Table des matières

AXIS-CSRM 01 : Le modèle d'IDENTIFICATION et de DÉNOMINATION' (Enregistrer & Représenter & Localiser) 1

URI (Universal Resource Identifier) 1

Configuration Management Services 2

Le système MuseumID 3

Configuration Management Resolving Services : 5

Construction des URN : 5

PURL : Persistent Uniform Resource Locator 6

AXIS-CSRM 01 : Le modèle d'IDENTIFICATION et de DÉNOMINATION' (Enregistrer & Représenter & Localiser)

Les identifiants sont des noms (termes, représentations formelles) choisis avec l'intention qu'ils soient non-ambigus, uniques dans le contexte courant. Les identifiants universels sont des noms construits pour être uniques dans le contexte de leurs règles de constructions.

En Belgique, l'identification officielle des personnes humaines est basée sur la date de naissance suivie d'un numéro de séquence et d'un code de validation, le tout codé suivant une syntaxe très stricte : exemple : 37.12.25-317.85 [Guy Maréchal est le 317ème inscrit au Registre des nés en Belgique le 25 décembre 1937 ; 85 est le code de validation basé sur un calcul cryptographique]. Il est significatif que cet exemple implique un Registre et de la Cryptographie. Ces deux concepts sont essentiels dans la mise en place de 'bons' systèmes d'identification et de nommage.

Les identifications et les dénominations comme telles ne localisent pas ce qu'elles désignent (objets, concepts, représentations …), mais ces localisateurs peuvent être, eux aussi, identifiés comme objets eux-mêmes.

Dans AXIS-CSRM, chaque 'Chose' reçoit un nom et un identifiant : ce sont deux propriétés distinctes de cette Chose. Le nom d'une chose est plutôt destiné à la lecture par des humains et l'identifieur est plutôt destiné à l'usage par des machines.

Les identifiants font partie intégrante des univers des représentations textuelles, audiovisuelles, structurelles … des objets et des concepts. Le caractère unique qu'offre un identifiant est extrêmement important pour toute une série de domaines, dont celle de la gestion des droits. Il s'agit de déterminer avec un haut degré de fiabilité qu'une expression audiovisuelle est effectivement unique. Les identifiants peuvent être délivrés par des tierces parties certifiantes (TTP) qui constituent également des ressources importantes pour la déclaration d'existence d'un contenu.

URI (Universal Resource Identifier)[1]

Il existe de nombreux mécanismes d'identification d'objets dont certains ont été normalisés comme ceux du «web». Fondamentalement sur le web, les objets ne sont pas nommés, ils sont simplement localisés ! C'est la raison pour laquelle le concept d'identification connu sous le nom d'URI (Universal Resource Identifier) couvre deux réalités disjointes : le nommage de l'objet URN (Universal Resource Name) et la localisation de l'objet URL (Universal Resource Locator).

[1] Une introduction plus détaillée est disponible à http://en.wikipedia.org/wiki/Uniform_resource_identifier.

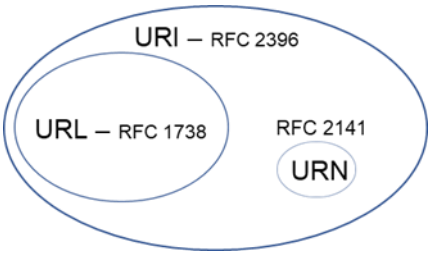

Figure Erreur ! Il n'y a pas de texte répondant à ce style dans ce document.**1 : L'URI couvre deux réalités : URL & URN**

Pour que l'identification des objets puisse être faite de manière décentralisée, un système hiérarchique a été mis au point afin d'identifier des domaines (et sous-domaines) et les objets dans un domaine particulier. Pour assurer une identification distincte à chaque objet, chaque domaine possède une fontaine à Identifieurs et un registre des Identifieurs distribués. De tels services sont intitulés UIS « Unique Identification Services ». Les UIS peuvent desservir plusieurs domaines ou sous-domaines.

