Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère Culture Communication

IDENTIFIANTS PÉRENNES
POUR LES RESSOURCES CULTURELLES

VADE-MECUM POUR LES PRODUCTEURS DE DONNÉES

VERSION 1.0

Ministère de la Culture et de la Communication
Stratégie « Métadonnées culturelles et transition Web 3.0 »

Publié en décembre 2015

A PROPOS DE CE DOCUMENT

Ce vade-mecum a été rédigé dans le cadre de la feuille de route stratégique « *Métadonnées culturelles et transition web 3.0* » du Ministère de la culture et de la communication par les membres du groupe de travail spécialisé n°1 : *Identifiants pérennes pour les ressources culturelles*.

Membres du groupe : Roselyne Aliacar, Emmanuelle Bermès, Katell Briatte, Francisca Maria Cabrera, Jean Davoigneau, Sébastien Peyrard, Matthieu Rivallin, Frédéric Rolland, Claire Sibille-De Grimouard, Louis Vignaud.

Il a été validé dans la présente version le 24 novembre 2014.

A QUI S'ADRESSE CE DOCUMENT ?

Ce document s'adresse à tous les producteurs de données du secteur culturel (services du Ministère de la culture et de la communication, établissements publics, administrations territoriales, associations etc.) et même au-delà. On entend par producteur de données toute personne ou organisation chargée de la création, de l'alimentation ou de la gestion d'une base de données ou d'un service en ligne décrivant, à l'aide de métadonnées, des documents, des ressources, des contenus, que ceux-ci soient disponibles sous forme numérique ou analogique.

L'attribution d'identifiants uniques et pérennes aux entités que l'on décrit, quelle que soit leur nature, est absolument nécessaire pour garantir la bonne gestion, l'accessibilité et la réutilisabilité des données et des métadonnées que l'on produit.

L'objectif du vade-mecum *Identifiants pérennes pour les ressources culturelles* est de guider les producteurs de données, à partir de douze questions simples, dans la mise en place de ces identifiants.

1. QU'EST-CE QU'UN IDENTIFIANT PÉRENNE ?

Un identifiant pérenne est une chaîne de caractères alphanumérique qui a pour fonction d'identifier de manière stable un document, une ressource ou une entité quelle que soit sa nature.

La plupart des producteurs de données gèrent déjà des systèmes d'identifiants. Par exemple, les identifiants peuvent être des cotes issues du cadre de classement d'un service d'archives ou d'une bibliothèque, mais aussi des identifiants de fichiers numériques, des identifiants d'enregistrements dans des bases de données, ou encore des URL de pages web.

Dans le web sémantique, toute ressource doit bénéficier d'un identifiant qui respecte une syntaxe particulière et qui peut être utilisé pour accéder à la ressource. Ces identifiants sont les **URI (Uniform Ressource Identifier - Identifiant uniforme de ressource)**. Ils partagent la même syntaxe que les URL mais avec, de surcroît, une exigence forte de pérennité. *Sans URI, aucune ressource ne peut être produite ou utilisée dans le web des données liées.*

On parle d'**identifiant globalement unique** lorsque cette unicité se vérifie sur l'ensemble des ressources, quelle que soit leur origine géographique ou leur nature (voir Q.3 et Q.4). C'est le cas des URI.

La **pérennité d'un identifiant** est généralement déterminée par le mode de gouvernance du système dans lequel il a été attribué, c'est-à-dire par l'organisation choisie par le producteur de données pour attribuer et maintenir ses identifiants (voir Q.2).

2. QUE DOIS-JE IDENTIFIER ?

Dans le web des données liées, on identifie soit des objets documentaires (une notice, une page HTML, une image, un site web …), soit des représentations virtuelles d'entités du monde réel : des objets physiques, des œuvres, des acteurs, des lieux, des événements, des concepts…

Pour savoir si je dois identifier un objet de manière pérenne, je dois m'interroger sur :

- **la pérennité** de la ressource ou de l'entité :

  Par exemple, on peut choisir d'identifier de manière pérenne la version finale d'un document, mais pas ses versions intermédiaires. Dans un thésaurus, on doit identifier de manière pérenne un concept (le concept de « chat ») mais pas nécessairement le terme qui le représente, susceptible de varier suivant les contextes (« chat », « minou », « Felis silvestris catus ») ou la langue (« Cat », « Katze »).

