Réalisation d'un annuaire de sources de données génomiques en vue de la collecte et de l'intégration de données sur le Web

Shazia Osman

Université Bordeaux I

Université Victor Segalen, Bordeaux II

Diplôme préparé : **Master Professionnel Sciences et Techniques**
mention Informatique, spécialité Bio-informatique.

Promotion 2003-2004

Mémoire de stage de :

Shazia OSMAN

soutenu le 24 septembre 2004

Réalisation d'un annuaire de sources de données génomiques en vue de la collecte et de l'intégration de données sur le Web

Stage effectué au :

Laboratoire Lorrain de Recherche en Informatique et ses Applications
(LORIA)
UMR 7503
Campus Scientifique BP 234
Vandoeuvre-lès-Nancy
54600 Cedex, France.

Responsables :
Dr. Marie-Dominique DEVIGNES
Dr. Malika SMAIL-TABBONE

Tutrice :
Dr. Marie BEURTON-AIMAR

Remerciements

Je tiens à remercier :

- Marie-Dominique DEVIGNES et Malika SMAIL de m'avoir accueillie au sein de leur équipe, de leur aide, de leur patience et de leurs précieux conseils qui ont permis le bon déroulement de ce stage. Ce fut un grand plaisir pour moi de travailler avec elles.
- Marie AIMAR d'avoir acceptée d'être ma tutrice de stage, de s'être déplacée jusqu'à Nancy afin d'avoir un suivi de mon travail ainsi que de ses bons conseils.
- Sylvain TENIER, thésard au LORIA, pour ses conseils en XML.
- Hervé de PALMA, Mohamed EL YAMANY et Nizar MESSAI pour les bons moments de rigolades lors des pauses café et pour leurs aides durant ce stage.
- Tous les autres membres du LORIA que j'ai cotoyé, notamment, Stéphanie BILLAUT-BONNE, Damien EVEILLARD et Philippe SEBIRE pour leur amitié.
- Olivier TOQUE, Nathalie et Christophe BUISSON pour leur aide du point de vue rédactionnel.
- Mes amis du master qui ont rendu cette année très agréable.
- Enfin, je tiens à remercier en dernier mais pas le moindre, mes parents, ma sœur et mon frère de m'avoir toujours soutenu et encouragé lors de mes études. En guise de remerciement, je leur dédie ce travail…

Abstract

Genomics as a field has matured at an astonishing pace over the last decade. Moreover, the emergence of high throughput biological technologies in the post genomic era has given rise to a huge amount of complex and heterogenous genomic data. As a result, numerous Web data sources have been created to gather this information but unfortunately, these sources are themselves heterogenous and dispersed. Consequently, biologists today still remain confronted with the problem of rapidly finding the relevant information from such diverse sources to improve and accelerate their search. Hence, bio-informatic tools that would assist the biologist in his quest have to be developed.

The 'BioRegistry' project discussed in this report aims at addressing this issue with the objective of creating a directory of genomic data sources. Consequently, a model needs to be developed to enable the construction of such a directory.
To do so, relevant meta-data related to genomic sources were identified and collected in the first instance. These were then organised and classified according to domain ontologies to deal with the semantic dimension of the information before being incorporated into the directory.
Once developed, the directory was exploited in two ways : firstly, to allow the visualisation of its contents and secondly, to allow the extraction of relevant, source-related information. This last step will eventually be used for the classification and the querying of the 'BioRegistry'.

Keywords: biological data sources, meta-data, ontologies, Web directory.

Résumé

La génomique est un domaine qui a mûri à une vitesse étonnante durant cette dernière décennie. L'émergence de technologies biologiques à 'haut débit' dans l'ère post-génomique a donné lieu à de vastes quantités de données de natures complexes et hétérogènes. En conséquence, des sources de données Web ont été créées pour rassembler ces informations. Cependant, ces sources sont elles-mêmes hétérogènes et dispersées. De ce fait, les biologistes se retrouvent toujours confrontés avec le problème de recherche d'informations afin d'améliorer et d'accélérer leur recherche. Le développement d'outils bio-informatiques est donc nécessaire pour assister le biologiste dans sa recherche.

Le projet 'BioRegistry', présenté dans ce rapport, vise à traiter ce problème dans le but de créer un annuaire de sources de données génomiques. En conséquence, un modèle doit être créé pour permettre la construction de ce type d'annuaire.
Dans un premier temps, des méta-données pertinentes associées aux sources génomiques ont été identifiées et collectées. Celles-ci ont ensuite été organisées et classifiées selon des ontologies de domaine afin de s'occuper de la dimension sémantique de l'information, avant d'être injectées dans l'annuaire. Une fois ce dernier achevé, l'annuaire créé a été exploité de deux façons : d'une part pour visualiser ses contenus et d'autre part, pour extraire les informations pertinentes associées aux sources. Cette dernière étape sera éventuellement utilisée pour permettre la classification et l'interrogation de 'BioRegistry'.

Mots clés : sources de données biologiques, méta-données, ontologies, annuaire Web.

TABLE DES MATIÈRES

TABLEAU DES ANNEXES

TABLEAU DES FIGURES

Introduction

Ce stage est effectué dans le cadre du Master de Bio-informatique, au sein de l'équipe 'Orpailleur' du Laboratoire Lorrain de Recherche en Informatique et ses Applications (L.O.R.I.A) situé à Vandoeuvre-lès-Nancy, France. Le LORIA est une unité mixte de recherche (UMR 7503) qui a pour principal objectif de développer des recherches fondamentales et appliquées en informatique.
Proposé par Marie-Dominique DEVIGNES (CR1[1], CNRS[2] Sciences de la Vie) et Malika SMAIL (MC[3] Informatique, Université de Nancy I), le sujet de ce stage constitue une première mise en œuvre sujette à confirmation par d'autres travaux à venir.

Actuellement, le volume de données génomiques sur le Web est en croissance exponentielle et un accès intelligent à ces données devient crucial. L'intérêt de tous les acteurs de la recherche fondamentale ou appliquée est de savoir exploiter au mieux toutes ces sources biologiques qui ont pour caractéristiques d'être à la fois volumineuses, complexes, hétérogènes, versatiles et dispersées. Cependant, du fait de la croissance et de l'évolution très rapide de ces informations, il devient de plus en plus difficile pour le biologiste de savoir où chercher l'information souhaitée tout en évitant la redondance ou l'insuffisance des données collectées, ainsi qu'en faisant face à la diversité des formats, des résultats et de leur évolutivité. Des procédures automatiques deviennent nécessaires afin d'aider le biologiste dans cette tâche.
Il se fait donc une nécessité de regrouper et d'organiser les informations concernant ces sources. Ainsi, une structure en annuaire offrirait un moyen de choisir plus aisément les sources à interroger selon le type d'information que le biologiste souhaite collecter.

L'objectif de ce travail est la réalisation d'une architecture à base d'annuaire (projet 'BioRegistry') qui s'appuierait sur des méta-données pour constituer une sorte de carte d'identité de chaque source.
Les fiches descriptives des sources pourraient être organisées de façon à permettre une navigation et/ou une interrogation selon divers points de vues ou ontologies du domaine. En effet, la mise en forme et la classification de méta-données selon des vocabulaires contrôlés ou des taxonomies conduiraient à un réel gain en efficacité et en qualité dans le choix des sources et l'interprétation des données recueillies en réponse à une question donnée.

Ce rapport décrira l'analyse et la réflexion menées pour la conception de l'annuaire et pour une première mise en oeuvre. Le premier chapitre décrira l'environnement de travail dans lequel s'est déroulé le stage. Nous présenterons une vue générale du laboratoire et de son équipe d'accueil ainsi que le contexte du stage et ses principaux objectifs.
Dans le second chapitre seront présentés un état de l'art sur les données biologiques et sur les sources de données elles-mêmes ainsi que les solutions existantes d'accès aux sources. Puis nous traiterons des méta-données et de leurs rôles ainsi que des ontologies, des vocabulaires contrôlés et des taxonomies.
Dans un troisième chapitre, nous aborderons une analyse du problème ainsi que sa conception.
Quant au quatrième chapitre, il traitera de la réalisation et des résultats obtenus.
Pour terminer, nous concluerons en montrant les perspectives découlant de cette étude.

[1] Chargé de Recherches
[2] Centre National de Recherche Scientifique
[3] Maître de Conférences

Chapitre 1 : Environnement de travail et contexte du stage

1.1. Présentation du LORIA

1.1.1. Historique

Le LORIA est une UMR commune aux organismes suivants :

- l'I.N.R.I.A (Institut National de Recherche en Informatique et Automatique)
- Le C.N.R.S (Centre National de Recherche Scientifique)
- l'I.N.P.L (Institut National Polytechnique de Lorraine)
- l'Université Henri Poincaré, Nancy I
- l'Université Nancy II.

La création de cette unité a été officialisée en décembre 1997 par la signature d'un contrat avec le Ministère de l'Education Nationale, de la Recherche et de la Technologie et par une convention entre les cinq partenaires. Le LORIA succède ainsi au Centre de Recherche en Informatique de Nancy (C.R.I.N), avec des équipes communes au CRIN et à l'Unité de Recherche INRIA de Lorraine.

1.1.2. Activités

Dirigé depuis janvier 2001 par Hélène KIRCHNER, le laboratoire regroupe plus de 300 personnes dont des chercheurs, doctorants, thésards, ingénieurs, techniciens et administratifs. Le LORIA est composé de 25 équipes de recherche ayant des activités axées autour de cinq thématiques principales 'transversales':

- Calculs, réseaux et visualisation à hautes performances
- Télé-opérations et assistants intelligents
- Ingénierie des langues, du document et de l'information scientifique et technique
- Qualité et sûreté des logiciels et systèmes informatiques
- Bio-informatique et applications à la génomique

1.1.3. L'équipe 'Orpailleur'

Les thèmes de recherche dans l'équipe Orpailleur, sous la direction du Dr. Amedeo NAPOLI, portent principalement sur les processus d'extraction de connaissances (fouille de données ou 'data mining') dans les bases de données, et la mise en œuvre de structures pour représenter les connaissances extraites.
L'extraction de connaissances consiste à faire émerger à partir de larges volumes de données des éléments d'information susceptibles de devenir des éléments de connaissances exploitables dans un système intelligent. Elle s'appuie sur l'application de méthodes de fouilles de données (ex : méthodes de classifications par des treillis ou par des modèles de Markov cachés d'ordre 1 et 2, l'extraction de motifs fréquents et de règles d'association) mais aussi sur l'exploitation des connaissances du domaine des données pour guider la fouille de données.

Pour ce qui est de la représentation et de la manipulation des connaissances, il existe divers formalismes de représentation et de modes de raisonnement, parmi lesquels se distinguent les représentations de connaissances par objets et les logiques de descriptions, avec le raisonnement par classification et le raisonnement à partir de cas.

Les données textuelles sur le Web constituent un des matériaux privilégiés pour la fouille de textes, mais aussi pour la mise en œuvre du 'Web sémantique'. Le Web sémantique est une extension du Web actuel visant à rajouter de la signification à l'information de telle sorte qu'elle devient automatiquement interprétable par des machines [BL01]. Il constitue un terrain d'investigation de première importance pour Orpailleur. Les domaines d'applications d'importance pour Orpailleur

appartiennent à la biologie (fouille et analyse de séquences génomiques), la médecine, la synthèse en chimie organique, l'analyse de textes scientifiques, la classification de signaux temporels et la sidérurgie.

Enfin, pour la petite histoire, un orpailleur est un artisan qui recueille par lavage, à travers un tamis, les paillettes d'or dans les fleuves et les terres aurifères. Ici, l'or c'est la connaissance, et c'est cette connaissance que l'équipe essaie d'extraire de différentes façons, à partir de diverses sources de données, pour l'intégrer dans des systèmes, lesquels seront dotés d' 'intelligence'…

1.1.4. Web sémantique et bio-informatique chez Orpailleur

Un domaine dans lequel un accès intelligent aux données du Web devient crucial est celui de la génomique. Des travaux antérieurs (Xmap) [DSS02] ont porté sur la conception de systèmes orientés utilisateur pour la collecte et l'intégration de données biologiques. Une solution générique (Xcollect)[4] [DS04[5]] a été proposée qui permet à partir d'un scénario élaboré par l'utilisateur et représenté selon un modèle générique, d'exécuter de façon automatique pour chaque étape du scénario les actions suivantes : formulation de la requête, soumission à la source choisie, extraction des données utiles à partir du document de réponse, stockage des informations sous forme structurée dans un document de session. L'automatisation du processus permet en particulier à l'utilisateur de rejouer le scénario ultérieurement afin de tenir compte d'éventuelles mises à jour dans le contenu des sources interrogées. Du fait de l'apparition fréquente de nouvelles sources et des modifications également fréquentes des sources disponibles, ce travail a conduit à une réflexion sur le problème de l'identification et de la caractérisation des sources pertinentes pour une question donnée.

1.2. Le projet 'BioRegistry' et les objectifs du stage

1.2.1. Le projet 'BioRegistry' dans son ensemble

Le projet 'BioRegistry' vise à élaborer une structure d'annuaire permettant de découvrir ou d'identifier les ressources du web pertinentes pour une question biologique donnée et de composer l'accès à ces ressources sous la forme de scénarios (ou workflow). En plus de l'exécution proprement dite du scénario, le système devra contribuer à la synthèse et à l'intégration des réponses à la question posée. La dimension sémantique sera particulièrement étudiée : le rôle des ontologies et l'utilisation de formalismes permettant le raisonnement seront explorés.

Concrètement, ce projet repose sur des collaborations avec des équipes de biologistes impliqués dans l'identification de gènes candidats pour certaines pathologies multifactorielles, dans la localisation de régions promotrices des gènes d'intérêt pour le suivi des cancers et dans l'identification de gènes responsables de maladies génétiques rares.

1.2.2. Objectifs du stage

Le travail qui m'a été confiée concerne la réalisation d'un modèle d'annuaire des sources de données biologiques pouvant servir de base au projet 'BioRegistry'. Les objectifs étaient les suivants :
a) explorer et identifier les informations (méta-données) pertinentes qui devront être associées aux sources dans l'annuaire.
b) explorer les ontologies utilisables pour organiser et classifier ces méta-données.
c) modéliser et construire l'annuaire.
d) visualiser les informations contenues dans l'annuaire.
e) extraire de l'annuaire les informations nécessaires pour classifier et interroger l'annuaire, en lien avec un travail effectué en parallèle par un étudiant de DEA de l'équipe [MES04].

[4] http://www.loria.fr/~devignes/DOC_XcollectProject.htm
[5] http://www.iscb.org/ismbeccb2004/short%20papers/66.pdf

Chapitre 2 : Etat de l'art

2.1. Caractéristiques des données biologiques

Il importe de bien préciser ce que l'on entend par données biologiques. Prenons pour exemple la banque protéique, Swiss-Prot[6] (voir annexe 1). La donnée de la séquence protéique en elle-même ne forme qu'une petite partie de l'entrée. Par contre, les annotations de la séquence qui décrivent la protéine ainsi que des commentaires sur la séquence entière (ex : fonction, maladies, espèce…) constituent la plus grande partie d'une entrée. Stevens ***et al.*** [SWLG04] considèrent que toutes ces informations qui sont aussi des données, peuvent être vues comme le composant connaissance de la base de données. Dans la source Swiss-Prot, ces connaissances sont exprimées sous forme de texte, qui décrivent les données le plus souvent par du texte libre, parfois par des mots-clés et vocabulaires contrôlés et très rarement par des expressions numériques. Cette forme de représentation de données est adaptée pour l'humain mais elle rend difficile le traitement par la machine.

Les données biologiques ont pour caractéristique d'être [SWLG04] :

- *Volumineuses* – Les projets de séquençage de divers génomes, les techniques de micro-arrays et la biologie haut-débit en général, font que les données sont produites en quantités et à des vitesses gigantesques.
- *Complexes* - Il est difficile de représenter directement les entités biologiques sous forme numérique. Par ailleurs, ces entités ont la particularité d'avoir de multiples relations, ce qui rend leur représentation encore plus complexe. Par exemple, une protéine possède à la fois une séquence, une fonction, une structure 3D, agit dans un ou plusieurs processus…, peut être impliquée dans une maladie…
- *Versatiles* – Les connaissances sur les entités biologiques changent et par conséquent les annotations au sein d'une source de données changent aussi.
- *Hétérogènes* – les données biologiques sont à la fois syntaxiquement et sémantiquement hétérogènes. Par exemple, le concept 'gène' a différentes interprétations qui sont toutes valables. De fait, la synonymie et l'homonymie pour les labels utilisés en biologie sont sources de beaucoup de controverses [SWLG04].
- *Disséminées* – À travers tout le Web, on retrouve plus de 500 sources de données et de nombreux outils d'analyse bio-informatiques. [GAL04, BioCat[7]]

2.2. Les sources de données biologiques

L'expression 'base de données' et celle de 'banque de données' sont employées de façon indifférente en biologie, mais il est utile de faire ressortir leurs différences du point de vue architecture : une **base de données** est un ensemble structuré logique d'informations qui bannit théoriquement toute redondance; elle est normalement pilotée par un système de gestion de base de données (SGBD) d'où sa robustesse, ex : les bases relationnelles telles que Ensembl. Une **banque de données** est un ensemble d'informations semi-structurées et généralement regroupées sous forme de fichiers plats (fichiers texte) ex : GenBank. Ainsi, le terme **source de données** sera plutôt employé pour représenter ces deux termes lorsque leur distinction n'est pas nécessaire.

[6] http://www.expasy.org/sprot/
[7] http://www.ebi.ac.uk/biocat/

2.2.1. Historique

Une des premières banques de données biologiques était fort probablement le livre de Margaret Dayhoff, '*Atlas of Protein Sequences and Structures*'. Cet ouvrage publié pour la première fois en 1965 et édité annuellement, contenait toutes les séquences protéiques connues. Les données de ce livre devinrent par la suite, la fondation pour la source 'Protein Information Ressource' (PIR) de l'Université de Georgetown, USA.
Au fur et à mesure de la croissance des données, la biologie devint une science riche en données d'où la nécessité de stocker et échanger de larges ensembles de ces données. Ainsi, plusieurs organismes ont mis en place des structures de collecte et de gestion de ces données. Le NCBI au NIH[8] aux Etats-Unis, l'EMBO[9] en Europe ont participé à des projets afin de développer des banques de séquences nucléiques regroupant des données issues du monde entier. Ainsi sont nées GenBank[10] en 1979 et l'EMBL[11] en 1980. Une collaboration a fini par s'établir entre ces deux banques, renforcée en 1987 par la DDBJ[12] au Japon. Parallèlement, se sont mises en place des banques de séquences protéiques comme PIR et Swiss-Prot.

De par l'explosion de la quantité et de l'hétérogénéité de données dans les sources généralistes, des sources spécialisées sont apparues. Ces dernières sont construites autour d'une thématique biologique particulière, ou autour des séquences d'un organisme spécifique afin de compléter et corriger les annotations trouvées dans les banques généralistes.

Ainsi, le nombre de bases ou banques de données disponibles actuellement sur Internet est impressionnant. La publication '*The molecular biology database collection : 2004 update*' [GAL04] qui recense plus de 500 références à des sources de données en biologie est là pour en témoigner.

2.2.2. Analyse du contenu des sources

2.2.2.1. Les banques généralistes

Une banque généraliste est, en fait, une bibliothèque de fiches descriptives de séquences nucléiques ou protéiques. Ces dernières proviennent de n'importe quel organisme et peuvent être de n'importe quelle nature : ADN, ADNc, ARN, protéine. Ces fiches contiennent des informations variées sous forme de commentaires structurés issus d'expertises biologiques ou d'analyses bio-informatiques. On peut sub-diviser ces banques généralistes en deux catégories : les banques de séquences nucléotidiques et les banques de séquences protéiques.

a. Les banques de séquences nucléotidiques

Les trois principales banques nucléotidiques, GenBank, EMBL et DDBJ, coexistent et coopèrent. En effet, ces trois banques échangent et mettent à jour leurs données quotidiennement. Elles collectent des informations de séquences principalement par soumission directe des auteurs (≈ 95 %) mais aussi à partir de la littérature scientifique.
Même si les données dans ces banques ne sont pas représentées de la même façon selon les organisations qui les maintiennent, la philosophie est identique en ce qui concerne la structuration des caractéristiques biologiques. En effet, chaque entrée ou enregistrement correspond à une séquence nucléotidique. Cet enregistrement peut être aisément divisé en quatre parties (voir annexe 2) :

---

[8] National Institute of Health
[9] European Molecular Biology Organisation
[10] http://www.ncbi.nlm.nih.gov/Entrez/
[11] European Molecular Biology Laboratory : http://www.ebi.ac.uk/embl/
[12] DNA Data Bank of Japan : http://www.ddbj.nig.ac.jp/

- L'en-tête : il contient les informations d'ordre général sur la séquence (identifiant, numéro d'accession, définition, mot-clé, taxonomie)
- Les informations de bibliographie associées à la séquence se composant du numéro de la référence bibliographique dans la banque Medline, des auteurs de l'article, du titre de l'article, et du journal dans lequel celui-ci est publié.
- Les caractéristiques ou 'features' correspondant aux zones d'intérêts de la séquence déterminées lors de l'annotation syntaxique. Ils concernent par exemple les séquences codantes et leur traduction, les zones promotrices, les introns et les exons. La collaboration entre ces banques a permis la définition d'un guide d'annotation pour la description des caractéristiques biologiques des séquences nucléiques (voir 'feature table'[13] EMBL/GenBank/DDBJ).
- La séquence proprement dite.

b. Les banques de données protéiques

Les banques protéiques sont organisées sensiblement de la même manière. Comme dans les banques nucléotidiques, chaque entrée sur une protéine contient la séquence elle-même, ainsi que des informations biologiques et biochimiques sur celle-ci.

PIR-PSD[14] International : PIR, en association avec le MIPS[15] et le JIPID[16] distribuent une collection de séquences qui se veut exhaustive et non redondante, et organisée selon des critères taxonomiques et d'homologies. Les données sont issues de la littérature, des soumissions directes, et de la traduction des séquences nucléiques issues de banques étudiées ci-dessus (EMBL, GenBank, DDBJ). Les séquences homologues ou très similaires sont regroupées au sein d'une même entrée. La séquence canonique permet de retrouver les séquences initiales qui ont mené à cette description. Des familles de protéines sont construites par similarité de séquence mais également de fonction.

Swiss Prot : C'est sans conteste la banque de séquences protéiques la moins redondante et le mieux annotée disponible actuellement. Cette banque créée en 1986 à Genève est maintenant le fruit d'une collaboration entre l'EMBL et l'Institut Suisse de Bio-informatique. Le format des entrées est sensiblement le même que celui de l'EMBL. Cette banque se distingue surtout par sa qualité d'annotation (voir annexe 1). Afin que ces annotations soient le plus juste possibles, elles sont rédigées à partir des articles de publication des séquences mais également d'articles de revue. De plus, un réseau de spécialistes volontaires annote et corrige les séquences qui ressortent de leur domaine d'expertise.

Afin que les séquences apportées par les projets génomes soient mises à la disposition du public sans pour autant nuire à la qualité de l'annotation dans Swiss-Prot, la banque TrEMBL[17] a été mise en place. TrEMBL et GenPept[18] contiennent les séquences protéiques conceptuelles obtenues par traduction automatique des séquences dites codantes (CDS[19]) dans l'EMBL et GenBank. TrEMBL, distribuée par l'EBI[20], contient la traduction des parties codantes des séquences nucléotidiques stockées dans EMBL à l'exception de celles déjà présentes dans Swiss-Prot. De la même façon, GenPept est la version protéique de GenBank. Il est important de noter que ces deux banques contiennent des séquences non vérifiées, dont les annotations découlent de celles des banques nucléotidiques. En effet, de plus en plus de

[13] http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html
[14] Protein Information Resource-Protein Sequence Database : http://pir.georgetown.edu/home.shtml
[15] Munich Information Center for Protein Sequences : http://mips.gsf.de/
[16] Japan International Protein Information Database
[17] Translated EMBL : http://www.expasy.org/sprot/
[18] http://bip.weizmann.ac.il/databanks/genpept.html
[19] Coding Sequences
[20] European Bioniformatics Institute

séquences codantes sont déterminées par des programmes de prédiction qui ne sont malheureusement pas sans faille. Ceux-ci conduisent à des erreurs de prédiction des parties codantes des séquences, donc à des erreurs de séquences protéiques que l'on retrouve dans ces deux banques.

2.2.3. Les banques de données spécialisées

Les banques généralistes présentent des avantages de par leur exhaustivité et malheureusement des limites de par leurs trop fréquentes imprécisions. Les banques de données spécialisées sont un moyen de résoudre ces problèmes. En effet, regrouper au sein d'une même structure un sous-ensemble de données, dans un domaine spécifique ou sur un organisme donné, semble être le meilleur moyen pour répondre aux points faibles évoqués ci-dessus, surtout si cela est fait par des experts du domaine considéré.

2.2.3.1. Les banques thématiques

Ces banques (ex : FlyBase[21]) peuvent réunir au sein d'une même structure des séquences nucléotidiques ou protéiques sélectionnées selon un critère précis (structure moléculaire, appartenance à un organisme modèle, présence d'un motif…) mais il existe également des banques (ex : KEGG[22]) qui abordent des aspects de la biologie moléculaire non directement liés aux séquences (métabolisme, réseaux de régulations, données d'expression…).

2.2.3.2. Les banques comparatives

Le développement de la génomique comparative a conduit à la création de sources de données relatives à la classification des protéines ou des gènes protéiques à travers les espèces (ex : Génolevures[23], Pfam[24]).

2.3. Solutions existantes d'accès aux sources

Les **moteurs de recherche** comme 'Google[25]' et 'AltaVista[26]' sont certes très utiles lorsqu'il s'agit de rechercher une information biologique par exemple, sur le Web. Cependant, le bruit associé aux résultats renvoyés suite à une requête peut être faramineux. Ceci vient du fait que ces moteurs indexent des documents par la fréquence des mots qu'ils contiennent sans tenir compte de la structure ou des fonctionalités associées à ces documents. Si, par exemple, on recherche l'interface d'interrogation de 'FlyBase' sans connaître 'FlyBase' proprement dite, et si par exemple, les mots '*chromosome aberrations*', '*drosophila*', '*nucleic sequence*' *'database'* sont rentrés dans 'Google', les résultats retournés font le lien avec des portails qui permettent d'accéder à la source 'FlyBase', et avec d'autres documents traitant cette source de données. L'utilisateur n'est pas directement mené à la source elle-même (requête effectuée le 07/09/2004).
Pour trouver une source sur le Web, l'utilisateur s'adresse plutôt à des **portails**. Les portails sont des documents Web, présentant une ou plusieurs listes de sources biologiques rendues accessibles grâce à des liens. L'utilisateur choisit donc la source qu'il souhaite interroger et est redirigé vers celle-ci pour

[21] http://flybase.bio.indiana.edu/
[22] http://www.genome.jp/kegg/
[23] http://cbi.labri.fr/Genolevures/
[24] http://www.sanger.ac.uk/Software/Pfam/
[25] www.google.com
[26] http://www.altavista.com/

l'interroger directement. Quelques exemples de ces portails sont le serveur ExPASy[27] du SIB[28] et le Deambulum[29] d'Infobiogen.

