Réutilisation des informations publiques et formats

Position du GFII en vue de la rédaction des recommandations du COEPIA

29/03/2011

Ce document rassemble les réflexions du GFII sur le format des données publiques :

- Le format des données
- La forme des données (accès sur internet, quel protocole, doit-on fournir un support, *etc.*)
- Le délai de fourniture des données : la valeur de certaines données dépend de leur fraîcheur et/ou de leur mise à jour.

1. Quelques remarques sur la situation actuelle

Relations difficiles

Les relations restent parfois difficiles entre fournisseurs publics de données et ré-utilisateurs industriels. Les ré-utilisateurs regrettent un manque de dialogue, l'absence d'un interlocuteur responsable clairement identifié du côté de l'éditeur public et de manière générale un manque de concertation entre fournisseurs de données publiques et ré-utilisateurs.

Physique et conceptuel

Nous avons distingué deux niveaux :

1. le niveau physique permettant notamment la présentation et l'affichage du document/données
2. le niveau du modèle conceptuel de données permettant de compréhension et l'analyse du contenu du document

Le premier niveau est générique et ne dépend pas du domaine traité par le document.

Les formats disponibles à ce jour sont :

- pour le texte/numérique : ascii, pdf texte, xls, csv, xml, ODF, html, *etc*.
- pour l'image : jpeg, *etc*.
- pour la video: mpeg-4, *etc*.

Le deuxième niveau est en général dépendant du domaine traité par le document (par exemple domaine géographique, transport, juridique ou économique). Il s'appuie sur un modèle de données prenant en compte les métiers concernés.

GFII - 25 rue Claude Tillier - 75012 Paris - Tél. : 01 43 72 96 52 - Fax : 01 43 72 56 04
mél : gfii@gfii.asso.fr - site web : http://www.gfii.asso.fr

- Par exemple, dans le domaine de l'information géographique, la mise en œuvre de la directive INSPIRE conduit à définir un modèle conceptuel dont l'utilisation va devenir obligatoire pour les échanges dans l'Union européenne. Jusqu'à maintenant, chaque grande ville française avait son propre système et modèle de données. La directive INSPIRE a mis de l'ordre dans cette surabondante diversité en réunissant tout le monde au niveau européen. 32 thèmes ont été définis (eg hydrographie est une section) et chacune a fait l'objet d'un modèle.
- Une approche similaire a été appliquée pour le domaine du transport avec l'initiative ITS (Information and transports system) européenne : http://ec.europa.eu/transport/its/road/action_plan/action_plan_en.htm
- Dans le domaine financier, une approche similaire est suivie autour de XBRL.
- Dans le domaine de la culture, le protocole OAI est assez communément adopté pour le transfert de données des bibliothèques, archives, musées (http://www.openarchives.org)

Grande hétérogénéité des situations

L'ouverture des données publiques recouvre des domaines très variés qui ont chacun leur problématique métier propre (par exemple les domaines juridique, géographique ou généalogique ont clairement leur spécificités propres qui impactent les formats les plus appropriés pour la réutilisation des données publiques). Il y a donc une grande hétérogénéité des situations du fait de la diversité des acteurs industriels, des acteurs publics, des modèles économiques, de la nature, du format des données ou des contraintes juridiques. Il est donc difficile de faire des recommandations générales sur les formats.

Dématérialisation

La problématique des formats des données publiques s'inscrit de façon générale dans le processus de dématérialisation des procédures administratives.

2. Recommandations

Recommandation 1 : des formats ouverts

Nous proposons que la mise à disposition soit faite *a minima* sous un format ouvert, et sous un format pour lequel l'ouverture a permis la disponibilité d'une offre diverse et parfois gratuite d'outils de présentation et de réutilisation.

Recommandation 2 : des métadonnées standard par métier

Le GFII recommande l'élaboration, l'adoption et la mise en œuvre de standards de métadonnées et données dans toutes les thématiques métiers (juridique, économique, SHS,…) Ces standards doivent être fondés sur les normes déjà largement admises comme DublinCore, ISO et AFNOR et s'inscrire dans le cadre du référentiel d'interopérabilité et bien entendu prendre en compte l'importance des méta-données dans la description générale des données. Ces métadonnées ne doivent pas s'entendre comme éléments d'identification de la donnée mais en tant qu'éléments de description complémentaires des données.

Le GFII recommande la définition par métier de normes de représentation de données prenant en compte les spécificités du métier. Ces normes doivent être établies par les représentants des filières métier, avec la participation des éditeurs publics concernés, et en prenant en compte la définition de normes similaires au niveau européen.

