Spécification d'un outil de normalisation

Projet SYNTAX – janvier 2005

Passage d'un format text standard à
un format XML normalisé

http://syntax.loria.fr

Sommaire

- La méthode d'approche
- La phase d'analyse
- La phase de normalisation
- Conclusion
- Perspectives

Problèmatique

• Besoin de normaliser une ressource au format XML (selon Modélisation + catégories de données)

• Trouver une méthodologie de normalisation fiable et fonctionnelle

La courbe du soleil (1)

Méthode systémique : la courbe du soleil (MERISE)

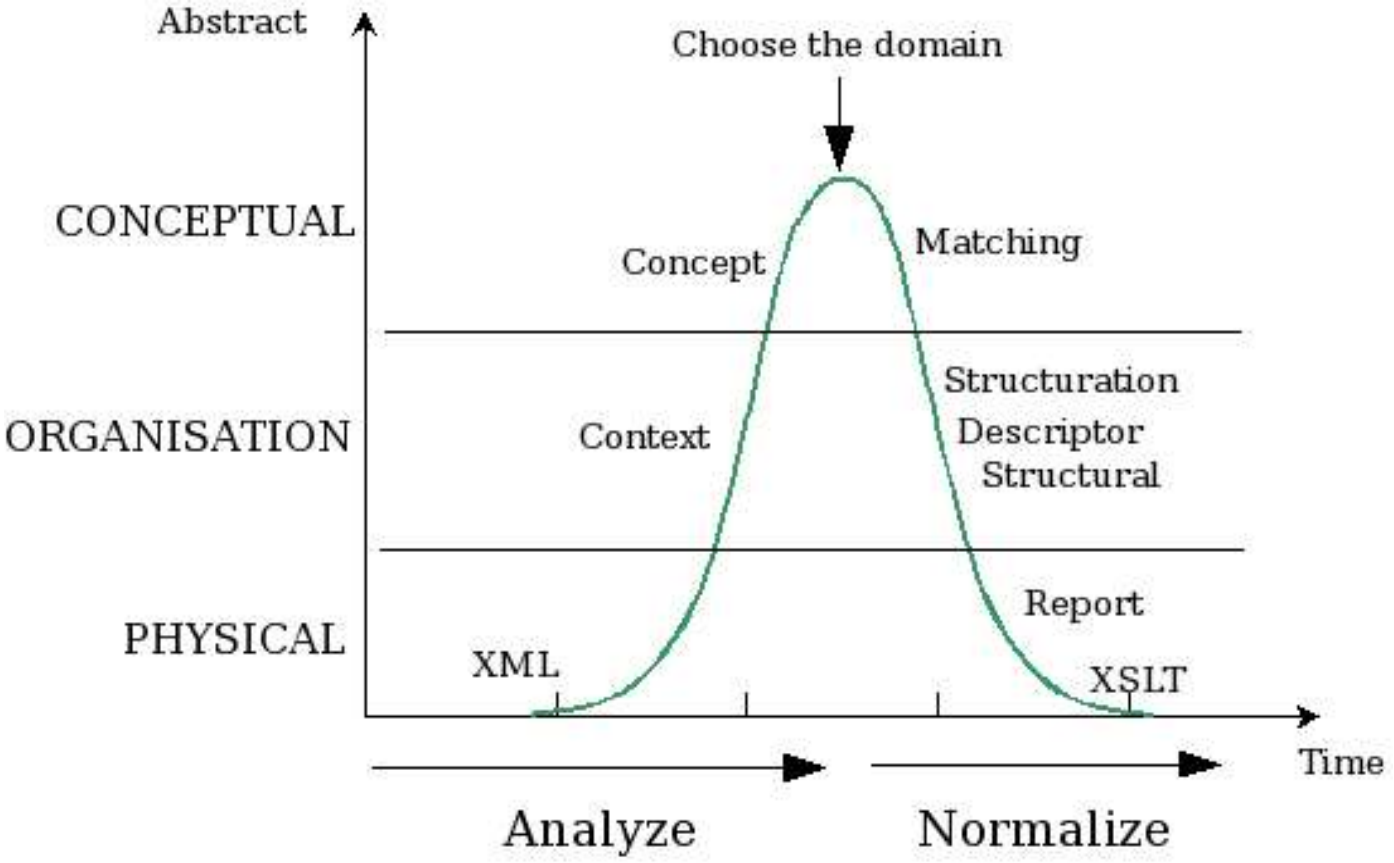
2 phases :

