



World Library and Information Congress: 69th IFLA General Conference and Council

1-9 August 2003, Berlin

Code Number: 142-F
Meeting: 86. Bibliography
Simultaneous Interpretation: Yes

Pérennisation des bibliographies nationales dans le nouvel environnement virtuel de l'information

Marcelle Beaudiquez

Bibliothèque nationale de France
Paris, France

Résumé

L'auteur s'interroge sur la possibilité d'appliquer aux sites webs les principes du contrôle bibliographique universel et les recommandations concernant la rédaction des bibliographies nationales, établies dans les années 70 pour les documents traditionnels. Elle identifie ensuite plusieurs scénarios pour « créer » une bibliographie nationale des sites et insiste sur la nécessité de pérenniser dans les recensements des sites web les fonctions des bibliographies nationales comme éléments de la mémoire nationale. Elle termine en demandant une implication forte de l'IFLA dans l'établissement de nouvelles recommandations.

La question de l'avenir des bibliographies nationales ⁽¹⁾ dans le contexte d'un contrôle bibliographique universel du 21^{ème} siècle ⁽²⁾ a déjà été abordée en 2001 à Boston. La présentation d'aujourd'hui peut donc être considérée comme le 2^{ème} volet de cette réflexion. En deux ans seulement, l'évolution des différentes composantes (« documents » à recenser, législations de dépôt légal, environnement Internet, ...) s'est accélérée. Les décisions à prendre par les bibliothèques nationales sont maintenant imminentes. Cette communication contribue à la réflexion internationale en ce domaine et s'appuie également sur certaines réflexions en cours à la Bibliothèque nationale de France ⁽³⁾

* * *

⁽¹⁾ La réflexion ne s'applique qu'aux bibliographies produites par une bibliothèque nationale ou un organisme officiel bénéficiaire du dépôt légal

⁽²⁾ « Usages et utilité des bibliographies nationales : quelles perspectives ? » in IFLA Journal, vol. 28 (2002), n° 1 (<http://www.ifla.org/IV/ifla67/papers/114-119f.pdf>)

⁽³⁾ en France, la prochaine loi sur le dépôt légal des sites web devrait être votée au parlement avant la fin 2003, suivie par son décret d'application. On sait déjà qu'un délai important (sans doute 36 mois) de mise en œuvre sera accordé pour permettre aux différents acteurs concernés de se préparer, selon leurs ressources humaines et financières.

Depuis plusieurs années déjà, les contenus des sites Internet sont considérés comme un élément constitutif de la mémoire patrimoniale, digne d'être collecté, identifié, conservé, communiqué dans le cadre des missions des bibliothèques nationales.

Ainsi, les nouvelles législations de dépôt légal prennent en compte désormais le dépôt des « signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication en ligne »⁽⁴⁾, de nombreuses expérimentations sont faites dans différents pays pour tester la collecte et l'archivage des sites. Pour les sites web⁽⁵⁾, désormais inclus dans les lois de dépôt légal, le recensement devient obligatoire, et ils rentrent ainsi dans le système du contrôle bibliographique universel qui inclut l'établissement d'une bibliographie nationale courante. A partir de ce postulat, il faut s'interroger et définir de quel contrôle bibliographique universel parlons-nous ? Y-a-t-il évolution dans sa mise en œuvre à partir du moment où les sites web sont concernés au même titre que les documents traditionnels ? Quelle bibliographie nationale faut-il établir ? Pour quels usages ?

Pour tenter de définir s'il est pertinent de continuer à établir des bibliographies nationales courantes dans un contexte Internet et l'usage qui leur serait réservé, il m'a semblé utile de raisonner par rapport aux principes et recommandations déjà établis pour les documents traditionnels.

1 - Les principes du CBU et les sites web

- Le contrôle bibliographique universel est basé sur le principe de la **responsabilité nationale** en matière de collecte et de recensement. Ce principe de nationalité n'est pas automatiquement transposable dans un monde virtuel comme celui de l'Internet dont le champ de diffusion n'a pas de frontière, mais des repères approchant peuvent être trouvés avec la notion de domaine-pays (comme .fr, .uk, .se...) à condition de s'interroger sur le sort de ce qui échappe ainsi au recensement « national » dans le .com, org, .net...