Configuration Management Services

Pour que l'on puisse désigner un objet indépendamment par son URN ou son URL, il faut pouvoir faire appel à des services traduisant un URN vers un URL ou vice versa ! De tels services sont intitulés « *Resolving Services* » qui peuvent desservir plusieurs domaines ou sous-domaines. Souvent les UIS et RS sont groupés avec d'autres services sous le nom de «Configuration Management Services». Ils font l'objet d'une définition détaillée à la section 6.8.
Chaque ensemble nommé possède un URN. Il peut se rapporter à un domaine entier. Un sous-domaine de ce domaine peut être nommé par un URN qui, usuellement, est construit au départ du nom de domaine. L'ensemble de tous les services et protocoles, systèmes associés à cet URN est connu sous le nom de URN Namespaces.

L'écriture de ces documents et normes a dû se faire alors que des systèmes disparates préexistaient. L'approche actuelle sous-jacente à ces documents suggère l'hypothèse qu'un système cohérent unique, calqué sur l'approche cohérente du Web, surgira progressivement, et que les systèmes existants disparaîtront progressivement[2].

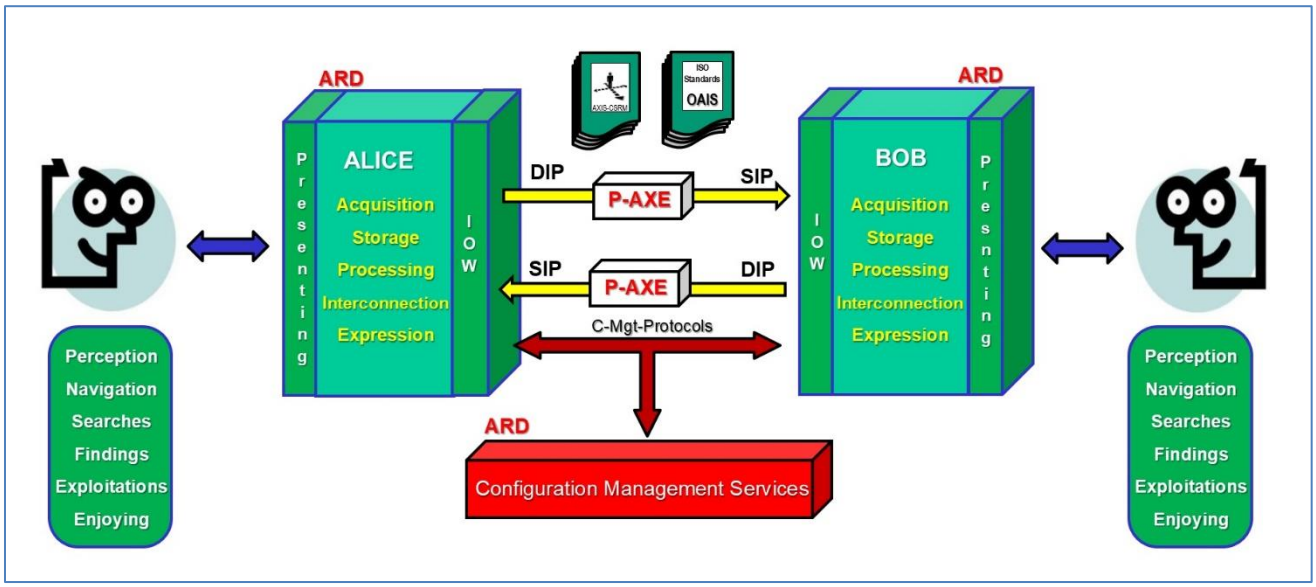

Figure 2 : Interopérabilité globale suivant AXIS-CSRM/OAIS

Le concept d'ARD [Autonomous Resource Domain] a été introduit à la section précédente.

[2] L'exemple flagrant est celui de l'identification des livres pour lequel le domaine racine est ISBN (International Standard Book Number) et qui possède des sous-domaines parfois assez incohérents (exemple : Benelux !)

Dans AXIS-CSRM, le principe de base est que tous les systèmes d'identification pertinents pour un objet peuvent cohabiter. Dans chacun des domaines d'AXIS-CSRM, un objet est identifié avec un identifieur propre au domaine.