- L'analyse du document à normaliser (l'actuel)
- La normalisation proprement dite (le futur)

3 niveaux d'approche :

- Physique (le plus concret)
- Organisationnel
- Conceptuel (le plus abstrait)

La courbe du soleil (2)



La phase d'analyse

But : extraire les concepts véhiculés par une ressource (ex : fichier XML) : phase d'abstraction

2 niveaux :

- la contextualisation
- la conceptualisation

La contextualisation

1) Trouver tous les contextes possibles d'une instance XML (extraction).

Definition : Un contexte est un chemin issu du parcours d'un arbre XML en considérant ou non les noeuds (elements) et leurs valeurs (#PCDATA), les attributs et leurs valeurs (#CDATA) comme autant de choix de parcours possibles.

La contextualisation (2)

```
<text-structure>
  <block id=1>
    <line type=nom>pierre</line>
    <line type=age>24</line>
    <line type=note>12</line>
  </block>
  <block id=2>
    <line type=nom>jean</line>
    <line type=age>18</line>
    <line type=note>8</line>
  </block>
</text-structure>
```

extraction

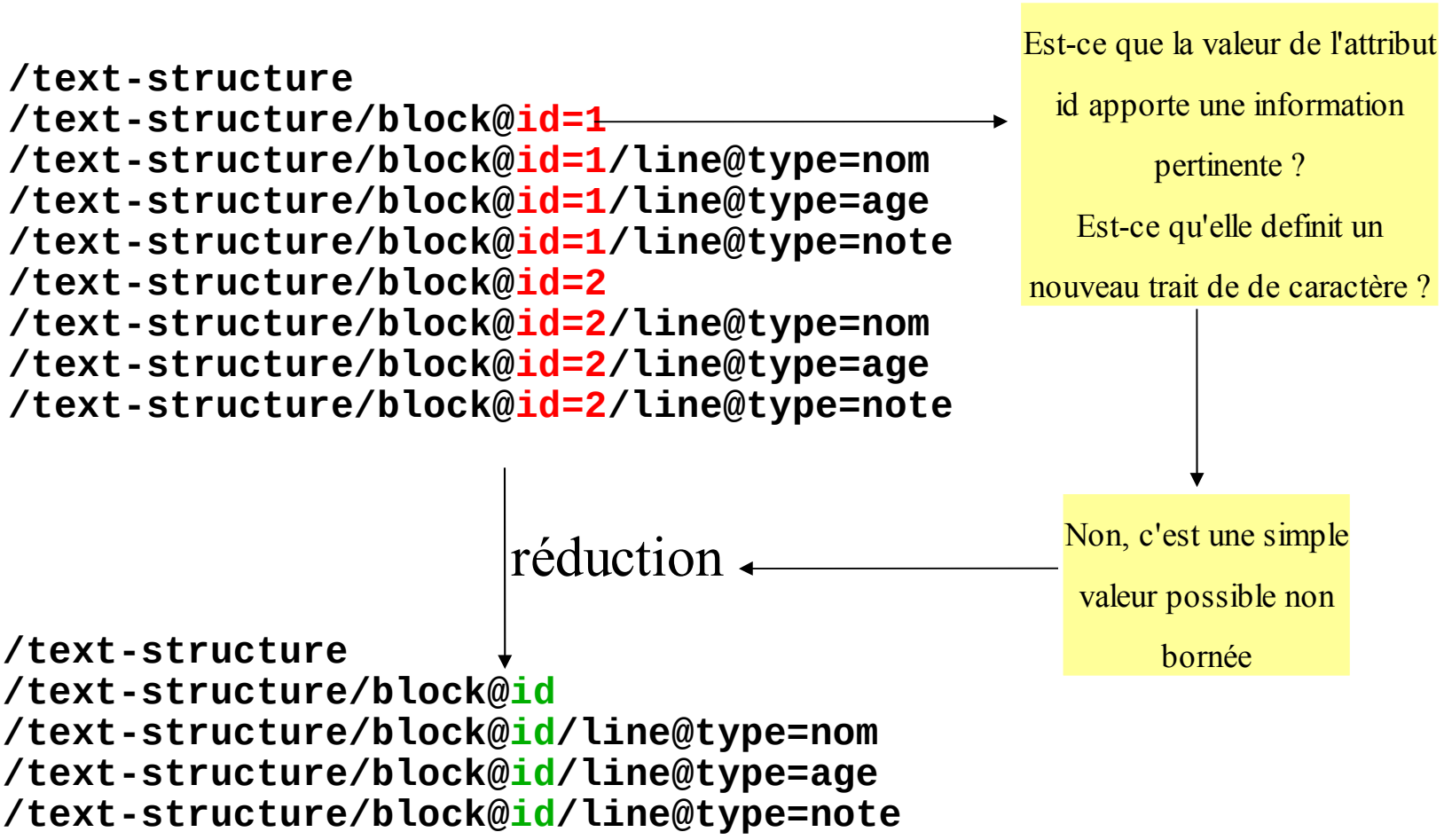
Paramètres : ouverture de la contextualisation à toutes les valeurs d'attribut (#CDATA), fermeture aux valeurs d' élément (#PCDATA).

```
/text-structure
/text-structure/block@id=1
/text-structure/block@id=1/line@type=nom
/text-structure/block@id=1/line@type=age
/text-structure/block@id=1/line@type=note
/text-structure/block@id=2
/text-structure/block@id=2/line@type=nom
/text-structure/block@id=2/line@type=age
/text-structure/block@id=2/line@type=note
```

