LE DIGITAL OBJECT IDENTIFIER

LE SYSTÈME DU DOI

CATHERINE LUPOVICI

Jouve Digitalisation des Informations

clupovici@jouve.fr

Le Digital Object Identifier (DOI) est un système développé par le CNRI (Corporation for National Research Initiatives) pour le compte de l'Association of American Publishers (AAP)[1]. Le système s'appuie sur la mise en place d'un répertoire international d'identifiants de portions de publications électroniques. Son objectif est de permettre la collecte des droits d'utilisation des documents publiés électroniquement sur une base définie par le détenteur des droits. L'identifiant DOI est compatible avec les différentes numérotations internationales de document et avec les principes de l'identification des documents sur Internet à laquelle travaille l'IETF (Internet Engineering Task Force).

Le DOI est la dernière initiative du monde de l'édition traditionnelle pour créer un numéro identifiant générique adapté à la gestion des droits dans le contexte de la distribution électronique des publications sur Internet. Ce projet est construit sur l'héritage des numéros d'identification des documents et des œuvres qui existent déjà.

Les numéros d'identification des documents

L'identification exacte d'un document a toujours été un élément important dans l'activité de création, de commercialisation et d'échange pour effectuer le lien entre le créateur de l'information et son utilisateur. Les identifiants sont une courte séquence numérique ou alphanumérique attribuée à une unité d'information et servant de manière univoque et permanente à désigner cette unité.

L'identifiant s'applique à une unité d'information qui a tout d'abord été considérée comme une entité physique (un livre, une publication en série, un disque). On parle alors de numéro ou de code. Puis, on a considéré les parties logiques composant cette unité physique, et pour certains usages, on a souhaité identifier ces parties comme des documents individuels (chapitre d'un livre, article dans une publication scientifique, chanson particulière sur un disque). On a alors utilisé l'expression *Document Like Object* (DLO) pour désigner ces parties de documents composantes d'une unité physique.

Enfin, dans le contexte de l'édition électronique, il devient nécessaire d'identifier des parties composantes

1. Appel à propositions lancé le 13 mars 1996.

d'une unité logique documentaire, de telles unités pouvant d'ailleurs appartenir, par le biais de liens, à plusieurs documents à la fois (par exemple un schéma, un tableau, etc.). Les objets de granularité plus ou moins fine par rapport à la publication sont alors caractérisés par un identifiant.

Les numéros d'identification d'unités physiques

Les systèmes d'identification des documents en usage depuis une vingtaine d'années servent tout au long de la chaîne de production et de distribution des publications et font partie, à des degrés divers selon les pays, des pratiques du monde de l'édition, des libraires et agences de diffusion et des bibliothèques et centres de documentation. Un certain nombre d'entre eux sont l'objet d'une norme internationale[2] qui définit la structure de l'identifiant et l'organisation internationale des agences d'attribution, voire l'organisme international lui-même. Des identifiants plus anciens comme le CODEN pour les publications en série sont encore en usage dans certaines communautés.

Les schémas d'enregistrement que nous connaissons sont souvent construits sur un organisme international, relayé par des agences responsables des publications par zones géographiques (pays ou zone de langue selon le cas).

Les bibliothèques nationales sont souvent l'agence nationale d'attribution de l'identifiant sur la base d'un système de dépôt légal ou de dépôt volontaire des publications nationales. Ces identifiants font partie de la description bibliographique d'un document, ainsi que des éléments figurant dans la citation d'un document dans la bibliographie d'une publication. Ces systèmes d'identification considèrent la publication dans son intégralité physique comme un livre, un disque.

DIFFÉRENTS SCHÉMAS D'IDENTIFICATION ONT ÉTÉ DÉFINIS SELON LES TYPES DE DOCUMENTS EN FONCTION DES USAGES ET DES BESOINS DU DOMAINE

Les identifiants les plus anciens en usage dans les bibliothèques sont l'ISBN (International Standard Book Number) et l'ISSN (International Standard Serial Number) qui servent d'identifiant d'une publication pour les acquisitions et la recherche bibliographique.

