Project Inserm
Interopérabilité

Documentation de spécification du format commun d'échange de données

Rédacteurs :	Nicolas Malservet
Date de création :	15 octobre 2012
Date de mise à jour :	12 novembre 2014
Version du document :	0.5
Nombre de pages :	9
Information de contact :	n.malservet@inserm.fr

Table des matières

1.Objectifs

1.1.Objectif de document

Ce document a été écrit pour présenter le format commun de données du projet poru les biobanques en France.

1.2.Objectifs de la plate-forme

Les objectifs de la plate-forme sont multiples.
Le premier est d'augmenter les interactions possibles entre chercheurs et biobanqueurs. De nombreuses études peuvent être réalisées en utilisant le matériel biologique stocké dans les biobanques et les interviews menées montrent que la recherche de matériel biologqiue peut prendre beaucoup de temps.
Le second est l'harmonisation des voies d'échanges de données autour des biobanques.

2.Références

Ce document fait référence aux autres initiatives de standardisation et d'harmonisation des formats de données.

2.1.Les recommandations de l' OCDE

Dans son document nommé " OECD Best Practices Guidelines for BRCs 2007", OCDE dresse une liste d'items à utiliser pour échanger des informations entre biobanques et décrire un échantillon. Il convient d'utiliser cette liste pour établir la première version du format commun d'échange. Cette liste ne décrit cependant pas le niveau d'implémentation que nous allons utiliser. D'autres implémentations de cette liste OCDE sont donc possibles.

2.2.Le projet MIABIS du regroupement BBMRI

Le regroupement européen BBMRI a produit un set de données pour représenter le minimum d'information requises pour activer les échanges d'échantillons biologiques et de données entre biobanques et chercheurs. Ce projet porte le nom de MIABIS pour « Minimum Information About Biobank data Sharing ».
La documentation de ce projet est accessible sur Internet via ce lien : http://bbmri-wiki.wikidot.com/en:dataset

Le niveau du jeu de données fourni est assez haut, c'est à dire qu'il synthétise le contenu des collections d'échantillons biologiques. Ce jeu de données ne fournit pas les détails liés à l'échantillon.

Il peut être util d'utiliser la normalisation de ce projet pour fournir les méta-informations de la biobanque, mais avec nos objectifs, de décrire les échantillons et construire un catalogue dynamique, ce set de données n'est pas assez détaillé.

2.3.Le projet INCa

INCa a fourni un set de données minimum pour les biobanques humaines, spécialisé dans le Cancer, et ce afin de produire un catalogue commun.
Le résultat de ce travail a produit la TVN (Tumorothèque Virtuelle Nationale), un catalogue en ligne contenant les échantillons de chaque biobanque participante.
La première livraison de ce travail a été réalisée avec le groupe thématique du cancer du poumon.
Une présentation de ce travail est disponible via cette URL: http://www.e-cancer.fr/recherche/les-ressources-biologiques/la-tumorotheque-virtuelle-nationale

le jeu de données OCDE peut être normalisé en suivant ce jeu de données INCa et étendu via celui-ci.
Le jeu de données INCa est cependant très spécifiques aux tumeurs et il ne peut être un format source applicable pour toute biobanque.

3.Définition du format commun

3.1.Langage

Le format commun doit être lisible par touts les acteurs atour des biobanques. Il doit être simple et extensible pour faciliter la communication

Le format commun est implémenté en XML (Extensible Markup Language), car ce langage est très utilisé en sciences informatiques, ouvert et simple d'utilisation. Par ce langage, il est aisé de créer des meta-données utilisables par une machine et lisible par un utilisateur averti. Ce langage XML est un meta-langage, il est possible de définir de nombreuses meta-données ce qui est un atout majeur dans l'échange de données structurées.

3.2.Le jeu de données

Nous définirons le jeu de données en utilisant les recommandations fournies par l'OCDE.
Le jeu de données est séparés en deux catégories matériel humain et micro-organismes.

