

Les archives et le web sémantique



Coordination :

Violette Lévy
 Experte fonctionnelle
 Archives nationales

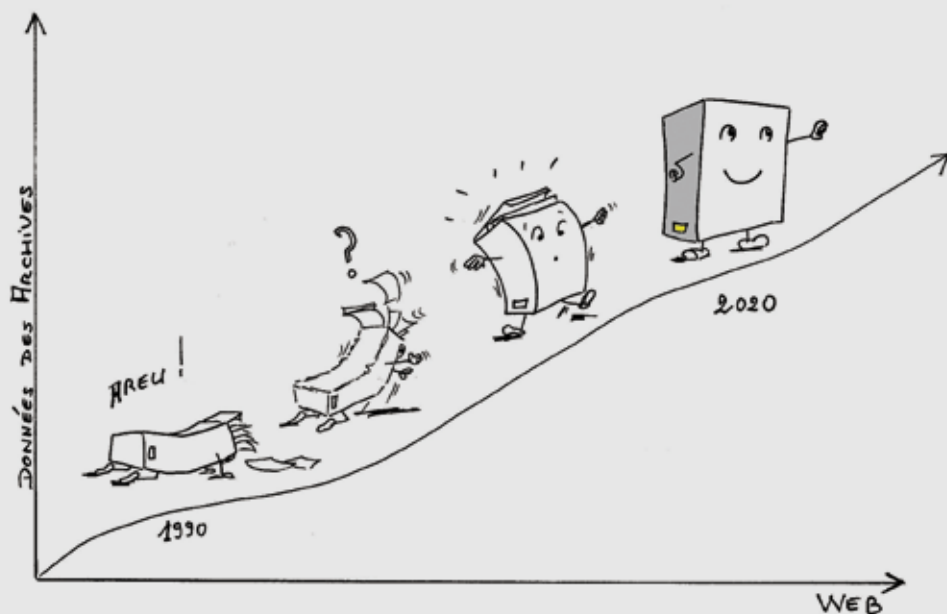


Introduction

Consacrer un dossier d'*Archivistes !* aux enjeux du web sémantique pour les archives pouvait sembler compliqué. Entre les experts susceptibles d'être déçus par ce qu'ils considéreraient – à juste titre – comme un survol du sujet, et les novices potentiellement effrayés par son caractère technique, comment intéresser la majorité de nos lecteurs ?

Un objectif pédagogique a été poursuivi : proposer un état de la réflexion clair et synthétique, afin de montrer pourquoi et comment des archivistes s'emparent aujourd'hui des technologies du web sémantique. Avec des exemples concrets de mise en œuvre, l'idée est également de démontrer que chacun de nous peut monter en compétences sur ce sujet, pour peu qu'il sache pourquoi s'orienter dans cette voie. Le dossier laisse aussi s'exprimer des points de vue divergents, qui formulent certaines critiques à l'encontre de ces technologies, et ouvre les perspectives à d'autres secteurs culturels et à l'international.

Puisse chaque adhérent de l'AAF, en lisant les pages qui suivent, se faire sa propre idée de l'intérêt que peut présenter le web sémantique pour les archives, selon son propre contexte.



À l'âge du web sémantique © Soaz



Lexique

Web sémantique :
 extension du web qui vise à faciliter l'exploitation des données structurées, pour donner du sens au contenu des pages web, en permettant leur interprétation par des machines.

RDF (resource description framework) :
 modèle de graphe destiné à décrire formellement les ressources web et leurs métadonnées, afin de permettre le traitement automatique de ces descriptions. Langage de base du web sémantique, développé par le W3C (world wide web consortium).

Triplet :
 unité de données la plus petite contenue dans un graphe de type RDF, au sein d'une base de données de type *triplestore* (voir ci-dessous). Association d'un sujet (qui représente la ressource à décrire), d'un prédicat (qui représente un type de propriété applicable à cette ressource) et d'un objet (qui représente une donnée ou une autre ressource).

SPARQL (SPARQL protocol and RDF query language) :
 langage de requête et protocole permettant de rechercher, d'ajouter, de modifier ou de supprimer des données RDF disponibles à travers Internet.

Triplestore :
 base de données spécialement conçue pour le stockage et la récupération de données RDF.

SKOS (simple knowledge organization system) :
 recommandation du W3C pour représenter des thésaurus documentaires, s'appuyant sur le modèle de données RDF.

Source : Wikipédia

Pourquoi et comment les archivistes s'emparent-ils de ces technologies ?

Archives et web sémantique : qu'en pensez-vous ?

Pour ouvrir ce dossier, on a souhaité interroger les archivistes et les usagers des archives pour savoir ce que leur inspirait l'association des termes « archives » et « web sémantique ». Voici quelques réponses à cette question.



Violette Lévy
@archiviolette

...

Avis aux [#archivistes](#) et [#usagers](#) des archives. Si l'on vous dit « Archives et web sémantique », qu'est-ce que cette association vous inspire ? Merci d'avance de vos retours qui alimenteront le prochain dossier de la revue Archivistes ! de [@Archivistes_AAF](#) à paraître en janvier 😊



Julien Pomart @JulienPom... · 8 déc. ...
En réponse à [@archiviolette](#) et [@Archivistes_AAF](#)

Qu'on parle beaucoup de structuration des données, moins qualité et circuit de production. Il faudrait tirer le bilan de l'utilisation (ou non) de l'EAD par les archivistes pour ne pas répéter les mêmes erreurs



Élise W. @_onemorez · 8 déc. ...
En réponse à [@archiviolette](#) et [@Archivistes_AAF](#)

Accessibilité !? En tout cas c'est ce qui me vient en premier (la ponctuation compte).



Maud Profizi Jouve @Maud... · 8 déc. ...
En réponse à [@archiviolette](#) et [@Archivistes_AAF](#)
[@benjegouzo](#) 😊



Thinking Twice @TwiceThin... · 8 déc. ...
En réponse à [@archiviolette](#) et [@Archivistes_AAF](#)
Un long chemin



Christelle B. @BruantC · 8 déc.

En réponse à [@archiviolette](#) et [@Archivistes_AAF](#)

Mapping mapping mapping... identifiants identifiants identifiants... balises balises balises... triplets triplets triplets... 🤖



Christelle B. @BruantC · 8 déc.

Des fonds au fond du fond du fond... des rebonds, des rebonds, j'suis perdue j'suis perdue j'suis perdue... autre version 😊



Nadia @Nadia_Fafi · 9 déc. ...

En réponse à [@archiviolette](#) et [@Archivistes_AAF](#)

Extraction des données situées dans les archives et organisation de ces données en ontologies = puit de données à interroger en sparkle.



Célia Cabane @CabaneCelia · 9 déc. ...

En réponse à [@archiviolette](#) et [@Archivistes_AAF](#)

Interopérabilité avec d'autres logiciels/systèmes... et travail en coopération avec les autres acteurs culturels, a minima français ! [#rejoigneznous](#) [#onadescookiesetonestgentils](#)



Simon Remy @Simmmbull · 8 déc. ...