Différents schémas d'identification ont été définis selon les types de documents en fonction des usages et des besoins du domaine. Les numéros dont nous disposons actuellement sont soit signifiants comme l'ISBN, qui identifie une publication par zone de langue, par éditeur, et par numéro chez l'éditeur, soit non signifiants comme l'ISSN, qui est un simple numéro séquentiel dans la tranche de numéros dont dispose l'agence nationale.

Les identifiants que nous connaissons ont été définis pour un usage particulier et leur extension à des usages connexes ou à des environnements techniques qui s'élargissent posent des problèmes. Par exemple, un ISBN différent est attribué à la version brochée et à la version reliée d'une publication et permet de gérer l'acquisition avec un prix différent pour l'une ou l'autre version. Cependant, cette distinction est gênante dans le cas d'utilisation de l'ISBN comme identification dans une citation, dès lors que les deux versions ont une pagination identique. A l'inverse, un seul ISSN, lié au titre-clé d'une publication en série, est attribué quel que soit le support de publication : papier, microforme ou cédérom. Cet identifiant pose à l'inverse des problèmes d'identification par rapport à un support pour la gestion de la description physique de la publication et des matériels éventuels nécessaires à son utilisation.

On voit donc ainsi le principe de base de l'identifiant qui est son unicité remis en question en fonction de l'utilisation, dès lors que l'on étend l'identifiant à d'autres usages que celui pour lequel il a été défini et adopté.

Les identifiants d'unités logiques

D'autres systèmes ont été mis en place pour la numérotation de portions logiques de publications formant un tout, comme par exemple l'identification d'un article à l'intérieur d'un fascicule d'une publication en série ou l'identification d'un chapitre d'un ouvrage collectif. Ces identifications plus précises sont utilisées à la fois pour la gestion de la production par l'éditeur et pour la gestion de la fourniture de documents ou du prêt entre bibliothèques, dès lors que la gestion des transactions est informatisée ou que l'on dispose d'une forme électronique du document permettant d'en sélectionner des parties composantes logiques à l'intérieur d'une unité physique.

Une norme ISO a été définie à cet usage, en extension de l'ISSN et de l'ISBN. Elle a été très longue à finaliser et vient d'être annulée, car elle n'est pas utilisée et ne correspond pas aux besoins des publications

2. – ISO 2108 : 1992 Information et documentation. – *Système international pour la numérotation des livres* (ISBN).
– ISO/DIS 3297 Information et documentation. – *Numérotation internationale normalisée des publications en série (ISSN) (Révision de l'ISO 3297 : 1986)*.
– ISO 3901 : 1986 Documentation. – *Code international normalisé des enregistrements* (ISRC).
– ISO 10444 : 1994 Information et documentation. – *Numéro international normalisé des rapports* (ISRN).
– ISO 10957 : 1993 Information et documentation. – *Numéro international normalisé de la musique* (ISMN).

électroniques[3]. C'était un identifiant complexe s'appuyant sur l'identifiant du contenant physique et comprenant l'année de publication, la tomaison, le fascicule, et les pages extrêmes, c'est-à-dire pratiquement les éléments d'une citation dans une bibliographie.
Les autres identifiants qui ont été développés et utilisés par les éditeurs, dans le contexte de l'édition traditionnelle, pour les unités logiques composant des unités physiques, sont le Publisher Item Identifier (PII) et le Serial Item and Contribution Identifier (SICI).

Publisher Item Identifier

Le PII a été défini en 1995 par un groupe d'éditeurs scientifiques. C'est un identifiant simple, utilisé pour les traitements et échanges dans un environnement électronique, identifiant de manière unique le document logique (sans référence aux différents supports sur lesquels on peut le trouver). Il est attribué par chaque éditeur et applicable à tout support (livre, CD ou autres), dès l'introduction de la contribution dans la chaîne de production, et n'est pas modifié en fonction de la date ni de la forme finale de la publication. Il est donc non signifiant et ne contient pas de métadonnées (informations sur le document). Compatible avec les normes existantes, puisqu'il intègre l'ISSN et l'ISBN des publications hôtes, il est composé de dix-sept caractères dont un caractère de contrôle.
Le PII a été utilisé dès 1996 par la plupart des éditeurs STM (Sciences, Techniques et Médecine), dont American Chemical Society, American Institute of Physics, American Physical Society, Elsevier Science, IEEE, American Mathematical Society, Springer Verlag, INSPEC, ADONIS. Il est essentiellement utilisé avant la publication.