- **la réutilisabilité de la ressource**, de la portion de la ressource ou de l'entité : est-elle destinée à être réutilisée à l'intérieur d'autres ressources ou pour d'autres usages ? Est-elle susceptible d'être citée ? Constitue-t-elle un objet autonome ou potentiellement autonome ? Est-elle destinée à être exposée dans le web des données liées ? Est-elle susceptible d'être liée à d'autres données ?

  Par exemple, une même image pourra servir à illustrer un article en ligne et une notice d'une base de données ; une même personne peut à la fois être l'auteur d'une ou plusieurs œuvres, être le sujet représenté sur un tableau, être citée dans des publications.

La pérennité de la ressource conditionne sa réutilisabilité.

La granularité définit le plus petit élément d'une ressource qui respecte ces deux critères d'identification, c'est-à-dire qui est à la fois pérenne et utile. Il peut être nécessaire d'attribuer des identifiants à différents niveaux de granularité.

L'élément à identifier peut être en relation essentielle avec la ressource à laquelle il appartient (par exemple, un événement dans une série d'événements, une page dans un livre, un article dans un périodique, les noms successifs d'une organisation). On privilégiera, dans ce cas, un schéma d'identification qui rende compte de cette dépendance.

*Par exemple, avec le système ARK, il est possible d'identifier, en s'appuyant sur la même base d'identifiant. la version numérisée du livre « Wheler, George. Voyage de Dalmatie, de Grèce et du Levant : **http://gallica.bnf.fr/ark:/12148/bpt6k85329c***
et la quatrième « vue » de cette ressource:
http://gallica.bnf.fr/ark:/12148/bpt6k85329c/f4.

L'élément à identifier peut, *a contrario*, être lié de manière conjoncturelle à la ressource, par exemple une illustration, une citation, une référence bibliographique, un descripteur. Dans ce cas, le composant et la ressource à laquelle il appartient peuvent se voir attribuer des identifiants complètement différents.

*L'identifiant **http://catalogue.bnf.fr/ark:/12148/cb37367035f** se rapporte à la notice bibliographique de l'édition 1857 des « Fleurs du Mal » tandis que http://gallica.bnf.fr/ark:/12148/bpt6k1057740n se rapporte au document numérisé de cette édition. Ces deux identifiants sont liés l'un à l'autre fonctionnellement mais ils sont complètement distincts.*

Le schéma d'identification doit être **extensible**, de manière à permettre de répondre à de nouveaux besoins, par exemple de nouveaux objets à identifier ou une exigence de granularité plus fine des contenus.

3. PUIS-JE RÉUTILISER DES IDENTIFIANTS EXISTANTS ?

La définition et la mise en oeuvre d'une politique d'identification pérenne passent d'abord par une analyse des pratiques existantes au sein de mon institution ou organisation.

Quels sont les identifiants déjà disponibles pour mes ressources ? Quelles entités identifient-ils ? Sont-ils complets ? Toutes les entités que j'ai besoin d'identifier ont-elles déjà un identifiant ? Enfin et surtout, mes identifiants sont-ils **localement uniques** ? Si mes identifiants sont localement uniques, c'est-à-dire si je peux garantir leur unicité à l'intérieur d'un domaine que je maîtrise, je pourrai les réutiliser pour construire des identifiants globalement uniques.

A l'intérieur de ma structure (institution, établissement, communauté d'intérêt), existe-t-il un **plan de nommage** des ressources qui permette d'éliminer tout risque de conflit entre deux identifiants, même s'ils sont issus d'applications métiers différentes ? Si oui, je dispose déjà d'identifiants localement uniques ; il est donc recommandé de s'intégrer à cette organisation et de respecter ses préconisations.