La plupart des items d'une jeu de données n'ont pas le même niveau de priorité pour être implémenté, donc en suivant les recommandations de l'OCDE, nous utiliserons trois niveaux d'intégration. Le premier niveau est le jeu de données minimal, (exactement le même que celui décrit par l'OCDE), le second niveau d'intégration est le jeu de données recommandés, (exactement le même que celui décrit par l'OCDE) et enfin le troisième niveau est le jeu de données étendu (incluant quelques items INCa et BBMRI).

OCDE dans son document ne propre pas un format numérique mais seulement les items nécessaires, nous avons donc effectués la traduction de ce document en langage informatique pour proposer une implémentation. Cette implémentation est amenée à évoluer, la référence à la version dans le document permettra de faciliter la lecture des données.

- La colonne Item indique quel item doit être stocké.
- La colonne level indique le niveau d'intégration requis.(MDS = Je de données Minimal = Minium Data-Set, RDS = Jeu de données recommandés, Recommended Data-Set, EDS = jeu de données étendu, Extended Data Set)
- La colonne source indique la provenance de cet item parmi les initiatives d'harmonization.
- La colonne « type of material » indique pour quel type de matériel l'item doit être implémenté. Les types sont : DNA, Tissue = tissues and isolated cells, Cells = cell line, primary cultured cells and transformed.
- Ce document est en anglais car sa portée devrait être internationale.