ExPASy ('Expert Protein Analysis System') est un portail orienté protéomique qui permet d'accéder à différents types d'informations uniquement sur les protéines.
Le Deambulum est un portail de la biologie et de la bio-informatique. Il permet de réaliser une exploration thématique des différents domaines de la biologie : la biologie moléculaire, la bio-informatique, le génome humain et les organismes, les données bibliographiques…
Deambulum est soumis à une catégorisation beaucoup plus extensive qu'ExPASy. Les banques de données regroupées dans le Deambulum sont classées par thème et chaque thème est soumis à une hiérarchisation. De plus, les deux portails offrent l'accès à des outils et des logiciels pour l'analyse des séquences (protéiques dans le cas d'ExPASy), ex : l'alignement de séquences, la recherche de motifs, la phylogénie… Par ailleurs, les deux portails offrent l'accès à diverses informations dans les domaines de la biologie en général et de la protéomique, tels que, les actualités scientifiques, les services, l'éducation…
Cependant, les informations fournies par ces portails (et par les portails en général) sur les sources sont très variables et ne permettent pas toujours à l'utilisateur de choisir entre les différentes sources qui semblent répondre à son besoin. De plus, les portails renvoient aux sources et ne permettent pas de les interroger directement. C'est à cette dernière limite que répond le logiciel **SRS**[30].
Ce logiciel, très connu des biologistes, permet à partir d'une interface unique d'accéder à des sources de données biologiques et de réaliser des interrogations croisées sur plusieurs sources à la fois, grâce à un langage de requête commun. L'interface va alors diriger les requêtes vers les différentes sources indexées par le système et reliées entre elles par des liens également indexés. Cependant, l'utilisateur doit avoir une connaissance des sources qu'il souhaite interroger car la sélection de celles-ci se fait avant d'établir la requête sur la base d'une simple catégorisation des sources interrogeables.

Aucune des solutions existantes ne permet donc de résoudre de façon satisfaisante le problème de la sélection d'une ou de plusieurs sources de données pertinentes par rapport à une question posée pouvant comporter plusieurs critères. L'analyse du problème a conduit l'équipe qui m'accueillait pour ce stage à proposer une solution d'annuaire de sources de données biologiques. Cette proposition est basée sur le recensement des méta-données connues sur les sources et sur leur organisation en structures fortement guidées par des ontologies. La structuration des méta-données devrait permettre la navigation selon divers critères tels que le contenu des sources (si ce sont des séquences nucléiques ou protéiques…) et la qualité des sources (la date de dernière mise à jour, l'existence d'une révision manuelle ou pas…)

2.4. Les méta-données et leurs rôles

Les informations que décrivent un ensemble de données sont connues comme les **méta-données**. Ce ne sont pas les données elles-mêmes mais plutôt des données, à un niveau d'abstraction supérieur, à propos de données d'un niveau inférieur. Le but est que l'éventuel utilisateur puisse accéder à un ensemble de données sans avoir à accéder et à étudier les données elles-mêmes.
Le 'Dublin Core' est un standard de méta-données en recherche documentaire.

2.4.1. Le Dublin Core (DC)

Le standard de méta-données du Dublin Core (voir annexe 3) est un ensemble d'éléments permettant de décrire des ressources en ligne. Le standard du DC comprend 15 éléments dont la sémantique a été établie par un consensus international provenant de diverses disciplines. Chaque élément est optionnel

[27] Expert Protein Analysis System : www.expasy.org/
[28] Swiss Institute of Bioinformatics
[29] http://www.infobiogen.fr/services/deambulum/fr/
[30] Sequence Retrieval System : srs.ebi.ac.uk/

et peut être répété (ex : l'élément 'rights' ou droits peut ne pas exister pour une source mais il peut aussi exister plusieurs fois, indiquant qu'il existe plusieurs droits relatifs à une source). Le Dublin Core est un standard de méta-données parmi plusieurs. Il est à noter qu'il ne décrit pas la façon de représenter les méta-données.

2.4.2. Les rôles des méta-données

Les rôles des méta-données sont multiples [SWLG04, RRSW04]:

- fournir un moyen de découvrir qu'un ensemble de données existe et comment cet ensemble pourrait être obtenu ou comment y accéder.
- rôle naturel d'aide à la structuration et à la recherche 'intelligente' d'informations sur le Web
- faciliter l'échange de données ainsi que la médiation et l'unification de sources de données.
- permettre de croiser des domaines ex : récupérer pour un domaine, la structuration d'un concept lié à un autre domaine
- permettre de documenter la qualité et les caractéristiques du contenu d'un ensemble de données et ainsi donner une indication de son aptitude à être utilisé.
- qualité et traçabilité des informations : croisement des données liées aux résultats obtenus

Les méta-données donnent un sens structurel et cognitif à l'information (du Web par exemple) et elles sont appliquées au catalogage, à la recherche et à l'indexation de ressources… Un annuaire basé sur des méta-données décrirait de façon compréhensive des ensembles de données et fournirait simultanément un moyen de les découvrir.

2.5. Les ontologies, vocabulaires contrôlés et les taxonomies

2.5.1. Les ontologies proprement dites

Le mot ontologie vient du grec '*Ontos*' qui signifie 'l'être'. L'ontologie signifie donc « étude de l'être, des types de choses qui existent ». Le concept d'ontologie a pour origine la philosophie et plus précisément la théologie, les discours concernant Dieu.
Le terme 'ontologie' recouvre toute proposition visant à définir dans un formalisme utilisable par des programmes, un ensemble de concepts propres à un domaine ainsi que les relations qui existent entre ces concepts[GRU93]. Un corpus de connaissances représenté formellement est basé sur une conceptualisation : les objets, concepts et autres entités qui sont présumés existés dans un certain domaine d'intérêt (un domaine étant une partie d'un sujet ou en ensemble de sujets) et les relations qui existent entre eux. Une conceptualisation est donc une vue simplifiée du monde que nous voulons représenter pour dans certain but, par exemple, le partage des connaissances. Une ontologie est donc une spécification explicite d'une conceptualisation [GSM01].
En théorie, les ontologies sont formalisées, réutilisables, et non instantiables. Celles qui sont réutilisables doivent être de haut niveau, en d'autres termes, les concepts proposés ne doivent pas être trop spécialisés [CHA04].
En pratique, la définition d'une ontologie est controversée. Dans la littérature existante, les ontologies sont définies de diverses façons, avec des degrés de formalisation différents [GSM01] (voir ci dessous 2.5.2).

2.5.2. Quelques définitions

Le mot "**ontologie**" est utilisé de parfois de façon abusive pour parler de vocabulaires contrôlés, taxonomies et de connaissances. Goble *et al.* [GSM01] ont tenté de donner une définition aux termes suivants :

Un **vocabulaire contrôlé** est une ontologie qui est simplement une liste de termes ou d'expressions qui sont utilisés dans un but précis. La signification des termes n'est pas forcément définie et il n'y a pas nécessairement d'organisation logique des termes entre eux. Les vocabulaires contrôlés sont complétés par un **thésaurus**. Celui-ci correspond à une liste de synonymes pour un terme.

Une **taxonomie** est une série de termes organisés en classifications hiérarchiques et qui spécifient les relations généralisation-spécialisation entre les termes. Une taxonomie ne définit ni les attributs de ces termes, ni les autres relations qui peuvent exister entre eux. Le lien précis entre un terme du vocabulaire et ses enfants a une signification unique typiquement celle de la relation '*is a*'. Les relations '*is a*' définissent les types de concepts (ex : « *vertebrate* ***is a*** *organism* » le vertébré est un type d'organisme).

Une **ontologie** dans le sens le plus riche du terme, peut exprimer une grande variété de relations ainsi que la négation, la disjonction et des contraintes plus sophistiquées.

Les ontologies sont utilisées par l'Homme, les bases de données et les applications qui doivent partager une information de domaine. Elles offrent ainsi un moyen aux biologistes d'améliorer la représentation de connaissances contenus dans les sources. Parmi les bio-ontologies, on retrouve les ontologies 'Gene Ontology[31]' (GO), 'TAMBIS'[32] (TaO), le 'MeSH[33]', BAO/BACIIS[34], le réseau sémantique de l'UMLS, le 'Tree of Life' et la taxonomie de NCBI.

2.6. Les Bio-ontologies

2.6.1. L'ontologie TAMBIS : TaO

TAMBIS[35] [BAK99, BAK98] ('*Transparent Access to Multiple Biological Information Resources*') est un projet ayant pour but de fournir un seul point d'accès aux ressources biologiques accessibles sur le Web. Son objectif est de récupérer et de filtrer l'information de façon transparente à partir de ces ressources. Les requêtes sont traduites dans les termes de l'ontologie TaO ('TAMBIS Ontology') et le système les convertit alors pour accéder aux ressources correspondantes. TaO permet de guider l'utilisateur à formuler les requêtes de haut niveau plutôt que de les formuler directement dans le langage de chaque source.

Le but de l'ontologie TaO est de capturer les connaissances biologiques dans un langage qui permet l'interprétation des concepts et leurs relations par la machine.
TaO peut être divisée en deux parties :

- Une hiérarchie de concepts de **haut niveau** d'abstraction : niveau qui fait la distinction entre le domaine des concepts et le domaine des rôles.
  Le domaine de concepts concerne les concepts généraux (`structures, processus, substances` et `fonction`).
  Le domaine des rôles définit des relations génériques telle que la localisation.
- Une hiérarchie de **bas niveau** qui représente les connaissances utilisateur dans le domaine biologique. La description des concepts tels que `Protéines, Acides Nucléiques` y sont décrits ainsi que les concepts qui en découlent tels que `DNA, RNA` et `enzyme`.

On retrouve les relations '*is a kind of*', '*is homologous to*', '*has component*', '*is a component of*' et '*functions in process*' entre autres.

[31] http://www.geneontology.org/
[32] http://imgproj.cs.man.ac.uk/tambis/
[33] Medical Subject Headings : http://www.nlm.nih.gov/mesh/meshhome.html
[34] Biological and Chemical Ontology/ Biological And Chemical Information Integration System
[35] http://imgproj.cs.man.ac.uk/tambis/

2.6.2. Gene Ontology (GO)

L'objectif de GO[36] [GO04, ASH00] est de mettre en place un vocabulaire contrôlé et structuré afin de permettre la description de certains domaines de la biologie moléculaire et cellulaire de manière assez générique. GO est composée de trois ontologies différentes couvrant trois domaines de connaissances différents : les processus biologiques, les fonctions moléculaires et les localisations ou composants cellulaires. Le but est de décrire ces phénomènes de façon identique chez tous les organismes vivants. Même si les concepts dans GO ne cessent de croître, il reste cependant un bon nombre de domaines biologiques non couverts : structures tridimensionnelles, données d'expression, structure des domaines protéiques, évolution…

L'ontologie des 'fonctions moléculaires' décrit les activités individuelles des produits de gènes (ex : l'activité ATPase). Celle des 'processus biologiques' décrit les processus biologiques généraux (ex : le métabolisme des pyrimidines, la méiose) et l'ontologie des 'composants cellulaires' décrit les structures, localisations sub-cellulaires et les complexes macromoléculaires (ex : noyau, télomères).

L'ontologie GO est organisée en Graphes Orientés Acycliques (DAG) où un terme fils (terme plus spécialisé) peut avoir plusieurs parents (termes moins spécialisés). Chaque terme GO possède un identifiant unique et une définition. Par exemple, le terme '*DNA Binding'* (liaison à l'ADN) a pour numéro d'accession GO:0003677 et pour définition '*interacting selectively with DNA*' (interagissant de façon sélective avec l'ADN). Le parcours dans l'arbre du terme se compose ainsi :

GO:0003673 : Gene Ontology
GO:0003674 : molecular function
GO:0005488 : binding
GO:0003676 : nucleic acid binding
GO:0003677 : DNA binding

Les relations sont de types '*is a*' et '*part of*'.
GO est une ontologie utile pour l'annotation de produits de gènes mais trop spécialisée pour l'indexation de sources de données.

2.6.3. L'ontologie 'Biological and Chemical Ontology for Information Integration' (BAO/BACIIS)

BAO[37] [BEN02]est une ontologie de domaine pour BACIIS[38] qui est un système d'intégration d'informations biologiques et chimiques.
Le but de BAO est de :

- Guider l'utilisateur à faire des requêtes efficaces
- Faciliter la résolution des variabilités entre différents formats de données
- Faciliter l'intégration de base de données web biologiques et chimiques

Le système BACIIS permet un accès aux sources de données hétérogènes. De plus, BACIIS est basé sur la compréhension sémantique du domaine de connaissance. BAO est une ontologie à trois dimensions car elle définit les structures hiérarchiques pour trois classes : les objets, les propriétés et un réseau de relations. De ce fait, la relation entre deux objets se distingue de la propriété d'un objet. Ceci permet d'avoir un meilleur isolement des concepts.
Les relations sont de type '*is a subset of*', '*has property*', '*source of*' …

[36] http://www.geneontology.org/
[37] Biological And chemical Ontology
[38] http://baciis.engr.iupui.edu/

2.6.4. Réseau sémantique de l'UMLS du National Library of Medicine (Unified Medical Language System Semantic Network)

Le but de UMLS[39] [McC03] est de fournir un accès intégré à une vaste panoplie de ressources biomédicales tout en unifiant les vocabulaires qui sont utilisés pour accéder à ces ressources. Les concepts englobés touchent les domaines cliniques et les sciences de la vie.
Le NLM produit les « UMLS Knowledge Sources » (UMLSKS). Celles-ci comprennent un Métathésaurus®, un réseau sémantique ainsi qu'un lexique SPECIALIST.
Le Métathésaurus® contient des concepts biomédicaux ainsi que des concepts de noms provenant d'une centaine de vocabulaires contrôlés et de classifications utilisées dans des données sur des patients, des bases de données bibliographiques …

Le rôle du réseau sémantique[40] est de fournir un squelette de haut niveau où tous les concepts ont une représentation logique et sémantiquement cohérente. Ce réseau se compose actuellement de 134 types sémantiques et de 54 relations. Au plus haut niveau du réseau, on retrouve deux hiérarchies : une pour les entités et une autre pour les évènements. Chaque type sémantique est lié à son père par un lien de type '*is a*'. De plus, chaque type sémantique a une définition textuelle. Cette dernière s'avère utile pour assigner et interpréter les types sémantiques qui sont liés aux concepts du Métathésaurus®. Par exemple : le type sémantique 'Mammifères' est défini comme un vertébré ayant une température de corps constante et est caractérisé par la présence de poils, de glandes mammaires et de glandes de sueur. Avec chaque définition, on a des exemples d'instances du type retrouvées dans le Métathésaurus®, pour l'exemple choisi : 'bears'(ours), '*Rattus norvegicus*', 'whales' (baleines)…
Par ailleurs, ce réseau sémantique est d'avantage défini par des relations associatives qui forment une hiérarchie par elles-mêmes. Au plus haut niveau, on retrouve les relations associatives '*physically related to*', '*spatially related to*', '*functionally related to*' … Il est à noter que les descendants du même type héritent des relations associatives.

2.6.5. Medical Subject Headings (MeSH) du National Library of Medicine

Le MeSH[41] est un vocabulaire contrôlé utilisé pour indexer, cataloguer et rechercher des informations et des documents concernant le domaine bio-médical et la santé. La seule relation retrouvée est de type '*is a*'. Le MeSH peut être accédé de plusieurs façons en-ligne : le 'MeSH Browser' qui contient le contenu complet de tous les vocabulaires ; au niveau de la base de données 'Entrez' afin d'assister ceux qui font des recherches sur la base MEDLINE/PubMED et via le 'Metathesaurus®' de l'UMLS qui est le thésaurus de base de MEDLINE.
Le MeSH, mis à jour annuellement, contient environ 19 000 concepts et le vocabulaire contient plusieurs types de termes pour chaque concept:

- Les Descripteurs (les principaux titres) : caractérisent le sujet ou le contenu.
- Les Qualifiants : utilisés avec les descripteurs, ils offrent un moyen de regrouper les documents ayant une similarité avec un sujet.
- Les Termes d'entrées ('Entry Terms') : sont des synonymes ou des termes proches qui sont associés aux descripteurs.

Le MeSH est organisé en structure arborescente : Les descripteurs sont organisés en 15 catégories, ex : Catégorie A pour les termes de l'anatomie, catégorie B pour les organismes… A l'intérieur de chaque catégorie, les descripteurs sont rangés de façon hiérarchique du plus général au plus spécifique. Un descripteur apparaît à au moins un endroit dans les arbres. Au niveau des arbres, chaque descripteur est suivi du ou des numéros indiquant sa ou ses positions dans l'arbre, ex : le descripteur '*Metabolism*' = G06.535.

[39] Unified Medical Language System
[40] http://www.nlm.nih.gov/research/umls/META3.HTML
[41] http://www.nlm.nih.gov/mesh/

Un parcours possible dans l'arbre (version 2004) du descripteur '*Metabolism*' est :

Biological Sciences [G] +

Biochemical Phenomena, Metabolism, and nutrition [G06] +

Metabolism [G06.535] +

2.6.6. Taxonomie des espèces vivants

2.6.6.1. La Taxonomie du NCBI

La source 'NCBI Taxonomy'[42] est un ensemble de noms et de classifications révisés pour tous les organismes représentés dans la source GenBank. Lors de la soumission de nouvelles séquences à GenBank, si elles font mention de nouveaux noms d'organismes, ceux-ci sont alors classés et rajoutés à la source de données 'Taxonomy' [NCBIHb].

2.6.6.2. 'Tree of Life'

Le 'Tree of Life'[43] (ToL) ou l'arbre de la vie, est un document concernant la diversité des organismes vivants. Il contient des informations sur l'évolution et les caractéristiques des organismes, et présente l'arbre de l'évolution dans son ensemble.

[42] http://www.ncbi.nih.gov/Taxonomy/

[43] http://tolweb.org/tree/phylogeny.html

Chapitre 3 : Analyse du problème et conception

3.1. Démarche suivie

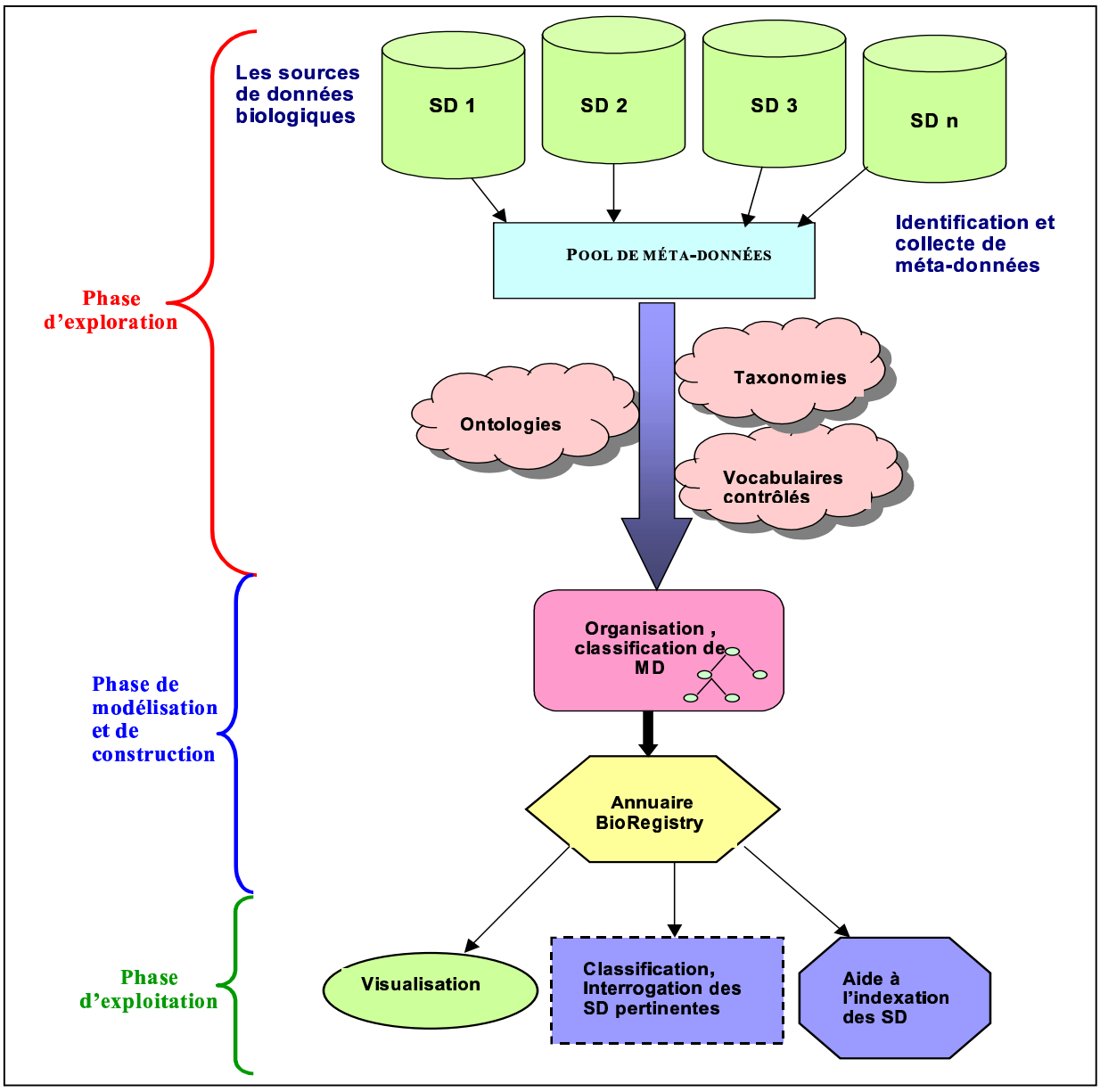


Figure 1 : Schéma général décrivant la démarche de construction suivie d'un annuaire pour les sources de données biologiques (SD = Sources de données, MD = Méta-données)

La **figure 1** schématise la démarche suivie pour la construction d'un annuaire. Cette démarche est constituée de trois phases : une phase d'exploration, une phase de modélisation et de construction, et une phase d'exploitation.

La **phase d'exploration** comporte deux volets :

- une exploration des sources de données où il s'agit dans un premier temps, d'explorer une série de sources, d'identifier les méta-données pertinentes et de les collecter. Ainsi, un pool de méta-données associées aux sources a été obtenu.

- une exploration des ontologies, des vocabulaires contrôlés et des taxonomies pour les associer à certaines méta-données afin de les organiser et de les classifier.

La **phase de modélisation et de construction** se compose de :
- la modélisation des méta-données en structure arborescente
- l'utilisation des méta-données organisées, classifiées et associées aux sources pour alimenter ou construire l'annuaire.

La **phase d'exploitation** est constituée de :
- la visualisation, sous format HTML, des méta-données associées aux sources.
- l'aide à l'indexation de certaines méta-données associées aux sources.
- l'extraction de l'annuaire des informations nécessaires pour classifier et interroger l'annuaire, en lien avec un travail effectué en parallèle par un étudiant de DEA de l'équipe [MES04].

La construction de l'annuaire et son exploitation seront décrites dans le chapitre 4.

3.2. Exploration des sources et de leurs méta-données

3.2.1. Choix de quelques sources de données

Une soixantaine de sources génomiques ont été sélectionnées comme point de départ. Cette sélection a été basée sur la publication '*The Molecular Biology Database Collection: 2004 update*' [GAL04] qui recense 548 bases de données biologiques organisées de façon hiérarchique et thématique. Les sources ont été sélectionnées de façon à avoir une bonne couverture des différents domaines de la génomique impliquant des organismes de tous types. Ainsi, une catégorisation des sources selon trois types a été établie : nucléotidiques (ex :EnsEMBLHuman[44]), protéiques (ex : PIR-PSD) et spécialisées (ex : FlyBase). Dans une première phase, les sources sélectionnées ont été explorées afin d'identifier les méta-données, à partir de la documentation (voir liste en annexe 4).

3.2.2. Choix des méta-données

3.2.2.1. Travaux voisins

Pour arriver à indexer les sources en termes de méta-données, nous nous sommes inspirées de plusieurs travaux :
- Un travail réalisé par le NLM, le '**Information Sources Map**' ou la carte des sources d'information de l'UMLS [UMLS97]. Ce projet qui a été abandonné en 1998, avait pour but de répertorier toutes les sources de données bio-médicales.
- Le catalogue **DBCat**[45] (voir annexe 5) qui est un catalogue de sources de données biologiques, nous a été utile car il nous a permis de voir en quoi il n'était pas suffisant pour trouver une source pertinente en réponse à une question posée. En effet, ce catalogue est constitué de fichiers plats et de texte libre.
- Les éléments du '**Dublin Core**' (voir annexe 3) qui pourraient être pertinents pour des sources biologiques.
- Un travail approuvé par le *'United States Federal Geographic Data Committee'* (FDGC)[46] en 1998, le '*Content Standard for Digital Geospatial Metadata*' (CSDGM) a été également une source d'inspiration malgré l'éloignement thématique. Les objectifs de ce standard étaient de fournir un ensemble commun de terminologies et de définitions pour la documentation de données digitales géospatiales.

[44] http://www.ensembl.org/Homo_sapiens/
[45] http://www.infobiogen.fr/services/dbcat/
[46] http://www.fgdc.gov/metadata/contstan.html

3.2.2.2. Cinq catégories de méta-données

Suite à une étude approfondie des ces travaux, cinq catégories de méta-données ont été identifiées :

- Les méta-données d'**identification** de sources de données :

Elles concernent les informations qui permettent d'identifier les sources. Cela concerne les différents noms que peut avoir une source, les contacts, les publications faisant référence à la source, une description de la source et son statut de maintenance (ex : l'occurrence d'un évènement important et la date de celui-ci dans le cycle de vie de la source), la période de temps à laquelle correspond les données de la source et l'organisation responsable pour la mise en ligne de la source.

La **Table 1** ci-dessous montre un exemple des différents noms que peut avoir une source :

Sources de données	Nom complet	Nom familier	Acronyme
1) IMGT[47]	International ImMunoGeneTics information system	IMGT	IMGT
2) TAIR[48]	The Arabidopsis Information Ressource	TAIR	TAIR

Table 1 : un exemple de méta-données d'identification pour les sources IMGT et TAIR.

- Les méta-données de **qualité** des sources (table 2) :

Ces méta-données incluent tout ce qui permet d'estimer la qualité du contenu d'une source. Par exemple, l'existence d'une révision manuelle des données, si les données sont conformes à un standard, les informations fournies (grâce à une documentation) concernant les origines des données dans la source, la couverture de la source et l'existence de références croisées vers des données d'autres sources.

Source de données	Révision manuelle des données	Qualité de la documentation en ligne	L'existence de références croisées avec d'autres sources
1)EnsEMBL Human	Oui	Riche	Oui
2) TIGR[49]	Non	Riche	Non

Table 2 : Quelques méta-données de qualité pour les sources EnsEMBL Human et TIGR.

- Les méta-données liées aux **contenus** des sources (Table 3) :

Celles-ci permettent de prendre connaissance des différents sujets et organismes couverts par la source. Les méta-données liées aux contenus sont très utiles car elles permettent d'estimer si la source est plutôt généraliste ou spécialisée.

Source de données	Sujets	Organismes
1) MosDB[50]	Clones, contigs, EST…	Riz ou *Oryza sativa*
2) EnsEMBL Human	Chromosomes, clones, exons, transcrits…	Humain ou *Homo sapiens*

Table 3 : Quelques méta-données liées aux contenu des sources

[47] http://imgt.cines.fr/
[48] http://www.arabidopsis.org/index.jsp
[49] http://www.tigr.org/
[50] http://mips.gsf.de/cgi-bin/proj/rice/search/

- Les méta-données liées au **traçage** d'autres méta-données (Table 4, ex : le traçage de la méta-donnée 'Sujets') :
  Les méta-données liées au traçage d'autres méta-données permettent de prendre connaissance du créateur de celles-ci et de sa date de création. De plus, elles permettent de savoir si la méta-donnée a été révisée; le réviseur et la date de révision sont compris aussi.

