Des identifiants pérennes pour les ressources numériques

Emmanuelle Bermès
Bibliothèque nationale de France, Département de la Bibliothèque numérique
Quai François Mauriac, 75012 PARIS
Emmanuelle.bermes@bnf.fr

Résumé

La création de ressources numériques en ligne soulève la question de l'identification fiable et durable de ces données sur le réseau. Pour les collections numériques, ces difficultés ont été souvent résolues en exploitant d'une part les métadonnées, d'autre part les systèmes d'identifiants.

L'objectif de la mise en place d'identifiants pérennes est de faciliter la citabilité et le référencement des documents numériques, mais aussi leur bonne gestion en vue de la conservation à long terme. Il existe différents systèmes permettant de créer des identifiants pérennes, et différentes approches permettant de les rendre actifs sur le Web.

Le système choisi par la BnF, l'Archival Resource Key (ARK), a permis de favoriser une bonne gestion des objets numériques, indépendamment des changements d'architecture informatique qui ont affecté, et affecteront encore la bibliothèque numérique. Appliqués aussi bien aux métadonnées qu'aux objets numériques, les identifiants pérennes ouvrent d'intéressantes perspectives pour l'exploitation des ressources de la bibliothèque sur le Web.

Mots clefs

Identifiants, URL, URI, ARK

1 Introduction

La création des ressources numériques en ligne, qu'il s'agisse de numérisation, d'une édition ou tout simplement d'un billet ou un commentaire de blog, soulève la question de l'identification fiable et durable de ces données sur le réseau. Les bibliothèques sont depuis longtemps déjà confrontées aux problèmes de numérotation, qu'il s'agisse des numéros attribués par les éditeurs, comme les ISBN et les ISSN, ou des cotes qui permettent de disposer et retrouver les livres dans la collection [1].

Sur le Web, les URL qui servent à localiser les ressources sont également utilisées comme référence pour les identifier et les citer, ce qui n'est pas sans poser un certain nombre de problèmes : gestion de versions et de granularité, changements d'architecture et évolutions logicielles, coexistence de tous les formats et genres… Ce type de problèmes apparaît de manière cruciale lorsqu'on gère un site comme Gallica, la bibliothèque numérique de la BnF créée en 1997, et qui donne aujourd'hui accès à près de 100 000 imprimés et autant d'images numérisées. Face à dix ans d'évolution de sa bibliothèque numérique, la BnF a dû développer des stratégies de gestion fiable de collection numérique qui incluent l'identification pérenne des ressources produites et mises en ligne par l'institution. En 2006, un nouveau système d'identifiants pérennes, les ARK (Archival Resource Key) a été mis en place sur l'ensemble de la collection numérique. Nous relatons ici les grandes étapes du choix d'un système d'identifiants, les difficultés rencontrées et les atouts que présente aujourd'hui ce système pour la bonne gestion des collections numériques de l'établissement.

2 Les identifiants sur le Web : objectifs, fonctions, systèmes

« *Les URIs ne changent pas : Ce sont les gens qui les changent.* » Ainsi Tim Berners Lee introduit-il avec justesse la problématique de la pérennité des adresses Web, dans un texte devenu célèbre, *Cool URIs don't change* [2]. Dans cet essai fondateur, il explique par des exemples triviaux l'une des principales raisons qui font que les adresses URL attribuées aux pages Web ne peuvent être considérées comme un support pérenne pour identifier des ressources : elles contiennent *trop* d'informations qu'il est impossible de pérenniser. Des informations de localisation sur un serveur, des noms d'auteurs, des chemins d'accès basés sur une organisation à un moment donné… Jusqu'à un certain point, le problème de l'identification pérenne n'est pas un problème technique, mais un problème de bonnes pratiques qui donnent accès aux ressources. Le cadre de normalisation fourni par le W3C autour des notions d'URL, URN et URI est un bon point de départ pour aborder ces bonnes pratiques préalablement à l'adoption d'un système d'identifiants.

2.1 URL, URN et URI

Les trois notions d'identification qui sous-tendent l'architecture du Web posent souvent des problèmes d'ambiguïté car elles ont évolué avec le temps. Aujourd'hui clarifiée par une spécification du W3C [3], leur relation peut être comparée à un annuaire téléphonique : si l'on ne connaît une personne que par son adresse, et que cette personne déménage, on n'a plus de moyens de la retrouver. C'est le problème qui s'est posé dès les années 1990 avec les URL, Uniform Resource Locator, qui identifiaient les pages Web en utilisant leur emplacement. Les difficultés liées aux URL ont été prises en compte et une initiative visant à identifier les ressources par leur nom propre (comme une personne dans un

annuaire) s'est créée sous le nom d'URN (Uniform Resource Name). Si la définition de principe des URN [4] peut être considérée comme un guide pour la mise en œuvre d'identifiants pérennes, les URN ne se confondent pas avec la notion d'identifiants telle qu'on la manipule aujourd'hui. Ils sont seulement un type d'identifiants. C'est la notion d'URI (Uniform Resource Identifiers), en tant que concept abstrait englobant à la fois les URL et les URN, qui correspond à la notion d'identifiants pérennes telle que nous la présentons ici : URL et URN sont en fait des types différents d'URI. Pour poursuivre la métaphore de l'annuaire, une URI correspondrait à une entrée d'annuaire écrite suivant une syntaxe prédéterminée.

Les identifiants pérennes ont donc une syntaxe commune qui est basée sur la spécification des URI [5]. Cette syntaxe est composée de trois parties :

- un « schème », ou préfixe, qui indique le contexte dans lequel l'identifiant est attribué (par ex. *http:*, *ftp:*, *urn:*, etc.)
- un élément qui permet de désigner l'« autorité nommante » qui a attribué l'identifiant au sein de ce système ; cette autorité est désignée par une convention définie au sein du système (un nom de domaine dans le cas d'une URL),
- enfin, le « nom » lui-même, c'est-à-dire une chaîne de caractères qui identifie la ressource de manière unique, au sein de ce système et pour cette autorité.

