



Office des publications

Direction Transformation objectif 2012  
Unité Architecture d'entreprise

# CELLAR et New ProCAT

## Note introductive

### Version 1.3

#### *Historique du document*

Version	Date	Auteur(s)	Remarques
0.1	24/06/2009	Peter Schmitz (PS)	Document de travail
0.2	25/06/2009	PS	Révision des schémas détaillés
1.0	29/06/2009	PS	Intégration des commentaires et remarques reçus
1.1	01/07/2009	PS	Intégration des commentaires d'Yves Steinitz
1.2	02/07/2009	PS	<u>Adaptations suite aux remarques du <i>steering committee</i>:</u> Schéma 1: enlèvement du terme 'management' pour éviter de la confusion; Schéma 2: ajout accès au contenu et aux métadonnées (flèche), adaptation des libellés au schéma 1 Schéma 3: séparation de la production et du repository commun des métadonnées Schéma 4: séparation de la production et du repository commun des métadonnées Ajout annexe 1: exemple d'une notice codée et décodée
1.3	15/07/2009	PS	Remplacement "New EUDOR" par "CELLAR"

---

Office des publications

---

2, rue Mercier L-2985 Luxembourg — Tél. +352 2929 - 42159 Fax +352 2929 - 42901

---

## 1. INTRODUCTION

La réalisation des projets CELLAR<sup>1</sup> et New ProCAT<sup>2</sup> s'inscrit dans le cadre du programme de transformation. En particulier, ils servent pour l'implémentation de la couche "Content and Metadata Manangement" qui représente une partie essentielle de l'architecture cible du programme. Le projet CELLAR concerne un repository commun pour tout contenu en format électronique publié par l'Office. Le projet New ProCAT concerne la refonte du système de gestion des métadonnées actuel pour tenir compte des évolutions des standards internationaux et de la technologie. Il vise à adapter les fonctionnalités aux besoins du programme de transformation, tout en simplifiant l'architecture du système. Il est évident que les deux projets sont complémentaires et doivent être cohérents.

Ce document présente une proposition de description fonctionnelle de l'architecture cible du programme de transformation et une ébauche de l'implémentation technique, en particulier en ce qui concerne la couche "Content and Metadata Manangement".

À partir du deuxième niveau de description fonctionnelle, les schémas montrent également le périmètre des projets CELLAR et New ProCAT en termes de fonctionnalité.

La proposition est complétée par des questions concernant l'orientation stratégique des systèmes.