Lorsque deux domaines s'échangent des objets, dans AXIS-CSRM, ces objets sont encapsulés (wrapper) suivant une représentation structurée appelée P-AXE (**P**ackaged **A**utonomous e**X**changed **E**ntities). Dans l'illustration ci-dessus, quand Bob reçoit dans son Guichet d'interopérabilité (IOW) un P-AXE d'Alice, tous les objets reçus vont être dotés d'une identification nouvelle spécifique et unique dans le domaine de Bob. Toutes les identifications des objets préexistantes seront conservées, en particulier celles qui étaient en vigueur lorsque l'objet était sous la gestion d'Alice. De telle sorte qu'Alice puisse continuer à gérer l'objet dans son domaine et qu'indépendamment Bob puisse faire évoluer le fichier qu'Alice lui a envoyé. De proche en proche, si le système Alice avait reçu copie de cet objet via un tiers, il en aurait gardé les identifiants d'origine (y compris ISBN)[3].

Dans AXIS-CSRM, le concept d'identifiant est générique. La structuration des données se fait sur la base d'identifiants, d'objets, ou d'identifiants de modèles d'objets, ainsi que tous les détails de ces modèles.

Pour pouvoir traiter le problème correctement, il faut donner un identifiant unique au niveau mondial soit à des objets dont on déclare l'existence (objets informatiques ou physiques), soit à des objets informatiques qui vont servir à modéliser les objets dont on a déclaré l'existence.

Il faut différencier une déclaration d'existence (nommer et identifier un objet qu'on prétend existant) du ou des représentations de cet objet nommé, identifié, ainsi que ses évolutions.

Le premier critère est de donner une identification à un objet (informatique ou physique) indépendamment de sa matérialisation effective. Donc il faut que l'informatique ait les moyens de représenter par une chaîne de bits, ou de caractères, un identifiant qui est associé à cet objet ou à un modèle de cet objet. L'objectif n'est nullement de localiser au travers de cet identifiant l'endroit où l'objet ou la ressource informatique se trouve. Cet identifiant n'est pas lié au temps ni lié à une technologie particulière ni porteur de sens; il est simplement un identifiant unique appelé URI dans le domaine web (URN ou URL.) Il est essentiel qu'il existe un nom indépendant du temps, de l'espace et de la façon dont il est matérialisé.

Au niveau conceptuel, le modèle de représentation, d'identification et de nomination va imposer le choix d'un système qui va assurer que l'identifiant de l'objet est unique, indépendant de l'endroit où l'objet se trouve.

La technologie des UUID (Universal Unique Identifier[4]) a été exploitée de manière astucieuse par le système allemand (MuseumID) en vue de garder le lien avec l'identifiant de l'objet muséal, le code-barre qui figure sur l'objet existant. Ce qui permet une exploitation informatique large lors de la fabrication de l'UUID, et faire en sorte que l'utilisateur puisse retrouver l'identifiant dans le musée par un resolving automatique. Une modification du système d'identification de l'objet (lors d'une vente), peut entrainer un changement du code-barres, mais l'UUID pourrait rester identique ! Dans ces cas néanmoins, malgré la perte de traçabilité, il est rare en pratique que l'on ne change pas aussi l'UUID lorsque l'on recode le code-barres suivant les règles locales.

Le système MuseumID

Le système proposé par MuseumID offre donc une solution astucieuse et satisfaisante pour identifier un objet muséal de manière unique mondialement, en utilisant comme semence l'identifieur de cet objet dans l'institution muséale qui le possède. Cet identifieur est tel qu'il peut être «résolu» pour retrouver l'identification semence et tel qu'il puisse servir de base pour générer une URL pointant vers la documentation de cet objet. Mais, ce système ne règle pas trois des cas classiques de 'resolving', à savoir repérer l'endroit où se trouve l'objet muséal (dans quelle ville, dans quel bâtiment, dans quelle salle …), la restructuration des sites Web et la traçabilité de l'objet muséal et de sa page Web lorsqu'il est mis à disposition, cédé ou vendu à une autre institution ou personne.

1. Pour le premier cas, l'on peut se contenter d'imaginer que la page Web pointée par l'URL mentionnera l'endroit où se trouve l'œuvre ; l'expérience montre que les responsables des institutions négligent souvent de mettre à jour la page ou même préfèrent ne pas la mettre à jour pour ne pas devoir signaler que l'œuvre

[3] Ce point sera précisé dans le « data model ». La traçabilité des échanges peut être assurée par un service du Configuration Management.

