CNES clearinghouse prototype

Anne HERMETZ[1], Frédéric CHAUVIN[2],
CRIL TECHNOLOGY
2, impasse Henri Pitot
31500 TOULOUSE
Tel: 05.62.16.79.79
Fax: 05.61.20.47.89
anne.hermetz@criltechnology.com[1], frederic.chauvin@critechnology.com[2]

English abstract

Context

The French space agency CNES has decided to build a clearinghouse devoted to earth observation and its environment. CRIL TECHNOLOGY was chosen to develop, in a R&D context, a prototype to ensure feasibility and to investigate XML technology, CCSDS OAIS and a future metadata ISO standard.

The aim of a clearinghouse is to put together information about a special interest of a community in order to give easy access and discovery of pertinent numerical data.

The specification of such a system answers questions as What is a data? What is information about a data? How can we find this kind of information? How can we deliver information? How can we deal with information? The design of such a system solves issues as information format, information input, information storage, information retrieving, information delivering and information localization.

Description

The system that has been developed manipulates information as metadata, that is « data upon data ». It answers:

- **Metadata format**: metadata can follow different field structure for numerical data (eg. FGDC CCSDS) or ISO standards (eg. ISO 19115 Geographic information — Metadata) or for document indexing (eg. TEI or DocBook). To achieve this goal, data is represented by a XML file validated by a related DTD from the pool of accepted DTD's. XML permits to be independent of the underlying data representation.
- **Metadata input**: for easy ingestion, metadata are manipulated with specialized structured editors integrated into the clearinghouse.
- **Metatata storage**: efficient storage of a huge set of metadata is crucial for such a system. New evolutions of the market leader RDBMS product allow native XML support within high performance.
- **Metadata access**: the Web interface ensures easy access for the whole scientific community targeted by the clearinghouse. Access to multiple metadata formats is allowed by an internal XML pivot metadata format. Thematic user profiles speed fields selection for the request.
- **Metadata delivering**: selected profile metadata fields or the whole original metadata can be restored. Users can browse within the metadata in order to access other links (eg. to the pointed raw data).
- **Metadata localization**: requests can be launched on multiple distributed clearinghouses without any difference in the user point of view.
- **System administration** : consists of management of metadata and profiles.

The technologies involved are: Apache server, Tomcat, Apache SOAP, JSP, Java Servlet, Java Applet, Oracle 9i, XML/XSL, TurboXML, XMetaL.

Résumé en français

Contexte

Le CNES a décidé de construire un bureau des métadonnées dans le but de stocker et de restituer des informations concernant les données géographiques terrestres et leurs environnements. CRIL TECHNOLOGY a été choisi pour développer, dans un contexte de R&D, un prototype permettant de valider la faisabilité d'un tel outil mettant en œuvre des nouvelles techniques de type XML, CCSDS OAIS ou norme ISO.

L'enjeu du bureau des métadonnées est de rassembler des informations sous le même horizon, pour en restituer l'intelligibilité dans les meilleures conditions de confort pour les utilisateurs.

Afin de réaliser un outil capable de répondre à ces attentes, il faut se pencher sur des questions simples telles que, qu'est-ce qu'une donnée, qu'est ce qu'une information, comment exploiter l'information, comment trouver l'information, où trouver l'information, comment communiquer l'information, comment gérer l'information ? Par la suite, les interrogations sont plus de l'ordre de la faisabilité d'un tel outil. Elles portent sur le format des informations, sur leur stockage, les moyens d'accessibilité, le niveau technique atteint et mis à disposition.

Description

Le Bureau des Métadonnées est un projet R & T dont le but a été de trouver des réponses à ces questions. Il gère des métadonnées c'est-à-dire des « données sur les données ».

- **Format d'une métadonnée :** Pour qu'une métadonnée soit exploitable, il faut que ces informations soient accessibles aisément. C'est pourquoi, les métadonnées suivent des normes. Le format le plus approprié est le format XML. Il permet le stockage des informations sans se soucier de sa représentation.

  Une métadonnée est un donc fichier XML correspondant à un format précis (DTD).