Le degré de normalisation de chacun de ces trois éléments est un élément clé de l'unicité des identifiants et de leur pérennité. Les schèmes sont enregistrés auprès de l'IANA[1], ce qui garantit leur unicité : un schème qui ne serait pas enregistré risquerait d'être utilisé par différentes communautés ou à différentes époques sans que l'on s'en rende compte, ce qui menacerait l'unicité des identifiants. URN fait partie des schèmes enregistrés à l'IANA.

Une URL est un type d'URI actionnable sur le Web, c'est-à-dire qu'on dispose d'une méthode pour accéder directement à la ressource en utilisant par exemple le protocole 'http'. La plupart des autres schèmes, et parmi eux le schème URN, ne sont pas actionnables directement par un navigateur et différentes méthodes doivent être utilisées comme la mise en place de greffons/plugs-ins que l'utilisateur ajoute lui-même à son navigateur, ou de résolveurs. Le rôle du résolveur d'identifiants est de faire correspondre au nom de la ressource son adresse réelle. Le résolveur peut être interne à l'institution qui donne les noms, ou externe et géré par une autorité indépendante[2].

La désignation des autorités nommantes est propre à chaque système, mais le nom lui-même est souvent laissé à la discrétion de l'autorité nommante : c'est justement cette liberté qui est à l'origine des difficultés pointées dans le texte de Tim Berners Lee. Alors que la pérennité des schèmes est garantie par leur enregistrement à l'IANA, et celle des autorités nommante par leur gestion centralisée dans le cadre d'un schème donné, les noms eux-mêmes ne font généralement l'objet d'aucune régulation. La définition de règles de bonnes pratiques est donc une étape primordiale dans la mise en place des identifiants.

2.2 Identifiants opaques et signifiants

La citabilité d'une ressource peut être facilitée par le choix d'identifiants dits sémantiques ou signifiants, c'est-à-dire qui portent en eux-mêmes du sens. Ce sens repose en général sur les métadonnées de la ressource qu'ils décrivent, par exemple son titre. Ce type d'identifiants a donc été assez largement adopté sur le Web, notamment par les outils de gestion de plates-formes de blogs. Cependant, ces identifiants signifiants, s'ils facilitent la compréhension et le référencement de la ressource, peuvent poser un certain nombre de problèmes. Tout d'abord, si la nature de la ressource change, le lien sémantique entre l'identifiant et la ressource peut être brisé. De plus, les identifiants signifiants sont profondément liés à la structure des documents qu'ils décrivent, ce qui peut rapidement dans le cas d'une collection de masse poser des problèmes de cohérence et de généricité. Enfin, d'une langue à l'autre, d'une époque à l'autre, la signification d'un mot ou d'un sigle peut changer, elle peut être explicite à une époque et incompréhensible un siècle plus tard, elle peut être anodine à un moment donné et devenir offensante en moins d'une décennie [6]. Lorsqu'on travaille sur le très long terme et à l'échelle internationale, il peut donc être important de réfléchir à ces contingences et d'envisager le choix d'un système de nommage opaque.

Les identifiants opaques sont en principe générés par des machines à l'aide de logiciels. Parmi les normes existantes, nous pouvons citer UUID[3] (Universally Unique Identifier), un identifiant construit par un algorithme normalisé sur la base d'informations techniques (quel ordinateur le génère) et temporelles (à quel instant précis on le génère). De tels identifiants soulèvent une toute autre difficulté : la nécessité de préserver un lien entre l'identifiant et la ressource décrite, puisque l'identifiant ne porte pas, en lui-même, d'informations sur le contenu de cette ressource. Par ailleurs, l'automatisation des UUID a permis à des logiciels de générer des identifiants non plus sur 16 bits, mais sur 128 bits ou plus, ce qui les rend impossibles à exploiter pour des utilisateurs humains. Le système NOID[4] (Nice Opaque Identifiers) a été développé par la California Digital Library pour permettre de générer facilement des identifiants opaques compatibles avec n'importe quel type d'URI.

---

[1] Internet Assigned Numbers Authority. La liste des schèmes enregistrés est accessible en ligne : http://www.iana.org/assignments/uri-schemes.html

[2] Ex. : la Library of Congress dispose de son résolveur qui fonctionne pour les identifiants DOI et Handle. On dispose d'un identifiant comme : doi:10.1045/january2005-fox, et il suffit de le faire précéder par l'adresse du résolveur pour accéder à la ressource : http://hdl.loc.gov/doi:10.1045/january2005-fox. Il s'agit d'une simple fonction de résolution ; la ressource prise en exemple n'a aucun rapport avec la Library of Congress.

[3] *A Universally Unique IDentifier (UUID) URN Namespace*, RFC n° 4122 [en ligne] <http://www.ietf.org/rfc/rfc4122.txt> consulté le 5 mai 2006

[4] http://www.cdlib.org/inside/diglib/noid/

L'expérience a toutefois montré que dans la plupart des cas, une combinaison d'éléments signifiants et opaques pouvait être utilisée de façon à atteindre les objectifs visés. Si on prend l'exemple des URL, c'est le choix d'identifier les autorités nommantes de façon signifiante, par des noms de domaines, qui a permis le succès grand public du Web (une adresse IP étant beaucoup plus difficile à retenir qu'un nom de domaine). Toutefois, les noms de domaines ne sont pas exempts de problèmes et ils représentent une difficulté majeure lorsqu'il s'agit de pérenniser des identifiants sur le long terme, parce qu'une institution ou une entreprise peuvent changer de nom, entraînant généralement avec elles leur nom de domaine.

