Documents et Applications : CMS nouvelle génération

Jean-Marc Lecarpentier (1)
jml@info.unicaen.fr

Hervé Le Crosnier (1)
herve@info.unicaen.fr

Jacques Madelaine (1)
jacques@info.unicaen.fr

(1) GREYC (Groupe de recherche en Informatique, Image, Automatique et Instrumentation de Caen) – CNRS UMR 6072 – Université de Caen Basse-Normandie

Mots-clés : création document, document multimédia, document multilingue, FRBR - Functional Requirements for Bibliographic Records, CMS - Content Management System

Keywords: document creation, multimedia, multilingual documents, FRBR - Functional Requirements for Bibliographic Records, CMS - Content Management System

Résumé : L'internet est devenu la plate-forme de prédilection pour la création de documents via l'utilisation des CMS (Content Management System ou Système de Gestion de Contenu). Or, trop souvent les CMS sont conçus comme des outils de production de sites web. Nous imaginons la réalisation de nouvelles plate-formes de création et de gestion de documents qui épousent le web d'aujourd'hui : un web multilingue, multimédia, sémantique et social. L'objectif de cet article et de proposer une architecture de production et gestion des documents numériques basée sur l'expérience des bibliothèques avec le logiciel Sydonie (SYstème de gestion de DOcuments Numériques pour l'Internet et l'Édition).

Abstract : Most Web Content is nowadays published with Content Management Systems (CMS). However, most CMS consider each document as an independent entity which matches one web page. New systems need to consider today's web : multilingual, multimedia, semantic and social. Based on the experience of libraries, this article aims to propose an architecture to author and manage digital documents, with Sydonie (SYstème de gestion de DOcuments Numériques pour l'Internet et l'Édition).

Introduction

La notion de « page web » n'est plus synonyme de « document web » comme au début de l'Internet. Une page web est désormais un ensemble composite de documents, de données et d'applications. Les pages présentées sont souvent composées d'un ou plusieurs documents, qui sont détenus en propre par le serveur, ou obtenus à partir de services distants. Nous avons décrit cette évolution dans un article antérieur [1]. Il s'agit maintenant d'en tirer des conclusions opérationnelles pour la gestion des documents numériques au travers du web.

L'internet est devenu la plate-forme de prédilection pour la création de documents, via l'utilisation des CMS (*Content Management System* ou Système de Gestion de Contenu). Or trop souvent les CMS sont conçus comme des outils de production de sites web, susceptibles d'intégrer des « documents » à l'intérieur des « pages » pré-organisées et conçues par un graphiste. Nous imaginons la réalisation de nouvelles plate-formes de création et de gestion de documents qui puissent épouser toute la complexité d'un réseau. En particulier, nous avons en ligne de mire trois ingrédients du web :

– un web multilingue : il s'agit de penser le document comme l'ensemble de ses expressions linguistiques disponibles, sur le serveur local (problématique de CMS traditionnel) comme entre serveurs distants (nécessité d'un schéma de numérotation global) ;

– un web multimédia, dans lequel les objets numériques ne sont pas de simples « fichiers numériques » destinés à un *player*, mais bien des documents à part entière, associant les métadonnées et annotations autour du fichier binaire proprement dit. Cette conception impose de voir tout document comme

« composite », et de concevoir les objets numériques (images, vidéo, sons, animations,...) comme étant eux-mêmes des documents ;

– un web sémantique et social, dans lequel des données et informations sémantiques rendues disponibles en divers points de la toile peuvent être utilisées pour annoter les documents, et en sens inverse dans lesquels des documents, données et informations organisées peuvent être proposées aux autres acteurs du web sémantique, suivant la logique des « *Linked Data* » [2].

L'objectif de cet article est de proposer une architecture de production et de gestion des documents numériques répondant à ces contraintes.

1 CMS : création de documents et applications web

L'utilisation d'un CMS pour la création de documents est aujourd'hui fortement liée à la présentation du document créé *via* ce même CMS. Autrement dit, la création et la présentation sont intégrées au sein d'un même processus, visant à créer un document *pour* un logiciel et un site particuliers. Bien entendu des méthodes d'exportation des données du CMS existent en général, mais l'export (le plus souvent au format XML) ne permet pas une exploitation immédiate du document, surtout s'il est composite.

