeu sociétal majeur

Comme nous venons de le relever, une multitude de contenus d'intérêt général ou relevant du patrimoine culturel anime le web. Les préoccupations suscitées

par la préservation de ce riche patrimoine culturel en ligne sont d'autant plus prégnantes que de tels contenus ne sont pas forcément répliqués dans l'univers papier.

Les initiatives d'archivage du web s'inscrivent nécessairement dans une optique prenant en compte la protection des droits de l'Homme et des libertés fondamentales, reconnaissant ainsi leur dimension sociétale majeure. La liberté d'expression est consacrée et protégée par plusieurs instruments juridiques, tant au niveau international, européen que national[3]. Il s'agit d'une liberté fondamentale qui consacre le droit de communiquer des idées, des opinions et des informations mais également le droit de les recevoir ; elle a donc un pôle actif et un pôle passif.

Au niveau du Conseil de l'Europe, la liberté d'expression est protégée par l'article 10 de la Convention. Outre la liberté fondamentale de pouvoir s'exprimer librement à titre individuel, cette disposition protège d'autres applications de la liberté d'expression telles que la liberté de la presse, le droit à l'information, la protection des sources journalistiques, la protection des lanceurs d'alerte ou encore, comme nous allons le voir, la constitution d'archives sur Internet mais également, par extension, la phase préalable de préservation.

La Cour européenne des Droits de l'Homme (ci-après "La Cour") a déjà eu l'occasion d'établir un lien entre les archives du web et la protection accordée au titre de l'article 10 de la Convention. Dans un arrêt *Times Newspapers Limited contre Royaume-Uni*, la Cour a ainsi déclaré que "*[…] l'article 10 garantit non seulement le droit de communiquer des informations mais aussi celui, pour le public, d'en recevoir. Grâce à leur accessibilité ainsi qu'à leur capacité à conserver et à diffuser de grandes quantités de données, les sites Internet contribuent grandement à améliorer l'accès du public à l'actualité et, de manière générale, à faciliter la communication de l'information. La constitution d'archives sur Internet représentant un aspect essentiel du rôle joué par les sites Internet, la Cour considère qu'elle relève du champ d'application de l'article 10*"[4]. La Cour relève par ailleurs que l'offre aux citoyens d'archives sur Internet "*contribue grandement à la préservation et à l'accessibilité de l'actualité et des informations*" et qu'elle constitue également un outil précieux pour l'enseignement et les recherches historiques[5].

Nous pouvons dès lors relever le rôle primordial que joue la constitution d'archives sur Internet – mais également par extension la phase préalable de préservation de l'information – pour le droit à l'information de tout un chacun. L'existence d'initiatives d'archivage du web garantit la mise en œuvre de ce droit fondamental en facilitant la recherche et l'accès à l'information tant pour les chercheurs que pour le grand public.

Toutefois, la liberté d'expression et le droit à l'information ne sont pas absolus. Cela signifie que, même si la constitution d'archives sur Internet et l'archivage du web en général sont susceptibles d'être protégés par l'article 10 de la Convention, des restrictions peuvent être appliquées au cas par cas en raison de la protection qui serait due aux intérêts en présence. Ainsi, le droit au respect de la vie privée pourrait justifier une mesure de restriction à la liberté d'expression profitant aux archives web ou à la constitution d'archives accessibles sur Internet. Il pourrait s'agir, par exemple, de l'obligation, pour un média d'information, d'ajouter une mention dans l'archive en ligne d'un article de presse pour indiquer que l'article papier (identique à l'archive disponible en ligne) a fait l'objet d'une action en diffamation devant les cours et tribunaux, sans pour autant devoir retirer cet article des archives disponibles en ligne[6].

Zoom sur quelques aspects légaux de l'archivage du web

Nous proposons maintenant de nous pencher de façon plus approfondie sur certains aspects légaux liés à l'archivage de la mémoire du web effectué par des institutions nationales de préservation du patrimoine culturel.

La répartition des compétences, rôles et responsabilités entre les institutions nationales de préservation du patrimoine culturel chargées de l'archivage du web

Avant même de pouvoir entamer le processus d'archivage du web, il est primordial de résoudre la question de l'institution nationale de préservation du patrimoine culturel compétente pour effectuer une telle mission d'intérêt public. Dans la majorité des pays tant européens que non européens, cette compétence est le plus souvent dévolue à la Bibliothèque Nationale sur base d'un mécanisme de dépôt légal qui est parfois expressément élargi au web[7]. Les Archives Nationales d'un pays jouent également un rôle pour l'archivage des sites web du secteur public selon les législations nationales relatives aux archives. Dans une minorité des cas, d'autres institutions chargées de l'archivage des productions audiovisuelles mènent également des activités d'archivage du web, comme l'Institut National de l'Audiovisuel en France.

En Belgique, il ressort des arrêtés royaux déterminant leurs missions respectives que tant la Bibliothèque

Royale de Belgique que les Archives générales du Royaume peuvent poursuivre des activités d'archivage du web. Pour la Bibliothèque Royale de Belgique, l'arrêté royal portant constitution en établissement scientifique de la Bibliothèque royale de Belgique liste, à côté de la mission de dépôt légal "*quel que soit le support utilisé*", la mission de collecte et d'inventaire "*des sites web en rapport avec* [ses missions] à l'exception des blogs et des sites internet privés"[8]. Même si l'on peut regretter la formulation de l'exception choisie par le législateur belge, il en ressort que la Bibliothèque Royale de Belgique a reçu pour mission d'intérêt public de procéder à l'archivage du web. Du côté des Archives générales du Royaume, l'arrêté royal déterminant les missions des Archives générales du Royaume et Archives de l'État dans les provinces indique en son article 2 qu'elles sont chargées de "*veiller à la bonne conservation et à la gestion des archives, quel que soit le support, produites et gérées par les autorités publiques, de collecter, conserver et éventuellement détruire les archives publiques*"[9]. L'utilisation des termes "*quel que soit le support*" implique que les Archives générales du Royaume sont aussi chargées de l'archivage des sites web des autorités publiques belges. À côté de l'obligation de conservation des archives publiques, notons qu'elles ont également la possibilité de conserver les archives privées "*qui peuvent intéresser le patrimoine de l'État fédéral ou l'histoire de la Belgique*"[10]. Il est indubitable que de telles archives peuvent provenir des divers sites web ou blogs personnels tenus en ligne[11]. Par ailleurs, l'article 7 de l'arrêté royal impose aux Archives générales du Royaume la mise en œuvre d' "*un plan "archives numériques" qui porte à la fois sur la numérisation des fonds d'archives, ainsi que sur l'acquisition d'archives créées sous forme numérique et la mise à disposition de ceux-ci en ligne et hors ligne*"[12]. À notre estime, il ne fait nul doute qu'une telle formulation accorde aux Archives générales du Royaume la mission d'archivage du web.

L'on constate donc que chez nous, en Belgique, deux institutions fédérales de préservation du patrimoine culturel ont pour mission d'intérêt public la préservation du web, à savoir la Bibliothèque Royale de Belgique et les Archives générales du Royaume. Il convient de relever que leurs missions respectives sont susceptibles de se recouper et qu'un même site web pourrait, selon des objectifs différents, à la fois tomber dans le champ d'action des deux institutions. Dans ce contexte, la collaboration entre institutions est primordiale.

La définition du web national

Après avoir déterminé les institutions fédérales en charge de la préservation de la mémoire du web, encore faut-il pouvoir délimiter le web belge... En effet, pour pouvoir entamer les démarches d'archivage du web, la définition légale du "web national" est déterminante, notamment pour les politiques de sélection des contenus à archiver. Il importe de délimiter ce que l'on entend par "web belge" grâce à l'élaboration de différents critères légaux. Ces critères permettront de considérer qu'un contenu en ligne déterminé ressort du "web national" d'un pays et non d'un autre, même si un même contenu peut parfois tomber dans la définition du web national de plusieurs pays en même temps.

Dans le cadre du projet de recherche *PROMISE*, une analyse détaillée des critères utilisés à l'étranger dans les législations relatives au dépôt légal nous a permis de dresser les grandes tendances et de proposer les critères les plus pertinents et les mieux adaptés à la situation belge. Les critères existants pour délimiter le web national peuvent ainsi être classés en trois catégories principales : les critères basés sur les noms de domaine (qu'il s'agisse de *Generic Top Level Domains* (gTLDs) ou de *Country-Code Top level Domains* (ccTLDs)), les critères liés à un indice de pertinence du contenu en ligne et les critères fondés sur la territorialité ou sur la nationalité.

La première catégorie basée sur les noms de domaine comprend des critères tels que le fait que :

- Le site web soit enregistré sous le nom de domaine national du pays (ccTLDs) ou sous des noms de domaine relatifs à des parties du territoire national, à des territoires d'outre-mer ou à des territoires dépendants du pays ;
- Le site web soit enregistré sous d'autres noms de domaine (gTLDs ou ccTLDs d'autres pays), à condition que :
    - Il soit enregistré auprès de l'organisme national de gestion des noms de domaine ; ou
    - Il soit enregistré par un citoyen ; ou
    - Son contenu soit destiné au "public national"[13].

La seconde catégorie liée à la pertinence du contenu en ligne regroupe des critères tels que le fait que :

- Le contenu du site web ait un lien avec le pays ou soit lié au pays, à ses citoyens ou à ses ressortissants ;
- Le contenu du site web présente un intérêt pour le patrimoine culturel national ;
- Le contenu du site web soit rédigé dans la (les) langue(s) nationale(s) du pays.

La troisième catégorie fondée sur la territorialité ou la nationalité vise des critères tels que le fait que :

- Le site web ait été réalisé sur le territoire national ;
- L'auteur du contenu en ligne ou le titulaire du site web possèdent la nationalité de l'État ;

- Le site web ait été publié sur le territoire national, avec les sous-critères suivants :
    - L'éditeur réside ou a son établissement au sein du territoire national ; ou
    - L'éditeur réside ou a son établissement à l'étranger si le site web a été réalisé au sein du territoire national ;
- L'État a financé ou supporté la production du site web ;
- Le site web a été mis à la disposition du public par une personne dont les activités de création, de production, de publication ou d'édition ont été effectuées au sein du territoire national.

Précisons que certaines législations nationales contiennent une combinaison de plusieurs critères pour délimiter leur "web national" et donc pour définir leur champ d'application. Par exemple, selon la loi danoise sur le dépôt légal, le "contenu danois" regroupe, d'une part, les contenus en ligne publiés sous le nom de domaine national (.dk) et, d'autre part, les contenus en ligne publiés sous d'autres noms de domaine s'ils sont destinés au public danois[14].

À notre estime, plusieurs critères doivent être utilisés en Belgique pour définir le "web belge". Tout d'abord, il ne fait aucun doute que le futur cadre légal belge devra permettre l'archivage des sites web enregistrés à la fois sous le nom de domaine fédéral belge (.be) mais également sous les noms de domaine relatifs aux communautés, aux régions, aux provinces ainsi qu'aux communes (par exemple, .Brussels, .Vlaanderen, etc.). Ensuite, il est également primordial d'envisager la possibilité d'archiver les sites web enregistrés sous les gTLDs (.com, .org, .net, etc.), sous le ccTLD de l'Union européenne (.eu) ou sous les ccTLDs nationaux d'autres pays tels que la France et les Pays-Bas (.fr et .nl)[15]. En effet, de nombreux contenus en ligne intéressants pour le patrimoine culturel belge ou concernant la Belgique, son histoire, ses évènements ou ses citoyens sont publiés sur des sites web enregistrés sous ces autres types de noms de domaine[16]. C'est la raison pour laquelle il est important de combiner le critère de définition du web belge basé sur les noms de domaine avec un critère de pertinence du contenu en ligne pour la Belgique ou pour la société belge en général (qui inclut – sans pour autant s'y limiter – le patrimoine culturel national). Enfin, une remarque demeure en ce qui concerne les critères relatifs au fait que le site web ait été enregistré auprès de l'organisme national de gestion des noms de domaine ou par un citoyen de l'État. Même si ces critères sont totalement pertinents et intéressants, l'obtention de ce type d'information nécessiterait que les institutions nationales de préservation du patrimoine culturel collaborent efficacement avec *DNS Belgium* et avec les organismes gestionnaires d'autres noms de domaine.

À l'inverse, certains critères utilisés dans les législations étrangères paraissent inadaptés à la situation belge ou plus difficiles à mettre en œuvre. D'une part, le fait que le contenu en ligne soit rédigé dans l'une des langues nationales n'est pas un critère pertinent pour la Belgique. En effet, les trois langues nationales de notre pays (allemand, français et néerlandais) correspondent également aux langues nationales d'autres pays (Allemagne, France, Luxembourg, Pays-Bas, Suisse, voire même certains pays d'Afrique du Nord). Un tel critère s'avère indéniablement pertinent pour les pays qui possèdent une langue nationale "unique" (par exemple, le Danemark) mais pas pour d'autres pays. D'autre part, même si le critère basé sur le fait que le contenu en ligne ait été créé, produit, édité ou publié sur le territoire national est intéressant, il n'en reste pas moins délicat à appliquer en pratique, sans compter les possibles imprécisions des résultats obtenus[17].

