9th International Symposium on Electronic Theses and Dissertations
IXe Symposium international sur les thèses et mémoires électroniques
June 7 – 10 Juin, 2006, Quebec City / Québec, Canada

TEF
METADONNEES POUR LES THESES NUMERIQUES FRANÇAISES

Yann NICOLAS
ABES (Agence Bibliographique de l'Enseignement Supérieur)

INTRODUCTION

TEF (Thèses Electroniques Françaises) est un ensemble de métadonnées pour les thèses numériques françaises. C'est une recommandation, émise par un groupe de travail de l'Agence française de normalisation (AFNOR). Cette recommandation ne prétend pas se substituer à d'autres vocabulaires comme le Dublin Core ou ETD-ms. TEF a une fonction très précise, qui trouve son sens dans un contexte réglementaire très précis. Il s'agit d'échanger des métadonnées descriptives et des métadonnées de gestion entre partenaires du même dispositif national. C'est ce contexte précis et cette fonction d'échange qui expliquent la richesse du vocabulaire TEF, la nature très spécifique de certaines métadonnées et l'importance donnée aux outils de validation XML.

Pourtant, tout au long du processus de normalisation, les concepteurs se sont efforcés de rendre TEF assez souple et ouvert pour qu'il puisse se prêter à des usages qui ne soient pas directement prévus par le dispositif national. Les moyens employés pour tendre vers une plus grande interopérabilité sont notamment la réutilisation de vocabulaires connus (Dublin Core, METS, MADS, METS Rights), la modélisation qui permet de prendre du recul, l'exploitation de Schematron qui offre une grande souplesse de validation ou encore le recours aux fichiers d'autorité.

1. LES THESES DANS LE CONTEXTE FRANÇAIS

Les thèses sont des documents à part : ils possèdent une double nature, à la fois scientifique et administrative. En tant qu'elles délivrent un diplôme national, les thèses françaises sont l'objet de texte réglementaires *ad hoc*.[1] Jusqu'en 2000, ces textes réglementaires ne concernaient que les thèses imprimées. Depuis cette date se met progressivement en place une politique générale de promotion des thèses numériques, dont les principes ont été clairement affirmés en 2005 et sont en voie d'application en 2006.

Cette politique affirme trois objectifs incontournables. Tout le reste est à la discrétion de chaque université qui délivre des thèses. Ces trois points cardinaux sont les suivants :

- Le référencement dans la bibliographie nationale des thèses, intégrée au catalogue collectif des bibliothèques universitaires, le Sudoc[2].
- La diffusion la plus large et la plus rapide possible, par l'université de soutenance elle-même ou par des tiers.
- La conservation à long terme, assurée par un organisme national, le CINES.

L'université est responsable de la validation scientifique et administrative. Le reste des opérations qui constituent le cycle de vie de la thèse peuvent (et certaines doivent) être prises en charge par des tiers.

Afin de faciliter la collaboration entre l'université de soutenance et ses partenaires (agence bibliographique, diffuseurs, archiviste), un outil de logistique a été commandé par le Ministère et

[1] http://www.abes.fr/abes/DesktopDefault.aspx?tabid=426
[2] http://www.sudoc.abes.fr/

réalisé par l'ABES. STAR est un simple intermédiaire, un outil de transit, un échangeur. Il sera en service à la fin de l'année 2006.
En *entrée*, STAR recueillera les thèses et leurs métadonnées auprès des établissements de soutenance, seuls habilités à garantir que le document transmis est bien conforme à la version validée par le jury.

En *sortie*, la thèse sera systématiquement orientée vers le système de conservation à long terme du CINES. De même, les métadonnées seront converties en UNIMARC pour enrichir la bibliographie nationale des thèses (catalogue Sudoc).

A côté de ces débouchés réglementaires, STAR proposera aux établissements de soutenance des services complémentaires :

- diffusion sur différents serveurs (dont HAL[3], plateforme national d'autoarchivage gérée par le CNRS) ;
- signalement et indexation en texte intégral dans le portail documentaire Sudoc[4] ;
- assignation d'un identifiant pérenne (URI) et service de résolution pour donner accès à la thèse, quelle que soit sa localisation en ligne.
- Exposition OAI-PMH des métadonnées, sous différents formats (DC, ETD-MS, MARCXML, TEF).