Globalement, l'usage d'identifiants signifiants facilite la citabilité et le référencement, mais fragilise la pérennisation. Le choix de l'une ou l'autre méthode dépend donc des priorités que l'on se fixe en termes de fonctionnalités.

2.3 Fonctionnalités des identifiants

Les fonctionnalités des identifiants ont été définies dans la spécification des URN [4] dès 1994. Aujourd'hui encore, elles ne sont pas remises en cause en tant que telles ; c'est plutôt la manière de les mettre en œuvre qui fait débat. On peut les regrouper en deux grands ensembles : les fonctionnalités organisationnelles d'une part, et les fonctionnalités techniques d'autre part.

Les fonctionnalités organisationnelles des identifiants pérennes relèvent avant tout de bonnes pratiques et de coordination locale et internationale : ce sont l'unicité globale et la pérennité. Un identifiant est supposé caractériser une ressource et une seule ; en retour, la même ressource, même située à différents endroits, devrait avoir le même identifiant. On parle alors d'identifiants « globalement uniques », ce qui peut supposer une organisation plus ou moins centralisée à l'échelle internationale. La pérennité est la clef de la stabilité de la référence et la principale problématique de l'utilisation des identifiants : il s'agit de garantir qu'ils ne changeront pas et continueront à identifier la même ressource. Cependant, il est désormais reconnu que la pérennité n'est en rien un problème technique, et qu'elle n'est garantie par aucun système connu d'identification[5]. C'est la gouvernance qui assure la pérennité, c'est-à-dire qu'elle est garantie par la pérennité de l'institution qui donne les identifiants. Les institutions ou acteurs appelés à durer, qui deviennent des autorités nommantes à l'échelle d'une organisation locale, d'un pays, ou du monde, sont dès lors détentrices d'un pouvoir, celui de nommer, mais aussi d'une lourde responsabilité, celle de pérenniser leurs identifiants.

L'indépendance de l'autorité nommante doit dès lors être discutée : pour une institution, il peut être raisonnable de déclarer que le système le plus pérenne est celui qui lui impose le moins de contraintes, ou celui dont les contraintes correspondent le mieux aux besoins de l'institution. Ainsi un établissement comme la Bibliothèque nationale de France accorde une grande valeur à l'indépendance à la fois technique et budgétaire de son système d'identifiants, car cette indépendance lui garantit une liberté de mise en oeuvre qui va être favorable, dans le contexte d'un tel établissement, à la pérennité. En revanche, un petit éditeur pourra préférer un système très contraint, car cette contrainte va lui apporter un confort technique (en lui fournissant des outils) et une sécurité globale (en proposant par exemple un système de continuité si l'éditeur disparaît) qui sont indispensables à une véritable pérennité. Par ailleurs, la pérennité comme l'unicité posent un problème d'échelle. Il est désormais admis que des identifiants peuvent être réaffectés, voire détruits sur le réseau. La pérennité ne se définit donc pas par « éternellement », mais par « suffisamment longtemps » à l'échelle des besoins de l'institution qui gère la ressource. Pour prendre un exemple, un identifiant comme http://www.lemonde.fr/ est tout à fait stable mais son contenu change quotidiennement. Peut-on alors parler d'identifiant pérenne ? Une distinction est à faire entre des ressources « abstraites », éventuellement mouvantes, et des ressources « concrètes », stables et uniques, toutes deux ayant besoin d'être identifiées.

Du point de vue des fonctionnalités techniques, les systèmes d'identifiants doivent se montrer aptes à répondre aux besoins spécifiques qui ont été définis par chaque producteur en fonction de ses propres moyens, et peuvent relever comme nous l'avons souligné soit de la citabilité, soit de la pérennité, soit des deux. Ces fonctionnalités portent sur la capacité des identifiants à s'adapter à des systèmes existants et aux ressources qu'ils décrivent, quelle que soit leur nature, leur quantité et leur évolution dans le temps. Ainsi, les identifiants doivent être applicables à n'importe quel niveau de granularité de la ressource : la ressource elle-même mais aussi la collection dont elle fait partie, les articles qu'elle rassemble, etc., ainsi que différentes versions d'une même ressource. Comme dans toute gestion de collection numérique, un choix initial est nécessaire entre granularité logique et physique. Il faut donc définir les différents niveaux de granularité de l'information qui doivent être identifiés, et comment ils vont se décliner dans le système d'identification : le choix peut aller de l'attribution d'identifiants complètement indépendants à chaque niveau (mais il faut alors gérer des liens entre ces niveaux d'identifiants, grâce à une carte de structure), jusqu'à un système hiérarchisé qui reflète l'organisation de la collection. Les identifiants doivent être capables d'intégrer des modèles préexistants pour le fournisseur qui les utilise : par exemple, les ISBN et ISSN[6], les cotes d'une bibliothèque, un système de nommage préexistant utilisé pour des URL ou des fichiers. La pérennité repose en outre sur la capacité à s'adapter aux changements de l'environnement, et il est nécessaire de pouvoir étendre les identifiants et les adapter au fur et à mesure de l'apparition de nouvelles ressources, des évolutions du réseau, des standards du Web, des capacités des navigateurs. Enfin le contexte de l'identifiant pérenne doit permettre de savoir à quoi celui-ci correspond et

[5] Voir notamment [7], [8] et [9] pour retrouver l'historique des débats débouchant sur ce constat sur la pérennité.

[6] Systèmes de référence utilisés pour identifier les livres et les revues.

d'accéder à la ressource elle-même. Soit on dispose d'autres informations sur la ressource, incluant éventuellement sa localisation, sous forme de métadonnées, soit on consulte une base de référence qui va donner l'adresse correspondant à l'identifiant, en passant par un service de résolution. Une solution n'exclut pas l'autre ; on peut avoir un identifiant associé à des métadonnées et en plus un résolveur qui va donner l'URL correspondante. En outre, comme nous l'avons déjà observé, les identifiants qui se basent sur des URL sont directement actionnables dans un navigateur Web.