---

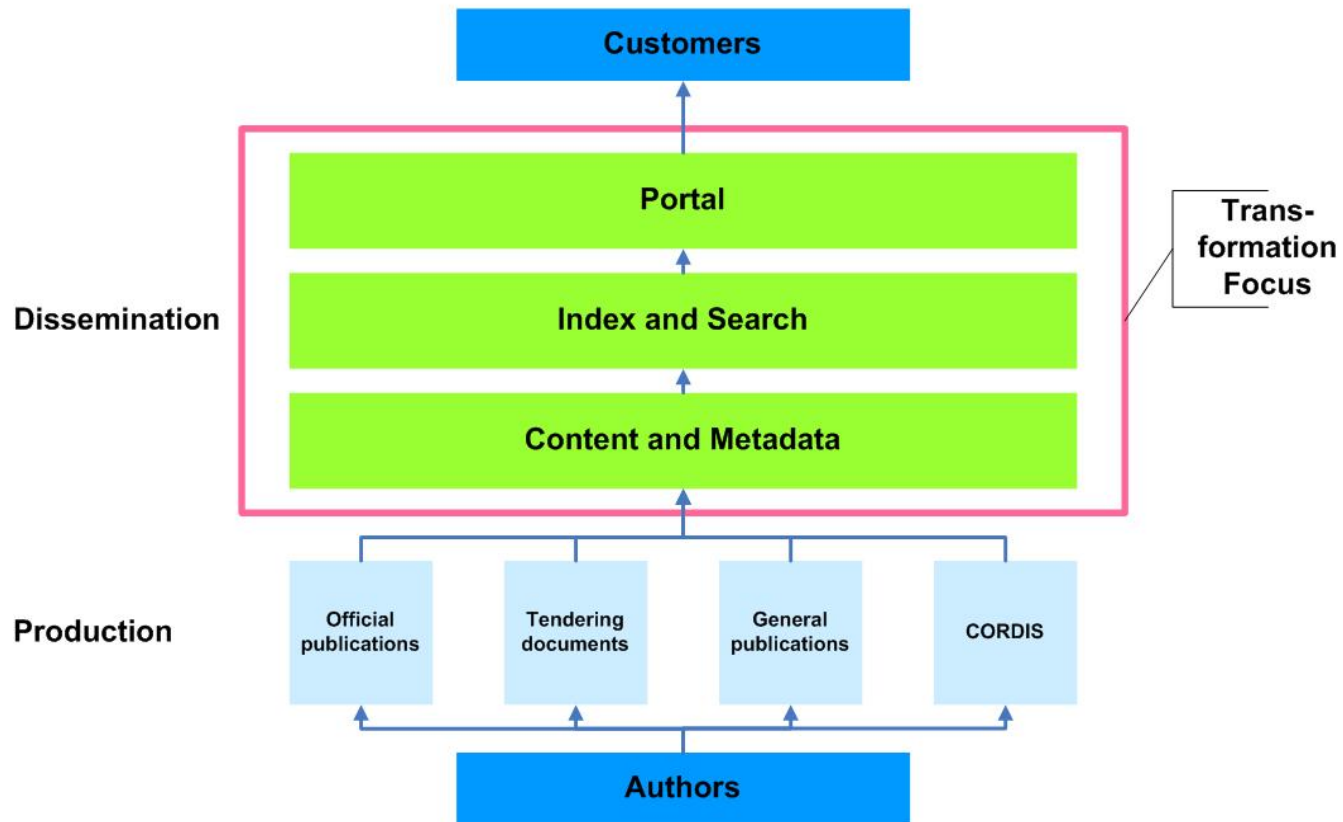
<sup>1</sup> Initialement le projet CELLAR a été appelé New EUDOR.

<sup>2</sup> Pour éviter des confusions avec des systèmes existants, un changement de nom du projet s'impose.