En réponse à [@archiviolette](#) [@LadyNamite2](#) et [@Archivistes_AAF](#)

C'était le sujet de mon mémoire du master pro de Mulhouse... en 2014.



Violette Lévy @archiviolette · 8 déc. ...

Super intéressant ! Votre mémoire est-il consultable ? Si oui, où ? Merci !



Simon Remy @Simmmbull · 8 déc. ...

Il doit l'être à la Bu fonderie à Mulhouse. Mais je l'ai en pdf si besoin. Je préfère préciser tout de suite que ce mémoire n'est pas fifou. Je l'ai fait en pur néophyte, je n'y connaissais rien en web sémantique avant de me lancer là dedans. 😊

La stratégie sémantique du ministère de la Culture

Le ministère de la Culture s'est très vite saisi, dès 2012, de la question de l'application du web sémantique aux données culturelles. Certains établissements culturels tels que la Bibliothèque nationale de France¹ ou le Centre George Pompidou avaient ouvert la voie en expérimentant l'application des standards et technologies du web sémantique à leurs collections et œuvres. Ces premiers retours d'expérience ont donné le ton pour envisager une généralisation à l'ensemble du secteur culturel.

L'objectif de la stratégie sémantique du ministère et de la feuille de route pour les « Métadonnées culturelles et transition web 3.0 », publiée en 2014, est donc de sensibiliser, former et mobiliser l'ensemble des acteurs culturels pour la mise en œuvre du web sémantique pour leurs données. Cet objectif sert plusieurs enjeux stratégiques. Au-delà du simple gain en visibilité des fonds et collections, c'est leur « découvrabilité » qui est en jeu. Dans un écosystème essentiellement anglophone, il est crucial que le patrimoine culturel français et francophone soit visible et découvrable : le recours au web sémantique est un moyen de répondre à cet enjeu. Le ministère de la Culture a mis en place un programme de formation pour permettre une montée en compétences des professionnels et tendre vers une harmonisation et une mutualisation des efforts pour des données structurées riches et interconnectées. Ce réseau de données interconnectées constituerait à long terme le « graphe culture », un graphe de connaissances culturelles.

Aujourd'hui le chantier autour de la feuille de route est toujours en cours et connaît un second souffle, puisqu'elle entre dans une phase d'actualisation. Les standards du web sémantique se sont plus largement déployés ; les technologies ont évolué dans la mesure

où les solutions logicielles intègrent tout ou partie de ces standards ; les systèmes informatiques étant plus performants, les applications sémantiques gagnent également en performance ; et, enfin, les acteurs culturels ont gagné en maturité sur le sujet : ce contexte s'avère favorable pour réancrer la stratégie ministérielle pour le web des données liées dans des problématiques actuelles et aller plus avant sur de nouveaux axes de

travail. La construction du « graphe culture », la question des interfaces et de la visualisation des données liées ou encore la traçabilité des données numériques² sont quelques-uns de ces nouveaux axes.

Les archives occupent une place essentielle dans cette stratégie de par la grande diversité des documents qu'elles manipulent ainsi que par leur caractère transverse. Le niveau de structuration de ces données est déjà très fort et le réseau des professionnels des services d'archives est très solide avec des formations complémentaires à celle du ministère organisées régulièrement, si bien que la transition vers le web

« Les archives occupent une place essentielle dans cette stratégie de par la grande diversité des documents qu'elles manipulent ainsi que par leur caractère transverse. »



État de l'art sur la traçabilité des données numériques

© Publication du ministère de la Culture

sémantique se fait assez naturellement. Les archives ont aussi une expertise notable sur la question de la traçabilité des données numériques.

Une première étape pour les services d'archives qui n'ont pas encore franchi le cap du web sémantique est de se former via le parcours proposé par le ministère. Une seconde étape consiste à solliciter le département de l'innovation numérique qui pourra orienter et conseiller sur la marche à suivre en s'appuyant sur le Service interministériel des Archives de France ou encore sur les Archives nationales, qui sont pleinement investis dans les travaux de la feuille de route sémantique.



Marie-Véronique Leroi

Chargée de mission web sémantique pour les données culturelles
Département de l'innovation numérique
Ministère de la Culture

1. Mise en place de data.bnf.fr

2. Traçabilité des données numériques : expliciter la provenance des données numériques et documenter avec des métadonnées spécifiques tout changement ou modification de ces métadonnées.

L'orientation web sémantique de FranceArchives

Un des axes majeurs de la feuille de route de FranceArchives jusqu'en 2023 est d'en faire le portail de référence pour les métadonnées archivistiques françaises dans le web de données. Pour cela, il est confronté à plusieurs enjeux d'ordre technique et économique.

The screenshot shows the FranceArchives website interface. At the top, there's a search bar with the text 'Rechercher'. Below it, a blue header contains the FranceArchives logo and a 'MENU' button. The main content area displays a biographical notice for 'Bougainville, Louis Antoine de (1729-1811)'. The notice includes fields for 'Date de naissance' (12 novembre 1729), 'Date de décès' (31 août 1811), 'Description' (Louis-Antoine de Bougainville (1729-1811) se fit d'abord recevoir avocat, puis entra comme aide-major dans le régiment de Picardie (1753). Il fut peu après envoyé à Londres comme secrétaire d'ambassade, et la publication de son *Traité du calcul intégral* (1754-1756) l'y fit...), 'Activité' (explorateur ; militaire ; sénateur), 'Source' (Archives nationales : Bougainville, Louis Antoine de (1729-1811)), and 'Autres ressources' (Wikidata : Louis-Antoine de Bougainville, data.bnf.fr : Louis Antoine de Bougainville (1729-1811)).

Exemple d'alignement d'une notice biographique vers Wikidata et data.bnf.fr © FranceArchives

FranceArchives, portail agrégateur, a pour mission de proposer à tout utilisateur, néophyte ou chercheur habitué des archives, un accès fédéré à l'ensemble des métadonnées archivistiques françaises. Les standards du web sémantique lui apportent quelques réponses aux défis que représentent l'hétérogénéité et la volumétrie des données à traiter : onze millions de descriptions produites par près de quatre-vingt-dix services contributeurs fin 2020.

Faire de FranceArchives le portail de référence des métadonnées archivistiques françaises dans le web de données est un des axes forts du projet depuis son origine. Cet axe intègre, en particulier, la traduction en RDF des métadonnées en utilisant l'ontologie RiC-O et la mise en place d'un serveur SPARQL.