Le droit d'auteur

À côté de leurs missions respectives d'archivage du web, la Bibliothèque Royale de Belgique et les Archives générales du Royaume doivent également rendre leurs collections accessibles pour le public[18]. Pour ce faire, le respect des règles du droit d'auteur est de mise, ce qui peut drastiquement limiter les possibilités d'un accès élargi du public. L'archivage du web implique en effet souvent des actes protégés par le droit d'auteur, à savoir des actes de reproduction (la collecte d'un site web et sa préservation) et des actes de communication au public (permettre l'accès aux archives web). Ces actes nécessitent en principe l'autorisation préalable du titulaire de droit, sauf possibilités d'exception[19].

Pour l'accès aux archives web (actes de communication au public), le législateur belge, sur base de la directive 2001/29[20], a introduit une exception en faveur des bibliothèques accessibles au public, des établissements d'enseignement et scientifiques, des musées et des archives qui ne poursuivent pas d'avantage commercial ou économique. L'article XI.190, 13° du Code de droit économique prévoit en effet que ces derniers peuvent rendre accessibles au public, à des fins de recherches ou d'études privées, les œuvres qui font partie de leurs collections en salles de lecture (plus précisément, grâce à des terminaux spéciaux accessibles dans leurs locaux)[21,22]. Autrement dit, sans l'obtention préalable de l'autorisation des titulaires de droit, la législation européenne en matière de droit d'auteur permet uniquement aux institutions nationales de préservation du patrimoine culturel de donner accès à leurs collections d'archives du

web sur place à des fins de recherche au moyen de terminaux dédiés[23]. Pour l'heure, il est donc impossible d'étendre l'accès aux collections des institutions nationales de préservation du patrimoine culturel sur base des seules exceptions légales au droit d'auteur. La seule solution disponible actuellement consiste à obtenir l'autorisation des titulaires de droit pour élargir les possibilités d'accès (en ligne, dans d'autres lieux que les salles de lecture, pour d'autres fins que celles liées à la recherche ou aux études privées, etc.). Dans la pratique, il est évident qu'il est complexe pour les institutions nationales de préservation du patrimoine culturel d'obtenir le consentement des titulaires de droit pour tous les types de publications. Toutefois, l'on peut opter pour une limitation de l'élargissement de l'accès à certaines thématiques telles que l'actualité et les articles de presse en concluant des accords avec les représentants des éditeurs de presse.

Pour la collecte et la capture des sites web à archiver (actes de reproduction), l'obtention des autorisations des titulaires de droit n'est pas tâche aisée pour les institutions nationales de préservation du patrimoine culturel[24,25]. Outre l'innombrable quantité de sites web susceptibles de rentrer dans la définition du "web national", il faut bien se rendre compte qu'un site web peut contenir de nombreux éléments protégés par le droit d'auteur (textes, titres, photos, logos, images, illustrations, mises en page, etc.) qui en plus peuvent posséder des titulaires de droit différents[26]... Malgré le principe de l'*opt-in* sur lequel se basent les règles du droit d'auteur au sein de l'Union européenne, certaines bibliothèques nationales ont développé des approches d'*opt-out* pour pallier ces difficultés pratiques. C'est par exemple le cas de la *Koninklijke Bibliotheek* ("KB") aux Pays-Bas qui notifie au titulaire du site web son intention de l'archiver en lui laissant la possibilité de s'y opposer durant un certain délai[27]. Si aucune objection n'a été formulée à l'issue de ce délai, la KB considère que le titulaire du site web a implicitement consenti à l'archivage de son site[28]. Généralement, les bibliothèques nationales qui se basent sur des raisonnements de ce type mettent en place des politiques de retrait très efficaces pour rapidement retirer les sites web des personnes mécontentes de leur base de données.

Certains auteurs appuient également ces raisonnements d'*opt-out* en considérant que, par exemple, la non-utilisation de mesures de protection des sites web contre les robots moissonneurs peut signifier que le titulaire du site ne s'oppose pas à son archivage[29]. Néanmoins, le problème de telles approches, outre le fait qu'elles ne respectent pas le principe de l'autorisation préalable, est que le titulaire d'un site web n'est pas automatiquement le titulaire des droits d'auteur sur les contenus qui alimentent son site. La jurisprudence belge a déjà eu l'occasion de se pencher sur l'absence d'autorisation préalable des titulaires de sites web dans une affaire *Copiepresse contre Google*. Dans cette affaire, Google estimait que les actes de reproduction des sites web des plaignants avaient été implicitement (voire même explicitement) autorisés, étant donné qu'ils n'avaient pas mis en place de mécanismes de Robot TXT pour bloquer le moissonnage de leur site web. Le moteur de recherche faisait ainsi valoir que les plaignants avaient la possibilité d'adapter leurs sites web pour interdire certaines actions effectuées par des robots et que la non-utilisation de Robot TXT signifiait que les éditeurs avaient consenti aux actes de reproduction et de communication au public. À cet égard, la Cour a indiqué que le droit d'auteur n'était pas un droit d'opposition mais un droit d'autorisation préalable, ce qui signifie que l'autorisation des titulaires de droit doit être, avec certitude, obtenue préalablement à l'utilisation envisagée. Par conséquent, selon la jurisprudence belge, l'absence de la mise en place d'un bloqueur de robot ne constitue pas une condition inconditionnelle de référencement[30]. Par ailleurs, dans un arrêt du 5 mai 2011, la Cour d'appel de Bruxelles a souligné que le droit de reproduction est un droit exclusif et absolu et que la protection par le droit d'auteur ne doit pas dépendre de l'adoption préalable de moyens techniques par le titulaire du droit. La Cour a donc, dans cet arrêt, conclu qu'il n'y avait aucune raison d'accepter les hypothèses selon lesquelles "*tout ce qui n'est pas interdit est permis*" ou que "*l'auteur ne peut être protégé s'il n'a pas mis en œuvre un procédé technique*"[31]. À côté de ces réflexions issues de la jurisprudence belge, attirons toutefois l'attention du lecteur sur l'acceptation timide, par la Cour de justice de l'Union européenne, du consentement implicite en matière de numérisation d'œuvres orphelines et d'œuvres indisponibles. Dans l'arrêt *Soulier et Doke contre Premier Ministre*, la Cour de justice souligne les conséquences attachées au caractère exclusif des droits de reproduction et de communication au public en précisant que l'exercice de tels actes par un tiers nécessite, sauf exception, l'autorisation préalable du titulaire de droit[32]. La Cour de justice constate par ailleurs que la directive 2001/29 ne précise pas les modalités d'obtention de ce consentement préalable et en déduit qu'il peut être explicite ou implicite[33]. Toutefois, la Cour de justice estime que l'information effective des auteurs sur les utilisations futures de leurs œuvres et les moyens dont ils disposent pour les interdire sont de la plus haute importance pour que les titulaires de droit puissent se positionner sur ces utilisations et y consentir implicitement[34]. En l'espèce, dans cet arrêt *Soulier et Doke contre Premier Ministre*, la législation nationale ne prévoyait pas de mécanismes garantissant l'information effective et individualisée des auteurs. La Cour de justice a

donc décidé que l'absence d'opposition du titulaire de droit ne pouvait pas constituer une autorisation implicite à l'utilisation de son œuvre[35].

La protection des données à caractère personnel

Pour les institutions nationales de préservation du patrimoine culturel, l'archivage du web implique souvent des "traitements de données à caractère personnel" et donc le respect du Règlement Général sur la Protection des Données (ci-après "RGPD") [36]. Sur les sites web, sont en effet susceptibles de se retrouver toute une série de données à caractère personnel. Pensons notamment aux noms et prénoms de personnes physiques, aux informations de contacts (adresse postale, adresse email, numéro de téléphone) tant personnel que professionnel de personnes physiques, aux données bibliographiques ou biographiques, à la photo d'une personne, aux croyances religieuses ou aux préférences culturelles des personnes physiques (goûts littéraires, cinématographiques, artistiques, etc.), aux revenus économiques d'une personne ou à son niveau de vie, etc. Dans leurs activités d'archivage du web, les institutions nationales de préservation du patrimoine culturel effectuent donc un traitement secondaire de données à caractère personnel prenant la forme d'une conservation à longue, voire très longue, durée (conservation à long terme). Relevons par ailleurs qu'il en est de même de la collecte des sites web, de la consultation, de l'utilisation et de la communication des archives web ou encore de leur effacement ou de leur destruction.

Les institutions nationales de préservation du patrimoine culturel ont dès lors l'obligation de respecter le cadre juridique – tant européen que national – applicable à la protection des données à caractère personnel. Il convient néanmoins de préciser que le RGPD met en place un régime dérogatoire pour les traitements de données effectués à des fins archivistiques dans l'intérêt public. Ce dernier admet ainsi plusieurs dérogations à certains principes clés du traitement, à certaines obligations imposées au responsable du traitement ainsi qu'à certains droits conférés aux personnes concernées. Alors que certaines de ces dérogations sont directement prévues par le RGPD, d'autres sont simplement permises par le texte du Règlement mais doivent être mises en place par le législateur national ou européen.

Pour bien comprendre le champ d'application d'un tel régime dérogatoire, il convient d'expliquer ce qu'il y a lieu d'entendre par traitement de données "à des fins archivistiques dans l'intérêt public". À cet égard, le considérant 158 du RGPD mentionne que "*les autorités publiques ou les organismes publics ou privés qui conservent des archives dans l'intérêt public devraient être des services qui, en vertu du droit de l'Union ou du droit d'un État membre, ont l'obligation légale de collecter, de conserver, d'évaluer, d'organiser, de décrire, de communiquer, de mettre en valeur, de diffuser des archives qui sont à conserver à titre définitif dans l'intérêt public général et d'y donner accès*"[37]. Même s'il ne possède pas de force contraignante, ce considérant nous éclaire sur la portée de cette finalité en listant les cinq conditions cumulatives à rencontrer pour bénéficier du régime dérogatoire[38]. Premièrement, le responsable du traitement à des fins archivistiques dans l'intérêt public doit être une autorité publique, un organisme public ou un organisme privé. Deuxièmement, les traitements ainsi effectués doivent viser la conservation d'archives dans l'intérêt public. Bien que le texte du RGPD ne définisse à aucun moment la notion d'"archives", nous pouvons relever la large définition existant dans la législation belge sur les archives. Constituent ainsi des archives, "*tous les documents qui, quels que soient leur date, leur forme matérielle, leur stade d'élaboration ou leur support, sont destinés, par leur nature, à être conservés par une autorité publique ou par une personne privée, une société ou une association de droit privé, dans la mesure où ces documents ont été reçus ou produits dans l'exercice de ses activités, de ses fonctions ou pour maintenir ses droits et obligations*"[39]. Troisièmement, l'activité de conservation d'archives dans l'intérêt public doit être légalement imposée, que ce soit par le droit national ou européen. Le responsable du traitement est alors soumis à une obligation légale de conservation d'archives dans l'intérêt public. Il s'agit indubitablement du cas des institutions fédérales de préservation du patrimoine culturel ayant pour mandat de conserver des archives dans l'intérêt public comme les Archives générales du Royaume et la Bibliothèque Royale de Belgique dans ses activités de dépôt légal. Quatrièmement, l'obligation légale à laquelle est soumis le responsable du traitement doit porter sur les opérations[40] relatives aux archives conservées dans l'intérêt public suivantes : la collecte, la conservation, l'évaluation, l'organisation, la description, la communication, la mise en valeur, la diffusion et l'accessibilité. À nos yeux, il est inopportun qu'un responsable du traitement soumis à une obligation légale de conservation d'archives dans l'intérêt public doive effectuer l'ensemble de ces actions pour bénéficier du régime dérogatoire. En effet, la finalité archivistique dans l'intérêt public devrait l'emporter sur la réalisation concrète de la totalité de ces opérations, un considérant ne possédant de toute façon aucune force contraignante. Ainsi, par exemple, dans l'exercice de sa mission légale d'archivage du web, la Bibliothèque Royale de Belgique doit bénéficier de ce régime particulier, d'autant plus que cette activité correspond à la philosophie poursuivie par le législateur européen dans la mise en place du régime dérogatoire pour les traitements

de données à des fins archivistiques dans l'intérêt public[41]. Cinquièmement et enfin, cette obligation légale concerne des archives qui doivent être conservées "à titre définitif dans l'intérêt général". Cet"intérêt général" rejoint indubitablement l'idée d'intérêt public de la finalité archivistique : l'on vise des archives qui présentent une certaine valeur culturelle, patrimoniale ou encore historique pour la société.

Les responsables du traitement qui rempliront ces conditions de la finalité archivistique dans l'intérêt public pourront bénéficier d'un régime plus souple, tant au niveau des dispositions prévues par le RGPD que des dispositions prévues par le législateur belge dans la loi du 30 juillet 2018[42]. Sous réserve du respect des conditions prévues, des dérogations sont ainsi possibles pour le principe de limitation des finalités, le principe de limitation de la conservation, l'obligation d'information en cas de collecte indirecte de données, le droit à l'effacement, l'interdiction de principe de traiter certaines catégories particulières de données, le droit d'accès, le droit de rectification, le droit d'opposition, le droit à la limitation, le droit à la portabilité ainsi que pour les obligations de notification en cas de rectification ou d'effacement de données ou en cas de limitation du traitement[43].

