

Introduction à XML

- Définitions
- Applications
- Règles
- Affichage et transformation
- Documentation mathématique
- Métadonnées
- En conclusion

Elizabeth Cherhal
Ecole Thématique
"Documentation en
mathématiques"

Luminy,
Octobre 2004

XML n'est pas...

- Une norme de description bibliographique
- « *Enfin le moyen pour que les documentalistes s'approprient la documentation en ligne* » (entendu aux journées INIST 2002)
- La solution aux problèmes de l'hétérogénéité des catalogues du RNBM
- Un gestionnaire de bases de données
- Un langage de programmation

Définitions

Applications
Règles
Affichage et
Transformations
Documentation
mathématique
Métadonnées
En conclusion

Généalogie de la famille «ML»

Définitions

Applications
Règles
Affichage et
Transformations
Documentation
mathématique
Métadonnées
En conclusion

- Le père: SGML (system generalized markup language), né vers 1980, flexible et compliqué.
- L'enfant: HTML (hypertext markup language), basé sur SGML, né au CERN début 1990, langage simple, facile à encoder, et lisible. HTML s'est échappé du CERN pour conquérir le monde
- **Le deuxième enfant: XML** (extensible markup language)... Né vers 1996 il a de nombreux petits frères (MathML CML p ex).

XML, Quand? Pourquoi? Quoi ?

■ 1 Quand?

- Mis au point par le W3C en 1996
- Spécifications de la version 1.0 écrites en 1998
- Il ne devait jamais y avoir de version supplémentaire, mais...
 - ♦ dernière mise à jour en février 2004
 - ♦ Une version 1.1 en 2004 avec des changements mineures
- C'est un langage reconnu (par le w3C) depuis 1998

Définitions

Applications
Règles
Affichage et
Transformations
Documentation
mathématique
Métadonnées
En conclusion

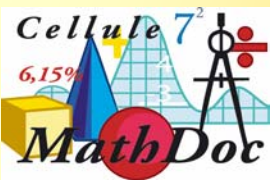
XML, Quand? Pourquoi? Quoi?

■ 2 Pourquoi?

- La grande limite d'HTML est qu'il est quasiment impossible de réutiliser l'information.
- SGML était considéré comme une norme trop lourde et inadaptée au traitement des documents pour le web.
- XML essaie de combiner la flexibilité de SGML avec la simplicité de HTML.

Définitions

Applications
Règles
Affichage et
Transformations
Documentation
mathématique
Métadonnées
En conclusion



XML, Quand? Pourquoi? Quoi?

- 3 Quoi ? extensible markup language
- Markup
 - Balisage: basé sur des balises (ou éléments) ouvrantes et fermantes
- Language
 - Comporte des règles (de grammaire) strictes.
- Extensible
 - On peut inventer ses propres balises

Définitions

Applications
Règles
Affichage et
Transformations
Documentation
mathématique
Métadonnées
En conclusion

Elizabeth Cherhal
Ecole Thématique
"Documentation en
mathématiques"

Luminy,
Octobre 2004


XML/HTML: les différences

2 principes de base restent valables: simplicité, **lisibilité** (par une machine, et par un humain)

Définitions

Applications
Règles
Affichage et
Transformations
Documentation
mathématique
Métadonnées
En conclusion

- **HTML**
- Langage figé: défini par le W3C
- Exprime d'avantage la forme que le contenu
- Est interprété par un navigateur
- Laxisme
- **XML**
- Langage extensible
- Exprime uniquement le contenu
- Doit subir un retraitement pour être visible à travers un navigateur
- rigueur

 Notons l'existence du XHTML, du HTML avec la rigueur de XML

Que fait-on avec XML?...

- **Des documents structurés:** Livre, article, etc.... Les projets « revues.org », « euclid », « cyberthèses », « sparte »... sont basés sur des documents en XML. (ou convertis en XML)
- **Des métadonnées:** De très nombreux projets et applications échangent des métadonnées en XML. L'exemple le plus « universel » est OAI.
- **Des sites web:** On peut imaginer un site dont tout le contenu est en XML, et la conversion en HTML se fait « à la volée ».
- XML passe pour être un « format pérenne ».

Créer un document XML ?