Source de données	**Item**	**Créateur de la méta-donnée**	**Date de création** *(Format :AAAA-MM)*
1) MosDB	Sujets	Shazia Osman	2004-06
2) EnsEMBL Human	Sujets	Shazia Osman	2004-05

Table 4 : Quelques méta-données permettant le traçage de la méta-donnée 'Sujets'

- Les méta-données de **disponibilité** des sources (Table 5) :

Cette catégorie concerne les informations relatives à la disponibilité d'une source de données. On y trouve des informations sur l'accès au site, son statut, sa version, les modalités d'interaction avec le site (ex : si un langage de requête est requis pour l'utilisation du site), les contraintes d'accès (ex : site payant, nécessite d'avoir un mot de passe…) au site pour les académiques et pour les industriels.

Source de données	**URL du site d'accès**	**Statut du site** *('principal'ou 'alternatif')*
1) TIGR *Arabidopsis thaliana* Gene Index	http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=arab	Site principal
2) WormBase	http://www.wormbase.org/	Site principal

Table 5 : Quelques méta-données de sur la disponibilité des sources

3.2.3. Choix des ontologies

Suite à une étude comparative des différentes bio-ontologies (voir section 2.6) pour structurer certaines méta-données, nous avons décidé d'utiliser le vocabulaire contrôlé du MeSH pour décrire le contenu d'une source de données.
L'ontologie TAMBIS n'a pas été retenue du fait de la trop fréquente indisponibilité de son site Web. Par contre, elle couvrait les termes recherchés dans un premier temps.
L'ontologie BAO fut rejetée de par l'indisponibilité de l'ontologie. En effet, aucun site Web n'est disponible pour accéder à l'ontologie. L'étude a été réalisée à partir de la bibliographie.
Le réseau sémantique de l'UMLS offre une bonne couverture des concepts et des termes pertinents pour cette étude. Cependant, certains termes importants comme 'metabolism' ou 'pathway' n'y sont pas.
L'avantage du réseau sémantique est qu'il est gratuit, téléchargeable et disponible sous divers formats. Par contre, son utilisation nécessite une licence.
Le MeSH a été adopté pour un début, de par sa disponibilité et de sa couverture relativement complète. De plus, aucune licence d'utilisation n'est requise. En outre, la navigation dans le thésaurus via le 'MeSH Browser' rend son utilisation agréable.

En ce qui concerne les organismes couverts par une source de données, nous avons fait appel à la taxonomie du NCBI. Il s'agit de la taxonomie la plus couramment utilisée par les biologistes et la majorité des sources de données biologiques sur le Web font des références croisées avec cette taxonomie. L'existence d'un 'browser' rend son utilisation facile et agréable.

En structurant les sujets et les organismes couverts par les sources de données grâce aux ontologies 'MeSH' et 'NCBI Taxonomy', nous permettrons au biologiste de naviguer dans ces ontologies afin de construire sa requête.

3.3. Modélisation de l'annuaire

Après l'identification des méta-données et leur organisation selon deux ontologies, la prochaine étape était de trouver un modèle générique pour ces méta-données. Ce modèle servirait à structurer les méta-données des sources qui alimenterait l'annuaire ainsi que les informations sur les ontologies utilisées et les relations entre les sources de données. Le modèle complet figure en annexe (voir annexe 6). Une partie de ce modèle arborescent (à une profondeur de 3, à partir de chaque partie) est présentée en figure 2 (page 25) :

Figure 2 : Content for BioRegistry Meta-data

Key:

- DS= Data Source
- *Definitions*
- 'possible values'
- **[+]** = can be repeated unlimited times
- Optional field
- Dublin Core (DC) Element Name Version 1.1 (if applicable)
- *DC Definition*
- (Pointer)

❖ **BioRegistry Identifier** of DS (DC Identifier : *An unambiguous reference to the resource within a given context*)

PART A. SOURCE META-DATA

Section A1 : Identification Information

Basic information about the data source

A1.1 : Data Source

A1.1.1: DS Name (DC Title : *A name given to the resource, typically the formal name*)
A1.1.2 : DS Contact **[+]**

A1.2 : Citation : *information to be used to reference the data source* **[+]**

A1.2.1 : Author(s)
A1.2.2 : Publication Year: *Year during which article is published*
A1.2.3 : Title: *Title of article*
A1.2.4 : Journal Information
A1.2.5 : PubMed Identifier
A1.2.6 : On-line link (URL)

A1.3 : Description (DC Description : *An account of the content of the resource*)

A1.3.1 : Abstract: *characterisation of the DS, including its intended use and limitations*
A1.3.2 : Purpose: *intentions with which the DS was developed*
A1.3.3 : Supplemental information: *other descriptive information about the DS*
A1.3.4 : Entry Sample : *a sample of an entry in the DS*

A1.4 : Time Period of Content : *time period(s) for which the DS corresponds to the currentness reference.*

A1.5 : Maintenance Status

A1.5.1 : Update (DC Date : *A date of an event in the lifecycle of the resource*)
A1.5.2 : Release (DC Date : *A date of an event in the lifecycle of the resource*)
A1.5.3 : Maintenance Organisation **[+]** (DC Publisher: *An entity responsible for making the resource available ex: a person, an organisation*)

A1.5.3.1 : Maintenance Name
A1.5.3.2 : Maintenance Address
A1.5.3.3 : Maintenance Country
A1.5.3.4 : Maintenance e-mail
A1.5.3.5 : Maintenance Telephone

Section A2 : Topic Information

A2.1: Keywords : *Words or phrases summarizing an aspect of the data source*
(DC Subject : *A topic of the content of the resource*)

A2.1.1 : Subject **[+]**
A2.1.2 : Organism

A2.2 : Use Constraints [+] : (DC Rights : *Information about rights held in and over the resource*) *Restrictions and legal prerequisites for using the data source after access is granted. These include any use constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on using the data source.*

A2.3 : Help Desk :

A2.3.1 : e-mail
A2.3.2 : Name
A2.3.3 : Telephone

Section A3 : Meta-Data Tracking Information

A3.1 : Meta-data tracking [+] : *Information about meta-data origin*

A3.1.1 : Meta-data Item: *any possible item such as 'Manual Revision', 'Description', 'All Meta-data items'…*
A3.1.2 : Creator: *name of person responsible for creating the meta-data*
A3.1.3 : Creation Date: *date at which the meta-data was first created*
A3.1.4 : Reviewer: *name of person undertaking the review of the meta-data*
A3.1.5 : Review Date: *date at which the meta-data is reviewed*

A3.1.6 : Status: 'Reviewed', 'Not Reviewed', 'Ongoing'

Section A4 : Data Quality Information : *A general assessment of the quality of the data source.*

A4.1 : Manual Revision: *Information concerning the manual revision of the DS*

A4.1.1 : Status: 'Yes', 'No', 'Not Found'

A4.1.2 : Supplemental Information

A4.2 : Standard compliancy [+] : *ex: MIAME compliant, ASN.1 compatible*

A4.2.1 : Standard Name

A4.2.2 : Standard Reference (URL)

A4.2.3 : Status: 'Total', 'Partial', 'Not Documented', 'Verified', 'Not Verified'

A4.2.4 : Supplemental Information

A4.3 : Data Reference Information : *information about data origin in the DS*

A4.3.1 : On-line Documentation Existence **[+]** : *description of any on-line reference information*

A4.3.2 : Existence of Reference to Publication in DS Entries

A4.3.3 : Supplemental Information

A4.4 : Source Coverage [+] : *Number of different types of entities in the data source*

A4.4.1 : Entity Name : *type of entity (ex: Genes, EST, EST Cluster, Contig etc. possibly associated to the organism)*

A4.4.2 : Number of Entities : *number of entities of this type in the DS*

A4.4.3 : Organism Name (with reference to Section A2.1.2.2.1)

A4.5 : Cross Reference [+] : (DC Relation : *A reference to a related resource*) *List of data sources cross referenced in the DS*

Section A5 : Availability Information : *Information about the availability of the DS*

A5.1 : Site of Access :

The site from where the DS may be accessed

A5.1.1 : Site Identification **[+]**

A5.1.2 : DS version : *version of the DS in the site of access*

A5.1.3 : Interaction Modalities: *Requirements for source usage (ex: Query language, Web services…)*

A5.2 : Access Constraints:

A5.2.1 : Access Constraints for Academics: 'Free' or 'Password' or 'Registration fees'

A5.2.2 : Access Constraints for Industrials: 'Free' or 'Password' or 'Registration fees'

PART B. ONTOLOGIES

Section B1 : Ontology Identification Information

B1.1 : Thesaurus / Ontology/Controlled Vocabulary/Taxonomy [+]

B1.1.1 : Thesaurus Name *ex: MeSH* (cf. Part A, Section A2.1.1.1)

B1.1.2 : Thesaurus Version *ex:2004*

B1.1.3 : Thesaurus URL

B1.1.4 : Supplemental Information

PART C. RELATIONSHIPS

Section C1 : Data Source Relationships

C1.1 : DS content coverage [+] *Relationships, in terms of content coverage, of the DS with other data sources*

C1.1.1 : Source Name 1 : *Actual DS*

C1.1.2 : Source Name 2 : *Source from which the DS draws information from or contributes information to.*

C1.1.3 : Type of Relationship : 'Draws from', 'Contributes to'

C1.1.4 : Relationship Extent : 'Complete', Partial', 'Unknown'

C1.1.5 : Field Map: *Field in the DS entry corresponding to the relation source*

Le modèle est constitué de trois grandes parties :
'Partie A : Source MetaData' 'Partie B : 'Ontology' et 'Partie C : Relationships'.
L'élément 'SourceMetaData' se compose de cinq sections (A1 à A5)
Dans le modèle, la section 'identification' a été décomposée en cinq sections :
Chaque méta-donnée d'identification est elle-même ramifiée en plusieurs sous-éléments c'est-à-dire, un élément fils dans l'arbre contient lui-même des fils et des feuilles.
Ces différentes méta-données ont été rajoutées pour amener plus de précision concernant l'identification d'une source. Par exemple, pour la méta-donnée 'release' (voir annexe 6), la fréquence de celle-ci, les dates de première et dernière 'release' ainsi qu'un identifiant ont été rajoutés.

L'élément 'topic information' se compose de mots-clés pour les sujets et les organismes couverts par la source. Les termes correspondants des ontologies ainsi que leurs identifiants pour les sujets et les organismes y figurent aussi. Les contraintes d'utilisation légale aux données de la source ainsi qu'une aide en ligne relative à ces données sont incluses aussi.

Il nous a paru utile d'inclure des méta-données permettant de retracer l'origine tous les autres méta-données du modèle. Celles-ci sont situées dans la section 'Meta-data tracking Information'.

Par ailleurs, les méta-données liées à la qualité des données dans les sources se composent aussi de plusieurs sous-éléments. Par exemple, la méta-donnée qui indique si les données d'une source sont conformes à un standard (ex : ASN.1[51]) : d'autres méta-données associées ont été rajoutées tels que la référence à l'URL du standard, le statut de la compatibilité des données, etc.

L'information que comprend la catégorie 'disponibilité de l'information' est : le ou les sites d'accès à une source et si ces sites sont des sites miroirs ou pas, l'organisme responsable de mettre la source en ligne, les contraintes d'accès pour les académiques et/ou les industriels…

La partie B regroupe les informations permettant d'identifier les ontologies utilisées pour décrire certaines méta-données dans les sources.
Enfin, la partie C concerne les relations entre une source et d'autres sources, les types de relations et leurs degrés.

En outre, à partir de la version 1.1 du 'Dublin Core' (DC), une correspondance a été faite entre certains éléments établis pour le modèle de l'annuaire et ceux du 'Dublin Core' quand ceci était possible, car le DC ne décrit pas tous les champs qui nous concernent. Ce modèle générique a ensuite été implémenté sous la forme d'un schéma XML.

[51] Abstract Syntax Notation one : http://www.asn1.org/

Chapitre 4 : Réalisations et résultats

4.1. Langages et Outils Utilisés

4.1.1. XML (eXtensible Markup Language)

XML [RHM02] est le standard soutenu par le W3C[52] pour le balisage de documents. Il définit une syntaxe générique utilisée pour formater des données avec des balises simples et compréhensibles par l'Homme. Ce format est suffisamment souple pour être adapté à des contextes aussi variés que les sites Web, l'échange de données électroniques…
XML est un langage souple car il permet de définir de nouvelles balises. Ce langage, grâce aux balises, permet de structurer des documents**.** En XML, l'unité de base de données et son marqueur est appelé 'element'.

Un **schéma XML** est un document XML qui contient une description formelle du contenu d'un document XML **valide** (une instance). Il est composé d'un ou de plusieurs éléments. Chaque élément a un type (ex : une date, un string…), un identificateur et une valeur selon son type. Un élément peut être simple ou complexe (composé de plusieurs éléments eux aussi simples ou complexes).
Différents langages de schémas XML existent (ex : RELAX NG) mais dans le cadre de ce travail, seul le langage de schéma recommandé par le W3C est utilisé.

XML s'avère être pertinent pour notre travail car il permet une structuration hiérarchique des éléments, ce qui est idéal pour représenter notre modèle. En outre, XML sert de base pour la définition de nouveaux langages tels que RDF[53] et OWL[54]. Ceux-ci sont les langages utilisées parmi d'autres pour le Web sémantique.

4.1.2. XSL et XSLT

XSL[55] est divisé en deux parties : **XSLT**[56]et **XSL-FO**[57]. Pour les besoins de ce travail, nous ferons appel qu'à XSLT.
XSL permet de définir des feuilles de style pour les documents XML. Ces feuilles de style sont constituées d'un ensemble de règles de transformation applicables au document XML et permettent de générer d'autres types de documents (PostScript, HTML, ...) ou bien un fichier XML de structure différente.

XSLT : le processeur XSLT, composant logiciel chargé de la transformation, crée une structure logique arborescente à partir du document XML et lui fait subir des transformations selon les règles contenues dans une feuille XSL pour produire un arbre résultat représentant, par exemple, la structure d'un document HTML.

[52] World Wide Web Consortium
[53] Ressource Description Framework
[54] Ontology Web language
[55] eXtensible Stylesheet Language
[56] XSL Transformations
[57] XSL Formatting Objects

4.1.3. XPath (XML Path Language)

XPath est un langage non-XML utilisé pour identifier des parties de documents XML. XPath voit un document XML comme un arbre de nœuds. Le langage permet de désigner un ou plusieurs nœuds dans un document XML, à l'aide d'**expressions de chemin.**
(ex : `<xsl:template match="BioRegistry/SourceMetaData">`) pour la sélection de règles à appliquer sur les nœuds.

4.1.4. Editeurs XML

oXygen 4.2©[58] est un éditeur XML utilisé pour la manipulation de documents XML et XSL. Cette application multi-plateformes, couplée avec les technologies de transformation comme XSLT, supporte la sortie vers de multiples formats cibles : HTML, XML, TXT…
Pour la génération de vues à partir des schémas, un autre éditeur, **XMLSpy 4.3**©[59], a été utilisé à cause de l'absence de cette fonctionnalité sous oXygen 4.2©.
Tout le travail réalisé a été effectué sous Windows© 2000 et sous Windows© XP.

4.2. Constitution de l'annuaire 'BioRegistry'

Après la conception du modèle des méta-données pour l'annuaire, la prochaine phase était d'établir un schéma XML basé sur ce modèle. Comme évoqué précédemment, XML est un langage souple qui permet de définir ses propres balises et permet une structuration hiérarchique d'où le choix de XML pour implémenter le modèle de méta-données conçu.

4.2.1. Le schéma de l'annuaire 'BioRegistry'

Le but d'utiliser un schéma XML (voir annexe 7) pour représenter le modèle décrit au chapitre 3 était de structurer l'annuaire sous forme d'un arbre et de profiter de tous les outils associés à XML. L'élément racine, 'BioRegistry', est composé de deux types d'éléments : l'élément 'Source Meta-data' et l'élément 'Ontologies' (voir figure 3). Par manque de temps, nous n'avons pas implémenté la partie C du modèle ('Relationships'). La date de dernière mise à jour ('BRUpdate') de l'annuaire est un attribut de la racine.
Les éléments 'Ontology' représentés sont référencés à l'aide de clés dans l'élément 'SourceMetaData'. Le nombre d'éléments 'SourceMeta-data' et 'Ontology' que peut comporter l'annuaire est illimité. La figure 3 illustre les principales sections de l'annuaire. Les '+' indiquent que l'élément contient des fils, sinon l'élément est une feuille dans l'arbre XML.

[58] www.oxygenxml.com
[59] www.xmlspy.com

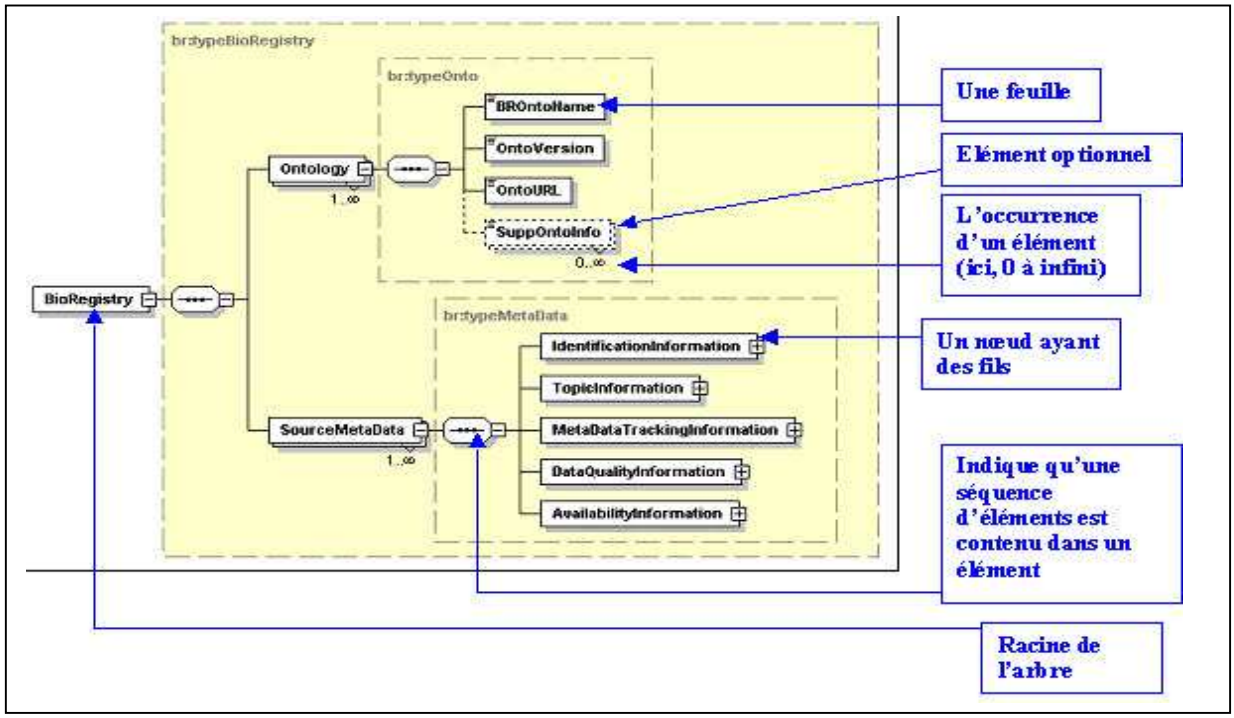


Figure 3 : Plan général du schéma "BioRegistry"

4.2.2. L'élément 'Ontology'

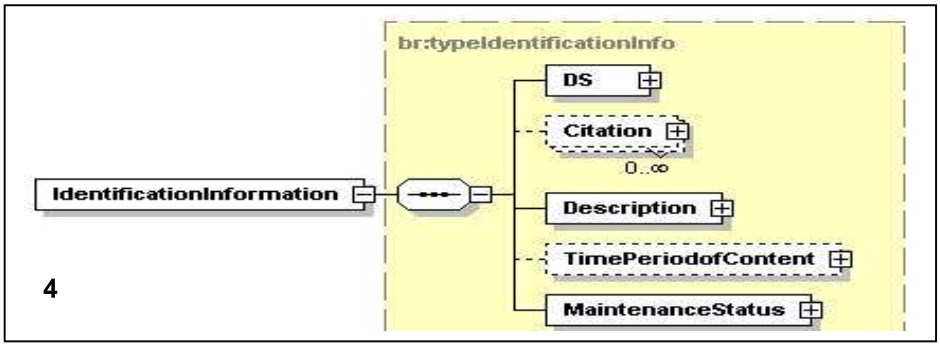
L'élément 'Ontology' est composé d'une séquence comportant le nom de l'ontologie, sa version, l'URL permettant d'y accéder ainsi que des informations complémentaires sur l'ontologie, le cas échéant (élément optionnel).

4.2.3. L'élément 'SourceMetaData'

Cet élément distinct pour chaque source a pour attribut le numéro (ou l'identifiant) d'une source dans l'annuaire.

4.2.3.1. L'élément 'Identification Information'

Cet élément illustré dans la figure 4, est constitué d'une séquence comportant les noms de la source, les citations concernant la source (optionnelles), une description du contenu, de la période de temps du contenu de la source (optionnelle) et de la maintenance. Pour l'élément citation (figure 4a), tous les informations concernant les publications y sont : auteur, nom du journal, titre de l'article, identifiant 'PubMed'… Tous les coordonnées de l'organisme de maintenance figurent dans cette partie aussi.



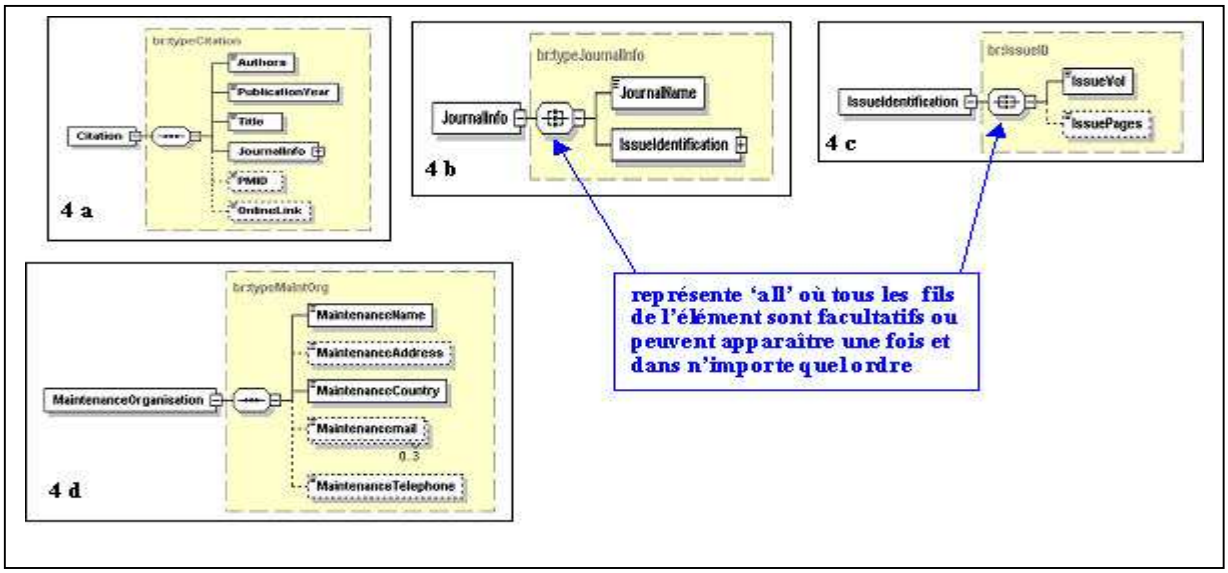


Figure 4 : 'Identification Information' ; 4a, 4b, 4c : 'Citation' ; 4d : 'Maintenance Organisation'

4.2.3.2. L'élément 'Topic Information'

Cet élément comprend la liste des termes permettant d'indexer la source. Il est constitué d'une part, de mots-clés avec un terme et un identifiant qui sont regroupés dans un élément 'Subject' caractérisé par une référence à une ontologie. Il en est de même pour les organismes concernés (voir Figure 5).

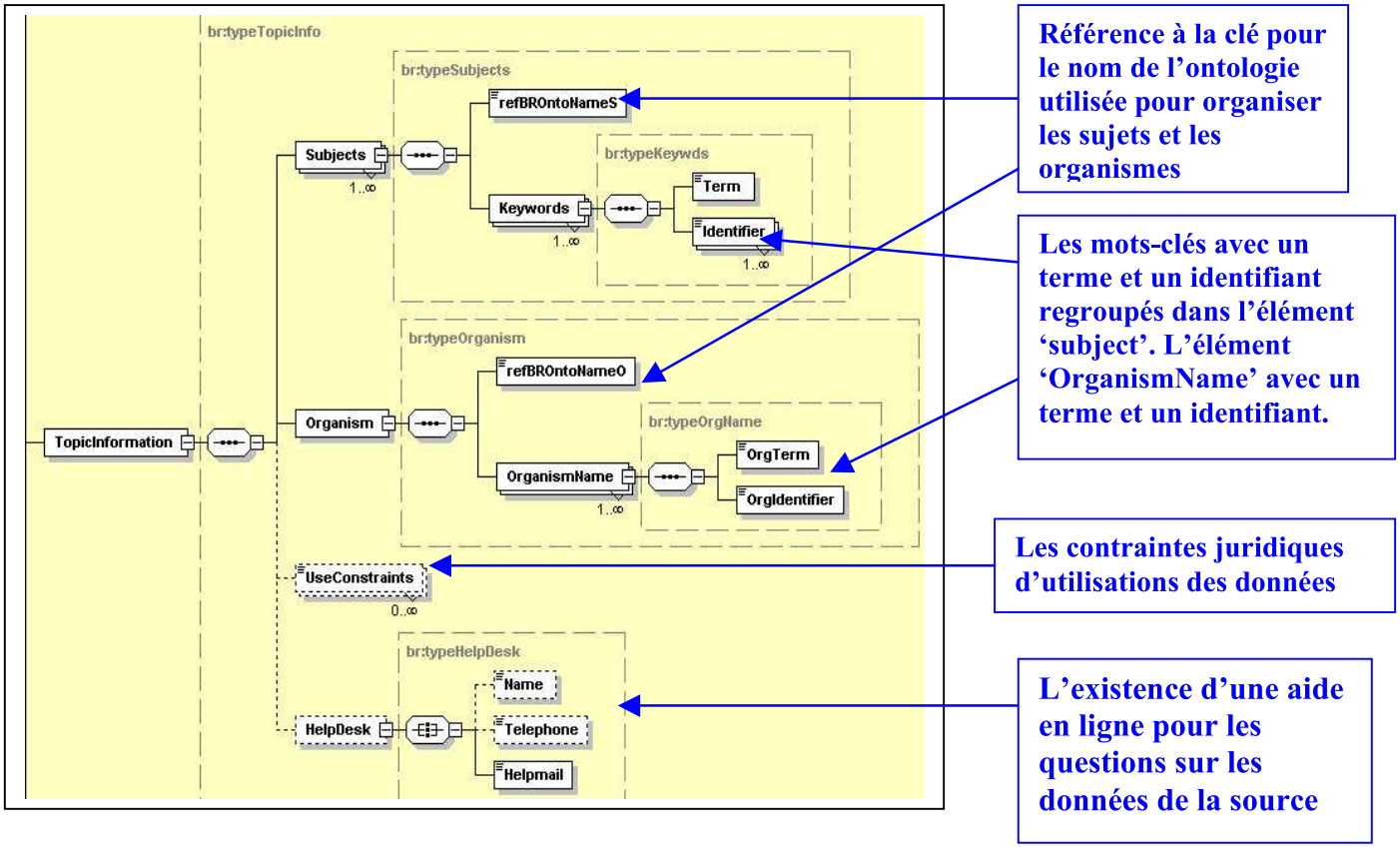


Figure 5 : L'élément 'TopicInformation' de 'BioRegistry'

4.2.3.3. L'élément 'Meta-data Tracking Information'

Cet élément (Figure 6) concerne le traçage de toutes les méta-données à une certaine profondeur (profondeur 2) du sous-arbre 'SourceMetaData'. Les noeuds sélectionnés sont tous au même niveau dans l'arbre : Data Source, Citation, Description, Time Period of Content, Maintenance Status, Subject, Organism, Use Constraints, Help Desk, Manual Revision, Standard Compliancy, Data Reference Information, Source Coverage, Cross Reference, Site of Access, Access Constraints. En outre, un champ recouvrant tous les méta-données ('allMetaDataItems') a été inclus.

