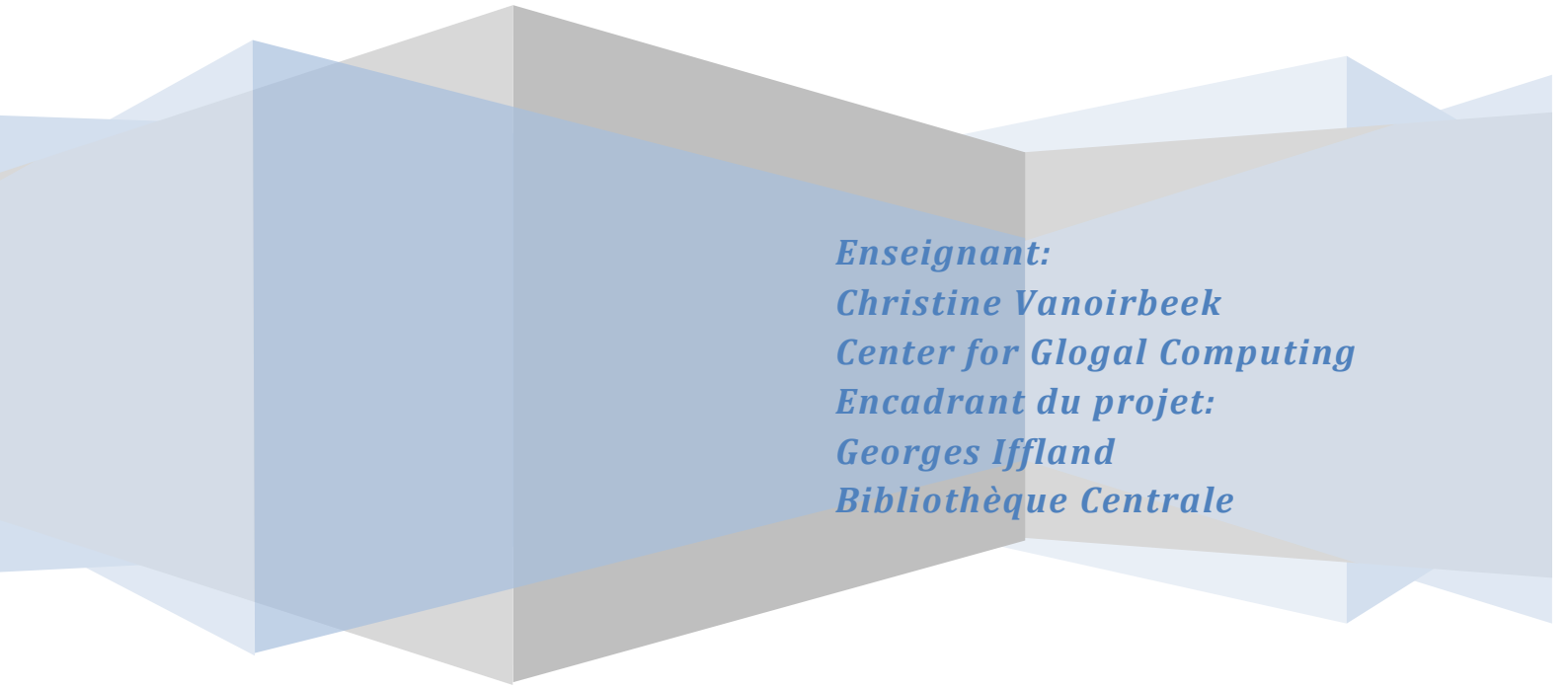
Projet de semestre

Analyse et optimisation du processus de publication électronique des thèses de l'EPFL et de récolte de métadonnées

Bogdan STEFANESCU



Enseignant:
Christine Vanoirbeek
Center for Glogal Computing
Encadrant du projet:
Georges Iffland
Bibliothèque Centrale

Sommaire

Contexte

La Bibliothèque Centrale est responsable du dépôt légal des thèses à l'EPFL. Elle maintient un catalogue complet qui est enrichi par des données obtenues du Service Académique.

La Bibliothèque propose sur son site Web le texte intégral sous forme de fichiers PDF, de la totalité des fonds (plus de 3800 thèses). La thèse est accessible sur le Web dans le monde entier, avec l'accord de l'auteur, sinon en intranet seulement. Dans ce dernier cas, la Bibliothèque fournit à la demande d'un exemplaire numérique sur une procédure manuelle. La Bibliothèque assure un archivage de toutes les thèses, mais selon un accord national, c'est la Bibliothèque Nationales Suisse qui se charge de l'archivage à long terme.

1. L'analyse du workflow actuel de publication et de diffusion des thèses

Le workflow de publication des thèses de l'EPFL est un processus complexe, processus qui implique plusieurs acteurs et étapes sur une période de temps d'approximativement 50 jours, période entre le premier dépôt de la thèse et sa publication électronique.

1.1 Les acteurs de la thèse

Dans le processus de publication de la thèse 4 acteurs ont été identifiés : le Doctorant, le Service Académique, le Service de la Reproduction et la Bibliothèque Centrale. Chaque acteur a un rôle précis, rôle qui est distribué à différentes personnes de chaque établissement, sauf le doctorant qui est une seule entité. En considérant la période de publication, on peut observer que les étapes sont conditionnées par des actions séquentielles mais la structure du workflow n'est pas séquentielle parce que les acteurs s'intercalent afin de minimiser le temps de publication.

1.2 Les étapes principales du workflow de publication

Les principales étapes de la thèse sont analysées en fonction de chaque acteur. Afin de valider une étape et de passer à l'étape suivante, l'acteur doit finaliser l'ensemble des actions qui lui sont distribuées.

Les principales étapes de la thèse ainsi que les actions que chaque acteur doit les réaliser sont synthétisées dans le tableau suivant.

Etape	Acteur	Action	Période
1. Premier Dépôt	Doctorant	-Premier dépôt de la thèse 22 jours avant l'examen final	Jour 1
2. Processus du suivi	SAC	-Avis au SISB de la date de l'examen -Saisie dans fichier XL «Thèses en cours »	Jour 1
3. Saisie des MD	SISB	-Saisie dans NEBIS d'une notice	Jour 1

		d'acquisition -Saisie des MD dans la base des thèses FileMaker -Proposition de diffusion Internet envoyé au doctorant	
4.Soutenance de la these	Doctorant	-Soutenance publique -Dépôt version finale au Service de la Reproduction -Accord de diffusion	Entre 3 semaines et 6 mois après l'examen oral
5. Traitement de la thèse	Service de la reproduction	-Contrôle de la version finale -Ajout de la page de titre officielle -Envoi du bon à tirer au doctorant -Dépôt sur le serveur Répro des fichiers PDF à l'attention du SISB	Environ 1 mois après
6. Administration+ Traitement de la thèse	SISB	-Récupération des fichiers -Création dossier par thèse -Saisie des MD dans la base FM -Event. Redimensionnement A4-A5 -Renumérotation PDF -Suppression MD non-pertinentes du PDF -Extraction d'abstract destine à la diffusion -Saisie dans FM des problèmes rencontrés -Mise a disposition des fichiers PDF « abstracts » pour traitement HTML	Entre 1 et 15 jour après
7. Processus d'impression	Service de la Reproduction	-Impressions légales	1-6 mois après l'étape 4
8. Encodage HTML	SISB	-Langue des résumes -Réalisation de l'encodage HTML -Vérification du codage HTML -Saisie de mots-clés dans la base FM	1-3 jours après l'étape 6
9. Publication sur le WEB	SISB	-Transfert des fichiers sur le serveur -M-à-J NEBIS, FM -Avis email au doctorant de fin de travaux et demande de feedback	1-2x mois avec possibilité de traitement urgent
10.Catalogage	SISB	-Accusé de réception SAC -indexation et codes d'acquisition -Dépôt légale M-a-J fichier XL « Thèses en cours » -Expédition des exemplaires a la Bibliothèque Nationale	16-22 jours après

Source: Tableau synoptique processus thèses EPFL/FSchmitt - juillet 2008

Le workflow de publication des thèses est présenté dans l'image suivante.

A gauche on peut observer les acteurs, en centre les étapes et à droite les actions attribuées à chaque étape.