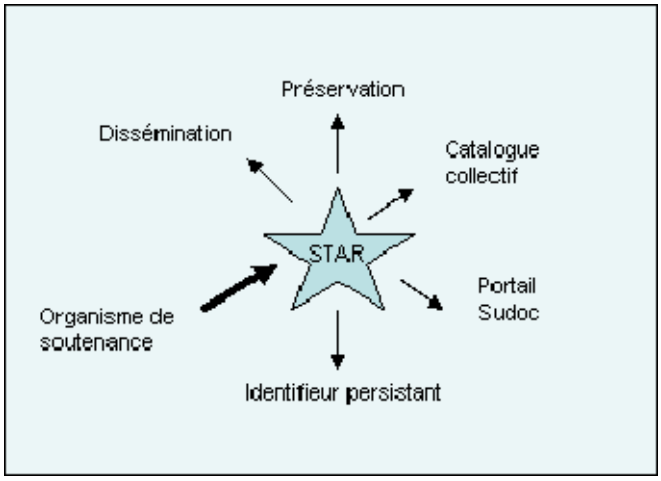

Figure 1. Entrée et sorties de STAR

En un seul dépôt, chaque établissement sera en mesure d'assurer la conservation à long terme de ses thèses, leur diffusion par de multiples canaux et leur signalement, en ayant la certitude qu'il s'agit de la bonne version, validée scientifiquement et administrativement. Le seul fait qu'une thèse française soit accédée via l'identifiant pérenne national indiquera qu'il s'agit d'une version officielle, garantie par l'établissement. L'URI servira à la fois d'identifiant unique, d'URL d'accès en ligne et de tampon de validité.

STAR ne se substitue par aux outils ou processus que les établissements ont pu mettre en place localement. Certes, pour ceux qui ne disposent pas d'outil local pour gérer les thèses, STAR offrira une interface Web pour déposer les fichiers et saisir les métadonnées. Pour les autres, STAR pourra importer les fichiers et les métadonnées produits localement. Ces métadonnées seront encodées selon le format d'échange national TEF.

[3] http://hal.ccsd.cnrs.fr/
[4] http://www.portail-sudoc.abes.fr/

2. FONCTION ET ANATOMIE DE TEF

On comprend maintenant que les métadonnées TEF sont solidaires d'un contexte national très précis. La mission de TEF est bien de regrouper en un **format d'échange** toutes les métadonnées nécessaires pour qu'un établissement B puisse signaler, diffuser et conserver une thèse soutenue dans un établissement A. Le signalement suppose des métadonnées descriptives et administratives. La diffusion suppose des métadonnées descriptives et des métadonnées de droits. La conservation suppose de surcroît des métadonnées techniques. TEF organise de manière modulaire ces différents types de métadonnées.

Métadonnées descriptives – TEF exige les métadonnées nécessaires et suffisantes pour générer une notice bibliographique UNIMARC conforme aux règles de catalogage françaises. Cela comprend la gestion des points d'accès et du lien aux notices d'autorités. Ce bloc s'appuie essentiellement sur les métadonnées Dublin Core et sur MADS.

Métadonnées administratives – Ce bloc comprend les métadonnées relatives à la thèse en tant que telle, à la soutenance, au jury. Là encore, le lien aux autorités est possible. Là encore, le Dublin Core sert de noyau, même s'il est complété par des éléments originaux relevant de l'espace de noms TEF.

Métadonnées de droits – En utilisant le vocabulaire METS Rights, TEF peut exprimer des informations juridiques relativement fines : autorisation du jury, autorisation de l'auteur, autorisation du chef d'établissement, période de confidentialité éventuelle, autorisation des ayants droit dont l'œuvre est réutilisée dans la thèse (images, textes, schémas…). Ce degré de finesse permet d'automatiser certaines opérations, comme la levée de la période de confidentialité.

Métadonnées de conservation – TEF ne comprend pas toutes les métadonnées nécessaires à la conservation, mais seulement celles que le CINES ne génère pas lui-même à partir des fichiers. Le schéma XML correspondant à ce bloc est extensible (xsd:any), dans l'éventualité où un établissement confirait à un autre organisme que le CINES la mission de conserver ses thèses et que ce dernier exigerait d'autres métadonnées.

Certaines informations sont à cheval sur différents blocs. Ainsi, l'auteur de la thèse étant à la fois un auteur, un étudiant et un ayant droit, il relève à la fois des métadonnées descriptives, des métadonnées administratives et des métadonnées de droits. Pourtant, les informations sur l'auteur ne sont mentionnées qu'une fois.