En effet, la machine est maintenant un intermédiaire indispensable entre l'archiviste auteur de la description et l'utilisateur final. Le principal enjeu est de proposer une alternative à l'affichage des instruments

de recherche linéaires et à l'approche classique, hiérarchisée, des archives pour favoriser les rebonds d'une entité à l'autre et la sérendipité de la recherche. Un internaute ignorant la complexité inhérente à la description archivistique arborescente et ne sachant pas formuler précisément ce qu'il

cherche pourrait construire sa requête pas à pas grâce à une interface d'interrogation simple et intuitive, tandis qu'un utilisateur expert pourrait manipuler à sa guise d'énormes masses de données structurées. FranceArchives participerait ainsi pleinement à la politique de diffusion et d'ouverture des données publiques, quel que soit le profil de l'internaute.

Mettre en valeur des lieux, des dates, des thèmes ou des agents implique un important travail d'harmonisation, de désambiguïsation et d'enrichissement des données, à l'instar de celui effectué, de longue date, par les bibliothèques. La puissance et la souplesse des

outils du web sémantique présentent des opportunités intéressantes pour sortir de la logique de silo des descriptions propres à chaque institution : par exemple, fusionner les entrées d'index relatives à Henri Beyle, dit Stendhal, et indiquer à la machine et à l'utilisateur final qu'il s'agit du même personnage, puis le relier à des bases externes pour en récupérer des informations biographiques¹. À son tour, par le liage des données, FranceArchives pourra enrichir d'autres bases en proposant des liens vers des ressources archivistiques. Le choix de l'ontologie RiC-O pour traduire en RDF des métadonnées décrites en EAD et en EAC s'est imposé, car elle permet de sémantiser les entités et leurs relations temporelles, hiérarchiques ou d'équivalence. Il est ainsi possible de rendre compte plus subtilement de la contextualisation des archives si chère à tout archiviste. Apparue plus tard que les modèles des bibliothèques (FRBR en 1992 fondu dans IFLA-LRM en 2017) et des musées (CIDOC-CRM en 1996), cette ontologie répond à l'enjeu d'interopérabilité et est un pas en avant pour sortir les archives de leur isolement technologique par rapport aux autres institutions culturelles. Le prototype PIAAF² en 2018 et l'outil RiC-O : *Convertir* mis au point en 2019 par les Archives nationales en ont montré tout l'intérêt et la faisabilité.



Manonmani Restif

Cheffe de projet FranceArchives
Service interministériel des Archives
de France

« La machine est maintenant un intermédiaire indispensable entre l'archiviste auteur de la description et l'utilisateur final. »

1. Pour en savoir plus : Manonmani Restif et Coline Silvestre, « FranceArchives : portail agrégateur et laboratoire de la qualité des données sur le web », *Archivistes* 1, juillet-septembre 2020, n° 134, p. 23.

2. « Archives et web de données. Mise en ligne du prototype "PIAAF" », FranceArchives, mis en ligne le 2 mars 2018 [https://francearchives.fr/fr/actualite/87098542].

RiC aux Archives nationales de France : perspectives pour 2021

Alors que le groupe d'experts sur la description archivistique (EGAD) du Conseil international des archives (ICA) s'apprête à publier la version 0.2 de l'ontologie *Records in Contexts-Ontology* (RiC-O), voici quelques informations sur les travaux de mise en œuvre de RiC prévus en 2021 aux Archives nationales.

Pourquoi choisir de « passer à RiC » et spécialement à RiC-O ?

Si vous souhaitez gagner en précision et en utilisabilité (donc en qualité), réduire les redondances et décloisonner les silos de métadonnées de description archivistique qui sont juxtaposés dans votre système d'information, en regroupant les métadonnées par objet décrit et en exprimant précisément les relations entre ces objets, le modèle RiC-CM constitue un cadre conceptuel de référence extrêmement utile. Il aide à définir rigoureusement les grandes lignes de ce qui pourra être votre futur système de gestion des métadonnées, à évaluer la distance qui sépare l'existant de ce système et à élaborer une stratégie pour réaliser pas à pas cette évolution.

RiC-O est la transposition formelle de RiC-CM, qui permet dès aujourd'hui de produire des jeux de données RDF conformes à ce cadre conceptuel afin de les exposer dans le web de données liées. Il propose des composants et mécanismes plus précis que ceux de RiC-CM. Il est aussi flexible : il offre parfois deux solutions, l'une simple, la deuxième plus précise et complexe, pour représenter une information, ce qui est important dans un contexte de transition où l'on part souvent de métadonnées plus ou moins structurées et contrôlées d'un instrument de recherche à un autre.

Les jeux de données RDF peuvent être interrogés et explorés en dépassant les limites techniques des systèmes d'information archivistiques actuels, notamment en parcourant le graphe RDF produit (comme le montre l'image ci-après). Il est aussi possible d'enrichir ces jeux de données en établissant des relations vers les données RDF d'autres institutions ou en inté-

grant dans ces jeux de données locaux des données RDF présentes dans d'autres systèmes d'information. RiC-CM ou RiC-O peuvent être utilisés en partie seulement, étendus (on peut définir des sous-classes et sous-propriétés de RiC-O pour servir des besoins spécifiques) ou encore employés en combinaison avec d'autres modèles, comme l'ontologie PREMIS pour les métadonnées de préservation¹. Des *mappings* entre RiC-O et les ontologies décrivant des domaines voisins seront d'ailleurs établis en 2021 afin de faciliter ces combinaisons.

Enfin, RiC est un standard international, dont l'adoption progressive par les services d'archives et portails devrait faciliter l'interconnexion entre jeux de données archivistiques ou avec d'autres jeux de données, ainsi que le partage d'outils ou de savoir-faire. Il

a vocation à évoluer avec les besoins de ces services et projets. Il est important que ce soit l'ICA, une organisation non gouvernementale internationale, qui le maintienne sur le long terme, par le biais des représentants du monde des archives et du *records management* qui forment le groupe EGAD, avec et pour ces communautés. Ce travail est mené dans la continuité des quatre normes précédentes, elles-mêmes élaborées par l'ICA et transposées en schémas XML (EAD et EAC-CPF), si bien que tout élément d'information présent dans des fichiers EAD ou EAC-CPF trouve un composant qui lui correspond dans RiC-CM et RiC-O. À l'inverse, à côté de RiC-O, on devrait pouvoir disposer d'ici quelques années de nouvelles versions des schémas EAD et EAC-CPF, plus proches de RiC, élaborées par la Society of American Archivists en liaison avec le groupe EGAD de l'ICA.

« RiC est un standard international, dont l'adoption progressive par les services d'archives et portails devrait faciliter l'interconnexion entre jeux de données »

1. www.loc.gov/standards/premis/ontology/index.html



Les ressources à consulter

Sur RiC-O

- La dernière version officielle en date (RiC-O 0.2, janvier 2021) [www.ica.org/standards/RiC/ontology].
- Le dépôt public sur GitHub [<https://github.com/ICA-EGAD/RiC-O>].
- Le site web d'information [<https://ica-egad.github.io/RiC-O/>].