Insistons toutefois sur le fait que le législateur belge détermine des obligations spécifiques à respecter pour pouvoir bénéficier du régime dérogatoire à des fins archivistiques dans l'intérêt public. Il en va ainsi de la désignation d'un délégué à la protection des données si les traitements envisagés peuvent engendrer un risque élevé pour les droits et libertés des personnes physiques, de mentions obligations à inscrire dans le registre des activités de traitement, d'informations supplémentaires à fournir à la personne concernée en cas de collecte directe de données, de la conclusion d'une convention en cas de traitement ultérieur de données et de précisions relatives à la diffusion et à la communication des données à caractère personnel traitées à des fins archivistiques dans l'intérêt public[44].

Conclusion

Comme nous l'avons vu, les initiatives de préservation de la mémoire du web permettent l'exercice du droit à la liberté d'expression et, plus précisément, du droit à l'information de tout un chacun en rendant plus aisées les démarches de recherche et d'accès à l'information. Elles bénéficient, à ce titre, d'une protection par l'article 10 de la Convention européenne des Droits de l'Homme. Toutefois, ces activités impliquent moult aspects légaux tout au long de leur processus (sélection, collecte, préservation, archivage et mise à disposition du public) dont le respect est primordial pour le développement et la mise en œuvre, par les institutions nationales de préservation du patrimoine culturel, de stratégies et de politiques d'archivage du web durables.

Sur le plan juridique, la préservation de la mémoire du web est encore loin d'avoir révélé tous ses secrets. À côté de l'archivage des sites Internet, les réseaux sociaux gérés par les plateformes en ligne regorgent de contenus profitables au patrimoine culturel. Leur préservation représente également un enjeu sociétal majeur auquel les institutions nationales de préservation du patrimoine culturel doivent se montrer attentives, tout en ayant à l'esprit les interrogations juridiques qu'elle soulève…

Alejandra Michel
Université de Namur
Rue de Bruxelles, 61
5000 Namur
alejandra.michel@unamur.be

Avril 2020

Notes

1. La présente contribution a pour vocation de valoriser une partie des résultats de recherche relatifs aux aspects légaux de l'archivage du web obtenus dans le cadre du projet de recherche *PROMISE* mené entre juillet 2017 et décembre 2019. Le projet de recherche *PROMISE* a été financé par BELSPO dans le cadre du programme BRAIN-be.
2. La Cour européenne des Droits de l'Homme souligne d'ailleurs dans sa jurisprudence la primordialité d'Internet pour l'exercice des droits conférés par l'article 10 de la Convention européenne des Droits de l'Homme. Dans un arrêt *Ahmet Yildirim contre Turquie*, elle a ainsi déclaré que " *l'internet est aujourd'hui devenu l'un des principaux moyens d'exercice par les individus de leur droit à la liberté d'expression et d'information : on y trouve des outils essentiels de participation aux activités et débats relatifs à des questions politiques ou d'intérêt public*". Voy. Cour eur. D.H. (2e section), arrêt *Ahmet Yildirim c. Turquie*, 18 décembre 2012, req. n° 3111/10, §54. Dans le même ordre d'idées, la Cour a également indiqué dans un *arrêt Times Newspapers Limited contre Royaume-Uni* que "*grâce à leur accessibilité ainsi qu'à leur capacité à conserver et à diffuser de grandes quantités de données, les sites Internet contribuent grandement à améliorer l'accès du public*

à l'actualité et, de manière générale, à faciliter la communication de l'information". Voy. Cour eur. D.H. (4e section), arrêt *Times Newspapers Limited c. Royaume-Uni*, 10 mars 2009, req. n°s 3002/03 et 23676/03, §27.

3. Au niveau international, la liberté d'expression est consacrée par l'article 19 du Pacte International relatif aux Droits Civils et Politiques ("PIDCP") ; au niveau du Conseil de l'Europe, elle est protégée par l'article 10 de la Convention de sauvegarde des Droits de l'Homme et des Libertés fondamentales ("CEDH") ; au niveau de l'Union européenne, elle est consacrée par l'article 11 de la Charte des droits fondamentaux de l'Union européenne et ; au niveau interne belge, la Constitution protège en son article 19 la liberté d'expression.

4. Voy. Cour eur. D.H. (4e section), arrêt *Times Newspapers Limited c. Royaume-Uni*, 10 mars 2009, req. n°s 3002/03 et 23676/03, §27. Précisons que dans cette affaire il était question de la constitution d'archives web relatives aux articles de presses d'un média. La Cour a par ailleurs précisé au §45 de l'arrêt qu'à côté de sa fonction principale de chien de garde de la démocratie, la presse exerce une fonction accessoire de par la constitution d'archives "à partir d'informations déjà publiées et en les mettant à la disposition du public [...]". Même si dans la présente affaire la Cour a conclu, au vu des circonstances de l'espèce, à la non violation de l'article 10 de la Convention, cet arrêt lui a donné l'occasion de prononcer des attendus intéressants concernant la constitution d'archives web. Ces attendus de principe ont par la suite été rappelés dans un arrêt *Wegrzynowski et Smolczewski contre Pologne* de 2013 dans lequel la Cour a eu une nouvelle fois l'occasion de se prononcer, entre autres, sur les archives web relatives à des articles de presse d'un média. Précisons que cette affaire a cette fois été portée devant la Cour sous l'angle d'une prétendue violation du droit au respect de la vie privée et que la Cour a conclu à la non violation de l'article 8 de la Convention. Voy. Cour eur. D.H. (4e sect.), arrêt *Wegrzynowski et Smolczewski c. Pologne*, 16 juillet 2013, req. n° 33846/07, §59. Sur ces attendus de principes, voy. plus récemment Cour eur. D.H. (5e sect.), arrêt *M.L. et W.W. c. Allemagne*, 28 juin 2018, req. n°s 60798/10 et 65599/10, §90 et §102.

5. Cour eur. D.H. (4e section), arrêt *Times Newspapers Limited c. Royaume-Uni*, 10 mars 2009, req. n°s 3002/03 et 23676/03, §45 ; Cour eur. D.H. (4e sect.), arrêt *Wegrzynowski et Smolczewski c. Pologne*, 16 juillet 2013, req. n° 33846/07, §59. Voy. aussi plus récemment Cour eur. D.H. (5e sect.), arrêt *M.L. et W.W. c. Allemagne*, 28 juin 2018, req. n°s 60798/10 et 65599/10, §90.

6. Sur ce type de mesure, voy. Cour eur. D.H. (4e section), arrêt *Times Newspapers Limited c. Royaume-Uni*, 10 mars 2009, req. n°s 3002/03 et 23676/03. Voy. également, Cour eur. D.H. (4e sect.), arrêt *Wegrzynowski et Smolczewski c. Pologne*, 16 juillet 2013, req. n° 33846/07, §59.

7. Parmi les pays qui ont élargi leur législation sur le dépôt légal pour également couvrir de manière explicite le web ou qui possèdent des législations sur le dépôt légal technologiquement neutre, l'on peut notamment citer la France, le G.D. de Luxembourg, le Royaume-Uni, le Danemark, le Canada ou encore la Suisse.

8. Arrêté royal du 19 juin 1837 portant constitution en établissement scientifique de la Bibliothèque royale de Belgique, *M.B.*, 8 juillet 1837, art. 3.

9. Arrêté royal du 3 décembre 2009 déterminant les missions des Archives générales du Royaume et Archives de l'Etat dans les provinces, *M.B.*, 15 décembre 2009, art. 2, §1.

10. Arrêté royal du 3 décembre 2009 déterminant les missions des Archives générales du Royaume et Archives de l'Etat dans les provinces, *M.B.*, 15 décembre 2009, art. 4 : "*Les Archives de l'Etat peuvent acquérir, à titre onéreux ou gracieux, et conserver des archives privées qui peuvent intéresser le patrimoine de l'Etat fédéral ou l'histoire de la Belgique, en provenance de personnes, physiques ou morales, non soumises à la loi sur les archives. L'Archiviste général du Royaume détermine les conditions de transfert de ces archives*".

11. Pensons par exemple au blog qui serait tenu par un historien sur les guerres et conflits du 20ème siècle (première et seconde guerres mondiales).

12. Arrêté royal du 3 décembre 2009 déterminant les missions des Archives générales du Royaume et Archives de l'Etat dans les provinces, *M.B.*, 15 décembre 2009, art. 7.

13. Plusieurs sous-critères permettent d'identifier ce que l'on entend par "public national". Il en va ainsi du fait que le contenu du site web soit rédigé dans la (ou les) langue(s) nationale(s) du pays ; que le contenu du site web soit lié au pays ou ait un impact sur le pays ; que l'auteur du contenu du site web soit un citoyen de l'État ; ou encore que la personne qui a enregistré le nom de domaine soit domiciliée dans l'État.

14. Danish Act n° 1439 on Legal Deposit of Published Material of 22nd December 2004, §8 (2).

15. Il apparait en effet que de nombreux citoyens belges choisissent d'enregistrer leur site web sous le ccTLD de la France (.fr) ou sous le ccTLD des Pays-Bas (.nl).

16. Cela permet ainsi de couvrir, dans l'archivage du web, tous les sites web liés à la Belgique dans tous ses aspects et dimensions (historique, politique, culturelle, sociale, etc.) mais également ceux dont la production a été financée ou supportée par l'État belge.

17. En pratique, des solutions ont été mises en place pour tenter de résoudre ces difficultés d'application. Une première solution technique existante consiste à faire appel à des techniques de géolocalisation (*Geo-IP localisation*) pour obtenir des informations sur les serveurs qui sont physiquement localisés sur le territoire national. Une seconde solution, cette fois plutôt pragmatique, s'intéresse aux numéros de téléphones et aux adresses postales mentionnés sur les sites Internet.

18. Voy. Arrêté royal du 19 juin 1837 portant constitution en établissement scientifique de la Bibliothèque royale de Belgique, *M.B.*, 8 juillet 1837, art. 3 ; Arrêté royal du 3 décembre 2009 déterminant les missions des Archives générales du Royaume et Archives de l'Etat dans les provinces, *M.B.*, 15 décembre 2009, art. 6 et 7.

19. Ce principe ne s'applique bien évidemment qu'en présence de contenus protégés par le droit d'auteur. Lorsqu'il s'agit d'un contenu non protégé par le droit d'auteur ou qui est tombé dans le domaine public, aucune autorisation préalable n'est nécessaire. Par ailleurs, des licences d'autorisation peuvent également permettre la reproduction ou la communication au public de contenus protégés sous réserve du respect de certaines conditions déterminées dans la licence.

20. Directive 2001/29/CE du Parlement européen et du Conseil du 22 mai 2001 sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information, *J.O.U.E.*, 22 juin 2001, L 167/10, art. 5, para. 3, point n : "*Les États membres ont la faculté de prévoir des exceptions ou limitations aux droits prévus aux articles 2 et 3 dans les cas suivants : lorsqu'il s'agit de l'utilisation, par communication ou mise à disposition, à des fins de recherches ou d'études privées, au moyen de terminaux spécialisés, à des particuliers dans les locaux des établissements visés au paragraphe 2, point c), [des bibliothèques accessibles au public, des établissements d'enseignement ou des musées ou par des archives, qui ne recherchent aucun avantage commercial ou économique direct ou indirect] d'œuvres et autres objets protégés faisant partie de leur collection qui ne sont pas soumis à des conditions en matière d'achat ou de licence*".

21. CDE, art. XI.190, 13°.

22. À côté de l'exception formulée à l'article XI.190, 13° du CDE, il existe également une autre exception à l'article XI.191/1, para. 1, 4° du CDE : "*Lorsque l'œuvre a [été] explicitement divulguée, et sans préjudice de l'application éventuelle des articles XI.189, § 3 et XI.190, 2°, 2/1°, 10°, 12°, 13°, 15°, 16° et 17°, l'auteur ne peut interdire : 4° la communication au public d'œuvres à des fins d'illustration de l'enseignement ou de recherche scientifique, par des établissements reconnus ou organisés officiellement à cette fin par les pouvoirs publics et pour autant que cette communication soit justifiée par le but non lucratif poursuivi, se situe dans le cadre des activités normales de l'établissement, soit sécurisée par des mesures appropriées et ne porte pas préjudice à l'exploitation normale de l'œuvre*". Il faut toutefois rester attentif au principe de stricte interprétation des exceptions en droit d'auteur et veiller à ne pas conférer à cette exception une portée qui dépasserait l'intention du législateur. Cette exception ne s'applique en effet qu'à des fins d'illustration de l'enseignement ou d'illustration de la recherche scientifique. Même si des discussions ont existé sur la formulation de cette exception, l'exposé des motifs de la loi du 22 décembre 2016 modifiant certaines dispositions du livre XI du Code de droit économique semble confirmer la nature distributive du terme "illustration" qui s'appliquerait tant à l'enseignement qu'à la recherche scientifique. En effet, dans l›exposé des motifs, le législateur utilise à plusieurs reprises les termes "à des fins d›illustration de l›enseignement et de la recherche scientifique" ou "à des fins d›illustration de l›enseignement ou de la recherche scientifique", ce qui démontre le caractère distributif de la condition d'"illustration". Contrairement à l'exception prévue à l'article XI.190,13° du CDE qui permet la mise à disposition des œuvres qui font partie des collections à des fins de recherche dans leur intégralité en salles de lecture, cette seconde exception permet uniquement une mise à disposition de l'œuvre limitée à ce qui est nécessaire à des fins d'illustration.