Toutes ces fonctionnalités constituent un cadre complexe et parfois contradictoire. Les identifiants les plus pérennes ne sont pas forcément les plus granulaires ou les plus faciles à rendre actionnables. Les identifiants les plus performants techniquement ne sont pas nécessairement les plus faciles à pérenniser. Autour de ces différentes fonctionnalités, des systèmes se sont mis en place, visant à faciliter la mise en œuvre d'identifiants globalement uniques. En vue de choisir on système, la BnF a procédé à une revue comparative des différents systèmes, dont nous présentons ici les conclusions.

2.4 Les systèmes d'identifiants pérennes aujourd'hui

Les systèmes d'identifiants pérennes actuels reposent sur l'existence d'une autorité nommante mondialement reconnue, qui dispose de la liberté et de l'indépendance nécessaires pour attribuer à des ressources des identifiants pérennes, uniques, et adaptables, que les navigateurs interprètent soit directement soit à l'aide d'un résolveur. Tous reposent sur la syntaxe des URI.

Une première catégorie d'identifiants pérennes a été créée dans la mouvance des URN pour pallier aux insuffisances des URL. Ils ont pour principe d'adopter un autre schème que 'http:' puis de définir une organisation pour assigner des autorités nommantes à l'intérieur de ce schème. Le système URN en est un bon exemple : le schème 'urn:' est enregistré comme schème d'URI à l'IANA. L'URN contient ensuite un identifiant d'espace de nom (Namespace Identifier) dont l'attribution est gérée par l'IETF[7] : cela correspond au codage d'autorités nommantes comme par exemple l'ISBN dans l'identifiant suivant : <urn:isbn:0747591059>. Ce type de système d'identifiants est purement organisationnel et ne correspond pas à une implémentation technique. Tout comme le schème 'info:'[8], créé pour garantir l'unicité globale de systèmes d'identifiants existants qui ne sont pas enregistrés auprès de l'IANA comme schèmes URI[9], ils visent à stabiliser les identifiants dans le temps, et non à les rendre actionnables.

Au contraire, un système tel que le DOI (Digital object identifier) combine des fonctionnalités organisationnelles et des fonctionnalités techniques. Pour l'organisationnel, il a été créé une entité l'International DOI Foundation qui a la responsabilité de désigner les autorités nommantes et de maintenir les règles de bonnes pratiques de nommage. L'adhésion au système et l'attribution des identifiants sont payantes, mais en retour l'organisation peut garantir leur pérennité au-delà de la durée de vie du producteur. Du point de vue technique, le DOI utilise une suite logicielle de gestion d'identifiants nommée Handle[10], qui fournit le mécanisme de résolution des identifiants. Grâce à Handle, les identifiants sont donc actionnables en passant par un proxy ou en installant un plug-in. Tout résolveur Handle sera capable de résoudre aussi bien des identifiants Handle et DOI locaux que ceux d'autres institutions, car le système est globalement cohérent. Le DOI est surtout répandu dans le domaine de l'édition électronique.

D'autres systèmes actionnables sont directement basés sur les URL et le protocole http comme le système PURL (Persistent URL) d'OCLC[11]. Un système comme PURL consiste schématiquement à maintenir une table d'équivalence entre des identifiants (PURL ID) et des adresses URL de localisation réelle. Le résolveur utilise une redirection http pour faire suivre la requête du PURL ID jusqu'à l'emplacement réel de la ressource. Les utilisateurs du système PURL ont le choix entre l'utilisation du résolveur d'OCLC, ou l'installation d'un résolveur local : c'est cette dernière option qui a été choisie par la bibliothèque nationale du Portugal pour sa bibliothèque numérique[12].

Tous ces systèmes présentent différents avantages et inconvénients, mais aucun ne concourt réellement à garantir la pérennité elle-même du lien qui relie l'identifiant à la ressource : les URN ne sont pas actionnables, les DOI et les PURL nécessitent la mise à jour d'une table de correspondance entre l'identifiant et la localisation. C'est donc l'institution nommante qui doit assumer la maintenance de ce lien. Ce constat a conduit certaines grandes bibliothèques à faire marche arrière quant au choix d'un de ces systèmes et à opter pour des « URL gérées » (*managed URLs*), c'est-à-dire des URL simples pour lesquelles un effort de conception et de pérennisation technique est garanti par l'institution. Ainsi la National Library of Australia a développé son propre système d'identifiants pérennes, sous la forme d'URL signifiantes qui décrivent la granularité de ses collections numériques [10]. Ce type de système présente surtout le risque de poser des problèmes d'extensibilité puisqu'à chaque nouvelle collection, à chaque nouveau type d'objet, il faut créer de nouveaux paramètres ce qui peut rapidement devenir difficile à gérer.

---

[7] Internet Engeneering Task Force, l'un des organes de normalisation du Web.

[8] http://www.ietf.org/rfc/rfc4452.txt

[9] Les règles d'enregistrement de nouveaux schèmes URI étant devenues extrêmement restrictives, très peu de systèmes d'identifiants peuvent effectivement être considérés comme des URI au sens strict du terme. 'doi:', 'ark:' etc. ne sont pas enregistrés par l'IANA. Ainsi, au sens strict de la norme, l'identifiant <doi:10.1045/july2007-rieger> n'est pas une URI valide, alors que l'identifiant <info:doi:10.1045/july2007-rieger> en est une.

[10] http://www.handle.net/

[11] http://purl.oclc.org/

[12] http://purl.pt

Confrontée à ces mêmes questions, la California Digital Library a elle aussi parié sur son propre système, s'appuyant sur les URL, mais elle en a profité pour faire un important effort de modélisation et de normalisation mis à disposition sous le nom de ARK (Archival Resource Key) [13]. C'est ce système qui a finalement fait l'objet du choix de la BnF.