Serial Item and Contribution Identifier

Les travaux sur cet identifiant ont commencé en 1983 au SISAC (Serials Industry Standardisation Advisory Committee) aux États-Unis. Il est ensuite devenu une norme américaine officielle ANSI/NISO Z39.56-1991. Sa révision, qui date de septembre 1996, permet la gestion de trois niveaux de granularité dans les parties composantes de la publication en série identifiée par son ISSN : le fascicule, l'article ou la partie d'article (table des matières, index, illustration par exemple).

LE CONCEPT D'IDENTIFICATION RECOUVRE DÉSORMAIS À LA FOIS L'IDENTIFICATION ET LA LOCALISATION

Le SICI est un identifiant unique pour chaque expression physique d'un même document logique (un même article aura des SICI différents selon le support de publication). C'est un identifiant de longueur variable qui peut avoir trois formes selon qu'il identifie un fascicule, une contribution dans un fascicule ou une contribution dans un autre composant.
Dans les deux premières formes, il s'appuie sur l'ISSN, dans la troisième sur un autre identifiant (cela peut être le PII). Le BIC (Book Industry Communication), au Royaume-Uni, a fait une proposition d'extension aux contributions dans les livres. L'identifiant correspondant sera le BICI (Book Item and Contribution Identifier). Le degré de granularité que le BICI couvrira à l'intérieur d'un livre est encore en discussion.

Le PII et le SICI (1996) sont utilisés en complémentarité par les éditeurs pour identifier l'article virtuel avant sa publication et la ou les formes physiques une fois l'article publié. La version 1996 du SICI a pour objectif de définir un identifiant utilisable pour la gestion des périodiques ou des articles qu'ils contiennent dans des fonctions telles que la commande, la réclamation des fascicules manquants, la collecte des royalties, la gestion des droits d'usage, la recherche en ligne, la navigation hypertexte entre les banques de données, la fourniture de documents.
Les identifiants SICI sont utilisés dans des applications d'échanges de données électroniques (EDI), dans le système de code à barres américain pour les publications en série (SISAC), dans les numéros utilisables pour une recherche dans la norme Z39.50. Ils peuvent être introduits dans les URN (Uniform Resource Names) en cours de définition et sont donc potentiellement utilisables en environnement réseau.

L'identifiant dans l'environnement réseau

L'identification d'une œuvre telle qu'elle était pratiquée jusqu'à maintenant était un outil fondamental, mais passif. Avec le réseau et les fonctionnalités des applications Web, l'identification devient un élément actif utilisable directement pour passer de l'identification à l'œuvre elle-même. L'utilisateur final peut ainsi agir directement *via* l'identifiant qui peut être transparent et présenté au travers d'un bouton. Le concept d'identification recouvre désormais à la fois l'identification et la localisation.
La communauté Internet a tout d'abord utilisé le lien actif des documents HTML comme élément d'identification-localisation-accès sur le Web. Ces liens sont les URL (Uniform Resource Locators) qui permettent de naviguer sur le réseau. Tous les premiers travaux de citation des documents sur le réseau utilisent l'URL comme composante

3. ISO 9115 : 1987 Documentation. – *Bibliographic Identification (BIBLID) of Contributions in Serials and Books*. Annulée en 1996.

