Le dépôt légal du web cartographique

Rencontre avec Guillaume Lebailly et Peter Stirling – 7 novembre 2013

Quelles sont vos attributions aux Cartes et Plans ?

Guillaume Lebailly. Je suis responsable du dépôt légal des documents cartographiques ce qui implique de suivre la totalité de la chaîne du dépôt légal. Sur le site François Mitterrand, le dépôt légal des livres et des périodiques est scindé en différentes étapes alors qu'ici, sur le site Richelieu, nous sommes une petite équipe qui traite le dépôt légal de l'ouverture des paquets jusqu'au catalogage, en passant par l'enregistrement, l'attribution d'un numéro de dépôt légal, la redistribution du deuxième exemplaire, la mise en magasin, la bibliographie nationale.

A côté de cela, je suis depuis 2009 coordinateur du dépôt légal du web pour le domaine cartographique au département des Cartes et Plans.
Le principe est qu'il y a un réseau de correspondants sur la totalité des départements de collection de la BN pour qu'existe au delà de l'équipe technique, un réseau de relais de gestionnaires de collections à même de pouvoir compléter des suggestions de collecte.

De manière anecdotique, je participe actuellement au commissariat de la grande exposition "Eté 1914" qui ouvrira les commémorations de la Grande Guerre (BnF, site François-Mitterrand, 25 mars-3 août 2014).

Peter Stirling. J'appartiens au département du Dépôt légal qui s'occupe des livres et des périodiques, et depuis quelque temps, des sites web. Nous travaillons avec la DSI, le département des Systèmes d'information, chargée de l'expertise technique. Notre modèle de collecte est composée de deux parties :
la collecte large est une collecte massive, par des robots, une fois par an, principalement de tout ce qui est en .fr.
la collecte ciblée est faite par un réseau de correspondants : les bibliothécaires de chaque département et sur certains projets, des partenaires externes, identifient des sites qui ne sont pas en .fr, ou doivent être collectés plus en profondeur ou plus souvent.

Combien de personnes travaillent ici sur le dépôt légal ?

Guillaume Lebailly. Cela varie. Nous sommes 5 actuellement à travailler sur le dépôt légal des documents cartographiques. Deux collègues sont chargés de l'enregistrement des documents et des redistributions du deuxième exemplaire, et les deux autres collègues cataloguent les documents de manière précise dans notre base (en indiquant par exemple l'échelle et les coordonnées géographiques des cartes et atlas). L'une de ces catalogueurs est également correspondante pour le dépôt légal du web et en particulier, chargée de la partie sur les sites liés à la thématique du voyage.

Combien de sites français dans cette collecte large ?

Peter Stirling. Cette année, 4 millions de domaines. Mais tous ces domaines ne possèdent pas de contenu. Beaucoup sont enregistrés dans l'attente d'un site ou pour les redirections (d'un site en .fr vers un .com par exemple).

Vous faites donc une veille pour trouver de nouveaux sites

Oui et la veille doit se faire sur l'existant pour vérifier que la collecte a bien été effectuée et que les sites existent toujours. Le grand intérêt du dépôt légal du web est d'archiver des sites qui peuvent disparaître du web actif. C'est un prolongement compliqué à mettre en place mais qui correspond à tout ce qui a été fait pour le dépôt légal des livres, des journaux, des cédéroms etc.

Comment retrouvez-vous ces nouveaux sites ?
Guillaume Lebailly. Nous avons le réseau de nos déposants traditionnels : je suis abonné à toutes leurs newletters et autres listes d'information, à des blogs, à des sites et je me suis constitué des systèmes d'alerte avec un compte Netvibes, qui me permet de suivre l'actualité et de déborder du domaine car le dépôt légal du web ne doit concerner que le web français mais pour comprendre le web français, il faut faire une veille plus large et la mienne porte sur toute l'actualité cartographique.

Certaines choses nous échappent mais nous avons l'essentiel. Cela nous permet de mieux comprendre en termes de productions éditoriales quelles sont les marottes du moment, quelles seront les incidences sur ce qu'on recevra sous forme papier et ce qui sera collecté sous la forme web. Il y a des modes chaque année. Ainsi Google Maps avec sa technologie « indoor » (Street View) a recensé des intérieurs de musées, de centres commerciaux et a fait apparaître de nombreux plans de bâtiments. En 2012, la mode portait plutôt sur la cartographie humanitaire. Avec les « printemps arabes », pour obtenir des informations, on a proposé aux gens sur place de cartographier leurs informations, de signaler les bâtiments détruits en punaisant les lieux sur des cartes Google Maps ou autres.

