https://www.transition-bibliographique.fr/2018-02-14-serendipite-usages-interfaces-outils-recherche-frbrisation/

Sérendipité et usages : interfaces et outils de recherche dans le cadre de la FRBRisation

Cet article est le compte-rendu de la session parallèle n°2 « Sérendipité et usages : interfaces et outils de recherche dans le cadre de la FRBRisation » de la 2e journée professionnelle du groupe Systèmes & Données, Métadonnées en bibliothèques : attention, travaux !, qui s'est tenue le mardi 14 novembre 2017 à la BnF.

Le support, les notes explicatives et les réponses aux questions posées lors de la session, ont été réalisés par les membres du groupe Systèmes & Données : Claire BELHADI-CHAVANNE, Pierre BOURNERIE (co-coordinateur du groupe), Xavier GUILLOT, Alix LIGOZAT, Didier THEBAULT, Anna SVENBRO (co-coordinatrice du groupe).

Si malgré toute l'attention apportée à l'écriture de ce billet, certaines informations étaient erronées, vous pouvez nous le signaler en nous contactant via ce formulaire.

Dans quelle mesure les interfaces de recherche, jusqu'alors dévolues essentiellement à un simple catalogue de ressources, sont-elles amenées à évoluer ? L'émergence du web sémantique dans le cadre des modèles FRBR et IFLA LRM apporte avec lui des possibilités étendues et rationnelles en matière d'interface et d'enrichissement des données. Le catalogue peut alors devenir un véritable outil de découverte.

En complément de cette présentation et lorsque cela s'avérait nécessaire pour la compréhension, les membres du groupe Systèmes & Données ont tenté de synthétiser les compléments d'information à certaines diapositives ou les éléments de réponse aux questions posées, dans les notes ci-dessous :

Diapo 2 :

Parce que le modèle IFLA LRM unifie les relations qui régissent les entités (auteurs, sujets, œuvres…), les manipuler pour les exposer dans une interface de recherche est rendu moins ardu. L'unité conceptuelle des relations qui en découle permet d'envisager l'interface de recherche de manière unifiée, sans qu'une hiérarchisation des données exposées ne s'impose a priori.

Diapo 3 :

Mais ces classes de « choses » (une statue, un lieu géographique…), ces propriétés et ces relations (a créé, est contemporain de, est né en, est situé à…) entre les « choses », comment les exprimer de manière non ambiguë pour des machines ?

Diapo 4 :

RDF (Resource Description Framework) est un modèle destiné à décrire les ressources issues du web de données. C'est le langage de base du web sémantique.
C'est un modèle qui peut s'exprimer dans plusieurs formats (n3, nt, xml, json-ld…)
Sa structure est constituée de triplets « sujet » « prédicat » « objet ».
Le prédicat correspond à l'étiquette d'un champ dans une base de données.

Diapo 5 :

Dans un graphe, chaque entité est appelée un nœud (node en anglais).
Un nœud peut être objet et/ou sujet selon le contexte établit par le prédicat (Les Rougon-Macquart dans l'exemple).
Un graphe est le résultat d'une requête particulière. Par exemple ici : Quelle entité personne a écrit Les Rougon-Macquart / Quelle entité oeuvre contient Les Rougon-Macquart / Quand est né l'entité personne.

Diapo 6 :

Le graphe est la manifestation des entités exprimées dans leurs interrelations dans le contexte du web sémantique. Il est comme l'écosystème qui sous-tend le concept.
Toute interface qui rend visibles ces entités et leurs relations est par conséquent un graphe.
Comment exprimer ce graphe dans une interface de recherche, et l'épurer suffisamment pour en rendre la substance ?
Dans le cadre d'une recherche documentaire, l'œuvre apparait comme étant un point d'entrée privilégié, car suffisamment large et pouvant encapsuler un maximum d'information (par rapport à l'expression/manifestation).
Comme nous le verrons, dans une recherche orientée, des entités auteur ou sujet peuvent également présenter le cadre de l'encapsulation des données.
Définition de l'ontologie : modèle de données représentant un ensemble de concepts et les relations entre eux.

Diapo 7 :

Une simple image pour illustrer la complexité potentielle d'un graphe.
Il s'agit ici de Balzac dans data.bnf.fr, et c'est un graphe partiel.
Les entrées de date par exemple (1834 : dateOfWork) ou de langue (fre : languageOfThePerson, language) peuvent potentiellement s'étendre vers un nombre de relations extrêmement important.