de la citation bibliographique. Malheureusement, cette information n'est pas pérenne et nombre d'URL sont modifiés en cas de changement de machine hôte. Le document peut aussi être supprimé.
Enfin, cet identifiant n'est pas unique, puisque le même document peut se trouver sur plusieurs sites.
La communauté Internet, soucieuse d'offrir des identifiants pérennes et uniques pour les publications effectuées sur le réseau, a donc travaillé à la normalisation de l'Uniform Resource Names (URN)[4] pour l'identification unique des ressources sur le réseau et de l'Uniform Resource Characteristics (URC) composé de métadonnées sur les ressources associées à l'URL pour la localisation.
En parallèle avec les travaux de l'IETF, OCLC a développé le PURL (Permanent URL). Celui-ci pointe vers un système de localisation intermédiaire, qui tient les changements d'URL à jour et réoriente les demandes vers la localisation actualisée. Le PURL s'appuie sur deux des fonctions requises pour la mise en œuvre de l'URN qui est le système des répertoires d'enregistrement des ressources d'une part et la résolution de l'URN d'autre part (traduction dans les URL correspondants).
Enfin, les éditeurs ont créé le système DOI, expliqué plus en détail dans la section suivante, lui aussi compatible avec les travaux de normalisation Internet de l'URN et, en particulier, avec la fonction de répertoire et la fonction de résolution des identifiants.
La communauté Internet a sa propre terminologie et appelle l'identifiant de toute ressource sur le réseau un «nom», comme par exemple les noms de domaines pour les différents types d'utilisateurs d'Internet. Les éditeurs gardent leur terminologie traditionnelle et parlent donc d'identifiant d'objet numérique pour le nom des ressources sur Internet.

URN (Uniform Resource Names)

Le RFC (Request for Comments) 1737 pour les besoins fonctionnels des URN a été lancé sur Internet en décembre 1994. Il permet l'identification d'une ressource globale ou une unité élémentaire d'information. Il peut identifier par exemple un contenu intellectuel, une présentation particulière d'un contenu intellectuel ou tout ce qu'une autorité d'attribution de nom définit comme entité pouvant recevoir un nom. Un URL identifie la localisation ou le contenant d'une instance de ressource identifiée par un URN. La ressource identifiée par un URN peut exister à un ou plusieurs endroits (il existe plusieurs URL), changer de localisation (changer d'URL) ou ne plus être disponible du tout (ne plus avoir d'URL).
Les travaux du groupe de travail URN de l'IETF portent sur la structure de l'URN, les mécanismes nécessaires pour établir un système global et pérenne, ainsi que sur la compatibilité avec les systèmes d'identification existants. Il est également nécessaire qu'au moins un système de résolution des URN existe.
La structure de l'URN est en cours de finalisation. Il sera composé de l'identifiant du domaine, du nom et de l'identifiant dans le domaine, par exemple : URN : ISBN : 2-7355-0353-3.
Le projet de norme a été distribué en mars 1997[5]. L'URN permet d'identifier n'importe quelle publication électronique formant un tout à n'importe quel niveau de granularité d'une publication hôte, et d'en rendre recevable l'enregistrement dans un répertoire.
Plusieurs expérimentations d'implémentation d'URN sont en cours pour contribuer à l'évolution du cadre général de définition de l'URN et faire aboutir le plus vite possible l'adoption du standard sur Internet.

PURL (Permanent URL)

Le PURL est une solution pragmatique intermédiaire proposée par le Département Recherche d'OCLC en attendant la finalisation de l'URN, aux travaux duquel ce département participe activement au sein de l'IETF. Cette solution s'appuie sur une fonction standard de redirection du protocole HTTP. Le PURL est une adresse composée du protocole, d'un organisme de résolution et d'un «nom», par exemple : http://purl.oclc.org/OCLC/PURL/INET96.
Ce système permet de dissocier la localisation de l'identification et de conserver l'identification si la localisation change. OCLC propose également un mécanisme d'enregistrement pour les utilisateurs et les groupes d'utilisateurs d'un tel service. OCLC offre ses services en tant que centre de résolution des PURL qu'il a enregistrés et pour lesquels il reçoit les mises à jour de localisation dans sa base de données qui fait fonction de répertoire.
Le PURL peut être introduit dans une page Web et l'utilisateur qui s'en sert est alors certain de trouver son document sur le Web, bien entendu si le serveur hôte du document a effectué la mise à jour de la localisation auprès de l'organisme de résolution.
Tous les programmes développés pour enregistrer et mettre à jour les informations d'un centre de résolution sont distribués gratuitement par OCLC pour des organismes qui voudraient eux-mêmes devenir un tel centre, comme par exemple des agences gouvernementales, des éditeurs, des bibliothèques, des universités.
Des PURL ont été attribués par les services OCLC à des notices de catalogage de ressources Internet faites dans le cadre de l'Internet Catalog Project subventionné par l'US Department of Education.
Le Service OCLC PURL a été ouvert en janvier 1996, et le 28 mars de la même année, il avait enregistré 5 500 PURL pour 5 000 utilisateurs différents. Il avait également servi 178 000 demandes de résolution.