3. MODELISER

En analysant ces différentes métadonnées, il apparaît qu'elles ne parlent pas toujours de la même "chose". Quand on mentionne le sujet de la thèse, il concerne la thèse en tant qu'*œuvre*, indépendamment de la version du texte. Par contre, la thèse en tant que telle, en tant qu'elle est validée par le jury, correspond à une *version* unique, canonique, officielle. Enfin, quand on indique le nombre de pages de la thèse, on parle d'une *édition* particulière (papier ou PDF).

En termes FRBR[5], on parlerait d'*œuvre*, d'*expression* et de *manifestation*. La recommandation comprend un modèle conceptuel qui s'appuie sur le modèle FRBR de l'IFLA. L'objectif de ce modèle est avant tout de débrouiller la notion de thèse, d'expliciter les différentes entités dont on parle confusément et enfin d'identifier à quelle entité se rapporte chacune des métadonnées. Les

[5] http://www.ifla.org/VII/s13/wgfrbr/wgfrbr.htm

métadonnées sont alors vues comme des propriétés (ou des relations) de ces entités. Ce travail de conceptualisation a facilité le travail de structuration en XML et il rend presque trivial la structuration en RDF.

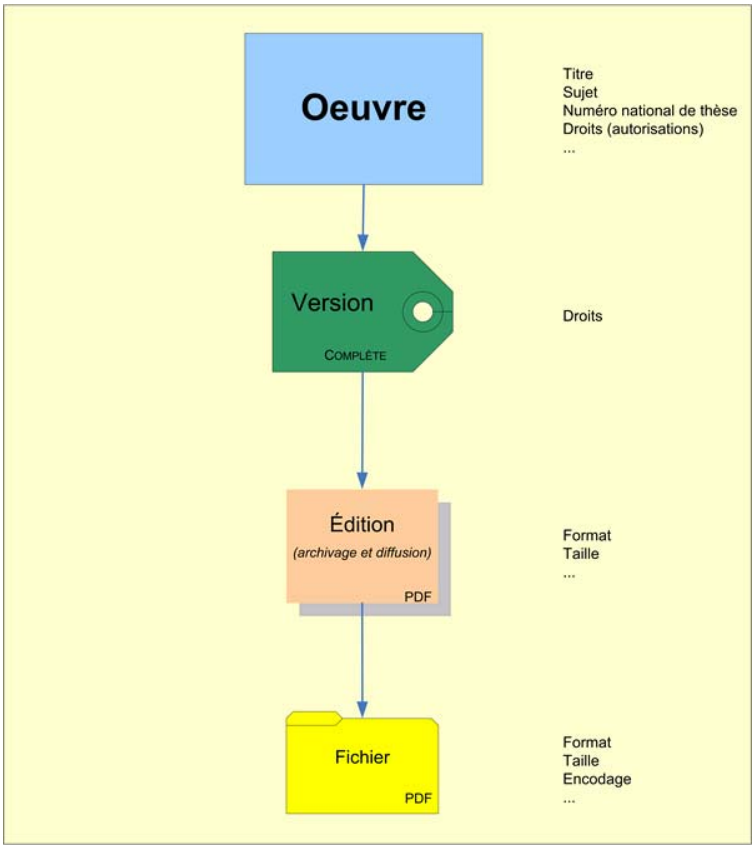

Figure 2. Modèle TEF – le cas simple

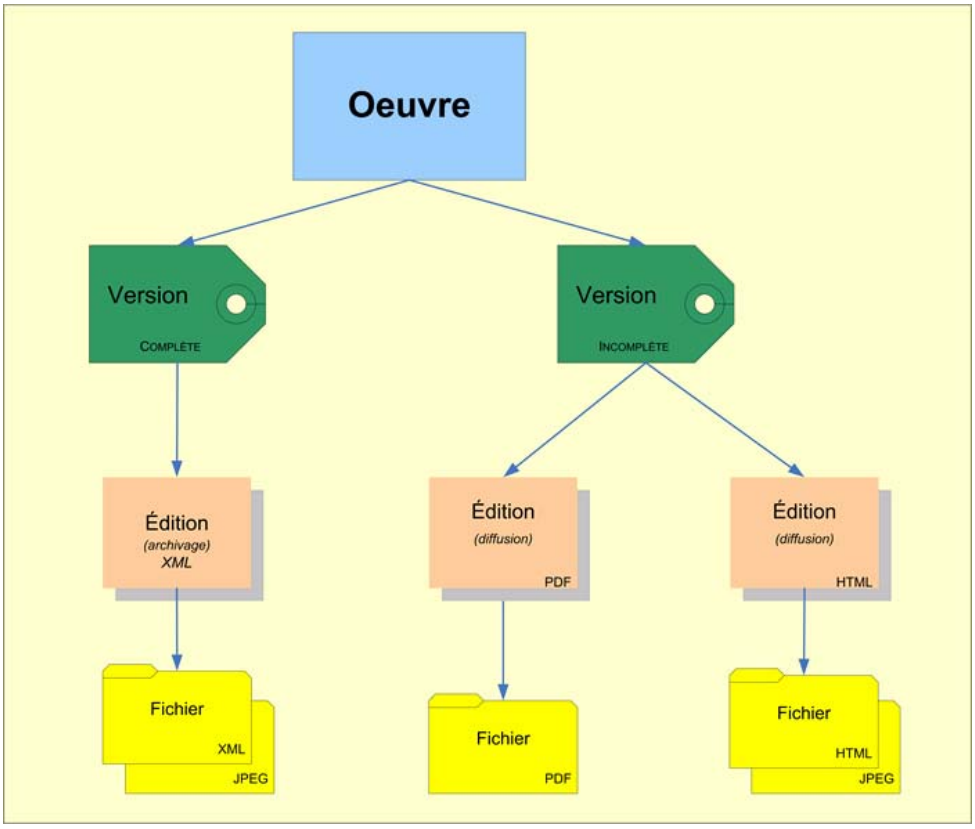

Figure 3. Modèle TEF – Différentes versions et différentes éditions

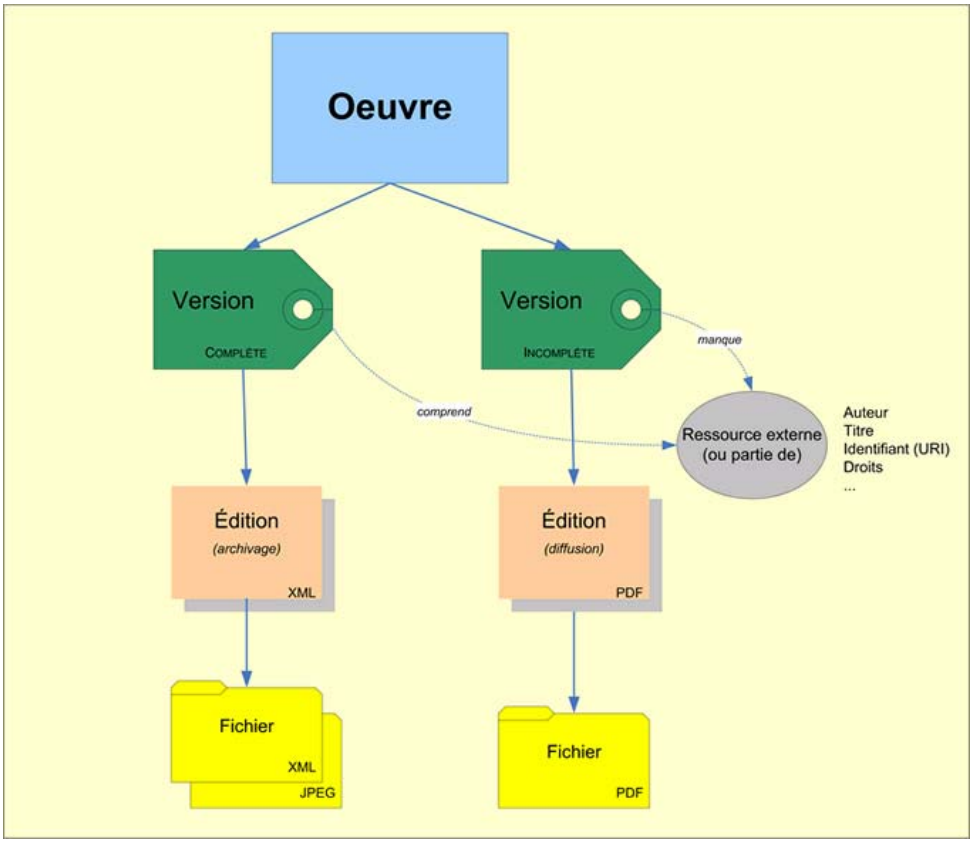

Figure 4. Modèle TEF – présence de ressources externes

En croisant le type de métadonnées et l'entité FRBR auquel elles se rapportent, on obtient un certain nombre de blocs de métadonnées homogènes. Les blocs sont représentés par des croix.