La contextualisation (3)

2) Réduire la liste de contextes manuellement (reduction) afin d'obtenir une liste pertinente et suffisante pour caractériser l'organisation du fichier XML en cours d'analyse.

La contextualisation (4)



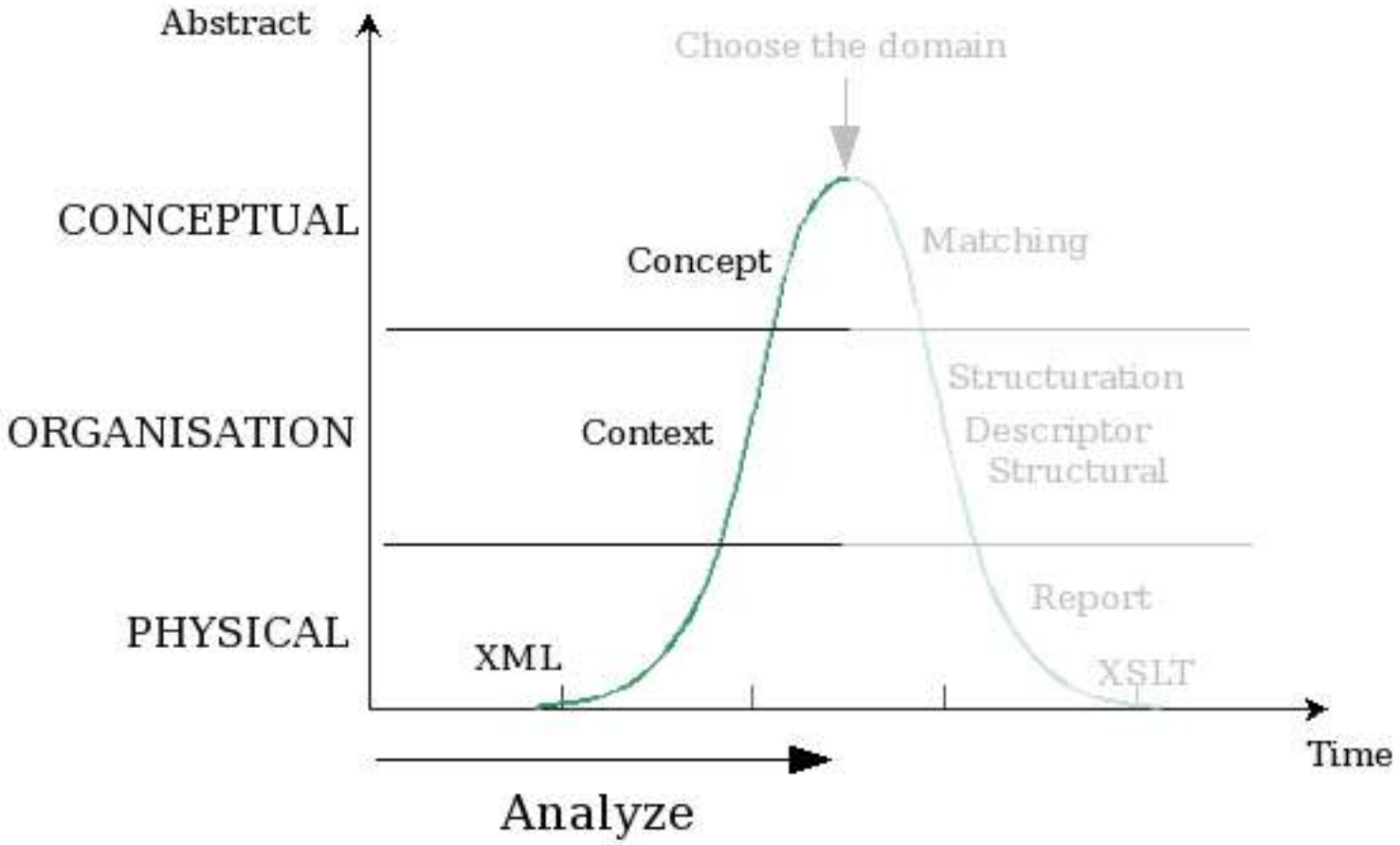
La conceptualisation (1)

But : caractériser les éléments terminaux des contextes (attributs ou elements terminaux d'un parcours d'arborescence XML) avec un concept sous la forme d'une description en langage naturel.

La conceptualisation (2)

Contextes pertinents	Concepts en langage naturel
/text-structure	
Text-structure	un relevé de notes
/text-structure/block@id	
Block	un élève
@id	le numéro de l'élève
/text-structure/block@id/line@type=nom	
Line	un enregistrement
@type=nom	le nom de l'élève
/text-structure/block@id/line@type=age	
Line	un enregistrement
@type=age	l'age de l'élève
/text-structure/block@id/line@type=note	
Line	un enregistrement
@type=note	la note de l'élève sur 20

Résumé


Abstract
CONCEPTUAL
ORGANISATION
PHYSICAL
Choose the domain
Concept
Matching
Structuration
Context
Descriptor
Structural
Report
XML
XSLT
Time
Analyze

La phase de normalisation

But : faire le lien entre l'existant, représenté par des concepts, et la norme : phase de conception

5 étapes :

- Specification du méta-modèle
- La collecte des catégories de données (matching)
- La structuration
- La normalisation des descripteurs
- La normalisation des structurants
- La création du rapport et de la XSLT

La spécification du méta-modèle

Un choix doit être arrêté sur le profile normatif que l'on souhaite mettre en place (TMF, MAF...) afin de guider le reste du processus.

- Acceptation d'un méta-modèle normalisé pré-existant,

ou bien

- Selection d'un méta-modèle personnel (construit par l'utilisateur).

La collecte des catégories de données (1)

But : faire correspondre un maximum de description de concepts issus de l'analyse avec des concepts normés (= catégories de données)

- Utilisation du DCS spécifique au méta-modèle,

ou bien

- Utilisation d'un DCS sélectionné par l'utilisateur.