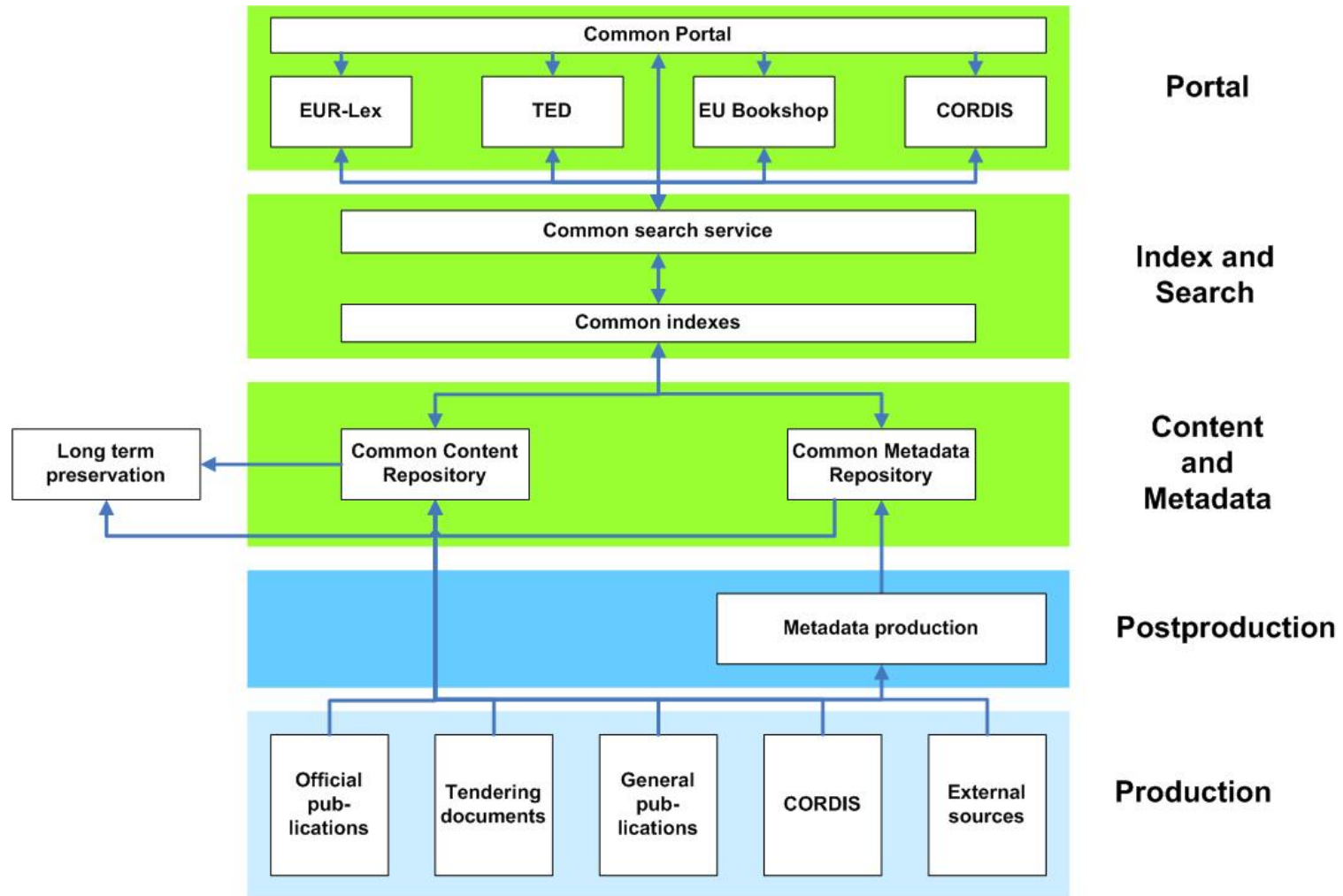
2. DÉCOMPOSITION FONCTIONNELLE DE L'ARCHITECTURE CIBLE DU PROGRAMME DE TRANSFORMATION

Pour rappel, le schéma 1 qui suit montre l'architecture cible du programme de transformation tel que défini par le management de l'Office.

Transformation program – target architecture



Le schéma 2 montre une première décomposition fonctionnelle du périmètre de transformation.