Diapo 8 :

Synthèse partielle des relations en jeu dans un graphe, il s'agit d'un graphe décrivant une situation « fictionnelle ».
Un auteur contributeur d'une manifestation d'une œuvre d'un autre auteur, dont il serait sujet par ailleurs…
Notion d'ontologie (dc, foaf, frbr-rda, skos, owl… => frbr-rda dans webvowl : http://visualdataweb.de/webvowl/#iri=http://rdvocab.info/uri/schema/FRBRentitiesRDA/).

Diapo 9 :

…Mais encore pourquoi pas : Sujet qui encapsule auteur(s), œuvre(s), manifestation(s)…. C'est tout l'intérêt du modèle, il n'y a pas de hiérarchisation –a priori-

Diapo 10 :

Adaptation au parcours de recherche des tâches FRBR « utilisateur », avec la 5ème tâche (explorer) apportant la possible contextualisation du « discours ».
Source http://library.ifla.org/1084/7/207-riva-fr.pdf page 4 (traduction Mélanie Roche):
Trouver : Rechercher tout critère pertinent afin de rassembler des informations sur une ou plusieurs ressources présentant un intérêt.
Identifier : Comprendre clairement la nature des ressources trouvées et faire la distinction entre des ressources similaires.
Sélectionner : Déterminer l'adaptation de la ressource trouvée et choisir (en acceptant ou rejetant) des ressources spécifiques.
Obtenir : Accéder au contenu de la ressource.
Naviguer (explorer*) : Utiliser les relations qui existent entre une ressource et une autre pour les situer dans un contexte (alignements*).
Les quatre premières tâches (trouver, identifier, sélectionner, obtenir) se conçoivent aisément comme des généralisations des quatre tâches FRBR portant les mêmes noms. Les tâches trouver et identifier apparaissent de même aussi bien dans FRAD que dans FRSAD ; FRSAD inclut aussi sélectionner. La tâche naviguer provient de FRSAD, mais est définie dans le modèle intégré de manière à pouvoir inclure la tâche FRAD contextualiser. Dans la mesure où la dernière tâche de FRAD (justifier) relève du travail du personnel de bibliothèque, elle n'entre pas dans le domaine d'application de FRBR ou d'IFLA LRM.
* Notes groupe 2

Diapo 11 :

[TROUVER / IDENTIFIER] Ici, sur l'OPAC de la bibliothèque de Vaulx-en-Velin (solution Progilone / Syrtis, actuellement en production), la non-congruence entre des données qui proviennent de silos différents* a obligé de séparer d'emblée les ressources numériques des documents physiques, ce qui permet indirectement de réduire le bruit dans le contexte d'une recherche documentaire. Un tri par « pertinence » est appliqué par défaut.**

* Impossibilité d'obtenir les métadonnées des ressources numériques proposées en abonnement par le prestataire.
Sans ces métadonnées, il est impossible d'appliquer les algorithmes de FRBRisation pour l'ensemble des ressources (données du SIGB et données du fournisseur de ressources électroniques).
** Note : il pourrait être intéressant de documenter le tri par pertinence (qu'entend-t-on par pertinence, ou par tri par défaut ? A chaque prestataire de le dire et de le documenter).
Dans ce cas, le tri par pertinence est ajustable dans les paramètres du SIGB, mais l'usager n'en a pas connaissance.

Diapo 12 :

[SELECTIONNER] L'auteur « encapsule » les œuvres… A droite, les formes rejetées apparaissent comme des entrées de filtrage (est-ce une bonne idée ?) => il faudrait peut-être plutôt agréger les résultats par autorité, les formes rejetées étant plutôt des données informationnelles. (skos:prefLabel vs skos:altLabel)

Diapo 13 :

[OBTENIR] Ici l'œuvre est exposée et encapsule les notes, les manifestations (éditions) qui encapsulent les items (exemplaires).
Suite à un choix de la bibliothèque de Vaulx-en-Velin, l'œuvre n'est considérée comme une autorité qu'en partie…

Diapo 14 :

Une tendance dans les SIGB aujourd'hui : les interfaces de recherche professionnelles et destinées au public ont des ergonomies similaires.
Celle destinée au public est en « lecture seule », et c'est là ce qui les différencie.