23. Dans le cadre du projet *PROMISE*, nos recherches ont permis de démontrer que la notion de "terminaux dédiés" variait d'un pays à l'autre, modifiant donc les conditions d'accès. Alors que certains législateurs sont plus souples en étendant l'accès à une liste légalement établie de bibliothèques partenaires ou en permettant aux chercheurs accrédités d'utiliser leur propre ordinateur portable pour se connecter aux archives web en salles de lecture (France), d'autres législateurs sont plus restrictifs en ne permettant qu'à un seul utilisateur à la fois d'accéder à une œuvre déterminée sur les ordinateurs disponibles en salles de lecture (Royaume-Uni).

24. L'article XI.190, 12° du Code de droit économique prévoit une exception à l'autorisation préalable du titulaire de droit pour les actes de reproduction justifiés par un but de préservation du patrimoine culturel et scientifique. Pour qu'une telle exception ait vocation à s'appliquer, quatre conditions cumulatives doivent être respectées. Premièrement, la reproduction doit être limitée à un nombre restreint d'exemplaires déterminé en fonction de et justifié par l'objectif de préservation du patrimoine culturel et scientifique. Deuxièmement, la reproduction doit être faite par des bibliothèques accessibles au public, des musées ou des archives qui ne recherchent aucun avantage économique ou commercial. Troisièmement, la reproduction ne peut pas porter atteinte à l'exploitation normale de l'œuvre. Quatrièmement, la reproduction ne peut pas porter un préjudice injustifié aux intérêts légitimes de l'auteur. L'on pourrait penser qu'une telle exception permettrait aux institutions nationales de préservation du patrimoine culturel effectuant l'archivage du web de ne pas avoir à obtenir l'autorisation préalable des titulaires de droit pour reproduire un site web à des fins de préservation du patrimoine culturel. Toutefois, la portée de cette exception est extrêmement débattue : pour certains cette exception se limite à la numérisation de documents papiers et non au moissonnage du web et pour d'autres cette exception pourrait viser l'archivage du web également. Cette exception a récemment été modifiée par la directive 2019/790 du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique. Contrairement à la directive 2001/29, cette nouvelle directive rend, en son article 6, cette exception obligatoire en faveur des institutions de gestion du patrimoine culturel pour conservation (et non plus "préservation") du patrimoine culturel. Nous pouvons relever plusieurs différences entre la directive 2001/29/CE (et la transposition qui en a été faite par le législateur belge) et la nouvelle directive de 2019. Premièrement, auparavant on parlait d'" actes spécifiques " alors que dans la nouvelle directive on vise les copies " dans la mesure nécessaire à cette conservation ". Deuxièmement, la nouvelle directive précise, contrairement à l'ancienne situation, que les copies peuvent se faire " *sous quelque forme ou sur quelque support que ce soit* ",

ce qui autorise les copies digitales par une formulation technologiquement neutre. Troisièmement, il n'y a plus la limite du nombre restreint de copies dans la nouvelle directive. Quatrièmement et enfin, dans la nouvelle directive, l'on parle de copies " *de toute œuvre ou tout autre objet protégé qui se trouve à titre permanent dans leurs* [des institutions du patrimoine culturel] *collections* ". Ce dernier point pose inévitablement la question de savoir ce qu'il faut entendre par " collection permanente ". La question de savoir si les institutions nationales de préservation du patrimoine culturel peuvent se baser sur cette nouvelle exception pour collecter les sites web à des fins de moissonnage du web reste incertaine... Peut-on en effet considérer que ces sites web se trouvent dans la collection permanente de ces institutions alors que justement le but du moissonnage est de les collecter pour qu'ils intègrent la collection des archives du web... ? Le considérant 27 de la directive de 2019 précise d'ailleurs que "*les actes de reproduction effectués par les institutions du patrimoine culturel à des fins autres que la conservation des œuvres ou autres objets protégés de leurs collections permanentes devraient rester soumis à l'autorisation des titulaires de droits*"... L'élargissement permis par la nouvelle formulation empruntée dans la Directive de 2019 ne semble donc pas pertinent par rapport au moissonnage. Voy. Directive (UE) 2019/790 du Parlement européen et du Conseil du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique et modifiant les directives 96/9/CE et 2001/29/CE, *J.O.U.E.*, 17 mai 2019, L 130/92, art. 6 et cons. 27.

25. Certains législateurs se sont montrés attentifs à ces préoccupations et ont introduit une nouvelle exception au droit d'auteur pour faciliter le travail des institutions nationales de préservation du patrimoine culturel. Par exemple, en France, une exception a été introduite pour couvrir les actes de reproduction et de communication au public liés au dépôt légal du web. Ainsi, le Code du patrimoine français prévoit en son article L132-4 que "*L'auteur ne peut interdire aux organismes dépositaires, pour l'application du présent titre : 1° La consultation de l'œuvre sur place par des chercheurs dûment accrédités par chaque organisme dépositaire sur des postes individuels de consultation dont l'usage est exclusivement réservé à ces chercheurs ; 2° La reproduction d'une œuvre, sur tout support et par tout procédé, lorsque cette reproduction est nécessaire à la collecte, à la conservation et à la consultation sur place dans les conditions prévues au 1°*". Cette exception existe également pour les artistes-interprètes, les producteurs de phonogrammes ou de vidéogrammes, les entreprises de communication audiovisuelle et les producteurs de bases de données. Autre exemple, au Royaume-Uni, avec la Section 44A du *Copyright, Designs and Patents Act* qui prévoit que, sous réserve de certaines conditions, il n'y a pas de violation du droit d'auteur lorsqu'une *deposit libary* (en ce compris ses employés) reproduit une œuvre d'Internet.

26. Attirons l'attention sur le fait que le titulaire d'un site web n'est pas forcément la personne qui détient les droits d'auteur sur l'ensemble du contenu l'alimentant...

27. KB Nederland, "Legal issues", disponible sur <https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving/legal-issues>, consulté le 24 avril 2020.

28. *Ibidem*.

29. Beunen A. & Schiphof, T. (2006). *Legal aspects of web archiving from a Dutch perspective* (report commissioned by the National Library in The Hague), p. 17.

30. Civ. Bruxelles (prés). 13 février 2007, *R.D.C.*, 2007/4, p. 390.

31. Bruxelles (9e ch.), 5 mai 2011, *D.I.*, 2011, p. 280.

32. C.J.U.E. (3e ch.), Soulier & Doke c. Premier Ministre, 16 novembre 2016, aff. C-301/15, point 33. Sur le détail complet de cet arrêt, voy. A. DELFORGE, N. GILLARD, M. KNOCKAERT, M. LOGNOUL, B. MICHAUX, A. MICHEL, Z. ROSIC et T. TOMBAL, "Chronique de jurisprudence : Droits intellectuels", *R.D.T.I.*, n° 68-69/2017, pp. 60 à 62.

33. C.J.U.E. (3e ch.), *Soulier & Doke c. Premier Ministre*, 16 novembre 2016, aff. C-301/15, point 35.

34. C.J.U.E. (3e ch.), *Soulier & Doke c. Premier Ministre*, 16 novembre 2016, aff. C-301/15, points 38 à 40.

35. C.J.U.E. (3e ch.), *Soulier & Doke c. Premier Ministre*, 16 novembre 2016, aff. C-301/15, point 43.

36. La "donnée à caractère personnel" est définie comme "toute information se rapportant à une personne physique identifiée ou identifiable". Le "traitement" est défini comme "toute opération ou tout ensemble d'opérations effectuées ou non à l'aide de procédés automatisés et appliquées à des données ou des ensembles de données à caractère personnel". Voy. Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données), *J.O.U.E.*, L 119 du 4 mai 2016, p. 1, art. 4, 1° et 2°.

37. RGPD, cons. n° 158.

38. Sur ces cinq conditions, voy. O. VANRECK, "Impacts du Règlement général sur la protection des données dans le domaine de l'archivage", in *Le Règlement général sur la protection des données (RGPD/GDPR) – Analyse approfondie*, C. de Terwangne et K. Rosier (coord.), Bruxelles, Larcier, 2018, pp. 851 à 852.

39. Arrêté royal du 18 août 2010 portant exécution des articles 1er, 5 et 6bis de la loi du 24 juin 1955 relative aux archives, *M.B.*, 23 septembre 2010, art. 1er, al. 2 ; Arrêté royal du 18 août 2010 portant exécution des articles 5 et 6 de la loi du 24 juin 1955 relative aux archives, *M.B.*, 23 septembre 2010, art. 1er, al. 2.

40. Comme le relève Odile Vanreck, l'utilisation de la conjonction "et" induit un caractère cumulatif de l'ensemble des actions composant l'obligation légale. Voy. O. VANRECK, *op. cit*. (voy. note 38), p. 852.

41. Sur ce point, voy. également A. MICHEL, "Les traitements de données à des fins archivistiques dans l'intérêt public", *DPO News*, 2019, n° 5, pp. 6 à 9.
42. Loi du 30 juillet 2018 relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel, *M.B.*, 5 septembre 2018.
43. Pour une analyse des spécificités de l'archivage au regard du RGPD, nous renvoyons le lecteur à la contribution de O. VANRECK, *op. cit*. (voy. note 38). Pour une vue d'ensemble des dérogations permises aux dispositions du RGPD et de la loi belge du 30 juillet et des conditions prévues pour leur application, voy. A. MICHEL, *op. cit*. (voy. note 41), pp. 6 à 9.
44. Voy. Loi du 30 juillet 2018 précitée, art. 187 à 208. Pour un détail de ces obligations spécifiques et des exceptions possibles, voy. A. MICHEL, *op. cit.* (voy. note 41), pp. 8 et 9.

EXPLORING THE 20-YEAR EVOLUTION OF A RESEARCH COMMUNITY: WEB-ARCHIVES AS ESSENTIAL SOURCES FOR HISTORICAL RESEARCH

Niels BRÜGGER[1]
Professor of Media Studies at the School of Communication and Culture, Head of the Centre for Internet Studies and of NetLab, Aarhus University, Denmark.
Valérie SCHAFER[2]
Professor of Contemporary European History at C²DH (Luxembourg Centre for Contemporary and Digital History), University of Luxembourg, Luxembourg.

Interview edited by:
Friedel GEERAERT, KBR,
Web-archiving scientific assistant
Nadège ISBERGUE, KBR,
Periodicals manager
Sally CHAMBERS, Ghent Centre for Digital Humanities,
Digital Humanities Research Coordinator

■ The *PROMISE* project could not be conceived without tackling the question of access to Belgian web archives. Indeed, if it is important to save them, it is because of their importance as sources of information for research. Therefore, during the conference "Saving the Web: the *Promise* of a Belgian Web Archive", the access and the use of this specific type of archives were widely discussed. Two of the speakers at this conference, Valérie Schafer and Niels Brügger, agreed to share their experiences regarding the use of web archives, some of which, go back to the early 2000s.

■ Het *PROMISE*-project kon niet worden bedacht zonder de kwestie van toegang tot Belgische webarchieven aan te pakken. Het belangrijk is om de Web te bewaren, omdat het belangrijk is als informatiebron voor onderzoek. Tijdens de conferentie "Saving the Web: the *Promise* of a Belgian Web Archive" werd daarom veel aandacht besteed aan de toegang en het gebruik van deze specifieke archieven. Twee van de sprekers op deze conferentie, Valérie Schafer en Niels Brügger, waren bereid om hun ervaringen te delen met betrekking tot het gebruik van webarchieven, die soms teruggaan tot begin jaren 2000.

■ Le projet *PROMISE* ne pouvait se concevoir sans aborder la question de l'accès aux archives du web belge. En effet, s'il est important de les sauvegarder, c'est en raison de leur importance comme sources d'information pour la recherche. Par conséquent, lors du colloque "Saving the Web: the *Promise* of a Belgian Web Archive", l'accès et l'utilisation de ces archives spécifiques ont été largement abordés. Deux des intervenants lors de ce colloque, Valérie Schafer et Niels Brügger, ont accepté de partager leurs expériences respectives quant à l'utilisation d'archives web, celles-ci remontant parfois au début des années 2000.

Introduction

The history of the World Wide Web already spans more than a quarter of a century[3]. The Internet Archive, as well as several National Libraries and Archives, have already been archiving the web for over 20 years[4]. Yet, the use of the archived web as an object of research remains at the fringes of (digital) humanities research[5]. Although many researchers in the humanities and social sciences still need to explore the potential of these archives, the research community around web-archives has been steadily evolving alongside these developments. Particular examples include the Big UK Domain Data for the Arts and Humanities (BUDDAH) project[6], the research being undertaken by members of RESAW, the Research Infrastructure for the Study of Archived Web Materials[7] and most recently WARCNet: Web ARChive studies NETwork researching web domains and events[8].

Within the context of the Belgium web-archiving project *PROMISE*[9], piloting access to the Belgian web archive for scientific research has been considered a key task from the outset. During the project, a survey with almost 400 respondents was conducted which aimed to understand "What are the requirements and needs of potential users of web archives?" and "How do they, or do they want to, access, use and consult web archives?". The results from this survey helped the project team to understand to what extent web archives can be seen as a data resource for digital scholars[10]. A particular challenge in this area is providing access to archived web-resources. Many web-archives remain solely accessible through dedicated computers inside (national) libraries due to legal restrictions. The article "Tour d'horizon sur les aspects légaux de l'archivage du web" by Alejandra Michel in this volume explores this issue in depth.