La collecte des catégories de données (2)

Etape décisive :

- Mise en evidence des lacunes dans le méta-modèle (impossibilité de transcrire des informations structurantes),
- Révèle la non-complétude du DCR (pas de catégorie de donnée pour normer un concept),
- Conditionne le reste de la normalisation (besoin d'un méta-modèle et d'un DCS avec une interopérabilité maximum).

La structuration (1)

But : trier les concepts en fonction du méta-modèle normatif choisi. Deux types de concepts possibles :

•Type descripteur : définit ou décrit un concept.

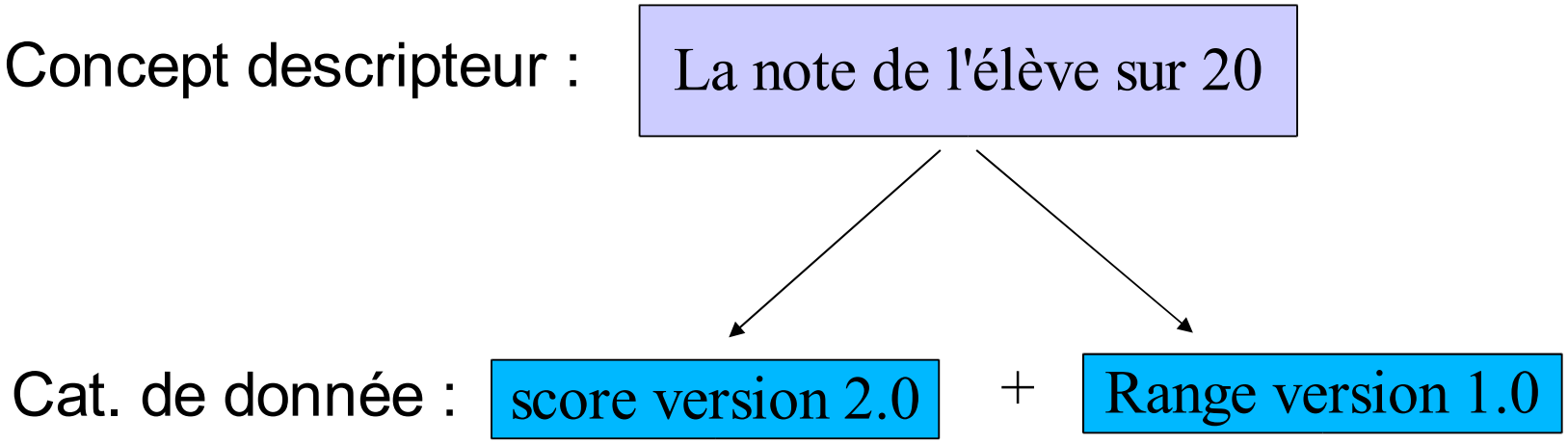
•Type structurant : est directement associé à un composant du méta-modèle choisi (ex : “un usage dans une langue” -> LanguageSection dans Terminological Markup Framework ISO16642:2003),

La structuration (2)

Description du concept	Descripteur	Structurant
Un relevé de notes		X
Un élève		X
Le numéro de l'élève	X	
Un enregistrement		X
Le nom de l'élève	X	
L'age de l'élève	X	
La note de l'élève sur 20	X	