  De plus, une métadonnée peut être associée à un ensemble de fichiers complétant ses informations.
- **Stockage d'une métadonnée :** Le stockage des métadonnées doit être simple. Une métadonnée doit être identifiable, modifiable. Une métadonnée doit vivre dans le temps. Le choix du stockage a été ORACLE. Ce SGBD permet le stockage de XML sous forme XML et de faire des interrogations sur les balises.
- **Accès aux métadonnées :** L'accès aux métadonnées doit également être simple. La norme des métadonnées n'étant pas unique, il est nécessaire d'extraire les informations typiques de chaque norme pour que l'accès soit unique quelque soit la norme. L'utilisateur doit être guidé dans sa requête, c'est pourquoi, le Bureau des Métadonnées classe les métadonnées suivant des profils d'utilisateur et fournit des critères de sélection pour affiner la recherche.
- **Restitution d'une métadonnée :** La restitution de la métadonnée doit être possible dans son intégralité sous forme conviviale ou sous forme brute et intégrale (métadonnée et donnée)
- **Communication entre serveur de données :** La communauté scientifique étant très vaste, les informations peuvent être stockées sur des serveurs distants. C'est pourquoi, le Bureau des Métadonnées doit communiquer avec d'autres serveurs et transmettre les demandes utilisateurs à ces serveurs. Il doit bien sur restituer les informations trouvées.
- **L'administration du système** se compose de la gestion des métadonnées et de la mise à jour des profils.

Les technologies utilisées sont : serveur Apache, Tomcat, Apache SOAP, JSP, Java Servlet, Java Applet, Oracle 9i, XML/XSL, TurboXML, XMetaL.

Le Bureau des Métadonnées

Le Bureau des Métadonnées (BdM) du CNES a pour vocation de stocker et de restituer des informations. C'est le but de n'importe quel serveur de données. Mais qu'entend-on par donnée ? Une donnée peut prendre n'importe qu'elle forme : un article, une image, un fichier. Difficile de stocker des informations qui n'ont pas de structure homogène. C'est pourquoi le bureau des métadonnées stocke, non pas des informations hétérogènes, mais des informations qualificatives (métadonnées) sur les informations ciblées. Il s'agit alors de déterminer le format des informations qualificatives.
Le but du CNES n'est pas juste de stocker des informations, mais consiste à l'exploitation de celles-ci par la communauté scientifique. C'est pourquoi le Bureau des Métadonnées définit des profils parmi lesquels les scientifiques se reconnaissent. Ces profils permettent de filtrer les métadonnées.
Lors d'une requête, l'ensemble des métadonnées est fourni sous forme de liste. Puis chaque métadonnée peut être visualisée dans sa totalité ou téléchargée.
Dans un premier temps, nous préciserons quelques définitions puis nous aborderons l'aspect fonctionnel, en parcourant le cycle de vie de la métadonnée, et enfin l'aspect technique.