To conclude the *PROMISE* project, on 18 October 2019, KBR organised the colloquium "Saving the Web: the *Promise* of a Belgian Web Archive"[11]. This colloquium was the opportunity for the researchers of *PROMISE*'s project to present their work regarding web archiving. In addition to Neil's Brügger's keynote: "National web archives: the land of *promise* for researchers"[12], which is included as an article in this volume, the final session of the colloquium was dedicated to research use of web-archives. Presentations during this session included Eveline Vlassenroot: "User Requirements for a Web Archive"[13], based on the experiences of the *PROMISE* project; "Coordination of web archiving in the Netherlands"[14] by Jesse de Vos from the Netherlands Institute for Sound and Vision. Furthermore, Patricia Blanco, a Masters Student in Digital Humanities at KU Leuven and an intern on the *PROMISE* project described her experiences of using web archives for research in "Saving the Belgian Web: Web archiving practices, research opportunities and limitations"[15], which is also elaborated as an article in this special issue.

This event provided the opportunity to welcome international names in the field: Niels Brügger[16], Professor of Media Studies at the School of Communication and Culture, Aarhus University, Head of the Centre for Internet Studies and of NetLab and Valérie Schafer[17], Professor of Contemporary European History at C^{2}DH (Luxembourg Centre for Contemporary and Digital History), University of Luxembourg. As a result, the interview that follows, provided the occasion not only to delve into a deeper discussion around the key themes, ideas and questions evoked during the *PROMISE* colloquium, but also to reflect on how these issues can help pave the way for future research in this area, both in Belgium and beyond.

An interview with Niels Brügger and Valérie Schafer

How did you become involved in studying the archived web?

Niels Brügger (**NB**): I have an academic background in French Language and Culture, and the History of Ideas, but in 1997 I moved from the French Department to Media Studies. At the time only one of my new colleagues studied the internet, but without having a clear historical perspective. Thus, there was an uncharted territory to go into for someone like me who was interested in the web and its very short history. But I soon found out that my object of study disappeared for my very eyes — the online web was changed or deleted in a way that was not familiar to anyone within media studies. So, my first step as a wannabe web historian was to ensure that I had a stable object to study. I therefore started to tinker with web archiving myself, but quickly realised that a professional organisation was needed here to make this happen at scale. Around the same time — in the Autumn of 2000 — I co-founded the Centre for Internet Studies at Aarhus University, and in our mission statement we mentioned that one of our aims was to have a national web archive established, without having any idea about how feasible that was and if that could be done at all. We sent out a press release about the new centre and it was well received by staff at the two national Danish libraries: the State Library in Aarhus and the Royal Library in Copenhagen, and together with them we got funding for a one-year pilot project aiming at investigating how a national Danish web archive could be established. To make a long story very short: the final report[18] from the project published in 2002 served as the basis for the discussion in the Ministry of Culture which later led to the revision of the Legal Deposit law passed in December 2004 where 'computer networks' were included, and in June 2005 the Danish web archive Netarkivet became a reality.

Valérie Schafer (**VS**): Thanks to Niels Brügger! In 2011 I edited an issue of the French journal Le Temps des Médias dedicated to the history of the Internet[19] and Niels proposed an article to us dedicated to Web archives. I started to extend my research to the history of the Web about a year later, after work more devoted to the history of the Internet itself, its infrastructure, protocols, etc. My first contact with the Wayback Machine was a revelation! I was immediately fascinated by Web archives and it hasn't stopped since. It was also the beginning of stimulating collaborations with Niels Brügger and colleagues who were working on these issues like him and whom he was able to bring together. The date of 2011 may seem very late as the Wayback Machine has been available online since 2001, but I can assure you that researchers who were already interested in Web archives such as Louise Merzeau, Fabienne Greffet, Dana Diminescu and a few others were still very rare in France at the time. Archivists and librarians were ahead of the game in this field.

What research projects related to the archived web are you currently involved in?

NB: I am currently heading two large projects. The first started some 3-4 years ago, and it aims at mapping the entire Danish web domain .dk and its development from 2005-2015. This is only possible since we have access to the copies of the Danish domain in Netarkivet where the entire web domain is archived four times per year. In addition, we have had access to Netarkivet's High Performance Computer 'The Cultural Heritage Cluster'[20] which has made it possible to run large scale quantitative analyses.

Among other things we have investigated the number of specific file types, and we now know that the ratio between text and image files only evolves slowly: we get more pictures, but we also get more text. However, over 10 years, the number of pictures grows a bit more than the amount of text. Also, we investigated the billions of hyperlinks on the Danish web and could map how many links to social media were on the Danish web in the period (these results can be found in a recent article "Big data experiments with the archived Web"[21]). The second project is a network, called WARCnet, 'Web ARChive studies network researching web domains and events', the research aim of the project is to investigate how national web domains and events on the web have developed over time. The project includes researchers and web archives from Denmark, France, Luxembourg, the Netherlands, Germany, the UK, and Belgium. The events we will focus on could be terrorist attacks, sport events, or elections, and definitely the COVID-19 pandemic will be studied (read more about WARCnet on the project website[22]).

VS: From 2014 to 2018 I was involved in two research projects that came to an end with my transfer from the CNRS[23] (France) to the University of Luxembourg. One of these projects was dedicated to the Web of the 1990s in France (Web90[24]) and the other to the digital traces and born-digital heritage of the 2015 attacks (ASAP) in partnership with the Bibliothèque nationale de France (BnF) and the Institut national de l'audiovisuel (Ina). This second project had a strong focus on digital social networks. Since then, I have dedicated myself in particular to developing in the field of teaching, at master's level. This included a training course on Web archives, in the framework of a week-long winter school involving Web researchers and archivists. I have also had the pleasure of publishing the book; Qu'est-ce qu'une archive du web?[25] with three colleagues from the Web90 project, mentioned above. Additionally, I broadened my reflection on the sustainability of Digital studies and Digital Humanities, by crossing several facets of digital heritage, whether digitised or born-digital. Finally, I'm involved in the WARCnet project, launched in 2020, as previously mentioned by Niels.

Why is archiving the (national) web indispensable for research and for society at large?

NB: A society that does not preserve sources to document its own history is a poor society. History continues to play a pivotal role in our understanding of the present and for our ways of anticipating the future, and if we want to base our history writing on scholarly ground. We need to have as many sources as possible. Today this also includes the online web, just as we have previously collected handwritten documents, print media, film, radio and television.

VS: Web archiving is included in the legal deposit framework in several European countries where Web contents are considered as publications. While newspapers are kept in national libraries, how can we imagine not keeping online news as well? The same applies to the audio-visual sector. But more broadly, the Web is today a medium of expression and communication for political, administrative, academic, economic and social life, for our personal and professional lives. It would no longer make much sense to keep only paper documents. Moreover, this archiving is a treasure trove of billions of pages in all sectors of collective intelligence, but also a reflection of the controversies, crises and challenges of our societies. National archives are not redundant as a result of the Internet Archive, they complement it, refine it nationally and keep track of "national heritage". National repositories also guarantee the sustainability, the legal nature of the framework for access and citation of these sources, and their durability. Libraries and archiving institutions also do a remarkable job related to digital literacy and training. They develop specific tools and are in direct contact with researchers' requirements as a result of collaborative projects.

What are the main research approaches that you use to study the archived web?

NB: I have used methods in both ends of the continuum of close and distant reading, to use the terms coined by Franco Moretti[26]. My first web history study focused on one website only, namely the website of the national Danish Public Service Broadcaster (DR), and I based the study on a large number of internal documents, but also archived versions of the website. I used classical historiographical methods such as document analysis and source criticism. In my latest projects, like the one mentioned above about the entire Danish web domain, I have used more quantitative based methods, simply because of the scale of the material. In some sub-projects of that study I have also used network analysis to map the hyperlink network of the Danish web. As is always the case one has to adopt the approaches, methods and tools that are the best fit to help answer the research question one pursues, and to enable the opening up of the available sources.

VS: I have used the web archives both as a research object and as a source for research. With Francesca Musiani, for example, we approached the politics of web archiving and explicitly the question 'Do Web Archives have Politics?' through an STS (Science and Technology Studies) approach, studying web archives

as boundary objects[27], their modes of governance, and the multiple agencies and stakeholders that contribute to them.

I also used web archives as a source, for example for my habilitation to direct research and the book Under Construction. La fabrique française d'Internet et du Web dans les années 1990 (Ina Éditions, 2019), and the Web90 project already mentioned. There my approach was globally rather qualitative. This did not prevent me from conducting experiments in distant reading, for example on the plethora of born-digital sources of the 2015 attacks in France. I am convinced that a scalable and "medium" reading (a notion I developed, that incorporates both qualitative and quantitative approaches, that is sensitive to the medium (Web, platforms, multimedia and transmedia) and to the context of production) is necessary. As noted by Jane Winters (2017: 239), "For most humanities scholars it will be a very long time before they transition to using solely digital sources, let alone solely born-digital sources, and for many this will never be the case. They will continue to mix and match, to compare and contrast, and to work with overlapping sets of material which contain subtly different information and are designed for subtly different audiences."

What do you consider to be the most exciting aspects of studying the historical web?

NB: I think one of the driving forces in my academic career has been to enter uncharted territory. In a way this is always what researchers do — if they knew the answers to their questions there was basically no need to do the research — but to me there is a big difference between, on the one hand, doing yet another study of social media use, and, on the other hand, venturing into a field of study and a type of sources where no one has been before. I am definitely intrigued by the latter, despite the fact that you literally have to start from scratch with not much to guide you and also you may feel a bit alone until more people start to see that what you are doing may be interesting. When I started studying the history of the web around 2000, one could count the researchers in the field on one hand, and it was not until 2010-11 that a researcher interest in the archived web emerged and I had someone to play with. But for now, things have improved a lot, and the field of web archive based studies is maturing more and more, in particular with many early career scholars entering the field.

VS: I first started by focusing my thesis on the history of Internet infrastructures[28]. My approach was essentially that of a historian of innovation, engineering cultures and standards. With the history of the Web, it's a more bottom-up approach, a history of digital cultures for the general public. This is how I went from "pipes" to content. I also like the possibility of integrating visual studies and maybe in the future also sound studies. And then, I found with Web archives, a research community, initially modest but growing, full of sharing and friendly spirit, as evidenced, for example, by the biennial RESAW conferences[29], mixing researchers and archivists. I am also passionate about returning to the question of heritage, of the archive, as in the collective work Qu'est-ce qu'une archive du Web? and about working in constant interaction with librarians and archivists. The digital hermeneutics dimension, which is one of the core research areas of my laboratory, the C2DH[30], at the University of Luxembourg, also remains for me a field of reflection that is constantly being renewed and of which I never tire.

What are currently the most significant challenges for research use of web archives?

NB: There are several, but the need for documentation of what is in web archives, and the need for having content from web archives extracted would be high on my list. As to the first because of the scale and the way web archiving works web archives tend to be black boxes where no one, not even the web archives themselves, know exactly what is in their collection and why. As to the latter we are facing the problem of Research Data Management, that is: who should manage our research data, should it be the web archives or the research institutions? In the first case researchers then have to comply with the research tools offered by the web archives, in the latter case researchers can use the tools and methods that are actually the best fit for their study. Therefore, I think better and more available documentation, and the possibility of being able to extract content from web archives are key to taking the next steps in web archive studies.

VS: Undoubtedly, one of the most important challenges is to enable transnational studies to be conducted smoothly. While the Internet Archive allows access to its collections online, this is not the case for many national collections, especially in Europe, due to the limitations imposed by copyright, legal deposit, etc. Legal deposit, which was, for example, introduced for the Web Archive in 2006 in France, is an opportunity that has made it possible to formalise and define the missions of institutions that collect the Web, but it limits access to the Web Archive. It is necessary to travel physically to consult these archives, for example at the Bibliothèque nationale de France (BnF). Allowing remote transnational access to European Web archives or at least the metadata for example, would be a step forward. The interoperability of data and metadata is obviously also a major issue here.

Another challenge is the uneven progress in archiving between European countries; some have an experience of a decade or more, while other countries are only now beginning to address this issue. Finally, there is a major educational challenge, to enable students but also their teachers to fully grasp the potential of the archived Web.

What future challenges in web archiving and/or in studying web archives do you foresee?

NB: In continuation of the above: if web archives are not well-documented and if researchers are not able to get material out of the web archives this may be a challenge for pushing web archive studies further. But also raising awareness about the existence of web archives is important, and that goes for fellow researchers and the wider public. It would be sad if web archives were to experience cutbacks because no one could see their relevance.

VS: Transnational studies remain a major challenge, as I have already said. There is also the capacity to capture audiences on the web and to conduct diachronic studies on different online platforms by being able to relate constantly changing indicators of participation. Another challenge is to develop research uses of web archives from the bachelor's level onwards, or even before. Finally, I also believe that there is also a big effort to be made towards public engagement with web archives. Web archives should not be reserved for researchers only, they are a fertile ground for a much wider public!

What future developments on the web do you anticipate and how do you think web archiving institutions can prepare for these changes?