Notre objectif est de séparer la création du document de sa publication (web ou autre). Dans cette optique, il nous semble nécessaire de penser un système composé de deux outils logiciels, actuellement en cours de développement :

– un outil de création de documents : **Sydonie** (SYstème de gestion de DOcuments Numériques pour l'Internet et l'Edition)

– un outil de création d'applications web (**Aglae**) permettant d'intégrer les documents créés, des applications externes (web services), voire d'autres documents respectant les protocoles d'échange XML de diverses professions, tels PRISM (*Publishing Requirements for Industry Standard Metadata*[1]) ou newsML[2]

L'objectif de Sydonie est de permettre le stockage des documents produits avec des outils divers (en base de données, dans des fichiers XML, avec des fiches de métadonnées en RDF,...). Sydonie produit ensuite des formats physiques qui correspondent aux différentes méthodes de visualisation (ou audition) par les lecteurs humains (HTML, pdf, impression, synthèse vocale…) ou non-humains (représentations XML ou RDF selon les besoins, intégration des données dans les documents dans la logique de GRDDL).

2 Les frontières du document

La double contrainte de séparation entre le document et la page web d'une part et de conception d'un document composite multilingue d'autre part nous impose de préciser les « frontières d'un document ». Quelle relation particulière entretient une traduction avec l'original ? Le texte encodé en HTML et la version mise en page pour l'impression en pdf sont-ils un même ou bien deux documents différents ?

2.1 Functional Requirements for Bibliographic Records

Pour éclairer toutes ces questions, nous avons étudié la façon dont les bibliothèques approchent dorénavant la notion de « Document ». Alors que le catalogue traditionnel des bibliothèques ne connaissait que « l'exemplaire dans les mains du bibliothécaire », tendant à multiplier les fiches pour diverses éditions ou traductions d'une même œuvre, la tendance actuelle est l'équivalent d'une révolution copernicienne : au cœur de la nouvelle description documentaire est « L'Œuvre » (*Work*) dont les descriptions des ouvrages présents dans la bibliothèque vont dépendre. La logique catalographique reprend alors les pratiques des lecteurs : le moment de recherche globale d'une œuvre précède le choix d'une édition puis d'un exemplaire. Une place est accordée au travail intellectuel préalable à la réalisation d'une édition (une « manifestation » en FRBR-lingo), à la fois dans la définition du « travail » de création de l'original, mais aussi dans la conception de traductions ou les corrections liées aux diverses éditions d'une œuvre.

[1] PRISM, http://www.prismstandard.org/about/

[2] NewsML, http://newsml.org

Cette transformation a été menée par une réflexion de l'IFLA[3] sous l'intitulé « FRBR – *Functional Requirements for Bibliographic Records* ». Le rapport final du groupe de travail des années 90 [3][4] définit un modèle pour les notices bibliographiques basé sur des relations entre entités. Trois groupes d'entités sont définis :

- le groupe 1 définit les entités *Work*, *Expression*, *Manifestation* et *Item*, qui représentent les différents aspects de ce qu'un utilisateur peut trouver dans les produits d'une activité intellectuelle ou artistique ;
- le groupe 2 définit les entités *Person* et *Corporate Body* qui représentent les personnes physiques ou morales qui ont la responsabilité du contenu intellectuel ou artistique, de la production matérielle et de la distribution, ou de la gestion juridique des entités du premier groupe ;
- le groupe 3 définit les entités *Concept*, *Object*, *Event* et *Place* qui représentent les langages documentaires et les contenus d'indexation (relation « a pour sujet »).

Nous nous sommes pour l'instant intéressés de plus près au groupe 1 (*Work*, *Expression*, *Manifestation* et *Item*). Notons dès à présent que nous avons volontairement choisi les dénominations anglaises de ces entités. En effet la traduction française produite par la BnF utilise le terme « Document » pour désigner l'entité *Item*, ce qui nous semble en contradiction avec la définition élargie de Roger T. Pédauque [5] de ce qu'est un document numérique. Pour notre part, nous désignons « Document » l'ensemble de l'arbre composite. La figure 1 (extraite du rapport FRBR) explicite les relations entre ces quatre entités.