3 Un système d'identifiants pour la BnF

Comparer ou évaluer les différents systèmes est une tâche complexe, dès lors que, comme nous l'avons vu, aucun n'est supérieur aux autres en termes de pérennité ; ce sont plutôt les besoins des producteurs de ressources qui sont à prendre en compte dans le choix d'un système. A la BnF, ce choix a été réalisé en prenant en compte les différentes fonctionnalités évoquées ci-dessus, et en les priorisant par rapport aux besoins de l'établissement concernant d'une part la visibilité de ses ressources sur le Web et la construction de services associés, et d'autre part la mise en place du système de gestion cohérente des documents numériques au sein d'une archive compatible OAIS[14], appelée Système d'Information numérique, en vue de leur préservation sur le très long terme.

3.1 Le choix d'un système : ARK

Le système d'identifiants pérennes devait répondre à la fois aux besoins du Système d'Information numérique et à ceux de la communication et de la consultation des documents, dans le respect de l'existant mais en préfigurant les futures extensions (intégration du Dépôt légal du Web, dépôts de masters électroniques par des partenaires, numérisation de masse…). La plus grande attention a été accordée à l'indépendance du système choisi, à la fois sur le plan budgétaire pour ne pas obliger l'établissement à s'engager dans un processus coûteux qu'il serait susceptible de vouloir abandonner un jour, et sur le plan technique afin que le système puisse être immédiatement intégré à l'architecture existante. Les systèmes payants comme le DOI, et les systèmes contraignants en termes de choix logiciels, comme Handle, ont donc été écartés d'office. Le système devait fonctionner dans l'état actuel des techniques du Web, et favoriser l'accès aux documents et la citabilité. Il convenait d'éviter aux usagers la manipulation d'outils techniques spécifiques, et de s'intégrer dans leurs pratiques documentaires sur le Web. Pour ces différentes raisons, un système basé sur URL semblait la solution la plus appropriée. Toutefois, un système de redirection tel que PURL ne semblait pas souhaitable, car l'un des enjeux de l'implémentation des identifiants était de rationaliser le système complexe de redirections qui était déjà utilisé pour gérer l'historique des évolutions d'architecture de Gallica, afin notamment d'assurer une meilleure visibilité et compréhension des contenus numériques pour les usagers et pour les robots d'indexation des moteurs de recherche. L'utilisation de PURL n'aurait fait que rajouter davantage de complexité à une situation déjà problématique. Nous avons donc envisagé d'opter, comme la National Library of Australia, pour des URL gérées qui s'afficheraient directement dans la barre d'adresse du navigateur, ce qui constitue le meilleur gage de citabilité et d'usabilité du point de vue utilisateur. Finalement, l'indépendance du système ARK et la qualité de sa modélisation technique, qui permettait d'envisager cette dernière option, nous ont conduits au choix de le mettre en place sur les collections de la bibliothèque numérique.

La particularité du système ARK réside dans deux éléments ajoutés aux trois principaux éléments de la syntaxe des URI : les autorités d'adressage (Name Mapping Autority Hostports - NMAH) et les qualifieurs. Le NMAH est l'adresse du serveur qui va résoudre l'identifiant ARK. Comparativement au rôle joué par l'autorité nommante, on pourrait parler d'autorité d'adressage. Cette partie figure dans l'URL mais ne fait pas vraiment partie de l'identifiant, elle est optionnelle et peut-être changée. De fait, il est possible d'avoir plusieurs autorités d'adressage pour un même identifiant, ce qui permet de résoudre le problème de la transitivité des noms de domaines : qu'un document soit visibles dans différents sous-sites de la BnF (par exemple, http://catalogue.bnf.fr et http://gallica.bnf.fr), qu'un sous-site change de nom, l'identifiant reste le même, seule l'autorité d'adressage change. Le qualifieur est une chaîne de caractères placée à la suite de l'identifiant qui permet de qualifier ce que l'on veut de l'objet. Il est optionnel. L'autorité nommante est libre de développer une hiérarchie et rendre visible des variantes : il devient ainsi possible de développer la granularité d'un document, de faire référence à des versions distinctes, ou d'appeler des services particuliers. Contrairement au nom lui-même, le qualifieur n'est pas soumis à une garantie de pérennité ; il peut donc être signifiant et il peut également être associé à des services susceptibles de changer d'apparence ou de disparaître.

On obtient donc pour une URL basée sur ARK la syntaxe suivante :

- l'autorité d'adressage : par exemple un nom de domaine ou de serveur comme http://gallica.bnf.fr/
- le schème : ark:/
- un code d'autorité nommante attribué de manière globale par la California Digital Library, qui maintient le système : 12148 dans le cas de la BnF
- le nom lui-même, une chaîne de caractères attribuée par la BnF
- les qualifieurs, introduits par des barres obliques ou des points, définis localement par la BnF.

3.2 Les choix techniques et fonctionnels

Le choix des identifiants ARK a été l'occasion pour les équipes chargées des évolutions de Gallica de poser un certain nombre de questionnements sur le fonctionnement de l'application et la gestion des objets numériques dans la bibliothèque en ligne. En effet, pour favoriser une

[13] Consulter la spécification du format ARK : http://www.cdlib.org/inside/diglib/ark/arkspec.pdf

[14] L'Open Archival Information System, reconnu comme norme ISO 14721, est un modèle conceptuel pour la mise en place d'un système ouvert de conservation à long terme des objets numériques.

implémentation rapide, il a été décidé de concevoir dans un premier temps le système d'identification pérenne comme une abstraction des différents types de documents et fonctions de visualisation de Gallica. De cette façon, aucune évolution profonde de l'application n'était nécessaire, mais on procédait par réécriture des URL grâce à un résolveur. Cet exercice de modélisation, portant sur la désignation de chaque entité qu'on voulait pouvoir manipuler dans le système, s'est révélé extrêmement riche d'expérience pour le développement d'une collection de documents numériques bien gérée.