Définitions

Applications

Règles

Affichage et

Transformations

Documentation

mathématique

Métadonnées

En conclusion

- En partant de zéro: des outils existent. On trouve des listes un peu partout, par exemple <http://www.xmlsoftware.com/editors.html>
- La plupart du temps, par conversion d'un autre format: bases de données, fichiers structurés divers...

Documents bien formés /documents valides

Contrairement à SGML, XML comporte la notion de document bien formé

Définitions
Applications
Règles
Affichage et
Transformations
Documentation
mathématique
Métadonnées
En conclusion

■ Un Document **bien formé** répond aux règles de base de XML

- Balises (éléments) ouvrantes /fermantes avec texte au milieu
- Balises « vides »
- Pas d'imbrication (un élément ne peut pas contenir un élément du même nom)

■ Un document **valide**

- Doit se conformer à une DTD ou un schéma
- Doit respecter exactement la DTD ou le schéma (ne pas changer l'ordre des éléments, par exemple)

Exemple fictif de document bien formé

Définitions
Applications
Règles
Affichage et
Transformations
Documentation
mathématique
Métadonnées
En conclusion

« Rencontre »
est un élément

Ainsi que «
organisateur »,
« responsables
», etc...

```

<?xml version="1.0" encoding="ISO-8859-1" ?> ← entête
<?xml-stylesheet type="text/css" href="ecolthem.css"?>
<rencontre type="ecole"> ← « type » est un attribut
  <titre>Documentation en mathématiques</titre>
  <organiseurs>
    <organisateur>Réseau National des Bibliothèques de
      Mathématiques</organisateur>
    <organisateur>Cellule MathDoc</organisateur>
  </organiseurs>
  <dates>
    <debut>2004-10-11</debut>
    <fin>2004-10-15</fin>
  </dates>
  <responsables>
    <resp_scientifique>
      <nom>Bost, Jean-Benoît</nom>
      <email>bost@math.u-psud.fr</email>
    </resp_scientifique>
    <resp_admin>
      <nom>Marchand, Monique</nom>
      <email>mmarchan@ujf-grenoble.fr</email>
    </resp_admin>
  </responsables>
</rencontre>

```

DTD et Schéma

Il arrive souvent que l'on ait besoin de définir un type de document (figer l'ensemble des éléments et attributs possibles)

- **DTD**
 - Document type definition
 - Hérité de SGML
 - Syntaxe différent de XML
 - Pas vraiment de typage des éléments (on ne peut pas dire que l'élément année contient un entier)
 - Compris par tous les parseurs/vérificateurs
- **Schéma**
 - écrit en XML (même syntaxe qu'un document XML)
 - Permet un typage fin des éléments
 - Pas encore beaucoup de vérificateurs disponibles

Définitions
Applications
Règles
Affichage et
Transformations
Documentation
mathématique
Métadonnées
En conclusion

Elizabeth Cherhal
Ecole Thématique
"Documentation en
mathématiques"

Luminy,
Octobre 2004

Des DTD ou schémas publiques

Définitions
Applications
Règles
Affichage et
Transformations
Documentation
mathématique
Métadonnées
En conclusion

- Il est courant d'inventer une DTD pour les besoins d'une application, (ex NUMDAM), mais il y a aussi nombre de DTD publics:
- Livres, articles, etc...(pour l'écriture de documents)
 - **DOCBOOK**: A pour éléments des choses comme: [book](#), [author](#), [preface](#), [chapter](#), [section](#), etc...
 - **TEI(lite)**: Text Encoding Initiative
 - **ERUDIT**: (semi public, disponible sur demande)
- Références bibliographiques
 - **BIBLIOML**: DTD pour les références bibliographiques (UNIMARC ->XML) A pour éléments des choses comme [AbbreviatedTitle](#), [CollectiveUniformTitle](#), [CreationDate](#)
 - **OAI_DC**: schéma pour les métadonnées exposées via OAI, a pour éléments: [creator](#), [title](#), [publisher](#)...

