



# Les identificateurs persistants du patrimoine culturel

Il est bien connu que les ressources de l'Internet tendent à avoir la vie courte ; leur identification et leur localisation posent des problèmes complexes qui touchent à de nombreuses questions organisationnelles et technologiques qui concernent la citation, la recherche et la préservation des ressources culturelles et scientifiques. Ceci est loin d'être un problème technique isolé : l'identification à travers le temps des objets numériques comprenant les textes, la musique, la vidéo, les images fixes, les documents scientifiques et beaucoup d'autres, reste un problème majeur qui empêche d'utiliser l'Internet actuel en tant que plateforme présentant toutes les garanties pour la recherche et la dissémination des contenus scientifiques et culturels.

## Pourquoi avons-nous besoin d'un 'Identificateur persistant' ?

La préservation à long terme, la dissémination et l'accès aux objets culturels sont aujourd'hui au cœur des missions des institutions culturelles telles que les universités, les dépôts d'archives, les musées, et les bibliothèques. L'usage des URL ne peut pas être considéré comme un moyen fiable pour résoudre tous les problèmes dus à l'instabilité structurelle des liens (par exemple un domaine indisponible) et les ressources qui leurs sont liées (relocalisation ou mise à jour). L'utilisation courante des URL comme moyen d'accès augmente le risque de perdre des documents culturels ou de sous utiliser des fonds disponibles. Dans le domaine du patrimoine culturel il est essentiel non seulement d'identifier une ressource mais aussi de lui garantir un accès permanent.

Une solution sûre est d'associer un identificateur persistant (PI) à une ressource numérique qui serait immuable, quel que soit l'endroit où la ressource se trouve.

Voici les principales étapes pour implémenter un système de PI :

- 1) Sélectionner les ressources qui nécessitent un PI
- 2) Assigner un nom à chacune des ressources et enregistrer ces créations
- 3) Relier l'IP à l'URL associée
- 4) Maintenir le répertoire qui associe les URL et les PI et qui est le garant d'un accès continu aux ressources.

La première étape est de la prérogative de chaque institution culturelle alors que les autres peuvent être déléguées à d'autres autorités afin de garantir une meilleure longévité fonctionnelle et un meilleur coût de service.

### Les conditions requises d'un système d'IP

Une institution patrimoniale et culturelle devrait choisir une infrastructure d'IP en se basant sur l'ensemble des conditions requises ci-après :

- Unicité globale
- Persistance
- Capacité à résoudre
- Fiabilité
- Habilitation
- Flexibilité
- Interopérabilité
- Coûts

### Unicité globale

Nous considérons l'identificateur comme une étiquette qui est associée à un objet dans un certain contexte. Le « contexte » comprend le type de standard utilisé pour la syntaxe du nom (par exemple URN :NBN :IT :xxx-xxxx) et l'identification de l'autorité (sous-espace de nom) qui assigne cette étiquette.

**Objet:** Toute entité ayant un intérêt dans une transaction de propriété intellectuelle est définie par des métadonnées et une terminologie tirée d'un dictionnaire de données afin de s'assurer que «ce que vous signifie est bien ce que je signifie» (interopérabilité). Les objets peuvent être physiques, numériques, ou abstraits, par exemple des personnes, des organisations, des accords, etc.

**Service de Résolution (dé référencement):** Le processus dans lequel un identifiant est l'entrée (la requête) dans un réseau de communication et reçoit en retour une sortie spécifique (une ressource, des métadonnées, etc).

**Autorité de nommage:** Une autorité indépendante qui distribue les noms et garantit leur unicité et leur rémanence. Un service de résolution de noms correspond à une autorité de nommage ; il effectue la résolution du nom. Un système distribué d'Identificateurs permanents prévoit que la responsabilité de génération et de résolution de noms peut être déléguée à d'autres organismes appelés autorités de nommage de sous domaines ; elles gèrent une portion du nom de domaine ou de l'espace.

**Espace de noms:** Un conteneur abstrait qui fournit le contexte des éléments qu'il contient et qui permet de différencier les éléments ayant le même nom (résidant dans différents espaces de noms).