*Il existe par exemple des **règles de nommage** de fichiers XML/EAD pour des inventaires d'archives :*
L'identifiant du fichier FR-FRAD000_APE_0008_1M comprend différents éléments séparés par des tirets bas (underscore) :
- *FR-FRAD000 : code institutionnel du service d'archives (AD pour Archives départementales, AR pour archives régionales, AC pour Archives municipales)*
- *APE pour les fichiers XML au format apeEAD (format pivot du portail européen des archives)*
- *code opération (sur quatre chiffres) interne au service d'archives*
- *identifiant de la série correspondant à l'instrument de recherche encodé*

Si mon application métier produit déjà des identifiants que je souhaite réutiliser, il sera peut-être nécessaire d'ajouter des éléments qui rendent ces identifiants uniques, non seulement à l'intérieur de l'application mais à l'intérieur de mon institution, organisation ou domaine d'activité.

Par exemple, ajouter à un identifiant alphanumérique « 99710246Z » le code du service producteur d'une ressource iconographique « IVR26 » peut suffire à transformer l'identifiant propre à la base de données en identifiant localement unique « IVR26_99710246Z ».

4. COMMENT TRANSFORMER MES IDENTIFIANTS LOCALEMENT UNIQUES EN IDENTIFIANTS GLOBALEMENT UNIQUES ?

Pour rendre globalement unique un identifiant local, il faut lui ajouter un élément globalement unique. Par exemple, l'identifiant de la ressource peut être complété par l'identifiant de **l'institution nommante**, qui à son tour doit être globalement unique.

Pour assurer l'unicité globale de l'identifiant de l'institution nommante, il existe deux possibilités :

- soit faire appel à un **organisme tiers** pour l'attribution d'un identifiant (le NAAN d'ARK, l'espace de noms IANA).

> *Par exemple, dans le système ARK, chaque institution nommante se voit attribuer un identifiant unique, le NAAN (Naming Assigning Authority Number) : « 67717 » est le NAAN du ministère de la culture, « 12148 » est le NAAN de la BnF.*
> *Cela permet de déterminer que l'identifiant* ***http://data.culture.fr/thesaurus/page/ark:/67717/T1-307***
> *a été attribué par le ministère de la culture alors que l'identifiant* ***http://data.bnf.fr/ark:/12148/cb12009547n***
> *a été généré par la BnF.*

- soit se **regrouper en association** pour gérer un annuaire commun.

> *Par exemple, pour identifier de manière unique les ressources multimédia produites par le Ministère de la culture et ses partenaires, on utilise un code producteur géré dans un annuaire national : « M5050 » pour le Centre Pompidou, « IVR41 » pour le Service Régional de l'Inventaire général du patrimoine culturel de Lorraine…*

5. MES IDENTIFIANTS DOIVENT-ILS ÊTRE OPAQUES OU SIGNIFIANTS ?

On parle d'**identifiant signifiant** ou portant de la sémantique quand la chaîne de caractères qui constitue l'identifiant est formée à partir des métadonnées qui décrivent la ressource qu'on identifie. Au contraire, un **identifiant opaque** est constitué d'une suite de lettres et de chiffres sans rapport avec le contenu de la ressource.

La question du choix entre identifiants signifiants et identifiants opaques est liée à l'utilisation qui est faite de l'identifiant.

> *Pour désigner « Manhattan », un identifiant de type* ***http://d.opencalais.com/genericHasher-1/c1008719-bca3-3c48-966a-192869743423.html****, comparé à «* ***http://dbpedia.org/resource/Manhattan*** *», peut à première vue apparaître difficile à lire, à saisir ou à interpréter par un utilisateur humain.*
> *Si l'identifiant doit être lu ou manipulé par un utilisateur humain, on aura tendance à vouloir le rendre signifiant. En revanche, si l'identifiant est principalement manipulé par des machines, la longueur de l'identifiant et son opacité importent peu.*
> *On peut aussi combiner les deux approches : par exemple, dans les identifiants* ***http://exemple.com/personne/123456*** *ou* ***http://exemple.com/geo/123456****, l'identifiant local de la ressource est opaque, mais l'URI comporte un élément signifiant, « personne » ou « geo », qui permet à un utilisateur humain de prédire la nature de la ressource référencée.*

Il faut être conscient qu'embarquer de la sémantique dans l'identifiant expose à être confronté aux questions d'ambiguïté de la langue. Il sera éventuellement nécessaire de faire porter par l'identifiant des éléments de désambiguïsation pour en garantir l'unicité.