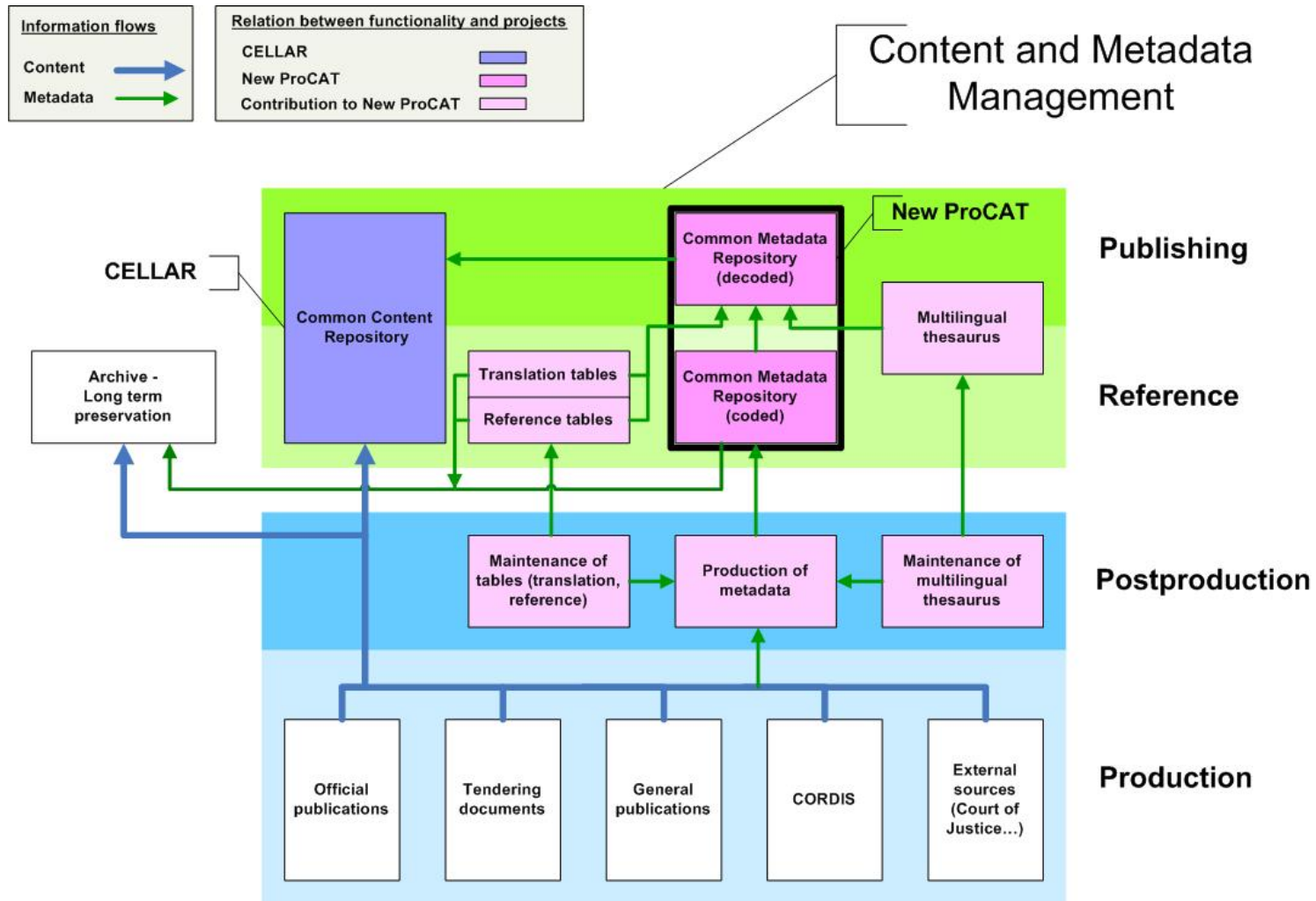
Au niveau de la gestion des portails on trouve un nouveau composant, le portail commun de l'Office, qui va permettre l'accès direct à l'ensemble du contenu publié par l'Office ainsi que l'accès aux portails spécifiques de l'Office. Les portails spécifiques restent bien évidemment directement accessibles pour faciliter l'accès aux publics spécialisés. Les portails spécifiques existants doivent être adaptés pour supporter la nouvelle architecture ainsi que pour intégrer des composants génériques qui seront choisis ou respectivement développés pour supporter des fonctionnalités standards d'un portail: gestion des utilisateurs et des souscriptions, gestion du site et du contenu éditorial (web content management), réseaux sociaux, analyses et statistiques sur la fréquentation des sites.

La couche de recherche et d'indexation peut être décomposée en deux parties. D'abord il y a un service de recherche commun qui sera accessible par une interface standardisée permettant de formuler des requêtes indépendantes des moteurs de recherche propriétaires. La requête sera exécutée par un moteur de recherche adéquat. Pour répondre à des requêtes complexes, la combinaison de différentes méthodes de recherche doit être possible. Des types de recherche complémentaires doivent être supportés: la recherche dans le texte intégral (*full text search*) et celle basée sur les relations sémantiques identifiées au niveau des métadonnées (*semantic search*). La recherche s'appuie sur un ensemble d'index communs qui seront générés sur base du contenu et des métadonnées. La synchronisation en temps réel entre index et données doit être garantie.

Les couches portail et index se basent sur une couche de gestion et de stockage de contenu et des métadonnées. À ce niveau, il y a un repository commun de contenu et un repository commun pour les métadonnées. Ces bases de données communes seront complétées par une archive pour assurer la préservation à long terme du contenu et des métadonnées.

La production des métadonnées est assurée par l'Office et se situe dans la phase de postproduction.

Le schéma 3 montre une deuxième décomposition fonctionnelle de la couche de gestion du contenu et des métadonnées.