Sur RiC-O Converter

- La présentation faite en décembre 2020 dans le cadre de SemWebPro 2020 [<https://peertube.semweb.pro/videos/watch/c195e540-4691-4b38-9619-0a59300d1adc>].
- L'enregistrement vidéo du webinaire du 25 juin 2020 [www.dailymotion.com/playlist/x6x1d0].
- Le code source et la documentation du logiciel [<https://github.com/ArchivesNationalesFR/rico-converter>].
- les *mappings* EAD 2002 vers RiC-O et EAC-CPF vers RiC-O [<https://github.com/ArchivesNationalesFR/rico-converter/tree/master/ricoconverter/ricoconverter-doc/src/main/resources>].

Ce sont les raisons pour lesquelles les Archives nationales ont décidé de se servir de RiC-CM et de RiC-O dans leur démarche d'évolution progressive des métadonnées descriptives qu'elles ont accumulées, produisent et mettent à jour, vers un graphe d'entités liées. Elles ont déjà converti l'ensemble de leurs instruments de recherche, notices d'autorité (via le logiciel RiC-O *Converter*) et vocabulaires en jeux de données RDF/RiC-O, sans modifier les fichiers sources.

Le programme de travail sur RiC aux Archives nationales en 2021

Parmi les travaux prévus aux Archives nationales en 2021, on peut citer :

1. la poursuite des travaux de modélisation de la future version des référentiels, y compris de nouveaux référentiels (comme celui des relations) ;
2. la poursuite des travaux de construction collaborative d'un référentiel des fonctions et d'enrichissement des référentiels de lieux via leur alignement avec des référentiels externes sémantisés ;
3. la publication sous licence libre des référentiels dans une version RDF utilisant RiC-O et SKOS comme modèles de référence ;
4. la mise en conformité du logiciel libre RiC-O *Converter* avec RiC-O 0.2 ;
5. la poursuite des réflexions sur le « passage à RiC » des métadonnées de description des documents, en incluant les métadonnées des archives nativement numériques ;
6. la réalisation, en collaboration avec la société Sparna, la BnF et le ministère de la Culture, d'améliorations du composant JavaScript Sparnatural, qui, une fois intégré à une page web, permet à un utilisateur final d'éditer de manière simple et intuitive des requêtes SPARQL, sans qu'il ait besoin de connaître ce langage de requête ni le modèle ontologique auquel les données interrogées sont conformes ;
7. la mise en place d'un démonstrateur Sparnatural pour un jeu significatif de données RDF/RiC-O des Archives nationales ;
8. la réalisation, avec l'équipe ALMAAnaCH d'Inria, d'un logiciel libre configurable de détection, de classification et de résolution d'entités nommées (par l'utilisation de référentiels tels que Wikidata) présentes dans des lots d'instruments de recherche en EAD.



Florence Clavaud

Responsable de la mission Référentiels
aux Archives nationales

Membre exécutif du groupe EGAD de l'ICA

Responsable de l'équipe de développement de RiC-O



Un exemple de requête SPARQL parcourant un graphe d'entités RDF/RiC-O, construite avec le composant Sparnatural, par une simple série de clics © Florence Clavaud/Archives nationales (sous licence CC BY 4.0)



RiC aux Archives nationales : en savoir plus

- Les présentations faites lors de la journée d'étude du 28 janvier 2020 sur le thème « Les métadonnées archivistiques en transition : le nouveau cadre normatif, les enjeux et les premières réalisations » [<https://labarchiv.hypotheses.org/1495>].
- Florence Clavaud, « Transformer les métadonnées des Archives nationales en graphe de données : enjeux et premières réalisations », *La Gazette des archives*, n° 254 « Les Archives nationales, une refondation pour le XXI^e siècle » (2019-2), p. 59-88.



Les projets RiC-O à l'étranger

Vous souhaitez savoir ce qui se passe hors de France en ce qui concerne l'utilisation de RiC-O, qu'il s'agisse de preuves de concept ou de programmes de recherche, d'implémentations au sein de systèmes d'information archivistiques ou de développement d'outils ?

Consultez les pages consacrées à ces sujets sur le site web d'information relatif à RiC-O [<https://ica-egad.github.io/RiC-O/>]

Si vous souhaitez y faire référencer un projet d'utilisation de RiC-O en France ou à l'étranger, contactez Florence Clavaud.

Retrouvez en complément, sur notre site Internet, dans l'espace dédié à ce numéro, l'article « RiC-O et ses utilisations hors de France : une petite liste ».



Anne-Marie Bruleaux, directrice des Archives départementales de l'Yonne © DR

Se former au web sémantique : témoignage d'une archiviste départementale

Anne-Marie Bruleaux, directrice des Archives départementales de l'Yonne, a suivi en septembre 2020 le stage expert « Mettre en œuvre le web sémantique dans la description archivistique », proposé par l'AAF et dispensé par Florence Clavaud (Archives nationales) et Pascal Romain (Rhizome Data).

Qu'est-ce qui vous a amenée à choisir cette formation ?

J'ai la conviction que le web 3.0 va complètement révolutionner la mise à disposition et la recherche de l'information dans les prochaines années. Je souhaitais mieux connaître le modèle RiC et voir concrètement sa mise en œuvre pour préparer mon service à ces évolutions.

Que retirez-vous de ce stage ?

Sans aucun doute, une amélioration de mes connaissances théoriques et un aperçu des applications concrètes. Je regrette néanmoins que le temps de formation n'ait pas été suffisamment long et maîtrisé pour faire davantage d'exercices. J'aurais souhaité aussi bénéficier d'un « bac à sable » permettant de prolonger la formation par des expérimentations personnelles.

Si vous pouviez maintenant mettre en place une action concrète, quelle serait cette action ?

Je reste assez dubitative à ce sujet. Le renforcement du travail d'indexation me semble essentiel dans un premier temps. Mais j'ai l'impression que nous resterons encore longtemps bloqués par nos progiciels de gestion d'archives et de diffusion qui ne sont pas adaptés, nos fournisseurs n'ayant pas tous, loin de là, perçu la nécessité urgente de faire évoluer leur offre en intégrant le modèle RiC. À la question « Que pouvons-nous faire concrètement pour préparer notre service ? », il nous a été répondu de ne plus faire de tableaux de gestion, mais des profils d'archivage selon le SEDA, même pour les versements papier (qui vont se poursuivre encore pendant plusieurs années). Je n'ai pas compris cette réponse et j'aurais aimé avoir un peu de temps pour creuser la question avec le formateur. Dans l'immédiat, il me semble que mon action va se borner à faire passer l'information que j'ai pu retenir auprès de mes collègues.

Propos recueillis par Violette Lévy

Des points de vue multiples

RiC et moi

Vu des archives locales, le web des données apparaît comme un horizon lointain sur lequel il est encore bien prématuré de se pencher. Il reste tant à faire et à décrire dans nos fonds, qu'il ne paraît pas franchement urgent de changer à nouveau le modèle d'encodage de nos inventaires.