Ces modes renouvellent les productions sur support et notre veille sur les sites web.

Tous les sites ne sont pas collectés : pourquoi ?
Peter Stirling. Nous rencontrons beaucoup de problèmes techniques en ce qui concerne les sites cartographiques et notamment sur Google Maps et assimilés car les technologies de collecte sont toujours un peu à la traîne par rapport au web. Nos robots collectent sans difficulté le web classique avec ses pages web à plat et ses liens statiques mais souvent les cartes des sites cartographiques reposent sur des bases de données qui restent inaccessibles. On a donc souvent des sites cartographiques dans leur contexte avec au milieu un cadre noir.

Guillaume Lebailly. Le robot récupère les fichiers des différents objets présents sur la page, clique sur ce qui est cliquable dans une page web et reconstitue l'architecture. Le robot ne peut pas saisir une requête dans un moteur de recherche. Dans un catalogue, s'il n'y a pas de liens directs, nous ne collecterons pas ces informations. Les SIG (systèmes d'informations géographiques) où il faut cocher et décocher des couches peuvent passer mais s'il faut saisir un toponyme, le robot n'obtiendra rien.

Peter Stirling. Sur le plan technique, les progrès sont constants. Ainsi, le flash passe de mieux en mieux. De même, d'énormes efforts ont été faits pour la collecte des vidéos et sur Dailymotion, les résultats sont bons. On travaille désormais sur d'autres plates-formes.

Une partie du web est inaccessible parce qu'il est sous mot de passe et payant.
Guillaume Lebailly. Fin 2011, un décret a complété les textes et nous a donné l'autorisation légale de passer outre.

Peter Stirling. Avec une précision : le dépôt légal touche seulement ce qui est publié donc les mots de passe pour protéger la vie privée sont exclus (sur un compte Facebook par exemple). En revanche, pour un mot de passe qui protège un accès payant, nous avons le droit de demander les codes d'accès. Nous avons travaillé sur cet accès cette année, notamment vers la presse où existent des titres uniquement en ligne. Cela se fait au cas par cas avec chaque éditeur.

Quelle pérennité sur l'information ?
Peter Stirling. La BNF possède un système de préservation pour toutes les informations numériques à long terme : préservation des données en l'état par des sauvegardes périodiques et préservation par une évolution des fichiers si la technologie de lecture devient obsolète. Ce n'est pas encore fréquent mais a déjà existé comme par exemple pour les formats des disques vidéos des années 80. On identifie donc précisément les formats des fichiers et on a une veille pour définir les fichiers à risque afin de prévoir leur migration sous un autre format.

Guillaume Lebailly. C'est important pour répondre à l'exigence des textes qui réclament une conservation pérenne : on ne pilonne rien, et les lecteurs doivent toujours pouvoir y accéder. Il en est de même pour les catalogues et pour les bibliothèques numériques comme Gallica dont la préservation est essentielle.

Peter Stirling. Il est intéressant de noter que pour beaucoup, le papier se conserve tout seul et que le numérique disparaît alors que le numérique peut être copié et est donc davantage à l'abri d'un incendie. Le risque est celui de l'évolution technologique.

Guillaume Lebailly. La grande force des archives du web telles qu'on les collecte et telles qu'on les met à disposition des lecteurs accrédités dans les salles de lecture est de reconstituer l'architecture pour permettre de naviguer sur le site comme si on était sur internet à la date donnée.

Comment classez-vous ces archives ?
Peter Stirling. Ce n'est pas catalogué ou classé mais indexé par URL.
La BNF n'a pas de catalogage à la pièce à cause de la masse de données et aussi, parce que la question se pose du site en tant qu'objet documentaire. On collecte des fichiers qui constituent un site mais qu'est-ce qu'un site ?
Nous n'avons malheureusement pas non plus d'indexation plein texte. On ne peut pas consulter les archives en saisissant un mot. La recherche se fait par URL donc il faut savoir quel est le site recherché. Nous aurons besoin d'outils plus intelligents. Les technologies existent mais faute de ressources nous n'avons pas encore pu les mettre en place.

