{BnF

Politique de qualité des données de la Bibliothèque nationale de France

Contenu

La politique de qualité des données de la Bibliothèque nationale de France (BnF) est un document de synthèse qui vise à fixer le cadre général et les engagements de l'établissement en matière de politique des données.

Elle a vocation à intégrer les enjeux d'agrégation, de production et de diffusion des données tant dans le contexte actuel qu'en lien avec les évolutions normatives, organisationnelles et techniques en cours d'implémentation.

Politique de qualité des données : définition

Les évolutions technologiques liées au Web positionnent la BnF dans un environnement global numérique avec des enjeux forts culturels et de recherche fondés sur la production, la diffusion et l'échange de données. De par la nature unique de ses collections, la masse et la profondeur thématique et chronologique de ses données, les compétences et l'expertise humaines qu'elle mobilise dans ses capacités de production et de traitement des données, la BnF constitue un opérateur innovant et incontournable du monde des données. Elle est en particulier un acteur de référence, à l'échelle nationale et internationale, au sein des réseaux de bibliothèques et des institutions bibliographiques, positionnement encore renforcé par son statut d'Agence bibliographique nationale et de co-pilote du programme national de la Transition bibliographique.

Dans ce contexte, la BnF s'est donnée comme objectif de se doter d'une politique de qualité des données qu'elle produit, afin de réaffirmer le rôle d'acteur de confiance qu'elle joue dans les domaines du signalement et du référencement.

Pour ce faire, elle se fonde sur les principes suivants :

La qualité des données désigne la capacité de l'ensemble des caractéristiques intrinsèques des données – accessibilité, conformité, cohérence, traçabilité, « localisation » – à répondre aux exigences de signalement des documents décrits et de réutilisation de ces données, en adéquation avec les moyens assignés à leur production.

La mise en œuvre de cette qualité est une mission fondamentale de la BnF, exercée de manière constante et dynamique : d'importants chantiers de correction, de mise en conformité et de curation des données sont mis en œuvre chaque année afin d'améliorer la qualité des données, de les enrichir, de les mettre à jour des évolutions normatives.

Politique de qualité des données : quel périmètre ?

La politique de qualité des données de la Bibliothèque nationale de France a pour périmètre l'ensemble des « données documentaires », *i.e.* les métadonnées bibliographiques et descriptives agrégées, produites et diffusées par la BnF, et qui permettent d'identifier et de qualifier la forme et le contenu des publications. La notion de données documentaires englobe le périmètre des données bibliographiques telles que prévu dans sa modélisation IFLA-LRM[1], implémentée à terme par le code de catalogage RDA-FR dans le cadre de la Transition bibliographique.

La politique de qualité des données inclut dans son champ d'application l'ensemble des bases bibliographiques et catalographiques de la BnF : Catalogue général, BnF Archives et manuscrits, Catalogue des Médailles et Antiques.

Si elle fixe avant tout un cadre général pour toutes les filières de la BnF en charge de l'agrégation, de la production et de la diffusion des données, la politique de qualité des données a également pour vocation de constituer un socle définissant les attentes et engagements de l'institution dans le cadre de coproduction de données avec d'autres établissements : Fichier national d'entités (FNE), International Standard Name Identifier (ISNI), Europeana…

Ce cadre posé, trois axes forts, qui se recoupent et convergent, structurent les attentes de la BnF en matière de signalement de ses ressources documentaires :

- La mission nationale confiée à la BnF de collecte du Dépôt légal, définie par le Code du patrimoine qui précise dans l'article L131-1[2] :
  « *Le dépôt légal est organisé en vue de permettre : […] La constitution et la diffusion des bibliographies nationales*… » ;
- Le rôle de la BnF en tant qu'Agence bibliographique nationale et tête du réseau français de lecture publique ;
- L'adoption progressive du modèle de catalogage IFLA-LRM, qui permet d'assurer la visibilité des collections de la BnF sur le Web de données tout en répondant pleinement aux tâches utilisateurs définies par ce modèle (trouver, identifier, sélectionner, obtenir et explorer).