> *Par exemple, il est nécessaire d'ajouter un élément de désambiguïsation dans l'identifiant pour distinguer l'objet « bidon » (****http://exemple.com/categorie/Bidon****) et la commune « Bidon » (****http://exemple.com/commune/Bidon****).*

On voit le type d'ambiguïté que peuvent porter des termes comme « avocat » (homonymie) ou « opéra » (polysémie). Ce problème se voit démultiplié dès lors qu'on aborde la question des noms propres, qu'il s'agisse par exemple de noms de lieux (« Montréal ») ou de personnes (« Jacques Martin »).

Par ailleurs, les possibles évolutions de la langue induisent un risque pour la pérennité.

> *Par exemple, le concept qui serait ainsi identifié : « **http://exemple.com/categorie/biodiesel** » devient d'une certaine façon obsolète, dès lors que le journal officiel du 22 juillet 2007 a préconisé l'utilisation de la forme française « biogazole » de préférence à « biodiesel ». Dans ce cas, ou bien l'identifiant du concept s'écarte du libellé du concept, ce qui réduit l'intérêt d'avoir un identifiant signifiant, ou bien on modifie l'identifiant du concept pour l'adapter à l'état de la langue, ce qui remet bien sûr en cause la pérennité de cet identifiant.*

L'avantage d'un identifiant opaque est également de fédérer les multiples expressions linguistiques (niveaux de langues ou multilinguisme) d'une même entité, évitant ainsi la prééminence d'une langue ou d'une pratique culturelle sur une autre.

> *Par exemple, le concept « déambulatoire » dans le Thésaurus de la désignation des œuvres architecturales et des espaces aménagés porte l'identifiant opaque **http://data.culture.fr/thesaurus/resource/ark:/67717/T96-146**, qui représente indifféremment la forme française « déambulatoire » ou la forme italienne « deambulatorio ».*
> *Pour désigner l'auteur latin Pétrone, l'identifiant opaque de VIAF, **http://viaf.org/viaf/95155909**, permet de fédérer les usages locaux : utilisation d'une forme traduite chez les uns (« Pétrone, 00..?-0066 » pour la BnF, « Petronio Árbitro » pour la Bibliothèque nationale d'Espagne), utilisation de la forme latine chez les autres (« Petronius Arbiter » à la Bibliothèque du Congrès aux USA ou à la Bibliothèque de la Diète au Japon).*

Enfin, que les identifiants soient opaques ou signifiants, les règles suivantes doivent être respectées :

- il ne faut jamais révéler dans l'identifiant des détails d'implémentation technique, forcément sujets à changement.

Par exemple, soit l'URI http://www.exemple.fr/index.asp?id=»123456» : si l'interrogation en langage ASP est remplacée par une interrogation en PHP, l'URI va changer.

- les extensions de fichiers peuvent être ajoutés à l'URI pour accéder à différentes représentations (formats) d'une même ressource, mais elles ne font pas partie de l'identifiant pérenne.

> *Les URI suivantes :*
> ***http://www.exemple.fr/id/123456.html***
> ***http://www.exemple.fr/id/123456.rdf***
> ***http://www.exemple.fr/id/123456.pdf***
> *concernent la représentation d'une ressource identifiée **http://www.exemple.fr/id/123456** dont le format peut changer.*

6. COMMENT TRANSFORMER MES IDENTIFIANTS EN URI UTILISABLES SUR LE WEB ?

Concrètement, une URI est la version web d'un identifiant. Ainsi, on peut rendre une ressource « déréférençable », c'est-à-dire accessible sur le web, par l'intermédiaire de son identifiant.

> *Par exemple si l'identifiant d'une ressource est « 123456 », on peut construire une URI du type **http://www.exemple.fr/id/123456** pour donner accès à cette ressource sur le web.*

Plusieurs stratégies sont possibles pour construire et gérer ses URI.