[4] Wikipédia : Ces identifiants uniques sont codés sur 128 bits et sont produits en utilisant des composantes pseudo-aléatoires ainsi que les caractéristiques d'un ordinateur (numéro de disque dur, adresse MAC, etc.). Un UUID se présente habituellement sous cette forme : 110E8400-E29B-11D4-A716-446655440000)

est en réserve, prêtée ou en cours de restauration ; ce qui risquerait de décourager certains visiteurs potentiels.

2. Le système PURL a été imaginé et mis en place pour résoudre le second cas (voir ci-après).
3. Le troisième cas est doublement non résolu : l'objet muséal sera d'habitude réidentifié dans son nouvel environnement (le lien sera perdu) et une nouvelle page Web sera émise (muni d'une nouvelle URL) pour la documentation organisée dans le nouveau contexte.

AXIS-CSRM propose une généralisation de la dénomination et de la localisation des **Choses,** de leurs **Représentations** ainsi que des **liens diarisés**[5] qui les unissent. L'intention est que l'approche proposée soit ouverte, flexible, complète et évolutive. La généralisation consiste à couvrir l'expression des liens (potentiellement évolutifs) entre toutes formes de choses matérialisées et toutes formes de localisations.

Il faut noter que dans AXIS-CSRM l'expression de l'URL associé à un URN (des URL's associés à un URN) est toujours exprimée dans un document distinct, spécifiquement et uniquement destiné à porter le lien entre URN et URL (les liens entre URN et URL's). Le cas particulier de donner l'URL du « Resolving Service » sera traité dans le modèle de données (Section 6.5).

Au niveau conceptuel, l'exploitation des spécifications de l'URN est essentielle (même si ce concept a été créé pour représenter une ressource au sens informatique au sein du web). L'approche distingue strictement l'identifiant d'un objet au sens conceptuel du terme (dénommer l'objet) de la façon par laquelle cet objet va finalement être localisé physiquement par une URL (sur le web une adresse HTTP ou comme un fichier sur un CD-Rom ou comme une coordonnée GPS). C'est un Resolving Services qui localisera physiquement la ressource, qui permettra d'y accéder ou de la localiser.

Ce qui est important c'est de garantir l'unicité de ces URN et cela se fait souvent via un système central générant tous les URN. Mais ce système central peut se borner à générer un identifiant racine, ce qui organise la possibilité d'une allocation décentralisée des URN locaux. Par exemple, un numéro ISAN qui est complété par une chaîne de caractères liée à un pays, un organisme, une version … Cet exemple implique une fonction de Configuration Management à l'intérieur du système pour pouvoir gérer l'URN.

Il existe des systèmes tels qu'ils n'ont pas besoin de service de configuration management (tel qu'une fontaine d'URN) pour générer des URN uniques. C'est le cas du système des UUID ! Ils sont uniques avec une probabilité « femtoscopique »[6] de collusions. Ils sont basés sur des mécanismes cryptographiques qui génèrent des séquences.

On peut générer les URN par une concaténation du nom de l'ARD avec un UUID nommant l'objet (la ressource) dans un mécanisme mixte d'encapsulation. Par exemple, lors de la création d'un l'URN, on écrit l'identifiant du domaine (par ex RTBF) suivi du ''UUID généré''. ''RTBF-UUID'' indique spontanément une origine RTBF. Cet exemple présuppose que RTBF, le nom de l'ARD, est unique au niveau mondial ; ce qui n'est de toute évidence pas le cas (ex. Radiotélévision Burkina Faso).

C'est au niveau du PROFILE qu'il faut déclarer comment sont codés les URN, en fonction de quelle règle, par exemple le nom de domaine de la RTBF suivi de l'UUID généré en fonction des normes. Cette règle permet aussi qu'un même objet puisse être indexé avec plusieurs identifications différentes.

Pour toute une série d'applications, il est nécessaire de connaître l'identification (par exemple l'ISBN d'un livre). Lors d'un traitement par un utilisateur, il est parfaitement possible que ce dernier utilise un identifiant généré (exemple RTBF-UUID), plus pragmatique pour le retrouver, l'archiver, identifier toutes les traductions, toutes les éditions, qui chacune ont un ISBN différent.