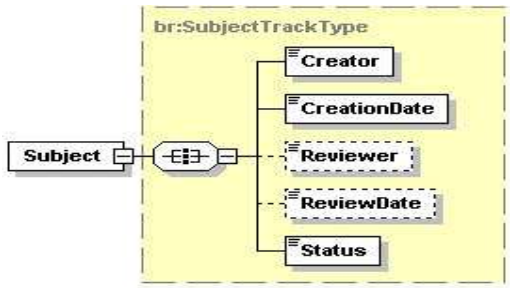


Figure 6 :Traçage d'un item, ici, 'Subjects'. Pour chaque item de méta-données, il y a une séquence du nom du créateur, de la date de création de l'item, un re-lecteur, la date de révision, ces deux derniers étant facultatifs, et le statut du traçage (ex : 'en cours' ou 'ongoing').

4.2.3.4. L'élément 'DataQuality Information' (figure 7)

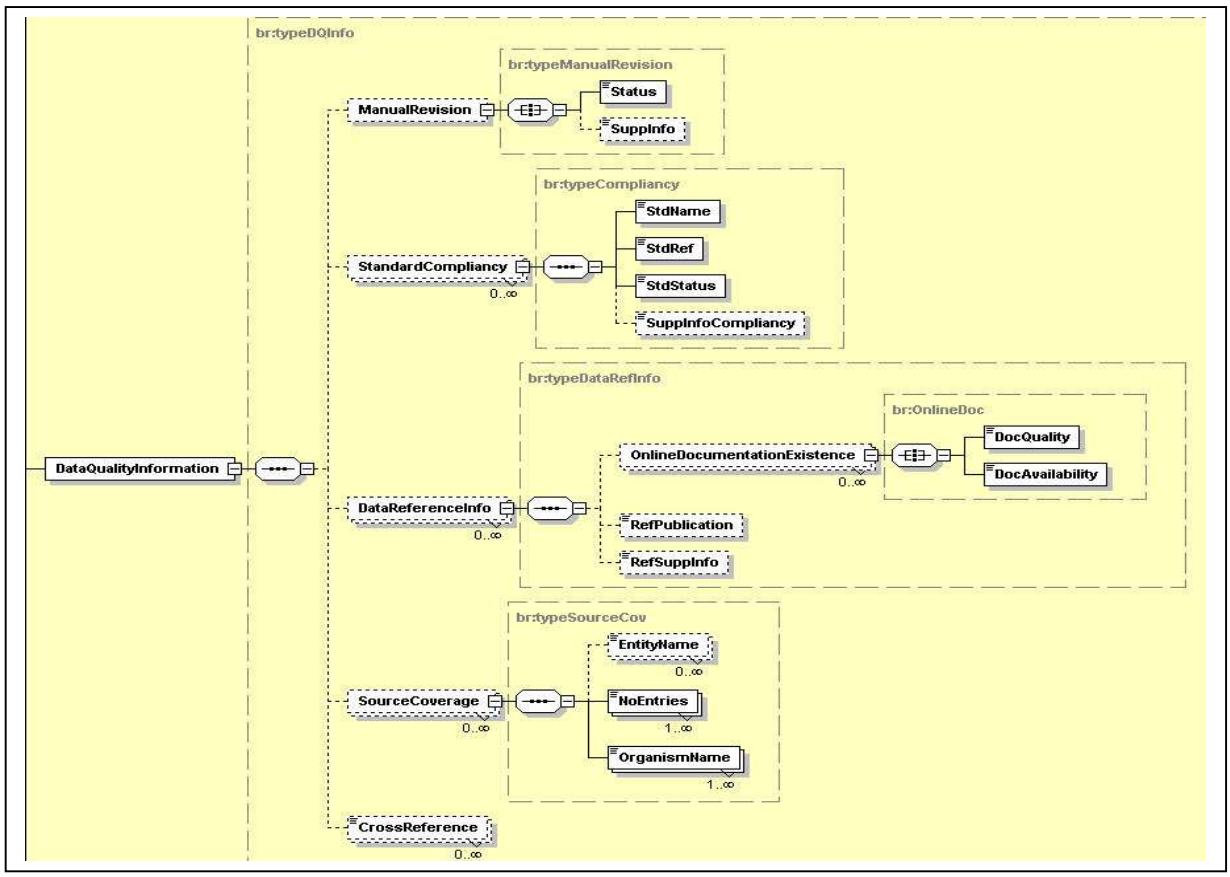


Figure 7 : Les différents composants de l'élément 'DataQualityInformation'

4.2.3.5. L'élément 'Availability Information' (figure 8)

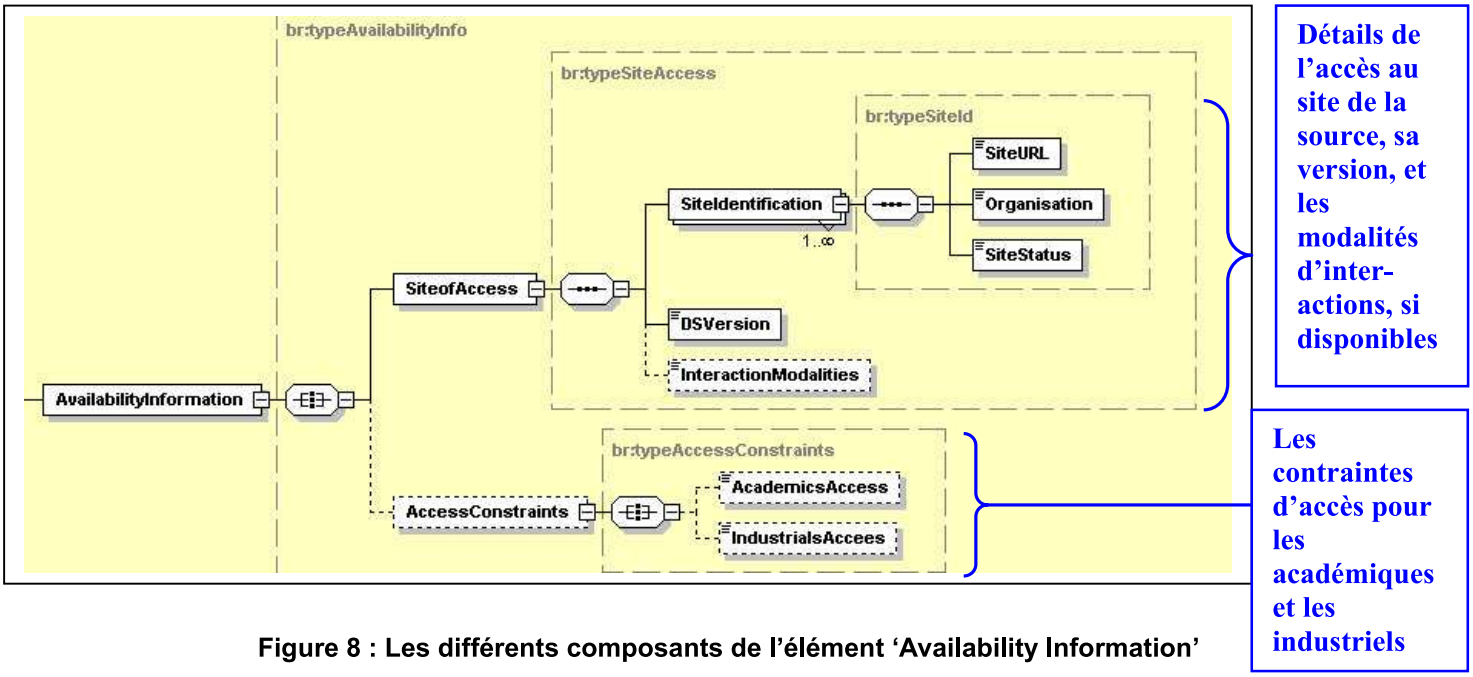


Figure 8 : Les différents composants de l'élément 'Availability Information'

4.3. Exploitation du schéma XML

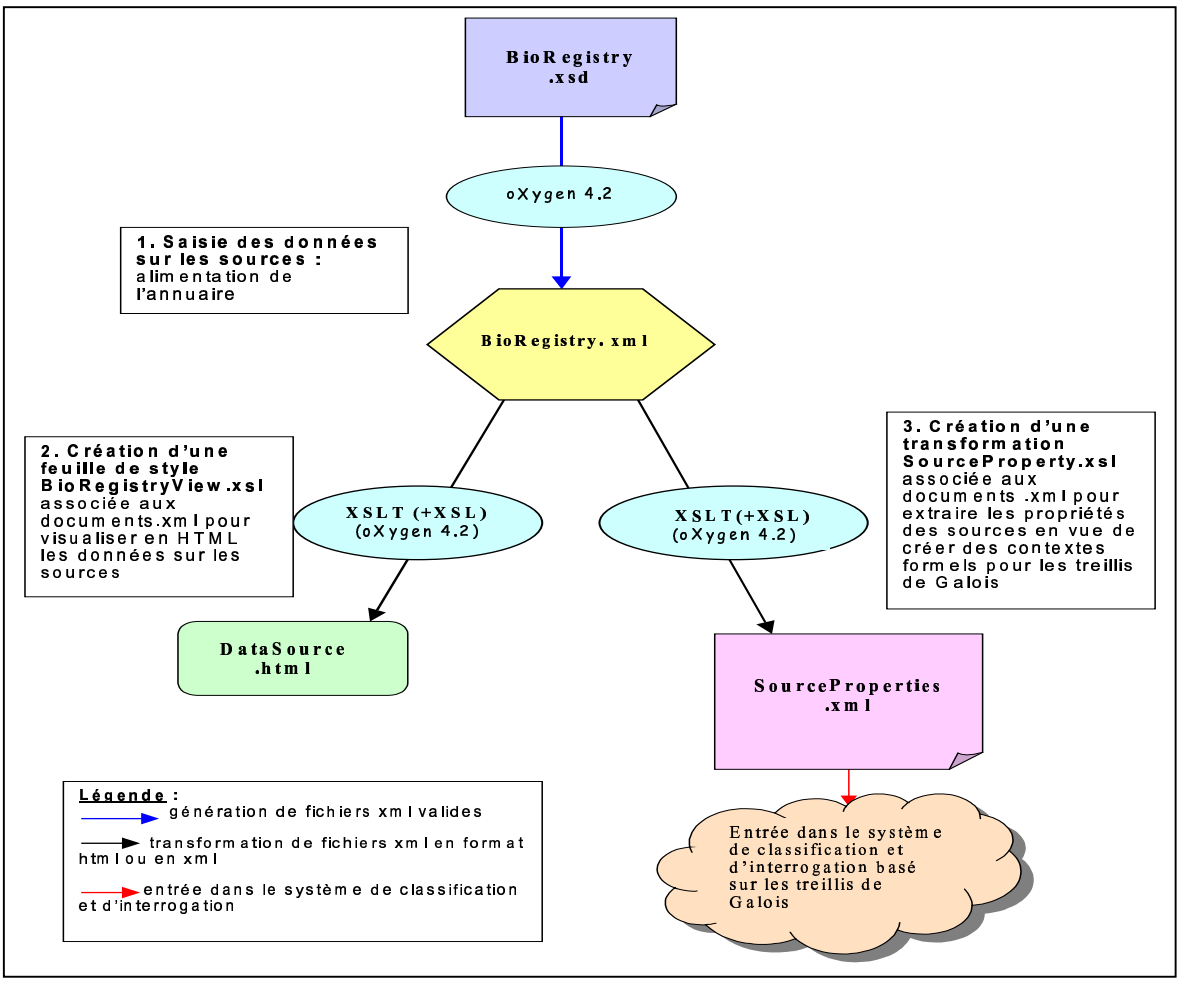


Figure 9 :Travail réalisé autour du schéma XML

Le schéma XML 'BioRegistry.xsd' a été exploité en trois temps comme le représente la figure 9. Tout d'abord (1. sur figure 9), une instance de schéma a été créée (BioRegistry.xml), en saisissant les méta-données relatives aux sources explorées au début de ce travail. Ensuite (2. sur figure 9), une transformation du document XML en format HTML a été programmée pour permettre la visualisation du contenu de l'annuaire. Pour terminer (3. sur figure 9), une autre transformation a été programmée en vue d'interroger l'annuaire.

4.3.1. Alimentation de l'annuaire

Les données sur quelques sources (voir annexe 8) ont été saisies manuellement sous oXygen 4.2©. La saisie sous oXygen 4.2© présente un avantage car la validité des données saisies peut être vérifiée grâce au schéma XML. De plus, l'ordre dans lequel les balises ouvrantes et fermantes doivent apparaître dans le fichier XML est respecté par rapport au schéma.
Par ailleurs, en moyenne, la vérification et la saisie des données pour une source s'étalait sur environ deux jours. Un document source a une taille de 20 Ko en moyenne.

4.3.2. Visualisation des données dans l'annuaire

Afin de visualiser les entrées pour les sources à partir du fichier XML, une feuille de style ('BioRegistryView.xsl', voir annexe 9) a été créée sous oXygen 4.2© pour générer un fichier HTML. Cette feuille de style créée devait traiter tous les nœuds du document XML afin de permettre l'affichage des données. De plus, des liens hypertextes ont été prévus avec le document HTML de la source, les identifiants PubMed, les contacts etc.
La figure 10 ci-dessous montre une partie du fichier HTML généré pour la source 'FlyBase' :

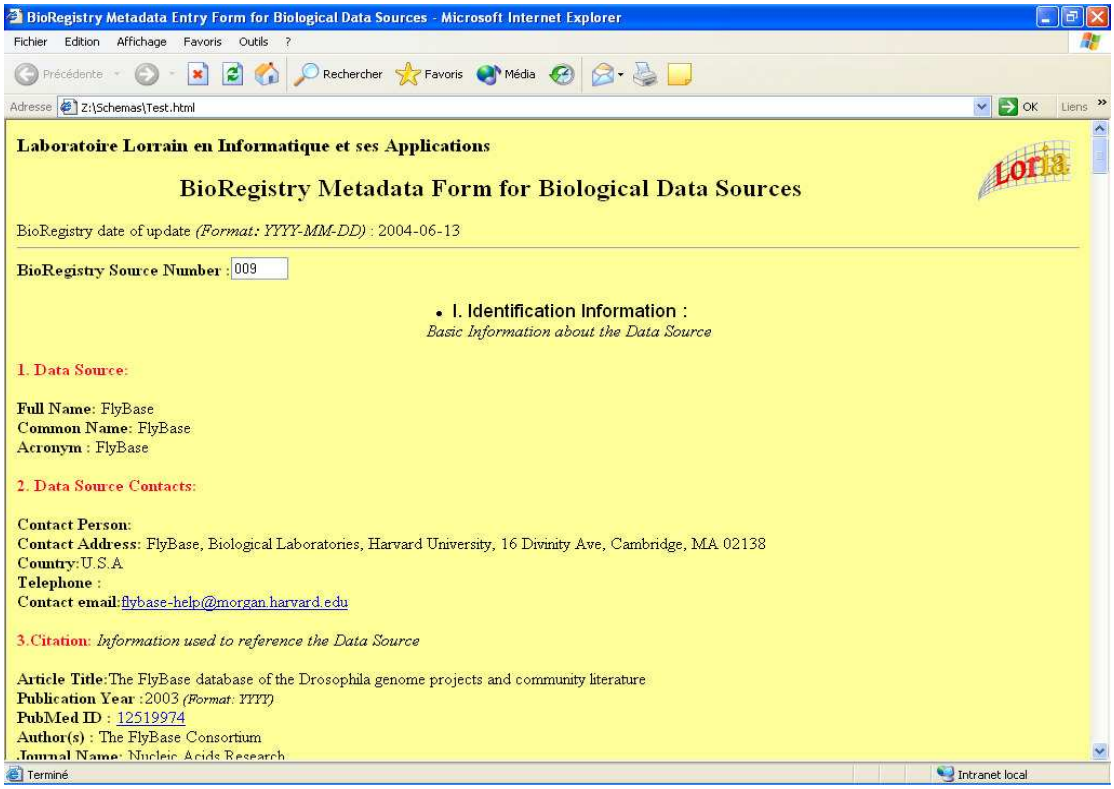


Figure 10 : Fichier HTML généré montrant les méta-données de la source 'FlyBase'

4.3.3. Vers une interrogation de l'annuaire

Une méthode de classification et de recherche des sources génomiques pertinentes pour une question donnée, basée sur les treillis de Galois, a été développée par un étudiant en DEA [MES04]. Elle consiste à construire le treillis à partir des propriétés binaires des sources. Ces propriétés sont extraites à partir d'un ensemble de méta-données associées aux sources. Comme point de départ de ce travail, une feuille de style, 'SourceProperties.xsl' (voir annexe 10) a dû être créée pour ensuite réaliser une transformation sous oXygen 4.2©, qui n'extrait que certaines propriétés dans un autre fichier xml ('SourceProperties.xml' voir annexe 11). Les propriétés extraites sont la date de mise à jour de la source, le numéro de la source dans l'annuaire, le nom familier de la source, la fréquence de mise à jour et des 'releases' de celle-ci, et les sujets et organismes couverts par la source avec leurs ontologies associées ainsi que leurs versions.

Conclusion et perspectives

Le travail réalisé est au cœur du projet 'BioRegistry' qui vise à mettre en place une structure en annuaire de sources Web de données biologiques. Ces sources sont complexes à cause de leurs diversités (plus de 500), de leurs contenus et de leurs structures hétérogènes. Des méthodes sont donc nécessaires pour faciliter leur identification. De ce fait, il m'a été demandé de réaliser un modèle d'une structure en annuaire pouvant servir de base au projet.
Dans la première partie du travail, j'ai dû identifier et collecter des méta-données pertinentes associées à ces sources. Ensuite, elles ont été organisées selon des ontologies de domaine. Un modèle pour l'annuaire a alors été etabli.
La deuxième phase du projet consistait à construire l'annuaire selon le modèle établi. Une fois celle-ci achevée, j'ai exploité l'annuaire afin de permettre la visualisation de son contenu ainsi que l'extraction des informations pertinentes qui serviront à permettre l'interrogation de ces sources.

Ce projet illustre ainsi l'étroite collaboration qui peut exister entre biologistes et informaticiens : les premiers étant confrontés au problème de recherche d'informations pertinentes dans les sources de données, les seconds étant capables de développer des applications capables d'automatiser et d'optimiser ces recherches. Ce type de collaboration, qui a pour but d'accélérer la découverte scientifique, a conduit à la naissance d'une discipline de recherche relativement récente : la bio-informatique.

Les perspectives de ce travail sont multiples et prometteuses.
La définition des relations (typées) entre les sources reste à faire. Ce travail, non négligeable, serait important car il permettrait éventuellement à l'utilisateur de prendre conscience de ces diverses relations (recouvrement total, partiel, dérivation…).

Eventuellement, il serait très utile de mettre en place une base de données pour le stockage des informations de l'annuaire, une fois que les tables et les relations entre elles seront bien définies. Une structure en base de données permettrait le stockage de ces données et l'interrogation de l'annuaire avec plus d'aisance.

De plus, il faudrait construire un système qui prendrait en compte la maintenance de l'annuaire et son peuplement. En effet, les données sur d'autres sources devraient être insérées dans l'annuaire, les données sur les sources devenues obsolètes devront être éliminées et celles sur les sources existantes devront être mises à jour. Il serait ainsi utile de développer une interface permettant la saisie et la mise à jour de ces données.
En outre, une système d'interrogation de l'annuaire basé sur un formulaire permettra au biologiste d'effectuer des requêtes multi-critères pour découvrir les différentes sources de données susceptibles de répondre à son besoin.

Quant à moi, je constate que ce stage s'est déroulé dans de bonnes conditions et m'a permis de parfaire mes connaissances dans le domaine de la bio-informatique ainsi que de compléter ma formation. Il comportait dans mon esprit plusieurs objectifs qui sont les suivants : la réussite de mon intégration au sein d'une équipe, l'enrichissement de mes connaissances théoriques et techniques en informatique, la compréhension de mon travail et l'obtention de résultats concrets. De plus, ce stage et l'environnement du LORIA m'ont permis de m'immerger profondément dans le domaine de l'informatique qui était jusqu'alors un domaine relativement mystérieux pour moi.

En conclusion, j'ai le sentiment d'avoir atteint les objectifs que je m'étais fixé. Ce stage, formant partie intégrante du cursus de Master de bio-informatique, participe réellement à notre formation et nous prépare au développement de qualités professionnelles requises dans le monde du travail.

Bibliographie

[ASH00] **Ashburner *et al.*,** «Gene ontology: tool for the unification of biology. The Gene Ontology Consortium», *Nature Genetics*, **25**(1):25-9, May 2000.

[BAK99] **Baker, P., Goble, C., Stevens, R., Bechhofer, S., Paton, N., Brass, A.,** «An ontology for bioinformatics applications », *Bioinformatics*, **15**(6): 510-520, 1999.

[BAK98] **Baker, P.G., Brass, A., Bechhofer, S., Goble, C., Paton, N., Stevens, R., "**TAMBIS**--**Transparent Access to Multiple Bioinformatics Information Sources" *Proc Int Conf Intell Syst Mol Biol*.**6**:25-34, 1998.

[BEN02] **Ben-Miled, Z.** *et al.*, «BAO, A biological and chemical ontology for information integration », Bioinformatics, **1**:60-73, 2002.

[BL01] **Berners-Lee, T., Hendler, J., Lassila, O.,** «The Semantic Web», Scientific American, 284(5), 34-43, May 2001.

[BoDS00] **Boudjlida, N., Devignes, M.D., Smaïl, M.,** « Services for open distributed environment » in XEWA 2000: IEEE Workshop on XML-Enabled Wide Area Search in Bioinformatics, League City, Texas.

[CHA04] **Chabalier, J.,** « Acquisition incrémentale et représentation des systèmes intégrés bactériens par une approche orienté-objet», Thèse, Université Marseille I, 2004.

[DEP02] **de Palma, H.,** « Création d'un annuaire de ressources Web de données biologiques sous forme d'un service Web», *rapport de stage de licence professionnelle en Informatique*, 2002.

[DS04] **Devignes, M.D., Smaïl, M.,** « Integration of Biological Data from web resources : management of multiple answers through metadata retrieval» *Short paper, ISMB-ECCB*, 31 july- 4 august, 2004.

[DS^{+}03] **Devignes, M.D., Smaïl, M., Norsa, Y., Collet, P., Domenjoud, L., Dauça, M.,** « A generic solution for automated collecting and integration of biological data from web sources » *Short paper ECCB*, 2003.

[DSB02] **Devignes, M.D., Smaïl, M., Boudjlida, N.,** « Collecte de données à partir de sources multiples et hétérogènes : Vers une structure de médiation conviviale et orientée source » In *Journées scientifiques sur le web sémantique*, Paris, 2002.

[DSS02] **Devignes, M.D., Schaaff, A., Smaïl, M.,** « Collecte et intégration de données biologiques hétérogènes sur le web », *Ingénierie des systèmes d'information* **7**(1-2) : 45-61, 2002.

[GAL04] **Galperin, M.Y.,** «The molecular biology database collection : 2004 update», *Nucleic Acids Research*, **32**: D3-D22, 2004.

[GO04] **Gene Ontology Consortium,** «The Gene Ontology (GO) database and informatics resource», *Nucleic Acids Research*, **32**: D258-D261, 2004.

[GSM01] **Goble, C., Stevens, R., McEntire, R.,** *et al.* « A market place for ontologies and ontology-based tools and applications in the life sciences », http://www.omg.org/docs/lsr/01-07-38.pdf , 2001.

[GRU93]**Gruber T.R.,** «A translation approach to portable ontologies» *Knowledge Acquisition*, **5**(2):199-220, 1993.

[JEF98] **Jeffrey, K.,** « Metadata : An overview and some issues », *ERCIM News* No. 35, October 1998.

[McC03] **McCray, A.T.,** « An upper-level ontology for the biomedical domain », *Comparative and Functional Genomics*, **4**: 80-84, 2003.

[MES04] **Messai, N.,** « Treillis de Galois et ontologies de domaine pour la classification et la recherche de sources de données génomiques », *rapport de DEA Informatique, Ecole doctorale IAEM Lorraine*, 2004.

[NEY03] **Neyrinck, N.,** « Prise en compte de critères de qualité dans un scénarion de collecte et d'intégration d'annotations fonctionnelles de gènes », *rapport de stage de D.U.T d'Informatique*, 2003.

[RRSW04] **Rojas I., Ratsch E., Saric J., Wittig U.,** « Notes on the use of ontologies in the biochemical domain», *In Silico Biology*, **4**(1): 89-96, 2004.

[SWLG04] **Stevens, R., Wroe, C., Lord, P., Goble C.,** «Ontologies in Bioinformatics», *In Handbook on Ontologies*, International Handbooks on Information Systems, Staab,S., Studer,R., Editors, 2004.

[UMLS97] «UMLS Information Sources Map (ISM) », Section 5, In *UMLS Knowledge Sources Documentation*, 8th Edition, January 1997.

[RHM02] **Rusty Harold, E., Means, W.S.,** «XML in a nutshell », 2ème Edition, Editions O'Reilly, ISBN 2-84177-223-3, 2002.