N'oublions pas, cependant, que la base de recensement des bibliographies nationales pour les documents traditionnels peut également être linguistique ou documentaire (concernant le pays) : ces mêmes bases peuvent également être adoptées pour le recensement des sites web.

- Le contrôle bibliographique universel est également largement fondé sur le **principe d'exhaustivité**. Appliqué aux sites web considérés comme une nouvelle mémoire patrimoniale, ce principe multiplie par 5 ou 6 le nombre de « dépôts » et donc le nombre de « recensements » potentiels.

Cependant, il ressort de toutes les enquêtes que déjà, pour le document traditionnel, la poursuite de l'exhaustivité est devenue une course sans fin, que l'on parle désormais d'« exhaustivité raisonnée », sinon « raisonnable », et que l'on a déjà introduit le principe de sélection de la collecte dans certaines législations de dépôt légal pour certains types de documents (émissions de radio et de télévision). Il n'est donc pas contraire à l'orthodoxie du contrôle bibliographique universel d'envisager une collecte sélective des sites Internet. Il faut veiller toutefois à ne pas introduire trop fortement la notion d'intérêt documentaire dans les critères pour ne pas porter de jugement de valeur sur la collecte, principe qui a été généralement respecté jusqu'à présent.

Mais on peut déjà s'appuyer sur des critères de nature variée : par exemple, ne sélectionner que les sites web qui, en tout ou partie, assurent la continuité des collections

⁽³⁾ C'est la formule de la future loi française sur le dépôt légal Internet

⁽⁴⁾ On emploiera dans le texte cette formule simple de « site web » par commodité, tout en sachant que la réalité de l'Internet est plus complexe

existantes, soit parce qu'ils sont ou qu'ils contiennent des types de documents déjà collectés sur d'autres supports ; soit parce qu'ils s'inscrivent dans les domaines où la bibliothèque nationale est déjà un pôle d'excellence.

2 - Les principes de la BnC et les sites web

- La bibliographie nationale courante, quant à elle, évoluant avec les législations de dépôt légal, a progressivement recensé des types de documents de plus en plus diversifiés, édités ou produits sur des supports également diversifiés. En tant que document, la bibliographie nationale a elle-même changé de supports "d'édition", qu'elle a d'abord cumulés (papier + cédérom), puis substitués l'un à l'autre (cédérom seul), depuis une dizaine d'années seulement. Aujourd'hui, on s'interroge déjà sur la nécessité ou non de continuer à produire un support cédérom, alors qu'on propose également la bibliographie nationale en ligne. Il faut cependant souligner que cette mise en ligne d'un certain nombre de bibliographies nationales courantes n'est souvent que le transfert en ligne de la structure de la bibliographie préexistante sans exploitation, dans la quasi totalité des cas, de la technologie nouvelle du web pour la recherche ou les interfaces, bien que les index soient cependant plus développés⁽⁵⁾. Ces bibliographies nationales mises en ligne, recensent presque toujours des documents sur des supports « traditionnels », allant du papier au DVD. Ainsi donc, la bibliographie nationale s'adapte aux évolutions éditoriales dont elle témoigne sans pour cela faire de véritable rupture.
- La bibliographie nationale courante est établie sur la base du dépôt légal et la recommandation est de **décrire dans la bibliographie tout ce qui est déposé**. Cependant, on sait que déjà pour les documents traditionnels ce principe de dépôt exhaustif n'est plus toujours appliqué et que l'exhaustivité « raisonnée » du recensement va de pair avec l'exhaustivité raisonnée du dépôt. Les critères de sélection des sites web pourraient être complétés de critères de sélection dans leur recensement bibliographique, si cela s'avère nécessaire.
- La bibliographie nationale courante des documents traditionnels repose sur un autre grand principe, celui du **traitement à l'unité** (bien que des traitements par lots existent déjà pour les « ephemera » ou les images) pour une description normalisée complète qui sert de fiche d'identité au document. Avec les sites web, cette notion « d'unité bibliographique » semble devoir être réservée à l'ensemble d'un site, alors même que les moteurs de recherche permettent une granularité plus fine en identifiant une page, par exemple.
- Avec les documents traditionnels, la bibliographie nationale courante assure immédiatement une **identification pérenne** car l'unité physique décrite est statique, tout changement de contenu entraînant une nouvelle édition, un nouveau dépôt, une nouvelle

⁽⁵⁾ La première évolution importante de la bibliographie nationale courante qui pourrait survenir rapidement est indépendante de la « virtualisation » des documents qu'elle recense : elle mettrait en œuvre la modélisation qui rencontre un nombre toujours plus grand de partisans parmi des bibliothécaires.