	thèse	version	édition	fichier	Ressource externe
Métadonnées descriptives	x	x Si version incomplète	x		x
Métadonnées administratives	x				
Métadonnées de droits	x	x			x
Métadonnées de conservation	x			x	

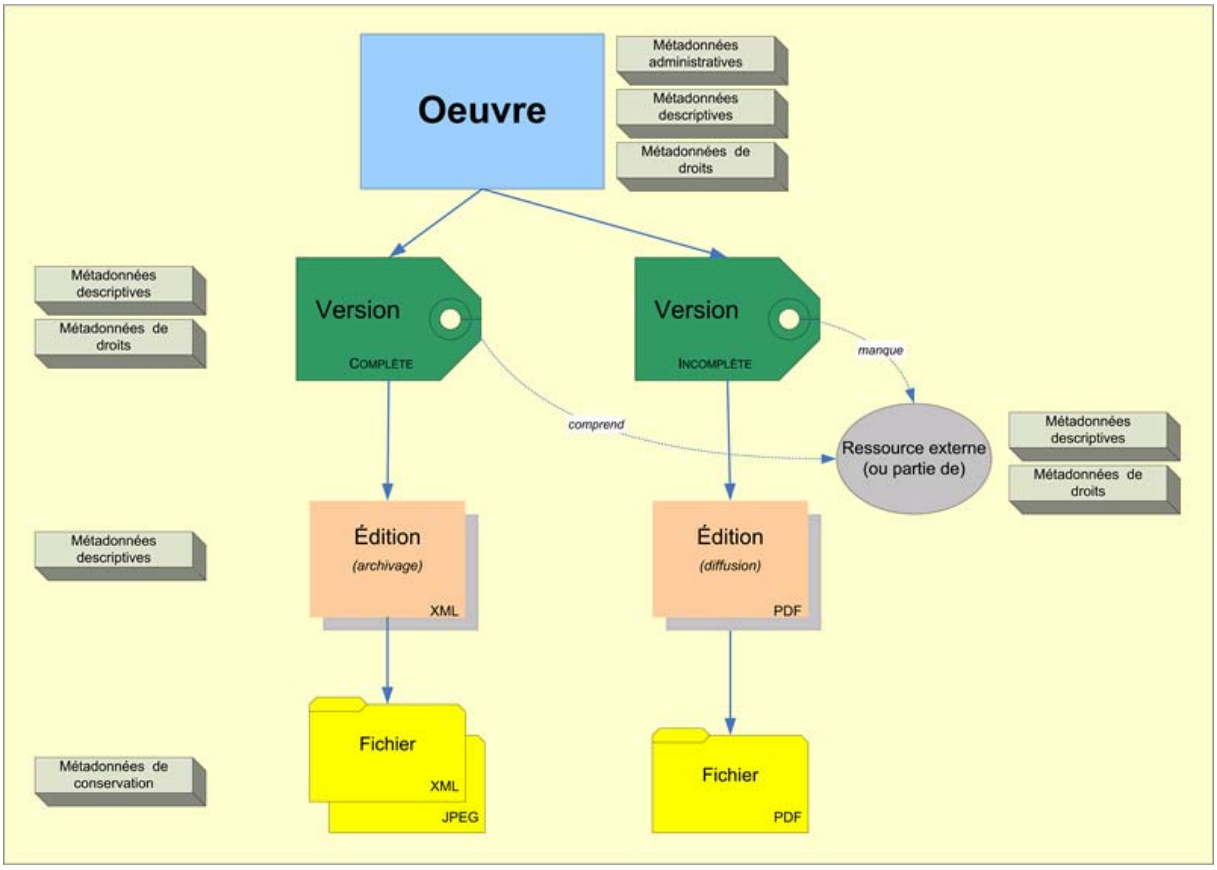

Figure 5. Les entités du modèle TEF et leurs métadonnées respectives

4. STRUCTURER EN XML

Dans la perspective d'un échange de métadonnées bien spécifiques entre des partenaires identifiés, XML s'est imposé comme le meilleur format pour encoder les métadonnées TEF. Grâce à sa malléabilité et à ses outils de validation, XML permet de définir exactement la structure voulue et de contrôler la conformité des instances TEF à cette structure.