Et que font les autres établissements soumis au dépôt légal ?
Guillaume Lebailly. La notion de dépôt légal varie selon les pays. Au Québec, ils ne collectent que les documents gouvernementaux.

Peter Stirling. Au Royaume-Uni, la loi a changé. Jusqu'à cette année, il n'y avait pas de loi sur le dépôt légal : il n'y avait que des collectes ciblées et il fallait demander l'autorisation des producteurs. La loi a changé en 2013 et ils ont lancé leur première collecte large.

Peter Stirling. Il existe IIPC, un organisme international de 46 membres dont l'une de fonctions est de se mettre ensemble et de communiquer sur les technologies. Chaque institution a des contraintes différentes. L'une peut avoir travaillé sur la préservation, l'autre sur les collectes larges ou sur Javascript et ces informations peuvent être partagées.

Nous consultons l'écran de recherche affiché pour le lecteur accrédité, en salle de lecture.

Guillaume Lebailly. Sur cette page, voilà ce qui est affiché pour le lecteur accrédité, en salle de lecture.
Les 3 manières d'interroger apparaissent : la recherche par un bouton URL avec le bouton remonter le temps
une interrogation plein texte qui reste anecdotique
et enfin, le parcours guidé qui regroupe des idées de collecte sur des sujets donnés tels que l'administration en ligne, avec un découpage de parties thématiques et des propositions de consultation de sites à certaines dates. C'est une logique de corpus, un travail sur le fond.

Peter Stirling. Au début, c'était pour pallier l'absence de recherche plein texte et pour proposer une recherche plus conviviale que les URL. La logique reste celle-ci mais elle a évolué en valorisation des collectes ciblées ou de projets qu'on a monté avec des partenaires externes. C'est une forme d'exposition des archives.

Guillaume Lebailly. Le projet sur les carnets de voyage sera prêt en 2014 : la collecte a été faite et il y a eu sélection de sites et rédaction des textes.

Peter Stirling. Nous signalons la présence dans nos collections des URL présentes sur data.bnf, par exemple sur une fiche auteur pour montrer que nous avons collecté le site de l'auteur à telle date. Nous allons aussi intégrer les sites de la collecte ciblée dans les pages thématiques de data (qui utilise la classification Rameau). Ainsi, en plus des informations livres ou documents audiovisuels, nous aurons les sites web archivés. L'objectif est de donner plus de visibilité puisque ces informations ne sont pas dans le catalogue général et qu'on ne peut pas y accéder de l'extérieur. Ceci dit, nous espérons aussi ajouter des « permaliens » pour permettre un accès direct aux archives depuis data quand le site est consulté sur place dans les salles de recherche.

L'accréditation ne peut pas être différenciée selon le contenu ?
Peter Stirling. Non car c'est le statut du dépôt légal : la consultation est uniquement sur place pour les chercheurs accrédités. Ces restrictions empêchent le vol et la dégradation des documents. Pour le numérique, la raison concerne avant tout le droit d'auteur et éventuellement la protection des données personnelles. Que quelque chose soit gratuit ne veut pas dire qu'on peut le copier. La BNF copie en collectant, la loi le permet mais le citoyen n'a pas ce droit. L'accès est donc restreint. Certains sites mis à jour n'aimeraient pas que leur ancienne version soit toujours visible. De plus un site collecté par la BNF pourrait apparaître en premier dans les moteurs des recherche aux dépens du site récent. Il y a aussi la question des liens publicitaires toujours actifs sur des sites archivés.

Guillaume Lebailly. Cette logique de dépôt légal est bien comprise une fois expliquée. Pour un document papier, l'accréditation est réclamée pour consulter un petit plan gratuit d'office de tourisme entré par dépôt légal comme pour un document scientifique.

La collecte d'une plate-forme qui propose des livres numériques permet-elle ensuite l'accès au contenu de ces livres ?
Peter Stirling. Aujourd'hui, ce que l'on peut collecter concerne l'autoédition ou des petites plates-formes où il y a juste les PDF ou les fichiers e-books à télécharger gratuitement. Sur les plates-formes payantes qui nécessitent une inscription, les technologies derrière sont plus compliquées, et on n'arrive pas à les collecter. On est donc en train d'étudier la possibilité d'un dépôt comme sur le modèle du papier, mais ça change pas mal de choses, il faut mettre en place d'autres technologies que celles de l'archivage actuel du web pour les stocker, leur donner accès et les cataloguer.