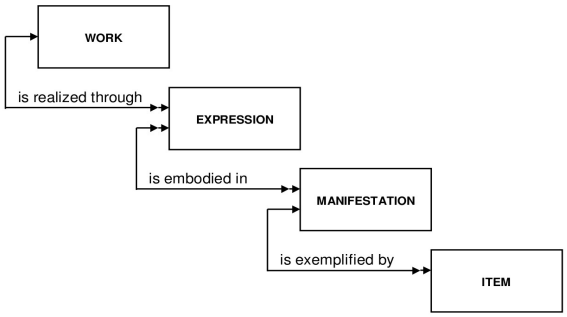

Figure 1 : Entités du groupe 1 et leurs relations

Les entités *Work* et *Expression* expriment le contenu intellectuel de l' œuvre. Ils sont abstraits, et ne comportent donc que des métadonnées documentaires :

- L'entité *Work* (Œuvre en français) représente une création intellectuelle ou artistique déterminée ;
- Une *Expression* est la réalisation d'une œuvre. Par exemple les versions linguistiques, ou les divers enregistrements d'une même œuvre sonore, ou diverses éditions revues et corrigées.

Les entités *Manifestation* et *Item* représentent la forme matérielle de l'œuvre :

- L'entité *Manifestation* exprime la représentation matérielle d'une *Expression* ;
- L'entité *Item* représente un exemplaire « physique » d'une *Manifestation.*

Par exemple le travail intellectuel *Madame Bovary* est représenté par une entité *Work*. L'*Expression* « originale » est la version française du roman. Une *Manifestation* pourrait être l'édition publiée par Flammarion, une autre *Manifestation* serait une édition de Poche, ou une édition électronique en PDF et une autre en format ePUB par exemple. Chaque *Manifestation* possède un ou plusieurs *Item* qui sont les exemplaires « physiques » présents sur les rayonnages, ou les sites de référence distribuant les versions électroniques. Une autre *Expression* serait par exemple une traduction en anglais, avec les métadonnées concernant la traduction (traducteur, année, etc.), et qui aurait elle-même plusieurs *Manifestations*...

FRBR permet donc de représenter des documents ayant différentes versions linguistiques, différents formats, mais plus intéressant encore, FRBR offre un point d'accès unique à une œuvre via l'entité *Work.*

2.2 FRBR et documents numériques

FRBR a été imaginé pour gérer des documents « physiques » présents sur des étagères de bibliothèque. Dans le cadre d'un Système de Gestion de Contenu ou d'une bibliothèque numérique, ce n'est manifestement pas le cas. De plus les catalogues de bibliothèques ne stockent que des métadonnées alors qu'un CMS doit stocker à la fois les métadonnées et le contenu d'un document. Comment alors adapter ce modèle pour gérer des documents numériques présents sur le Web ?

[3] International Federation of Library Associations and Institutions, http://www.ifla.org

Les entités *Work* et *Manifestation* représentant le travail intellectuel, ces deux entités peuvent être utilisées sans problème dans le cadre de documents numériques, les informations stockées étant des métadonnées. L'entité *Manifestation* exprime la représentation matérielle d'une *Expression.* Nous pouvons alors considérer que le représentation au format HTML (par exemple) est une *Manifestation* et que la représentation au format PDF en est une autre. Pour l'entité *Item*, qui représente un exemplaire avec un emplacement précis sur les étagères permettant de le trouver, nous utiliserons par analogie un URI qui permet de spécifier l'emplacement d'un document sur le web. Cette analogie permet alors à une *Manifestation* d'avoir plusieurs URIs associés, par exemple dans les cas suivants :

– document présent sur un serveur miroir (servant exactement les mêmes pages) ;

– document présenté avec un « environnement » différent, c'est-à-dire avec des informations périphériques dépendant du contexte d'affichage ;

– document publié par des sites tiers.