Plutôt que d'imaginer un nouveau format, il a été décidé d'utiliser METS[6] pour organiser les différents blocs de métadonnées TEF. METS présente quatre caractéristiques particulièrement intéressantes pour TEF :

- La carte de structure (**mets:structMap**) qui permet d'inventorier les différentes composantes logiques ou physiques d'un objet numérique complexe. TEF définit une carte logique où sont positionnées les œuvres, les versions, les éditions et les ressources externes.

```
−<mets:structMap TYPE="logical">
 −<mets:div TYPE="THESE" ADMID="b999 a121 a122" DMDID="a111" CONTENTIDS="ark:99999/star/ISAL/linck/the
  −<mets:div TYPE="VERSION_COMPLETE" ADMID="a221" CONTENTIDS="ark:99999/star/ISAL/linck/vc">
   −<mets:div TYPE="EDITION" DMDID="xx311" CONTENTIDS="ark:99999/star/ISAL/these/linck/vc/ed1">
      <mets:fptr FILEID="FGrID1"/>
    </mets:div>
   </mets:div>
  </mets:div>
 </mets:structMap>
```

- La section des fichiers (**mets:fileSec**) qui permet d'inventorier les fichiers et de leur attacher des métadonnées de conservation.
- Le fait que METS ne prescrive aucun langage de métadonnées particulier. C'est une enveloppe vide dans laquelle tout type de métadonnées peut être utilisé.
- L'organisation modulaire des métadonnées : à chaque bloc de métadonnées correspond un bloc XML particulier.

Puisque METS n'est qu'une enveloppe qui doit s'adapter à son contenu et à l'usage qu'on veut en faire, le schéma XML de METS est très tolérant. Les contraintes qu'il impose sont minimales. Certes les schémas de métadonnées qu'on utilise à l'intérieur du fichier METS apportent des contraintes supplémentaires. Toutefois, comment exprimer que telle composant de la carte de structure doit être décrit avec telles métadonnées… ? En général, ces contraintes sont rédigées dans un document appelé "profil METS". Bien qu'il soit en XML, ce document est destiné aux hommes et non aux machines ; il ne permet pas un contrôle automatisé des règles qu'il prononce. C'est pourquoi nous avons choisi de compléter la validation par les schémas XML par l'utilisation de règles Schematron[7].

Schematron est un langage de validation XML qui permet d'exprimer des contraintes très fines et de rédiger comme on le souhaite les messages qui décrivent le résultat de la validation – en particulier les erreurs.

```
−<sch:pattern name="tef_desc_these -- général">
 −<sch:rule context="/mets:mets/mets:dmdSec/mets:mdWrap[@OTHERMDTYPE='tef_desc_these']">
  −<sch:assert test="mets:xmlData/tef:thesisRecord">
     La racine du bloc md de type "tef_desc_these" doit être "tef:thesisRecord".
   </sch:assert>
  </sch:rule>
```

Une règle Schematron permet aussi d'aller chercher dans un fichier XML distant des valeurs pour remplir une zone. TEF utilise cette fonction pour vérifier que l'élément tef:oai_setSpec a bien pour valeur une division Dewey.

Chaque règle est indépendante des autres. Cela implique qu'il est facile d'amender les règles exprimées dans un schéma Schematron pour s'adapter à des circonstances et des contraintes particulières.

[6] http://www.loc.gov/standards/mets/
[7] http://www.schematron.com/

Schematron a le double mérite de garantir la validité des données échangées et d'être assez souple pour permettre des aménagements locaux.

On le voit, TEF est moins un nouveau format, qu'un profil METS : un convention particulière pour rédiger un document METS conforme à des besoins spécifiques et explicites.