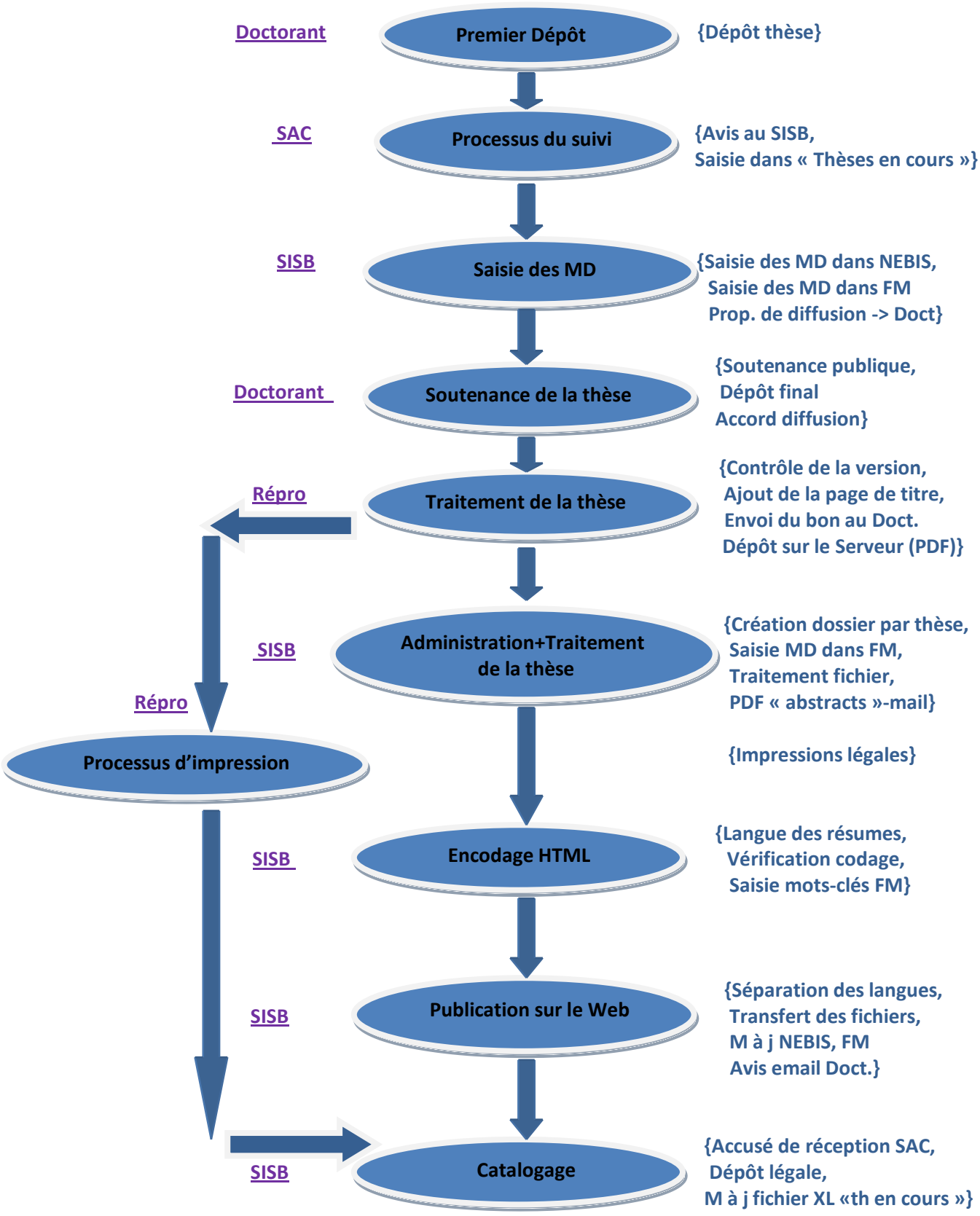


Image : Workflow de publication des thèses de l'EPFL

1.3 Les serveurs du processus de publication

Afin que les thèses soient publiées plusieurs serveurs sont utilisés dans les workflow de publication.

Le serveur IS-Academia est utilisé par le SAC et contient les informations principales sur le doctorant.

Le serveur de la Reproduction est utilisé par le Service de la Reproduction pour déposer les fichiers PDF à l'attention du SISB.

Le serveur d'administration FM est utilisé par la Bibliothèque Centrale en local pour la publication en Intranet de thèses. Toutes les métadonnées de la thèse sont stockées sur ce serveur. Le serveur d'administration permet aussi l'export vers InfoScience.

Le serveur WEB de publication (library.epfl.ch) est utilisé par la Bibliothèque Centrale pour la publication des thèses sur Internet.

Les relations entre le serveur de la Reproduction, le serveur d'administration et le serveur WEB sont présentées dans l'image suivante :

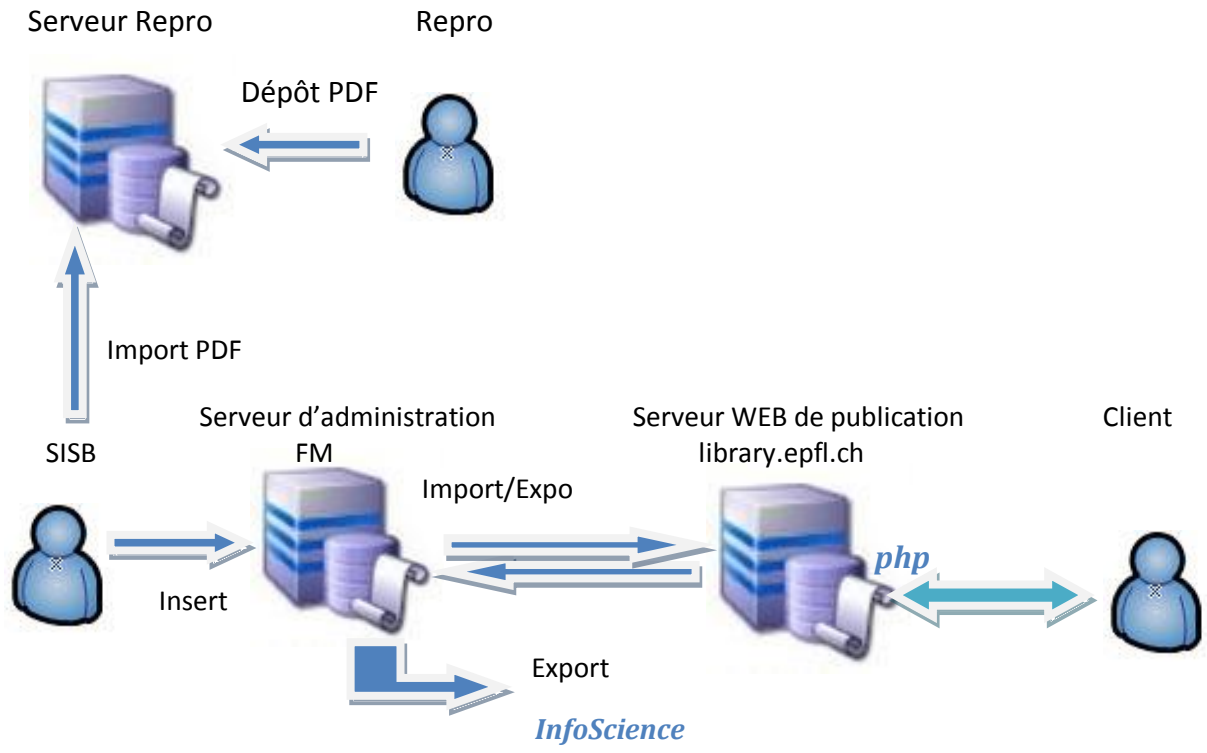


Image : Les relations entre les différents serveurs

1.4 Le workflow de diffusion de thèses

Le workflow de diffusion de la thèse concerne un seul acteur, la Bibliothèque centrale. Après la réception de l'email de commande sur l'adresse theses.bc@epfl.ch il y a une vérification de la validité de la demande. La personne qui demande la thèse doit être d'accord avec les conditions juridiques. La thèse est envoyée en pièce jointe à l'adresse d'email du demandeur. Les emails sont marqués comme envoyés, retournés ou terminés. Après l'envoie d'email il y a une saisie dans la base FM des données du client et de numéro de la thèse commandée. Le nombre de commandes est enregistré dans un fichier Excel.