L'approche FRBR (Functional Requirements for Bibliographic Records) identifiant l'œuvre et l'expression convient bien à l'identification sur la durée d'une édition nationale au plan intellectuel, tandis que la manifestation traduit la matérialité de la collecte des supports. Les bibliographies nationales rétrospectives, tous types de documents confondus, pourraient ainsi évoluer dans leur ergonomie et la « ferbérisation » gagnerait ensuite naturellement l'ensemble du catalogue. L'adoption ou non de la modélisation de type FRBR dans les bibliographies courantes sera plutôt liée, en premier lieu, à l'évolution des normes de catalogage (le résultat des travaux des ateliers sur le code international de catalogage nous en dira plus le moment venu ; c'est pourquoi je tenais à faire figurer cette évolution possible dans la réflexion générale).

identification. Les sites web sont des ressources dynamiques, les « continuing resources », très souvent modifiées et mises à jour ; ils peuvent également être éclatés dans une architecture telle que tout ou partie du site n'a d'existence que par l'affichage d'une recherche (site « à la volée » ou deep-web). Toutefois, pour la durée de son existence (75 jours en moyenne), c'est le site lui-même qui est identifiable et qui est le plus proche équivalent de l'unité « physique » à identifier.

- Pour un document sur support, la bibliographie nationale courante restitue une **notice bibliographique** sans lien avec le lieu de conservation du document identifié, même s'il est évident que la bibliographie nationale courante étant rédigée sur la base du dépôt légal, il existe au moins un endroit où l'on peut consulter le document identifié même très longtemps après sa publication, c'est la bibliothèque nationale du pays concerné. La bibliographie nationale courante atteste donc de l'existence d'unités physiques stables dont un exemplaire au moins est conservé à la bibliothèque nationale, mais qui ont été produites à des milliers d'exemplaires qui sont conservés dans d'autres organismes, chez l'éditeur lui-même et chez les particuliers qui les ont achetés. Pour le site web, la description, la localisation et la conservation qui peuvent être faites par la bibliothèque nationale correspondent à son état à un « instant t » seulement, celui de la capture du site et dont la trace n'est pas forcément conservée par une autre institution, ni même par son « créateur », au même instant, et encore moins dans sa globalité. Les modes de capture et de dépôt, les modes d'archivage retenus pour restituer une communication dynamique du site avec ses liens d'origine par exemple pèseront nécessairement sur les possibilités ou les obligations d'identification et d'utilisation des sites dans la durée.
- En matière de **rapidité d'annonce**, et depuis longtemps, les bibliographies nationales courantes s'appuyant sur un dépôt effectif dans un établissement bénéficiaire, rivalisent avec les bibliographies courantes d'origine commerciale qui s'appuient souvent sur les informations des documents « à paraître ». Pour tout utilisateur d'une bibliographie nationale courante, la recherche le conduit à une notice bibliographique qui lui permettra dans un deuxième temps de consulter le document là où il est obligatoirement conservé, la bibliothèque nationale. Pour un internaute aujourd'hui, sa recherche le mène directement au site. Cette immédiateté de l'accès au site rend « inutile » toute bibliographie nationale courante des sites, tant que ceux-ci sont vivants (75 jours en moyenne) et gratuits. La simultanéité de la recherche / identification et de la visualisation du site constitue un critère marquant de l'exploitation documentaire du « patrimoine ». Elle induit un repositionnement du rôle des bibliographies nationales courantes : leur rapidité d'annonce n'a plus le même enjeu, car il n'y a guère de valeur ajoutée pour la recherche courante. C'est, par contre, dans la recherche rétrospective que la bibliographie nationale prendra tout son sens si elle accompagne un archivage pérenne.