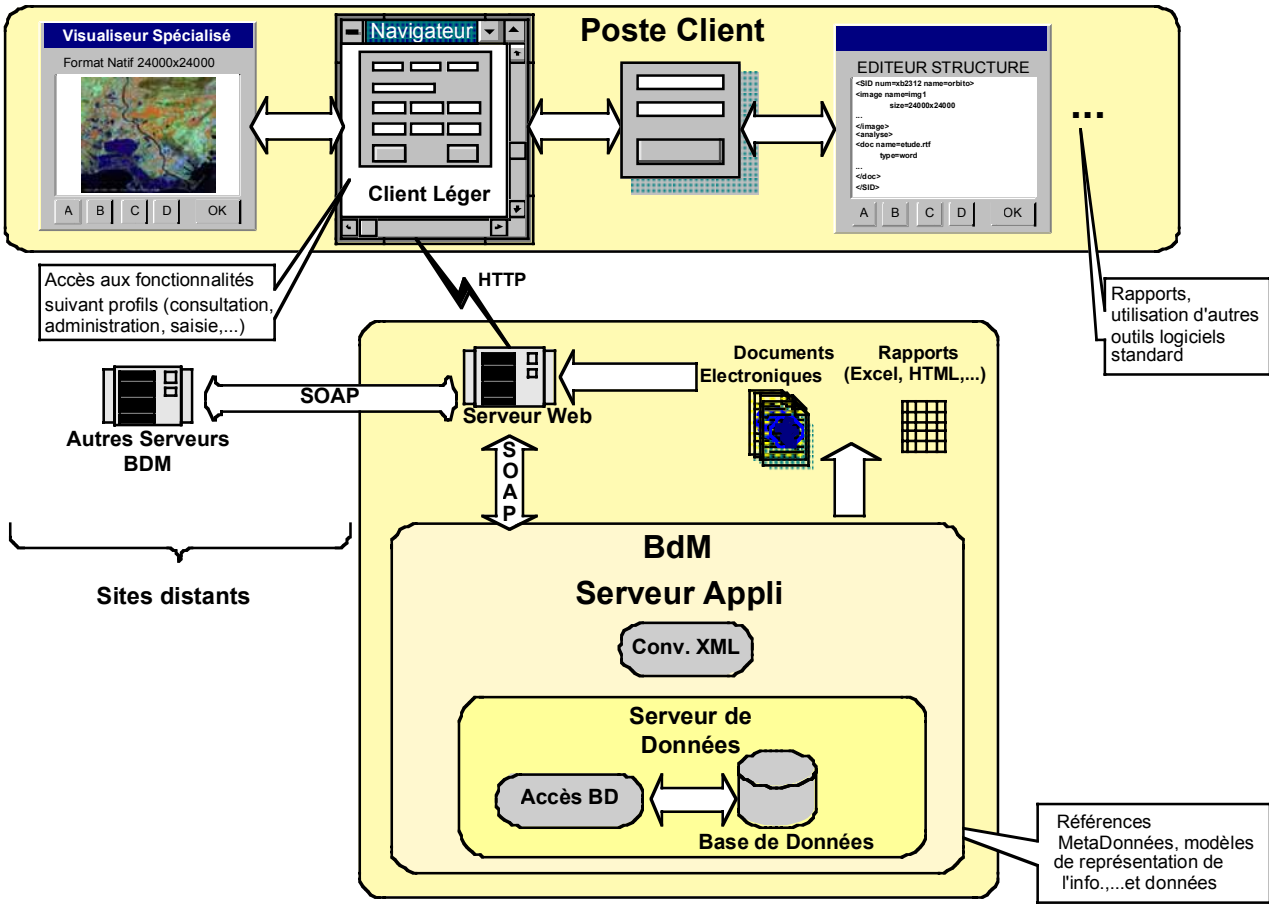

Figure 1 : Schéma des logiciels nécessaires au BdM

Définition

« Métadonnée » signifie donnée sur les données. Une métadonnée est donc un fichier d'un format connu qui décrit des données.

Une donnée peut être de nature différente : une image, un article, un livre. Une donnée n'est pas forcément stockée au sein du BdM mais peut se trouver sur un autre site.

Le format des métadonnées choisi dans le cadre du prototype du Bureau des Métadonnées est celui de la norme ISO 19115. Cette norme permet de décrire les informations géographiques et terrestres. Nous avons également envisagé d'autres normes telles que la norme FGDC. Le bureau des métadonnées n'étant qu'un prototype, seule la norme ISO 19115 a été intégrée. Par contre le BdM a été conçu dans un souci de permettre la gestion de n'importe quelle norme décrivant des informations géographiques et terrestres.

La description de la norme ISO 19115 peut se faire sous la forme d'une DTD (Document Type Definition) ou d'un XML schéma. La DTD décrit la structure du fichier mais ne fournit pas le type de la donnée (caractères, chaine, entier, …) ce que fournit le schéma. Chacune de ces descriptions fournit une représentation au format XML.
Valider un fichier XML consiste à vérifier sa cohérence avec sa DTD.
L'avantage des fichiers XML est de ne contenir que de l'information. Chaque information est placée entre une balise ouvrante et une autre fermante.

Aspects fonctionnels

Le schéma suivant montre le processus d'ingestion et de restitution d'une métadonnée.