Le modèle conceptuel d'identification et de nomination distingue objet et modèle d'objet ou parties de modèle d'objet par un système d'URI, car il faudra pouvoir un jour désigner la ou les matérialisations physiques de ces objets.

---

5 Processus de partitionnement d'un flux audio en segments homogènes en fonction de l'identité du locuteur

6 Le terme « femtoscopique » désigne une valeur tellement petite que l'on peut considérer son impact opérationnel comme négligeable en pratique dans le contexte considéré. Dans de tels systèmes soit les cas de collision seraient détectés *ab initio* et seraient contournés ; soit l'impact d'une collision serait limité et contrôlable ultérieurement. Ce mot provient de préfixe 'femto' qui en mathématique et en physique désigne 10-15 (un millionième de milliardième).

Il faut dès lors disposer d'un mécanisme qui, au départ de l'URN, permet de retrouver la matérialisation de l'objet qu'il soit externe (un lieu géographique…), ou interne (une ressource informatique : bits codés sur un DVD). Cela implique un identifiant de l'objet encapsulé, de l'objet devenant physiquement accessible, de l'objet devenant interopérable dans le sens où il a une existence matérielle. Son identifiant ayant les caractéristiques telles que l'on puisse accéder à sa ou ses matérialisations.

Configuration Management Resolving Services :

C'est l'objet des Resolving Services du Configuration Management. À chaque matérialisation d'un objet informatique ou à chaque déclaration d'existence d'un objet physique, le Configuration Management Services stocke l'information de l'endroit où se trouve l'objet physique représenté par cet URI. Le Configuration Management Service doit pouvoir relayer cette information, ou demander l'accès à une des copies de tels objets informatiques.

Il existe aussi d'autres types de Resolving Services qui doivent pouvoir répondre à des questions indirectes au sujet de la localisation du modèle d'un objet. C'est un resolving d'une autre nature, car il implique non seulement de retrouver l'objet, mais également d'exploiter les URI associés à l'un des modèles de cet objet.

Le Resolving Services ne porte donc pas uniquement sur le fait de pouvoir transformer un identifiant en une localisation, mais peut résoudre des choses extrêmement complexes comme retrouver la ou les modélisations d'un objet. Ou l'inverse, avoir le modèle et rechercher l'objet qui est modélisé, ou l'évolution, l'enchaînement entre les modèles.
Le modèle d'identification est donc essentiel pour conférer au système d'identification les propriétés indispensables à une bonne organisation.

AXIS-CSRM définit une forme de structuration de telle sorte que les propriétés de l'organisation soient respectées : un URN distinct, disjoint de l'URL ou de l'activation de service qui rende possible de répondre aux questions des Resolving Services.

Construction des URN :

L'approche serait de construire les URN comme une séquence de trois champs :

1. Le nom de l'ARD serait construit sur base du nom de domaine tel qu'utilisé sur le Web. Dans le cas où l'on veut identifier un département du domaine, le nom de cette ressource serait concaténé avec celui du domaine. L'intérêt de cette approche est qu'elle garantit l'unicité sans avoir besoin de créer un nouveau service de Configuration Management. Cette partie peut donc aussi servir comme désignation des 'Namespaces' intervenants dans les expressions XML, OWL, RDF et similaires ...

   Pour la syntaxe, voir : http://tools.ietf.org/rfc/rfc2141.txt

   Exemple : le nom de l'ARD pourrait lui-même être constitué de deux parties :
   - d'un nom de domaine Web (unique par définition). Par exemple : **www.titan.be**
   - d'une chaine hiérarchique de sigles exprimant la structure dans le domaine. Ce n'est pas un sous domaine au sens du Web : il s'agit simplement de pouvoir nommer distinctement les divers ARD associés au domaine unique du Web.
   - Le caractère concaténateur entre le nom de domaine Web et entre les divers sigles serait, par exemple, le tiret simple (hyphen) « **-** »

   Exemple de nom d'ARD : **www.titan.be-rd-s5**

2. Un caractère concaténateur : dans l'exemple, nous suggérons d'utiliser :
   a. le double point « **:** » pour les URN des objets modélisant et le double dot ":"
   b. le *underscore* « **_** » pour les URN des objets modélisés the underscore "_"

3. Le nom interne de l'objet dans le domaine : il serait construit en se servant de l'approche UUID où l'unicité est garantie par une approche cryptographique normalisée qui ne nécessite pas de service de Configuration Management externe. L'identifieur est de taille fixe. Il peut être présenté suivant divers modes dont le plus courant est une chaine de caractères hexadécimaux, ce qui facilite la lecture par des humains.