Recommandation 3 : Anonymisation

Le sujet de l'anonymisation est traité de façon plus complète par une autre commission du GFII. Nous recommandons ici spécifiquement un point lié au format : les données fournies par l'éditeur public doivent conserver les données à anonymiser, ces données étant entourées d'un balisage spécifique indiquant clairement la partie du texte devant être anonymisée en vue d'une diffusion commerciale, la conservation des noms et des adresses permettant au ré-utilisateur meilleure compréhension du texte d'origine.

Recommandation 4 : des données source à côté des PDF

PDF est un excellent format de données pour présenter un document, notamment au grand public.

Il est en revanche souvent un très mauvais format pour la réutilisation, car il oblige le réutilisateur à un exercice contre-productif de rétro-ingénierie.

Dans la majorité des cas, le fichier PDF texte résulte du passage en PDF de documents plus exploitables (word, excel, powerpoint, etc.). Donner accès aux données source qui ont servi à produire le PDF (quand ils sont disponibles) ne présente aucun coût supplémentaire pour l'éditeur.

Nous recommandons donc que dans le cas de publication d'un fichier PDF texte, le fichier source du PDF, s'il est disponible à l'éditeur, soit systématiquement publié en même temps que le fichier PDF.

Recommandation 5 : privilégier le format XML

Le GFII considère que l'adoption de XML comme langage de publication de données est un pas positif de la part des éditeurs publics. C'est actuellement le format qui est à privilégier. Cette adoption doit s'accompagner d'une réflexion sur le choix des formalismes XML choisis (voir recommandation 2). Dans cet esprit, tout opération de balisage de données textes brutes en relevant le niveau sémantique est à encourager.

Recommandation 6 : mettre en place des structures de dialogue Editeurs / Réutilisateurs

A ce jour, il n'existe pas suffisamment de structures de dialogue dédiées chez les éditeurs de données publiques en France. Nous recommandons que chaque éditeur pourvoyeur d'un grand nombre de données mette en place une telle structure, dont la fonction serait au minimum d'informer de l'existant, des plans d'évolutions futures, d'expliquer et de motiver les choix d'évolution, d'écouter les suggestions et de répondre aux questions des industriels réutilisateurs. Cette structure pourrait-être formalisée en transposant ce qui a été fait dans le domaine de l'information géographique avec le Conseil national de l'information géographique. L'extranet réalisé par la DILA pour les réutilisateurs du BODACC est également un autre exemple de bonnes pratiques.

Recommandation 7 : Définitions de formats d'échange

A ce jour, de nombreux éditeurs publics fournissent les données dans le format interne des données donc lié à l'usage qu'ils en font en interne. Nous recommandons la définition d'un format d'échange, centré sur la fonction diffusion. Ces formats d'échanges pourraient être élaborés conjointement par des éditeurs de données publiques et les industriels et utilisateurs de ces données. De nombreux formats d'échanges ont été définis par des DTD XML dans des métiers divers. Ces formats d'échange garantissent l'interopérabilité des systèmes qui utilisent ces données.

3. Résumé

1. **Pour des formats ouverts des données physiques**
2. **Pour le développement de normes métier pour le niveau sémantique des données**
3. **Pour un processus d'anonymisation ne détruisant pas le sens des données**
4. **Pour une publication des sources des fichiers PDF texte**
5. **Pour un usage généralisé de XML comme format de données**
6. **Pour un dialogue entre éditeurs publics et ré-utilisateurs**
7. **Pour la définition de formats d'échange**

4. Glossaire des termes

Ouvert ou propriétaire

Format ouvert : format publié, non propriétaire (donc libre de droit), pour lequel le développement d'outils d'édition/visualisation/présentation est ouvert à toute entreprise.

Format propriétaire : format non nécessairement publié et protégé par des droits, pour lequel le développement d'outils d'édition/visualisation/présentation est restreint.

Exemple de formats physiques ouverts : html, odf et csv
Exemple de formats physiques propriétaires : pdf, doc et xls

Données brutes et données raffinées

Les données brutes sont les données telles quelles sans traitement particulier ou valeur ajoutée.
Les données raffinées sont des données auxquelles on a appliqué un traitement spécifique pour les améliorer (correction, changement de format, traitement sémantique, anonymisation, intégration, rectification et correction, etc.).
Ces deux notions sont bien entendu relatives : le raffinage peut être un processus itératif, donc la donnée raffinée de l'un peut être la donnée brute de l'autre.