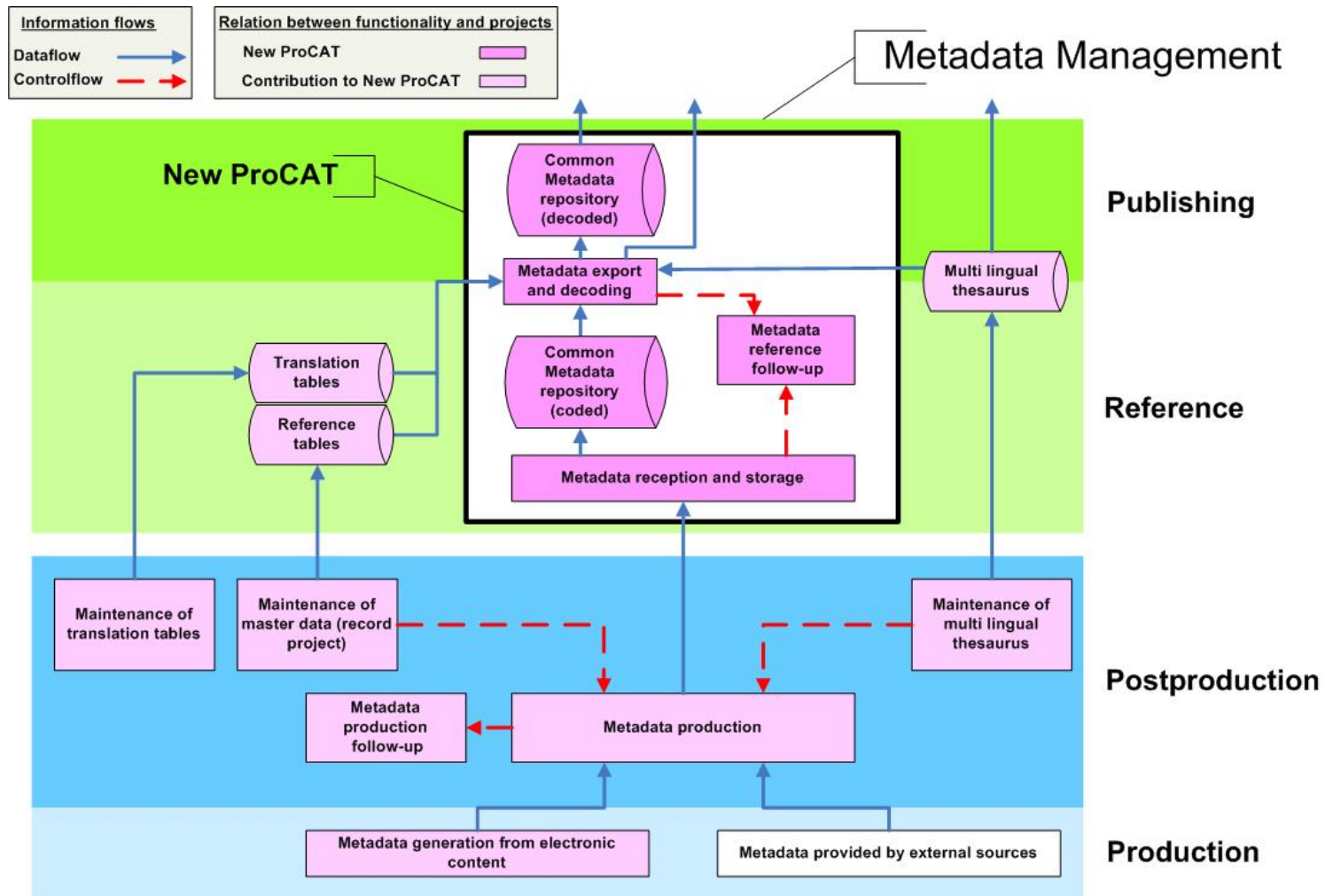


Le schéma montre que la production des métadonnées s'appuie sur des données de référence (projet RECORD) et sur le thesaurus multilingue (EUROVOC). L'information de base vient des piliers de la production. Les résultats de la production des métadonnées sont stockés dans un repository commun, qui contient tout le fond documentaire de l'Office. À ce stade, certaines données restent codées (auteurs, langues etc.). Le repository de diffusion est alimenté à partir du repository commun. Le processus d'alimentation du repository de dissémination assure également le décodage des données sur base des tables de référence, des tables de traduction et du thesaurus multilingue. Les codes seront remplacés par des libellés dans toutes les langues à supporter par le système (c.-à-d. au minimum toutes les langues officielles de l'Union Européenne). Les métadonnées décodées sont également attachées au contenu. Les études techniques à mener détermineront s'il faut implémenter un repository dédié pour le stockage de métadonnées codées ou si la présence des métadonnées décodées dans le repository commun du contenu est suffisante. Les métadonnées codées et décodées doivent être archivées (archive long terme).