**Registre:** table de référence d'association de noms entre des URN et une ou plusieurs URL.  
**Un dépôt :** L'endroit où des ressources numériques sont déposées (DSace, Fedora, Codex, etc) avec ou sans système de gestion (gestion de fichiers).

**URI:** Un Identificateur uniforme de ressources est l'ensemble générique de tous les noms/adresses et qui sont en fait de courtes chaînes de caractères qui se réfèrent à des ressources.

**URL:** Un localisateur unifié de ressources est un URI qui, en plus d'identifier une ressource, donne les moyens d'agir sur, ou d'obtenir, une représentation de la ressource en décrivant son mécanisme d'accès de base ou son 'adresse' de réseau.

## Unicité globale

Nous considérons l'identificateur comme une étiquette qui est associée à un objet dans un certain contexte. Le « contexte » comprend le type de standard utilisé pour la syntaxe du nom (par exemple URN :NBN :IT :xxx-xxxx) et l'identification de l'autorité (sous-espace de nom) qui assigne cette étiquette.

## Persistence

La persistance désigne la durée de vie permanente d'un identificateur. Il n'est pas envisageable de réassigner l'IP à d'autres ressources ou de le supprimer. Ainsi l'IP sera globalement unique pour toujours, et pourra ainsi être utilisé en tant que référence d'une ressource bien au-delà de la durée de vie de la ressource identifiée ou de l'implication de l'autorité de nommage. La persistance est évidemment un sujet spécifique dans la politique ou le fonctionnement d'une institution culturelle. La seule garantie de l'utilité et de la persistance des systèmes d'identification est l'engagement montré par les organisations qui délivrent, gèrent et résolvent les identificateurs.

## Capacité à résoudre

La capacité à résoudre se réfère à la possibilité de retrouver une ressource seulement si elle est publiée. Il est important de distinguer le concept d'identification de celui de résolution. Le choix d'espace de nom de l'identification n'implique pas nécessairement de choisir une architecture de résolution correspondante.

## Fiabilité

Pour assurer la fiabilité d'un système d'IP, deux aspects doivent être pris en compte : l'infrastructure supportant les IP doit toujours être active (redondance de service, service de back-up...) et le répertoire maintenu à jour (au travers de routines automatiques).

## Autorité

La seule garantie de l'utilité et de la persistance des systèmes d'identification est l'engagement montré par les organisations qui délivrent, gèrent et résolvent les identificateurs. Dans le domaine du patrimoine culturel, la tendance est d'utiliser les services fournis par les institutions publiques telles les bibliothèques nationales, les archives d'Etat etc. Les besoins tels que l'habilitation et la crédibilité d'un système de gestion d'IP devraient être soigneusement évalués avant d'adopter une solution.

## Flexibilité

Un système de gestion d'IP sera plus performant s'il est capable de prendre en charge les besoins spécifiques à différents types de matériaux ou de collections. Par exemple, un système de gestion d'IP devrait être capable de gérer différents niveaux de granularité parce que ce sur quoi un 'identificateur' doit pointer est très variable, selon les différents champs d'applications des utilisateurs.

## Interopérabilité

Cet aspect est fondamental si l'on veut pouvoir garantir la diffusion et l'accessibilité aux objets numériques culturels. De nombreuses techniques et conceptions existent et certaines sont adaptées à des besoins spécifiques sectoriels. Parmi les différents systèmes, l'interopérabilité doit exister au minimum au niveau du service offrant à l'utilisateur des interfaces communs et faciles. L'interopérabilité d'un système peut reposer sur l'utilisation de standards ouverts.

## Les technologies actuelles

**PURL (persistant URL):** Un PURL est un localisateur persistant et unifié de ressources. Fonctionnellement, un PURL est un URL. Cependant, au lieu de pointer directement sur une ressource de l'Internet, un PURL pointe sur un serveur intermédiaire de résolutions (transformations) et qui, en utilisant les possibilités standard de redirection du serveur web, peut donc ainsi rediriger les requêtes sollicitant des ressources possédant un identificateur persistant, vers la localisation courante du document ou de la ressource en général.