La première stratégie est la **réécriture d'URL**. De nombreux outils de publication (notamment les CMS) proposent des pages web dynamiques, c'est-à-dire que les pages qu'ils affichent sont une réponse à une requête dans une base de données propre au site. L'URL réécrite se base généralement sur un ou plusieurs éléments du contenu de la page (titre, date de création, numéro d'enregistrement, etc.).

Par exemple, une URL telle que ***http://www.exemple.fr/index.php?doc=123456&language=en*** *demande à la base de données interne au site web la version en anglais du document 123456. On peut demander au serveur de réécrire cette URL pour la rendre plus stable et indépendante de l'implémentation. Ainsi l'URI* ***http://www.exemple.fr/index.php?doc=123456&language=en*** *resterait purement interne et l'URI réécrite, par exemple* ***http://www.exemple.fr/doc/123456-en****, serait la seule visible et donc citable par les internautes.*

Cette méthode est facile à implémenter, mais la pérennité des URL réécrites est tributaire de la pérennité des éléments utilisés . Ainsi, si on expose le titre dans l'URL, toute modification du titre « cassera » l'URL. Il est donc important de choisir quels champs on réutilise et appliquer des règles de réécriture cohérente à travers le temps.

Il faut de plus indiquer explicitement aux moteurs de ne pas indexer les URL natives et de ne prendre en compte que les URL réécrites.

La redirection d'URL consiste à définir une URI destinée à la citation pérenne de la ressource, parfois appelée « permalien », généralement dans un nom de domaine prévu à cet effet, puis à effectuer une redirection vers une page temporaire.

Par exemple, une URL telle que ***http://purl.org/dc/terms*** *est toujours accessible et renvoie au même contenu (le dictionnaire de données Dublin Core terms). Mais elle redirige systématiquement vers l'URL de la version la plus récente, soit, à la date de rédaction de ce document,* ***http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms****.*

Ainsi, si on utilise un nom de domaine tiers pour construire les permaliens, on peut maintenir l'accès à la ressource même si le site change de nom de domaine. Ce choix est donc indiqué si la durée de vie du nom de domaine apparaît inférieure à la durée de vie de la ressource à laquelle on donne accès.

Cependant, l'internaute aura tendance à sauvegarder et citer l'URL vers laquelle il est redirigé. Cela ne dispense donc pas d'une réflexion sur la stabilité de ces dernières URL.

Le résolveur interne est un logiciel qui permet de faire correspondre une URL à un objet numérique par l'intermédiaire de son identifiant.

Par exemple, soit l'identifiant de ressource ***ark:/12148/cb39836284g****, renvoyant à une notice de photographie. Par l'intermédiaire du résolveur* ***http://ark.bnf.fr****, l'URI* ***http://ark.bnf.fr/ark:/12148/cb39836284g*** *redirige vers* ***http://catalogue.bnf.fr/ark:/12148/cb39836284g****, qui est l'application par défaut actuellement utilisée pour consulter cette ressource.*

Ce choix offre une grande souplesse de gestion : un résolveur interne permettra de gérer de façon souple des mises à jour de noms de domaine ou des règles de réécriture. Cependant, le déploiement d'un tel service demande un travail important d'implémentation et donc des compétences et des ressources informatiques.

7. POUR DÉFINIR MES URI, SUR QUELS SYSTÈMES PUIS-JE M'APPUYER ?

Afin d'aider les producteurs de données à résoudre le problème de l'identification de leurs ressources, des **systèmes d'identifiants pérennes** ont été créés. Ils s'appuient sur la syntaxe des URI et proposent des modèles organisationnels plus ou moins centralisés ainsi que, dans certains cas, des logiciels « clé en main ».

Quelques exemples :

- *DOI (****http://www.doi.org/****) est un système payant centralisé au niveau international, principalement utilisé par des éditeurs, qui fournit des règles de création des URI, une suite logicielle pour*

les faire fonctionner et une gouvernance assurant la reprise en cas de défaillance du producteur.
- *ARK (**http://www.bnf.fr/fr/professionnels/issn_isbn_autres_numeros/a.ark.html**) est un système centralisé au niveau international qui attribue des identifiants d'autorités nommantes uniques et fournit des règles garantissant la pérennité des URI. Il n'impose aucun logiciel particulier et est entièrement gratuit.*
- *PURL (**https://purl.oclc.org**) est un système de redirection qui permet de rediriger sur le web une URI pérenne vers une URL non pérenne.*

Plusieurs stratégies sont possibles :
- adopter le système le plus adapté à mes besoins,
- adopter le système le plus répandu dans ma communauté de métier ou de pratique,
- ne pas adopter de système et utiliser mes propres URL.