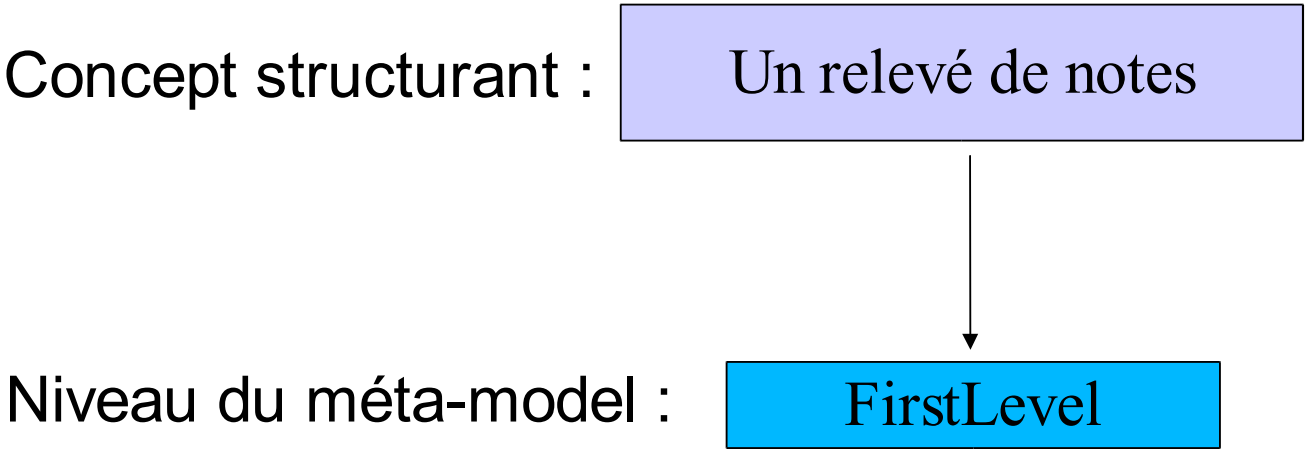
La normalisation des descripteurs

But : faire correspondre un concept descripteur avec une ou plusieurs catégories de données issues d'un DCS construit au préalable.



La normalisation des structurants

But : faire la correspondance entre les concepts structurants avec un niveau de meta-modèle choisi (ex : dans TMF -> TE, LS, TS, TCS).