2.3 Gestion de documents composites et Sydonie

FRBR permet donc de modéliser des documents numériques en se concentrant sur l'aspect intellectuel du document via l'entité *Work*. Au travers des entités définies par FRBR, un document peut être représenté sous forme arborescente, comme illustré sur la Figure 2. Le modèle que nous choisirons est alors basé sur le même principe : nous considérons qu'un document est l'arbre complet tel que nous l'avons construit par analogie avec FRBR.

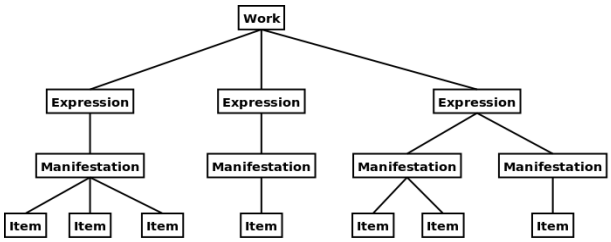

Figure 2 : Structure arborescente du Document

Ce choix permet à un document de connaître toutes ses versions linguistiques et formats de fichiers. Parmi les *Expressions* présentes, l'une d'entre elles a un rôle particulier puisqu'il s'agit de l'*Expression originale*, à partir de laquelle les autres *Expressions* ont été créées. De même pour chaque *Expression*, l'une des *Manifestations* a la qualité de *Manifestation référence* à partir de laquelle les autres *Manifestations* seront créées (que ce soit automatiquement ou non).

Pour la mise en œuvre informatique, un document est une entité abstraite composée de plusieurs objets suivant le schéma présenté Figure 3.

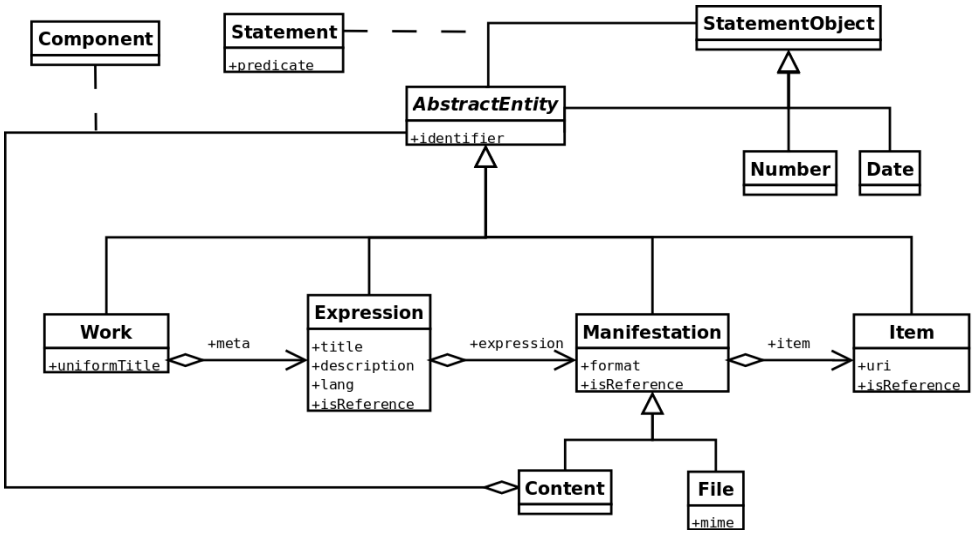

Figure 3 : Objets composant un arbre Document

Le logiciel de création de documents Sydonie en cours de développement utilise ce modèle pour stocker et gérer les documents créés. Nous devons y représenter les documents de façon générique (texte, image, vidéo, actualité, article, anthologie, etc.) grâce à un arbre en dérivant la classe abstraite *DocumentType* qui contient les propriétés et méthodes de base à tout type de document. Les spécificités d'un type de document sont gérées par un processus d'assertions indépendantes, ce qui favorise la modélisation en RDF des documents. Une telle conception du document permet d'approcher la notion de document composite. Par exemple, une image intégrée dans une page n'est plus considérée comme un simple lien vers un fichier numérique (modèle HTML) mais bien comme un document intégré, disposant de ses propres métadonnées, et pouvant éventuellement lui-même avoir plusieurs *Expressions*, cas d'un graphique avec du texte intégré, par exemple.