5. IMPLEMENTER

Que veut dire implémenter TEF ? Cela ne signifie pas offrir une interface de saisie pour renseigner des métadonnées TEF. On ne produit pas du TEF comme on produit une notice MARC. Dans l'esprit de ses concepteurs, TEF est un format d'échange qui permet de rassembler des métadonnées diverses, aux origines et aux fonctions variées. Au sein du système d'information d'un établissement, les métadonnées de thèse proviennent de sources différentes : elles peuvent être extraites du document lui-même (notamment de la page de titre), d'une application administrative de suivi des étudiants, d'une application dédiée à la gestion des thèses, d'un formulaire en ligne rempli par l'auteur… Il ne faut pas imaginer le catalogueur seul devant un formulaire TEF – et encore moins devant son éditeur XML. Le recueil des métadonnées de thèse, pour être efficace, nécessite de coordonner différentes applications et différents métiers au sein d'un établissement. TEF peut contribuer à rationaliser ce processus *interne* à l'établissement. En effet, TEF étant construit de manière modulaire, on peut imaginer que certains blocs de métadonnées TEF circulent au sein du système d'information, pour être réutilisés ou agrégés. C'est le cas notamment des informations portant sur le doctorant. Mais on peut aussi n'utiliser TEF qu'à titre de format d'échange, pour *l'export* vers STAR par exemple.

Pour aider les établissements à exporter du TEF, le groupe AFNOR et l'ABES s'efforcent de les accompagner. Il ne suffit pas de "lâcher" une norme dans la nature. TEF possède son site web, où l'on trouve un guide d'utilisation détaillé, des exemples, un FAQ. Il existe aussi un blog consacré à TEF. L'essentiel du travail d'accompagnement consiste à travailler avec les différents systèmes locaux de gestion des thèses, à expertiser leur schéma de métadonnées interne, à le comparer aux exigences de TEF et, enfin, à écrire ensemble le script XSLT qui assure la conversion de ce schéma interne vers TEF. L'idée n'est pas d'effectuer ce travail pour chaque établissement, mais pour chaque outil (Eprints, DSpace, OGET, Castore, ORI…).

6. LES METADONNEES AU GRAND LARGE

Au-delà des échanges de point à point prévus par le cadre réglementaire français, les métadonnées de thèse doivent pouvoir circuler librement sur le Web. Le protocole OAI-PMH est aujourd'hui universellement reconnu comme un bon moyen d'exposer des métadonnées sur le Web, de les rendre disponibles et exploitables de manière ouverte, anonyme et simple. Il est probable que grâce au nouveau dispositif national, les métadonnées des thèses françaises seront massivement exposées en OAI-PMH, et ce sous différents formats : TEF, mais aussi Dublin Core, ETD-MS voire marcxml (bientôt normalisé en MarcXchange).

Ce mode d'exposition des métadonnées a ses limites. Si un tiers moissonne les métadonnées en TEF, alors, pour les exploiter, ce tiers devra déployer des efforts importants pour analyser les fichiers, étudier les spécifications de TEF, écrire les outils de traitements adaptés. Si, au contraire, il moissonne les métadonnées en DC, il perdra une grande partie de la richesse initiale des métadonnées TEF. RDF peut nous sortir de ce dilemme.

Dans le cas de TEF, il permettra à terme une exploitation multiple des notices TEF en l'état, sans les convertir dans un vocabulaire plus répandu comme le Dublin Core. Il faut pour cela associer les notices TEF à un schéma RDF ou OWL qui précise les relations sémantiques entre les éléments propres à l'espace de noms TEF et, par exemple, les éléments du Dublin Core ou les propriétés des

FRBR. Par ailleurs, formaliser TEF en RDF permet de proposer un encodage très proche du modèle conceptuel de TEF, puisque la logique de RDF est d'exprimer les métadonnées sous la forme de propriétés et de relations qui s'appliquent à des entités bien identifiées.

La logique RDF est de décomposer l'information en une série de faits élémentaires (" ceci a pour titre cela", "ceci a pour créateur X", "X a pour prénom 'abc'"…), en principe indépendants les uns des autres. Ce qui importe en RDF n'est pas qu'un ensemble de métadonnées soit complet, autosuffisant et conforme à un schéma prescriptif, mais au contraire qu'il puisse être complété, enrichi par un autre ensemble de métadonnées RDF, accessible sur le Web. RDF doit permettre de mettre en relations les thèses avec des personnes, des organismes, des laboratoires, des programmes et des projets, des disciplines et bien sûr d'autres documents (articles, recensions, ressources citées ou réutilisées dans la thèse…), de situer les thèses dans le contexte des activités et des résultats de la science vivante. C'est aussi dans cette perspective que TEF donne tant d'importance aux fichiers d'autorité. A condition de les porter à l'échelle du Web (notamment leurs identifiants), les fichiers d'autorités utilisés dans les bibliothèques sont un bon moyen d'identifier une entité quelconque de manière univoque, donc de la réidentifier d'un contexte à un autre.