Item	Level	Source	DNA	Tissues	Cells	variable name	Values format
Identification of depositor	MDS	OCDE	Y	Y	Y	id_depositor	Text
Identification number of the family	MDS	OCDE	Y	N	N	id_familiy	Text
Identification number of the donor	MDS	OCDE	Y	Y	N	id_donor	Text
Identification number of the biological mater	MDS	OCDE	Y	Y	Y	id_sample	Text
Consent/Approval by ethical commitee	MDS	OCDE	Y	Y	Y	consent_ethical	Fixed values(Yes, No, Unknown) : Y,N,U
Gender of donor	MDS	OCDE	Y	Y	Y	gender	Fixed values(Male, Female, Hermaphrodite, Unknown) : M, F, H, U
Age of donor	MDS	OCDE	Y	Y	Y	age	Integer
Pathology of family with OMIM number	MDS	OCDE	Y	N	N	pathology	Integer (unique six-digit number)
Status of the biological materiel	MDS	OCDE	Y	Y	N	status_sample	Fixed Values(Affected, Non-affected, indication of suspected diagnosis, indication of grade of tumor) :
date of collect of the material	MDS	OCDE	Y	Y	Y	collect_date	ISO-standard (8601) time format
nature of the human biological material whe	MDS	OCDE	Y	N	N	nature_sample_dna	fixed values (affected, non affected)
preservation or storage conditions	MDS	OCDE	Y	Y	Y	storage_conditions	fixed values (liquid nitrogen, -80C, room temperature)
quantity of biological material	MDS	OCDE	Y	N	N	quantity	for dna concentration µg/µl and number of µl
Disease diagnosis	MDS	OCDE	N	Y	N	disease_diagnosis	Text
Origin of the biological material	MDS	OCDE	N	Y	Y	origin	Organ and tissue
hazard status	MDS	OCDE	N	Y	N	hazard_status	Text
nature of the human biological material	MDS	OCDE	N	Y	N	nature_sample_tissue	fixed values (tissue, slide,cells, pellet)
documentation on processing method	MDS	OCDE	N	N	N	processing_method	Text
nature of the cells	MDS	OCDE	N	N	Y	nature_sample_cells	fixed values(epithelia, fibroblast,myphocyt)
culture condition	MDS	OCDE	N	N	Y	culture_condition	medium and subculture routine
consent	RDS	OCDE	Y	Y	Y	consent	Fixed values(Yes, No, Unknown) : Y,N,U
family tree	RDS	OCDE	Y	N	N	family_tree	Fixed values(Yes, No, Unknown) : Y,N,U
samples from relatives available	RDS	OCDE	Y	N	N	available_relatives_samp	Fixed values(Yes, No, Unknown) : Y,N,U
form of supply	RDS	OCDE	Y	Y	Y	supply	Text
maximum delay for delivery	RDS	OCDE	Y	Y	Y	max_delay_delivery	Integer (hours)
karyotype	RDS	OCDE	Y	N	N	karyotype	Text
quantity of families and subjects available fo	RDS	OCDE	Y	N	N	quantity_families	Integer
detail information of treatment/medications	RDS	OCDE	Y	Y	Y	detail treatment	Text
information on disease outcome	RDS	OCDE	Y	Y	Y	disease_outcome	Text
associated clinical data	RDS	OCDE	Y	Y	N	associated_clinical_data	Fixed values(Yes, No, Unknown) : Y,N,U
associated molecular data (ref associated c	RDS	OCDE	Y	Y	N	associated_molecular_d	Fixed values(Yes, No, Unknown) : Y,N,U
associated imagin data (ref associated clin	RDS	OCDE	Y	Y	N	associated_imagin_data	Fixed values(Yes, No, Unknown) : Y,N,U
information on life style	RDS	OCDE	Y	Y	N	life_style	Text
information on family history	RDS	OCDE	Y	Y	N	family_history	Text
dna fingerprinting or another method of aut	RDS	OCDE	Y	Y	Y	authentication_method	Text
hazard status	RDS	OCDE	Y	N	Y	hazard_status	Text
details of diagnosis	RDS	OCDE	N	Y	Y	details_diagnosis	Text
related biological material	RDS	OCDE	N	Y	Y	related_biological_mater	Fixed Values : DNA Biopsy, tissue,serum, dna
Quantity available	RDS	OCDE	N	Y	Y	quantity_available	Text
Concentration available	RDS	OCDE	N	Y	Y	concentration_available	Text
characteristics of the sample	RDS	OCDE	N	Y	N	samples_characteristics	Text (sample composition, content tumour cells)
delay of freezing	RDS	OCDE	N	Y	N	delay_freezing	Integer(hours)
characterization of cells	RDS	OCDE	N	N	Y	cells_characterization	Text (doubling time, tumorigenicity, karyotype etc)
number of passage	RDS	OCDE	N	N	Y	number_of_passage	Text
morphology and growth characteristics	RDS	OCDE	N	N	Y	morphology_and_growth	Text
reference paper	RDS	OCDE	N	N	Y	reference_paper	Text
biobank id	MDS	BBMRI	Y	Y	Y	biobank_id	Textual string of letters starting with the country code (according to standard ISO1366 alpha2) followed by the under
biobank name	RDS	BBMRI	Y	Y	Y	biobank_name	Text
Date of Entry	MDS	BBMRI	Y	Y	Y	biobank_date_entry	ISO-standard (8601) time format when data about the biobank was reported into a database.
Sample Collection/Study ID*	MDS	BBMRI	Y	Y	Y	biobank_collection_id	Text : Textual string depicting the unique ID or acronym for the sample collection or study
Collection/Study name*	RDS	BBMRI	Y	Y	Y	biobank_collection_name*	Text : Textual string of letters denoting the name of the study in English
birth date patient	EDS	INCA	Y	Y	Y	patient_birth_date	Date format aaaammjj
tumor diagnosis	EDS	INCA	Y	Y	Y	tumor_diagnosis	Text : CIM 10 format

4.Guide d'utilisation pour l'implémentation du fichier XML

Cette section décrit comment produire le fichier XML pour utiliser le format commun.

4.1.Syntaxe

La syntaxe du fichier XML doit suivre les recommandations de bonne pratique d'usage du XML
Un fichier XML est un arbre structuré d'informations. Chaque nœud de l'arbre contient une information.
Chaque nœud est défini par un nom, pour lequel une valeur est associée.
Chaque nœud utilise la même syntaxe avec un marqueur de début et marqueur de fin.
Un marqueur de début est représenté par le nom de l'attribut, préfixé de < et suffixé de >.
Un marqueur de fin contient le nom de l'attribut préfixé de </ puis suffixé de >.
La valeur de l'attribut peut être écrite entre les marqueurs.
Par exemple :

```
<myitem>myvalue</myitem>
```