Ainsi, les autres volets de la chaîne des missions liées au dépôt légal évoluent eux aussi quand il s'agit du dépôt légal des sites web :

- la collecte devient une activité volontariste de l'établissement gestionnaire au sens où il doit obligatoirement mettre en place les moyens techniques de capture de sites ou d'exploration des supports de collecte (fichiers FTP) ;
- la conservation / archivage est indissociable de la collecte qui en induit les choix techniques ;
- la communication, si on ne veut pas qu'elle réduise le « document virtuel » communicable à une partie statique et fermée du site web d'origine, doit pouvoir restituer l'accès au site dans son environnement (ses liens) d'origine, c'est-à-dire être prévue dès sa capture (donc effet induit sur la collecte et la conservation).

3 - Les particularités d'identification des sites web

3.1. quels sites ?

La lente évolution observée actuellement est sans doute normale si l'on considère que la matière première des bibliographies nationales reste « traditionnelle » : la masse de documents imprimés et audiovisuels à recenser reste largement supérieure à la masse des documents en ligne dans la mesure où le dépôt légal de ces derniers ne concerne encore qu'un nombre très limité de pays. Mais, en matière de dépôt légal, de nombreux signes montrent que la mutation s'accélère.

3.1.1. Continuité des typologies de documents collectés

Il faut cependant trouver rapidement des solutions pour les cas les plus urgents, c'est-à-dire pour les types de documents traditionnellement collectés et recensés sur supports et qui ne sont désormais diffusés qu'en ligne, comme partie prenante d'un site web. La continuité des collections du dépôt légal est une préoccupation légitime des gestionnaires et il est important de pouvoir repérer parmi les informations désormais sur le web celles qui se substituent à des éditions antérieures sur supports traditionnels (e-books ou publications en série en ligne par exemple).

Que ces documents virtuels soient, selon le pays, déjà soumis ou non à une obligation de dépôt légal, leur recensement, leur signalement dans une bibliographie doivent être étudiés, sinon la mémoire éditoriale traditionnellement répertoriée est brusquement tronquée.

On voit bien que cette place à part faite à la continuation d'une typologie traditionnelle de documents déjà soumis à dépôt comme les livres, périodiques, mais aussi éventuellement les images, le son, les programmes de télévision, les films..., (en considérant l'Internet comme le nouveau support du moment) peut constituer un critère fort dans la sélection des sites web soumis au dépôt.

Ainsi, en France l'Institut national de l'audiovisuel en France qui sera l'un des gestionnaires du dépôt légal Internet dès la promulgation de la loi a identifié 5 000 sites de télévision et radio en ligne, pour prolonger la collecte des informations (sélection d'émissions de télévision et de radio) dont il a déjà la charge.

De son côté, la Bibliothèque nationale de France étudie la faisabilité d'une identification des périodiques français en ligne sur le web pour éventuellement les collecter et les traiter en priorité parmi les sélections définies dans la futur loi.

3.1.2. Continuité des bases de recensement

Pour d'autres pays, l'intérêt documentaire national et patrimonial des sites web a été le premier critère de choix de capture ou de description. Ainsi la Suède depuis 1997 a capturé automatiquement 126 000 sites du web suédois pour obtenir « a true picture of the Swedish space at the time of archiving » ; l'Australie, à travers le projet PANDORA, a identifié et décrit environ 2500 sites « that have research value for studying the history of Australian society » avant même que l'obligation de dépôt ne soit inscrite dans la loi. Pour le Canada, c'est à l'occasion de la création du nouvel organisme Bibliothèque et Archives du Canada qu'il définit son champ de collecte sur « les documents Internet qui sont attribuables à une origine canadienne ou qui présentent un intérêt pour le Canada ».

Dans ces cas de sélection typologique ou thématique et de volumétrie faible, on considère que la volumétrie à gérer autorise encore un traitement bibliographique à l'unité.

Mais, au-delà de ces 2 cas, pour lesquels diverses solutions de traitement bibliographique peuvent être trouvées (voir 3.2) il semble évident qu'à partir du moment où le contenu des sites web est considéré comme partie constituante de la mémoire nationale au même titre que les documents traditionnels, il faudra toujours pouvoir répondre au questionnement de la recherche :

- quels sites (adresses et titres) ont été déposés / capturés à tel moment ou sur telles périodes
- quels « créateurs » ont déposé sur une période donnée ?
- quels sites ont été créés dans tel ou telle thématique ?