L'image suivante présente les principales étapes et les actions à faire pour chaque étape.

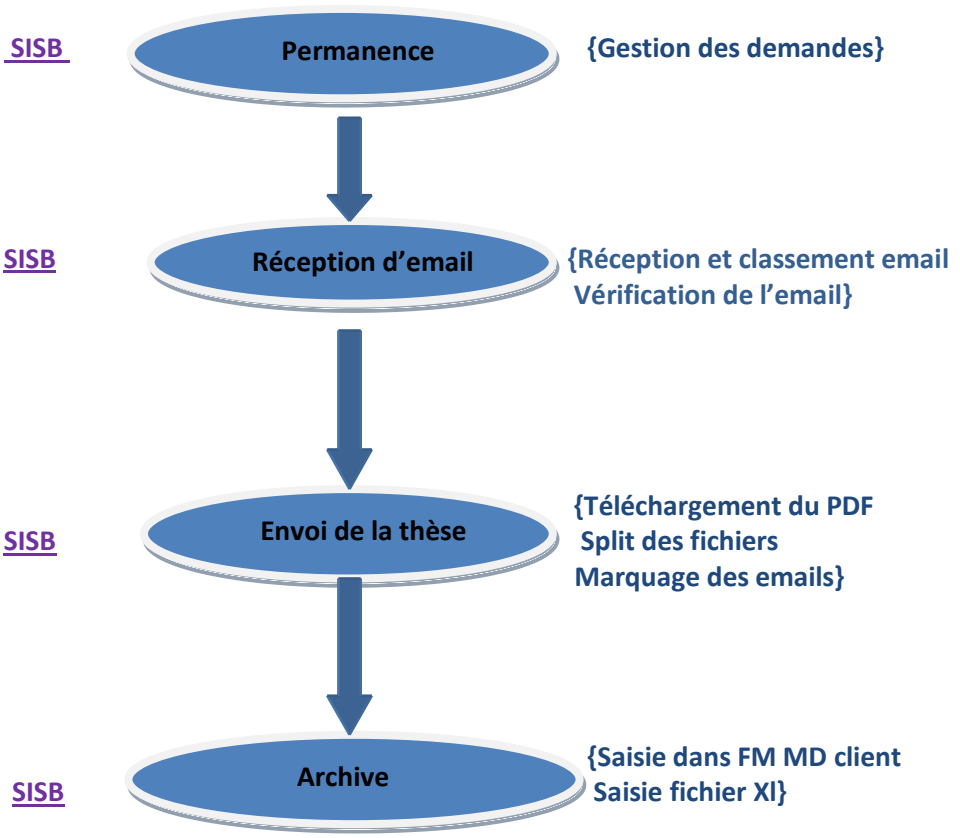


Image : Workflow de diffusion des thèses de l'EPFL

1.5 Les inconvénients du système

Le trajet de la thèse

L'observation principale qu'on peut faire sur le système actuelle est le travail redondant entre les différents acteurs de la thèse qui est traduit par une augmentation du temps de publication.

Une fois que le doctorant a fait le premier dépôt de sa thèse 22 jours avant la date de l'examen oral en complétant un formulaire les données sont recopiées sur le serveur IS-Academia. La Bibliothèque Centrale n'a pas accès au serveur du SAC, les métadonnées de la thèse sont transmises à l'aide d'un bordereau et les mêmes métadonnées sont recopies dans le serveur d'administration de la bibliothèque. La recopie des métadonnées de la thèse provoque des erreurs, erreurs qui sont traduits par une augmentation du temps parce que à chaque fois les données doivent être vérifiées. Un système centralisé avec un accès commun en fonction des droits spécifiques peut résoudre le problème.

Apres le dépôt de la thèse, le Service de la Reproduction refait la page de couverture. Le même travail est déjà fait par le doctorant mais sous son propre modèle. Si le doctorant a un modèle standardisé de la page de couverture il peut le faire lui-même en respectant le modèle.

La thèse arrive au Service de la Reproduction sous plusieurs formes. Elle est composée parfois de plusieurs fichiers, le Service de la Reproduction ajout la page de couverture qui est un fichier en plus et la thèse arrive à la Bibliothèque Centrale composée de plusieurs fichiers. C'est la Bibliothèque qui regroupe la thèse dans un seul fichier PDF, et qui fait un redimensionnement de la thèse si c'est le cas. Si c'est le doctorant qui fait le regroupement de la thèse dans un seul fichier et le redimensionnement, si nécessaire, alors le temps de publication peut être réduit.

Le même système centralisé peut résoudre aussi le problème de publication de la thèse sur le web. La mise à jour des thèses est faite sur le serveur d'administration en deux endroits différents : Internet et Intranet. Les thèses qui ont le droit de publication sont recopiées sur le serveur de publication Web. Au lieu d'avoir deux serveurs, l'un d'administration et l'autre de publication sur le Web on peut avoir un seul serveur avec des droits différents si la thèse est publiée en Intranet ou en Internet.

Un dernier problème concerne le workflow de diffusion de thèses. Apres la réception de l'email de commande à la Bibliothèque centrale, la personne responsable envoie la thèse en pièce jointe. Pour les gros fichiers PDF, il est nécessaire de faire un split de fichier afin que l'attachement soit possible. Le nouveau système qui va être mis en place consiste à envoyer directement le lien vers la thèse après la vérification de la validité de la demande.

2. Les métadonnées de la thèse

2.1 Classification et optimisation des métadonnées

En analysant les 220 métadonnées de la thèse utilisées dans le système File Maker, certains d'entre elles doivent être réorganisés afin de simplifier le processus de publication. Certains métadonnées ont été crées au cours du temps pour simplifier les processus de publication. Ces métadonnées ont été calculées à partir des métadonnées déjà indexées. Le processus de publication a changé au cours du temps et les métadonnées de la thèse aussi. Certains ont restées dans le système mais caractérisent en fait des thèses déjà publiées et maintenant elles ne sont pas utilisées.

Le nombre de métadonnées a été réduit en conservant seulement celles qui caractérisent maintenant le processus de publication de la thèse. Les métadonnées ont été reparties en 2 catégories :

Métadonnées Administratives

Métadonnées qui caractérisent les acteurs externes de la thèse.
Exemple : Auteur, Directeur de la thèse, Faculté, Institut

Métadonnées Descriptives

Métadonnées qui décrivent la thèse.
Exemple : Titre, Sous-titre, Abstract, Classification

Dans le processus de publication de la thèse ont été identifies 3 acteurs principaux qui saisissent des métadonnées. Les 3 acteurs sont : le Doctorant, le Service Académique, la Bibliothèque Centrale.
Note : Dans le processus de publication de la thèse 4 acteurs ont été identifiés : le Doctorant, le Service Académique, le Service de la Reproduction, la Bibliothèque Centrale.

2.2 Le standard XMetaDiss

XMetaDiss est le format d'échange commun pour les thèses en ligne des universités suisses transmises à la Bibliothèque nationale suisse. Ce standard est obligatoire pour tous les types de publications universitaires transmises à la Bibliothèque nationale suisse.

Certaines métadonnées utilisées maintenant dans le système File Maker peuvent être standardisées en utilisant le modèle proposé par la Bibliothèque nationale suisse. D'autres métadonnées ont été ajoutées pour permettre l'export vers la Bibliothèque Nationale. Il y a aussi des métadonnées locales qui ont été ajoutées pour faciliter la communication entre les différents acteurs.

Les espaces de noms

Les éléments qui commencent par „***dc***", sont issus du standard Dublin-Core-Set. Les métadonnées qui caractérisent les personnes commencent par „***pc***" et font partie du

standard XMetaPers, les métadonnées pour les institutions et les collectivités commencent par „***cc***" et représentent le standard « Metadata for corporate bodies ». Les noms des éléments nécessaires au „chemin" du document et des éléments pour l'archivage à long terme, commencent par "ddb" et sont issus du standard «Long-term preservation and administrative metadata ». Les éléments qui commencent par „thesis" font partie du standard « Metadata Standard for Electronic Thesis and Dissertations »

2.3 La distribution de métadonnées en fonction d'acteur

Le tableau suivant synthétise les métadonnées utilisées ainsi que les acteurs impliqués.