Nous n'avons pas terminé d'adopter le XML-EAD qu'il faudrait déjà passer à RiC, alors même que nos logiciels ne le permettent pas et qu'il n'existe aucun outil capable d'interpréter les données ainsi modélisées. L'intérêt d'un tel changement est d'ailleurs assez flou, puisque personne ne perçoit précisément les gains en matière de recherches. Cela n'a donc rien d'une priorité, et RiC est avant tout une histoire de spécialistes qui disposent d'assez de temps pour imaginer le futur.

Ce résumé est certes un peu caricatural, mais il rend compte d'un sentiment diffus bien réel dans les services. Il est urgent d'attendre, faute de pouvoir apprécier le rapport entre efforts et profits. Il est par ailleurs parfaitement usant d'avoir à remettre éternellement sur l'écheveau des travaux que l'on pensait clos. En résumé, le passage au web des données n'est pas une demande du terrain, sauf peut-être d'une poignée d'archivistes qui ont développé un intérêt personnel pour ces questions et qui en perçoivent mieux les enjeux.

Or ces enjeux sont importants et dépassent le seul changement de format. Le XML-EAD, qui est déjà d'un âge respectable selon les standards du web, correspond à une phase précoce de l'informatisation des données, au cours de laquelle nous avons transposé nos pratiques traditionnelles sur écran. Le XML-EAD est un format hiérarchisé non pas pour des questions de performances informatiques, mais pour entretenir les méthodes de travail issues d'ISAD(G). Il nous permet de traiter une partie des informations selon une logique de base de données, tout en conservant nos descriptions très littéraires – et donc peu exploitables par un ordinateur. En d'autres termes, nous avons créé un format qui ressemblait au plus près à nos inventaires papier.

RiC-O et le web sémantique d'une manière plus générale proposent une évolution plus profonde. Ils orientent notre travail vers la production de données structurées et non plus de « simples » descriptions. Autrement dit, ils proposent d'abandonner à terme une logique d'inventaire – outil conçu pour être lu – au profit d'une vaste base de données – outil conçu pour être interrogé. Il ne s'agit donc pas que d'une question de format : les éditeurs seront capables, une fois le modèle stable, de proposer des exports à partir

des données déjà existantes de nos systèmes d'information archivistique (SIA) ou du XML-EAD vers RiC. Il s'agit plutôt d'une évolution de notre manière de décrire nos fonds.

Il n'y a en fait rien de nouveau en matière d'informations : RiC s'appuie sur les normes de description déjà existantes (ISAD, ISAAR, ISDIAH, ISDF), et qui font déjà pleinement partie de notre paysage normatif. Mais il permet d'abolir les frontières entre les différents instruments de recherche en globalisant l'interrogation des données de toutes origines, et vient terminer le travail de normalisation débuté il y a des années en l'intégrant pleinement aux modèles de données actuels.

Alors, pourquoi changer ? Parce que le web sémantique ne nous offre pas un nouveau modèle de recherche dans nos fonds, mais la possibilité d'en créer une infinité. Ce que nous permettra RiC (ce que permet en partie seulement le XML-EAD), c'est de faire de la dentelle, plutôt que d'offrir une interface unique d'accès à des

inventaires. Nous avons créé des bases il y a des années de cela pour donner accès aux images numérisées : il s'agissait bien de données archivistiques, mais dont nous avions retravaillé et affiné la structuration. RiC doit nous permettre d'étendre cela à tous nos fonds en manipulant la manière de les interroger plutôt qu'en les retravaillant pour chaque usage.

Les gains sont donc encore à créer, et c'est à chacun de nous qu'il appartiendra de les imaginer. Mais, dans les faits, le changement est déjà en cours parce qu'il repose d'abord sur l'évolution de pratiques bien avant que de changer le format d'encodage : atomiser encore un peu plus l'information en de multiples champs selon un vocabulaire plus contrôlé, indexer plus massivement, renseigner de manière systématique certains champs jusque-là un peu oubliés.

En résumé, préparer le passage à RiC et au web sémantique, c'est continuer de faire ce que nous faisons depuis toujours : améliorer sans cesse la qualité de nos descriptions.



Gaël Chenard

Directeur

Archives départementales de la Vienne

Jean-Séverin Lair : « Le web sémantique n'est qu'une forme de description pour échanger de la donnée »



Intervention de Jean-Séverin Lair lors de la table ronde « Les enjeux du web de données pour la culture et les langues » (19 novembre 2012)

© TheSupermat, via Wikimedia Commons

Ancien directeur du programme Vitam puis du programme Tech.Gouv, actuellement directeur du système d'information de l'INSEE, Jean-Séverin Lair livre ici sa vision des technologies du web sémantique, et en démontre les limites, notamment dans le domaine de l'archivage numérique.

Pour commencer, quel est votre point de vue d'informaticien sur les technologies du web sémantique ?

Je pourrais, à mon habitude, vous donner un avis à l'emporte-pièce... « Une technologie considérée comme révolutionnaire dont l'usage n'a pas explosé en vingt ans dans le monde Internet est morte ! »

Celui qui le dit avec beaucoup plus d'arguments, c'est Gautier Poupeau, lui qui a pourtant été un des papes du web sémantique¹.

Alors que beaucoup ont rêvé que tout deviendrait lien RDF pour un univers hyperconnecté où toute l'information serait utilisable pour de l'inférence logique, il a fallu se rendre à l'évidence que cela ne serait pas le cas. Il y a un certain nombre de problèmes : la charge humaine pour créer de bons liens de l'information, l'incertitude des choix sur les liens faits par l'homme, la sensibilité des traitements de ces données aux er-

reurs contenues, la volumétrie des données à traiter, la capacité des solutions techniques à traiter de très grandes volumétries...

Qui plus est, il faut bien admettre que le fait de ne pas passer à cette vision hyperstructurée ne semble pas empêcher l'évolution numérique. Aujourd'hui, on s'appuie surtout sur l'usage de données massives brutes organisées automatiquement avec des algorithmes d'intelligence artificielle. L'IA actuelle ne s'appuie que peu sur les données structurées, contrairement à ce que l'on a cru à un moment.

Un point important à noter, la limitation de la volumétrie acceptable dans les systèmes d'information traitant le web sémantique semble difficilement, voire totalement, indépassable... En informatique, quand on veut traiter de grandes volumétries, il est possible soit de gonfler la machine (avec davantage de mémoire, de processeurs, etc.), ce qui plafonne très vite et fait monter les coûts de manière non linéaire, soit on multiplie les petites machines élémentaires (et virtuelles), selon les besoins. Le problème est que la manipulation des graphes sémantiques ne se divise pas bien sur plusieurs machines, car l'information est un ensemble hyperconnecté où il n'y a pas de frontières claires où faire le découpage.