NB: This is a very difficult question to answer. If someone had asked me this question in, say, 2004 I could not have answered Facebook or Twitter. In other words, it is hard to imagine what will come next, and we therefore tend to see the potential future as a mirror of our present +10%, that is: more of the same. However, what can be said is, first, that the development of the web is very rapid, changes happen all the time, and some of them are very fundamental game changers. Secondly, web archives are always lagging behind when it comes to the archiving of all the new web developments. It took years until web archives were able to collect and preserve Facebook and Twitter — if this is solved at all in a useful way even today — and there is no reason to believe that this tendency will stop or change.

VS: The historian is not always the best equipped for foresight. But let's give it a try! First of all, I imagine, and I hope, a broadening of audiences and research projects based on Web archives: the privileged period experienced by the pioneers, who were able to benefit from a tailor-made follow-up of their projects and the personalised help of librarians and archivists, will have to evolve. Also, there will surely be new literacies and pedagogical tools to be developed, and other new forms of support. There are lab projects under consideration in some institutions already. Another trend to follow is of course that of closed communities, private gardens, paywalls, and other obstacles to archiving which are constantly developing on part of the Web, and the emergence and rapid disappearance (think of Vine[31]) of new digital social networks. This calls for rapid adaptations. In the same way, to be able to seamlessly combine digitised and born-digital sources between institutions seems to me to be an important trend, in addition to the need for transnational studies as well as gateways, ad hoc infrastructures, etc.

What are the main ethical and legal issues related to web archiving and the study of the archived web?

NB: In my view there is nothing new under the sun in terms of ethical and legal issues, compared to other media forms. It is also important to keep the two separate. On the one hand, for countries with a web archive based on Legal Deposit legislation they are entitled to archive whatever has been made publicly available on the online web. I still find this a reasonable and good approach to ensuring that we have preserved an important part of our cultural heritage. But, second, the next question is whether one can use the archived material, and here copyright and privacy legal frameworks have to be taken into account, as they should with all other source types as well. But even if we can legally use material from web archives, we may not want to do this, and this is where research ethics comes in. So, the legal frameworks guide what web archives can archive and what researchers can use, whereas ethical considerations guide what we should do, which in some cases may be less than what we can do.

VS: I find Ian Milligan's presentation[32] on this subject very inspiring. I particularly like this reflection: "I feel similarly uncomfortable with leaving the voices of everyday people completely outside the historical record when there is ample opportunity to include them. Moving to a full opt-in process would likely lead to the historical record being dominated by corporations, celebrities and other powerful people, tech males, and those wanted their public face and history to be seen a particular way". There are many ethical and legal issues at stake, from copyright and the possibility of reproducing screenshots from web

archives to the application of General Data Protection Regulation (GDPR) and the question of anonymisation. One of the challenges highlighted by the National Forum on Ethics and Archiving the Web and the Documenting the Now[33] project, launched in 2016 following the Ferguson protests and riots and the Black Lives Matter movement, is also inclusiveness, not to reproduce existing biases, to take into account issues of gender and cultural diversity for example.

Use of web archives is in most cases rather limited. How do you think that the (research) use of web archives could be boosted?

NB: It is important to have some convincing showcases that can illustrate that one could not study phenomenon X if one did not include the archived web as a source. There may be a difference who the users should be: researchers or the wider public? The showcases should probably be different depending on the audience. And also, I think web archives and researchers studying the archived web should try to reach more out to the groups, institutions, or whatever is studied and try to create a community around the study. Personally, I would very much like to investigate the inclusion of the public in the research a bit more, to venture into "citizen science".

VS: There is a growing interest in the research world for web archives. This is not to say that teachers and students necessarily take the next step towards practice. But in my opinion, this is just a matter of a few years. The efforts we make within bachelor and master's programmes to familiarise students with these types of sources, but also initiatives such as datathons for more experienced audiences, seem to me to be a path to pursue. Putting tutorials online too. And of course, the availability of tools such as The Archives Unleashed Toolkit[34] developed at the University of Waterloo. The 'Saving the Web' colloquium[35] that KBR organised as part of the *PROMISE* project, which is the basis for this interview, and which involved a multitude of stakeholders, also seems to me to contribute fully to this need to make Web archives widely known.

Do any formal ties exist between the national web archive in your countries of residence and the research institutions you work for? Do your institutions have any (in)formal say in the way the selection policies are shaped? If yes, how is this organised on a practical level?

NB: The pilot project mentioned in the beginning of this interview, where two internet researchers collaborated with the two national libraries, has proven to be a blueprint of how such collaborations can take place. Since this first collaboration in 2000 the internet and web research community at Aarhus University has had a number of close collaborations with the national libraries and Netarkivet, which they run. All of my above-mentioned projects have built on this collaboration. In addition, it is stated in the Danish Legal Deposit law that an editorial committee is established with representatives from the web archive, the researcher communities, and the web content providers. I have been a member of this committee for a number of years, and we have discussed collection policies, access forms, and much more. Also in Denmark we have a national digital research infrastructure called Digital Humanities Lab (DIGHUMLAB, see more on dighumlab.org) in which four universities and the Royal Library participate. The research infrastructure for the study of the archived web, NetLab which I'm heading, is part of DIGHUMLAB. This organisational structure has played a pivotal role in sustaining the web archive and researcher contact over time, in contrast to establishing these contacts only based on time limited funding where a lot of knowledge is lost after the funding ends.

VS: The National Library of Luxembourg (BnL) and the team behind; Yves Maurer and Ben Els who are dedicated to Web archiving, are very dynamic (as their website[36] and the recent "Content at risk"[37] event they co-organised can testify) and focused on exchanges with the academic world. For the past two years (since I arrived at the University of Luxembourg), we have been able to work together on several occasions: Yves and Ben are involved in the Master's winter school that I coordinate to present Luxembourg's Web archives to students and help them to discover the content and tools developed at the BnL. The students also have the chance to have a presentation by Els Breedstraet and her team on the archiving of European institutional sites, which is also conducted in Luxembourg, within the framework of the European Publications Office[38]. In addition, in June 2021, we will organise in partnership between the University of Luxembourg and the BnL a "web archiving week" during which a datathon and two major conferences dedicated to web archives will be held; the International Internet Preservation Consortium (IIPC[39]) at the BnL and RESAW[40], a European research group on web archives created and federated by Niels Brügger, at the university. We are therefore working actively together, without playing an active role in the selection policies. However, we feel that the selection policies are very well done, and we know that if we had a suggestion it would be listened to.

What would be the research project of your dreams in the context of web archives if funding and organisational limitations were not an issue?

NB: I would very much like to replicate the studies we make of the Danish web domain on a European level. This is, in fact, the aim of the WARCnet network, but we only have two years and limited funding, but, who knows, maybe this network project can be a stepping-stone to something bigger.

VS: The project proposal that I submitted last year, but didn't make it! It deals with a diachronic study of viral content since the 1990s. It is likely to contribute fully to the study of European digital cultures, popular cultures on the Web and digital social networks. It will allow comparative studies between European countries, both in terms of their Web archiving strategies and the circulation of viral phenomena. But I continue to believe in this dream, a resubmission of the project has already been made. However, it is still difficult to make the specificities of archived Web research understandable: for example, it must be explained to the evaluators that we cannot provide them with a precise, turnkey corpus at the time of submission, because the creation of corpora is a challenge in Web archives that is fully part of the research process.

Do you have any particular advice for institutions wishing to start web archiving, for example which pitfalls to avoid at all cost? Do you have any specific lessons learned from your own experience?

NB: One of the most important lessons learned from the Danish case is that close collaborations between web archiving institutions and research communities are key, and that both parties benefit from this. Web archives will get invaluable information about researchers' interests and needs to help guide the archiving practices and make the collections relevant for future researchers. Researchers will benefit from having the best possible archival copies of important cultural heritage. However, there are also a couple of possible pitfalls to keep in mind. First, it is important that the feedback loop to researchers coversas many different research areas and approaches as possible with a view to making the web archive collections useful for a variety of research topics in the future. Secondly, do not forget to include the costs to these collaborations as a running cost in the budget of the web archive, just as important a budget item as buying hard- and software. If the researcher involvement in web archive activities is only temporary and unsystematic, for instance as part of a funded research project that ends after two years, both web archives and researchers tend to reinvent the wheel every time one research project is followed by a new project. As always when it comes to research infrastructure long term sustainability is pivotal.

VS: I am amazed by what institutions such as BnF and Ina have achieved in France since 2006. They have of course evolved their archiving strategies and their collection and consultation tools over the years to take into account the evolution of the Web, the arrival of social-digital networks, etc. And they have also chosen to work in collaboration with the world of research. It is important to listen to the end users of Web archives. Another piece of advice would be not to get locked into a preconceived vision of what is considered interesting content, often modelled in this case on the world of paper. On the Web, it is not only the online press or institutional sites that count! The culture of Internet users, their forms of expression, in short, an inclusive and non-elitist vision of content is fundamental. Inclusiveness also becomes an important issue: we must not perpetuate gender bias in the Web archives, for example, and we must not reinforce the invisibility of social groups. Countries embarking on web archiving now have the opportunity to build on previous experiences, which can be shared, for example, within the IIPC (International Internet Preservation Consortium). It is an important asset to be able to capitalise and build on these previous experiences.

Our next project will investigate the development of a sustainable social media archiving strategy for Belgium. What do you think it is essential for us to consider, to ensure that researchers use this social media archive in the future?

NB: Everything that has been said above also applies for a social media archive: documentation, possibility of extracting material, creating interest and relevance in research communities, and close collaborations with researchers. Belgium is in a fortunate position in the sense that you can take the best from already existing initiatives, without having to make all the mistakes that others have previously made.

VS: Digital social networks are evolving very rapidly (Vine and other digital social networks have already disappeared) and with them the need for the responsiveness of archiving institutions. Instagram or Periscope are less well archived than Twitter for example or YouTube. This is due to collection methods (public API, use of Heritrix, etc.) and collection choices, but there are also legal limitations on some digital social networks, such as Facebook for private content. An interesting point is that of retweets, which are frozen in time at the time of archiving. If the tweet has a lot of retweets after its archiving date, you can't see them. Another challenge is to collect good hashtags on Twitter. This requires monitoring, sometimes in real time, which is difficult to predict in advance, as shown by the real-time collections during the attacks of 2015 and 2016 in Europe, even if automated solutions are being

tested to more effectively follow the major trends. The citability of these archives is obviously an important point to take into account as well. Another challenge is to be able to cross-reference these results with other collections ("classic" websites, circulation of content between the various digital social media, for example from Twitter to Facebook and vice versa, and with audio-visual sites that include content linked to digital social networks or, indeed, the comments on them). Finally, documenting these collections in order to point out the choices, limitations, accounts or hashtags included or excluded, etc. is obviously also very important, especially when researchers are conducting quantitative studies on the data for which it is important that they know the representativeness, the gaps and the biases.

Conclusion

Even though the *PROMISE* project[41] formally came to an end in December 2019, the work does not stop there. The recommendations and scenarios to establish a federal strategy for the preservation of the Belgian web need to be implemented. An important step in this direction was the appointment in February 2020 of a permanent scientific assistant for web-archiving at KBR to ensure the continuity of this work. Already, metadata descriptions of over 2,400 websites that were harvested as part of the *PROMISE* web-archiving pilot have been provided as a first level of access to the Belgian web archive in the KBR Catalogue[42]. Belgium has much to learn from researchers such as Niels Brügger and Valérie Schafer, and the advice and guidance that they provided in this interview will be extremely valuable in the months to come.

Regarding future research, the work related to born-digital collections will continue in a follow-up project; BESOCIAL, the aim of which is to develop a sustainable social media archiving strategy for Belgium. This 24-month project will take a two-fold approach to social media archiving in Belgium; firstly, investigating the archiving and analysis of the social media channels used by Belgian newspapers included in KBR's collections, and secondly, the archiving of social media content related to events of national and historical importance (e.g. the fire at Notre-Dame in Paris or the terrorist attacks in Brussels). The results of BESOCIAL are intended to be a first major step towards implementing a long-term Social Media Archiving strategy for Belgium.

Furthermore, two additional initiatives at KBR will also contribute to the provision of research access to the Belgian web-archive. Firstly, February 2020 saw the start-up of the emerging Digital Research Lab at KBR[43] as part of a long-term collaboration with the Ghent Centre for Digital Humanities (GhentCDH). The aim of the KBR Digital Research Lab is to stimulate the study of Belgium's digitised and born-digital historical, literary and cultural heritage of the 19th - 21st centuries by: a) facilitating data-level access to KBR's digitised and born-digital collections, b) ensuring that the digitised and born-digital collections are embedded into the researcher's workflow in a user-friendly manner and c) optimising the digitised collections for using digital humanities research methods, such as text and data mining. Related to this, the DATA-KBR-BE project will facilitate data-level access to KBR Collections for digital humanities research through a new open data platform (data.kbr.be), which will make KBR's available as 'Collections as Data'[44] and ensure that they are compliant with the FAIR (Findable, Accessible, Interoperable and Reusable) principles of research data management.