Webographie

1. Article ISM : http://www.medinfo.rochester.edu/umls/doc

2. Article ontologies [GSM01]: http://www.omg.org/docs/lsr/01-07-38.pdf

3. DBCAT : http://www.inforbiogen.fr/services/dbcat/

4. Dublin Core :
http://www.bibl.ulaval.ca/DublinCore/usageguide-20000716fr.htm

http://www.openweb.eu.org/articles/dublin_core/

5. Gene Ontology :http://www.geneontology.org/

6. MeSH: http://www.nlm.nih.gov/mesh/
http://www.nlm.nih.gov/mesh/introduction2004.html

7. NCBI Handbook [NCBIHb]:
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=handbook.TOC&depth=2

8. Ontologie BAO : http://baciis.engr.iupui.edu/

10. oXygen 4.2 :
http://www.oxygenxml.com/doc/oXygenUserGuide-standalone-fr.pdf

11. Réseau Sémantqiue UMLS : http://www.nlm.nih.gov/research/umls/META3.HTML

12. Standard de Méta-données : http://www.fgdc.gov/metadata/contstan.html
http://biology.usgs.gov/fgdc.metadata/version2/

13. Swiss Prot : http://www.expasy.org/sprot/

14. TAMBIS :http://imgproj.cs.man.ac.uk/tambis/

15. Tree of Life:
http://tolweb.org/tree/learn/learning.html

16. XSL et XSLT :
http://www.commentcamarche.net/xml

GLOSSAIRE

Accession Number pour Numéro d'Accession : c'est un identifiant unique qui est donné à une séquence lorsqu'elle est soumise à une banque de données (EMBL, DDBJ, GenBank et Swiss Prot). Ce numéro peut être accompagné d'un numéro de version permettant de tracer les modifications successives apportées à l'enregistrement de cette séquence.
DDBJ : 'DNA Data Bank of Japan'
Dublin Core: Un standard de méta-données. Elles utilisent un standard défini formellement pour la description d'éléments de méta-données. Cette formalisation contribue à améliorer la cohérence avec d'autres communautés décrivant des méta-données et à augmenter la précision, la portée et la cohérence interne de la définition des éléments du Dublin Core.
EBI : 'European Bioinformatics Institute'
ENZYME : banque de données spécialisée sur la nomenclature des enzymes
EMBL : 'European Molecular Biology Laboratory'
EMBO: 'European Molecular Biology Organisation'
EXPASY : 'Expert Protein Analysis System' pour système expert d'analyse de protéines : le serveur protéomique du Swiss Institute of Bio-infomatics (SIB).
FTP: 'File Transfer Protocol'
GenBank: Banque de séquences nucléiques
GenPept : Séquences codantes de GenBank
GO : 'Gene Ontology'
Hyperlien : connexion entre une information et une autre, sur le Web, par le moyen d'URL
MedLine : 'MEDLARS onLINE' ('MEDical Litterature Analysis and Retrieval System'). Une banque de références ('abstracts') des articles publiés dans les journaux bio-médicaux et couvrant la biologie, la biochimie, la médecine clinique, la santé publique, l'éthique, l'économie, la pharmacologie, la psychiatrie, la toxicologie, la médecine vétérinaire etc… Son indexation utilise le thésaurus MeSH du 'National Library of Medicine', U.S.A.
NCBI : 'National Center for Biotechnology Information'
NIH : 'National Institute of Health'
NLM : 'National Library of Medicine'
PIR : 'Protein Information Resource'
PROSITE : banque de données spécialisée sur les familles protéiques et leurs domaines
PSD : 'Protein Sequence Database'
Requête : Un mot, une expression ou un groupe de mots employés pour passer des instructions à un moteur de recherche ou à un répertoire afin de localiser des pages sur le sujet recherché.
SRS : 'Sequence Retrieval System'. Logiciel développé par *Lion Bioscience Ltd* qui est une plate-forme robuste d'intégration de données. SRS permet d'accéder rapidement à des données biologiques et permet aussi de réaliser des interrogations croisées sur plusieurs banques de données. SRS contient aussi différents outils d'analyse de données biologiques.
Swiss Prot : banque de séquences protéiques
URL : 'Uniform Resource Locator'
XML: 'eXtensible Markup Language'
XPath: XPath voit un document XML comme un arbre de nœuds et définit une syntaxe (non-XML) pour identifier des noeuds et groupes de noeuds particuliers de documents XML.
XSL: 'eXtensible Stylesheet Language'
XSLT : 'eXtensible Stylesheet Language Transformation' : langage de programmation fonctionnel utilisé pour spécifier comment un document XML est transformé en un autre document texte, non nécessairement en un autre document XML.

Annexes

Annexe 1
Swiss-Prot: Q8CG79

NiceProt - a user-friendly view of this Swiss-Prot entry

En-tête

```
ID   ASP2_MOUSE     STANDARD;      PRT;  1088 AA.
AC   Q8CG79; Q8K2L5;
DT   10-OCT-2003 (Rel. 42, Created)
DT   10-OCT-2003 (Rel. 42, Last sequence update)
DT   05-JUL-2004 (Rel. 44, Last annotation update)
DE   Apoptosis stimulating of p53 protein 2 (Tumor suppressor p53-binding
DE   protein 2) (Fragment).
GN   Name=Trp53bp2; Synonyms=Aspp2;
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=FVB/N; TISSUE=Breast tumor;
RX   MEDLINE=22388257; PubMed=12477932 [NCBI, ExPASy, EBI, Israel, Japan];
DOI=10.1073/pnas.242603899;
RA   Strausberg R.L., Feingold E.A., Grouse L.H., Derge J.G.,
RA   Klausner R.D., Collins F.S., Wagner L., Shenmen C.M., Schuler G.D.,
RA   Altschul S.F., Zeeberg B., Buetow K.H., Schaefer C.F., Bhat N.K.,
RA   Hopkins R.F., Jordan H., Moore T., Max S.I., Wang J., Hsieh F.,
RA   Diatchenko L., Marusina K., Farmer A.A., Rubin G.M., Hong L.,
RA   Stapleton M., Soares M.B., Bonaldo M.F., Casavant T.L., Scheetz T.E.,
RA   Brownstein M.J., Usdin T.B., Toshiyuki S., Carninci P., Prange C.,
RA   Raha S.S., Loquellano N.A., Peters G.J., Abramson R.D., Mullahy S.J.,
RA   Bosak S.A., McEwan P.J., McKernan K.J., Malek J.A., Gunaratne P.H.,
RA   Richards S., Worley K.C., Hale S., Garcia A.M., Gay L.J., Hulyk S.W.,
RA   Villalon D.K., Muzny D.M., Sodergren E.J., Lu X., Gibbs R.A.,
RA   Fahey J., Helton E., Ketteman M., Madan A., Rodrigues S., Sanchez A.,
RA   Whiting M., Madan A., Young A.C., Shevchenko Y., Bouffard G.G.,
RA   Blakesley R.W., Touchman J.W., Green E.D., Dickson M.C.,
RA   Rodriguez A.C., Grimwood J., Schmutz J., Myers R.M.,
RA   Butterfield Y.S.N., Krzywinski M.I., Skalska U., Smailus D.E.,
RA   Schnerch A., Schein J.E., Jones S.J.M., Marra M.A.;
RT   "Generation and initial analysis of more than 15,000 full-length human
RT   and mouse cDNA sequences.";
RL   Proc. Natl. Acad. Sci. U.S.A. 99:16899-16903(2002).
```

Informations bibliographiques

```
CC   -!- FUNCTION: Regulator that plays a central role in regulation of
CC       apoptosis and cell growth via its interactions. Regulates p53/TP53
CC       by enhancing the DNA binding and transactivation function of
CC       p53/TP53 on the promoters of proapoptotic genes in vivo. Inhibits
CC       the ability of APPBP1 to conjugate NEDD8 to CUL1, and thereby
CC       decreases APPBP1 ability to induce apoptosis. Impedes cell cycle
CC       progression at G2/M (By similarity).
CC   -!- SUBUNIT: Binds to the central domain of p53/TP53 as well as to
CC       BCL2. Interacts with protein phosphatase 1. Interacts with RELA
CC       NF-kappa-B subunit. This interaction probably prevents the
CC       activation of apoptosis, possibly by preventing its interaction
CC       with p53/TP53. Interacts with APPBP1 (By similarity).
CC   -!- SUBCELLULAR LOCATION: Predominantly cytoplasmic; perinuclear
CC       region. Some small fraction is nuclear (By similarity).
CC   -!- DOMAIN: The ankyrin repeats and the SH3 domain are required for a
CC       specific interactions with p53/TP53 (By similarity).
CC   -!- SIMILARITY: Belongs to the ASPP family.
CC   -!- SIMILARITY: Contains 2 ANK repeats.
CC   -!- SIMILARITY: Contains 1 SH3 domain.
CC   -!- CAUTION: Ref.1 (AAH42874) sequence differs from that shown due to
CC       a frameshift in position 23.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; BC030894; AAH30894.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
```

Commentaires sur la séquence

```
DR   EMBL; BC042874; AAH42874.1; ALT_FRAME. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   PIR; PT0551; PT0551.
DR   HSSP; P08631; 1BU1. [HSSP ENTRY / SWISS-3DIMAGE / PDB]
DR   MGD; MGI:2138319; Trp53bp2.
DR   GeneLynx; Trp53bp2.
DR   Ensembl; Q8CG79. [Entry / Contig view]
DR   SOURCE; Trp53bp2.
DR   InterPro; IPR002110; ANK.
DR   InterPro; IPR001452; SH3.
DR   InterPro; Graphical view of domain structure.
DR   Pfam; PF00023; Ank; 2.
DR   Pfam; PF00018; SH3; 1.
DR   Pfam; Graphical view of domain structure.
DR   PRINTS; PR01415; ANKYRIN.
DR   ProDom; PD000066; SH3; 1.
DR   ProDom [Domain structure / List of seq. sharing at least 1 domain ]
DR   SMART; SM00248; ANK; 2.
DR   SMART; SM00326; SH3; 1.
DR   PROSITE; PS50088; ANK_REPEAT; 2.
DR   PROSITE; PS50297; ANK_REP_REGION; 1.
DR   PROSITE; PS50002; SH3; 1.
DR   HOVERGEN [Family / Alignment / Tree]
DR   BLOCKS; Q8CG79.
DR   ProtoNet; Q8CG79.
DR   ProtoMap; Q8CG79.
DR   PRESAGE; Q8CG79.
DR   DIP; Q8CG79.
DR   ModBase; Q8CG79.
DR   SMR; Q8CG79.
DR   SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW   ANK repeat; Apoptosis; Cell cycle; Repeat; SH3 domain; SH3-binding.
FT   NON_TER       1      1
FT   REPEAT      918    950       ANK 1.
FT   REPEAT      951    983       ANK 2.
FT   DOMAIN     1017   1079       SH3.
FT   DOMAIN      292    308       INTERACTION WITH APPBP1 (BY SIMILARITY).
FT   DOMAIN       92    133       Gln-rich.
FT   SITE        826    835       SH3-binding (Potential).
FT   CONFLICT    327    342       VKPALPDGSLLMQSAE -> DAWVAHASAHASAHAS (in
FT                                Ref. 1; AAH30894).
SQ   SEQUENCE   1088 AA;  120731 MW;  1023B229099BF3EC CRC64;
     NDCHLAEVWC GSERPVADNE RMFDVLQRFG SQRNEVRFFL RHERPPNRDI VSGPRSQDPS
     VKRNGVKVPG EHRRKENGVN SPRLDLTLAE LQEMASRQQQ QIEAQQQMLA TKEQRLKFLK
     QQDQRQQQQA AEQEKLKRLR EIAESQEAKL KKVRALKGHV EQKRLSNGKL VEEIEQMNSL
     FQQKQRELVL AVSKVEELTR QLEMLKNGRI DGHHDNQSAV AELDRLYKEL QLRNKLNQEQ
     NAKLQQQREC LNKRNSEVAV MDKRVSELRD RLWKKKAALQ QKENLPVSPD GNLPQQAVSA
     PSRVAAVGPY IQSSTMPRMP SRPELLVKPA LPDGSLLMQS AEGPMKIQTL PNMRSGAASQ
     SKGSKAHPAS PDWNPSNADL LPSQGSSVPQ SAGTALDQVD DGEIAVREKE KKVRPFSMFD
     TVDQCAAPPS FGTLRKNQSS EDILRDAQAV NKNVAKVPPP VPTKPKQIHL PYFGQTAQSP
     SDMKPDGNAQ QLPIAATSVG AKLKPAGPQA RMLLSPGAPS GGQDQVLSPA SKQESPPAAA
     VRPFTPQPSK DTFPPAFRKP … //
```

Références Croisées

Mots-clés

'Features'

SEQUENCE

Annexe 2

Une entrée de la banque EMBL (format EMBL) via SRS 7.1.1 (Serveur EBI)

```
ID   AF427192   standard; genomic RNA; VRL; 997 BP.
XX
AC   AF427192;
XX
SV   AF427192.1
XX
DT   14-NOV-2001 (Rel. 69, Created)
DT   11-AUG-2004 (Rel. 80, Last updated, Version 4)
XX
DE   HIV-1 isolate ccr54 from USA pol protein (pol) gene, partial cds.
XX
KW   .
XX
OS   Human immunodeficiency virus 1
OC   Viruses; Retroid viruses; Retroviridae; Lentivirus;
OC   Primate lentivirus group.
XX
RN   [1]
RP   1-997
RX   PUBMED; 15297515.
RA   Ellis G.M., Mahalanabis M., Beck I.A., Pepper G., Wright A., Hamilton S.,
RA   Holte S., Naugler W.E., Pawluk D.M., Li C.C., Frenkel L.M.;
RT   "Comparison of oligonucleotide ligation assay and consensus sequencing for
RT   detection of drug-resistant mutants of human immunodeficiency virus type 1
RT   in peripheral blood mononuclear cells and plasma";
RL   J. Clin. Microbiol. 42(8):3670-3674(2004).
XX
RN   [2]
RP   1-997
RA   Mahalanabis M., Pepper G., Wright A., Hamilton S., Beck I.A., Ellis G.,
RA   Naugler W.E., Frenkel L.M.;
RT   ;
RL   Submitted (01-OCT-2001) to the EMBL/GenBank/DDBJ databases.
RL   Laboratory Medicine, University Of Washington, Clinical Virology
RL   Laboratory, Children's Hospital, 4800 Sand Point Way NE, G-800A, Seattle,
RL   WA 98105, USA
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..997
FT                   /country="USA:Seattle, Washington"
FT                   /db_xref="taxon:11676"
FT                   /mol_type="genomic RNA"
FT                   /note="isolated from plasma"
FT                   /virion
FT                   /organism="Human immunodeficiency virus 1"
FT                   /isolate="ccr54"
FT   gene            <1..>997
FT                   /gene="pol"
FT   CDS             <1..>997
FT                   /codon_start=1
FT                   /db_xref="GOA:Q904N0"
FT                   /db_xref="UniProt/TrEMBL:Q904N0"
FT                   /note="contains protease and reverse transcriptase"
FT                   /gene="pol"
FT                   /product="pol protein"
FT                   /protein_id="AAL30537.1"
FT                   /translation="FSFPQITLWQRPIVTVKIGGQLKEALLDTGADDTVLEEMNLPGRW
FT                   KPKMIGGIGGFIKVRQYDQIXXEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFPI
FT                   SPIETVPVKLKPGMDGPRVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPVF
FT                   AIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKKKKSVTVLDVGDAYFSVPL
FT                   DEDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPFRKQNPDIV
FT                   IYQYMDDLYVGSDLEIGQHRAKIEELRKHLLAWGFTTPDKKHQKEPPFLWM"
XX
SQ   Sequence 997 BP; 375 A; 161 C; 217 G; 242 T; 2 other;
     tttagcttcc ctcagatcac tctttggcaa cgacccatcg tcacagtaaa gatagggggg        60
     cagctaaagg aagctctgtt agataccgga gcagatgata cagtactaga agaaatgaat       120
     ttgccaggaa gatggaagcc aaaaatgatt gggggaattg gaggttttat caaagtaaga       180
     cagtatgatc aaatatscat rgaaatctgt ggacataaag ctataggtac agtattagta       240
      ggacctacac ctgtcaacat aattggaaga aatctgttga ctc… //
```

En-tête

Informations bibliographiques

'Features'

SEQUENCE

Annexe 3

http://www.openweb.eu.org/articles/dublin_core/

Liste des éléments du Dublin Core

Élément	Description et liste des *raffinements*
title	**Titre du document** : il s'agit a priori du titre principal du document. Pour indiquer un autre type, on peut utiliser le *raffinement* suivant : • **alternative** : alternative pour le titre, par exemple une abréviation ou une traduction.
creator	**Créateur du document** : nom de la personne, de l'organisation ou du service à l'origine de la rédaction du document.
subject	**Sujet et mots-clefs** : mots-clefs, phrases de résumé, ou codes de classement. Il est préférable d'utiliser des mots-clefs choisis dans le cadre d'une politique de classement. Par exemple, il est recommandé d'utiliser les codages de la bibliothèque du congrès (LCSH et LCC), le vocabulaire médical (MESH), ou les notations décimales des bibliothécaires (DDC et UDC).
description	**Description du document** : résumé, table des matières, ou texte libre. Le type de description peut être précisé à l'aide des *raffinements* suivants : • **tableOfContents** : table des matières ; • **abstract** : résumé.
publisher	**Publicateur du document** : nom de la personne, de l'organisation ou du service à l'origine de la publication du document.
contributor	**Contributeur au document** : nom d'une personne, d'une organisation ou d'un service qui contribue ou a contribué à l'élaboration du document.
date	**Date d'un évènement dans le cycle de vie du document** : il peut s'agir par exemple de la date de création ou de la date de mise à disposition. Il est recommandé de spécifier la date au format W3CDTF (AAAA-MM-JJ). Pour préciser de quelle date il s'agit, on utilise les *raffinements* suivants : • **created** : date de création ; • **valid** : date ou période de validité ; • **available** : date ou période de mise à disposition ; • **issued** : date de publication ; • **modified** : date de modification ; • **dateAccepted** : date d'acceptation (par exemple, acceptation d'une thèse par une université, d'un article par un journal, etc.) ; • **dateCopyrighted** : date du copyright ; • **dateSubmitted** : date où le document a été soumis (par exemple, soumis à un comité de lecture s'il s'agit d'un article).
type	**Nature ou genre du contenu** : grandes catégories de document. Il est recommandé d'utiliser des termes clairement définis au sein de son organisation. Par exemple, le Dublin Core définit quelques types dans le vocabulaire DCMITypes.
format	**Format du document** : format physique ou électronique du document. Par exemple, type de média ou dimensions (taille, durée). On peut spécifier le matériel et le logiciel nécessaires pour accéder au document. Il est recommandé d'utiliser des termes clairement définis, par exemple les types MIME ou IMT . Les *raffinements* suivants sont disponibles : • **extent** : taille ou durée ; • **medium** : support physique.
identifier	**Identificateur non ambigu** : il est recommandé d'utiliser un système de référencement précis, par

Liste des éléments du Dublin Core	
Élément	**Description et liste des *raffinements***
	exemple les URI ou les numéros ISBN.
source	**Ressource dont dérive le document** : le document peut découler en totalité ou en partie de la ressource en question. Il est recommandé d'utiliser une dénomination formelle des ressources, par exemple leur URI.
language	**Langue du document** : il est recommandé d'utiliser un code de langue conforme au format RFC3066.
relation	**Lien vers une ressource liée** : il est recommandé d'utiliser une dénomination formelle des ressources, par exemple leur URI. On précise le type de lien avec des *raffinements* : • **isVersionOf** : on a affaire à une nouvelle version, une modification ou une adaptation du document lié. Les changements concernent le contenu et pas seulement la forme ; • **hasVersion** : réciproque d'isVersionOf. Le document lié est une version modifiée du présent document ; • **isReplacedBy** : le présent document a été remplacé par le document lié ; • **replaces** : réciproque de replaces. Le présent document remplace le document lié ; • **isRequiredBy** : on a besoin du présent document pour interpréter correctement le document lié ; • **requires** : réciproque d'isRequiredBy. Le présent document a besoin du document lié pour être correctement présenté, transmis, ou pour assurer sa cohérence ; • **isPartOf** : le document est une partie (physique ou logique) d'un autre document ; • **hasPart** : réciproque d'isPartOf : le document inclut le document lié, physiquement ou logiquement ; • **isReferencedBy** : le document courant est référencé, cité, ou lié par le document indiqué ; • **references** : réciproque d'isReferencedBy : le document courant référence, cite ou pointe vers le document indiqué ; • **isFormatOf** : le présent document a le même fond que le document indiqué, mais présenté sous une forme différente ; • **hasFormat** : réciproque d'isFormatOf : le présent document possède une variante sous une forme différente ; • **conformsTo** : référence à un standard établi auquel se conforme le présent document.
coverage	**Portée du document** : la portée inclut un domaine géographique, un laps de temps, ou une juridiction (nom d'une entité administrative). Il est recommandé d'utiliser des représentations normalisées de ces types de données. Le type de couverture peut être précisé : • **spatial** : couverture spatiale. On peut utiliser les codages Point (point géographique), ISO3166 (codes de pays à deux lettres), Box (régions géographiques), ou TGN (dictionnaire de noms de lieux) ; • **temporal** : couverture temporelle. On peut utiliser les codages Period (intervalle de temps) ou W3CDTF (dates).
rights	**Droits relatifs à la ressource** : permet de donner des informations sur le statut des droits du document, par exemple la présence d'un copyright, ou un lien vers le détenteur des droits. L'absence de cet élément ne présume pas que le document est libre de droits.
Audience	**Audience du document** : l'audience représente le groupe de personnes à qui le document est destiné. L'audience est déterminée par l'auteur, le publicateur, ou un tiers. On peut utiliser les *raffinements* suivants : • **mediator** : entité qui sert d'intermédiaire pour l'accès au document ; • **educationLevel** : position du niveau de l'audience par rapport à un contexte d'éducation ou

Liste des éléments du Dublin Core	
Élément	**Description et liste des *raffinements***
	de formation.

Annexe 4

Liste des sources explorées :

- **Sources Nulcéotidiques :**

1. EMBL
2. EnsEMBL Human
3. EnsEMBL Mouse
4. EnsEMBL Fly
5. EnsEMBL Fish
6. IMGT/HLA
7. IMGT/LIGM-DB
8. IMGT/ GENE-DB
9. TIGR Animal Gene Indices
10. TIGR Human Gene Index
11. TIGR Plant Gene Indices
12. TIGR *Arabidospsis* Gene Index
13. TIGR Protist Gene Indices
14. TIGR *P.falciparum* Gene Index
15. TIGR Fungal Gene Indices
16. TIGR *S.cerevisiae* Gene Index
17. UniGene
18. Xpro
19. GenBank

- **Sources Protéiques :**

1. EXProt
2. NCBI Protein DB
3. PDB
4. PIR-PSD
5. iProClass
6. PIR-NREF
7. PRF/LITTDB
8. PRF/SEQDB
9. PRF/SYNDB
10. UniProt
11. UniProt/Swiss-Prot
12. UniProt/TrEMBL

- **Sources Spécialisées :**

1. NCBI Taxonomy
2. NCBI RefSeq
3. DEG
4. KEGG Genes DB
5. KEGG Pathway DB
6. KEGG Ligand DB
7. KEGG Expression DB
8. BRITE
9. KEGG-SSDB
10. Transport DB
11. coliBase
12. Colibri
13. Essential genes in *E.coli*
14. EcoCyc
15. EcoGene
16. SGD
17. CYGD
18. Génolevures
19. Sputnik
20. TropGene DB
21. ARAMEMNON
22. MAtDB
23. TAIR
24. MosDB
25. RPD
26. WormBase
27. FlyBase
28. GPCRDB
29. GeneCards
30. MTB
31. ENZYME

Annexe 5

Un exemple de fiche **DBCat** décrivant la source 'ENZYME'

```
AC                DBC00005
NAME              ENZYME nomenclature database
DOMAIN            Proteins
DESCRIPTION       A repository of information relative to the nomenclature of
DESCRIPTION       enzymes
CHECKED  YES
AUTHOR            Bairoch A.
RA                Bairoch A.
RT                "The ENZYME data bank in 1995.";
RL                Nucleic Acids Res. 24:221-222(1996).
RX                SeqAnalRef: BAIA9603 .
ORIGINAL-SITE     Medical Biochemistry Department
ADDRESS  Centre Medical Universitaire
ADDRESS  1, Rue Michel Servet
ADDRESS  1211 Geneva 4
ADDRESS  SWITZERLAND
CONTACT           bairoch@medecine.unige.ch
SUBMIT            http://expasy.proteome.org.au/enzyme/enz_new_form.html
URL-FTP           ftp://ftp.expasy.ch/databases/enzyme
URL-WWW  http://www.expasy.ch/enzyme
URL-QUERY         http://www.expasy.ch/enzyme
RELEASE  4 times a year
UPDATES  -
COMMENTS -
OTHER-SITE        INFOBIOGEN
ADDRESS  523 place des Terrasses de l'Agora
ADDRESS            91034 Evry Cedex
ADDRESS            FRANCE
URL-FTP           ftp://ftp.infobiogen.fr/pub/db/enzyme/
URL-WWW  -
URL-QUERY         http://www.infobiogen.fr/srs
UPDATES  -
COMMENTS          Formats: swissprot, index SRS
OTHER-SITE        EMBL Outstation, EBI
ADDRESS          European Bioinformatics Institute
ADDRESS          Hinxton Hall
ADDRESS          Hinxton, Cambridge CB10 1SD
ADDRESS          U.K.
URL-FTP             ftp://ftp.ebi.ac.uk/pub/databases/enzyme
URL-WWW       -
URL-QUERY         -
UPDATES           -
COMMENTS         Formats: index srs
OTHER-SITE       NCBI, National Center for Biotechnology Information
ADDRESS          National Library of Medicine, 38A, 8N805
ADDRESS          8600 Rockville Pike
ADDRESS          Bethesda, MD  20894
ADDRESS          U.S.A.
URL-FTP           ftp://ncbi.nlm.nih.gov/repository/enzyme
URL-WWW  -
URL-QUERY         -
UPDATES           -
COMMENTS -
```

Annexe 6
Content for BioRegistry Meta-data

Key:

- DS= Data Source
- *Definitions*
- 'possible values'
- [+] = can be repeated unlimited times
- Optional field
- Dublin Core (DC) Element Name Version 1.1 (if applicable)
- *DC Definition*
- (Pointer)

Part A. Source Meta-Data

❖ BioRegistry Identifier of DS (DC Identifier : *An unambiguous reference to the resource within a given context*)

Section A1 : Identification Information:

Basic information about the data source

A1.1 : Data Source

A1.1.1: DS Name (DC Title : *A name given to the resource, typically the formal name*)
A1.1.1.1 : Full Name
A1.1.1.2 : Common Name
A1.1.1.2 : Acronym : *Acronym used to reference the DS*
A1.1.2 : DS Contact [+]
A1.1.2.1 : Contact Person
A1.1.2.2 : Contact Address
A1.1.2.3 : Country
A1.1.2.4 : Contact email
A1.1.2.5 : Contact Telephone

A1.2 : Citation : *information to be used to reference the data source* [+]

A1.2.1 : Author(s)
A1.2.2 : Publication Year: *Year during which article is published*
A1.2.3 : Title: *Title of article*
A1.2.4 : Journal Information
A1.2.4.1 : Journal Name
A1.2.4.2 : Issue Identification
A1.2.4.2.1 : Issue Volume
A1.2.4.2.2 : Issue Pages
A1.2.5 : PubMed Identifier
A1.2.6 : On-line link (URL)

A1.3 : Description (DC Description : *An account of the content of the resource*)

A1.3.1 : Abstract: *characterisation of the DS, including its intended use and limitations*
A1.3.2 : Purpose: *intentions with which the DS was developed*
A1.3.3 : Supplemental information: *other descriptive information about the DS*
A1.3.4 : Entry Sample : *a sample of an entry in the DS*

A1.4 : Time Period of Content

A1.4.1: Time Period Information: *time period(s) for which the DS corresponds to the currentness reference.*
A1.4.1.1: Beginning of Time Period
A1.4.1.2: End of Time Period

A1.5 : Maintenance Status

A1.5.1 : Update (DC Date : *A date of an event in the lifecycle of the resource*)
A1.5.1.1: Update Frequency
A1.5.1.2: Date of Last Update
A1.5.2 : Release (DC Date : *A date of an event in the lifecycle of the resource*)
A1.5.2.1 : Release Frequency
A1.5.2.2 : Date of First Release
A1.5.2.2 : Date of Latest Release
A1.5.2.3 : Latest Release Identifier

A1.5.3 : Maintenance Organisation [+] (DC Publisher: *An entity responsible for making the resource available ex: a person, an organisation*)

A1.5.3.1 : Maintenance Name
A1.5.3.2 : Maintenance Address
A1.5.3.3 : Maintenance Country
A1.5.3.4 : Maintenance e-mail
A1.5.3.5 : Maintenance Telephone

Section A2 : Topic Information

A2.1: Keywords : *Words or phrases summarizing an aspect of the data source*
(DC Subject : *A topic of the content of the resource*)

A2.1.1 : Subject [+]
A2.1.1.1 : Thesaurus/Ontology: *reference to a thesaurus name* (with reference to Part B, Section B1.1.1)
A2.1.1.2 : Subject keyword: *words defining the source contents* [+]
A2.1.1.2.1 : Term
A2.1.1.2.2 : Identifier [+]

A2.1.2 : Organism
A2.1.2.1 : Organism Thesaurus / Ontology: *reference to an approved thesaurus encompassing organisms*
A2.1.2.2 : Organism Names or Keywords : *words defining the organisms in the DS* [+]
A2.1.2.2.1 : Term (cf. Part A, Section A4.4.3)
A2.1.2.2.2 : Identifier

A2.2 : Use Constraints [+] :
(DC Rights : *Information about rights held in and over the resource*)
Restrictions and legal prerequisites for using the data source after access is granted. These include any use constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on using the data source.