La première question qui s'est posée a été celle des identifiants opaques ou signifiants. En effet, avant l'implémentation du système ARK, les documents de Gallica étaient nommés de façon partiellement signifiante par l'attribution d'un code barre dans une série différente selon qu'on avait affaire à des volumes reliés (monographies) ou à des lots d'images. Deux préfixes signifiants, NUMM (pour numérisation de monographies) et IFN (pour images fixes numérisées), désignaient dans les cotes numériques ces deux grands types documentaires. Avec le passage à la numérisation de masse et la bascule numérique[15], enclenchés simultanément avec le début de la mise en place du système d'archivage, ce principe était cependant problématique car la diversité des types de documents numérisés ne pouvait qu'être appelée à augmenter : il fallait désormais prendre en compte les documents sonores et audiovisuels notamment, mais aussi un certain nombre d'objets difficiles à classer, comme les manuscrits enluminés, qui peuvent entrer aussi bien dans la catégorie des « monographies » que dans celle des « lots d'images ». Le préfixe signifiant des identifiants, s'il était maintenu, risquait donc de conduire à la multiplication exponentielle du nombre de préfixes, ceux-ci devenant de plus en plus difficile à gérer de façon cohérente. Pour éviter cette dérive, l'aspect signifiant des identifiants de Gallica a été supprimé en remplaçant les préfixes NUMM et IFN par des suites de caractères alphanumériques non signifiantes (par exemple, « NUMM- 85329 » est devenu « bpt6k85329 »). Ainsi, le système des préfixes existe toujours dans les faits, mais n'étant plus lié à une signification, il ne nous oblige pas à rester en cohérence avec des choix documentaires devenus caducs. Tout en étant compatible avec la chaîne de numérisation actuelle, qui n'a pas eu à subir de modification, ce système permet d'envisager toutes les évolutions futures. Les combinaisons alphanumériques utilisées, au contraire des anciens codes-barres, n'ont pas un nombre de caractères fixe ce qui évite d'avoir à manipuler des suites de 0 au début de la séquence chiffrée. Elles utilisent tous les chiffres et toutes les consonnes[16] de l'alphabet avec la structure suivante :

- le premier caractère est réservé à l'identification de l'application nommante au sein de la BnF
- les caractères 2 à 5 sont réservés à la série de nommage, qui succède aux « NUMM » et aux « IFN » (707 281 séries possibles)
- l'identifiant proprement dit est attribué sur un maximum de 8 caractères
- le dernier caractère est un caractère de contrôle.

Ce système permettrait de nommer 3,5 millions de fois plus d'objets que n'en préserve la BnF dans ses collections actuelles (environ 20 millions).

La seconde question a porté sur les différentes fonctions existantes dans le visualiseur de Gallica : on peut effet, outre le feuilletage du document de page en page et directement par sélection d'un numéro de page, choisir de consulter un document avec sa notice, sa table des matières, ou son « chemin de fer » (les vignettes des pages) ou en télécharger plusieurs pages dans un fichier PDF. Toutes ces fonctions étaient appelées dans les URL par des paramètres de type : <http://visualiseur.bnf.fr/Visualiseur?Destination=Gallica&O=NUMM-85329&&E=27&M=chemindefer&Y=image>… L'enjeu de la mise en place des identifiants pérennes était de remplacer ces paramètres d'URL par des noms génériques attribués aux fonctions, qui pourraient être pérennisés même si les applications évoluaient et se comportaient de manière différente. Pour ce faire, deux types de fonctions ont été identifiés : d'une part, les structures qui correspondent à la granularité des objets numériques et sont peu susceptibles d'évoluer dans le temps, d'autre part les services qui permettent d'appeler un contexte de visualisation particulier mais dont la pérennité n'est pas garantie, ce contexte pouvant disparaître avec l'évolution de l'application. S'y ajoutent des fonctions d'appel de métadonnées et d'appel de versions différentes (un autre format, une autre résolution) qui ont été prévues mais ne sont pas encore toutes implémentées. Pour donner accès aux structures et aux services, on a utilisé le système de qualifieurs défini par le système ARK. Ainsi, des qualifieurs de structures introduits par une barre oblique permettent d'appeler une page précise à l'intérieur d'un livre ou d'un lot d'images, tandis que des qualifieurs de services, introduits par un point à la suite de l'identifiant et éventuellement du qualifieur de structure, désignent le contexte de visualisation. Pour ces derniers, ont été retenus des qualifieurs signifiants, plus faciles à manipuler et plus fiables en termes de citabilité ; dans la mesure où on ne s'engage pas sur leur pérennité, cela ne posait pas problème majeur. Les nouvelles URL de Gallica ont donc une forme de type : <http://gallica.bnf.fr/ark:/12148/bpt6k85329c/f27.cheminddefer>. On retrouve la dualité, évoquée plus haut, entre un identifiant abstrait qui est parfaitement pérenne et pointe vers un objet logique (l'identifiant ARK), et un identifiant concret qui désigne ou peut désigner une partie physiquement toujours identique de l'objet logique. Ainsi, l'identifiant <ark:/12148/bpt6k85329c> désigne l'objet logique qui correspond à la numérisation de « Wheler, George. Voyage de Dalmatie, de Grèce et du Levant » tandis que l'identifiant <http://catalogue.bnf.fr/ark:/12148/bpt6k85329c/f27.pagin

[15] La numérisation de masse, décidée avec le projet de bibliothèque numérique européenne, conduit à numériser 100 000 documents par an au lieu des 5 000 qui étaient traités annuellement pour Gallica. La « bascule numérique » désigne le fait de passer à des reproductions numériques au lieu des microfilms pour la reproduction de sauvegarde préventive. Ces deux projets survenus presque en même temps ont conduit à une croissance exponentielle, en quantité comme en diversité d'objets, de la collection numérique.