La gestion des métadonnées sera basée sur un concept de versions. Les versions sont nécessaires pour assurer la synchronisation entre les différents systèmes et pour déterminer le périmètre de la dissémination.

Les vues détaillées de la gestion des métadonnées et du contenu sont montrées par les schémas suivants.

Schéma 4: Vue fonctionnelle de la gestion des métadonnées



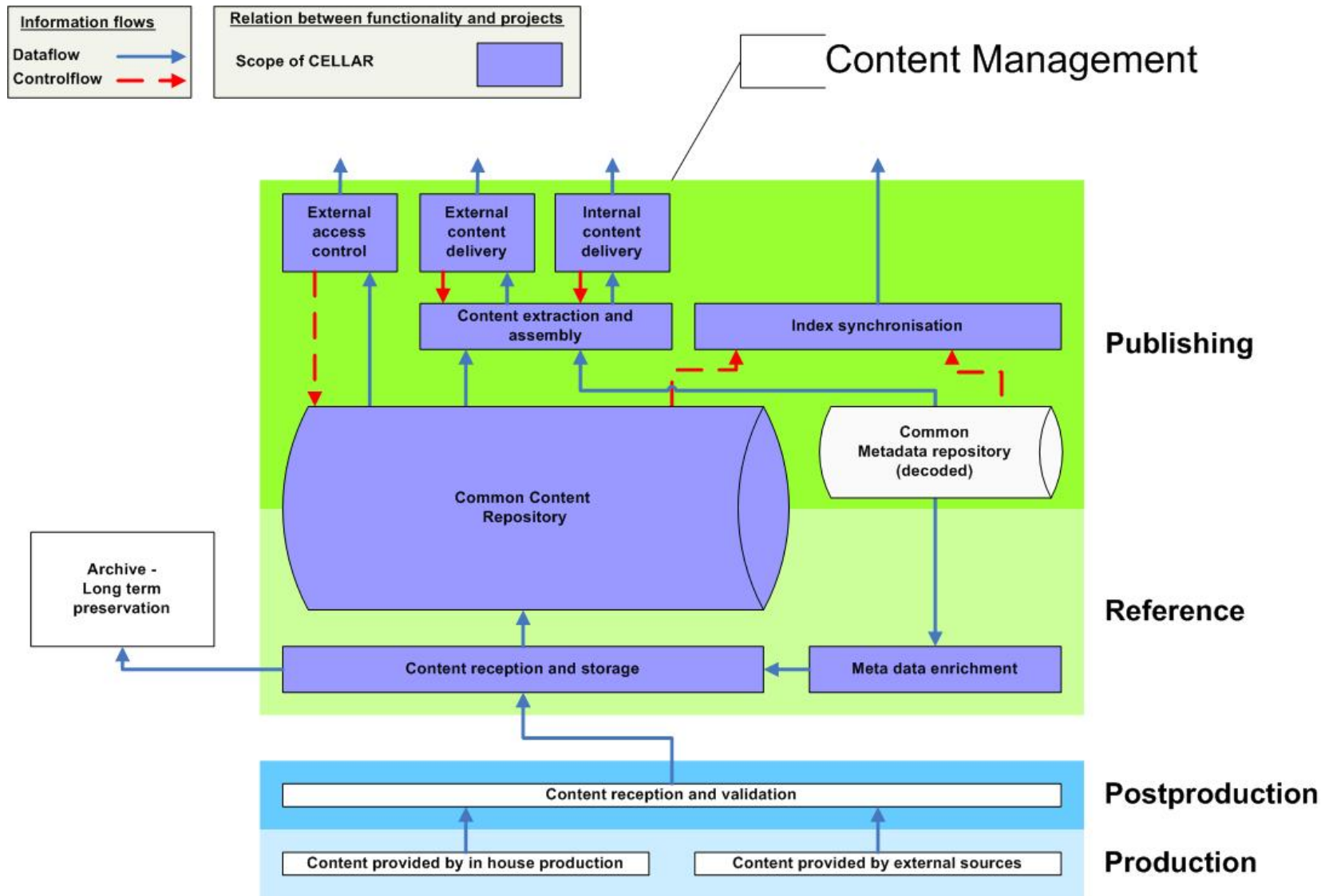
Les métadonnées sont générées à partir du contenu électronique produit par l'Office ou fourni par des sources extérieures. On peut distinguer des niveaux de qualité différents de métadonnées. Le niveau de qualité le plus bas est celui qui remplit les critères minimum pour assurer une première dissémination. Ensuite la qualité est améliorée successivement. Chaque niveau de qualité utile est disséminé et remplace la version antérieure. Les niveaux de qualité sont implémentés par un concept de version.