Description	MD actuelle BC FM	MD standardisé BN	Statut	Acteur
Titre	Title	dc:title	obligatoire	Doctorant
Sous-titre	Subtitle	dcterms:alternative	facultatif	Doctorant
Titre dans d'autres langues	title_translated	dc:title ddb:type="translated"	facultatif	Doctorant
Auteur	author_firstname author_lastname	pc:forename pc:surName	obligatoire	Doctorant
Date de naissance	author_birthday_sac	pc:dateOfBirth	facultatif	Doctorant
Lieu de naissance	author_birthplace	pc:placeOfBirth	facultatif	Doctorant
Adresse		pc:address	facultatif	Doctorant
Classification		dc:subject xsi:type="dcterms:DDC	obligatoire	Doctorant
Mot-matière	Keywords	dc:subject xsi:type="xMetaDiss:noScheme	obligatoire	Doctorant
Abstract (plein texte)		dcterms:abstract xsi: type="ddb:noScheme	facultatif	Doctorant
Abstract (URL)	abstract_pdf_url abstract_html_url	dcterms:abstract xsi:type=" ddb:contentISO639-2 " type="dcterms:URI	facultatif	SISB
Mention d'acceptation de l'unité de recherche qui accepte la thèse		ddb:note	facultatif	SAC
Nom de l'institution déposante		cc:name cc:place cc:address	obligatoire	Automatique
Numéro ID de l'institution déposante		ddb:contact	obligatoire	Automatique
Personne de contact de l'institution déposante (e-mail)		ddb:description	obligatoire	Automatique

Lieu de la haute école, haute école, faculté	doctoral_school faculty institut	cc:name cc:place cc:department	facultatif	Doctorant
Directeur de thèse	Advisor	pc:foreName pc:surName	facultatif	Doctorant
Date de l'examen oral/date de la thèse	date_oral_sac	dcterms:dateAccepted	obligatoire	SAC
Date de l'examen oral/date de la thèse		dcterms:dateSubmitted	Facultatif	SAC
Date de la remise valable des exemplaires, respectivement des fichiers du dépôt légal		dcterms:created	facultatif	SAC
Type de la publication universitaire		dc:type thesis:level	obligatoire	Automatique
Adresse Internet	fulltext_html_url fulltext_pdf	ddb:identifier	obligatoire	SISB
Institution exploitant le serveur et lieu		ddb:server	facultatif	SISB
Nombre de fichiers		ddb:fileNumber	obligatoire	SISB
Nom du fichier		ddb:fileProperties	obligatoire	SISB
Indications sur le propre ou un autre persistent identifier		ddb:identifier ddb:type="other"	facultatif	Automatique
Date de livraison des métadonnées		ddb:dateDelivered	facultatif	BN
Langue de la publication universitaire	lang_text	dc:language	obligatoire	Doctorant
Relations avec d'autres documents (texte libre)		dc:relation	facultatif	Doctorant
Conditions juridiques (plein texte)		ddb:rights	obligatoire	Automatique
Conditions juridiques (URL)		dcterms:accessRights	facultatif	Automatique

3. Comparaison avec ce qui se fait ailleurs

3.1 Le projet ORI-OAI

L'Outil de Référencement et d'Indexation (ORI-OAI) vise la mise en place d'un système ouvert, en open source, permettant :

- de **gérer** tous les documents numériques produits par les établissements universitaires,
- de les **partager** avec d'autres établissements,
- de les **valoriser** par une indexation professionnelle,
- de les **rendre accessibles**, à distance et selon les droits définis, dans des interfaces ergonomiques.

3.2 Le Workflow de thèses en France

La version 2.0 du projet, propose d'intégrer le nouveau format TEF (Thèses Electroniques Française) dans le cadre du projet ORI-OAI. L'outil va être utilisé dans plusieurs universités en France pour modéliser la publication électronique des thèses. Parmi les universités on peut énumérer : l'INSA de Lyon, l'Institut National Polytechnique de Toulouse, l'Université Bordeaux 1, l'Université de Rennes 1, l'Université de Valenciennes.

Le group de travail de thèses du projet ORI-OAI, propose un « workflow » complexe (Image : Workflow de publication des thèses en France)

Les différents acteurs qui interviennent dans le processus de publication de la thèse sont : le doctorant, le gestionnaire de la thèse, la scolarité, le directeur de la thèse, le cataloguer et le validateur final.

Le doctorant va saisir le premier niveau d'information qui va permettre d'initialiser le processus de validation.

Le gestionnaire des thèses va vérifier les fichiers déposés par le doctorant, il va vérifier les données saisies par l'auteur : autorisation de diffusion de la thèse, informations personnelles et il va poster les fichiers qui seront diffusés et/ou archivés.

La scolarité renseignera les données concernant la confidentialité, la diffusion de la thèse, les directeurs de thèses ainsi que la date de naissance et la nationalité du doctorant, des données complémentaires sur le jury, les rapporteurs de la thèse et les partenaires de recherche.

Le validateur scientifique intervient pour valider le dépôt de la version corrigée de la thèse suite à une demande de corrections par le jury (corrections mineures ou majeures).

Le catalogueur complète les métadonnées de description.

Le validateur final valide l'ensemble de la procédure.

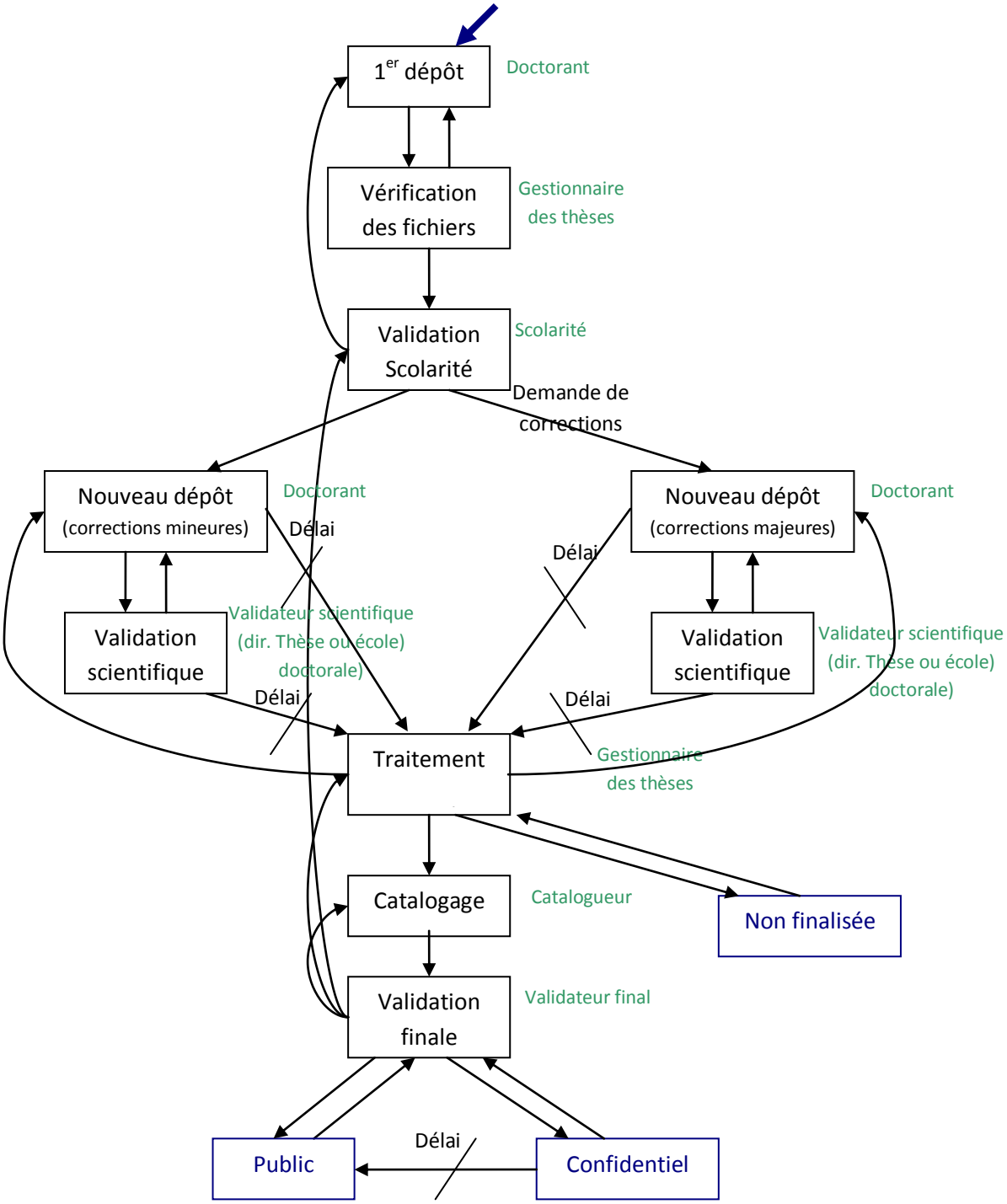


Image : Workflow de publication des thèses en France

3.3 Les Métadonnées de la thèse

En France, l'Agence Bibliographique de l'Enseignement Superior, a introduit une distinction entre quatre types de métadonnées :