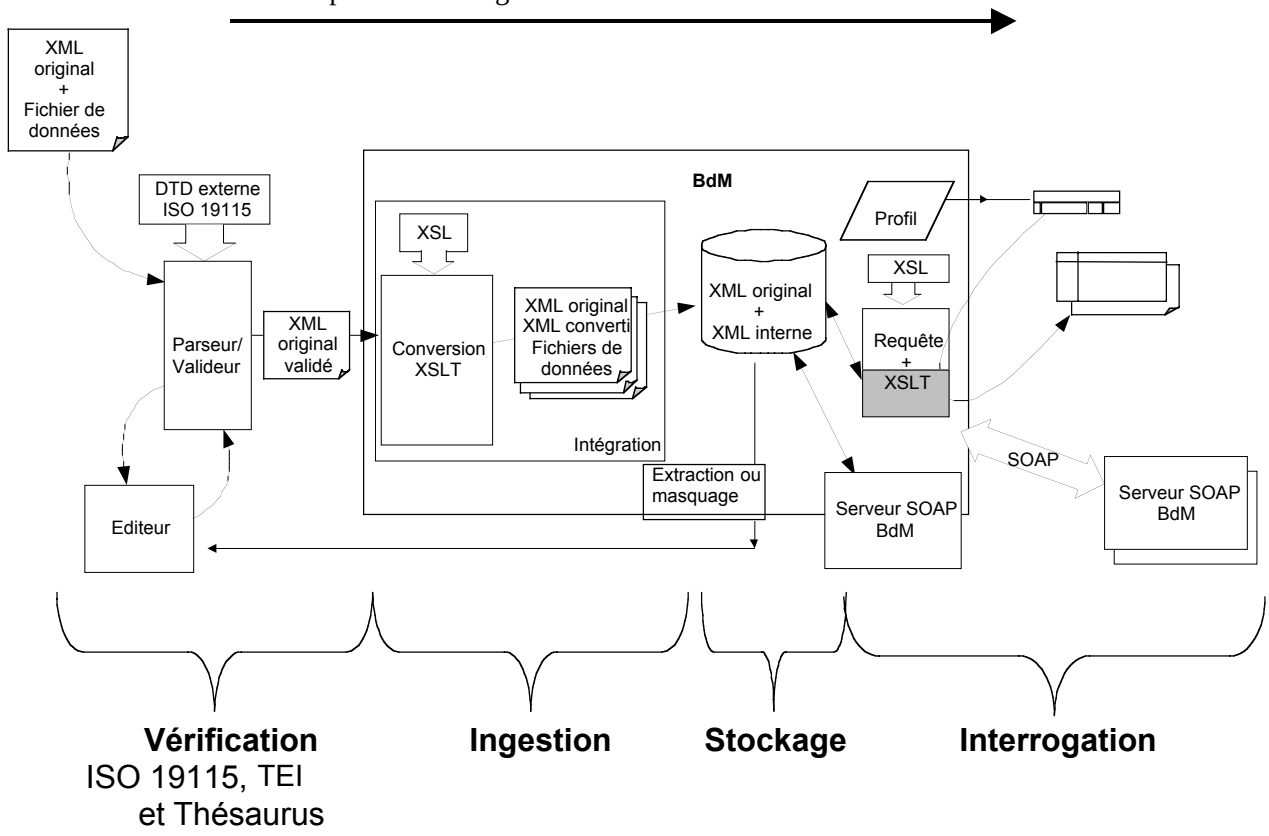

Figure 2 : Cycle de vie d'une métadonnée dans le BdM

Phase de vérification

Nous avons choisi d'utiliser la forme DTD de la norme, d'une part parce que celle-ci était disponible (elle représente tout de même 1000 lignes), d'autre part, les schémas ont été normalisés récemment mais les produits du marchés n'ont pas encore largement intégré la dernière spécification.
La phase de vérification consiste à valider la cohérence du fichier XML. Cette vérification s'effectue à l'aide d'un parseur. Le parseur est un outil de vérification des DTD, des fichiers XML et de la cohérence des

fichiers XML par rapport à leur DTD. Le parseur utilisé est également un éditeur permettant de générer des gabarits de fichiers XML.
Les éditeurs fournissent des aides à la création et à la visualisation des fichiers XML.

L'éditeur fournit une aide à la construction du fichier XML :

- Présentation des fichiers sous forme de tableaux s'ouvrant et se fermant au niveau des balises,
- Fourniture des balises possibles à insérer,
- Indication des erreurs.

Les éditeurs testés sont TurboXML de chez TIBCO et XMLSpy. Le choix s'est porté sur TurboXML qui permet à l'utilisateur de donner plusieurs vues au fichier XML. Sur chaque vue, il est possible de donner du relief à la métadonnée afin de faciliter la saisie des balises.