A2.3 : Help Desk :
A2.3.1 : e-mail
A2.3.2 : Name
A2.3.3 : Telephone

Section A3 : Meta-Data Tracking Information

A3.1 : Meta-data tracking [+] :
Information about meta-data origin

A3.1.1 : Meta-data Item: *any possible item such as 'Manual Revision', 'Description', 'All Meta-data items'...*
A3.1.2 : Creator: *name of person responsible for creating the meta-data*
A3.1.3 : Creation Date: *date at which the meta-data was first created*
A3.1.4 : Reviewer: *name of person undertaking the review of the meta-data*
A3.1.5 : Review Date: *date at which the meta-data is reviewed*
A3.1.6 : Status: 'Reviewed', 'Not Reviewed', 'Ongoing'

Section A4 : Data Quality Information

A general assessment of the quality of the data source.

A4.1 : Manual Revision:
Information concerning the manual revision of the DS

A4.1.1 : Status: 'Yes', 'No', 'Not Found'
A4.1.2 : Supplemental Information

A4.2 : Standard compliancy [+]: *ex: MIAME compliant, ASN.1 compatible*

A4.2.1 : Standard Name
A4.2.2 : Standard Reference (URL)
A4.2.3 : Status: 'Total', 'Partial', 'Not Documented', 'Verified', 'Not Verified'
A4.2.4 : Supplemental Information

A4.3 : Data Reference Information : *information about data origin in the DS*

A4.3.1 : On-line Documentation Existence **[+]** : *description of any on-line reference information*

A4.3.1.1 : Documentation Quality : 'Rich', 'Poor', 'Intermediate'

A4.3.1.2 : Documentation Availability : 'Available', 'Obsolete', 'Out-dated', 'Under construction'

A4.3.2 : Existence of Reference to Publication in DS Entries

A4.3.3 : Supplemental Information

A4.4 : Source Coverage [+] : *Number of different types of entities in the data source*

A4.4.1 : Entity Name : *type of entity (ex: Genes, EST, EST Cluster, Contig etc. possibly associated to the organism)*

A4.4.2 : Number of Entities : *number of entities of this type in the DS*

A4.4.3 : Organism Name (with reference to Section A2.1.2.2.1)

A4.5 : Cross Reference [+] :

(DC Relation : *A reference to a related resource*)

List of data sources cross referenced in the DS

Section A5 : Availability Information

Information about the availability of the DS

A5.1 : Site of Access :

The site from where the DS may be accessed

A5.1.1 : Site Identification **[+]**

A5.1.1.1 : URL

A5.1.1.2 : Organisation (ex: EBI, Infobiogen in the case of SRS…)

(DC Publisher : *An entity responsible for making the resource available*)

A5.1.1.3 : Status: 'Main site' or 'Alternative site'

A5.1.2 : DS version : *version of the DS in the site of access*

A5.1.3 : Interaction Modalities: *Requirements for source usage (ex: Query language, Web services…)*

A5.2 : Access Constraints:

A5.2.1 : Access Constraints for Academics: 'Free' or 'Password' or 'Registration fees'

A5.2.2 : Access Constraints for Industrials: 'Free' or 'Password' or 'Registration fees'

****

Part B. Thesaurus / Ontology / Controlled Vocabulary

Section B1 : Ontology Identification Information

B1.1 : Thesaurus / Ontology/Controlled Vocabulary [+]

B1.1.1 : Thesaurus Name *ex: MeSH* (cf. Part A, Section A2.1.1.1)

B1.1.2 : Thesaurus Version *ex:2004*

B1.1.3 : Thesaurus URL

B1.1.4 : Supplemental Information

****

Part C. Relationships

Section C1 : Data Source Relationships

C1.1 : DS content coverage [+]

Relationships, in terms of content coverage, of the DS with other data sources

C1.1.1 : Source Name 1 : *Actual DS*

C1.1.2 : Source Name 2 : *Source from which the DS draws information from or contributes information to.*

C1.1.3 : Type of Relationship : 'Draws from', 'Contributes to'

C1.1.4 : Relationship Extent : 'Complete', Partial', 'Unknown'

C1.1.5 : Field Map: *Field in the DS entry corresponding to the relation source*

****

Annexe 7

Schéma XML de BioRegistry (.xsd)

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema targetNamespace="BRSchema" elementFormDefault="qualified" xmlns:br="BRSchema"
xmlns:xsd="http://www.w3.org/2001/XMLSchema">
   <xsd:annotation>
      <xsd:documentation>XML Schema For BioRegistry</xsd:documentation>
   </xsd:annotation>
   <!--DC corresponds to the Dublin Core Elements (version 1.1) with their corresponding definitions,  where applicable  -->
   <!********************************************************************** -->
   <xsd:element name="BioRegistry" type="br:typeBioRegistry"/>
   <xsd:annotation>
      <xsd:documentation>Root Element</xsd:documentation>
   </xsd:annotation>
   <xsd:complexType name="typeBioRegistry">
      <xsd:sequence>
         <xsd:element name="Ontology" type="br:typeOnto" maxOccurs="unbounded">
            <!--cf PART B -->
            <!--declaration contraintes referentielles de BROntoName -->
            <xsd:key name="cleOnto">
               <xsd:selector xpath="BioRegistry/Ontology"/>
               <!--indique tous les elements concernes-->
               <xsd:field xpath="./BROntoName"/>
               <!--indique le champ concerne, BROntoName -->
            </xsd:key>
            <xsd:keyref name="refBROntoName1" refer="br:cleOnto">
               <!--referencie nom ontologie pour les subjets -->
               <xsd:selector xpath="BioRegistry/SourceMetaData/TopicInformation/Subjects"/>
               <xsd:field xpath="./refBROntoNameS"/>
            </xsd:keyref>
            <xsd:keyref name="refBROntoName2" refer="br:cleOnto">
               <!--referencie nom ontologie pour les organismes -->
               <xsd:selector xpath="BioRegistry/SourceMetaData/TopicInformation/Organism"/>
               <xsd:field xpath="./refBROntoNameO"/>
            </xsd:keyref>
         </xsd:element>
         <xsd:element name="SourceMetaData" type="br:typeMetaData" maxOccurs="unbounded">
            <!--declaration contraintes referentielles de BRSourceNumber -->
            <xsd:key name="cleSource">
               <xsd:selector xpath="BioRegistry/SourceMetaData"/>
               <!--indique les elements concernes -->
               <xsd:field xpath="@BRSourceNumber"/>
               <!--indique le champ concerne, BRSourceNumber -->
            </xsd:key>
         </xsd:element>
      </xsd:sequence>
      <xsd:attribute name="BRupdate" type="xsd:date" use="optional"/>
      <!--attributs -->
   </xsd:complexType>
   <xsd:complexType name="typeMetaData">
      <xsd:sequence>
         <xsd:element name="IdentificationInformation" type="br:typeIdentificationInfo"/>
         <xsd:element name="TopicInformation" type="br:typeTopicInfo"/>
         <xsd:element name="MetaDataTrackingInformation" type="br:typeMDTrackingInfo"/>
         <xsd:element name="DataQualityInformation" type="br:typeDQInfo"/>
         <xsd:element name="AvailabilityInformation" type="br:typeAvailabilityInfo"/>
      </xsd:sequence>
      <xsd:attribute name="BRSourceNumber" type="xsd:positiveInteger" use="required"/>
      <!--Part A -->
      <!--attributs -->
   </xsd:complexType>
   <xsd:complexType name="typeIdentificationInfo">
      <xsd:annotation>
         <xsd:documentation>Basic information about the Data Source</xsd:documentation>
      </xsd:annotation>
      <xsd:sequence>
         <xsd:element name="DS" type="br:typeDS"/>
         <xsd:element name="Citation" type="br:typeCitation" minOccurs="0" maxOccurs="unbounded"/>
         <xsd:element name="Description" type="br:typeDescription"/>
         <xsd:element name="TimePeriodofContent" type="br:typeTPContent" minOccurs="0"/>
         <xsd:element name="MaintenanceStatus" type="br:typeMaintenanceStatus"/>
      </xsd:sequence>
```

```
    <!–Section A1 –>
  </xsd:complexType>
  <xsd:complexType name="typeDS">
    <xsd:sequence>
      <xsd:element name="DSName" type="br:typeDSName"/>
      <xsd:element name="DSContact" type="br:typeDSContact" maxOccurs="unbounded"/>
    </xsd:sequence>
    <!–Section A1.1 –>
  </xsd:complexType>
  <xsd:complexType name="typeDSName">
    <xsd:annotation>
      <xsd:documentation>DCTitle : A name given to the resource, typically the formal name</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="FullName" type="xsd:string"/>
      <xsd:element name="CommonName" type="xsd:string"/>
      <xsd:element name="Acronym" type="xsd:string"/>
    </xsd:sequence>
    <!–Section A1.1.1 –>
  </xsd:complexType>
  <xsd:complexType name="typeDSContact">
    <xsd:sequence>
      <xsd:element name="ContactPerson" type="xsd:string" minOccurs="0" maxOccurs="3"/>
      <xsd:element name="ContactAddress" type="xsd:string" minOccurs="0" maxOccurs="3"/>
      <xsd:element name="Country" type="xsd:string" maxOccurs="3"/>
      <xsd:element name="Contactphone" type="xsd:string" minOccurs="0" maxOccurs="3"/>
      <xsd:element name="Contactmail" maxOccurs="3">
        <xsd:simpleType>
          <xsd:restriction base="xsd:string">
            <xsd:pattern value="(.)+@(.)+"/>
            <!–email utilisant un type anonyme –>
          </xsd:restriction>
        </xsd:simpleType>
      </xsd:element>
    </xsd:sequence>
    <!–Section A1.1.2 –>
  </xsd:complexType>
  <xsd:complexType name="typeCitation">
    <xsd:annotation>
      <xsd:documentation>Information to be used to reference the data source </xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Authors" type="xsd:string"/>
      <xsd:element name="PublicationYear" type="xsd:gYear"/>
      <!–xsd:gYear=gregorian yr forme CCYY et suffixe fuseau horaire optionnel hh:mm–>
      <xsd:element name="Title" type="xsd:string"/>
      <xsd:element name="JournalInfo" type="br:typeJournalInfo"/>
      <xsd:element name="PMID" type="xsd:string" minOccurs="0"/>
      <!–Identifiant PubMed –>
      <xsd:element name="OnlineLink" type="xsd:anyURI" minOccurs="0"/>
    </xsd:sequence>
    <!–Section A1.2 –>
  </xsd:complexType>
  <xsd:complexType name="typeJournalInfo">
    <xsd:all>
      <xsd:element name="JournalName" type="xsd:string"/>
      <xsd:element name="IssueIdentification" type="br:IssueID"/>
    </xsd:all>
    <!–Section A1.2.4 –>
  </xsd:complexType>
  <xsd:complexType name="IssueID">
    <xsd:all>
      <xsd:element name="IssueVol" type="xsd:string"/>
      <xsd:element name="IssuePages" type="xsd:string" minOccurs="0"/>
    </xsd:all>
  </xsd:complexType>
  <xsd:complexType name="typeDescription">
    <xsd:annotation>
      <xsd:documentation>DC Description : An account of the content of the resource</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Abstract" type="xsd:string"/>
      <xsd:element name="Purpose" type="xsd:string"/>
      <xsd:element name="SupplementalInfo" type="xsd:string" minOccurs="0"/>
      <xsd:element name="EntrySample" type="xsd:anyURI" minOccurs="0"/>
    </xsd:sequence>
```

```
    <!–Section A1.3 –>
  </xsd:complexType>
  <xsd:complexType name="typeTPContent">
    <xsd:annotation>
      <xsd:documentation> Time Period Information : time period(s) for which the DS
        corresponds to the currentness reference </xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Beginningtimeperiod" type="xsd:gYear"/>
      <xsd:element name="Endtimeperiod" type="xsd:gYear"/>
    </xsd:sequence>
    <!–Section A1.4 –>
  </xsd:complexType>
  <xsd:complexType name="typeMaintenanceStatus">
    <xsd:sequence>
      <xsd:element name="Update" type="br:typeUpdate"/>
      <xsd:element name="Release" type="br:typeRelease"/>
      <xsd:element name="MaintenanceOrganisation" type="br:typeMaintOrg" maxOccurs="unbounded"/>
    </xsd:sequence>
    <!–Section A1.5 –>
  </xsd:complexType>
  <xsd:complexType name="typeUpdate">
    <xsd:all>
      <xsd:element name="UpdateFrequency" type="br:typeUpdateFreq"/>
      <xsd:element name="datelastupdate" type="xsd:date" minOccurs="0"/>
    </xsd:all>
    <!–Section A1.5.1 –>
  </xsd:complexType>
  <xsd:simpleType name="typeUpdateFreq">
    <xsd:restriction base="xsd:string">
      <xsd:enumeration value="annual"/>
      <xsd:enumeration value="bi-annually"/>
      <xsd:enumeration value="6 times per year"/>
      <xsd:enumeration value="three times per year"/>
      <xsd:enumeration value="3-4 times per year"/>
      <xsd:enumeration value="every 3 months"/>
      <xsd:enumeration value="every 2 months"/>
      <xsd:enumeration value="monthly"/>
      <xsd:enumeration value="bi-monthly"/>
      <xsd:enumeration value="at least monthly"/>
      <xsd:enumeration value="quarterly"/>
      <xsd:enumeration value="fortnightly"/>
      <xsd:enumeration value="weekly"/>
      <xsd:enumeration value="bi-weekly"/>
      <xsd:enumeration value="daily"/>
      <xsd:enumeration value="unknown"/>
    </xsd:restriction>
  </xsd:simpleType>
  <xsd:complexType name="typeRelease">
    <xsd:sequence>
      <xsd:element name="ReleaseFrequency" type="br:RelFreq"/>
      <xsd:element name="DateFirstRelease" type="xsd:date" minOccurs="0"/>
      <xsd:element name="DateLatestRelease" type="xsd:date" minOccurs="0"/>
      <!–forme CCYY-MM-DD –>
      <xsd:element name="LatestReleaseID" type="xsd:string" minOccurs="0"/>
    </xsd:sequence>
    <!–Section A1.5.2 –>
  </xsd:complexType>
  <xsd:simpleType name="RelFreq">
    <xsd:restriction base="xsd:string">
      <xsd:enumeration value="annual"/>
      <xsd:enumeration value="bi-annually"/>
      <xsd:enumeration value="6 times per year"/>
      <xsd:enumeration value="three times per year"/>
      <xsd:enumeration value="3-4 times per year"/>
      <xsd:enumeration value="every 3 months"/>
      <xsd:enumeration value="every 2 months"/>
      <xsd:enumeration value="monthly"/>
      <xsd:enumeration value="bi-monthly"/>
      <xsd:enumeration value="at least monthly"/>
      <xsd:enumeration value="quarterly"/>
      <xsd:enumeration value="fortnightly"/>
      <xsd:enumeration value="weekly"/>
      <xsd:enumeration value="bi-weekly"/>
      <xsd:enumeration value="daily"/>
      <xsd:enumeration value="unknown"/>
```

```
    </xsd:restriction>
  </xsd:simpleType>
  <xsd:complexType name="typeMaintOrg">
    <xsd:sequence>
      <xsd:element name="MaintenanceName" type="xsd:string"/>
      <xsd:element name="MaintenanceAddress" type="xsd:string" minOccurs="0"/>
      <xsd:element name="MaintenanceCountry" type="xsd:string"/>
      <xsd:element name="Maintenancemail" minOccurs="0" maxOccurs="3">
        <xsd:simpleType>
          <xsd:restriction base="xsd:string">
            <xsd:pattern value="(.)+@(.)+"/>
            <!--email utilisant un type anonyme -->
          </xsd:restriction>
        </xsd:simpleType>
      </xsd:element>
      <xsd:element name="MaintenanceTelephone" type="xsd:string" minOccurs="0"/>
    </xsd:sequence>
    <!--Section A2 -->
  </xsd:complexType>
  <xsd:complexType name="typeTopicInfo">
    <xsd:sequence>
      <xsd:element name="Subjects" type="br:typeSubjects" maxOccurs="unbounded"/>
      <xsd:element name="Organism" type="br:typeOrganism"/>
      <xsd:element name="UseConstraints" type="xsd:string" minOccurs="0" maxOccurs="unbounded"/>
      <xsd:element name="HelpDesk" type="br:typeHelpDesk" minOccurs="0"/>
    </xsd:sequence>
    <!--Section A2.1 -->
  </xsd:complexType>
  <xsd:complexType name="typeSubjects">
    <xsd:annotation>
      <xsd:documentation>Words or phrases summarising an aspect of the DS; DC Subject : A
        topic of the content of the resource</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="refBROntoNameS" type="xsd:string"/>
      <!--cle etrangere -->
      <xsd:element name="Keywords" type="br:typeKeywds" maxOccurs="unbounded"/>
    </xsd:sequence>
    <xsd:attribute name="OntoNumber" type="xsd:string" use="optional"/> <!--declaration de l'attribut OntoNumber  -->
  </xsd:complexType>
  <!--Section A2.1.1 -->
  <xsd:complexType name="typeKeywds">
    <xsd:annotation>
      <xsd:documentation>Words defining the source contents</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="Term" type="xsd:string"/>
      <xsd:element name="Identifier" type="xsd:string" maxOccurs="unbounded"/>
    </xsd:sequence>
    <!--Subject Keywords -->
  </xsd:complexType>
  <xsd:complexType name="typeOrganism">
    <xsd:sequence>
      <xsd:element name="refBROntoNameO" type="xsd:string"/>
      <!--la cle etrangere -->
      <xsd:element name="OrganismName" type="br:typeOrgName" maxOccurs="unbounded"/>
    </xsd:sequence>
    <!--Section A2.1.2 -->
  </xsd:complexType>
  <xsd:complexType name="typeOrgName">
    <xsd:annotation>
      <xsd:documentation>Words defining the organisms in the data source</xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="OrgTerm" type="xsd:string"/>
      <xsd:element name="OrgIdentifier" type="xsd:string"/>
    </xsd:sequence>
    <!--Section A2.1.2.2 -->
  </xsd:complexType>
  <xsd:complexType name="typeHelpDesk">
    <xsd:all>
      <xsd:element name="Name" type="xsd:string" minOccurs="0"/>
      <xsd:element name="Telephone" type="xsd:string" minOccurs="0"/>
      <xsd:element name="Helpmail">
        <xsd:simpleType>
          <xsd:restriction base="xsd:string">
```

```
                <xsd:pattern value="(.)+@(.)+"/>
                <!--email utilisant un type anonyme -->
              </xsd:restriction>
            </xsd:simpleType>
          </xsd:element>
        </xsd:all>
        <!--SectionA2.3 -->
      </xsd:complexType>
      <xsd:complexType name="typeMDTrackingInfo">
        <xsd:annotation>
          <xsd:documentation>Information about meta-data origin</xsd:documentation>
        </xsd:annotation>
        <xsd:sequence>
          <xsd:element name="MDItem" type="br:typeMDItem" maxOccurs="unbounded"/>
          <!--enumeration des differents items -->
        </xsd:sequence>
        <!--SectionA3 -->
      </xsd:complexType>
      <xsd:complexType name="typeMDItem">
        <xsd:sequence>
          <xsd:element name="allMDItems" type="br:allMDTracktype"/>
          <xsd:element name="DataSource" type="br:DSTracktype"/>
          <xsd:element name="Citation" type="br:CitationTracktype" minOccurs="0"/>
          <xsd:element name="Description" type="br:DescriptionTracktype"/>
          <xsd:element name="ContentTimePeriod" type="br:ContentTracktype" minOccurs="0"/>
          <!--Time period of content tracking -->
          <xsd:element name="MaintenanceStatus" type="br:MaintenanceTracktype"/>
          <xsd:element name="Subject" type="br:SubjectTrackType"/>
          <xsd:element name="Organism" type="br:OrgTrackType"/>
          <xsd:element name="UseConstraints" type="br:ConstraintsTracktype" minOccurs="0"/>
          <xsd:element name="HelpDesk" type="br:HDTracktype" minOccurs="0"/>
          <xsd:element name="ManualRevision" type="br:RevisionTracktype" minOccurs="0"/>
          <!--Manual Revision Tracking-->
          <xsd:element name="StandardCompliancy" type="br:CompliancyTracktype" minOccurs="0"/>
          <!--Std Compliancy tracking -->
          <xsd:element name="DataReferenceInfo" type="br:ReferenceTracktype" minOccurs="0"/>
          <!--Data Reference Info tracking -->
          <xsd:element name="SourceCoverage" type="br:CoverageTracktype"/>
          <!--Source Coverage tracking -->
          <xsd:element name="CrossReference" type="br:CrossRefTracktype" minOccurs="0"/>
          <!--Cross reference tracking -->
          <xsd:element name="SiteofAccess" type="br:SiteAccessTrackType"/>
          <!--Site of Access tracking -->
          <xsd:element name="AccessConstraints" type="br:AccessTrackType" minOccurs="0"/>
          <!--Access Constraints tracking -->
        </xsd:sequence>
        <!--Section A3.1.1 -->
      </xsd:complexType>
      <!--definition de chaque type de tracking de metadonnes -->
      <xsd:complexType name="allMDTracktype">
        <xsd:all>
          <xsd:element name="Creator" type="xsd:string"/>
          <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
          <!--annee et mois format CCYY-MM -->
          <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
          <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
          <xsd:element name="Status" type="br:StatusTrack"/>
        </xsd:all>
      </xsd:complexType>
      <xsd:complexType name="DSTracktype">
        <xsd:all>
          <xsd:element name="Creator" type="xsd:string"/>
          <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
          <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
          <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
           <xsd:element name="Status" type="br:StatusTrack"/>
        </xsd:all>
      </xsd:complexType>
      <xsd:complexType name="CitationTracktype">
        <xsd:all>
          <xsd:element name="Creator" type="xsd:string"/>
          <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
          <!--annee et mois format CCYY-MM -->
          <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
          <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
          <xsd:element name="Status" type="br:StatusTrack"/>
```

```
    </xsd:all>
  </xsd:complexType>
  <xsd:complexType name="DescriptionTracktype">
    <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <!--annee et mois format CCYY-MM -->
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
    </xsd:all>
  </xsd:complexType>
  <xsd:complexType name="ContentTracktype">
    <xsd:all>
      <xsd:element name="Creator" type="xsd:string" minOccurs="0"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth" minOccurs="0"/>
      <!--annee et mois format CCYY-MM -->
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
    </xsd:all>
    <!--Time period of content tracking -->
  </xsd:complexType>
  <xsd:complexType name="MaintenanceTracktype">
    <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
    </xsd:all>
  </xsd:complexType>
  <xsd:complexType name="SubjectTrackType">
    <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
    </xsd:all>
  </xsd:complexType>
  <xsd:complexType name="OrgTrackType">
    <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
    </xsd:all>
  </xsd:complexType>
  <xsd:complexType name="ConstraintsTracktype">
    <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
    </xsd:all>
  </xsd:complexType>
  <xsd:complexType name="HDTracktype">
    <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
    </xsd:all>
  </xsd:complexType>
  <xsd:complexType name="RevisionTracktype">
    <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
    </xsd:all>
```

```
</xsd:complexType>
<xsd:complexType name="CompliancyTracktype">
   <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
   </xsd:all>
</xsd:complexType>
<xsd:complexType name="ReferenceTracktype">
   <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
   </xsd:all>
</xsd:complexType>
<xsd:complexType name="CoverageTracktype">
   <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
   </xsd:all>
</xsd:complexType>
<xsd:complexType name="CrossRefTracktype">
   <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
   </xsd:all>
</xsd:complexType>
<xsd:complexType name="SiteAccessTrackType">
   <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
   </xsd:all>
</xsd:complexType>
<xsd:complexType name="AccessTrackType">
   <xsd:all>
      <xsd:element name="Creator" type="xsd:string"/>
      <xsd:element name="CreationDate" type="xsd:gYearMonth"/>
      <xsd:element name="Reviewer" type="xsd:string" minOccurs="0"/>
      <xsd:element name="ReviewDate" type="xsd:gYearMonth" minOccurs="0"/>
      <xsd:element name="Status" type="br:StatusTrack"/>
   </xsd:all>
</xsd:complexType>

<xsd:simpleType name="StatusTrack">
   <xsd:restriction base="xsd:string">
   <xsd:enumeration value="Reviewed"/>
      <xsd:enumeration value="NotReviewed"/>
      <xsd:enumeration value="Ongoing"/>
   </xsd:restriction>
</xsd:simpleType>
<!–Fin Liste MD Tracking********************* –>
<xsd:complexType name="typeDQInfo">
   <xsd:annotation>
      <xsd:documentation>A general assessment of the quality of the Data Source</xsd:documentation>
   </xsd:annotation>
   <xsd:sequence>
      <xsd:element name="ManualRevision" type="br:typeManualRevision" minOccurs="0"/>
      <xsd:element name="StandardCompliancy" type="br:typeCompliancy" minOccurs="0" maxOccurs="unbounded"/>
      <xsd:element name="DataReferenceInfo" type="br:typeDataRefInfo" minOccurs="0" maxOccurs="unbounded"/>
      <xsd:element name="SourceCoverage" type="br:typeSourceCov" minOccurs="0" maxOccurs="unbounded"/>
      <xsd:element name="CrossReference" type="xsd:string" minOccurs="0" maxOccurs="unbounded"/>
   </xsd:sequence>
   <!– Section A4–>
```