Politique de qualité des données : quels usagers?

Une politique de qualité des données ne se conçoit, en particulier dans le cadre de l'action publique, que définie à l'aune des besoins de leurs usagers. Or, les profils des utilisateurs des données de la BnF et, partant, leurs attentes en matière de qualité liée au service proposé, sont de nature extrêmement variée :

- Des utilisateurs internes : tous les professionnels de la BnF qui fondent une partie de leur activité sur l'exploitation de ces données (gestion physique et scientifique des collections et des fonds, actions de valorisation des documents, pilotage et statistiques, activités de numérisation…)

[1] https://www.ifla.org/publications/node/11412.
[2] https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006074236&idArticle=LEGIARTI000006845515.

- Des utilisateurs externes : reversement de tout ou partie des données à des grands réservoirs nationaux (Sudoc) et internationaux (WorldCat, centre ISSN international, base ISNI…), récupération des notices bibliographiques par le réseau français de lecture publique ou par des éditeurs de SIGB, réutilisation dans le cadre du développement d'applications Web, travaux de recherche scientifique portant sur les données en tant que collection à part entière de la BnF, fouille de données…

Il résulte de cette grande variété de profils que la perception de la qualité des données et les notions que chaque utilisateur pourrait en faire découler (exhaustivité, précision…) sont nécessairement différentes suivant les attentes : la notion de qualité telle que définie par les utilisateurs est fortement liée à leur contexte d'utilisation de ces données.

Politique de qualité des données : quels critères ?

La qualité des données désigne la capacité de l'ensemble des caractéristiques intrinsèques des données – accessibilité, conformité, cohérence, traçabilité, « localisation » – à répondre aux exigences de signalement des documents décrits et de réutilisation de ces données, en adéquation avec les moyens assignés à la production de ces données.

- ***Accessibilité***

Le mouvement de l'*Open Data* s'est traduit en France par une politique gouvernementale d'ouverture des données publiques qui incite les administrations et établissements publics à assurer l'ouverture juridique et technique des données qu'ils produisent à la fois par souci de transparence vis-à-vis des citoyens et parce que cette démarche peut être génératrice de croissance économique. La BnF s'est voulue exemplaire dans l'application de cette politique et a adopté en 2014 la licence ouverte de l'État pour l'ensemble des métadonnées qu'elle produit. Selon les termes de cette licence, les métadonnées de la BnF sont désormais librement et gratuitement réutilisables, quel qu'en soit le format et le protocole de diffusion, pourvu que les réutilisateurs en mentionnent la provenance.

- ***Conformité***

En tant qu'Agence bibliographique nationale, la BnF assure un rôle de bibliothèque de référence pour le signalement des documents en sa possession, en particulier dans le domaine de l'édition et du patrimoine français. Dans ses processus de production de données, la BnF , dans le cadre de l'application de la législation en vigueur sur la production et la diffusion des données, met en œuvre un ensemble de standards, normes, codes et formats, généralement à vocation nationale ou internationale, qui régissent la structure de l'information bibliographique et les règles de description, d'indexation et d'interopérabilité de ces données. Ces formats sont précisés et, le cas échéant, adaptés pour certains contextes ou filières de production, par des consignes d'application au niveau de l'établissement. La BnF s'engage à ce que l'intégralité des documents formant le cadre normatif et fonctionnel de la description bibliographique soit mise à disposition pour consultation et référence.

En outre, l'ensemble des filières de description des documents entrés à la BnF par dépôt légal pratique un contrôle qualité post-production systématique des données produites.

- ***Cohérence***

La masse des données de la BnF est caractérisée par son hétérogénéité qui prend sa source dans la diversité des contextes historiques, normatifs et organisationnels de production et de provenance. La BnF poursuit un important travail de mise en cohérence de ces données produites dans des contextes extrêmement variés, à la fois par l'important travail de contrôle qualité évoqué plus haut et par des chantiers semi-automatisés de traitement de masse.