Tableau 1 : Exemple DTD, XML

DTD	XML
<!ENTITY % MD_Metadata "(fileIdentifier, language?, characterSet?, parentIdentifier**?**, hierarchyLevelCode*****, hierarchyLevelName*, contact, dateStamp, metadataStandardName, metadataStandardVersion?, distributionInfo?, identificationInfo+, applicationSchemaInfo* , portrayalCatalogueInfo*, metadataMaintenance?, metadataConstraints*, dataQualityInfo*, spatialRepresentationInfo*, referenceSystemInfo*, contentInfo*, metadataExtensionInfo*)"> <!ELEMENT Metadata %MD_Metadata;> <!ELEMENT fileIdentifier (CharacterString)> …	<?xml version="1.0" encoding="UTF-8"?> <Metadata> <fileIdentifier> <CharacterString> <String>24</String> </CharacterString> </fileIdentifier> <contact> <role value="custodian"/> </contact> <dateStamp> <DateTime/> </dateStamp> <metadataStandardName> <CharacterString> <String>ISO19115</String> </CharacterString> </metadataStandardName> <identificationInfo> <citation> <title> <CharacterString> <String>Titre</String> </CharacterString> </title> <date> <dateType value="creation"/> </date> <citation> <identificationInfo> </Metadata>

Avec :

- + : 1.. *,
- * : 0..*,
- ? : 0..1.

Phase d'ingestion et de stockage

La phase d'ingestion permet de stocker les fichiers de métadonnées et les fichiers de données associés. Elle prend la métadonnée en entrée et la convertit au format interne du bureau des métadonnées. Elle consiste en :

- Une vérification du format de la métadonnée par rapport à la DTD ISO 19115.
- Une conversion XSLT de la métadonnée pour en extraire les informations nécessaires au fichier interne de traitement.
- Un stockage dans la base.

La vérification de la métadonnée à l'entrée du bureau des métadonnées est une protection pour ne filtrer que les métadonnées de la norme choisie.
La traduction du XML en entrée (conversion XSLT) permet de récupérer certaines informations de la métadonnée afin de renseigner le fichier interne associé. C'est sur ce fichier que l'ensemble des fonctionnalités s'appuie. Ce fichier interne, proche de la norme ISO 19115, est au format XML. Il correspond à une DTD interne. Ce fichier possède deux intérêts :

- le premier est que son format est unique quelque soit la norme en entrée : en effet, la norme choisit en entrée du BdM est la norme ISO 19115 mais à terme d'autres normes pourront être prises en compte. Chaque fois que l'on ajoute une nouvelle norme, il suffit de développer le programme de conversion vers le format du fichier interne.
- le second est qu'il permet d'effectuer des requêtes plus rapides.

Les informations retenues pour le fichier interne sont :

- L'identifiant du fichier,
- La personne ou l'organisme qui publie la métadonnée,
- La date de création ou de modification de la métadonnée,
- Le résumé de la métadonnée,
- Le sujet de la métadonnée,
- Les mots clés représentés par des triplets mots clés, thesaurus d'appartenance, type,
- La catégorie de la métadonnée,
- Les dates de début et de fin d'application de la métadonnée,
- Les coordonnées géographiques,
- Les informations des catalogues,
- Les noms et types des fichiers de données associés à la métadonnées.

Phase d'interrogation

L'interrogation du Bureau des Métadonnées est basée sur la notion de profil. Aux profils, sont associés les critères de sélection et de restitution choisis parmi les informations du fichier interne.

Principe du profil

Un profil est un ensemble d'éléments caractérisant une population de scientifique.

Les profils sont gérés par l'administrateur du site. Un profil est décrit par :

- Les catégories choisies parmi une liste,
- Les mots clés liés à ce profil,
- Les critères de sélection,
- Les critères de restitution.

Critères de sélection

Les critères de sélection possibles sont :

- Les coordonnées géographiques : applet Java représentant une mappemonde sur laquelle l'utilisateur dessine un rectangle de sélection.
- Les mots clés présentés sous forme de tableaux et regroupés par type et par thesaurus. L'utilisateur choisit les mots clés dans des listes.
- Les dates de période (recherche entre 2 dates, date de début et de fin, recherche à partir d'une date ou après une date).
- Les saisons : une saison est caractérisée par un intervalle de temps dans l'année représentée par un jour et un mois de début et un jour et un mois de fin
- Les feature type : mots (à saisir) se trouvant dans les catalogues.
- Les termes se trouvant dans la métadonnée.