```
</xsd:complexType>
<xsd:complexType name="typeManualRevision">
   <xsd:annotation>
      <xsd:documentation>Information concerning the manual revision of the data source</xsd:documentation>
   </xsd:annotation>
   <xsd:all>
      <xsd:element name="Status" type="br:Statustype"/>
      <xsd:element name="SuppInfo" type="xsd:string" minOccurs="0"/>
   </xsd:all>
   <!–Section A4.1 –>
</xsd:complexType>
<xsd:simpleType name="Statustype">
   <xsd:restriction base="xsd:string">
      <xsd:enumeration value="Yes"/>
      <xsd:enumeration value="No"/>
      <xsd:enumeration value="NotFound"/>
   </xsd:restriction>
</xsd:simpleType>
<xsd:complexType name="typeCompliancy">
   <xsd:sequence>
      <xsd:element name="StdName" type="xsd:string"/>
      <xsd:element name="StdRef" type="xsd:anyURI"/>
      <xsd:element name="StdStatus" type="br:typeStdStatus"/>
      <xsd:element name="SuppInfoCompliancy" type="xsd:string" minOccurs="0"/>
   </xsd:sequence>
   <!–Section A4.2 –>
</xsd:complexType>
<xsd:simpleType name="typeStdStatus">
   <xsd:restriction base="xsd:string">
      <xsd:enumeration value="Total"/>
      <xsd:enumeration value="Partial"/>
      <xsd:enumeration value="NotDocumented"/>
      <xsd:enumeration value="Verified"/>
      <xsd:enumeration value="NotVerified"/>
   </xsd:restriction>
</xsd:simpleType>
<xsd:complexType name="typeDataRefInfo">
   <xsd:annotation>
      <xsd:documentation>Information about data origin in the data source</xsd:documentation>
   </xsd:annotation>
   <xsd:sequence>
      <xsd:element name="OnlineDocumentationExistence" type="br:OnlineDoc" minOccurs="0" maxOccurs="unbounded"/>
      <xsd:element name="RefPublication" type="xsd:string" minOccurs="0"/>
      <xsd:element name="RefSuppInfo" type="xsd:string" minOccurs="0"/>
   </xsd:sequence>
   <!–Section A4.3 –>
</xsd:complexType>
<xsd:complexType name="OnlineDoc">
   <xsd:annotation>
      <xsd:documentation>Description of any on-line reference information</xsd:documentation>
   </xsd:annotation>
   <xsd:all>
      <xsd:element name="DocQuality" type="br:typeDocQuality"/>
      <xsd:element name="DocAvailability" type="br:typeDocAvailability"/>
   </xsd:all>
   <!–Section A4.3.1 –>
</xsd:complexType>
<xsd:simpleType name="typeDocQuality">
   <xsd:restriction base="xsd:string">
      <xsd:enumeration value="Rich"/>
      <xsd:enumeration value="Poor"/>
      <xsd:enumeration value="Intermediate"/>
   </xsd:restriction>
   <!–Section A4.3.1.1 –>
</xsd:simpleType>
<xsd:simpleType name="typeDocAvailability">
   <xsd:restriction base="xsd:string">
      <xsd:enumeration value="Available"/>
      <xsd:enumeration value="Obsolete"/>
      <xsd:enumeration value="OutDated"/>
   </xsd:restriction>
   <!–Section A4.3.1.2 –>
</xsd:simpleType>
<xsd:complexType name="typeSourceCov">
   <xsd:annotation>
      <xsd:documentation>Number of different types of entities in the data source</xsd:documentation>
```

```
      </xsd:annotation>
      <xsd:sequence>
        <xsd:element name="EntityName" type="xsd:string" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element name="NoEntries" type="xsd:string" minOccurs="1" maxOccurs="unbounded"/>
        <xsd:element name="OrganismName" type="xsd:string" minOccurs="1" maxOccurs="unbounded"/>
        <!-- element ref="br:OrganismName" -->
      </xsd:sequence>
      <!--Section A4.4 -->
    </xsd:complexType>
    <xsd:complexType name="typeAvailabilityInfo">
      <xsd:annotation>
        <xsd:documentation>Information about the availability of the Data Source</xsd:documentation>
      </xsd:annotation>
      <xsd:all>
        <xsd:element name="SiteofAccess" type="br:typeSiteAccess"/>
        <xsd:element name="AccessConstraints" type="br:typeAccessConstraints" minOccurs="0"/>
      </xsd:all>
      <!--Section A5 -->
    </xsd:complexType>
    <xsd:complexType name="typeSiteAccess">
      <xsd:sequence>
        <xsd:element name="SiteIdentification" type="br:typeSiteId" maxOccurs="unbounded"/>
        <xsd:element name="DSVersion" type="xsd:string"/>
        <xsd:element name="InteractionModalities" type="xsd:string" minOccurs="0"/>
      </xsd:sequence>
      <!--Section A5.1 -->
    </xsd:complexType>
    <xsd:complexType name="typeSiteId">
      <xsd:sequence>
        <xsd:element name="SiteURL" type="xsd:anyURI"/>
        <xsd:element name="Organisation" type="xsd:string"/>
        <xsd:element name="SiteStatus" type="br:typeSiteStatus"/>
      </xsd:sequence>
      <!--Section A5.1.1 -->
    </xsd:complexType>
    <xsd:simpleType name="typeSiteStatus">
      <xsd:restriction base="xsd:string">
        <xsd:enumeration value="Main Site"/>
        <xsd:enumeration value="Alternative Site"/>
      </xsd:restriction>
    </xsd:simpleType>
    <xsd:complexType name="typeAccessConstraints">
      <xsd:all>
        <xsd:element name="AcademicsAccess" type="br:Academics" minOccurs="0"/>
        <xsd:element name="IndustrialsAccees" type="br:Industrials" minOccurs="0"/>
      </xsd:all>
    </xsd:complexType>
    <xsd:simpleType name="Academics">
      <xsd:restriction base="xsd:string">
        <xsd:enumeration value="Free"/>
        <xsd:enumeration value="Password"/>
        <xsd:enumeration value="Registration fees"/>
      </xsd:restriction>
    </xsd:simpleType>
    <xsd:simpleType name="Industrials">
      <xsd:restriction base="xsd:string">
        <xsd:enumeration value="Free"/>
        <xsd:enumeration value="Password"/>
        <xsd:enumeration value="Registration fees"/>
      </xsd:restriction>
    </xsd:simpleType>
    <!--Part B -->
    <xsd:complexType name="typeOnto">
      <xsd:sequence>
        <xsd:element name="BROntoName" type="xsd:string"/>
        <!--correspond au nom + année de l'ontologie -->
        <xsd:element name="OntoVersion" type="xsd:string"/>
        <xsd:element name="OntoURL" type="xsd:anyURI"/>
        <xsd:element name="SuppOntoInfo" type="xsd:string" minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
      <!--Section B1.1 -->
    </xsd:complexType>
  </xsd:schema>
```

Annexe 8

Fichier XML pour la source FlyBase

```
<?xml version="1.0" encoding="UTF-8"?>
<BioRegistry xmlns="BRSchema"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="BRSchema file:/Z:/Schemas/BioRegistry.xsd"
    BRupdate="2004-06-13">
   <Ontology>
      <BROntoName>MeSH2004</BROntoName>
      <OntoVersion>2004</OntoVersion>
   </Ontology>
   <Ontology>
      <BROntoName>NCBI Taxonomy2004</BROntoName>
      <OntoVersion>2004</OntoVersion>
   </Ontology>
   <SourceMetaData BRSourceNumber="009">
      <IdentificationInformation>
         <DS>
            <DSName>
               <FullName>FlyBase</FullName>
               <CommonName>FlyBase</CommonName>
               <Acronym>FlyBase</Acronym>
            </DSName>
            <DSContact>
               <ContactAddress>FlyBase, Biological Laboratories, Harvard University, 16 Divinity Ave, Cambridge, MA
02138</ContactAddress>
               <Country>U.S.A</Country>
               <Contactmail>flybase-help@morgan.harvard.edu </Contactmail>
            </DSContact>
         </DS>
         <Citation>
            <Authors>The FlyBase Consortium</Authors>
            <PublicationYear>2003</PublicationYear>
            <Title>The FlyBase database of the Drosophila genome projects and community literature</Title>
            <JournalInfo>
               <JournalName>Nucleic Acids Research</JournalName>
               <IssueIdentification>
                  <IssueVol>Vol. 31, No. 1</IssueVol>
                  <IssuePages>172-175</IssuePages>
               </IssueIdentification>
            </JournalInfo>
         </Citation>
         <Citation>
            <Authors>The FlyBase Consortium</Authors>
            <PublicationYear>2002</PublicationYear>
            <Title>The FlyBase database of the Drosophila genome projects and community literature</Title>
            <JournalInfo>
               <JournalName>Nucleic Acids Research</JournalName>
               <IssueIdentification>
                  <IssueVol>Vol.30</IssueVol>
                  <IssuePages>106-108</IssuePages>
               </IssueIdentification>
            </JournalInfo>
         </Citation>
         <Citation>
            <Authors>The FlyBase Consortium</Authors>
            <PublicationYear>1998</PublicationYear>
            <Title>FlyBase: a Drosophila database</Title>
            <JournalInfo>
               <JournalName>Nucleic Acids Research</JournalName>
               <IssueIdentification>
                  <IssueVol>Vol.26</IssueVol>
                  <IssuePages>85-88</IssuePages>
               </IssueIdentification>
            </JournalInfo>
         </Citation>
         <Citation>
            <Authors> The FlyBase consortium</Authors>
            <PublicationYear>1997</PublicationYear>
            <Title>FlyBase: a Drosophila database</Title>
            <JournalInfo>
               <JournalName>Nucleic Acids Research</JournalName>
               <IssueIdentification>
                  <IssueVol>Vol.25</IssueVol>
                  <IssuePages>63-66</IssuePages>
```

```
            </IssueIdentification>
          </JournalInfo>
        </Citation>
        <Citation>
          <Authors>The FlyBase consortium</Authors>
          <PublicationYear>1996</PublicationYear>
          <Title>FlyBase: the Drosophila database</Title>
          <JournalInfo>
            <JournalName>Nucleic Acids Research</JournalName>
            <IssueIdentification>
              <IssueVol>Vol.24</IssueVol>
              <IssuePages>53-56</IssuePages>
            </IssueIdentification>
          </JournalInfo>
        </Citation>
        <Citation>
          <Authors>The FlyBase consortium</Authors>
          <PublicationYear>1994</PublicationYear>
          <Title>FlyBase--the Drosophila database</Title>
          <JournalInfo>
            <JournalName>Nucleic Acids Research</JournalName>
            <IssueIdentification>
              <IssueVol>Vol.22</IssueVol>
              <IssuePages>3456-3458</IssuePages>
            </IssueIdentification>
          </JournalInfo>
        </Citation>
        <Description>
          <Abstract>FlyBase is a database of genetic and molecular data for Drosophila. FlyBase includes data on all species
from the family Drosophilidae; the primary species represented is Drosophila melanogaster.
          FlyBase is produced by a consortium of researchers funded by the National Institutes of Health, U.S.A., and the
Medical Research Council, London.
          This consortium includes both Drosophila biologists and computer scientists at Harvard University, University of
Cambridge (UK), Indiana University, University of California, Berkeley, and the European Bioinformatics Institute.</Abstract>
          <Purpose>Drosophila sequences and genomic information</Purpose>
          <SupplementalInfo>FlyBase includes the following:
              * Information on genes and mutant alleles
              * Information about the expression and properties of transcripts and proteins
              * Information on the functions of gene products
              * Nucleic acid accession numbers linked from gene records
              * Protein sequence accession numbers linked from protein records
              * Information about natural and engineered transposons and other molecular constructs
              * Lists of genomic clones
              * Descriptions of chromosomal aberrations
              * Descriptions of Drosophila stocks held in stock centers and private labs
              * Images that illustrate Drosophila anatomy and development terms
              * A bibliography of Drosophila citations
              * An address book of Drosophila researchers
              * Drosophila genetic, cytological, and molecular map information
              * Berkeley Drosophila Genome Project data
              * European Drosophila Genome Project data
              * Allied databases
              * A searchable archive of bionet.drosophila postings</SupplementalInfo>
          <EntrySample>http://flybase.bio.indiana.edu/.bin/fbidq.html?FBgn0001085</EntrySample>
        </Description>
        <MaintenanceStatus>
          <Update>
            <UpdateFrequency>unknown</UpdateFrequency>
          </Update>
          <Release>
            <ReleaseFrequency>unknown</ReleaseFrequency>
            <DateLatestRelease>2004-03-01</DateLatestRelease>
            <LatestReleaseID>Release 3.2.0</LatestReleaseID>
          </Release>
          <MaintenanceOrganisation>
            <MaintenanceName>University of Indiana</MaintenanceName>
            <MaintenanceCountry>U.S.A</MaintenanceCountry>
          </MaintenanceOrganisation>
        </MaintenanceStatus>
      </IdentificationInformation>
      <TopicInformation>
        <Subjects>
          <refBROntoNameS>MeSH2004</refBROntoNameS>
          <Keywords>
            <Term>Chromosomes</Term>
            <Identifier>D002875</Identifier>
```

```
    </Keywords>
    <Keywords>
      <Term>Genetics</Term>
      <Identifier>Q000235</Identifier>
    </Keywords>
    <Keywords>
      <Term>Chromosome Aberrations</Term>
      <Identifier>D002869</Identifier>
    </Keywords>
    <Keywords>
      <Term>Chromosome Disorders</Term>
      <Identifier>D025063</Identifier>
    </Keywords>
    <Keywords>
      <Term>Genes</Term>
      <Identifier>D005796</Identifier>
    </Keywords>
    <Keywords>
      <Term>Alleles</Term>
      <Identifier>D000483</Identifier>
    </Keywords>
    <Keywords>
      <Term>Genome</Term>
      <Identifier>D016678</Identifier>
    </Keywords>
    <Keywords>
      <Term>Genome Components</Term>
      <Identifier>D040481</Identifier>
    </Keywords>
    <Keywords>
      <Term>Cytogenetic Analysis</Term>
      <Identifier>D020732</Identifier>
    </Keywords>
    <Keywords>
      <Term>Chromosome mapping</Term>
      <Identifier>D002874</Identifier>
    </Keywords>
    <Keywords>
      <Term>Phenotypes</Term>
      <Identifier>D010641</Identifier>
    </Keywords>
    <Keywords>
      <Term>DNA</Term>
      <Identifier>D004247</Identifier>
    </Keywords>
    <Keywords>
      <Term>Nucleic Acids</Term>
      <Identifier>D009696</Identifier>
    </Keywords>
    <Keywords>
      <Term>cDNA</Term>
      <Identifier>D018076</Identifier>
    </Keywords>
    <Keywords>
      <Term>Clones</Term>
      <Identifier>D002999</Identifier>
    </Keywords>
    <Keywords>
      <Term>Proteins</Term>
      <Identifier>D011506</Identifier>
    </Keywords>
    <Keywords>
      <Term>Transposons</Term>
      <Identifier>D004251</Identifier>
    </Keywords>
    <Keywords>
      <Term>Cytogenetics</Term>
      <Identifier>D003582</Identifier>
    </Keywords>
    <Keywords>
      <Term>Anatomy</Term>
      <Identifier>Q000033</Identifier>
      <Identifier>D000715</Identifier>
    </Keywords>
  </Subjects>
  <Organism>
```

<refBROntoNameO>NCBI Taxonomy2004</refBROntoNameO>
<OrganismName>
<OrgTerm>Drosophilidae</OrgTerm>
<OrgIdentifier>7214</OrgIdentifier>
</OrganismName>
<OrganismName>
<OrgTerm>Pomace flies</OrgTerm>
<OrgIdentifier>7214</OrgIdentifier>
</OrganismName>
<OrganismName>
<OrgTerm>Drosophila melanogaster</OrgTerm>
<OrgIdentifier>7227</OrgIdentifier>
</OrganismName>
<OrganismName>
<OrgTerm>Fruit fly</OrgTerm>
<OrgIdentifier>7227</OrgIdentifier>
</OrganismName>
<OrganismName>
<OrgTerm>Drosophila ananassae </OrgTerm>
<OrgIdentifier>7217</OrgIdentifier>
</OrganismName>
<OrganismName>
<OrgTerm>Drosophila buzzatii</OrgTerm>
<OrgIdentifier>7264</OrgIdentifier>
</OrganismName>
<OrganismName>
<OrgTerm>Drosophila virilis</OrgTerm>
<OrgIdentifier>7244</OrgIdentifier>
</OrganismName>
<OrganismName>
<OrgTerm>Drosophila subobscura</OrgTerm>
<OrgIdentifier>7241</OrgIdentifier>
</OrganismName>
</Organism>
<UseConstraints>

</UseConstraints>
<HelpDesk>
<Helpmail>flybase-help@morgan.harvard.edu</Helpmail>
</HelpDesk>
</TopicInformation>
<MetaDataTrackingInformation>
<MDItem>
<allMDItems>
<Creator>Shazia Osman</Creator>
<CreationDate>2004-06</CreationDate>
</allMDItems>
<DataSource>
<Creator>S.O</Creator>
<CreationDate>2004-06</CreationDate>
</DataSource>
<Citation>
<Creator>S.O</Creator>
<CreationDate>2004-06</CreationDate>
</Citation>
<Description>
<Creator>S.O</Creator>
<CreationDate>2004-06</CreationDate>
</Description>
<MaintenanceStatus>
<Creator>S.O</Creator>
<CreationDate>2004-06</CreationDate>
</MaintenanceStatus>
<Subject>
<Creator>S.O</Creator>

<CreationDate>2004-06</CreationDate>
</Subject>
<Organism>
<Creator>S.O</Creator>
<CreationDate>2004-06</CreationDate>
</Organism>
<UseConstraints>
<Creator>S.O</Creator>
<CreationDate>2004-06</CreationDate>
</UseConstraints>
<HelpDesk>
<Creator>S.O</Creator>
<CreationDate>2004-06</CreationDate>
</HelpDesk>
<ManualRevision>
<Creator>S.O</Creator>
<CreationDate>2004-06</CreationDate>
</ManualRevision>
<SourceCoverage>
<Creator>S.O</Creator>
<CreationDate>2004-06</CreationDate>
</SourceCoverage>
<SiteofAccess>
<Creator>S.O</Creator>
<CreationDate>2004-06</CreationDate>
</SiteofAccess>
<AccessConstraints>
<Creator>S.O</Creator>
<CreationDate>2004-06</CreationDate>
</AccessConstraints>
</MDItem>
</MetaDataTrackingInformation>
<DataQualityInformation>
<ManualRevision>
<Status>Yes</Status>
<SuppInfo>Information in FlyBase is curated from the scientific literature, including unrefereed sources such as abstracts, reviews, and personal communications. With the partial exception of map data, statements of fact are not vetted by FlyBase.
Users should remember that published information is not always correct information and are encouraged to use FlyBase as a guide to the literature rather than as a substitute for it.</SuppInfo>
</ManualRevision>
<DataReferenceInfo>
<OnlineDocumentationExistence>
<DocQuality>Rich</DocQuality>
<DocAvailability>Available</DocAvailability>
</OnlineDocumentationExistence>
<RefPublication>Yes</RefPublication>
<RefSuppInfo>see: http://flybase.bio.indiana.edu/docs/</RefSuppInfo>
</DataReferenceInfo>
<SourceCoverage>
<EntityName>BAC</EntityName>
<NoEntries>949</NoEntries>
<OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
<EntityName>CDS</EntityName>
<NoEntries>18746</NoEntries>
<OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
<EntityName>DNA motif</EntityName>
<NoEntries>5</NoEntries>
<OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
<EntityName>EST</EntityName>
<NoEntries>304257</NoEntries>
<OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
<EntityName>Abberation junction</EntityName>
<NoEntries>87</NoEntries>
<OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
<EntityName>cDNA clone</EntityName>

```
    <NoEntries>10204</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>enhancer</EntityName>
    <NoEntries>27</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>gene</EntityName>
    <NoEntries>13473</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>insertion site</EntityName>
    <NoEntries>424</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>mRNA</EntityName>
    <NoEntries>18810</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>mature peptide</EntityName>
    <NoEntries>8</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>ncRNA</EntityName>
    <NoEntries>65</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>oligonucleotide</EntityName>
    <NoEntries>193813</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>point mutation</EntityName>
    <NoEntries>416</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>polyA site</EntityName>
    <NoEntries>101</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>processed transcript</EntityName>
    <NoEntries>16748</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>protein</EntityName>
    <NoEntries>233812</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>protein binding site</EntityName>
    <NoEntries>85</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>pseudogene</EntityName>
    <NoEntries>39</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>rRNA</EntityName>
    <NoEntries>85</NoEntries>
    <OrganismName>Drosophilidae</OrganismName>
</SourceCoverage>
<SourceCoverage>
    <EntityName>region</EntityName>
    <NoEntries>28</NoEntries>
```

```
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>regulatory region</EntityName>
      <NoEntries>136</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>repeat region</EntityName>
      <NoEntries>3390</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>rescue fragment</EntityName>
      <NoEntries>135</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>segment</EntityName>
      <NoEntries>437</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>sequence variant</EntityName>
      <NoEntries>225</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>signal peptide</EntityName>
      <NoEntries>1</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>snRNA</EntityName>
      <NoEntries>28</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>snoRNA</EntityName>
      <NoEntries>28</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>so</EntityName>
      <NoEntries>16244</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>tRNA</EntityName>
      <NoEntries>288</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>transcription start site</EntityName>
      <NoEntries>16997</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>transposable element</EntityName>
      <NoEntries>1567</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
    <SourceCoverage>
      <EntityName>transposable element insertion</EntityName>
      <NoEntries>4566</NoEntries>
      <OrganismName>Drosophilidae</OrganismName>
    </SourceCoverage>
  </DataQualityInformation>
  <AvailabilityInformation>
    <SiteofAccess>
      <SiteIdentification>
        <SiteURL>http://flybase.org/</SiteURL>
        <Organisation>University of Indiana, USA</Organisation>
        <SiteStatus>Main Site</SiteStatus>
      </SiteIdentification>
      <SiteIdentification>
```

```
            <SiteURL>http://fbserver.gen.cam.ac.uk/</SiteURL>
            <Organisation>University of Cambridge, U.K</Organisation>
            <SiteStatus>Alternative Site</SiteStatus>
          </SiteIdentification>
          <SiteIdentification>
            <SiteURL>http://astorg.u-strasbg.fr:7081/</SiteURL>
            <Organisation>University of Strasbourg, France.</Organisation>
            <SiteStatus>Alternative Site</SiteStatus>
          </SiteIdentification>
          <SiteIdentification>
            <SiteURL>http://bioinfo.weizmann.ac.il/flybase</SiteURL>
            <Organisation>Weizmann Institute of Science, Israel</Organisation>
            <SiteStatus>Alternative Site</SiteStatus>
          </SiteIdentification>
          <SiteIdentification>
            <SiteURL>http://www.angis.su.oz.au:7081/ </SiteURL>
            <Organisation>Australian National Genomic Information Service (ANGIS), Australia</Organisation>
            <SiteStatus>Alternative Site</SiteStatus>
          </SiteIdentification>
          <SiteIdentification>
            <SiteURL>http://flybase.nhri.org.tw/</SiteURL>
            <Organisation>National Health Research Institute, Taiwan</Organisation>
            <SiteStatus>Alternative Site</SiteStatus>
          </SiteIdentification>
          <SiteIdentification>
            <SiteURL>http://shigen.lab.nig.ac.jp:7081/ </SiteURL>
            <Organisation>SHIGEN at National Institute of Genetics, Japan.</Organisation>
            <SiteStatus>Alternative Site</SiteStatus>
          </SiteIdentification>
          <DSVersion>2004</DSVersion>
        </SiteofAccess>
        <AccessConstraints>
        <AcademicsAccess>Free</AcademicsAccess>
        <IndustrialsAccees>Password</IndustrialsAccees>
        </AccessConstraints>
      </AvailabilityInformation>
    </SourceMetaData>
</BioRegistry>
```

Annexe 9

Feuille de Style XSL pour la visualisation de 'BioRegistry' (BioRegistryView.xsl)

```
<?xml version="1.0" encoding="UTF-8" ?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
   <xsl:output method="html"/>
  <xsl:template match="/">
      <!–appliquer le modele a tous les elements du noeud Racine –>
         <html>
         <head>
            <META NAME="DESCRIPTION" CONTENT="Laboratoire Lorrain en Informatique et ses Applications"/>
               <META NAME="AUTHOR" CONTENT="Shazia Osman"/>
               <META NAME="KEYWORDS" CONTENT="Biological Data Sources, BioRegistry, Metadata, Ontologies,
Taxonomies, Controlled Vocabularies"/>
            <title>
                BioRegistry Metadata Entry Form for Biological Data Sources
            </title>
         </head>
         <body BGCOLOR="#FFFF99">
            <a href="http://www.loria.fr">
               <img border="0" src="./Logos/logo_LORIA.gif" title="Loria Homepage" width="100"
                  height="69" align="right"/></a>
            <b> <font size="4">Laboratoire Lorrain en Informatique et ses Applications<div align="center"/></font></b>
            <br/><b><font size="5">
                  <div align="Center">BioRegistry Metadata Form for Biological Data Sources</div></font></b><br/>
            <p><xsl:apply-templates select="BioRegistry"/></p>
         </body>
         </html>
  </xsl:template>
   <xsl:template match="BioRegistry/Ontology">
      <!–appliquer le modele aux sous elements Ontology –>
      <b>Thesaurus Name and Version :</b>
      <xsl:value-of select="./BROntoName"/><br/>
     <b> Thesaurus Version :</b>
      <xsl:value-of select="./OntoVersion"/><br/>
       <b> Online Link :</b>
        <a href="{./OntoURL}">
            <xsl:value-of select="./OntoURL"/>
            </a><br/>
                  <b>Supplemental Information :</b>
      <xsl:value-of select="./SuppOntoInfo"/><p/>
   </xsl:template>
   <!–**********Partie B* ******–>
    <xsl:template match="BioRegistry/SourceMetaData/AvailabilityInformation/AccessConstraints">
      <p><b><font size="3" color="#FF0033">2. Access Constraints: </font></b></p>
      <b>Academics : </b>
      <xsl:value-of select="./AcademicsAccess"/><br/>
      <b>Industrials : </b>
      <xsl:value-of select="./IndustrialsAccess"/><br/>
   </xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/AvailabilityInformation/SiteofAccess/SiteIdentification">
         <b>URL : </b>
      <a href="{./SiteURL}">
            <xsl:value-of select="./SiteURL"/>
            </a> <br/>
      <b>Organisation </b><font size="2"><i>(ex. EBI, Infobiogen in the case of SRS) </i></font><b> :</b>
       <xsl:value-of select="./Organisation"/><br/>
      <b>Status :</b>
      <xsl:value-of select="./SiteStatus"/><p/>
   </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/AvailabilityInformation/SiteofAccess">
       <b>a) Site Identification : </b><p/>
      <xsl:apply-templates select="./SiteIdentification"/> <br/>
       <b>b) DS Version :</b>
      <xsl:value-of select="./DSVersion"/><p/>
      <b>c) Interaction Modalities</b>
      <font size="2"><i> (Requirements for source usage ex: Query Language, Web services...) </i></font><b>:</b>
      <xsl:value-of select="./InteractionModalities"/><br/>
   </xsl:template>
 <xsl:template match="BioRegistry/SourceMetaData/AvailabilityInformation"> <!–Section A5 –>
    <hr width="100%"/>
            <font face="Arial Baltic" size="4">
            <div align="center">V. Availability Information :</div></font>
         <div align="center"><dfn><font size="3"><i>Information about the availability of the DS</i></font></dfn></div><br/>
    <b><font size="3" color="#FF0033">1. Site of Access :</font></b>
```