Par exemple, la BnF a choisi ARK pour bénéficier des spécifications déjà existantes de ce système. Elle a également utilisé ce système comme un outil de sensibilisation interne qui a permis de dégager des moyens pour la mise en place d'un résolveur centralisé. Elle s'est ensuite impliquée dans la maintenance du registre des autorités nommantes, ce qui constitue un engagement institutionnel supplémentaire.

Pour choisir un système, il est donc nécessaire de définir des critères d'évaluation en fonction de ses propres objectifs et priorités.

Chacun peut établir sa propre grille d'évaluation en pondérant les critères suivants :
- *je souhaite que mon système d'identifiant pérenne soit gratuit*
- *je veux être libre d'utiliser les logiciels de mon choix*
- *je veux qu'on me propose un logiciel clé en main*
- *je souhaite utiliser un système largement adopté par ma communauté*
- *je veux être totalement libre de l'attribution de mes identifiants*
- *je veux pouvoir identifier n'importe quelle ressource*
- *je préfère un système qui me guide dans l'attribution des identifiants*
- *je souhaite un système compatible avec des identifiants existants*
- *je voudrais qu'on me propose un plan de reprise*
- *etc.*

L'utilisation d'un système d'identifiants n'est pas obligatoire : on peut techniquement atteindre les mêmes résultats en gérant soi-même ses URI. L'avantage principal de cette dernière option est la simplicité : je peux utiliser de simples outils web (voir Q.6). Toutefois, cela m'impose de bien veiller à la pérennité de mes URI et m'oblige donc à faire un travail de spécification attentif au début du projet, à maintenir ces URI dans le temps et à veiller à ce que cette maintenance reste une priorité (voir Q.9).

8. QUELLE STRATÉGIE PUIS-JE METTRE EN ŒUVRE POUR DÉPLOYER MES IDENTIFIANTS PÉRENNES ?

Une fois décidé quelles ressources je souhaite identifier et de quelle manière je vais construire mes identifiants, deux stratégies de mise en œuvre se présentent :

- la stratégie « big bang » : je génère en une seule fois des identifiants pour l'ensemble de mes ressources

- l'application progressive : j'avance par sous-ensembles homogènes (application métier, type de ressource …).

Un scénario de type « big bang » peut s'avérer pertinent si l'on a choisi d'utiliser des identifiants opaques (voir Q.5) et si l'on a un grand nombre d'identifiants à attribuer. En effet, la même structure d'identifiants peut être attribuée partout de manière centralisée. Le principal risque de ce scénario est de vouloir faire table rase de l'existant et d'« abandonner » d'anciennes URI construites sur des identifiants historiques (voir Q.9).

A l'inverse, le scénario d'application progressive peut s'avérer plus adapté à des identifiants signifiants (voir Q.5), car les règles de constitution d'identifiants signifiants dépendent davantage du type de ressources considéré (on ne structurera pas de la même manière un identifiant de document numérique ou un identifiant de personne, par exemple).

Le risque du scénario d'application progressive est la part d'inconnu qui réside dans les cas limites d'homonymies, qu'il faudra lever progressivement. Une telle maintenance doit faire l'objet d'un suivi et d'une cohérence dans le temps. Le choix de ce scénario a donc un impact non négligeable en ressources humaines.
Dans tous les cas, il est fortement recommandé de constituer un échantillon de cas représentatifs des types de données à identifier, selon les différents types de documents et (potentiellement) les différents types de procédures dans lesquelles l'identifiant est amené à être attribué et diffusé.

9. QUE FAIRE DE MES ANCIENNES URI ?

Dans certains cas, la mise en place d'identifiants pérennes est réalisée alors que les ressources étaient déjà accessibles via d'anciennes URI. Pour garantir une continuité de service, il est nécessaire de mettre en place une **redirection** des anciennes URI vers les nouvelles.