Affichage et transformation

Définitions
Applications
Règles
**Affichage et
Transformation**
Documentation
mathématique
Métadonnées
En conclusion

- Un document XML est structuré logiquement, il n'y a aucune structure physique.
- Pour rendre visible physiquement un document XML, il y a plusieurs méthodes:
 - CSS (Cascading Style Sheets) – feuille de style
 - XSL(extensible stylesheet language) T(ransformation) pour opérer de véritables transformations.
 - Un programme de conversion

Exemple de formatage avec une simple feuille CSS

On prend les données du fichier « exemple » précédent

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<?xml-stylesheet type="text/css" href="ecolthem.css"?>
<rencontre type="ecole">
  <titre>Documentation en mathématiques</titre>
  <organisateur>
    <organisateur>Réseau National des Bibliothèques de Mathématiques</organisateur>
    <organisateur>Cellule MathDoc</organisateur>
  </organisateur>
  <dates>
    <debut>2004-10-11</debut>
    <fin>2004-10-15</fin>
  </dates>
  <responsables>
    <resp_scientifique>
      <nom>Bost, Jean-Benoît</nom>
      <email>bost@math.u-psud.fr</email>
    </resp_scientifique>
    <resp_admin>
      <nom>Marchand, Monique</nom>
      <email>mmarchan@ujf-grenoble.fr</email>
    </resp_admin>
  </responsables>
</rencontre>
```

Documentation en mathématiques

Organisateurs: Réseau National des Bibliothèques de Mathématiques, Cellule MathDoc,

Dates: du 2004-10-11 au 2004-10-15

Responsables: (scientifique) Bost, Jean-Benoît (bost@math.u-psud.fr)

(administratif) Marchand, Monique (mmarchan@ujf-grenoble.fr)

Définitions
Applications
Règles
**Affichage et
Transformation**
Documentation
mathématique
Métadonnées
En conclusion

Elizabeth Cherhal
Ecole Thématique
"Documentation en
mathématiques"

Luminy,
Octobre 2004

CSS/XSLT

- CSS est à peu près compréhensible par un non informaticien, **mais** on ne peut opérer que des transformations simples à l'aide de CSS. (exemple, les rencontres du CIRM).
- XSLT nécessite de connaître des techniques simples de programmation et d'algorithmique.
- XSLT peut servir à générer de l'html, du XML, du pdf, du LaTeX, il est à la base de « sites web en XML »...

XML et la documentation Mathématique

- Quand XML essaie d'écrire des formules mathématiques, il s'appelle MathML.
- MathML est une recommandation du w3C depuis 1999. Il propose deux moyens d'encodage des math: présentation et contenu..

Définitions
Applications
Règles
Affichage et
Transformation
**Documentation
mathématique**
Métadonnées
En conclusion

```
<mrow>
<mrow>
<msup>
<mi>x</mi>
<mn>2</mn>
</msup>
<mo>+</mo>
<mrow>
<mn>4</mn>
<mo>&InvisibleTimes;</mo>
<mi>x</mi>
</mrow>
<mo>+</mo>
<mn>4</mn>
</mrow>
<mo>=</mo>
<mn>0</mn>
</mrow>
```

$$x^2 + 4x + 4 = 0 \quad (\$x^2+4x+4=0\$)$$

contenu →

← présentation