4. RFC 1737. Functional Requirements for Uniform Resource Names.

5. Ryan MOATS, *URN syntax, Internet Draft*, Internet Engineering Task Force. March 1997. ftp ://ds. internic.net/internet-drafts/draft-ietf-urn-syntax-05.txt

Le PURL n'est pas une identification normalisée ou normalisable à terme. C'est un projet pilote de démonstration d'une méthode d'identification sur Internet participant aux travaux sur la normalisation des URN.

Le système DOI

Les éditeurs STM ont ressenti le besoin d'étendre les systèmes d'identifiants existants, soit au niveau de l'unité physique, soit au niveau des parties composantes pour répondre aux besoins de l'environnement réseau.
Les objectifs consistent à définir un identifiant pour l'accès à des portions d'informations numériques afin de permettre aux éditeurs :
– la gestion et le recouvrement des droits,
– la distribution électronique de tout ou partie des œuvres qu'ils publient,
– la navigation hypertexte au travers des citations dans les œuvres,
– la gestion électronique du *copyright*,
– la découverte de documents,
– l'archivage d'objets électroniques,
– l'offre de la pérennité sur le Web.
Les portions d'information numériques concernées peuvent se situer à des degrés de granularité divers, déjà définis dans les extensions récentes du SICI ou du BICI et pouvant aller jusqu'au résumé, à la table des matières ou à des illustrations à l'intérieur d'une contribution.
Le dispositif DOI est compatible avec le cadre général des URN et est composé des trois éléments : un identifiant pour les objets, un répertoire offrant la résolution des identifiants, une base de données d'informations sur les objets. C'est une norme ISO potentielle et le projet a déjà été présenté pour information à différents niveaux de l'ISO en 1997.

L'identifiant

L'identifiant DOI est composé de deux parties séparées par un slash : un préfixe identifiant l'éditeur et un suffixe identifiant l'objet chez l'éditeur. L'identifiant de l'éditeur est attribué au niveau international par le ou les responsables de la gestion du répertoire. On envisage une répartition soit par pays soit par grands secteurs industriels, tels que l'édition, la photographie, la musique, les logiciels, etc. Les éditeurs pourront choisir de demander un seul préfixe pour l'ensemble de leurs secteurs d'activité ou plusieurs préfixes pour chaque ligne de produits. Pour l'instant, il y a un seul responsable du répertoire et tous les préfixes déjà attribués commencent par le chiffre 10.

L'IDENTIFIANT DOI EST COMPOSÉ DE DEUX PARTIES SÉPARÉES PAR UN SLASH : UN PRÉFIXE IDENTIFIANT L'ÉDITEUR ET UN SUFFIXE IDENTIFIANT L'OBJET CHEZ L'ÉDITEUR

Le suffixe est sous la responsabilité de l'éditeur et peut être un des identifiants déjà en usage tels que ISBN, ISSN, SICI, PII ou un identifiant totalement propriétaire et local. S'il s'agit d'un identifiant standard, les éditeurs sont encouragés à le faire précéder d'un code entre crochets carrés. Exemple : 10.1002/[ISBN]0-471-58064-3.
Le suffixe peut être attribué par l'éditeur à des objets de toute taille – livre, article, résumé, graphique –, et de tout type – texte, son, vidéo, image ou logiciel. Ainsi un objet global peut avoir un DOI et une de ses parties composantes un autre DOI, selon ce que l'éditeur a décidé de commercialiser sur Internet.

Le répertoire

Le répertoire est un système central unique de résolution des identifiants, qui les traduit dans les URL des systèmes des détenteurs de droit et redirige automatiquement l'utilisateur vers ces sites. L'utilisateur se voit alors offrir soit directement le document recherché, soit des explications sur les conditions d'acquisition.
Le répertoire conserve le DOI et effectue la mise à jour de l'URL correspondant au détenteur de droit en cas de cession des droits d'un document, ou des droits de toute une ligne de produits d'un éditeur. Le DOI reste inchangé, seule la localisation change.
Le répertoire est supporté par un système informatique développé par le CNRI. Le système fonctionne sur des sites en miroir aux États-Unis, en Europe et en Asie. Environ 250 000 DOI ont déjà été attribués par des organismes comme International Publishers Association, John Wiley & Sons, Academic Press, Springer-Verlag, le Copyright Clearing Center, Elsevier, l'American Medical Association, etc.