- métadonnées descriptives (bibliographiques)
- métadonnées administratives

- métadonnées de droits
- métadonnées de conservation

Le standard principal utilisé en France pour les métadonnées de thèse est TEF. Les autres standards sont: METS (Metadata Encoding and Transmission Standard), DC (Dublin Core), DCTERMS, MetsRights.

Il est important de préciser que TEF s'appuie sur METS, mais n'en exploite pas toutes les fonctionnalités.

La Bibliothèque Nationale Allemande a créé son propre standard XMetaDiss pour la publication électronique de thèses. Les autres differents standards utilisées dans ce format sont : XMetaPers, Corporate Core, Dublin Core, DCTerms.

XMetaDiss est le format d'échange commun pour les publications universitaires en ligne des universités suisses transmises à la Bibliothèque nationale suisse.

4. L'éditeur de métadonnées

4.1 Présentation de l'éditeur de métadonnées

Afin de pouvoir éditer les fiches des métadonnées dans les standards imposés par la Bibliothèque Nationale Suisse un éditeur de métadonnées a été mis en place pour faciliter le travail.

On peut mettre en évidence plusieurs fonctionnalités dans le cadre de l'éditeur des métadonnées :

- navigation intuitive grâce à la technologie Ajax ;
- "widgets" qui permettent de faciliter la création de formulaires ;
- ajouter ou supprimer des champs ;
- sauvegarde/chargement des XForms sous forme des fichiers xml ;
- création des vocabulaires spécifiques sous le format vdex (Vocabulary Definition Exchange);
- ajout des nouveaux formulaires

L'éditeur propose 3 types des formulaires attribués aux acteurs du workflow de la thèse qui intervient dans la saisie des métadonnées. A la fin, les métadonnées sont sauvegardées dans des fichiers XML en respectant le standard XMetaDiss.

4.2 Choix technique

La base du module thesis-editor est un projet esup-commons, qui est un framework de développement basé sur Spring, JSF et Hibernate proposé comme standard de développement d'applications dans le cadre du projet ESUP-Portail. Un éditeur des métadonnées a été aussi développé dans le cadre du projet ORI-OAI, éditeur qui a les mêmes fonctionnalités.

Le module thesis-editor utilise le moteur XForms Orbeon, les « sources » de l'éditeur correspondent à des fichiers XML utilisés directement par l'application Orbeon Forms. En sachant que XForms n'est pas supporté nativement par les navigateurs un moteur comme Orbeon est nécessaire.

Orbeon Forms est une application J2EE qui utilise XForms côté serveur pour proposer dans un navigateur des interfaces Web Ajax. XForms, recommandation W3C, est une technologie qui permet de décrire en XML des formulaires manipulant des fichiers XML. Le XForm est interprété par Orbeon pour générer une interface Ajax à l'utilisateur.

Le module est installé sur un serveur Apache-Tomcat. Pour les tests la version 5.5.25 a été choisie. Afin des différencier les différents services web le serveur a été configure avec les ports suivants (fichier conf/server.xml):

Shutdown	Non SSL	SSL	AJP 1.3	Proxy
8187	8186	8486	8386	8586

Tableau: Configuration des ports pour le serveur apache-tomcat.

4.3 Les différents formulaires

Après la répartition des métadonnées en fonction d'acteur impliqué dans la saisie de métadonnées, 3 fiches XML ont été crées :

- Thesis-doctorant-blank.xml (Annexe 1)
- Thesis-sac-blank.xml (Annexe 2)
- Thesis-sisb-blank.xml (Annexe 3)

La fiche « Thesis-doctorant-blanck.xml » est la plus complexe. C'est le doctorant qui va remplir la majorité des métadonnées, métadonnées qui seront utilisées dans le workflow de publication. Parmi les métadonnées que le doctorant est obligé de les saisir on peut énumérer : le titre de la thèse (le tire traduit), le sous-titre de la thèse (le sous-titre traduit), les mots-clef, le résume de la thèse qui va être utilisé sur le serveur de publication, la

classification de la thèse en fonction de la classification Dewey, les informations personnels, le directeur de la thèse les membres du jury, et les droits de diffusion de sa thèse.

L'obligation du doctorant est de fournir les métadonnées de sa thèse afin d'éviter plusieurs problèmes :

- le travail redondant d'insérer les mêmes métadonnées dans plusieurs systèmes ;
- les possibles erreurs qui peuvent survenir à la recopier des métadonnées ;
- deviner certaines métadonnées ;
- retravailler l'abstract de la thèse ;

Apres la soutenance de sa thèse, le service académique confirme la version finale du document, et identifie la thèse avec un numéro. Il est nécessaire aussi de spécifier la personne de contact de l'institution déposante, respectivement l'EPFL.

La Bibliothèque centrale saisie l'URL de l'abstract de la thèse, l'URL de la thèse, le nombre des fichiers PDF (normalement dans le « nouveau guide du doctorant » le doctorant est obligé de fournir un seul fichier PDF) et la date de la remise valable des exemplaires, respectivement des fichiers du dépôt légal.

Afin de modéliser certaines métadonnées utilisées dans le processus de publication plusieurs standards et attributs ont été rajoutés :

- pour modéliser le titre du doctorant, l'attribut «***gender*** » a été rajouté. L'attribut n'est pas supporte par la Bibliothèque National mais est inclus dans le standard XMetaPers « ***pc :person*** » proposé par la Bibliothèque Nationale Allemande.
- pour modéliser le rôle des personnes participant au procédé d'acceptation de la thèse la valeur «***committeeMember*** » pour l'attribut « ***thesis :role*** » a été rajoutée. Les possibles valeurs de l'attribut « ***thesis :role*** » sont : « ***advisor*** » pour spécifier le rôle de directeur de la thèse et « ***committeeMember*** » pour spécifier le rôle de jury de la thèse.
- pour modéliser les droits de l'auteur sur la publication de la thèse le standard MetsRights a été adopté. Dans le cas favorable, c'est-à-dire le doctorant accepte la publication de sa thèse sur Internet, il accepte aussi la copie, la duplication et l'impression de sa thèse. Les valeurs booléens sont modéliser à l'aide des attributs suivants : « ***DISCOVER="true" COPY="true" DISPLAY="true" DUPLICATE="true" PRINT="true"*** ». Dans le cas négatif les valeurs sont mises à « ***false*** ».
- pour modéliser le numéro de la thèse, numéro utilisé dans le cadre du workflow, l'élément « ***ddb :identifier ddb :type= 'other'*** » a été rajouté. Normalement cet élément est utilisé pour indiquer tous les identificateurs de la publication. Le type de l'identificateur est spécifié a l'aide de l'attribut « ***ddb :type*** ».

4.4 Structure technique de l'éditeur

Le cœur de l'éditeur est le moteur Orbeon. Sa principale fonctionnalité est la création des fiches et la sauvegarde sous forme de fichiers XML des métadonnées.