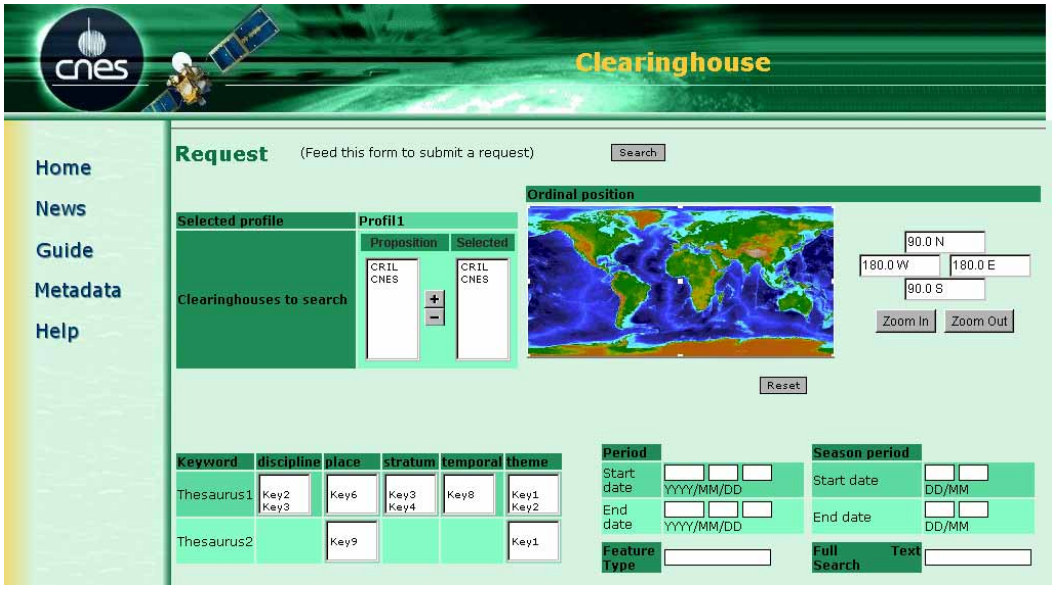

Figure 3 : Interface du module de requête

La recherche fournit, par serveur, la liste des métadonnées trouvées et les présente à l'aide des critères de restitution associés au profil choisi.

Critères de restitution

Les critères de restitution regroupent les informations présentées à l'utilisateur dans la liste des métadonnées trouvées. Les critères possibles sont pris dans le fichier interne.

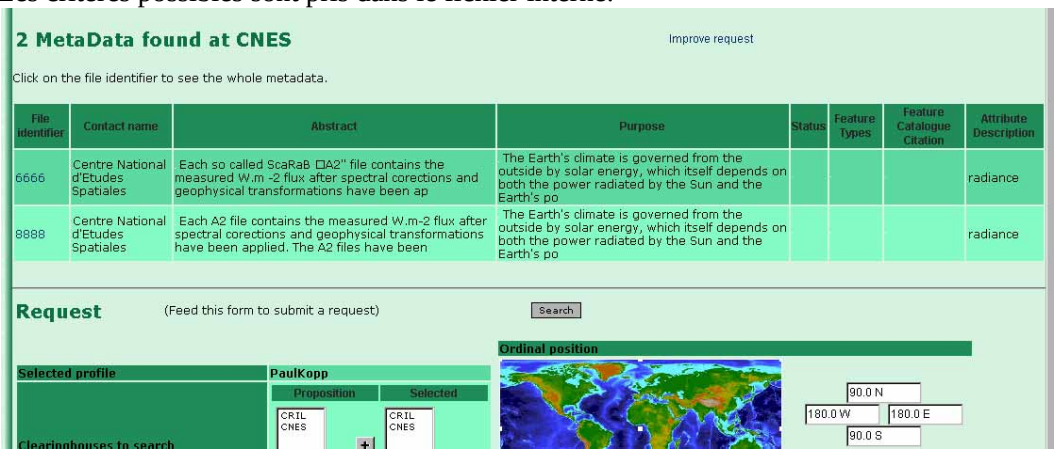

Figure 4 : Interface des métadonnées trouvées

Visualisation et récupération de la métadonnée

La métadonnée peut être téléchargée ou visualisée dans son intégralité.