Ce sont là les fonctions que remplit une bibliographie nationale dans la durée, par rapport à un corpus d'informations national ; mais remplir les fonctions ne présuppose pas la forme de production, ni de diffusion tenue pour cette bibliographie nationale courante.

3.2. Quelle bibliographie ?

Il reste donc à prévoir quel(s) mode(s) d'établissement d'une bibliographie nationale des sites web est (sont) possible(s) en fonction des spécificités du monde Internet.

Trois paramètres sont incontournables dans la prise de décision :

- le pourcentage de sites web à recenser potentiellement par rapport au pourcentage de documents traditionnels concernés et leur nombre, sachant qu'à moyen terme le nombre de documents traditionnels sur supports ne diminuera pas sensiblement et qu'ils devront continuer à être recensés dans les bibliographies nationales courantes que nous connaissons, même si celles-ci sont de plus en plus souvent accessibles en ligne ;
- la spécificité de l'environnement Internet et les référentiels qu'il contient déjà : une exploitation maximale des possibilités techniques du réseau et les automatisations que l'on peut en attendre à la place des interventions manuelles de création de « notices », toujours consommatrices de ressources humaines ;
- les conditions juridiques de communication sur place et à distance de ces sites en fonction de la protection des droits d'auteurs dans chaque pays. Certains protectionnismes en la matière pourraient entraîner des mises à disposition différées et donner une perspective temporelle différente en matière de bibliographie nationale ; de plus les difficultés juridiques et techniques liés au dépôt et à la communication des sites payants ou verrouillés devrait amener à les exclure de la collecte. Mais devrait-il en être de même pour la bibliographie nationale, qui serait alors volontairement et irrémédiablement tronquée (ce qui n'est pas le cas avec les échantillonnages).

Dans ce contexte, l'établissement de la bibliographie nationale peut connaître des variantes, et adoptera pour un temps encore, des solutions «mixtes», en parallèle :

- a) si le nombre de sites à recenser est faible, ou si les critères de sélection définis pour la collecte le rendent tel, on choisira d'inclure dans la structure de l'actuelle bibliographie nationale courante, en liste unique ou en listes séparées, des notices descriptives de sites web qui suivront les règles du dernier ISBD révisé à cet effet⁽⁶⁾ ;
- b) quel que soit le nombre de sites à recenser, on peut choisir de traiter leur description bibliographique avec les métadonnées de type Dublin Core, récupérés du site lui-même ou créés manuellement et incluant l'utilisation des numéros d'identification univoques développés pour les informations de toute nature

⁽⁶⁾ ISBD (CR) ...International Standard Bibliographic Description for serials and other continuing resources. Revised from ISBD (S)...IFLA-UBCIM, 2002

accessibles sur le web comme l'ISRC (International Standard Record Code), le GRID (global Release identifier) ou le DOI (Digital object Identifier) par exemple.

- c) lorsque la volumétrie des sites concernés par la loi de dépôt légal est élevée et dépasse les possibilités d'identifications manuelles, il faudra chercher le maximum de procédures automatiques et exploiter les référentiels et les outils du monde Internet. Le document virtuel n'est pas seulement un avatar supplémentaire des changements de support de production et de diffusion de l'information ; il a sa propre logique et ses propres outils. Il semble donc plus pertinent d'exploiter au mieux cet environnement Internet, plutôt que de calquer le modèle traditionnel de recensement.

4 – Etablissement de la bibliographie nationale propre aux sites Internet

* **Base du recensement :**

Le domaine-pays (.fr, .uk, .se) complété si possible par les .com ou .net

* **Description du site :**

Réutilisation des éléments d'identification figurant dans la base du gestionnaire national des noms de domaines, qui dans chaque pays maintient une base de gestion permettant d'**identifier le créateur du site** (organisme et / ou personne physique). Une coopération devrait s'établir entre ce gestionnaire et la bibliothèque nationale chargée du dépôt légal des sites Internet, afin que celle-ci puisse exploiter au mieux les éléments de la base de gestion.