[www.purl.org](http://www.purl.org)

**URN (Nom unifié de ressources) :** Un URN est un URI qui utilise le schéma URN et qui n'implique pas la disponibilité de la ressource identifiée. Les URN sont prévus pour servir d'identificateurs persistants de ressources, localisation-indépendants et sont conçus pour faciliter la représentation des autres espaces de noms (qui partagent les propriétés de ces URN) en espaces d'URN. De plus, la syntaxe des URN fournit le moyen d'encoder les caractères dans une forme qui peut être transmise dans des protocoles existants, qui existe sur la plupart des claviers, etc.

[www.ietf.org/rfc/rfc1737.txt](http://www.ietf.org/rfc/rfc1737.txt)

**Handle system.** Le système Handle est un service de noms global et complet qui permet la résolution des noms sécurisés et leur administration à travers les réseaux tel l'Internet. Le système Handle gère des noms de code, qui sont des noms uniques pour les objets numériques et les autres ressources de l'Internet. Une autorité de nommage a la charge de créer et de maintenir ces noms de code ; l'identifiant de chaque nom de code doit être unique pour cette autorité de nommage mais n'a pas de syntaxe obligée.

[www.handle.net](http://www.handle.net)

**XRI (OASIS Identifiant étendu de ressources) :** Le but de XRI est de définir un schéma URI et un URN de nom d'espace correspondant pour les services d'arborescences distribuées, afin de permettre l'identification d'entreprises, d'applications et de ressources (y compris des personnes et des organismes), ainsi que le partage de données au travers de domaines.

[www.oasis-open.org](http://www.oasis-open.org)

**ARK (Archival Resource Key) (proposition internet de IETF) :** Le but recherché est de faciliter le nommage persistant et la recherche d'objets informationnels. Un des principes fondamentaux de l'ARK est que la persistance est uniquement un problème de service et n'est ni inhérente à un objet ni conférée à un objet par une syntaxe de nommage particulière. Un identifiant de ce système résout jusqu'à trois niveaux : la ressource, ses métadonnées et les dispositions de préservation.

[www.cdlib.org/inside:dglib:ark](http://www.cdlib.org/inside:dglib:ark)

### Les coûts

Dans le domaine du patrimoine culturel, les systèmes de gestion d'IP devraient être gratuits ou au moins à un coût supportable parce que le rôle des institutions culturelles est d'offrir un accès gratuit aux ressources à travers le temps et d'éviter ainsi un fossé numérique.

## D'autres considérations

### Granularité

La granularité fait référence à un niveau de détail auquel un identificateur persistant devra être affecté. Les contraintes de la granularité auront un impact considérable sur le système d'identification qu'une institution adoptera.

Dans certains cas, il peut être nécessaire de citer une page du web qui sert d'accès à une collection de fichiers de l'Internet, ou de citer un article de journal, un élément ou un chapitre. Chaque institution devrait évaluer si un service d'IP lui fournit le bon niveau de granularité pour ses types de ressources.

### Les IP opaques ou sémantiques

Un identificateur persistant peut ne fournir aucune information sur l'objet qu'il identifie (opaque id) dû au fait qu'il est constitué de caractères aléatoires qui n'ont pas de sémantique associée. Un identificateur opaque nécessite une routine de résolution afin de pouvoir être identifié, bien qu'il puisse comporter une signification interne (id sémantique).

Il est généralement plus facile de mémoriser et d'utiliser des identificateurs reposant sur du mnémonique que ceux qui contiennent une séquence de caractères sans signification, quoique ceci ne soit pas pertinent pour un ordinateur.

### Les versions

Chaque nouvelle version d'une ressource devra posséder un identificateur persistant distinct. Une nouvelle version peut être considérée comme un nouvel objet car son contenu ou son format peut avoir changé. Les différentes versions peuvent être gérées par des règles de nommage ou à l'aide de métadonnées.