1. « Les technos du web sémantique ont-elles tenu leurs promesses ? », *Les Petites Cases*, 6 octobre 2018 [www.lespetitescases.net/les-technos-du-web-sémantique-ont-elles-tenu-leurs-promesses] ; « Why I don't use Semantic Web technologies anymore, even if they still influence me ? », *Les Petites Cases*, 12 décembre 2019 [www.lespetitescases.net/why-I-dont-use-semantic-web-technologies-anymore-even-if-they-still-influence-me].

Le meilleur exemple de la rencontre de cette limite est le système SPAR à la Bibliothèque nationale de France². Spar gère son graphe sémantique sur un gros serveur de calcul et pourtant ne peut pas faire de requêtes sur la base complète des métadonnées en moins de quelques heures, voire quelques jours... De ce fait, la BnF a fait une base réduite de métadonnées en limitant les champs présents et liés, qui est requêteable en quelques secondes.

En réalité, le web sémantique et les données liées sont intéressants sur des informations de volumétrie modérée, mais à forte valeur ajoutée. Il faut savoir mixer les modèles et sans doute avoir des référentiels sémantisés, à côté de grands gisements de données plus traditionnels (et efficaces !).

En tant qu'ancien directeur du programme d'archivage numérique Vitam, comment diriez-vous que le programme a abordé le sujet du web sémantique ?

Dans le projet Vitam justement, l'architecte qui a présidé à la conception nous a rapidement mis en garde contre les risques et les limitations des technologies sémantiques. Nous avions à cette occasion étudié beaucoup de solutions de base pour les très grandes volumétries, et nous avons fait le choix de solutions NoSQL qui ont depuis conquis une grande part du marché (MongoDB et Elasticsearch).

Il est intéressant de voir que pour atteindre de très grandes volumétries, on est obligé de limiter certaines exigences, pourtant jugées primordiales sur les bases SQL. À titre d'exemple, on ne peut plus être sûr, dans une base NoSQL, que l'on trouve en réponse une donnée dès qu'elle a été écrite. Parfois cela peut mettre quelques secondes pour être pris en compte et, pendant cette durée, certaines requêtes peuvent retourner cette donnée toute nouvelle et d'autres pas... Mais est-ce si grave ? Relâcher ces contraintes permet de distribuer facilement et sans surcoût excessif les calculs sur un nombre *n* de serveurs qui mettent un certain temps à se synchroniser³.

D'une manière générale, le traitement des très grandes volumétries oblige à faire moins structuré et avec moins d'exigence de perfection de traitement, un peu comme dans la nature... On est bien là dans une démarche contraire à celle du web sémantique.

Que vous inspirent les exemples d'expérimentations ou de réalisations en matière de sémantisation qui commencent à apparaître dans certains services d'archives ?

La sémantisation présente peut-être un intérêt sur des référentiels de volumétrie réduite, facilement gérable et actualisable par l'homme. Mais ceci reste à démontrer, et avant de rentrer dans une démarche trop formelle, donc difficilement comprise, et consommatrice de ressources humaines, il faudrait faire la

preuve de cette utilité. Tous les grands du numérique fonctionnent maintenant par preuve de concept⁴ et expérimentation.

J'ai peur que la démarche actuelle soit portée par des hyperspécialistes qui ne répondent plus à des objectifs concrets et sont loin du terrain. À titre personnel, j'ai vu une présentation sur les schémas applicables à des structures sémantiques faite par les Archives nationales. On était clairement dans de l'informatique théorique, qui avait plus sa place dans une thèse du domaine que dans une pratique professionnelle.

Pour terminer par une provocation de plus, je dirais : « Attention à ne pas refaire ce qui a été fait avec le XML ! » La profession archivistique s'est entichée de cette technologie, qui fut à une époque l'alpha et l'oméga de la transmission de données. Elle l'a enseignée à ses métiers, elle a poussé des prestataires à l'utiliser pour des bases archivistiques, mais finalement cela a gêné les dernières évolutions dans le domaine de l'archivage numérique... En effet, d'une part les systèmes d'information récents ne font plus du XML, mais du JSON, qui n'en est pas loin, mais avec des nuances complexes, ce qui a compliqué les développements Vitam. Il y a d'ailleurs fort à parier que dans cinq ans, il y aura encore autre chose que

du XML ou du JSON. D'autre part, le XML est un standard d'échange de données, et non un standard de stockage et de manipulation de la donnée. Les bases XML ont été des catastrophes en matière d'efficacité et de solidité et ont, sauf anecdotiquement, disparu. De la même manière, on pourrait dire que le RDF n'est qu'un format du moment et que le web sémantique n'est qu'une forme de description pour échanger de la donnée, qu'il

faut sans doute utiliser sous une autre forme pour des raisons d'efficacité.

J'aurais donc une requête à faire aux archivistes... Appuyez-vous sur les informaticiens pour choisir et faire évoluer les technologies utilisées, et soyez davantage centrés sur vos exigences et vos besoins fonctionnels, sur la métadonnée, sa nature et sa collecte. Je pense que c'est une grande part du succès du programme Vitam, au sein duquel il y avait un réel respect des rôles et des compétences de chacun entre les archivistes et les techniciens, et notamment entre les deux têtes fonctionnelle, Édouard Vasseur, et technique, Frédéric Brégier. On a donc une solution très efficace techniquement pour réaliser les processus essentiels des archivistes.

Propos recueillis par Violette Lévy

« En réalité, le web sémantique et les données liées sont intéressants sur des informations de volumétrie modérée, mais à forte valeur ajoutée. »

2. Système de préservation et d'archivage réparti [www.bnf.fr/fr/spar-systeme-de-preservation-et-darchivage-reparti].

3. https://fr.wikipedia.org/wiki/Propri%C3%A9t%C3%A9s_ACID

4. Une preuve de concept (de l'anglais « *proof of concept* », POC), ou démonstration de faisabilité, est une réalisation ayant pour vocation de montrer la faisabilité d'un procédé ou d'une innovation (source : Wikipédia).

Dix ans de web sémantique, et maintenant ?

En 2009, dans deux billets de blog intitulés « Les carcans de la pensée hiérarchique et documentaire »¹, j'invitais les professionnels de l'information à dépasser la forme documentaire et le modèle hiérarchique de la description des collections, en reprenant entre autres l'exemple de l'EAD.

Je présentais le modèle de graphe en général et son implémentation avec les technologies du web sémantique en particulier comme un moyen de dépasser ce paradigme afin d'offrir une plus grande souplesse de description des collections. Dix ans plus tard, alors que les enjeux auxquels doivent faire face les institutions patrimoniales ont évolué, on peut se demander si ces solutions restent à privilégier.

Entre 2008 et 2014, j'ai eu l'occasion de déployer les technologies du web sémantique dans différents contextes². Si, au terme de ces expériences, les promesses de souplesse, d'interopérabilité et de décentralisation des données sont globalement tenues, leurs apports sont contrebalancés par des limites tant au niveau technique, structurel que dans les perspectives d'évolutivité et de maintenabilité des solutions à cause de la faible appropriation de ces technologies par les développeurs. Au-delà de ces expériences, si quelques initiatives ont réussi à émerger, les technologies du web sémantique ne se sont pas imposées alors même que l'ouverture et l'échange des données, cas d'usage pour lesquels elles avaient été mises au point, ont constitué un des enjeux des quinze dernières années.