```
      <dfn><font size="3"><i> The site where the Data Source may be accessed</i></font></dfn><br/>
      <xsl:apply-templates select="./SiteofAccess"/>
      <xsl:apply-templates select="./AccessConstraints"/>
   </xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/DataQualityInformation/SourceCoverage">
      <b>Entity Name</b><font size="2"><i>(ex. Genes, EST, EST Cluster...)</i><b>:</b></font>
      <xsl:value-of select="./EntityName"/><br/>
      <b>No. of Entries :</b>
      <xsl:value-of select="./NoEntries"/><br/>
      <b>Organism Name :</b>
      <xsl:value-of select="./OrganismName"/><p/>
   </xsl:template>
   <xsl:template
match="BioRegistry/SourceMetaData/DataQualityInformation/DataReferenceInfo/OnlineDocumentationExistence">
      <xsl:apply-templates select="./text()"/>
      <b>a) Online Documetation Existence :</b>
      <dfn><font size="3"><i> Description of any on-line documentation reference information
      </i></font></dfn><br/>
      <b>Documentation Quality :</b>
      <xsl:value-of select="./DocQuality"/><br/>
      <b>Documentation Availability :</b>
      <xsl:value-of select="./DocAvailability"/><br/>
   </xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/DataQualityInformation/DataReferenceInfo">
   <xsl:apply-templates select="OnlineDocumentationExistence"/><br/>
   <b>b) Existence of Reference to Publication in the DS Entries :</b><xsl:value-of select="./RefPublication"/><br/>
  <b>c) Supplemental Information :</b><xsl:value-of select="./RefSuppInfo"/><br/><p/>
</xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/DataQualityInformation/StandardCompliancy">
      <b>Standard Name :</b><xsl:value-of select="./StdName"/><br/>
      <b>Standard Reference (URL) : </b>
      <a href="{./StdRef}">
            <xsl:value-of select="./StdRef"/></a><br/>
      <xsl:value-of select="./StdRef"/><br/>
      <b>Standard Status : </b><xsl:value-of select="./StdStatus"/><br/>
      <b>Supplemental Information :</b><xsl:value-of select="./SuppInfoCompliancy"/><br/>
   </xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/DataQualityInformation/ManualRevision">
      <b>Status :</b>
      <xsl:value-of select="./Status"/><br/>
      <b>Supplemental Information :</b>
      <xsl:value-of select="./SuppInfo"/><br/>
   </xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/DataQualityInformation">   <!–Section A4 –>
        <hr width="100%"/>
      <li><font face="Arial Baltic" size="4">
            <div align="center">
            IV. Data Quality Information : </div></font></li>
             <div align="center">
            <dfn><font size="3"><i>A general assessment of the quality of the data source
            </i></font></dfn></div><br/>
      <b><font size="3" color="#FF0033">1. Manual Revision: </font></b>
      <dfn><font size="3"><i>Information Concerning the manual revision of the DS</i></font></dfn><br/>
   <xsl:apply-templates select="./ManualRevision"/><p>
      <b><font size="3" color="#FF0033">2. Standard Compliancy: </font></b>
      <dfn><font size="3"><i>ex: MIAME compliant, ASN.1 compatible...</i></font></dfn><br/></p>
   <xsl:apply-templates select="./StandardCompliancy"/>
       <b><font size="3" color="#FF0033">3. Data Reference Information: </font></b>
   <dfn><font size="3"><i>Information about data origin in the Data Source
   </i></font></dfn><br/>
   <xsl:apply-templates select="./DataReferenceInfo"/>
      <b><font size="3" color="#FF0033">4. Source Coverage: </font></b>
      <dfn><font size="3"><i>Number of different types of entries in the Data Source
      </i></font></dfn><br/>
   <xsl:apply-templates select="./SourceCoverage"/>
      <b><font size="3" color="#FF0033">5. Cross Reference: </font></b>
   <dfn><font size="3"><i>List of Data Sources cross Referenced in the DS
   </i></font></dfn><br/>
      <b>Cross-Reference :</b>
   <xsl:value-of select="./CrossReference"/><br/>
</xsl:template>

    <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/AccessConstraints">
    <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
    <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
       <i><font size="2">(format YYYY-MM)</font></i><br/>
```

```
        <b> Reviewer : </b>
        <xsl:value-of select="./Reviewer"/><br/>
        <b> Review date : </b>
        <xsl:value-of select="./ReviewDate"/>
        <i><font size="2">(format YYYY-MM)</font></i> <br/>
        <b> Status : </b>
        <xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/SiteofAccess">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
         <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
         <i><font size="2">(format YYYY-MM)</font></i><br/>
         <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
         <b> Review date : </b> <xsl:value-of select="./ReviewDate"/>
        <i><font size="2">(format YYYY-MM)</font></i> <br/>
        <b> Status : </b>
        <xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/CrossReference">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
         <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
           <i><font size="2">(format YYYY-MM)</font></i><br/>
         <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
        <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
       <i><font size="2">(format YYYY-MM)</font></i> <br/>
        <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
     <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/SourceCoverage">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
         <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
           <i><font size="2">(format YYYY-MM)</font></i><br/>
         <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
        <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
       <i><font size="2">(format YYYY-MM)</font></i> <br/>
        <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
     <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/DataReferenceInfo">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
         <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
           <i><font size="2">(format YYYY-MM)</font></i><br/>
         <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
        <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
       <i><font size="2">(format YYYY-MM)</font></i> <br/>
        <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
     <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/StandardCompliancy">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
         <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
           <i><font size="2">(format YYYY-MM)</font></i><br/>
         <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
        <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
       <i><font size="2">(format YYYY-MM)</font></i> <br/>
        <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
<xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/ManualRevision">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
         <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
           <i><font size="2">(format YYYY-MM)</font></i><br/>
         <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
        <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
       <i><font size="2">(format YYYY-MM)</font></i> <br/>
        <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
      <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/HelpDesk">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
         <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
           <i><font size="2">(format YYYY-MM)</font></i><br/>
         <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
        <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
       <i><font size="2">(format YYYY-MM)</font></i> <br/>
        <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/UseConstraints">
 <b> Creator : </b>
           <xsl:value-of select="./Creator"/><br/>
         <b> Creation date : </b>
```

```
          <xsl:value-of select="./CreationDate"/>
          <i><font size="2">(format YYYY-MM)</font></i><br/>
        <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
       <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
      <i><font size="2">(format YYYY-MM)</font></i> <br/>
       <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/Organism">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
        <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
          <i><font size="2">(format YYYY-MM)</font></i><br/>
        <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
       <b> Review date : </b> <xsl:value-of select="./ReviewDate"/>
      <i><font size="2">(format YYYY-MM)</font></i> <br/>
       <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/Subject">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
        <b> Creation date : </b>
          <xsl:value-of select="./CreationDate"/>
          <i><font size="2">(format YYYY-MM)</font></i><br/>
        <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
       <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
      <i><font size="2">(format YYYY-MM)</font></i> <br/>
       <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/MaintenanceStatus">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
        <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
          <i><font size="2">(format YYYY-MM)</font></i><br/>
        <b> Reviewer : </b> <xsl:value-of select="./Reviewer"/><br/>
       <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
      <i><font size="2">(format YYYY-MM)</font></i> <br/>
       <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
  <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/ContentTimePeriod">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
        <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
          <i><font size="2">(format YYYY-MM)</font></i><br/>
        <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
       <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
      <i><font size="2">(format YYYY-MM)</font></i> <br/>
       <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
  <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/Description">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
        <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
          <i><font size="2">(format YYYY-MM)</font></i><br/>
        <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
       <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
      <i><font size="2">(format YYYY-MM)</font></i> <br/>
       <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
<xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/Citation">
 <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
        <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
          <i><font size="2">(format YYYY-MM)</font></i><br/>
        <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
       <b> Review date : </b> <xsl:value-of select="./ReviewDate"/>
      <i><font size="2">(format YYYY-MM)</font></i> <br/>
       <b> Status : </b>
        <xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/DataSource">
 <b> Creator : </b>
          <xsl:value-of select="./Creator"/><br/>
        <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
          <i><font size="2">(format YYYY-MM)</font></i><br/>
        <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
       <b> Review date : </b> <xsl:value-of select="./ReviewDate"/>
      <i><font size="2">(format YYYY-MM)</font></i> <br/>
       <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
</xsl:template>
  <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem/allMDItems">
       <b> Creator : </b><xsl:value-of select="./Creator"/><br/>
        <b> Creation date : </b><xsl:value-of select="./CreationDate"/>
```

```
          <i><font size="2">(format YYYY-MM)</font></i><br/>
         <b> Reviewer : </b><xsl:value-of select="./Reviewer"/><br/>
        <b> Review date : </b><xsl:value-of select="./ReviewDate"/>
       <i><font size="2">(format YYYY-MM)</font></i> <br/>
        <b> Status : </b><xsl:value-of select="./StatusTrack"/> <p/>
       </xsl:template>
       <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation/MDItem">
         <b><u> Meta-data Item</u> :</b><p/>
           <b> a) All Meta-Data Items : </b> <br/>
      <xsl:apply-templates select="./allMDItems"/>
          <b> b) Data Source : </b> <br/><xsl:apply-templates select="./DataSource"/>
          <b> c) Citation : </b> <br/><xsl:apply-templates select="./Citation"/>
          <b> d) Description : </b> <br/><xsl:apply-templates select="./Description"/>
          <b> e)Time Period of Content : </b> <br/><xsl:apply-templates select="./ContentTimePeriod"/>
          <b> f) Maintenance Status: </b> <br/><xsl:apply-templates select="./MaintenanceStatus"/>
          <b> g) Subjects: </b><br/><xsl:apply-templates select="./Subject"/>
           <b> h) Organisms : </b> <br/><xsl:apply-templates select="./Organism"/>
           <b> i) Use Constraints: </b> <br/><xsl:apply-templates select="./UseConstraints"/>
           <b> j) Help Desk : </b> <br/><xsl:apply-templates select="./HelpDesk"/>
           <b> k) Maunal Revision : </b> <br/><xsl:apply-templates select="./ManualRevision"/>
           <b> l) Standard Compliancy: </b> <br/><xsl:apply-templates select="./StandardCompliancy"/>
           <b> m) Data Reference Information : </b><br/>
        <xsl:apply-templates select="./DataReferenceInfo"/>
             <b> n) Source Coverage : </b> <br/><xsl:apply-templates select="./SourceCoverage"/>
           <b> o) Cross Reference : </b> <br/>
        <xsl:apply-templates select="./CrossReference"/>
           <b> p) Site of Access : </b> <br/><xsl:apply-templates select="./SiteofAccess"/>
           <b> q) Access Constraints: </b> <br/><xsl:apply-templates select="./AccessConstraints"/>
  </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/MetaDataTrackingInformation">   <!--SECTION A3 -->
          <hr width="100%"/>
       <li><font face="Arial Baltic" size="4">
             <div align="center">III. Meta-Data Tracking Information : </div></font></li>
      <div align="center"><dfn><font size="3"><i>Information about meta-data origin </i></font></dfn></div><br/>
       <b><font size="3" color="#FF0033">1. Meta-Data Tracking: </font></b><br/>
    <xsl:apply-templates select="./MDItem"/>
    </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/TopicInformation/HelpDesk">
       <b><font size="3" color="#FF0033">4. Help Desk: </font></b><br/>
       <b>Name :</b><xsl:value-of select="./Name"/><br/>
       <b>E-mail :</b><a href="mailto:{./Helpmail}">
       <xsl:value-of select="./Helpmail"/>
       </a><br/>
      <b>Telephone :</b><xsl:value-of select="./Telephone"/><br/>
    </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/TopicInformation/Organism/OrganismName">
         <b>Term:  </b><xsl:value-of select="./OrgTerm"/><br/>
       <b> Identifier: </b><xsl:value-of select="./OrgIdentifier"/><p/>
    </xsl:template>
     <xsl:template match="BioRegistry/SourceMetaData/TopicInformation/Organism">
       <br/>
       <b><font size="3" color="#FF0033">2. Organism Names or Keywords :</font></b>
       <dfn><i><font size="3">Words defining the organisms in the DS </font></i><p/></dfn>
        <b><u> Ontology Name and Version</u>: </b>
       <xsl:value-of select="./refBROntoNameO"/><br/>
       <xsl:apply-templates select="./OrganismName"/>
    </xsl:template>
     <xsl:template match="BioRegistry/SourceMetaData/TopicInformation/Subjects/Keywords">
       <b>Term :</b><xsl:value-of select="./Term"/><br/>
       <b>Identifier : </b><xsl:value-of select="./Identifier"/> <br/>
    </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/TopicInformation/Subjects">
       <hr width="100%"/>
       <li><font face="Arial Baltic" size="4">
             <div align="center">II. Topic Information :</div></font></li>
             <div align="center"><dfn><font size="3"><i>Words or phrases summarizing an aspect of the Data
Source</i></font></dfn></div><br/>
       <b><font size="3" color="#FF0033">1. Subject Keywords: </font></b> <br/>
        <i><font size="3">Words defining the source contents </font></i><p/>
       <b><u> Ontology Name and Version</u>: </b>
       <xsl:value-of select="./refBROntoNameS"/><br/>
       <xsl:apply-templates select="./Keywords"/>
       <b>Ontology No. :</b><xsl:value-of select="./@OntoNumber"/><br/>
    </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/TopicInformation">   <!--Section A2 -->
       <xsl:apply-templates select="./Subjects"/>
```

```
      <xsl:apply-templates select="./Organism"/>
      <p><b><font size="3" color="#FF0033">3. Use Constraints: </font></b>
        <dfn><i><font size="3">
          Restrictions and legal prerequisites for using the Data Source after access is
          granted. These include any use constraints applied to assure the protection of
          privacy or intellectual property, and any special restrictions or limitations on using the DS.
        </font></i></dfn></p>
      <xsl:value-of select="./UseConstraints"/><br/>
      <xsl:apply-templates select="./HelpDesk"/>
    </xsl:template>      <!–**********FIN SECTION A1********** –>
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/MaintenanceStatus/MaintenanceOrganisation">
      <p>
      <b><font size="3" color="#FF0033">7. Maintenance Organisation : </font></b>
      <dfn><font size="3"><i>  An entity responsible for making the DS available  ex: a person, an
organisation...</i></font></dfn><br/>
      </p>
    <b>Maintenance Name: </b><xsl:value-of select="./MaintenanceName"/><br/>
      <b>  Address:</b><xsl:value-of select="./MaintenanceAddress"/><br/>
      <b>Country : </b><xsl:value-of select="./MaintenanceCountry"/><br/>
      <b>E-mail : </b>
       <a href="mailto:{./Maintenancemail}">
         <xsl:value-of select="./Maintenancemail"/><br/>
      </a><br/><b>Telephone: </b><xsl:value-of select="./MaintenanceTelephone"/><br/>
    </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/MaintenanceStatus/Release">
      <p><u><b>
       b) Release Information:<div align="left"/></b></u></p>
       <b>Release Frequency : </b><xsl:value-of select="./ReleaseFrequency"/><br/>
      <b>Date of first release: </b><xsl:value-of select="./DateFirstRelease"/>
      <i><font size="2"> (Format YYYY-MM-DD) </font></i> <br/>
     <b>Date of latest release: </b><xsl:value-of select="./DateLatestRelease"/>
      <i><font size="2">(Format YYYY-MM-DD) </font></i><br/>
       <b>Release ID: </b><xsl:value-of select="./LatestReleaseID"/><br/>
    </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/MaintenanceStatus/Update">
       <p><b><u>
    a) Update Information:<div align="left"/></u></b></p>
       <b>Update Frequency: </b><xsl:value-of select="./UpdateFrequency"/> <br/><b>
      Date of last update: </b><i><font size="2"> (Format YYYY-MM-DD) </font></i>
      <xsl:value-of select="./datelastupdate"/>
    </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/MaintenanceStatus">
      <p><b><font size="3" color="#FF0033">6. Maintenance Status:</font></b><p></p></p>
    <xsl:apply-templates select="./Update"/>
    <xsl:apply-templates select="./Release"/>
    <xsl:apply-templates select="./MaintenanceOrganisation"/>
    </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/TimePeriodofContent">
       <xsl:value-of select="./Beginningtimeperiod"/>Beginning of Time Period:
       <xsl:value-of select="./Endtimeperiod"/>End of Time Period:
    </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/Description">
      <xsl:apply-templates select="./text()"/><p>
      <b><font size="3" color="#FF0033">4. Description :</font></b>
      <dfn><font size="3"><i> An account of the content of the Data Source</i></font><br/></dfn><p/>
     <b> Abstract:</b>
      <div align="left">
      <xsl:value-of select="./Abstract"/>
      </div><br/>
      <b>Purpose: </b>
      <xsl:value-of select="./Purpose"/><br></br><br></br>
      <b>Supplemental Information: </b><br/><xsl:value-of select="./SupplementalInfo"/></p>
      <p><b>Entry Sample Link: </b>
         <a href="{./EntrySample}">
           <xsl:value-of select="./EntrySample"/></a></p>
    </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/Citation/JournalInfo/IssueIdentification">
      <b> Volume Number: </b>
      <xsl:value-of select="./IssueVol"/><br/>
     <b>Pages: </b> <xsl:value-of select="./IssuePages"/>
    </xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/Citation/JournalInfo">
      <b> Journal Name: </b><xsl:value-of select="./JournalName"/> <br/>
       <xsl:apply-templates select="./IssueIdentification"/>
    </xsl:template
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/Citation">
```

```
      <p> <b> Article Title:</b><xsl:value-of select="./Title"/><br/>
      <b>Publication Year :</b><xsl:value-of select="./PublicationYear"/>
      <i><font size="2"> (Format: YYYY)</font></i> <br/>
      <b>PubMed ID : </b>
      <a
href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&amp;db=pubmed&amp;dopt=Abstract&amp;list_uids={./PMI
D}"><xsl:value-of select="./PMID"/>
      </a><br/><b>Author(s) : </b><xsl:value-of select="./Authors"/><br/>
      <xsl:apply-templates select="./JournalInfo"/><br/></p>
   </xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/DS/DSContact">
      <b>Contact Person:</b>
      <xsl:value-of select="./ContactPerson"/> <br/>
      <b>Contact Address: </b><xsl:value-of select="./ContactAddress"/> <br/>
     <b> Country:</b><xsl:value-of select="./Country"/> <br/>
      <b>Telephone : </b><xsl:value-of select="./Contactphone"/> <br/>
       <b>Contact email:</b>
      <a href="mailto:{./Contactmail}">
         <xsl:value-of select="./Contactmail"/>
      </a><br/>
   </xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/DS/DSName">
     <p><div align="left"><b>Full Name: </b><xsl:value-of select="./FullName"/><br/>
     <b>Common Name: </b><xsl:value-of select="./CommonName"/><br/>
     <b>Acronym : </b><xsl:value-of select="./Acronym"/></div><br/></p>
   </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/DS">
         <b><font size="3" color="#FF0033">
         <div align="left">1. Data Source:</div></font></b>
      <xsl:apply-templates select="./DSName"/>
       <p><b><font size="3" color="#FF0033">2. Data Source Contacts:</font></b></p>
      <xsl:apply-templates select="./DSContact"/>
      </xsl:template>
     <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation">
      <xsl:apply-templates select="./text()"/>
      <ul><li><font face="Arial Baltic" size="4"><div align="center"> I. Identification Information : </div></font></li>
         <dfn><font size="4"/><div align="center"><i> Basic Information about the Data Source</i></div></dfn><br/></ul>
      <xsl:apply-templates select="./DS"/>
        <p><b><font size="3" color="#FF0033">3.Citation: </font></b>
         <dfn><font size="3"><i>Information used to reference the Data Source </i></font></dfn><br/></p>
      <xsl:apply-templates select="./Citation"/>
      <xsl:apply-templates select="./Description"/>
        <b><font size="3" color="#FF0033">5. Time Period of Content :</font></b>
            <dfn><font size="3"><i>
            Information concerning the time period(s) for which the DS corresponds to the currentness reference</i></font>
               <i><font size="2"> (Format: YYYY)</font></i></dfn><br/>
      <xsl:apply-templates select="./TimePeriodofContent"/>
      <xsl:apply-templates select="./MaintenanceStatus"/>
   </xsl:template>
      <xsl:template match="BioRegistry/SourceMetaData">
          <b>BioRegistry Source Number :</b>
      <input type="text" maxlength="3" name="BRSourceNumber" size="5">
         <xsl:attribute name="value">
         <xsl:value-of select="@BRSourceNumber"/>
      </xsl:attribute>
      </input>
      <xsl:apply-templates select="./IdentificationInformation"/>
      <xsl:apply-templates select="./TopicInformation"/>
      <xsl:apply-templates select="./MetaDataTrackingInformation"/>
      <xsl:apply-templates select="./DataQualityInformation"/>
      <xsl:apply-templates select="./AvailabilityInformation"/>
   </xsl:template>
    <!–********************************************** –>
    <xsl:template match="BioRegistry">
       BioRegistry date of update  <i> (Format: YYYY-MM-DD)</i>  :
                <xsl:value-of select="./@BRupdate"/> <br/>
                <!–attribut de BR –> <hr width="100%"/>
      <xsl:apply-templates select="./SourceMetaData"/>
        <hr width="100%"/>
       <li><font face="Arial Baltic" size="4"><div align="center">VI. Ontology Identification Information :</div></font></li>
         <b><font size="3" color="#FF0033">1. Thesaurus/Ontology: </font></b><br/>
        <xsl:apply-templates select="Ontology"/>
      </xsl:template>
  </xsl:stylesheet>
```

Annexe 10

Feuille de style xsl (SourceProperties.xsl) pour l'extraction de propriétés des sources

```
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
 <xsl:output method="xml"/>
  <xsl:template match="/">
     <BioRegistry>
     <xsl:apply-templates select="BioRegistry"/>
        </BioRegistry>
     </xsl:template>
       <xsl:template match="BioRegistry/Ontology">
      <BROntoNameOrganism>
      <xsl:value-of select="./BROntoName"/>
         </BROntoNameOrganism>
     <Version>
      <xsl:value-of select="./OntoVersion"/>
     </Version>
  </xsl:template>
     <xsl:template match="BioRegistry/SourceMetaData/DataQualityInformation/ManualRevision">
        <Status>
     <xsl:value-of select="./Status"/>
           </Status>
  </xsl:template>
  <xsl:template match="BioRegistry/SourceMetaData/DataQualityInformation">
     <ManualRevision>
     <xsl:apply-templates select="./ManualRevision"/>
       </ManualRevision>
     </xsl:template>
     <xsl:template match="BioRegistry/SourceMetaData/TopicInformation/Organism/OrganismName">
        <OrganismTerm>
     <xsl:value-of select="./OrgTerm"/>
        </OrganismTerm>
        <OrganismIdentifier>
     <xsl:value-of select="./OrgIdentifier"/>
           </OrganismIdentifier>
  </xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/TopicInformation/Organism">
      <BROntoNameOrganism>
     <xsl:value-of select="./refBROntoNameO"/>
      </BROntoNameOrganism>
      <OrganismName>
     <xsl:apply-templates select="./OrganismName"/>
        </OrganismName>
  </xsl:template>
   <xsl:template match="BioRegistry/SourceMetaData/TopicInformation/Subjects/Keywords">
      <Term>
     <xsl:value-of select="./Term"/>
      </Term>
      <Identifier>
     <xsl:value-of select="./Identifier"/>
        </Identifier>
  </xsl:template>
  <xsl:template match="BioRegistry/SourceMetaData/TopicInformation/Subjects">
    <BROntoNameSujects>
     <xsl:value-of select="./refBROntoNameS"/>
    </BROntoNameSujects>
     <Keywords>
     <xsl:apply-templates select="./Keywords"/>
       </Keywords>
     <BROntoNumber>
     <xsl:value-of select="./@OntoNumber"/>
        </BROntoNumber>
  </xsl:template>
  <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/MaintenanceStatus/Release">
        <ReleaseFrequency><xsl:value-of select="./ReleaseFrequency"/>
        </ReleaseFrequency>
     </xsl:template>
  <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/MaintenanceStatus/Update">
     <UpdateFrequency><xsl:value-of select="./UpdateFrequency"/>
     </UpdateFrequency>
     </xsl:template>
     <xsl:template match="BioRegistry/SourceMetaData/TopicInformation">   <!–Section A2 –>
     <xsl:apply-templates select="./Subjects"/>
     <xsl:apply-templates select="./Organism"/>
  </xsl:template>
```

```
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/DS/DSName">
     <CommonName><xsl:value-of select="./CommonName"/></CommonName>
  </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation/DS">
     <DSName><xsl:apply-templates select="./DSName"/></DSName>
     </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData/IdentificationInformation">
     <xsl:apply-templates select="./DS"/>
        <MaintenanceStatus>
           <Update><xsl:apply-templates select="./MaintenanceStatus/Update"/></Update>
           <Release><xsl:apply-templates select="./MaintenanceStatus/Release"/></Release>
        </MaintenanceStatus>
  </xsl:template>
    <xsl:template match="BioRegistry/SourceMetaData">
        <BRSourceNumber>
         <xsl:value-of select="@BRSourceNumber"/>
        </BRSourceNumber>
        <IdentificationInformation>
     <xsl:apply-templates select="./IdentificationInformation"/>
        </IdentificationInformation>
        <TopicInformation><xsl:apply-templates select="./TopicInformation"/></TopicInformation>
        <DataQualityInformation><xsl:apply-templates select="./DataQualityInformation"/></DataQualityInformation>
  </xsl:template>
    <xsl:template match="BioRegistry">
        <BRDateUpdate>
           <xsl:value-of select="./@BRupdate"/>
           </BRDateUpdate>
        <SourceMetaData>
     <xsl:apply-templates select="./SourceMetaData"/>
        </SourceMetaData>
        <Ontology>
       <xsl:apply-templates select="Ontology"/>
           </Ontology>
     </xsl:template>
</xsl:stylesheet>
```

Annexe 11

Exemple de fichier xml 'SourceProperties.xml' généré pour l'extraction de propriétés de la source (EMBL)

```
<?xml version="1.0" encoding="utf-8" ?>
- <BioRegistry>
    <BRDateUpdate>2004-05-03</BRDateUpdate>
  - <SourceMetaData>
      <BRSourceNumber>001</BRSourceNumber>
    - <IdentificationInformation>
      - <DSName>
          <CommonName>EMBL</CommonName>
        </DSName>
      - <MaintenanceStatus>
        - <Update>
            <UpdateFrequency>daily</UpdateFrequency>
          </Update>
        - <Release>
            <ReleaseFrequency>quarterly</ReleaseFrequency>
          </Release>
        </MaintenanceStatus>
      </IdentificationInformation>
    - <TopicInformation>
        <BROntoNameSujects>MeSH2004</BROntoNameSujects>
      - <Keywords>
          <Term>DNA</Term>
          <Identifier>D004247</Identifier>
          <Term>RNA</Term>
          <Identifier>D012313</Identifier>
          <Term>Nucleic Sequence</Term>
          <Identifier>D030561</Identifier>
          <Term>RNA Sequence</Term>
          <Identifier>D001483</Identifier>
        </Keywords>
        <BROntoNumber />
        <BROntoNameOrganism>NCBI Taxonomy2004</BROntoNameOrganism>
      - <OrganismName>
          <OrganismTerm>any organism</OrganismTerm>
          <OrganismIdentifier>root</OrganismIdentifier>
        </OrganismName>
      </TopicInformation>
    - <DataQualityInformation>
        <ManualRevision />
      </DataQualityInformation>
    </SourceMetaData>
  - <Ontology>
      <BROntoNameOrganism>MeSH2004</BROntoNameOrganism>
      <Version>2004</Version>
      <BROntoNameOrganism>NCBI Taxonomy2004</BROntoNameOrganism>
      <Version>2004</Version>
    </Ontology>
</BioRegistry>
```