2.4 Publication Web

La séparation entre la création de documents via le web et la publication de documents impose l'utilisation de normes ouvertes afin de garantir l'interopérabilité entre les divers systèmes de publication. La norme PRISM [6] (Publishing Requirements for Industry Standard Metadata) définit un vocabulaire XML permettant la gestion et le stockage de documents composites pour la publication de magazines et de journaux. Son utilisation permet de gérer les documents composites au sens de Sydonie, tout en garantissant la possibilité d'importation par d'autres systèmes. Une *Manifestation* (c'est-à-dire le contenu d'un document) peut donc être stockée sous ce format, permettant ainsi l'échange de données. Diverses transformations peuvent ainsi être appliquées pour construire des présentations dans les divers formats professionnels tout en conservant le même lot de métadonnées prduites par Sydonie.

L'accès aux documents peut se faire à tout niveau dans l'arbre représentant le document mais la vue qui sera servie à un agent utilisateur (ou *User-Agent*[4]) est toujours la vue d'une *Manifestation*. Un utilisateur final accède donc à une *Manifestation* via un URI modélisé par un *Item*. Un système de gestion et de déréférencement des usages portant sur ces divers URI devra être mis en place, autour d'une logique de *Handler,* à l'image de ARK[5] ou du DOI[6], avec une numérotation unique. Nous envisageons la mise en place d'un modèle et d'un système associé à Sydonie (GONG : Gestion des Objets Numérique Généralisée).

De plus, une stratégie doit être mise en œuvre pour définir la *Manifestation* qui sera présentée quand une requête porte sur l'accès à un document au niveau de l'entité *Work* ou *Expression*. Le rapport *Cool URIs for the semantic web* [7], publiée par le W3C en décembre 2008, donne des indications sur ce type de stratégie. Le protocole HTTP permet à deux machines d'entrer en négociation pour déterminer le contenu adapté à une requête. Le serveur peut donc utiliser les *Language-Negotiation* et *Content-Negotiation* pour adapter la langue et le format de document à fournir au client :

- si un nœud de type *Work* est demandé, alors le processus *Language-Negotiation* permet au serveur de déterminer quelle *Expression* utiliser (en fonction des préférences de langues du client et des langues disponibles pour le document). Puis un *Content-Negotiation* permet ensuite de déterminer quelle manifestation servir au client (cf. ci après) ;
- si un nœud de type *Expression* est demandé, alors *Language-Negotiation* n'est pas nécessaire et seul le *Content-Negotiation* aura lieu. Un exemple type sera le cas d'un navigateur qui recevra du XHTML (c'est le type de contenu que ce client déclare préférer) alors qu'un robot recevra du XML ou du RDF ;
- si un nœud de type *Manifestation* est demandé, alors le serveur peut renvoyer directement le contenu sans avoir a effectuer de négociation.

Le système devra toujours préciser si la *Manifestation* servie dérive de l'*Expression* originale et si elle est une référence (telle que définie plus haut), et si ce n'est pas le cas indiquer un URI de la *Manifestation* référence pour l'*Expression* originale. Ce processus de négociation est réalisé en chaîne pour tous les composants d'un document, ce qui permet par exemple de servir les versions linguistiques adaptés pour les images ou vidéos.

3 Perspectives et travail futur

La conception d'un document comme un objet représentant un arbre complet d'entités FRBR correspondant à un travail intellectuel est très attirante. Un tel modèle peut être implémenté dans un Système de Gestion de Contenus pour répondre aux besoins de rédacteurs de *Manifestations* et être modélisé avec XML pour la publication et l'échange de documents.