Le fichier principal de configuration de chaque fiche est « ***main-form.xhtml*** ».La configuration de « widgets » est décrite dans ce fichier. L'exemple suivant permet de mettre en évidence l'interaction entre l'élément et la création du champ à saisir, la « relation avec d'autres documents » étant définie de façon suivante :

```
<fieldset>
        <widget:legend instance="i18n_doctorant-thesis" termIdentifier="8"/>
        <fieldset>
            <widget:legend instance="i18n_doctorant-thesis" termIdentifier="8.1"/>
            <widget:ori-block element="dc:relation[@xsi:type='ddb:noScheme']" minOccurs="0"
maxOccurs="unbounded" sortable="true"  preceding-elements="">
                <xforms:input ref="."/>
            </widget:ori-block>
        </fieldset>
</fieldset>
```

Fichier /WEB-INF/resources/apps/ori-md-editor/xforms/thesis-doctorant/main-form.xhtml:
Widget « dc :relation »

A l'aide d'un navigateur en cliquant sur la fiche du doctorant on obtient le résultat suivant :

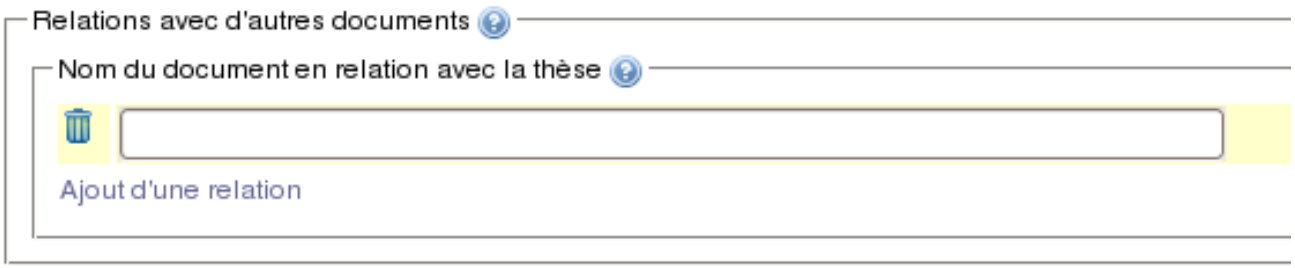


Image: Le champ à saisir « Relations avec d'autres documents »

La balise « fieldset » permet d'encadrer le champ à saisir et marque le début du widget. L'instance « ***i18n_doctorant-thesis*** » définie le vocabulaire utilisé dans la fiche « doctorant-thesis ». Le vocabulaire est défini dans le fichier ***« /WEB-INF/resources/apps/ori-md-editor/xml-vocab-local/thesis-editor_doctorant-thesis_i18n.xml*** ». Ce vocabulaire contient les textes utilisés dans le formulaire « Doctorant Thesis ». Chaque entrée dans le vocabulaire est identifiée par un « ***termIdentifier*** », identificateur qui est utilisé dans le « ***main-form.html*** » de chaque formulaire pour choisir le texte qui convient.

```
<vdex:term validIndex="false">
 <vdex:termIdentifier>1</vdex:termIdentifier>
 <vdex:caption>
  <vdex:langstring language="fr">Langue(s) de la thèse (*)</vdex:langstring>
```

```
</vdex:caption>
<vdex:description>
 <vdex:langstring language="fr">Une thèse peut être écrite en une ou plusieurs langues</vdex:langstring>
</vdex:description>
<vdex:metadata>
 <xforms:alert>Une thèse doit avoir au moins une langue</xforms:alert>
</vdex:metadata>
<vdex:metadata>
 <xforms:hint>Langue de la thèse</xforms:hint>
</vdex:metadata>
</vdex:term>
```

Fichier /WEB-INF/resources/apps/ori-md-editor/xml-vocab-local/thesis-editor_doctorant-thesis_i18n.xml: Vocabulaire "Langue de la thèse"

En passant avec le curseur sur le champ à saisir dans le formulaire le texte entre la balise ***« xforms:hint »*** va s'afficher. En cliquant sur le « help » à cote du chaque champs, un texte explicatif va s'afficher. Le texte est encadré par les balises ***« vdex:langstring ».***

Tous les éléments utilisés dans le formulaire sont définies dans le prototype du fichier XML. Dans le cas de la fiche thesis-doctorant on a le fichier ***« /WEB-INF/resources/apps /ori-md-editor/prototypes/thesis-doctorant-prototype.xml »***. Dans le fichier blank.XML (exemple « ***thesis-doctorant-blank.xml*** ») on a les éléments dont les champs vont s'afficher dans le formulaire, mais l'ajout des champs implique l'ajout des éléments, éléments qui sont dans le fichier prototype.xml (exemple ***« thesis-doctorant-prototype.xml »***). Pour mieux mettre en évidence la différence, on prend l'exemple du titre de la thèse. Le titre de la thèse est enregistre entre la balise « ***dc:title*** ». Si on ajoute encore un titre traduit dans le formulaire, l'attribut « ***ddb :type= 'translated'*** est ajouté. A la fin on a « ***dc:title xsi:type="ddb:titleISO639-2" lang="" ddb:type="translated"»***.

Dans le formulaire « thesis-doctorant » on a des vocabulaires spécifiques qui permettent de modéliser les langues, le titre de la personne, la classification de la thèse, le rôle du membre de la commission de l'examen (Annexe 4). A l'aide d'un menu déroulant, l'utilisateur choisira la valeur qui corresponde.

Chaque vocabulaire est identifié dans le fichier principal ***« main-form.xhtml »*** par un identificateur qui pointe vers le fichier XML où le vocabulaire est défini.

```
<xforms :instance id= "classification" src= "ori-md-editor/vocab/classification" xxforms :readonly="true"
xxforms:shared="application"/>
```

Code: Definition de l'instance id « classification »

L'id est utilisé après dans le widget « Classification de la thèse » pour définir la classification Dewey.

```
<xforms:itemset nodeset="xxfomrms:instance('classification')/vdex:term"
```

Code: Intégration de l'id dans la définition du widget « Classification de la thèse »

Le domaine « Data processing and computer science » est enregistré de façon suivante dans le fichier « ***/WEB-INF/resources/apps/ori-md-editor/xml-vocab-local/classification.xml*** » :

```
<vdex:term validIndex="true">
 <vdex:termIdentifier>004</vdex:termIdentifier>
 <vdex:caption>
  <vdex:langstring language="fr">Data processing and computer science</vdex:langstring>
  <vdex:langstring language="en">Data processing and computer science</vdex:langstring>
 </vdex:caption>
</vdex:term>
```

Fichier /WEB-INF/resources/apps/ori-md-editor/xml-vocab-local/classification.xml: Vocabulaire "Data processing and computer science"

Quand l'utilisateur choisira le domaine "Data processing and computer science" le moteur Orbeon va remplir dans la fiche XML la valeur entre la balise « ***vdex :termIdentifier*** » c'est-à-dire 004.

Le même raisonnement est valable pour les vocabulaires « gender.xml », « languages.xml », « thesis-role.xml ».

Une autre fonctionnalité importante de l'éditeur est le remplissage automatique du contenu de l'attribut. Si on prend l'exemple du titre de la thèse, l'attribut « ***lang*** » peut prendre les valeurs suivantes : eng, fre, ita, ger, valeurs définies dans le vocabulaire ayant l'id « languages_short ». Du point de vue technique, dans la définition du widget « titre de la thèse » on appelle l'element « cont :input-with-lang-short ». A l'aide de la technologie Extensible Binding Language (XBL) qui permet d'implémenter des components réutilisable, l'élément «cont :input-with-lang-short » est défini de façon suivante :

```
<xbl:binding id="ori-input-with-lang-short" element="cont:input-with-lang-short">
    <xbl:template>
        <xforms:select1 ref="@lang">
          <xforms:itemset nodeset="xxforms:instance('languages_short')/vdex:term">
            <xforms:label ref="vdex:caption/vdex:langstring[@language=substring-
before(xxforms:instance('i18n-choice')/lang, '_')]"/>
            <xforms:value ref="vdex:termIdentifier"/>
          </xforms:itemset>
        </xforms:select1>
    </xbl:template>
  </xbl :binding>
```

En sélectionnant l'attribut «lang » le moteur Orbeon interprète le vocabulaire « languages_short » pour remplir le contenu de l'attribut.