Si j'ai des URI historiques à gérer, il est essentiel d'éviter que les anciennes et les nouvelles URI n'entrent en collision. Deux méthodes sont possibles :

- vérifier que les structures de ces URI ne peuvent pas coïncider,
- « protéger » les nouveaux identifiants en ajoutant un préfixe entre le nom de domaine et l'identifiant.

Lors de la mise en place du nouveau schéma d'identification, il faut porter attention aux deux problèmes suivants :

- les doublons : lorsqu'une même ressource possède 2 identifiants ou plus, il faut dédoublonner les identifiants (**fusion**), et rediriger une URI « historique » vers l'URI conservée (voir Q.6).
- les collisions : lorsque le même identifiant a été attribué à deux ressources différentes (e.g. homonymes), il faut séparer les deux identifiants (**scission**). Il est alors recommandé de conserver les identifiants et URI historiques, et de faire une page de renvoi vers les deux identifiants « post-scission » et donc leurs URI correspondantes.

De manière générale, il est beaucoup plus facile de gérer des fusions que des scissions.

10. COMMENT LIMITER LA PROLIFÉRATION DES IDENTIFIANTS ?

Il y a souvent plusieurs identifiants pour une même entité identifiée.

> *Par exemple, une personne pourra avoir un ISNI, un numéro de description dans un catalogue, un numéro de sécurité sociale, un identifiant dans l'annuaire LDAP de l'établissement, un numéro ORCID en tant que chercheur universitaire etc. De même un périodique peut avoir à la fois un ISSN, un DOI, un identifiant ARK attribué par la BnF…*

La prolifération des identifiants peut complexifier la gestion des entités identifiées et entraîner des coûts, mais correspond souvent à des besoins particuliers :

- Ces numéros identifient en réalité des notions proches, mais distinctes
  Par exemple, l'ISNI identifie l'identité publique d'une personne, tandis qu'un numéro de notice d'autorité identifie une description de cette personne.

- Certains numéros sont spécifiques à une base de données particulière, qui est la seule à les connaître.
- Ces numéros n'œuvrent pas au même niveau de granularité.
  Par exemple, un identifiant A correspond à un fonds d'archives, tandis qu'un identifiant B correspond

à une pièce particulière de ce fonds.

Lorsque je veux lier mes données sur le web de données, je vais identifier des jeux de données sources qui me proposent des entités communes — personnes, œuvres, lieux, concepts etc. — qui existent également dans mon jeu de données. Je peux alors choisir de réutiliser directement les URI du jeu de données auquel je souhaite faire référence : cela m'évite de créer et de maintenir mes propres URI pour des entités qui sont déjà identifiées par ailleurs.

Si je ne souhaite pas réutiliser ces URI (besoin de gestion interne, indépendance du système d'information ...), le web de données me fournit des mécanismes d'alignement : des moyens techniques de déclarer que des entités disposant d'URI différentes sont équivalentes ou similaires.

Par exemple :

- *la propriété owl:sameAs signifie une équivalence exacte entre les deux ressources qu'elle relie. Dans ce cas toute assertion concernant l'une de ces ressources s'applique également à l'autre.*
- *la propriété rdf:seeAlso signifie « voir aussi ».*
- *différentes propriétés sont proposées par le vocabulaire SKOS dans le cas d'équivalence entre des thésaurus ou autres référentiels : skos:exactMatch pour une équivalence exacte, skos:closeMatch pour une équivalence approximative etc.*

11. QUE SE PASSE-T-IL SI JE SUPPRIME OU MODIFIE UNE RESSOURCE ?

Dans le cas d'une URI, une suppression pure et simple de la ressource identifiée aboutit à une erreur « http 404 » (la ressource n'est pas disponible à cette adresse). Dans un tel cas, l'internaute ne peut pas savoir la raison de cette réponse, qui peut être liée à quatre cas différents :

1. Rien n'a jamais été disponible à cette adresse, l'utilisateur a entré une adresse incorrecte.

2. Une ressource a été disponible à cette adresse, mais l'institution a décidé de la retirer pour diverses raisons (droits ou confidentialité, mauvaise qualité, information périmée). Dans ce cas il serait utile de fournir des informations de base sur la ressource supprimée et la raison de sa suppression.