[16] Les voyelles sont exclues pour éviter le risque d'apparition de syllabes signifiantes dans les identifiants.

ation> désigne l'objet physique correspondant à la 27e image de cet objet logique, dans le contexte du catalogue et avec le service d'affichage de la pagination. De ces deux identifiants, seul le premier est pérenne ; le second contient des informations difficilement pérennisables, y compris des informations sémantiques (« catalogue », « pagination ») qui ont seulement vocation à rendre un service. Par ailleurs, les anciennes URL continuent temporairement à fonctionner de façon à ne pas briser la continuité de service pour les utilisateurs qui les auraient référencées avant la mise en place des ARK.

C'est au moment des évolutions logicielles que tout l'intérêt des identifiants pérennes est mis au jour, et en ce sens, l'expérience de la mise en ligne d'Europeana[17] a été particulièrement intéressante. Europeana était en effet un site géré par la BnF, avec des documents numériques tirés de Gallica, mais sous un nom de domaine différent et avec des fonctionnalités d'accès différentes. La souplesse du système ARK nous a permis de réutiliser nos identifiants dans la nouvelle interface, seuls les qualifieurs de services étant différents. Ainsi, l'adresse <http://www.europeana.eu/ark:/12148/bpt6k2029102> et l'adresse <http://gallica.bnf.fr/ark:/12148/bpt6k2029102> donnent accès au même objet numérisé : dans le premier cas, on le visualise dans l'interface d'Europeana en PNG avec une fonction de recherche plein texte, et dans le second cas, on voit uniquement le mode image, en PDF dans l'interface de Gallica. Cet exemple montre bien le progrès que constitue l'implémentation des identifiants ARK pour la bonne gestion d'une collection numérique : la ressource numérique elle-même n'existe que sous une seule désignation, indépendamment des usages qu'on en fait, qui peuvent varier dans le temps suivant les besoins et les usages.

L'approche choisie pour les identifiants ARK ainsi constitués a déjà fait ses preuves en termes d'extensibilité, puisqu'il a été possible de l'appliquer à un type d'objets complètement différent : les notices bibliographiques. En effet, il était impossible de faire référence à une notice bibliographique du catalogue Bn-Opale Plus, en raison de la gestion des sessions qui donnent accès à ce catalogue. Or, pour améliorer la visibilité des notices sur le Web, il a été décidé d'exposer une version abrégée de ces notices en utilisant le protocole OAI-PMH ; dès lors, on avait besoin de pouvoir créer un lien pérenne entre les notices abrégées et les notices complètes. Le système ARK a donc été mis en place sur le catalogue, considérant chaque notice bibliographique comme une ressource à identifier. Une autre application nommante a été créée (ces identifiants commencent par « c » et non par « b ») et des règles indépendantes pour les qualifieurs ont été mises en place. Les noms eux-mêmes sont construits à partir des numéros de notices. En dehors de l'exposition des notices en OAI-PMH, l'implémentation des ARK sur le catalogue a prouvé son utilité, facilitant la citabilité des notices dans la vie quotidienne des agents et du public. Il est prévu d'étendre l'utilisation des ARK aux notices d'autorité, et à d'autres catalogues comme la base iconographique d'enluminures Mandragore, ou encore le catalogue en EAD des Archives et Manuscrits.

Le fait d'avoir identifié les notices bibliographiques nous a également permis de résoudre un problème qui était récurrent dans Gallica : celui de la navigation dans les périodiques. En effet, chaque volume de périodique, étant numérisé séparément, était traité comme une monographie avec un identifiant NUMM : les différents volumes d'une même revue n'avaient pas d'autre lien entre eux que leur rattachement à une même notice bibliographique. Il était donc possible, à partir de la notice, d'avoir une vue d'ensemble des volumes disponibles ; mais pas, à partir d'un volume, de passer directement au suivant sans repasser par la notice. En utilisant les identifiants ARK des notices de périodiques et en leur attribuant des qualifieurs spécifiques, nous avons rendu possible la navigation de date en date, de fascicule en fascicule à l'intérieur d'un titre, et la navigation non linéaire dans un titre (à partir de n'importe quel numéro, on peut revenir au calendrier de l'année ou à la liste des années disponibles). Cette fonctionnalité, principalement développée pour la presse numérisée à l'origine, a constitué un net progrès dans la présentation des périodiques de Gallica[18].

3.3 Perspectives : les identifiants pérennes dans le Système d'Information numérique

L'opération de rationalisation de l'ensemble des applications qui permettent de produire, manipuler et donner accès à des documents numériques à la BnF, démarrée simultanément avec le projet d'archivage à long terme, fait une place de choix aux identifiants pérennes. En effet, les objets numériques appelés à intégrer le SI numérique de la BnF en vue de l'archivage à long terme sont d'une grande diversité. Les processus de collecte ou de numérisation eux-mêmes sont variés : il existe trois chaînes de numérisation (une pour les livres, une pour les images et une pour l'audiovisuel) auxquelles on ajoutera le dépôt légal du Web et la production documentaire interne (*records management*). Chacun de ces processus dispose nécessairement de ses propres identifiants de production[19], qui logiquement sont tous différents puisqu'ils répondent à des besoins différents. Au moment où tous ces objets intègreront le SI numérique, il deviendra impossible de gérer ces différents identifiants ; ils seront donc enregistrés dans les métadonnées tandis que le système attribuera un nouvel identifiant ARK qui se trouvera dès lors au centre du système. Cet identifiant sera chargé de faire le lien entre les métadonnées et les objets eux-mêmes, et grâce au système de qualifieurs d'ARK, il est également possible de gérer à ce niveau la granularité des objets. C'est en utilisant ces identifiants que les différents modules du système pourront échanger entre eux des informations

[17] Prototype pour la contribution française à une bibliothèque numérique européenne, réalisé par la BnF et mis à disposition du public en mars 2007 à l'adresse : http://www.europeana.eu.