Les textes pour ajouter des champs sont définis en français et en anglais dans les fichiers « ***/WEB-INF/resources/apps/ori-md-editor/i118n/en_FR.xml*** » et respectivement ***« /WEB-INF/resources/apps/ori-md-editor/i118n/fr_EN.xml*** ». A l'aide des balises « add-element » on peut définir les textes, ou «l'element » peut avoir plusieurs valeurs : {creator, title, language …}. Le moteur Orbeon reconnait le nom de l'élément et ajoutera dans le formulaire le texte correspondant.

4.5 L'installation de l'éditeur

L'application est installée sous Linux. Dans la version de tests le système d'exploitation Ubuntu a été choisi. On considère la variable EDITEUR étant le chemin vers le lieu où on a choisi de faire l'installation. On peut considère EDITEUR étant « /usr/local/editeur ». Dans l'EDITEUR on va créer un directoire « tomcat-editor » où le serveur Tomcat est installé. Le deuxième directoire dans l'EDITEUR est le directoire des sources où l'application va être copiée. Dans notre exemple dans /usr/local/editeur/src on va copie directement l'application « thesis-editor ».

Le fichier « EDITEUR/src/thesis-editor/build.properties » doit-être modifié. On doit modifier le répertoire de déploiement :

deploy.home=EDITEUR/tomcat-editor/webapps

Dans notre exemple on a deploy.home = /usr/local/editeur/tomcat-editor/webapps

Maintenant on se place dans le directoire de l'application : EDITEUR/src/thesis-editor.

A l'aide de la commande « ant deploy » on va déployer l'application. Si l'application a été correctement déployée à la fin va s'afficher le message « BUILD SUCCESFULL »

On démarre le serveur Tomcat avec la commande « EDITEUR/tomcat-editor/bin/startup.sh ».

Depuis un navigateur web on accede la premier page de l'editeur:

http://NOM_SERVEUR:8186/thesis-editor/

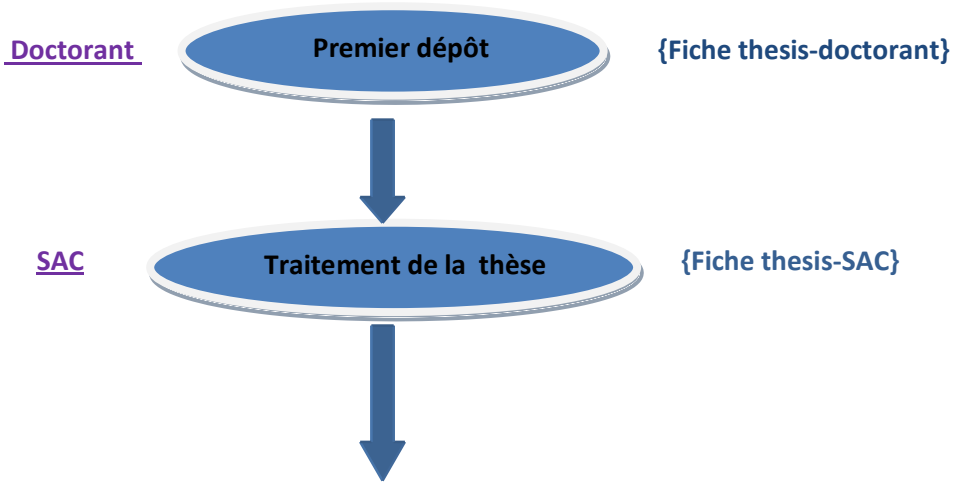
Si l'installation a été correctement faite on doit obtenir le resultat suivant :

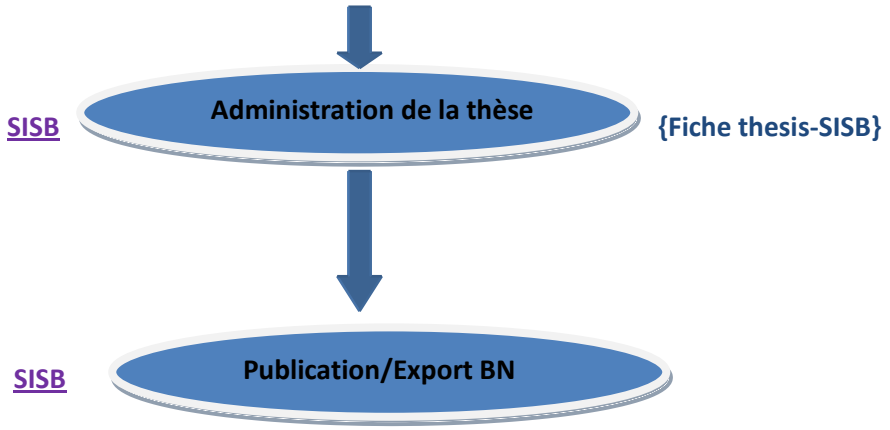
Image: Capture d'ecran de l'application

5. Les avantages du nouvel modèle

L'éditeur de métadonnées est une application qui sépare les métadonnées de la thèse en fonction des acteurs. On observe que le doctorant est l'acteur principal qui fournit la majorité de métadonnées. A 'aide d'un système centralisé les métadonnées peuvent être stockées et accessibles en fonction des droits de chaque acteur. A ce moment chaque acteur passe le workflow de la thèse d'une étape à l'autre et on a une indépendance entre les acteurs.

Le nouveau workflow des métadonnées de la thèse est le suivant :





Le système permet d'éviter certains problèmes déjà énoncés dans le chapitre 1.5, et en même temps en peut mettre en évidence la puissance des fiches XML.

Pour que l'export vers la Bibliothèque Nationale soit possible une simple transformation XSLT des trois fiches permet de créer le bon fichier XML à transmettre.

L'éditeur de métadonnées déjà construit peut être intégré dans un système plus complexe qui permet la sauvegarde de fiches XML dans une base, une indexation de mots clés plus efficace, un système de recherche de thèses qui est accessible au grand publique, l'export automatique des fiches XML vers la Bibliothèque Nationale. Un système semblable pour les documents numérique est le système ORI-OAI qui peut-être adapté dans le cas de thèses.

Conclusion

Ce projet m'a permis d'analyser le workflow de publication de thèses à l'EPFL, de mettre en évidence les principaux problèmes du système, et de concevoir une solution en utilisant plusieurs technologies afin d'optimiser le processus de publication. Les RDV avec les principaux acteurs de la thèse m'ont permis de me familiariser avec le workflow actuel de publication et de comprendre les besoins de chaque acteur.

La solution choisie change les rôles des différents acteurs ; la publication de la thèse étant concentrée sur les métadonnées fournies par le doctorant, un nouveau « Guide du doctorant » doit être rédigé pour familiariser le doctorant avec le système, dans le cas ou la solution soit retenue.

Je tiens à remercier toute l'équipe du laboratoire Center for Global Computing pour m'avoir intégré rapidement au sein de leur laboratoire et m'avoir accordé leur confiance.

Je tiens à remercier Madame Christine Vanoirbeek pour l'aide et les conseils concernant les travaux évoqués dans ce rapport.

Je remercie également Monsieur Georges Iffland, pour m'avoir intégré au sein de la Bibliothèque Centrale, pour le temps qu'il m'a consacré au long de cette période, sachant répondre a mes interrogations.

Monsieur François Schmitt pour les réponses promptes et exactes aux questions et pour les documents fournies, sans oublier sa participation au cheminent de ce rapport.

Monsieur David Aymonin, directeur de la Bibliothèque Centrale, pour son accueil et la confiance qu'il m'a accordé dès mon arrivé.