For both of these initiatives, KBR can draw on international experience in this area, such as the international GLAM (Galleries, Libraries, Archives and Museums) Labs Community[45] and the Special Web Archive Collections at the National Library of Luxembourg[46], in the KB Lab at the National Library of the Netherlands[47] and a dataset containing the descriptive metadata of over 2 million websites archived by the Austrian National Library available in ÖNB Labs[48], both of which are particular sources of inspiration. Such 'library labs' are ideal incubators for both increasing access to archived-web resources alongside other digital collections, such as digitised newspapers, and therefore stimulating their take-up and use in the (digital) humanities and social sciences research communities and beyond.

Edited by:
Friedel Geeraert
Nadège Isbergue
KBR
Boulevard de l'empereur 4
1000 Brussels
friedel.geeraert@kbr.be
nadege.isbergue@kbr.be
www.kbr.be
Sally Chambers
GhentCDH
St. Pietersnieuwstraat 35
9000 Ghent
sally.chambers@ugent.be
www.ghentcdh.ugent.be

May 2020

Notes

1. <https://www.netlab.dk/netlab/people/niels/>
2. <https://www.c2dh.uni.lu/people/valerie-schafer>
3. See for example, Brügger, N. (Ed.) (2017) Web 25: *Histories from the First 25 Years of the World Wide Web*, New York: Peter Lang Publishing.
4. Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. International Journal of Digital Humanities, 1(1), 85-111. <https://doi.org/10.1007/s42803-019-00007-7>
5. Winters, J. (2017) 'Coda: Web archives for humanities research: some reflections', in *The Web as History: Using Web Archives to Understand the Past and Present*, ed. Niels Brügger and Ralph Schroeder (London: UCL Press, 2017), pp. 238-248.
6. <https://buddah.projects.history.ac.uk>
7. <http://resaw.eu/about/>
8. <https://cc.au.dk/en/warcnet/>
9. <https://www.kbr.be/en/projects/*promise*-project/>
10. Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 1(1), 85-111. <https://doi.org/10.1007/s42803-019-00007-7>
11. <https://www.kbr.be/en/colloquium-saving-the-web/>
12. <https://www.kbr.be/wp-content/uploads/2019/11/2019-10-18_Saving-the-Web_Brugger.pdf>
13. <https://www.kbr.be/wp-content/uploads/2019/11/2019-10-18_Saving-the-Web_Using-web-archives-for-researchers.pdf>
14. <https://www.kbr.be/wp-content/uploads/2019/11/2019-10-18_Saving-the-Web_deVos.pdf>
15. <https://www.kbr.be/wp-content/uploads/2019/11/2019-10-18_Saving-the-Web_Using-web-archives-for-researchers.pdf>
16. <https://www.netlab.dk/netlab/people/niels/>
17. <https://www.c2dh.uni.lu/people/valerie-schafer>
18. <http://netarkivet.dk/publikationer/webark-final-rapport-2003.pdf>
19. <https://www.cairn.info/revue-le-temps-des-medias-2012-1.htm>
20. <http://en.statsbiblioteket.dk/kulturarvscluster/>
21. <https://firstmonday.org/ojs/index.php/fm/article/view/10384>
22. <https://cc.au.dk/warcnet/>
23. Centre national de la recherche scientifique, the French National Centre for Scientific Research.
24. <https://web90.hypotheses.org>
25. <https://books.openedition.org/oep/8713>
26. See for example: Moretti, F. (2005) Graphs, maps, trees: abstract models for a literary history. Verso and Moretti, F. (2013) Distant reading. Verso Books.
27. See for example: Nicolini, D., Mengis, J., & Swan, J. (2012). Understanding the role of objects in cross-disciplinary collaboration. Organization science, 23(3), 612-629.
28. Schafer, V. (2007). Des réseaux et des hommes. Les réseaux à communications de paquets, un enjeu pour le monde des télécommunications et de l'informatique françaises (des années 1960 au début des années 1980) (Doctoral dissertation, Paris 4).
29. <https://resaw.eu/events/>
30. Luxembourg Centre for Contemporary and Digital History (C²DH): <https://www.c2dh.uni.lu>
31. <https://en.wikipedia.org/wiki/Vine_(service)>
32. <https://ianmilli.wordpress.com/2018/03/27/ethics-and-the-archived-web-presentation-the-ethics-of-studying-geocities/>

33. <http://www.docnow.io/>
34. <https://archivesunleashed.org/aut/>
35. <https://www.kbr.be/en/colloquium-saving-the-web/>
36. <https://www.webarchive.lu>
37. <https://www.science.lu/fr/content-risk>
38. <https://op.europa.eu/fr/home>
39. <https://netpreserve.org>
40. <https://resaw.eu>
41. <https://www.kbr.be/en/projects/*promise*-project/>
42. For example, here is the metadata description of archives website of "Brusselse Bibliotheken", the Dutch-language public libraries in Brussels: <https://opac.kbr.be/Library/doc/SYRACUSE/20776949/brusselse-bibliotheken>
43. <https://www.kbr.be/en/projects/digital-research-lab/>
44. <https://collectionsasdata.github.io>
45. <https://glamlabs.io>
46. <https://www.webarchive.lu/what-we-have/>
47. <https://lab.kb.nl/datasets?f%5B0%5D=field_product_type%3A1>
48. <https://labs.onb.ac.at/en/datasets/>

BEHIND THE SCENES OF THE BELGIAN WEB ARCHIVE RESEARCH OPPORTUNITIES AND CHALLENGES

Patricia BLANCO
Internship at the Royal Library of Belgium during the PROMISE project
MSc in Digital Humanities at KU Leuven

■ When the *PROMISE* project was launched in Belgium to set up a national web archive, a researcher in Digital Humanities was allowed to participate in the different stages of the web archiving workflow: selection and harvesting of websites, quality control of the captures and access to the archived files to test research tools. The goal was to know the potential researchers' needs and expectations, but also the technical requirements and limitations of providing data-level access to the files. This article summarizes the best practices and the challenges that were identified during this researcher's experience.

■ Lorsque le projet *PROMISE* a été lancé en Belgique pour mettre en place une archive web nationale, un chercheur en *digital humanities* a été autorisé à participer aux différentes étapes du processus d'archivage web : sélection et récolte des sites web, contrôle de la qualité des captures et accès aux fichiers archivés pour tester les outils de recherche. L'objectif était de connaître les besoins et les attentes des chercheurs potentiels, mais aussi les exigences et les limites techniques pour fournir un accès aux fichiers. Cet article résume les meilleures pratiques et les défis qui ont été identifiés au cours de l'expérience de ce chercheur.

■ Bij de lancering van het PROMISE-project in België om een nationaal internetarchief op te zetten, mocht een onderzoeker in Digital Humanities deelnemen aan de verschillende fasen van de workflow voor internetarchivering: selecteren en verzamelen van websites, kwaliteitscontroles van die sites en toegang tot de gearchiveerde bestanden voor het testen van onderzoekstools. Het doel hiervan was om de potentiële behoeften en verwachtingen van onderzoekers te kennen, maar ook de technische vereisten en beperkingen van het verstrekken van toegang tot de bestanden op gegevensniveau. Dit artikel geeft een overzicht van de beste praktijken en van de uitdagingen die tijdens de ervaring van de onderzoeker naar boven kwamen.

Introduction

The first initiatives to preserve websites were launched in the mid-90s by nonprofit organizations and national libraries. They were aware that the information published online and its form were unique, but also prone to disappear without a trace. The main motivation for national libraries was to preserve it, both for the general public and researchers. However, many people today are still unaware of the existence of web archives and their use has not yet been consolidated among the research community. Capturing, accessing and analyzing archived web data is still plagued by unknowns and challenges, both technical and legal.

The *PROMISE* project[1] was initiated by the States Archives and the Royal Library of Belgium (KBR) in 2016 to build a national web archive. Raising awareness and promoting its use for research became one of their priorities. A Master' student was invited to participate in the project and offer some feedback on the potential users' needs. The student created a collection of websites around a specific topic in order to explore different selection methods; assisted during quality control tests to identify the problems with the web harvesting tools and ultimately, explored the use of computational tools (text and hyperlinks extraction and analysis) to understand the technical requirements of giving data-level access to the archived files.

This article aims to shed some light on web archives and show what happens behind the scenes of a national web archive.

Selection of web content

In countries with legal deposit laws, national archives are authorized to archive websites with patrimonial purposes without having to ask the author's permission. This allows them to launch large-scale crawls and systematically archive, for instance, every website with the country-code top-level domain (ccTLD): websites ending in ".be", ".brussels", ".vlaanderen", ".gent" for Belgium, ".es" for Spain, etc. However, there are other generic top-level domains (gTLD), such as ".net, ".edu", ".gov" or ".org", that might also contain information related to a country and its inhabitants.

For the creation of special collections, the selection of websites needs to be done manually. Under the topic "representation of minorities in Belgium", three collections were created in order to save websites related to the Spanish, the Italian and the Portuguese communities in the country.

Search engines were the main tool used to find relevant websites, with a special attention to those disseminating news, literature, cultural associations and events concerning these communities. The results were obtained through the combination of keywords that were translated into the official and co-official languages of Belgium, Spain, Portugal and Italy[2].

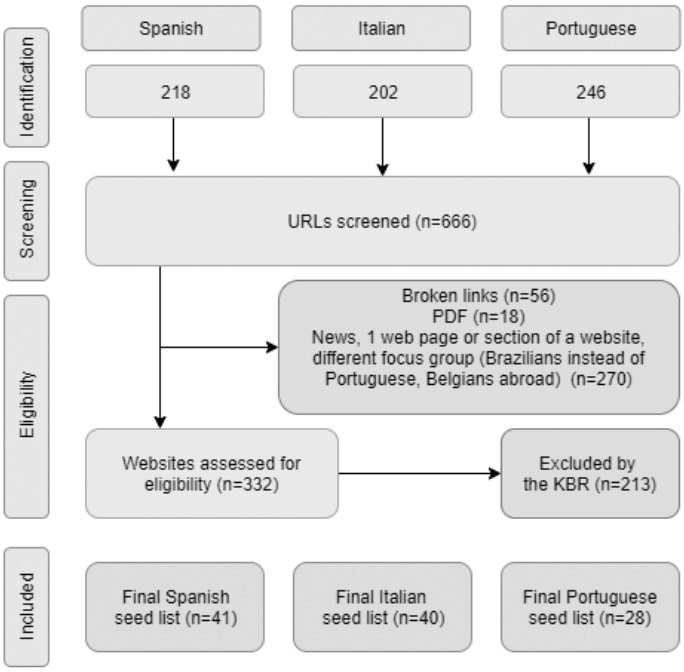

Fig. 1: Flow diagram illustrating the selection process of thematic collections.

Social media platforms were excluded from these collections. They were, however, used as a crowdsourcing tool to retrieve relevant websites. A call for participation was made in different Facebook groups of these communities: users were asked to share the websites they use to connect with their compatriots in Belgium, share experiences, promote their language and culture, organize events, etc. By encouraging participation, we also brought attention about the existence of our web archive.

From the initial 666 URLs screened, only 109 were selected for the crawl (fig. 1): some websites were no longer accessible and those excluded did not fit into the KBR's cultural heritage preservation mission (small businesses websites) or were not sufficiently relevant to be archived entirety (one-page news articles).

Research opportunities

These collections are not only a testimony of the presence, history, impact and evolution of foreign communities in Belgium, but a corpus from which other countries can benefit to carry out transnational research.

The largest existing text corpus is the Web. It is a gold mine for the field of computational linguistics. These collections contain text in ten different languages[3] and multiple dialectal variations. The text can be automatically indexed to allow full-text search and text analysis. Since it is machine readable, we can take advantage of Natural Language Processing (NLP) tools to extract data (Named Entity Recognition), and improve natural-language understanding and natural-language generation tools (machine translation). This multilingual corpus also allows the study of language varieties, such as dialects, sociolects and multiethnolects.

Although commercial pages have not been included, directories of companies and organizations have been kept. They contain addresses of small businesses and professionals, trade unions and cultural associations created by emigrants in Belgium.

With this information we can map their location and observe the geographical distribution of these communities across the country, explore their sectors of activity and study their socio-economic evolution.

Harvest and Quality control

There are a number of open-source tools we can use to harvest websites. However if we want to work with a large sample of websites, we still rely on institutions with the capacity to regularly archive and preserve millions of them. Some organizations, such as the Internet Archive or Common Crawl, opt for a discovery crawl: they feed the web crawler a list of websites to be harvested, let it follow the hyperlinks that they contain, add them to the list and continue to increase the scope. Their goal is to automatically discover new websites and preserve the maximum possible of the web sphere.Cultural institutions tend to limit the scope to the websites of their selection, particularly for thematic collections. Used by most

	Spanish	Italian	Portuguese
Initial seed list	41	40	28
URLs downloaded	1,772,554	2,610,369	2,604,372
URLs queued	87,438	7,938,266	14,166,405
Size	81 GB	163 GB	865GB

Tab. 1: Harvest progress: initial seed lists, URLs downloaded and queued, and size of the collections.

national libraries, Heritrix was also the web crawler used by the *PROMISE* project to capture extracts of the Belgian web. Heritrix visits all the hyperlinks in our list ("seed list"), capture their content and save it into multiple container files, with a maximum size of 1 GB each. These container files are known as WARC files (Web ARChive), a standard file format for web archival.