Au minimum, on fournit un message informant que la ressource a été supprimée et si possible, les métadonnées décrivant l'objet supprimé, accompagnées de la date et de la cause de la suppression.

3. Une ressource a été disponible à cette adresse, mais a changé d'adresse. Dans ce cas il est recommandé de mettre en place une redirection vers la nouvelle URI, afin que l'internaute soit ramené vers la nouvelle page (voir Q.6).

4. Dans le cas d'un dédoublonnage (voir Q.9), il faut rediriger la ressource supprimée vers la ressource avec laquelle elle a été fusionnée.

Même si une ressource est supprimée, il ne faut jamais ré-attribuer son identifiant à une autre ressource, car cet identifiant peut avoir été conservé par un utilisateur ou par un système client. Si l'identifiant renvoie désormais à quelque chose de complètement différent, l'utilisateur ou l'application cliente se retrouveront citer une ressource différente, sans même en avoir connaissance. Cela peut aboutir à des effets de bord variés qui peuvent diminuer la confiance des utilisateurs dans la stabilité de mes URI.

Afin d'éviter la suppression physique d'une ressource dont l'identifiant est potentiellement utilisé par un utilisateur ou une application cliente , il est parfois utile de jouer sur le statut logique de cette ressource.

Par exemple, il est recommandé de ne pas supprimer physiquement un concept potentiellement utilisé pour indexer des ressources, mais de préciser dans son statut qu'il est « obsolète » (son utilisation n'est plus conseillée) ou « prohibé » (son utilisation est désormais interdite pour indexer de nouvelles ressources).

Lorsqu'une ressource est mise à jour, je peux soit lui attribuer un nouvel identifiant, soit conserver l'identifiant existant. Cette question revient à me demander si mon identifiant concerne une ressource « abstraite » amenée

à évoluer dans le temps, ou une version particulière d'une ressource (voir Q.2).

Dans la plupart des institutions culturelles, les ressources que l'on identifie sont amenées à évoluer à divers titres, par exemple :

- *l'enrichissement de la description de l'objet par correction et/ou ajout de métadonnées,*
- *l'enrichissement de la ressource elle-même dans le cas d'un objet numérique, tels l'ajout d'OCR à un document numérisé ou la modification du fichier …*

Si l'on souhaite conserver un accès aux différents états d'une ressource donnée, il est nécessaire de définir, en complément de l'**URI abstraite** qui identifie la dernière version de la ressource, des **URI concrètes** qui identifient une version particulière de la ressource identifiée.

Par exemple, le W3C distingue la spécification HTML désignée de manière abstraite indépendamment de sa version, http://www.w3.org/TR/html/, de la version de la spécification en cours à la date de rédaction de ce document, http://www.w3.org/TR/2014/PR-html5-20140916/.
En général, l'URI abstraite renvoie vers la version la plus récente de la ressource.

12. QUELS SONT MES ENGAGEMENTS LORSQUE JE PUBLIE DES RESSOURCES IDENTIFIÉES DE MANIÈRE PÉRENNE SUR INTERNET ?

En publiant mes ressources sur Internet, je m'insère dans une toile de données décentralisée. **Je deviens fournisseur de services**. Dans la mesure du possible, je dois veiller à assurer la pérennité de mes URI, qui peuvent faire l'objet de réutilisations par des tiers dès le moment de leur publication sur le web. Mon action s'inscrit dès lors dans la disponibilité *canonique du web : 24*7*365*.

Je dois donc documenter mon système d'identifiants pour instaurer une relation de confiance avec l'utilisateur, notamment en publiant un « manifeste » explicitant ma « politique » d'identification : structure de mes identifiants, fonctions associées, fréquence de mise à jour et durée de vie pressentie pour mes données et pour mes identifiants.

Je dois enfin expliciter les changements apportés à mon système, programmer les changements majeurs pour permettre aux réutilisateurs de les anticiper, communiquer sur la date de bascule et si possible, déclencher des alertes à l'attention de mes utilisateurs.