À ce titre, l'application progressive du modèle IFLA-LRM représente un fort axe de mise en cohérence des données visant, par le biais en particulier de *data.bnf.fr*, à une FRBRisation progressive des données bibliographiques de la BnF, selon un modèle exploitable en conformité avec les standards du Web de données.

- ***« Localisation »***

Le concept de « localisation » fait écho à celui, anglo-saxon, de *findability*, fondé sur le constat que la disponibilité des données et leur accessibilité par le biais d'outils de requête ne sont pas nécessairement des gages suffisants pour que les usagers les atteignent. La BnF met en œuvre une diversité de mesures et de traitements afin d'assurer la localisation de ses données : emploi d'identifiants et de référentiels normés, établissements de liens pérennes pour les points d'accès, indexation analytique des contenus fixée dans le cadre de la politique du sujet de la BnF. Dans cette optique, la FRBRisation des données documentaires de la BnF est un levier puissant d'amélioration de leur qualité tout en garantissant leur meilleure exposition sur le web de données et une expérience utilisateur plus satisfaisante.

- ***Traçabilité***

La BnF est au cœur d'un écosystème riche et varié de production et de diffusion des données, avec une diversité de profils et de compétences (producteurs primaires de données, agrégateurs, données fortement standardisées ou non, bases et référentiels portés par des établissements et des institutions du monde la culture, de la recherche, de l'édition…). Dans ce contexte, la BnF se dote d'un système de traçabilité porté par des méta-métadonnées permettant de renseigner l'origine, la nature et l'historique de ces données, de leur collecte à leur diffusion, en passant par la nature de leur traitement.

La politique de qualité des données en trois actions

Le contexte national et international de production des données bibliographiques connaît depuis plusieurs années des évolutions majeures. Un axe fort de la politique de qualité des données de la BnF est de s'appuyer de plus en plus sur les données sources produites par un écosystème de producteurs et agrégateurs exogènes, et de favoriser au maximum l'alignement de référentiels. Parallèlement, la nature des collections couvertes par l'obligation de Dépôt légal a considérablement crû avec l'adjonction du dépôt légal de l'Internet et des documents dématérialisés en 2006. Découle de ce contexte une massification des flux de données traités par la BnF.

Ce changement d'échelle implique nécessairement un changement de pratiques et une réflexion forte, entre autres sur l'organisation du travail afférente aux processus de production de données, le contrôle qualité de ces données et, plus fondamentalement, la logique

d'approche, la stratégie de description globale des documents ainsi traités. Dans ce cadre, le recours plus systématique, dès que le contexte le rend pertinent, aux traitements automatisés et semi-automatisés, au sourçage des données exogènes, aux traitements par fonds et non plus nécessairement à la pièce semblent des pistes permettant de maintenir l'adéquation entre les objectifs en matière de politique bibliographique de la BnF et les moyens qu'elle peut y consacrer.

- ***Un impératif : l'exhaustivité du signalement des collections***

La BnF est engagée depuis des années dans des chantiers importants de conversion et de catalogage rétrospectifs, permettant de signaler au public l'intégralité des collections dont elle dispose. À ce titre, la description des ensembles de collections prime sur l'exhaustivité dans la description de chaque document. Pour de nombreux fonds, il découle de cette priorité l'application d'une logique de description de type archivistique, accordant la primauté aux fonds et aux recueils sur la description à la pièce, dans l'optique de permettre un premier niveau d'accès à l'ensemble des ressources de la BnF. Ce signalement pourra progressivement être complété, détaillé et enrichi au cours d'opérations ultérieures de reprise dans le cadre de chantiers ciblés spécifiques. Ce signalement rétrospectif s'appuie également, dès que possible, sur les possibilités offertes par l'automatisation partielle ou complète de certains processus : flux entrants de données, alignements de référentiels, outils de reconnaissance et d'indexation automatique des images… L'ensemble de ces préconisations et les modalités de leur mise en œuvre sont portées par le groupe de travail « Exhaustivité du signalement ».