[4] Définition de User-Agent sur Wikipedia : http://fr.wikipedia.org/wiki/User-Agent

[5] Archival Resource Key, http://www.cdlib.org/inside/diglib/ark/

[6] Digital Object Identifier, http://www.doi.org/

Le modèle proposé considère une *Manifestation* d'un document comme un composite avec des parties ou informations annexes qui sont elles-mêmes des documents basés sur le même modèle (c'est-à-dire eux-aussi représentés sous la forme d'un arbre d'entités FRBR). La mise en œuvre d'un tel modèle n'est pas simple, en particulier en termes d'ergonomie de l'interface de saisie. De notre point de vue, le modèle de saisie actuel de la plupart des CMS n'est pas adapté à cette conception. De nouvelles formes de saisies doivent être imaginées, combinant les possibilités des formulaires actuels, des éditeurs WYSIWYG en ligne (de type tinyMCE par exemple), des boîtes modales et des échanges Ajax avec le serveur.

Un mécanisme de saisie utilisant au mieux toutes ces possibilités devrait permettre de collecter des informations précises qui pourront être reprises dans des modèles de métadonnées comme RDF-a ou les microformats. La structure arborescente du modèle de document permet aussi d'éviter la duplication d'informations, lors de la saisie de nouvelles versions linguistiques par exemple, puisque les métadonnées au niveau *Work* sont déjà dans le système.

L'expérience des bibliothèques ouvrant de nouvelles perspectives dans la conception de systèmes de gestion de documents sur le web, le modèle présenté devra aussi intégrer les entités FRBR des groupes 2 (personnes) et 3 (sujets du document). Un groupe de travail sur l'harmonisation entre la classification FRBR des bibliothèques et la classification des objets de musée a publié un brouillon de $FRBR_{OO}$, une reformulation orientée objet du modèle entité-relations de FRBR sous la forme d'une ontologie [8]. Un tel accord montre le caractère fécond de l'approche FRBR qui peut s'étendre au-delà des catalogues de bibliothèque, ce qui nous semble justifier notre approche par analogie pour le document numérique.

L'expérience des médias imprimés, synthétisée dans la spécification de PRISM, permet d'établir un format pivot entre l'édition et la publication. Le modèle proposé doit pouvoir être étendu à d'autres formats professionnels, comme NewsML par exemple.

Conclusion

L'avenir de la publication sur Internet réside principalement dans la « re-publication » de tout ou partie de travaux intellectuels dans divers médias, sites, blogs, etc. Une architecture globale qui collecte les divers URIs présentant un document ouvre la possibilité de statistiques sur les usages du document. Couplé à l'utilisation de licences adéquates, cela permet d'envisager des négociations et transactions pour la publication des travaux d'auteurs.

Ni *framework* généraliste, ni CMS fermé, le projet Sydonie tente de fournir des outils adaptés à l'édition et la publication de documents numériques composites dans un environnement ouvert.

Bibliographie

[1] J-M. Lecarpentier, H. Le Crosnier et J. Madelaine, Évolutions de l'architecture du web et des documents numériques, *Traitements et Pratiques Documentaires. Vers un changement de paradigme? Actes de la deuxième conférence Document Numérique et Société*, pages 13-30, ADBS éditions, Paris 2008

[2] Chris Bizer, Tom Heath, Kingsley Idehen, Tim Berners-Lee, Linked Data on the Web (LDOW2008), *Proceedings WWW2008*, Beijing, China, 2008
http://www2008.org/papers/pdf/p1265-bizer.pdf

[3] IFLA Study Group, *Functional Requirements for Bibliographic Records*, K . G. Saur, München, 1998

[4] Groupe de travail IFLA, *Spécifications fonctionnelles des notices bibliographique,* BNF, Paris, 2001

[5] Roger T. Pédauque, *Le document à la lumière du numérique,* C&F éditions, 2006

[6] DEAlliance, *PRISM : Publishing Requirements for Industry Standard Metadata*, 2009,
http://www.prismstandard.org/specifications/2.1/

[7] W3C Interest Group, *Cool URIs for the semantic web*, W3C, 2008
http://www.w3.org/TR/cooluris

[8] WG on FRBR and CIDOC CRM hamonization, *FRBR object-oriented definition and mapping to* $FRBR_{ER}$,
http://cidoc.ics.forth.gr/docs/frbr_oo/frbr_docs/FRBR_oo_V0.9.pdf