Diapo 15 :

[EXPLORER] Notion d'alignement (exploration / navigation FRBR) : une source institutionnelle… Les « URIs » : une garantie (à comparer à l'alignement par chaîne de caractères)
Apport : L'idée d'aller plus loin et d'avoir la possibilité de contextualiser sa recherche, que ce soit par un apport purement esthétique (histoire des arts) ou informationnel est alléchant et va dans le sens de ce qu'est le web sémantique dans son essence.

Diapo 16 :

La « balade intuitive » proposée par Vaulx-en-Velin. Une interface sous forme de graphe, permettant une exploration des données.

Diapo 17 :

Il s'agit d'un prototype développé en interne destiné à explorer les possibles en matière d'exposition de données sémantisées (data.bnf.fr) par l'intermédiaire d'un auteur et de ses œuvres.
C'est une interface de découverte et d'exploration à partir d'un auteur.
Il ne s'agit pas de proposer une interface de recherche « plein texte », mais d'exposer des données à partir d'un URI. Le champ de saisie va dans ce sens.

Les « dumps » de data.bnf.fr sont intégrés dans un triple store sur un serveur à la bibliothèque.
A cause de la configuration matérielle réduite du serveur utilisé, seuls les dumps « auteurs » et « œuvres » ont été intégrés.
Il s'agit d'une preuve de concept, aucune donnée locale n'est alignée.
Enrichir et contextualiser l'ensemble avec des données issues de Wikipédia et d'Europeana.
Toutes les requêtes sont faites en « SPARQL » (langage de requête dans le contexte du web sémantique et de RDF) sur le triple store local et sur les points de terminaison SPARQL exposés par Wikipedia et Europeana (dans un premier temps).
Ne sont exposées qu'un échantillon aléatoire de 100 œuvres dont un auteur donné est le créateur ou le contributeur, afin de ne pas surcharger la représentation. (J.S. Bach a par exemple plus de 1000 œuvres référencées sur data.bnf.fr),
L'intégration du jeu de données complet de data.bnf.fr dans un triple store est très gourmand en mémoire à cause de la (pré)indexation des triplets => Le serveur utilisé est très limité et le nombre de requête « à concurrence » s'en trouve fortement diminué. Le strict minimum a été intégré pour le prototype (œuvres et auteurs) les sujets, expressions et manifestations n'ont pour le moment pas été intégrées.

Diapo 18 :

Les œuvres d'un auteur sont représentées sous forme de graphe, avec la notion d'agrégation (<http://www.openarchives.org/ore/terms/isAggregatedBy> ex. : « La comédie humaine » agrège « Études de mœurs », qui agrège « Scènes de la vie parisienne » qui agrège « Histoire des treize », etc…).
Les oeuvres dont l'auteur est considéré comme étant contributeur (<http://purl.org/dc/terms/contributor>) sont également représentées, mais les « codes fonction » (rôles => <http://data.bnf.fr/vocabulary/roles/>) n'ont pas encore été intégrés.
Au dessus du graphe, les dates d'édition sont exposées sous forme d'histogramme, le survol d'une date mettant en évidence les œuvres concernées.
Un clic sur une date filtre les œuvres et met à jour l'ensemble des ressources numérisées issues d'Europeana.
Un apport d'information par « pop-up » lors d'un survol de la souris sur une entité (un nœud dans le vocabulaire du graphe).
Un clic sur un nœud renvoie vers la source de l'entité (Gallica, data.bnf.fr…).

Diapo 19 :

A gauche du graphe, quelques auteurs nés la même année sont proposés.
S'il n'y en a pas, des suggestions sont faites de manière aléatoire.
Dans le cas de certains auteurs, et si il y en a, les influences deviennent un point d'entrée (source Wikipédia).
Les notes biographiques sont récupérées depuis Wikipédia s'il y en a (notes issues de data.bnf.fr à défaut).

Diapo 20 :

Plus bas sur la page, un ensemble de ressources numérisées issues d'œuvres publiées ou créées une année donnée, indexées par Europeana.
Ces œuvres proviennent de bibliothèques, archives et musées européens, dont Gallica.
=> Cela permet de contextualiser l'auteur dans son époque.
Il n'y a pas d'alignement à proprement parler de data.bnf.fr avec Europeana, il s'agit d'une requête à part, directement sur un point de terminaison SPARQL Europeana avec la date comme variable.