La base de données

Chaque éditeur maintient une base de données d'informations sur les documents enregistrés. Selon les éditeurs, cette base va contenir les documents eux-mêmes ou des informations sur la façon d'obtenir les documents. Chaque éditeur définit son offre et la base de paiement qui peut être un abonnement, une licence d'utilisation à l'unité ou un paiement à la visualisation. Le site Web du DOI permet de visiter les sites de différents éditeurs et de comparer les offres qui sont faites en utilisant des démonstrations. La base de données de l'éditeur met en œuvre le concept Internet d'URC (Uniform Resource Characteristics) dans une offre commerciale.

La Fondation internationale DOI

La Fondation internationale DOI est une association sans but lucratif responsable du système DOI. Elle a un bureau à Washington et un autre à Genève. Elle est responsable de l'organisation générale et de l'administration du réseau. Le président est Charles Ellis (John Wiley & Son), aux États-Unis. Le bureau comprend un représentant des éditions Vuibert, pour la France; de Springer-Verlag, pour l'Allemagne; de Routledge, pour la Grande-Bretagne; d'Elsevier Science BV, pour les Pays-Bas; de l'Association of American Publishers (AAP), pour les États-Unis; de l'Association japonaise des éditeurs et de la Fédération européenne des éditeurs.

La demande d'attribution d'un préfixe peut se faire en ligne. Elle sous-entend l'acceptation des conditions de participation qui comportent le respect du *copyright* et l'engagement de mettre à jour les URL permettant d'accéder aux informations sur les publications enregistrées. L'attribution d'un préfixe coûte 1000 dollars américains et la maintenance annuelle des données sera d'environ 0,01 dollar par DOI.

Le DOI émane exactement de la même initiative venant du monde de l'édition traditionnelle que celles qui ont conduit à la normalisation internationale de l'ISBN et de l'ISSN. Elle se situe d'emblée dans la normalisation et les pratiques d'Internet et y intègre les anciens numéros ISO des années 70, les identifiants non normalisés définis dans la décennie 80 et adaptés à la distribution électronique dans la décennie 90. Les éditeurs se sont positionnés comme gestionnaire des identifiants qui sont des outils de perception des droits d'utilisation.

L'utilisation à des fins d'identification bibliographique devra être examinée par la communauté des bibliothèques et des bases de données en parallèle avec les initiatives URN/URC. La communauté bibliographique anglo-saxonne participe déjà partiellement à ces débats en s'impliquant directement au sein de l'IETF, à l'ANSI/NISO ou au travers d'expérimentations. On peut espérer que la communauté bibliographique française investira concrètement avant l'apparition du projet de norme international.

Février 1998

BIBLIOGRAPHIE

1. Bossuat, Marie-Louise; Girard, Christine. – « Le Numéro international normalisé des publications en série (ISSN) ». – *Bulletin des bibliothèques de France*, 1974, n° 12, p. 557-562.

2. Digital Object Identifier System. URL : http://www.doi.org

3. Honoré, Suzanne. – « La Numérotation normalisée internationale du livre ». – *Bulletin des bibliothèques de France*, 1969, n° 8, p. 321-333.

4. Lynch, Clifford. – *Identifiers and Their Role in Networked Information Applications*. – 1997. URL : http://www.arl.org/newsltr/194/identifier.html

5. Paskin, Norman. – « Document identifiers : an update on current activities ». – *ICSTI Forum*, September 1996, n° 23, p. 3-8.

6. PURL. URL : http://purl.oclc.org

7. Santiago, Suzanne. – « Les Numérotations internationales normalisées ». – *Bulletin des bibliothèques de France*, 1993, t. 38, n° 5, p. 40-41.

8. SICI. – URL : http://sunsite.berkeley.edu/sici

9. « Uniform Resource Names : a progress report/the URN implementors ». – *DLIB Magazine*, February 1996. – URL : http://www.dlib.org.dlib/february96/02arms.html