```
<reln>
<eq/>
<apply>
<plus/>
<apply>
<power/>
<ci>x</ci>
<cn>2</cn>
</apply>
<apply>
<times/>
<cn>4</cn>
<ci>x</ci>
</apply>
<cn>4</cn>
</apply>
<cn>0</cn>
</reln>
```

Elizabeth Cherhal
Ecole Thématique
"Documentation en
mathématiques"

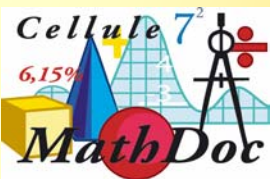
Luminy,
Octobre 2004

MathML

- MathML, vu sa verbosité, a besoin d'être généré par un programme.
- Il existe un certain nombre d'outils capables de générer du MathML (de présentation plus que de contenu), mais ils ne sont pas encore utilisés par les mathématiciens, qui connaissent un seul outil: (La)TeX.
- La flexibilité de (La)TeX (macros personnelles, styles etc...) rend le développement d'outils de conversion LaTeX-> MathML pas très simple à réaliser.
- L'affichage sur le web de formules de math n'est donc pas ce qu'il pourrait être.
- On peut cependant en voir [un exemple](#) dans « euclid »:

(La)TeX et PDF

- (La)TeX donne un résultat typographique sans comparaison avec ce qu'un navigateur web sait afficher.
- PDF, bien qu'étant un format propriétaire, peut être généré directement par (La)TeX.
 - Il ne s'agit pas d'un simple format « image du texte », car il est possible d'y insérer des liens, de générer des tables des matières, etc...
- Le format de « document mathématique numérique » le plus répandu est pdf. (généré directement par LaTeX).



XML et les métadonnées

Définitions
Applications
Règles
Affichage et
Transformation
Documentation
mathématique
Métadonnées
En conclusion

- XML sert souvent de format d'échange de métadonnées.
- Contrairement à des données en HTML, nous pouvons extraire les données d'un fichier XML, et les réutiliser (alimenter une autre base de données par exemple).
- Il existe des « parseurs » (analyseurs syntaxiques) pour chaque langage de programmation, permettant de décortiquer le XML, et le réutiliser aisément.

Elizabeth Cherhal
Ecole Thématique
"Documentation en
mathématiques"

Luminy,
Octobre 2004

Un mot sur OAIPMH

Définitions
Applications
Règles
Affichage et
Transformation
Documentation
mathématique
Métadonnées
En conclusion

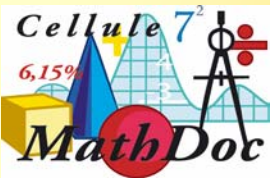
- Open Archives Protocol for Metadata Harvesting
- Basé sur des normes simples:
 - Dublin Core (oai_dc) est la description bibliographique minimale à la base de tout serveur OAI. Il est possible d'offrir en supplément d'autres formats.
 - Une syntaxe de requêtes très simplifiée permet de récupérer (pour retraitement) un fichier XML contenant des métadonnées.

Elizabeth Cherhal
Ecole Thématique
"Documentation en
mathématiques"

Luminy,
Octobre 2004

A quoi « nous » sert XML?

- Actuellement, à tout ce qui n'est pas le document « plein texte », c à d les métadonnées, ou il sert de format d'échange:
- **NUMDAM**: nous échangeons nos métadonnées sous forme d'XML avec nos prestataires. (la DTD NUMDAM sert de modèle à d'autres applications)
- **OAI**: Nous exportons les données de NUMDAM sur un **serveur OAI** en XML.



XML nous sert à récolter des métadonnées:

Définitions
Applications
Règles
Affichage et
Transformation
Documentation
mathématique
Métadonnées
En conclusion

- Différents services de la CMD sont basés sur la récolte de métadonnées:
- La base de prépublications et des thèses:
 - Alimenté entre autres par le CCSD via OAI
- La base LiNuM
 - Bientôt alimenté par le serveur OAI de Gallica
 - Alimenté indirectement par envoi et reconversion de fichiers XML (u-michigan, göttingen).
- Le nouveau projet « minidml »:
 - Alimenté entre autres par OAI (euclid, arXiv)

Elizabeth Cherhal
Ecole Thématique
"Documentation en
mathématiques"

Luminy,
Octobre 2004

En conclusion...

- XML est un standard pour:
 - Structurer un document « primaire »
 - Echanger des métadonnées
- XML exprime un contenu et est indépendant de l'apparence
- XML, pour être utile, doit être transformé:
 - Formatage pour produire un document affichable ou imprimable
 - Transformation pour réutilisation de l'information
- L'utilisation d'XML suppose l'acquisition de connaissances et compétences techniques.

Références

- XML: <http://www.w3.org/XML/>
 - RAY, Eric T, « Learning XML », O'Reilly, 2001
 - De nombreux « introductions » et « tutoriaux » en ligne
- XHTML: <http://www.w3.org/MarkUp/>
- CSS: <http://www.w3.org/Style/CSS/>
- XSL/XSLT: <http://www.w3.org/Style/XSL/>
- MathML: <http://www.w3.org/Math/>
- DocBook: <http://www.docbook.org/>
- TEI: <http://www.tei-c.org/>
- Biblioml: <http://www.biblioml.org/>
- OAI: <http://www.openarchives.org/>
- Et les sites: www.revues.org, www.cybertheses.org, sparte.abes.fr