## Comment les technologies peuvent elles nous aider ?

L'application de gestion des IP nécessite une base de données qui puisse garder trace de la localisation courante d'un objet digital, appelée 'base de données de résolution'. Une telle base de données 'donne' la localisation de la ressource et redirige l'utilisateur vers cette localisation courante. Une telle base de données et sa routine de résolution peuvent être implémentées de différentes façons : centralisée ou distribuée, basée ou non sur les DNS.

**Centralisée:** Cette architecture repose sur un point central qui génère les noms des ressources et qui assure leurs transformations et leur fiabilité dans la durée. Cette solution implique une centralisation des responsabilités et des coûts de gestion. Par conséquent, un service de transformations centralisé est un point faible.

**Décentralisée:** Cette architecture nécessite des répertoires et des services de transformations distribués pour chaque autorité de sous nommage obligée de gérer ses propres noms d'IP ; une autorité de niveau supérieur gère le processus de transformation des redirections vers les services appropriés.

## Des ressources

ERPANET workshop Persistent Identifiers

Thursday 17th - Friday 18th June 2004-University  
College

Cork, Cork, Ireland

[www.erpanet.org/events/2004/cork/index.php](http://www.erpanet.org/events/2004/cork/index.php)

DCC Workshop on Persistent Identifiers

30 June – 1 July 2005

Wolfson Medical Building, University of Glasgow

<http://www.dcc.ac.uk/events/pi-2005/>

URN:NBN

<http://www.ietf.org/rfc/rfc3188.txt>

URN:NBN:DE

<http://www.persistent-identifier.de>

URN:NBN:IT

<http://www.rinascimento-digitale.it>

DOI

<http://www.doi.org>

ARK

<http://www.cdlib.org/inside/diglib/ark/>

PADI

<http://www.nla.gov.au/padi/topics/36.html>

PILIN

<https://www.pilin.net.au/>

OpenURL

[http://www.niso.org/committees/committee\\_ax.html](http://www.niso.org/committees/committee_ax.html)

### Basée sur les DNS:

Le protocole http est sollicité pour 'activer' le lien de la citation sur la toile au travers d'une requête http vers une routine de transformation. Cette conception basée sur les DNS ne nécessite pas de clients spécifiques ou de plug-ins pour les navigateurs standards de l'Internet.

### Non basée sur les DNS:

De nouvelles implémentations ont aidé au développement d'un protocole spécifique pour la gestion du nommage et les transformations des IP (exemple DOI). Dans ce cas, un client spécifique (ou une extension au navigateur) est nécessaire pour résoudre un identificateur spécifique et accéder aux objets digitaux ou à leurs métadonnées. Cette solution peut comporter un proxy pour étendre le service jusqu'au protocole http.

## Les opportunités de recherches

Avec l'accroissement des sociétés des technologies de l'information, le problème de la stabilité des URL pour l'accès aux ressources sur Internet fait l'objet d'une attention croissante. Les systèmes d'identificateurs persistants sont une réponse relativement nouvelle à ce problème. Le contexte extrêmement dynamique dans lequel ces systèmes opèrent est la cause de l'émergence d'importants secteurs de recherches. Voici quelques aspects intéressants et non encore résolus à étudier plus en profondeur :

- La tendance actuelle est de se doter de systèmes qui sont liés au domaine de l'utilisateur (par exemple, le NBN dans le domaine des bibliothèques). Cependant une ressource peut faire partie d'un ou plusieurs domaines et peut être identifiée par plusieurs systèmes. Aussi est il nécessaire de garantir l'interopérabilité entre les différents systèmes d'identification et les implémentations basées sur le même espace de noms ;
- Les identificateurs persistants donnent accès aux ressources mais aussi à leurs métadonnées, qui sont fondamentales pour permettre à l'utilisateur d'identifier les contenus. Cependant, il est encore plus important de développer des gestionnaires de métadonnées évolués et des services aux utilisateurs, telles les recherches s'étendant sur plusieurs dépôts.
- Les relations sémantiques d'objets multimédia pourraient être prises en considération afin de définir des ontologies et une meilleure compréhension des ressources Internet.