* **Valeur ajoutée par la bibliothèque nationale :**

- On peut envisager pour optimiser cette identification et, si le contexte juridique local le permet, de développer les formulaires de déclaration de noms de domaine ou de s'appuyer sur, et de créer des liens avec, les fichiers d'autorité nationaux de collectivités et de marques commerciales ou d'autorité matière (sans pour cela envisager des créations systématiques de liens manuels naturellement).
- On peut également envisager le développement par la bibliothèque d'un moteur de recherche propre à la bibliographie nationale s'inspirant des indexations matière propres à nos établissements documentaires.

* **Pérennisation de la base de données :**

La base de données constituée par la bibliographie nationale par réutilisation des bases-domaines sera pérennisée de façon à pouvoir conserver dans la durée les données d'identification. C'est plus que ce que font aujourd'hui les moteurs de recherche de type Google ou même les bases de gestion des noms de domaines, qui ne prennent pas en compte la perspective temporelle.

* **Archivage intelligent des sites :**

La bibliographie nationale courante, devenue virtuelle, aura de plus la caractéristique de donner directement accès au site lui-même, restitué de manière dynamique avec ses liens d'origine, grâce à un archivage ad hoc.

5 - Conclusion

Cette réflexion conclut à l'évolution des différents volets d'un contrôle bibliographique universel du XXI^e siècle qui ne renie rien de sa création, mais évolue avec l'information qu'il contribue à véhiculer tout en laissant à chaque pays la possibilité d'évoluer au rythme de sa réalité éditoriale et législative.

Jusqu'à présent, pour répondre aux questions "qui ? quoi ? quand ?", on établissait une liste bibliographique qui était elle-même publiée et conservée comme tout autre document "traditionnel sur support". Tant que des documents "traditionnels" seront produits, la liste continuera à être produite selon les mêmes principes, même si elle est mise en ligne et si ses cumulatifs sont des fichiers informatiques.

Pour les informations sur le web recueillies par capture ou dépôt, il faudra abandonner ce système de listes de description à l'unité comportant le risque des adresses erronées vers des sites non conservés ou disparus. La bibliographie nationale courante sera alors gérée comme une base de données en ligne, interrogeable sur la durée, et qui fournira des sous-produits rétrospectifs autant que de besoin.

Il faudra de plus assurer un archivage intelligent des sites qui anticipera sur les besoins des chercheurs futurs et prévoir une bibliographie nationale rétrospective virtuelle qui répondra alors à leurs besoins.

* * *

L'IFLA, à l'origine du concept de contrôle bibliographique universel en 1973, pourrait conserver une part très active et décisive sur la définition du nouveau contrôle bibliographique universel.

En effet, si l'on admet le rapprochement avec les gestionnaires de noms de domaines et la nécessité pour chaque bibliothèque nationale de collaborer avec ceux-ci pour exploiter les éléments d'identification facilitant une recherche pérenne, l'IFLA pourrait se rapprocher du gestionnaire international de noms de domaines, l'IRCANN (Internet Corporation of Assigned Names and Numbers) qui assure et supervise les tâches liées au fonctionnement du Système des Noms de Domaines (DNS). Il s'agirait de s'assurer que la charte de délégation des domaines nationaux à des gestionnaires locaux mentionne bien la collecte d'éléments d'identification et la mise à disposition des informations auprès des établissements gestionnaires du dépôt légal Internet. En matière d'identification, parmi la prolifération des numérotations internationales univoques, l'IFLA pourrait également être prescripteur, en se rapprochant des organismes internationaux de type ISO et ses agences de maintenance. De même, en matière d'indexation, le travail de l'IFLA sur les « subject gateways » devrait aboutir à des recommandations dont l'utilisation dans les moteurs de recherche serait une véritable valeur ajoutée.

On connaît mieux maintenant les besoins et les possibilités techniques et bibliothéconomiques liés au dépôt légal Internet. Un certain nombre de bibliothèques ont déjà fait des expériences de réalisation. Il est bientôt temps de penser aux recommandations pour assurer une cohérence globale des traitements. C'est le rôle de l'IFLA en matière de description bibliographique depuis l'origine, c'est également le rôle de l'IFLA de fixer l'évolution de son concept de contrôle bibliographique universel.

Ne laissons pas se reconstituer une tour de Babel dans un paysage clarifié depuis 30 ans.