Pour faciliter la production des métadonnées et pour augmenter la qualité des données, il faut faire le maximum de la production sur base des formats XML standardisé (FORMEX). Un balisage adapté du contenu permettra à terme de générer la majorité des métadonnées directement à partir du contenu. Pour les contenus qui n'existeront pas en format XML mais en format PDF, une standardisation de la partie bibliographique des documents PDF au sein de l'Office facilitera également une génération automatique des métadonnées.

Il est prévu d'utiliser un système intégré de gestion de bibliothèques pour réaliser la gestion des métadonnées: production, enrichissement et validation. L'étape de validation comprend notamment la vérification de la conformité avec les données de référence telle que définie par le projet "RECORD". L'Office a sélectionné le logiciel Virtua ILS pour supporter ces fonctionnalités.

Si une nouvelle version d'une notice bibliographique a été créée, elle peut être chargée dans le repository commun des métadonnées. Le processus de chargement engendre une validation des données. Chaque interaction avec le repository commun est tracée par un module de suivi.

Un module d'exportation et de décodage assure la transformation de notices codées en notices décodées. Ce module doit également assurer des exportations en masse des notices (codés ou décodés) à toutes fins utiles.



Le contenu qui a été accepté par les processus de validation dans la phase de postproduction peut être chargé dans le repository commun de référence et de dissémination. Une partie du contenu sera également envoyée à l'archive long terme sur base des règles d'archivage à définir. Pour faciliter l'accès au contenu, la transformation du format XML de production (FORMEX) dans un format XML de dissémination (FORMEX lite) doit être envisagée. Ce format doit également remplacer les formats intermédiaires qui sont actuellement générés pour l'indexation du texte intégral (texte index).

Un module de synchronisation des index assurera la visibilité du contenu à partir des portails de l'Office ainsi qu'à partir de tout autre système qui est autorisé à accéder la couche des index communs de l'Office.

Un module d'extraction de contenu va assurer la mise à disposition du contenu aux systèmes internes et externes. En tant que services supplémentaires, des transformations du contenu peuvent être offertes. La liste des transformations à offrir reste à définir.

Un module de contrôle des accès externes facilitera l'accès au contenu à tout demandeur autorisé par l'Office (moteurs de recherche Internet, *information brokers*, etc.).

L'accès au contenu par des systèmes externes sera permis sans restriction. Néanmoins il faut définir et implémenter des contrôles pour éviter une dégradation du service causé par des accès externes.

À terme, le repository commun de contenu va inclure toutes les publications dans tous les formats électroniques publiés par l'Office. Il reste à analyser si chaque objet du contenu doit être traité au même niveau où s'il serait utile d'introduire une classification des objets pour augmenter la pertinence des index et des recherches. Cette classification peut également avoir un impact sur l'implémentation technique du repository (combinaison de plusieurs technologies de stockage).