Annexe 1 : Thesis-doctorant-blank.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<xMetaDiss:xMetaDiss xmlns:xMetaDiss="http://www.nb.admin.ch/standards/xmetadiss/xMetaDiss/"
xmlns:cc="http://www.nb.admin.ch/standards/xmetadiss/cc/"
xmlns:ddb="http://www.nb.admin.ch/standards/xmetadiss/ddb/"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcmitype="http://purl.org/dc/dcmitype/"
xmlns:dcterms="http://purl.org/dc/terms/" xmlns:pc="http://www.nb.admin.ch/standards/xmetadiss/pc/"
xmlns:urn="http://www.nb.admin.ch/standards/urn/"
xmlns:thesis="http://www.ndltd.org/standards/metadata/etdms/1.0/"
xmlns="http://www.nb.admin.ch/standards/subject/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:metsRights="http://cosimo.stanford.edu/sdr/metsrights/"
xsi:schemaLocation="http://www.nb.admin.ch/standards/xmetadiss/xMetaDiss/  xmetadiss.xsd">
        <dc:title xsi:type="ddb:titleISO639-2" lang="eng"/>
        <dcterms:alternative xsi:type="ddb:talternativeISO639-2" lang="eng"/>
        <dc:subject xsi:type="xMetaDiss:noScheme" lang ="eng"/>
        <dc:subject xsi:type="dcterms:DDC"/>
        <dcterms:abstract xsi:type="ddb:contentISO639-2" lang="eng" ddb:type="noScheme"/>
        <dcterms:dateAccepted xsi:type="dcterms:W3CDTF"/>
        <dcterms:dateSubmitted xsi:type="dcterms:W3CDTF"/>
        <dc:creator xsi:type="pc:MetaPers">
                <pc:person gender="m">
                        <pc:name type="nameUsedByThePerson">
                                <pc:foreName/>
                                <pc:surName/>
                        </pc:name>
                        <pc:dateOfBirth xsi:type="dcterms:W3CDTF"/>
                        <pc:placeOfBirth/>
                        <pc:address/>
                        <pc:email/>
                </pc:person>
        </dc:creator>
        <dc:contributor xsi:type="pc:Contributor" thesis:role="advisor" type="ISO3166" countryCode="ch">
                <pc:person>
                        <pc:name type="nameUsedByThePerson">
                                <pc:foreName/>
                                <pc:surName/>
                        </pc:name>
                </pc:person>
        </dc:contributor>
        <dc:language xsi:type="dcterms:ISO639-2"/>
        <dc:relation xsi:type="ddb:noScheme"/>
        <dc:relation xsi:type="dcterms:URI"/>
        <thesis:grantor>
                <cc:universityOrInstitution>
                        <cc:name/>
                        <cc:place/>
                        <cc:departement>
                                <cc:name/>
                        </cc:departement>
                </cc:universityOrInstitution>
        </thesis:grantor>

        <metsRights:Context CONTEXTCLASS="GENERAL PUBLIC">
                <metsRights:Permissions DISCOVER="true" COPY="true" DISPLAY="true"
DUPLICATE="true" PRINT="true" MODIFY="false" DELETE="false"/>
        </metsRights:Context>
```

```
</xMetaDiss:xMetaDiss>
```

Annexe 2: Thesis-sac-blank.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<xMetaDiss:xMetaDiss xmlns:xMetaDiss="http://www.nb.admin.ch/standards/xmetadiss/xMetaDiss/"
xmlns:cc="http://www.nb.admin.ch/standards/xmetadiss/cc/"
xmlns:ddb="http://www.nb.admin.ch/standards/xmetadiss/ddb/"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcmitype="http://purl.org/dc/dcmitype/"
xmlns:dcterms="http://purl.org/dc/terms/" xmlns:pc="http://www.nb.admin.ch/standards/xmetadiss/pc/"
xmlns:urn="http://www.nb.admin.ch/standards/urn/"
xmlns:thesis="http://www.ndltd.org/standards/metadata/etdms/1.0/"
xmlns="http://www.nb.admin.ch/standards/subject/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:metsRights="http://cosimo.stanford.edu/sdr/metsrights/"
xsi:schemaLocation="http://www.nb.admin.ch/standards/xmetadiss/xMetaDiss/  xmetadiss.xsd">
        <ddb:note></ddb:note>
        <dc:publisher xsi:type="cc:Publisher">
                <cc:universityOrInstitution>
                        <cc:name>EPFL</cc:name>
                        <cc:place>Lausanne</cc:place>
                </cc:universityOrInstitution>
                <cc:address>EPFL</cc:address>
        </dc:publisher>
        <ddb:contact ddb:contactID="F1111-1111"/>
        <ddb:description>E-mail:</ddb:description>
        <dc:type xsi:type="ddb:PublType">ElectronicThesisandDissertation</dc:type>
        <thesis:level>thesis.doctoral</thesis:level>
</xMetaDiss:xMetaDiss>
```

Annexe 3: Thesis-sisb-blank.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<xMetaDiss:xMetaDiss xmlns:xMetaDiss="http://www.nb.admin.ch/standards/xmetadiss/xMetaDiss/"
xmlns:cc="http://www.nb.admin.ch/standards/xmetadiss/cc/"
xmlns:ddb="http://www.nb.admin.ch/standards/xmetadiss/ddb/"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcmitype="http://purl.org/dc/dcmitype/"
xmlns:dcterms="http://purl.org/dc/terms/" xmlns:pc="http://www.nb.admin.ch/standards/xmetadiss/pc/"
xmlns:urn="http://www.nb.admin.ch/standards/urn/"
xmlns:thesis="http://www.ndltd.org/standards/metadata/etdms/1.0/"
xmlns="http://www.nb.admin.ch/standards/subject/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:metsRights="http://cosimo.stanford.edu/sdr/metsrights/"
xsi:schemaLocation="http://www.nb.admin.ch/standards/xmetadiss/xMetaDiss/  xmetadiss.xsd">
        <dcterms:abstract xsi:type="ddb:contentISO639-2" lang="fr" type="dcterms:URI"/>
        <ddb:identifier ddb:type="URL"/>
  <ddb:fileNumber/>
  <ddb:fileProperties ddb:fileName="dissertation.pdf" ddb:fileID="file1"/>
  <dcterms:created xsi:type="dcterms:W3CDTF"/>
  <dc:identifier xsi:type="urn:nbn">urn:ubn:ch:UNBEKANNT</dc:identifier>
  <ddb:server>SISB Lausanne</ddb:server>
  <ddb:dateDelivered xsi:type="dcterms:W3CDTF"/>
</xMetaDiss:xMetaDiss>
```

Annexe 4: languages-short.xml

```
<?xml version="1.0" encoding="utf-8"?>
<vdex:vdex xmlns:vdex="http://www.imsglobal.org/xsd/imsvdex_v1p0"
        xmlns:orioai="http://www.ori-oai.org/static/xsd/orioaivocab"
        xmlns:xforms="http://www.w3.org/2002/xforms"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.imsglobal.org/xsd/imsvdex_v1p0
http://www.imsglobal.org/xsd/imsvdex_v1p0.xsd"
        profileType="flatTokenTerms">
<vdex:vocabName>
 <vdex:langstring language="fr">langues</vdex:langstring>
</vdex:vocabName>
<vdex:vocabIdentifier isRegistered="false">languages</vdex:vocabIdentifier>
<vdex:term validIndex="true">
 <vdex:termIdentifier></vdex:termIdentifier>
 <vdex:caption>
  <vdex:langstring language="fr">Choose a language</vdex:langstring>
  <vdex:langstring language="en">Choose a language</vdex:langstring>
 </vdex:caption>
</vdex:term>
<vdex:term validIndex="true">
 <vdex:termIdentifier>eng</vdex:termIdentifier>
 <vdex:caption>
  <vdex:langstring language="fr">anglais</vdex:langstring>
  <vdex:langstring language="en">english</vdex:langstring>
 </vdex:caption>
</vdex:term>
<vdex:term validIndex="true">
 <vdex:termIdentifier>fre</vdex:termIdentifier>
 <vdex:caption>
  <vdex:langstring language="fr">français</vdex:langstring>
  <vdex:langstring language="en">french</vdex:langstring>
 </vdex:caption>
</vdex:term>
<vdex:term validIndex="true">
 <vdex:termIdentifier>ger</vdex:termIdentifier>
 <vdex:caption>
  <vdex:langstring language="fr">allemand</vdex:langstring>
  <vdex:langstring language="en">german</vdex:langstring>
 </vdex:caption>
</vdex:term>
<vdex:term validIndex="true">
 <vdex:termIdentifier>ita</vdex:termIdentifier>
 <vdex:caption>
  <vdex:langstring language="fr">italien</vdex:langstring>
  <vdex:langstring language="en">italian</vdex:langstring>
 </vdex:caption>
</vdex:term>
</vdex:vdex>
```