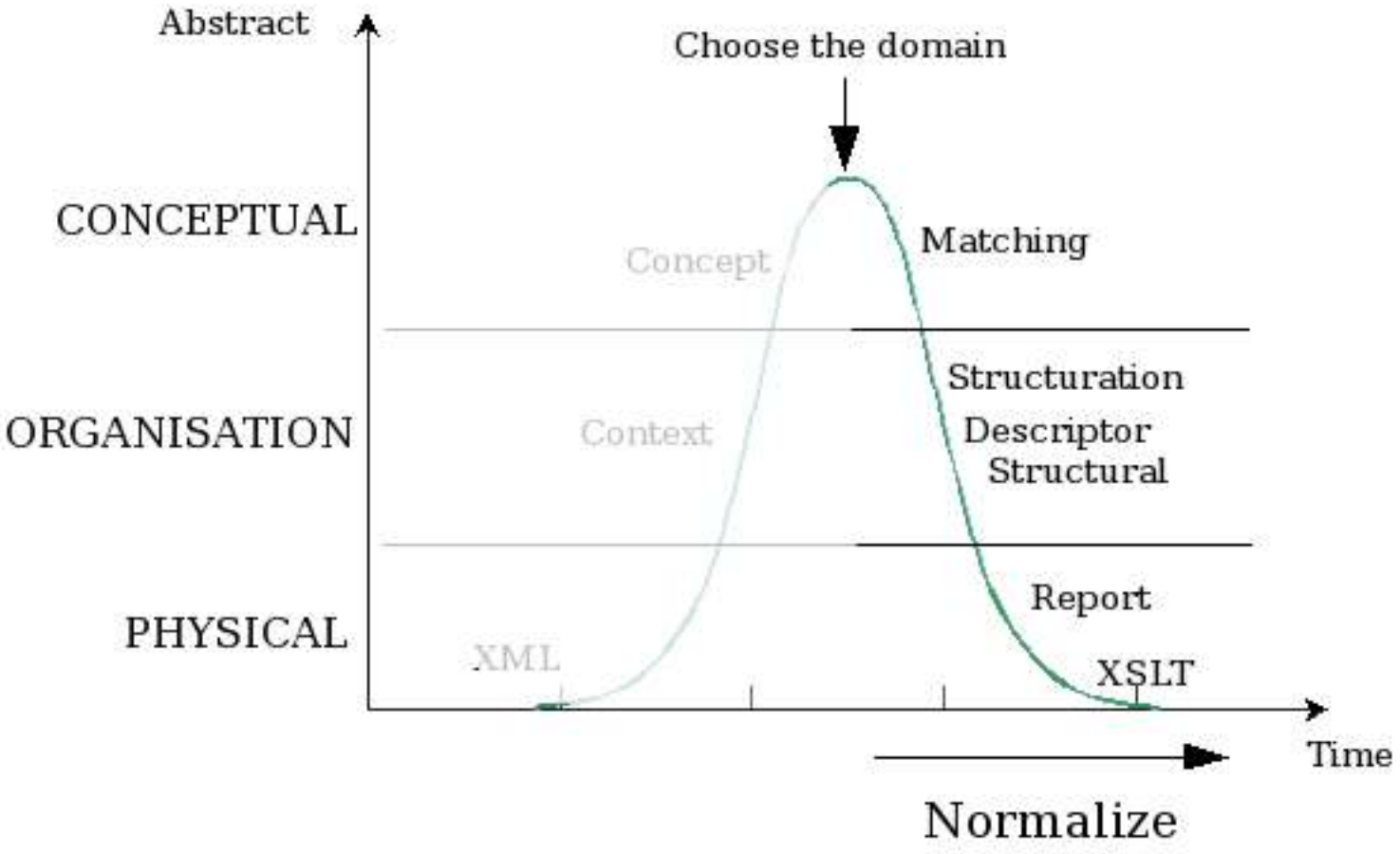


La création du rapport et de l'XSLT

- Générer un rapport retraçant tous les étapes de l'analyse et de la normalisation en n'oubliant pas de décrire le méta-modèle de la norme choisie ainsi que la sélection de catégories de données (DCS) utilisée.

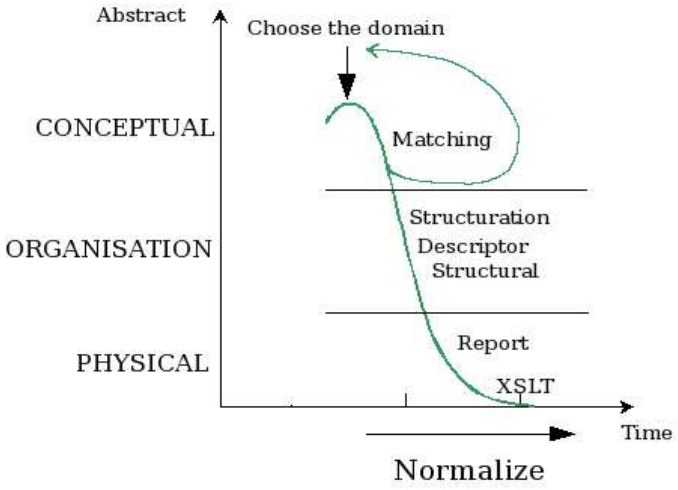
- Générer automatiquement un canevas de 2 feuilles XSLT qui permettent de convertir un fichier XML de départ vers XML normé (XSLT), et de convertir le XML normé en XML de départ (XSLT-1).

Résumé


Abstract
Choose the domain
CONCEPTUAL
Concept
Matching
Structuration
ORGANISATION
Context
Descriptor
Structural
Report
PHYSICAL
XML
XSLT
Time
Normalize

Conclusion

Peut être utilisé pour la création de norme par itérations au niveau conceptuel de la phase de normalisation



Révélateur de l'interopérabilité entre une ressource et une norme

Perspectives

Première tentative d'implémentation en PHP 4 afin de l'intégrer à Syntax Suite mais problèmes :

- Faible robustesse du parser XML (paquet de 4Mo max, sensibilité à l'UTF-8 et aux caractères spéciaux),
- Temps de traitements > 20 mn (contextualisation de n instances -> !n contextes),
- Capacité limitée de travailler sur de gros fichiers en client-serveur (maximum 10 Mo)

Afin de résoudre ces problèmes, une nouvelle implémentation est prévue en JAVA.