Un commentaire dans le fichier XML peut être inséré et configuré en utilisant <!-- avant le commentaire et terminé par - - > pour clore le commentaire.
Exemple :

```
<!-- ceci est un commentaire -->
```

Un fichier Xml doit contenir un entêtre permettent de spécifié quelle version d'XML est utilisée et quel jeu d'encodage appliquer.

```
<?xml version="1.0" encoding="utf-8" ?>
```

Puis spécifiquement au projet, nous pouvons décrire une section contenant les meta-données utiles pour décrire la version du format commun que nous utiliserons, ainsi que la date de dcréation du fichier.

```
<!-- date iso format 8601 -->
<date>2011-04-01T13:01:02</date>
<formatVersion>1</formatVersion>
```

Puis nous avons une section décrivant les informations de la biobanque.

```
<biobank>
            <biobank_id>44XYZ</biobank_id>
            <biobank_name>Lorraine Biobank</biobank_name>
</biobank>
```

Et enfin pour chaque échantillon, nous pouvons stocker ses données, en partant toujours du jeu de données minimal, puis celui recommandé et enfin celui étendu si possible.
Le jeu de données minimal est défini par le type d'échantillon, ADN, tissu, ou cellules.
Par exemple, appliqué à un échantillon de type tissu :

```
<sample>
            <id_depositor>my depositor</id_depositor>
...
            <gender>M</gender>
...
</sample>
```

◦ Champs non harmonisés

Dans le cadre de l'utilisation pour des réseaux thématiques, un certain nombre de variables sont utiles mais non harmonisées par défaut. L'harmonisation d'items spécifiques demanderait un effort important de concertation de la part des divers utilisateurs de ce type de données avec un intérêt plutôt limité dans le cadre de projets généraux.

Pour pourvoir ce besoin sans devenir une contrainte pour tous, nous avons opter pour un système semi-harmonisé, en permettant l'agrégation d'informations de type notes agrégées à un échantillon.

Des annotations sont alors possibles, pour lesquelles il faudra définir une clé et sa valeur. La clé correspond à l'intitulé du champs à stocker, et le champ valeur à la valeur à stocker en format texte libre.

```
<sample>
< !-- notes est une balise contenant les notes variables -->
      <notes>
      < !-- une annotation est contenu dans une balise note contenant 2 sous-
      balises une clé et une valeur-->
            <note>
                  <key>is_parkinson</key>
                  <value>yes</value>
            </note>
      </notes>
```

4.2.Exemples

Dans cette section, vous trouverez les exemples typiques d'usage pour les cas les plus utilisés.
Pour utiliser au mieux le fichier de définition du format commun, nous vous recommandons d'utiliser les filtres disponibles dans les entêtes de colonne afin d'obtenir automatiquement la liste des items à renseigner.

4.2.1Cas tissu humain – Minimum Data-Set

Pour décrire un échantillon de type tissu humain, en jeu de données minimal, nous obtiendrons un fichier contenant la structure suivante :

```
<?xml version="1.0" encoding="utf-8" ?>
<!-- date iso format 8601 -->
<date>2011-04-01T13:01:02</date>
<formatVersion>0.5</formatVersion>
<biobank>
        <biobank_id>3344XYZ</biobank_id>
        <biobank_name>My Biobank</biobank_name>
</biobank>
<samples>
    <sample>
        <id_depositor>d123456</id_depositor>
        <id_donor>do123456</id_donor>
        <id_sample>s123456</id_sample>
        <consent_ethical>Y</consent_ethical>
        <gender>M</gender>
        <age>56</age>
        <status_sample>Affected</status_sample>
        <collect_date>2006-04-01T13:01:02</collect_date>
        <storage_conditions>LN</storage_conditions>
        <disease_diagnosis>free text</disease_diagnosis>
        <nature_sample_tissue>T</nature_sample_tissue>
        <biobank_date_entry>2006-05-01T13:01:02</biobank_date_entry>
        <biobank_collection_id>collec_tumeurs_1</biobank_collection_id>
        <!-- champs variables libres-->
        <notes>
            <note>
            <key>IS_ALZHEIMER</key>
            <value>true</value>
            </note>
        </notes>
    </sample>
</samples>
```