## 3. ANNEXE 1: EXEMPLE DE NOTICE CODÉE ET DÉCODÉE

## 3.1. a) Notice de la publication KINA22040ENC codée

020			\$a EN 92-79-01414-5
029	7		\$a EN KI-NA-22-040-EN-C \$2 EN OPC
035			\$a EN Fabrice \$x EN 2006-04-12
041	0		\$a EN eng
084			\$a EN M06 \$2 EN OPC
099			\$a EN 2006/2295 \$2 EN GESCOM
245	1	n	\$a EN European SmartGrids Technology Platform \$b EN vision and strategy for Europe's electricity networks of the future
249			\$a EN EUR 22040 EUROPEAN TECHNOLOGY PLATFORM SMARTGRIDS - ELECTRICITY
260			\$a EN {LUXB} \$b EN OPL \$c EN 2006
300			\$a EN 37 NPAG \$b EN NFIG COL, NPHT \$c EN A4 \$d EN AG
440		n	\$a EN EUR_CE_C

			\$v EN 22040
520			\$a EN Whilst current electricity networks presently have fulfilled their function effectively, more of the same will not be sufficient to meet current challenges and policy imperatives. In this context, the European Technology Platform (ETP) SmartGrids was set up in 2005 to create a joint vision of European networks for 2020 and onwards. The platform includes representatives from industry, transmission and distribution system operators, research bodies and regulators. It has identified clear objectives and proposes a strategy for the development of future electricity networks.
540			\$a EN REPRO1 \$b EN CEE
540			\$a EN REPRO \$b EN Belpress.com \$3 EN NFIG
710	2		\$a EN CEU \$b EN RTD
773	1	8	\$t EN EUR_CE_C \$q EN 2006, NPER 22040
856	4	1	\$q EN PDF \$s EN 1,72 MB \$u EN <a href="http://europa.eu.int/comm/research/energy/pdf/smartgrids_en.pdf">http://europa.eu.int/comm/research/energy/pdf/smartgrids_en.pdf</a>
910			\$a EN GR
920			\$a EN 704

**3.2. Notice de la publication KINA22040ENC décodée**

=LDR 01880nam 22003134i 4500

=001

=005 20060404220334.0

=008 060412s2006\\\$eu\ado\fr\\\$i000\0\eng\ d

=020 \\\$a92-79-01414-5

=029 7\\\$aKI-NA-22-040-EN-C\$2LU-LuOPE

=040 \\\$aLU-LuOPE

=041 0\\\$aeng

=044 \\\$aeu

=084 \\\$aEnergy research\$2LU-LuOPE

=245 10\\\$aEuropean SmartGrids Technology Platform\$bvision and strategy for Europe's electricity networks of the future

=260 \\\$aLuxembourg\$bPublications Office\$c2006

=300 \\\$a37 p.\$bill. col., photos\$c21.0 × 29.7 cm\$dstapled

=440 \0\\\$aEUR / European Commission\$v22040\$x1018-5593

=520 \\\$aWhilst current electricity networks presently have fulfilled their function effectively, more of the same will not be sufficient to meet current challenges and policy imperatives. In this context, the European Technology Platform (ETP) SmartGrids was set up in 2005 to create a joint vision of European networks for 2020 and onwards. The platform includes representatives from industry, transmission and distribution system operators, research bodies and regulators. It has identified clear objectives and proposes a strategy for the development of future electricity networks.

=540 \\\$aReproduction is authorised provided the source is acknowledged\$bEuropean Communities

=540 \\\$aAll rights reserved\$bBelpress.com\$3ill.

=650 \7\$81.1\7\$aenergy research\$2Eurovoc

=650 \7\$81.2\7\$aenergy grid\$2Eurovoc

=650 \7\$81.3\7\$aelectrical energy\$2Eurovoc

=710 2\7\$aEuropean Commission\$bDirectorate-General for Research

=773 18\$tEUR. European Commission\$q2006, No 22040\$x1018-5593

=856

41\$qPDF\$s1,72

MB\$uhttp://europa.eu.int/comm/research/energy/pdf/smartgrids\_en.pdf\$uhttp://bookshop.eu.int/eGetRecords?Template=Test\_EUB/en\_publication\_details&CAT

NBR=KINA22040ENC

=910 \\\$aFree

=920 \\\$aScientific