- ***Mise en place d'un référentiel qualité des données de la BnF***

La BnF vise à se doter à moyen terme d'un référentiel qualité, exprimé au niveau de chaque entité du catalogue, qui définisse pour les usagers de manière lisible, documentée et pertinente, de grands corpus cohérents de données fondés sur la qualité de leurs données et la nature des traitements – humains ou automatiques – qu'elles ont, le cas échéant, subis.

Cette démarche vise à aboutir à une labellisation BnF de jeux de données issus de contextes variés de production (Dépôt Légal, projets de recherche, description « scientifique » de documents pour la préparation d'exposition…) mais qui correspondent tous à un engagement fort de l'établissement sur la qualité (exactitude et richesse des données) des données ainsi exposées. *A contrario*, les catalogues de la BnF ont vocation à accueillir aussi des données provenant de sources externes (dépôt dématérialisé en masse par exemple) qui, à moins de reprise ultérieure dans le cadre de chantiers de corrections, ne bénéficieront pas de ce label.

- ***Politique du catalogue et documentation***

La BnF a produit en 2016 un document fixant sa politique de catalogage pour le catalogue général[3] et, en 2018, son pendant pour la politique d'indexation[4]. Il lui incombe maintenant d'adopter cette démarche pour l'ensemble de ses bases bibliographiques, dont au premier chef la base BnF Archives et manuscrits, afin d'assurer une cohérence de traitement de l'ensemble de ses fonds dans le respect des principes édictés par le présent document.

[3] http://www.bnf.fr/fr/professionnels/catalogage_pratiques_bnf/a.politique_catalogage.html.
[4] http://www.bnf.fr/fr/professionnels/catalogage_pratiques_bnf/a.politique_indexation.html.

Parallèlement, un important travail de documentation des consignes, procédures et traitements spécifiques à chaque filière et chaque service de production de données de la BnF est en cours depuis déjà quelques années. Ce processus doit être mené à terme afin de fournir aux usagers une documentation complète et cohérente afin de leur permettre la meilleure appréhension possible de la diversité de nature, de format, de présentation, de richesse d'information des données fournies par la BnF, ainsi que des cohérences internes qui les structurent.

Conclusion

Les données documentaires irriguent tous les processus, physiques ou numériques, qui permettent à la BnF de conduire ses missions. Il s'agit d'un domaine aux contextes de production, de diffusion et d'utilisation extrêmement diversifiés, marqué par de profondes évolutions :

- *Évolution quantitative* : la massification des entrées de données, en particulier *via* l'augmentation exponentielle de la production culturelle et le traitement des supports dématérialisés pose la question de la place, du coût humain et des finalités du contrôle qualité dans la chaîne de traitement.
- *Évolution des modèles, fondés sur l'IFLA-LRM depuis 2017*, décliné en France par le code de catalogage RDA-FR qui est adopté progressivement dans le cadre du programme national de Transition bibliographique.
- *Évolution des modèles d'alimentation des bases bibliographiques* : récupération de données exogènes, alimentation par flux de données, co-production et alignement de jeux de données et de référentiels.
- *Évolution des usages des données exposées* : localisation et consultation de ressources, mais aussi récupération de notices, fouille de données. Dans ce contexte, les attentes en matière de « fraîcheur » des données, ainsi que de transparence et de documentation des processus de production, se font de plus en plus prégnantes.

Face à ces évolutions majeures et structurantes, la BnF, en tant qu'Agence bibliographique nationale, co-pilote de la Transition bibliographique et bibliothèque de référence à l'échelle internationale, maintient l'exigence de qualité de sa production de données en formalisant cette politique. Elle vise à présenter une vision globale et cohérente tenant compte de tous les contextes et filières de production et de toutes les interfaces de diffusion ; elle s'appuiera sur un référentiel qui permettra d'en exprimer les déclinaisons en totale transparence par rapport à ses partenaires et ses utilisateurs, en se fondant primordialement sur trois axes forts : le sourçage des données, la nature et le degré de complétude des traitements dont elles ont bénéficié, et les modalités du contrôle qualité auquel elles ont, le cas échéant, été soumises.

Octobre 2018