[18] Par exemple, voici un identifiant qui permet de naviguer dans le quotidien *Le Temps*, numérisé des origines à 1935 : http://gallica.bnf.fr/ark:/12148/cb34431794k/date.

[19] Par exemple : des ISSN, des ISBN, des cotes de bibliothèques, des URI, des DOI, des identifiants automatiques de type UID, etc.

concernant les ressources et faciliter leur manipulation. Pour prendre un exemple, le module chargé de gérer les droits d'accès aux ressources numériques est un système expert, au sein du système d'archivage, qui va calculer une licence d'usage pour chaque document. Pour calculer cette licence, il utilisera un certain nombre de métadonnées : celles-ci seront stockées à part dans une base de données. Pour faire le lien entre les objets stockés et leurs métadonnées de droits conservées à l'extérieur du système, on utilisera les identifiants ARK. De même, pour indiquer à quel document appartient la licence d'usage, et indiquer à l'application d'accès quelles restrictions elle doit apporter à la communication, c'est l'identifiant qui servira de référence. La granularité qu'on aura besoin de gérer pourra être très fine ; nous connaissons déjà des cas pour lesquels il sera nécessaire de savoir masquer uniquement les images, ou certaines images, à l'intérieur d'un journal ou d'une thèse. Pour cela, il sera également nécessaire de savoir les identifier : on envisage donc d'étendre le système des qualifieurs de structure à un degré plus fin que la page, en se basant sur des coordonnées à l'intérieur d'une image. Une telle fonctionnalité serait précieuse pour construire certaines applications d'accès, telles que les interfaces d'annotation collaborative qui ont été envisagées dans le cadre d'Europeana : on ouvrirait aux utilisateurs la possibilité de sélectionner une portion d'image pour la citer, l'annoter, la commenter ou la signaler à un autre utilisateur. Sur des documents de très grand format, comme la presse ou certaines cartes, il n'y a pas à douter que cette possibilité sera très appréciée.

Conclusion

A l'heure où le Web est en train de connaître des évolutions profondes notamment du point de vue des usages, avec ce que l'on a appelé le Web 2.0, l'utilisation des identifiants pérennes révèle son importance mais aussi son potentiel. Les lecteurs des bibliothèques, les enseignants, les chercheurs, les étudiants, ne sont plus passifs devant la collection numérique comme ils l'étaient il y a quelques années. Leurs pratiques changent en profondeur : alors qu'autrefois ils téléchargeaient massivement les ouvrages de Gallica sur leur disque dur, aujourd'hui ils prennent l'habitude de déplacer leur pratique vers le Web où ils effectuent directement un certain nombre d'opérations de lecture et d'annotation. Les utilisateurs du Web 2.0 sont des internautes experts, hautement familiarisés avec l'hypertexte, qui ont besoin de pouvoir citer des documents dans des articles en ligne ou des sites Web : dans ce contexte, l'URL est la « monnaie » du Web, ce qui donne de la valeur aux ressources [11]. Dans ces conditions, le fait de bien gérer les identifiants de ses ressources numériques est plus important que jamais : de plus en plus, une ressource qui ne peut être citée perd toute visibilité. Devant nous, les évolutions technologiques à venir renforcent cette tendance : les technologies du Web sémantique telles que le RDF et les ontologies reposent intégralement sur la notion d'URI, celles-ci n'identifiant plus seulement les ressources, mais également les personnes, les concepts, etc. Dans cet univers, ne pas avoir d'identifiant pérenne équivaudra à ne pas exister.

Bibliographie

[1] Catherine Lupovici, Le Digital Object Identifier : Le système du DOI. *Bulletin des Bibliothèques de France*, 3 : 49-54, 1998 (http://bbf.enssib.fr).

[2] Tim Berners Lee, *Cool URIs don't change*. W3C, 1998 (http://www.w3.org/Provider/Style/URI). Traduction française Karl Dubost (http://www.la-grange.net/w3c/Style/URI).

[3] *URIs, URLs, and URNs: Clarifications and Recommendations 1.0*. Report from the joint W3C/IETF URI Planning Interest Group. W3C, 21 septembre 2001 (http://www.w3.org/TR/uri-clarification).

[4] *Functional Requirements for Uniform Resource Names*. W3C, 1994 (http://www.w3.org/Addressing/rfc1737.txt).

[5] *Uniform Resource Identifier (URI): Generic Syntax*. W3C, janvier 2005 (http://www.ietf.org/rfc/rfc3986.txt).

[6] John Kunze, *Towards Electronic persistence using ARK identifiers*. California Digital Library, 2003 (http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf).

[7] *Persistent Identifiers*. University College Cork, Ireland, June 17-18, 2004 (http://www.erpanet.org/events/2004/cork/Cork%20Report.pdf).

[8] *DCC Workshop on Persistent Identifiers*. University of Glasgow, 30 June – 1 July 2005 (http://www.dcc.ac.uk/events/pi-2005/).

[9] *NISO digital identifiers roundtable*. US National Library of Medicine, Bethesda, Maryland, mars 2006 (http://www.niso.org/news/events_workshops/ID-06-wkshp.html).

[10] Diana Dack, *Persistent identification systems*. National Library of Australia, 2001 (http://www.nla.gov.au/initiatives/persistence/PIcontents.html).

[11] Lorcan Dempsey, An addressable knowledge base. *On libraries, services and networks*, 30 mars 2006 (http://orweblog.oclc.org/archives/000984.html).