   Exemple : **16fd2706-8baf-433b-82eb-8c7fada847da**

   Voir : http://fr.wikipedia.org/wiki/Universal_Unique_Identifier

En suivant la syntaxe et les choix de l'exemple, l'URN serait :

www.titan.be-rd-s5:16fd2706-8baf-433b-82eb-8c7fada847da

Le caractère concaténateur étant le double point, il s'agit de l'URN d'un «objet modélisant».

Il s'agit bien d'un URN et non d'un URL !

En pratique, à l'intérieur d'un ARD les objets sont identifiés par leur UUID propre précédé par le concaténateur distinguant les modélisés et les modélisants. Dans les exports, les objets à l'intérieur des P-AXE sont nommés par le nom complet concaténé. Objectivement, l'UUID suffirait pour identifier l'objet (la ressource). Néanmoins, pour de nombreuses raisons, le mécanisme d'autonomie du nommage des ARD et du nommage d'objets (resources) dans un ARD possède de très nombreux avantages. Une analyse très fine en est faite dans la référence du MuseumID (précité) :

http://museumid.net/documentation-20101206 .

Autre exemple : un utilisateur introduit l'adresse «dbpedia.org/resource/Bruxelles » pour une recherche sur des pages ciblées sur la ville de Bruxelles ! L'adresse introduite est une représentation formelle du concept, d'une ressource, mais elle n'est pas déréférençable !

« dbpedia.org/resource/Bruxelles » est l'URI de type HTTP qui identifie le concept de Bruxelles dans l'univers de dbpedia. Cet URI est déréférençable, car il est possible, en l'introduisant dans un navigateur, d'obtenir, ici par une redirection 303, la description de la ressource en question dans DBpedia, description qui est localisée par l'URL « dbpedia.org/page/Bruxelles ».

De plus, dans chaque ARD, il y a des représentations différentes d'un même concept ! La ville de Bruxelles est définie différemment dans DBpedia ou Freebase, … le fait de relier les termes (manuellement ou par inférence) par «*sameas*» globalise les Entités des deux bases de connaissances.

PURL : Persistent Uniform Resource Locator

Un localisateur persistant de ressources uniformes (PURL) est un localisateur de ressources uniformes (URL) (basé sur la localisation de l'identifiant de ressource uniforme ou URI) qui est utilisé pour rediriger vers l'emplacement de la ressource Web demandée. PURL redirige les clients HTTP en utilisant des codes d'état HTTP. Des PURL sont utilisés pour assurer le processus de résolution des URL, résolvant ainsi le problème de l'URI transitoire dans les régimes URI géodépendants tels que HTTP.

PURL met en œuvre une forme d'identification permanente de ressources virtuelles. D'autres schémas d'identification persistants comprennent les identificateurs d'objets numériques (DOI), les identifiants des Sciences de la vie (LSIDs) et des URI INFO. Tous les systèmes d'identifications persistantes fournissent des identificateurs uniques pour (éventuellement changer) les ressources virtuelles, mais tous les régimes n'offrent pas des possibilités de conservation (curation). La conservation des ressources virtuelles a été définie comme «la participation active des professionnels de l'information dans la gestion, y compris la préservation, de données numériques pour une utilisation future".

L'usage de PURL a été critiqué pour leur besoin de résoudre une URL, liant ainsi un PURL à un emplacement réseau. Les emplacements de réseau comportent plusieurs vulnérabilités, telles que les enregistrements de noms de domaine et les dépendances du système d'accueil. Un échec pour résoudre un PURL pourrait conduire à un état ambigu: Il ne serait pas clair si le PURL a échoué à résoudre en raison d'une défaillance du réseau ou parce qu'il n'existait pas.

Wikipédia : http://en.wikipedia.org/wiki/Persistent_uniform_resource_locator

ISO standard (ISO 26324) for the DOI system

DOI registration agencies: http://www.doi.org/registration_agencies.html

DOI Handbook: http://www.doi.org/hb.html