Guillaume Lebailly. S'il y a une filière de dépôt légal e-book qui nous amène à recevoir plusieurs milliers de fichiers en une seule fois, il faudrait récupérer des métadonnées pour faire des notices élémentaires dans le catalogue général. En amont, il faut donc une réflexion sur les moyens techniques et humains.

Le Département des Cartes et Plans gère les propositions de sites web pour la collecte ciblée du Dépôt Légal de l'Internet dans le domaine de la cartographie. Quels types de sites sont concernés par cette collecte ?
Guillaume Lebailly. Les 350 sites web collectés sont gérés sur la plate-forme professionnelle BCWeb, grâce à laquelle nous visons « l'exhaustivité raisonnée », c'est-à-dire qu'il ne s'agit pas d'avoir uniquement des sites très scientifiques, sérieux ou pointus, il s'agit d'être représentatifs de l'état de l'art sur Internet. On a réparti nos sites en quelques thèmes (aménagement du territoire, cartes, culture, géopolitique, sciences de la terre, tourisme, etc.) et à l'intérieur, on a tous types de sites, du blog de prof de géo au Géoportail, en passant par les jeux géographiques. C'est vraiment très varié.

Pour chaque site repéré, on fait une suggestion de collecte en indiquant les paramétrages pour le robot, notamment la fréquence de collecte en fonction des mises à jour du site, depuis une fois par jour jusqu'à une fois par an, et la profondeur du site (collecte de la totalité ou juste d'un segment). Ensuite on attribue un thème au site, des mots-clés, une petite description et des notes techniques pour vérifier la collecte. Le bouton « état » indique si la collecte est toujours active ou non, ce qui veut dire qu'on a sur cet outil également les sites qui ont été collectés mais qui n'existent plus.
Quels résultats obtient-on pour des sites de cartographie dynamique type Bison Futé ?
Guillaume Lebailly. On copie l'URL de Bison Futé dans notre outil de recherche de l'archivage du web, on clique sur le bouton « Remonter le temps », et voici comment les résultats se présentent : c'est une frise chronologique qui montre, année par année, ce qu'on a collecté. Si je veux voir à quoi ressemblait Bison

Futé en 2005, je choisis une date de collecte : vous voyez, la carte qui s'affiche était une image fixe donc aucun souci, mais quand on clique sur la carte, il manque la base de données derrière. En revanche, en 2012, il n'y a pas de problème. C'est du « mosaïquage » d'images fixes, et non du Google Maps, donc cela convient bien au robot. A 9h22 le 13 août 2012, on voit qu'il n'y avait pas d'embouteillages !

Peter Stirling. En règle générale, les collectes plus récentes sont de meilleure qualité.

C'est quoi la tendance en cartographie sur le web ?
Guillaume Lebailly. C'est la carte sur tout et n'importe quoi. C'est punaiser toutes sortes d'informations sur un fond de carte : il y a le passionné de gastronomie qui punaise sur un fond Google Maps les meilleurs vendeurs de saucisses en Allemagne. La tendance est aussi aux jeux géographiques, et au renforcement du lien entre l'art et la cartographie.

Jusqu'à quand peut-on « remonter le temps » dans les archives du web ?
Peter Stirling. Les premières expérimentations ont été faites en 2002 mais on avait un contrat avec Internet Archive qui nous a donné une extraction pour ce qui est du web français : nos premiers résultats peuvent donc remonter jusqu'en 1996. C'est un peu partiel, mais ça a le mérite d'exister. J'aime beaucoup cet exemple sur l'ancien web du site Total.com : en 1996 ils n'ont qu'un tout petit site avec une page qui explique « Pourquoi Total sur le web ? » comme pour se justifier, et qui met en avant les « possibilités nouvelles de communication offertes par Internet ».

Pour conclure, un petit mot sur les améliorations que vous aimeriez bien voir apportées ?
Guillaume Lebailly. Ce serait l'indexation plein texte des archives du web, et d'autre part le travail sur la collecte d'applications dynamiques. En attendant, ça reste important pour nous de continuer à proposer des sites qu'on n'arrive pas à collecter, car on peut au moins avoir le contexte du site, l'architecture, le mode graphique, les symboles, les explications, les pages d'aide, et c'est intéressant de voir comment le site se présentait à un moment donné. A nous aussi de continuer à trouver comment toujours mieux communiquer sur les archives du web.

Pour la Géofeuille de GéoRéseau