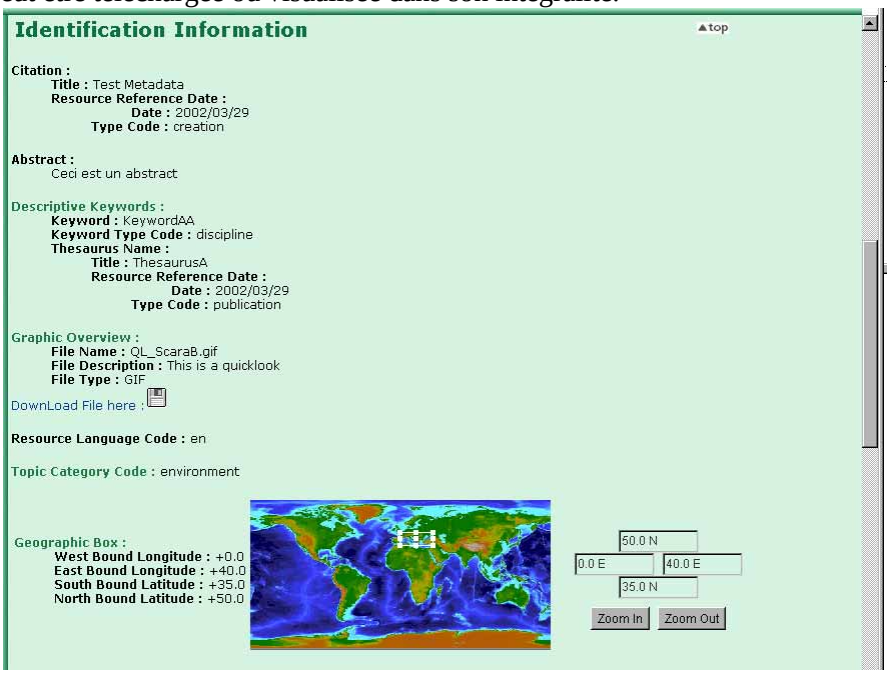

Figure 5 : Visualisation de la métadonnée

La visualisation est réalisée via une traduction XSLT de présentation fournissant une page HTML de la métadonnée.

Certaines informations de la métadonnée peuvent évoluer comme le Personal Identification (PI). En effet, cette information correspond à une personne ou à un organisme qui peut changer d'adresse par exemple. L'administrateur gère une liste de PI. C'est pourquoi, le Bureau des Métadonnées permet la gestion de ces informations.

A la restitution, lorsqu'un PI est trouvé dans la liste des PI, celui de la métadonnée est aussitôt modifié par celui se trouvant dans la base de données. Ainsi, la métadonnée d'origine est conservée mais la restitution fournit des informations à jour.

Aspects techniques du BdM.

Le Bureau des Métadonnées est un site Web écrit en HTML, JSP et JAVA. Le serveur TOMCAT prend en charge l'ensemble des pages HTML du site. Il traduit les pages contenant des JSP en classes Java qu'il exécute à chaque appel de page.

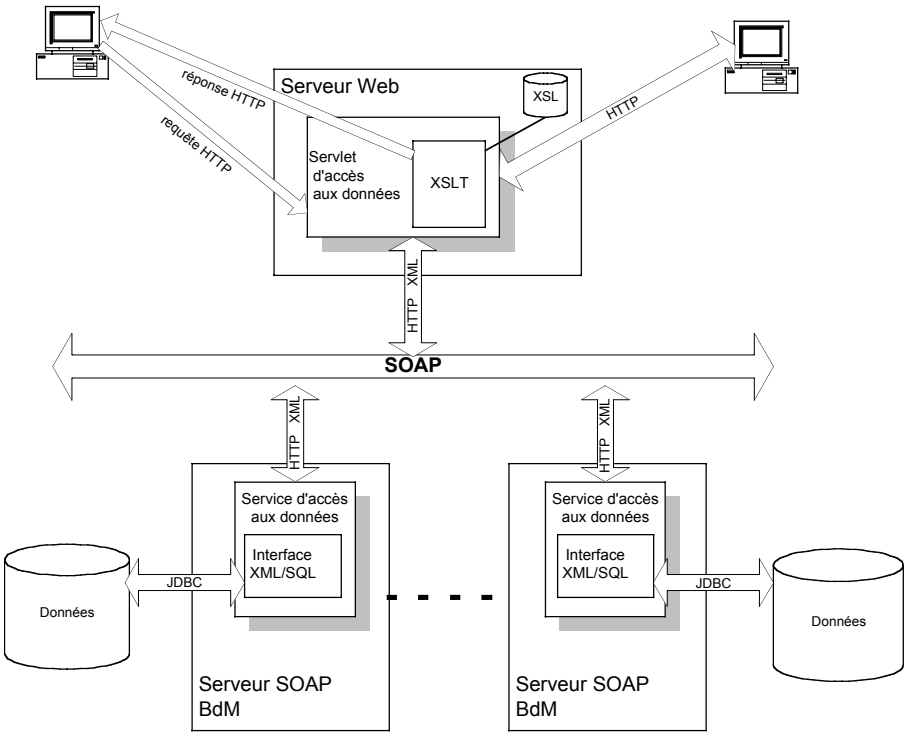

Figure 6 : Architecture technique

Interrogation sur plusieurs BdM

L'utilisateur a la possibilité de choisir les serveurs sur lesquels il souhaite lancer sa recherche.