Problems and possible solutions

We identified the first problems in the harvest when the size of the Portuguese collection, initially with less URLs, was increasing faster and slowing the download (Tab. 1). When we analyzed the WARC files, we noticed that many contents were videos from a Portuguese news channel in Belgium. This made us re-evaluate our selection strategy and the parameters set on web crawler. These can be configured to restrict the capture of certain files, such as audio and video, but that would also mean that our archived versions would be incomplete.

Can we calculate the size of a crawl in advance? As far as we know, there are no specific tools available. One of the solutions could be launching a crawl to collect only the metadata of the websites in our list, without archiving the contents. The metadata about the size of the files would provide an estimation of the space we will need. We could then decide if we restrict the capture of some files on certain websites or if we excluded them from the seed list.

We replay the websites to see if their appearance matches the live version and if all the contents had been archived correctly. We identified the following issues in a number of websites: variations in the fonts, the page layout had not been captured and the contents had been displaced, images or icons were replaced by symbols or disappeared, making some drop-down menus undiscoverable, etc.

The root of these problems stems from the limitations of the software itself, especially when capturing inline JavaScript and dynamic content. It also depends largely on the "archivability" of the website which determines how easy it is to archive. If standards are not respected in the design of a website, this will not be 100% archivable. There are tools to evaluate the archivability of a website[4], a practice that should be generalized among web builders and web archivists. Websites poorly designed slow down and even block the web crawlers. Others websites could not be captured due to the robots.txt file. This is a file, added by the author or the web designer, to prevent web crawlers from harvesting the entire website or sections of it.

Using web archives

The most straightforward way to access archived websites is by interacting with replayed versions of the web pages[5]. With some limitations, we can navigate the old version as we would do it with a browser. We can observe their appearance, read its contents, but we cannot process the data or do quantitative analysis.

For any organizations, the challenges of giving access to the web data are not only legal but also technical. The legal deposit serves to justify archiving websites for patrimonial purposes, but other legal restrictions play a part concerning the publication of this data[6].

There are tools that allow us to concatenate WARC files and extract only the data we need, such as text, images, metadata and hyperlinks. A sample of the three collections (Tab. 2) was placed on a remote server and we used the *Archives Unleashed Toolkit* (AUT), a command line-based tool, to explore it. One of its features is to identify the types of files and file formats captured. Thousands of files could not be recognized among the most common or standardized formats, which reveals the problems with long-term preservation (LTP) once these files have to be migrated.

We also used the AUT to extract hyperlinks and text filtered by language or by web domain.

	Spanish	Italian	Portuguese
Number of WARC files	22	11	36
Total size	~ 70 GB		

Tab. 2: Number of WARC files and size of the sample used to explore the collection.

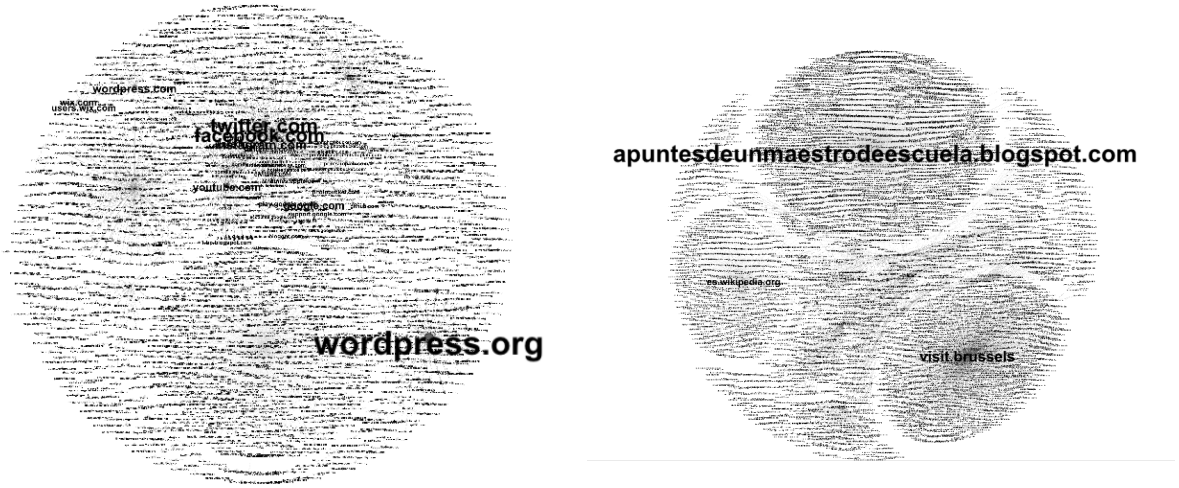

Fig. 2: Data visualizations to explore the hyperlink networks in our collections.

Hyperlinks can be visualized with tools such as *Gephi*[7]. We could observe which domains have more presence in our collection but also how they link to each other and towards other websites. The study of the hyperlinks can reveal strong connections between websites that we would not be able to appreciate with close reading methods.

For example, some researchers have studied hyperlinks to reveal connections between online newspapers and political parties[8].

The websites more frequently linked to from our collection were Wordpress pages and social media platforms, such as Facebook, Twitter and Youtube. Some of the websites we selected have links to their social media profile or Facebook group. These are widely used by the expats and are updated more frequently than the website, which sometimes ends up being completely replaced. We cannot overlook the importance of social networks and the need to archive some of their publicly accessible content too.

Fig. 3: Visualizing text from the website La Maison de l'Amérique Latine[9].

Extracting text was more challenging and the 36 WARC files from the Portuguese collection often caused the AUT software to close. The output text by language requires cleansing in order to be analyzed. However, we used text extraction tools to explore the most frequent words from each domain (fig 3). Some can be included in the description of the website in our web archive catalog as keywords or tags.

Conclusion

The development of a big data infrastructure proved to be necessary to let people download, manipulate, analyze or extract GBs of data to carry out their research. However, the priority today is to save the websites. We should first ensure their preservation, so that they can be used once the legislation and technology ease working with this re-born digital source.

And this is a task to which we can all contribute:

- We should use archived versions of websites more often in our references. Archived versions are like editions of a book. The archive date is even stamped on the hyperlink. We can find them in web archives or even create them ourselves with a simple click[10].
- Any organizations, large or small, should take advantage of open source tools to create their own web archive. This will allow more diversity in the sources available in the future.
- We must respect standards when designing websites to make our pages "archivable" and restrict the use of robots.txt files when it is possible. We can send our websites to web archives to be preserved, by sending the hyperlink or the files. Institutions can consider embargoes to give

access only when the author authorizes it, or when the page is no longer accessible online.
- National libraries can offer small samples and curated collections, tutorials to introduce researchers to the web archives and promote their use.

Saving websites asks for a collaborative effort. We can all help to save relevant information and contribute to create sustainable web archives.

Patricia Blanco
KULeuven
Oude Markt 13 -3000 LEUVEN
p.blanco.nunez@gmail.com

April 2020

References

ACKLAND, Robert; GIBSON. Hyperlinks and networked communication: a comparative study of political parties online. *International Journal of Social Research Methodology* [on line], April 2013, vol. 16, nº 3, p. 231-244. <https://doi.org/10.1080/13645579.2013.774179>

Archives Unleashed Project. *The Archives Unleashed Toolkit* [on line]. Archived on 27 May 2019. <https://web.archive.org/web/20190527054243/https://archivesunleashed.org/aut/>

BANOS, Vangelis; MANOLOPOULOS, Yannis. A quantitative approach to evaluate Website Archivability using the CLEAR+ method. *International Journal on Digital Libraries* [on line], June 2016, vol. 17, nº 2, p. 119-141. <https://doi.org/10.1007/s00799-015-0144-4>

BLANCO, Patricia. *Saving the Belgian Web: Web archiving practices, research opportunities and limitations*. KU Leuven, 2019. Master's thesis. MSc in Digital Humanities.

CHAMBERS, Sally; MECHANT, Peter; GEERAERT, Friedel. Towards a national web archive in a federated country: a Belgian case study. In Brügger, Niels; Laursen, Ditte (eds.) *The Historical Web and Digital Humanities*. New York: Routledge, 2019, p. 29-44.

Internet Archive Blogs. *If you see something, save something* [on line], 1 January 2017. Archived on 16 April 2019. <https://web.archive.org/web/20190416230839/https://blog.archive.org/2017/01/25/see-something-save-something/>

VLASSENROOT, Eveline; CHAMBERS, Sally; DI PRETORO, Emmanuel; GEERAERT, Friedel; HAESENDONCK, Gerald; MICHEL, Alejandra; MECHANT, Peter. Web archives as a data resource for digital scholars. *International Journal of Digital Humanities* [on line], April 2019, vol. 1, nº 1, p. 85-111. <https://doi.org/10.1007/s42803-019-00007-7>

Notes

1. *PROMISE* stands for PReserving Online Multiple Information: towards a Belgian StratEgy.
2. Knowledge of these languages was also instrumental in filtering the results. The methodology and list of keywords can be found in the student Master's thesis. In References, Blanco, Patricia, *Saving the Belgian Web: web archiving practices, research opportunities and limitations.*
3. French, Dutch, German, English, Italian, Portuguese, Spanish, Galician, Catalan and Basque.
4. Archive Ready <http://archiveready.com/> , tool developed with the *CLEAR+ method*. In References, Banos, Vangelis; Manolopoulos, Yannis. A quantitative approach to evaluate Website Archivability using the CLEAR+ method.
5. Examples of open web archives: Wayback Machine <https://archive.org/web/>, Portuguese Web Archive <https://arquivo.pt/>, UK Web Archive < https://www.webarchive.org.uk/>.
6. For instance, the Right to erasure, or "Right to be forgotten", included in the European General Data Protection Regulation (GDPR) policy.
7. Gephi <https://gephi.org/>
8. Ackland, Robert; Gibson. Hyperlinks and networked communication: a comparative study of political parties online.
9. The text extracted from *La Maison de l'Amérique Latine* (archived on 16 April 2019) <https://web.archive.org/web/20190416014159/https://www.america-latina.be/> was extracted with the Archives Unleashed Toolkit (AUT) and visualized with Voyant Tools <https://voyant-tools.org/>.
10. Internet Archive Blogs. *If you see something, save something*.

NOUVELLES PARUTIONS

NIEUWE PUBLICATIES

PRESSES DE L'ENSSIB

http://www.enssib.fr

ÉDUCATION CRITIQUE AUX MÉDIAS ET À L'INFORMATION EN CONTEXTE NUMÈRIQUE

▪ *Coordonné par Sophie JEHEL et Alexandra SAEMMER - Collection : Papiers - mars 2020– 320 p. - ISBN : 978-2-37546-126-6*

Interdiction des téléphones portables à l'école, contrôle des plateformes en ligne pour lutter contre la désinformation – l'éducation aux médias se retrouve au centre des politiques publiques numériques. Depuis 2013, la loi de refondation de l'école a inscrit dans ses missions fondamentales une éducation aux médias et à l'information.

Cet ouvrage présente le résultat de trois années de réflexion collective avec des chercheurs explorant l'économie politique de la communication, la sémiotique, la sociologie des usages, la critique des industries culturelles et créatives et la sociologie du genre.

Écrire pour les *Cahiers*

Les *Cahiers de la documentation* sont alimentés par leurs auteurs. Si vous souhaitez partager avec l'ensemble des membres de l'ABD votre expérience dans un domaine ou vos connaissances d'un sujet ou faire le compte rendu d'une conférence à laquelle vous avez assisté, n'hésitez pas à prendre contact avec le Comité de publication : <cahiers-bladen@abd-bvd.net>

Afin d'assurer une présentation cohérente de notre périodique, nous demandons aux auteurs de respecter les instructions aux auteurs disponibles sur <https://www.abd-bvd.be/fr/publications/cahiers-de-la-documentation/ecrire-pour-les-cahiers/>

Schrijven voor de *Bladen*

Bladen voor Documentatie bestaat dankzij de auteurs. Indien u uw ervaringen binnen een domein of uw kennis van een bepaald onderwerp wilt delen met alle BVD-leden of een verslag wilt maken van een studiedag waaraan u heeft deelgenomen, aarzel dan niet om het Publicatiecomité te contacteren via <cahiers-bladen@abd-bvd.net>

Om een coherente presentatie van ons tijdschrift te verzekeren, vragen wij de auteurs de auteursaanbevelingen te respecteren : <https://www.abd-bvd.be/nl/publicaties/bladen-voor-documentatie/schrijven-voor-de-bladen/>

ABD
BVD

ASSOCIATION BELGE
DE DOCUMENTATION

BELGISCHE VERENIGING
VOOR DOCUMENTATIE

asbl créée le 21 mars 1947
vzw opgericht op 21 maart 1947

Plus de 500 professionnels de
l'information et de la documentation

Meer dan 500 informatie- en
documentatiespecialisten

http://www.abd-bvd.be

Correspondance
c/o Bibliothèque royale de Belgique
Boulevard de l'Empereur 4
1000 Bruxelles
Belgique
abdbvd@abd-bvd.be

Briefwisseling
p/a Koninklijke Bibliotheek van België
Keizerslaan 4
1000 Brussel
België
abdbvd@abd-bvd.be

Tarif anciens numéros
Prix au numéro : 20 EUR
Prix par article : 10 EUR

Tarief vorige nummers
Prijs per nummer: 20 EUR
Prijs per artikel: 10 EUR

Commandes
tresorier-schatbewaarder@abd-bvd.net

Bestellingen
tresorier-schatbewaarder@abd-bvd.net