Malgré tout, la réflexion sur les technologies du web sémantique a indéniablement joué un rôle dans la gestion des données des institutions patrimoniales et a accompagné la réflexion sur la modélisation, en particulier dans l'abandon progressif de la forme documentaire de cette description au profit du modèle de graphe. C'est le cas aussi bien pour les bibliothèques avec le modèle FRBR-LRM, les musées avec le modèle CIDOC-CRM que les archives, plus récemment, avec le modèle RiC. Elles ont permis de placer les référentiels au cœur des systèmes de gestion des collections, afin de décroquer et

mettre en relation des silos de données. Enfin, alors que la question de la donnée devenait centrale au sein des organisations, elles ont démontré qu'il était possible de penser la structuration et le stockage des données en dehors de leurs usages, ouvrant ainsi la voie à une véritable gouvernance de la donnée.

Mais face à l'augmentation, sans précédent, des collections par l'irruption du numérique, à la démultiplication des applications pour répondre aux nouveaux besoins, à la montée en puissance de l'intelligence artificielle et à l'évolution des usages des collections, l'enjeu est plus que jamais de maîtriser le système d'information en général et les données en particulier. Dans ce cadre, si le modèle de graphe reste la voie à privilégier pour envisager la structuration de la description des collections, les technologies du web sémantique ne peuvent constituer le cœur de la solution technique à déployer en particulier de par les doutes qu'elles laissent planer sur leur évolutivité, leur robustesse et leur maintenabilité. Elles peuvent être envisagées pour l'exposition des données si tant est qu'elles trouvent un large usage, ce qui n'est pas le cas actuellement. L'enjeu est de dépasser la problématique technique de la description des collections pour aborder celle de la gouvernance de toutes les données des institutions.

«L'enjeu est plus que jamais de maîtriser le système d'information en général et les données en particulier.»



Gautier Poupeau
Data architect
Institut national de l'audiovisuel

1. Gautier Poupeau, « Les carcans de la pensée hiérarchique et documentaire », *Les Petites Cases*, 3 mai 2009 [www.lespetitescases.net/carcans-de-la-pensee-hierarchique-et-documentaire-1] et [www.lespetitescases.net/carcans-de-la-pensee-hierarchique-et-documentaire-2].

2. Pour une analyse détaillée de ces différentes expériences, vous pouvez consulter la série de quatre billets visant à faire un bilan de mon utilisation des technologies du web sémantique [www.lespetitescases.net/les-technos-du-web-sémantique-ont-elles-tenu-leurs-promesses] (consulté le 6 novembre 2020).

Archives en RDF, thésaurus et versionnement à SemWeb.Pro 2020

SemWeb. Pro 2020, la 9^e édition de la conférence annuelle des professionnels du web sémantique, a eu lieu au mois de novembre dernier dans un espace virtuel au décor champêtre, pour tenter d'oublier le confinement généralisé.

Comme tous les ans, de nombreuses présentations relevaient du domaine des archives, musées et bibliothèques et offraient un intérêt pour les archivistes. *Records in Contexts* (RiC) est un nouveau standard de description des archives adapté au web sémantique. Les Archives nationales sont très impliquées dans ce projet et ont dévoilé le

convertisseur EAD et EAC-CPF vers RDF/RiC-O qui a été mis en œuvre ces derniers mois. ScoLOMFR est un thésaurus exprimé en RDF/SKOS, pour décrire des ressources pédagogiques. Le réseau Canopé, l'opérateur public au service de la communauté éducative, a présenté l'effort particulier déployé pour faciliter aux uti-

lisateurs la mise à jour de leurs métadonnées lors du passage à cette version 6.

Le standard ELI est aujourd'hui largement diffusé au sein de l'Union européenne. Il permet de décrire en RDF les législations et leurs évolutions. Un site comme Légifrance l'utilise pour exposer les articles de loi, mais aussi toutes les versions antérieures de ces articles et les liens vers les décrets et les amendements qui les ont modifiés.

Du côté des sujets plus techniques, on notera un outil pour naviguer dans les entrepôts SPARQL sans écrire de requête,

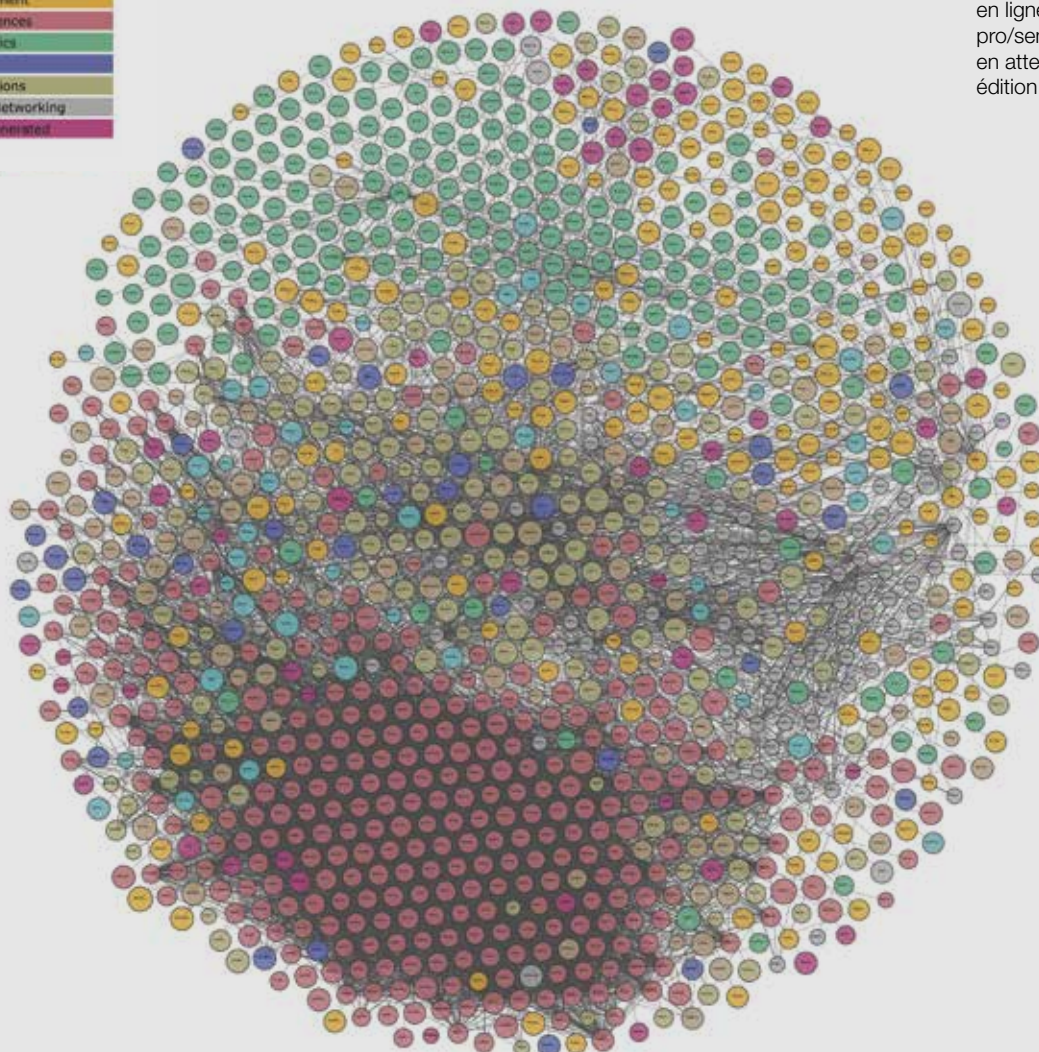
un type de données RDF pour exprimer les unités de mesure, le nouveau standard de décentralisation nommé Solid qui protège les données individuelles et plusieurs exemples d'utilisation du web sémantique pour décrire des installations industrielles.