La recherche s'effectue par le serveur Web, type Apache, sur lequel on installe l'extension permettant la gestion du protocole SOAP (SOAP Apache).
Il fournit les services d'accès aux données du bureau des métadonnées qu'il supporte.

La requête HTTP est transmise au serveur Web via une servlet qui dans le cas d'une recherche de données :

- Formalise la requête,
- Sélectionne les serveurs Bureaux de Métadonnées disponibles,
- Envoie cette requête XML via le protocole SOAP à l'ensemble des serveurs BdM sélectionnés à l'aide d'un format XML interface,
- Récupère les réponses XML des différents serveurs BdM ,
- Convertit ces réponses et prépare la mise en page au format HTML grâce à un traducteur XSLT (Xalan d'Apache).
- La page HTML ainsi préparée est renvoyée au navigateur client via HTTP.
- Les serveurs BdM sont ici des serveurs de données, répondant à une logique simple :
  - Réception d'une requête dans un format standardisable XML,
  - Réponse à cette requête dans un format standard XML supportant le schéma XML de la DTD interface des Bureaux de Métadonnées.

On voit qu'ainsi tous les serveurs BdM sont indépendants du serveur Web qui sert uniquement d'interface avec l'utilisateur.

Ce serveur déclenche sur réception d'une requête :

- L'analyse de celle-ci et sa transformation en requête SQL,
- Envoi de cette requête SQL sur le serveur de base de données via JDBC,
- Conversion de chaque enregistrement trouvé au format XML,
- Envoie de cette réponse au client demandeur (ici le serveur Web).

Le code embarqué dans les pages HTML nécessaire au bon fonctionnement et à l'efficacité de l'interface (contrôle de saisie, ...) est écrit en Javascript 1.1 pour être compatible avec tous les navigateurs.

Technologie XML, XSLT et XPath

Le bureau des métadonnées est innovant sur les aspects techniques utilisés. En effet, il s'agit d'une application entièrement réalisée sur des fichiers XML :

- En entrée, fichier XML norme ISO19115,
- A l'intégration, fichier interne XML norme DTD interne construit à l'aide d'un traducteur XSLT,
- Au stockage d'un fichier XML,
- A l'interrogation, construction d'un fichier XML de requête et prise en compte de ce fichier pour écrire et lancer la requête,
- A la restitution, fichier XML dont certaines balises sont filtrées et remplacées. Visualisation à l'aide d'une conversion XSLT vers un fichier HTML,

Les conversions XSLT sont écrites en XSL et en XPath pour pouvoir faire des tests sur les informations.

Base de données : Oracle 9i

La métadonnée et son fichier interne sont stockés au format XMLTYPE d'Oracle, c'est-à-dire tel quel (format XML). Ce format est un format CLOB sur lequel Oracle fournit des fonctions d'interrogation et d'extraction de balise. Oracle a donc choisi d'ajouter une sur-couche à sa base relationnelle pour pouvoir intégrer du XML.
Seule la version 9i d'Oracle permet ce type d'interrogation.
La base de données est attaquée via des procédures PL/SQL. Les requêtes sont relativement lourdes à écrire parce qu'il faut spécifier pour chaque balise concernée son chemin complet.
Afin d'accélérer les temps d'accès, il est possible d'indexer les balises des fichiers XML

Conclusion

La faisabilité du Bureau des Métadonnées est acquise. Le CNES va procéder à une phase d'expérimentation qui devrait permettre d'évaluer l'utilisation du site : génération des métadonnées, implication des scientifiques, aussi bien en amont pour la fourniture des métadonnées qu'en aval pour l'utilisation du site, pertinence des profils, …
Après cette phase, le CNES devrait passer un nouveau marché de réalisation d'un site beaucoup plus complet (Plusieurs normes en entrée, profils plus complets et plus ciblés, …) qui prend en compte le retour d'expérience.
Sur l'aspect technique, il apparaît que les techniques évoluent très rapidement. Certaines convergent (Z3950 et SOAP), de nouvelles apparaissent… Lors de cette prochaine réalisation, une phase d'évaluation devra être réalisée afin de cerner les techniques les plus appropriées du moment.