La formalisation RDF de TEF est en cours. Ce travail est grandement facilité par l'existence de la modélisation TEF, même incomplète, par les travaux actuels sur l'encodage du Dublin Core en RDF[8] et par l'ontologie FRBR en OWL publiée par Ian Davis et Richard Newman[9].

J'entrevoie au moins trois points plus délicats.

1. Comment exprimer en RDF des vedettes matière conformes à RAMEAU (les LCSH français) ? Le vocabulaire SKOS fait sans doute partie de la solution.[10]

2. Comment exprimer en RDF le lien aux autorités ? Si on transforme l'identifiant d'une notice d'autorité de personne en identifiant global (URI), à quoi réfère cet URI ? à la personne ou bien à sa notice d'autorité ? On doit considérer la relation "a-pour-autorité" comme une propriété unique, qui ne peut être partagée par deux personnes différentes. Le modèle FRANAR[11], qui modélise les données d'autorité comme les FRBR l'ont fait pour les données bibliographiques, est un acquis précieux pour aller dans cette direction.

3. Comment exprimer en RDF le fait qu'une thèse est un texte *validé par un jury* ? Il faut sans doute imaginer des formulations qui permettent de rapprocher cette validation par un jury de thèse des notions plus génériques de validation scientifique, d'évaluation, de recommandation voire de confiance, dernier étage du Web sémantique. On pourrait réutiliser à cet effet l'ontologie Trust de Jennifer Golbeck[12] et interpréter la validation par le jury comme une confiance accordée au doctorant *eu égard à la discipline* :

[8] http://dublincore.org/architecturewiki/DCRDFTaskforce
[9] http://vocab.org/frbr/core
[10] http://www.w3.org/2004/02/skos/
[11] http://www.ifla.org/VII/d4/wg-franar.htm
[12] http://trust.mindswap.org/trustOnt.shtml

```
− <rdf:RDF>
  − <tef:Jury rdf:ID="Jen">
    − <trust:trustsRegarding>
      − <trust:TopicalTrust>
          <trust:trustSubject rdf:resource="#discipline_XYZ"/>
          <trust:trustedPerson rdf:resource="#Dan"/>
          <trust:trustValue>8</trust:trustValue>
        </trust:TopicalTrust>
      </trust:trustsRegarding>
    </tef:Jury>
  </rdf:RDF>
```

La confiance du jury n'est pas absolue, ni *ad hominem*. Mais elle n'est pas non plus un simple satisfecit portant sur un texte. C'est plutôt, en théorie, la reconnaissance d'un pair dans une discipline. Il va de soi que ces suggestions sont loin de résoudre la question.

CONCLUSION

TEF veut répondre à deux objectifs qui peuvent apparaître comme contradictoires. D'un côté, il y a la nécessité d'un format d'échange qui permette le transfert de métadonnées strictement validées, dans des conditions assez contraignantes qui sont fixées par des textes juridiques nationaux. De l'autre, il y a la volonté de rendre les métadonnées plus faciles à réutiliser par quiconque et à agréger à d'autres métadonnées présentes sur le Web.

Pour satisfaire ces deux objectifs, la stratégie de TEF est de proposer deux modes d'encodage à partir d'une même modélisation des thèses et de leurs métadonnées : XML et METS pour les échanges stricts ; RDF pour la mise à disposition des métadonnées dans toute leur richesse.

Les métadonnées sont trop coûteuses et trop utiles pour n'être utilisées qu'une fois. Comme on l'a vu, les métadonnées TEF elles-mêmes ne partent pas de zéro. Dans un système d'information efficace, elles peuvent en grande partie être dérivées d'informations générées au cours du traitement administratif de la thèse. En aval, les métadonnées TEF exigées par le circuit national des thèses numériques peuvent aussi se prêter à de multiples usages, qu'il s'agisse d'infométrie ou d'évaluation de la recherche. D'une manière générale, si une organisation décide de partager ses données sur le Web de manière ouverte, il est important de les exposer sous le plus de formes possibles, pour ne pas préjuger de l'utilisation que d'autres organisations pourraient en faire.

Site TEF :
http://www.abes.fr/abes/documents/tef/index.html

Blog TEF :
http://tefsav.canalblog.com/