Nicolas Chauvat
Organisateur de la conférence SemWeb. Pro
Fondateur de Logilab



Pour les personnes intéressées, les vidéos sont toujours consultables en ligne [<https://semweb.pro/semwebpro-2020>] en attendant la prochaine édition fin 2021.



Presque dix ans pour data.bnf.fr : quel regard sur le web sémantique ?

En 2011, la BnF lançait une nouvelle application nommée data.bnf.fr. Parmi les présentations du site, on pouvait lire : « Depuis 2008, la BnF expérimente les standards du web sémantique permettant l'utilisation et l'interconnexion des données structurées de ses différents catalogues en ligne et participe aujourd'hui au web des données en rendant accessibles à tous de manière libre, des données publiques tout en les reliant avec d'autres ressources disponibles elles aussi librement, telles que les contenus de Wikipédia. »

Pourquoi la BnF s'est-elle lancée dans un tel projet ?

À l'époque, le constat est sévère : les données de la BnF sont peu accessibles sur le web, elles sont peu liées avec des données extérieures à l'établissement, ce qui amplifie leur manque de visibilité, et elles sont difficiles à réutiliser, entre autres, parce que les formats de diffusion de données sont des formats purement bibliothéconomiques.

Le projet data.bnf.fr est aussi largement encouragé par deux facteurs externes. En 1997, le monde des bibliothèques voit la publication du modèle conceptuel FRBR (puis IFLA-LRM en 2017) modifiant la modélisation des données bibliographiques et ayant vocation à encourager et à faciliter l'intégration des données des bibliothèques au web sémantique. En 2011, en France, la licence Etalab voit le jour.

De quoi s'agit-il concrètement ?

Data.bnf.fr permet d'agréger sur une même page des données provenant de différents réservoirs de la BnF en indiquant toutes les informations nécessaires à leur description et à la construction de liens vers d'autres entités et vers des données extérieures à l'établissement. Les données sont diffusées en RDF en utilisant des ontologies qui font universellement référence.

En outre, les données sont structurées dans des pages HTML renvoyant à des entités non ambiguës (l'utilisateur cherchera intuitivement une œuvre et non une notice bibliographique), ce qui leur permet d'être identifiées et cherchées efficacement sur les différents moteurs de recherche.

Entre 2011 et 2021, qu'avons-nous appris ?

En principe, le web sémantique a d'énormes avantages. On y trouve des données fiables, structurées et décrites selon des règles et des ontologies largement utilisées par les acteurs du web sémantique, ce qui encourage la réutilisation et l'interopérabilité. Ainsi, la BnF expose ses données dans un format appréhendable pour une communauté beaucoup plus large

que celle des bibliothèques et, en parallèle, elle peut aussi profiter de données produites par d'autres établissements.

Mais, en réalité, l'interopérabilité réelle n'est pas complètement au rendez-vous, puisque la compréhension des ontologies peut être différente selon les établissements, le coût de maintien des alignements entre les données peut être important puisqu'elles évoluent. La notion de fiabilité des données est contextuelle et dépend des politiques de gestion des données de chaque établissement, rendant parfois difficiles les raisonnements déductifs et les inférences. Ces technologies nécessitent une médiation très importante pour être réellement opérantes. De plus, le monde ne peut pas être réduit à des triplets RDF ; les nuances et la rigueur de la description s'en trouvent parfois amoindries.

Plus qu'une solution miracle à un problème de visibilité, le web sémantique apparaît comme un moyen de diffuser les données à conjuguer avec d'autres. Les établissements peuvent évaluer la pertinence

de propulser leurs données dans le web sémantique au regard de leurs objectifs, de leurs moyens et de leurs données. La réalité est donc nuancée ; néanmoins, le web sémantique promeut des principes fondamentaux de gestion des données : identifiants pérennes, évaluation de la qualité des données, descriptions via des standards largement diffusés, etc. Plus fondamentalement, le web sémantique est une part d'un

écosystème complexe visant à contenter les usages multiples des utilisateurs. Récemment, la BnF s'est lancée, seule et avec d'autres institutions, dans des projets touchant à la production ou coproduction de données – NOEMI¹ et le FNE² –, et l'établissement maintient toujours au sein des multiples objectifs de ces nouveaux projets de faciliter l'intégration de ses données au web sémantique.

« Les établissements peuvent évaluer la pertinence de propulser leurs données dans le web sémantique au regard de leurs objectifs, de leurs moyens et de leurs données. »



Tiphaine-Cécile Foucher
Cheffe de produit data.bnf.fr

1. Pour « Nouer les œuvres, les expressions, les manifestations et les items ».

2. Fichier national d'entités.

Conclusion

La stratégie ministérielle pour les archives est clairement orientée vers le web sémantique et déclinée par le Service interministériel des Archives de France à travers les évolutions à venir du portail FranceArchives. Les Archives nationales apparaissent comme pilotes en la matière, car elles ont fait le choix du modèle conceptuel et de l'ontologie *Records in Contexts* pour faire évoluer leurs métadonnées descriptives.

La maturité sur le sujet est néanmoins variable selon les services d'archives, qui devront se former pour prendre le cap du web sémantique. Les retours d'expérience d'experts ou d'informaticiens mettent aussi en lumière les limites de ces technologies, en matière d'évolutivité, de robustesse et de maîtrise. Leur principal intérêt semble résider dans l'interopérabilité des données qu'elles permettent.

Le web sémantique nous oblige surtout à réfléchir à la structuration de nos métadonnées. Avec la prise en compte de ces problématiques, les archivistes sont sans doute à la veille de réaliser leur « transition archivistique », pendant de celle des bibliothèques. Une affaire à suivre...

Coordination : Violette Lévy