L’alignement a posteriori avec les « auteurs liés » se fait avec une requête croisée sur data.bnf.fr à partir des URIs des imagettes disponibles (foaf:depiction <ark://>).
Cela devient une sorte d’inférence (un peu forcée) = alignement déduit.

Diapo 21 :

Dans le cas d’un musicien, les genres musicaux sont également exposés (<http://musicontology.com/genre>), et permettent, à l’instar des dates d’édition, de filtrer l’exploration.
De plus, un survol à la souris de certaines œuvres musicales, qui possèdent un extrait disponible, offre une écoute.
Problème : L’alignement vers la ressource étant une manifestation, c’est la première plage de la manifestation concernée qui est proposée à l’écoute par défaut (plage = interprétation = expression). Ce qui peut conduire à une écoute qui ne correspond pas à l’œuvre décrite.
Et peut même dans certains cas correspondre à une œuvre qui n’est pas de l’auteur (cas des recueils).

Diapo 22 :

Ouverture vers la transition bibliographique
Dans nos catalogues traditionnels, les valeurs des champs sont statiques et descriptives.
Avec le web sémantique, ces mêmes valeurs sont décrites indirectement sur le web de données, sur des ressources pointées par des URIs.
Au-delà, toujours dans le cadre du « linked data », seuls les URIs demeurent, les propriétés des champs elles-mêmes devenant des URIs (prédicat).
Cette nouvelle modélisation et l’ouverture possible vers le web de données engendre une exploration accrue et une contextualisation des résultats de recherche.
S’il est trop tôt pour avoir le recul nécessaire en matière d’usage, il est dores et déjà possible de prédire que cela favorisera la découverte et la sérendipité.

Question : Comment la notion de série est-elle intégrée dans le modèle ?

Réponse :

Les dernières évolutions de RDA, en cours de publication, intègrent la notion de « ressources continues ». Dans ce contexte l’oeuvre est vue comme une entité diachronique, dont la publication se poursuit dans le temps, et est envisagée comme étant le plan qui expose la manière dont le contenu évolue.

En effet, les expressions et manifestations d’une telle oeuvre ne peuvent pas être décrites tant que l’oeuvre dans son ensemble n’est pas complète.
3 relations principales sont définies dans IFLA LRM :

LRM-R19 : une oeuvre précède / succède à une autre oeuvre
Définition : Ceci est la relation de deux oeuvres où le contenu de la seconde est une suite logique de la première.
Exemples :
– Le film « Autant en emporte le vent » de Margaret Mitchell précède « Scarlett » d’Alexandra Ripley et « Le clan de Rhett Butler » de Donald McCaig.

– « Autant en emporte le vent » de Margaret Mitchell succède au « Voyage de Ruth » de Donald McCaig.
– La série télévisée « Better Call Saul! » précède la série télévisée « Breaking Bad ».
– « Le sorcier de Terremer » précède « Les tombeaux d'Atuan », lui-même précède « L'ultime rivage », l'ensemble est inclus dans le « Cycle de Terremer » par Ursula K. Le Guin.

LRM-R22 : est une transformation de / transformé en
Définition : Cette relation indique qu'une nouvelle oeuvre a été créée en changeant la portée ou la politique éditoriale (comme dans une oeuvre sérielle ou agrégative), le genre ou la forme littéraire (dramatisation, novalisation), le public cible (adaptation pour les enfants) ou le style (paraphrase, imitation, parodie) d'un travail antérieur.
Exemples :
– « Cymbeline » de Mary Lamb, dans « Contes de Shakespeare » de Charles et Mary Lamb, est une transformation de « Cymbeline » de William Shakespeare.
– « Orgueil et préjugés et zombies » de Seth Grahame-Smith est une transformation de « Orgueil et préjugés » de Jane Austen.
– Le périodique intitulé « Le Patriote de Saône-et-Loire » (ISSN 1959-9935) a été transformé en la nouvelle revue intitulée « Le Démocrate de Saône-et-Loire » (ISSN 1959-9943) après que le premier ait été supprimé par la censure en 1850 [remplacement définitif].
– Les périodiques distincts « Animal research » (ISSN 1627-3583), « Animal science » (ISSN 1357-7298) et « Reproduction nutrition development » (ISSN 0926-5287) ont été transformés en un périodique intitulé « Animal » (ISSN 1751-7311) [une fusion].

LRM-R25 : a été agrégé par / a agrégé
Définition : Cette relation indique qu'une expression spécifique d'une oeuvre a été choisie dans le cadre d'une expression agrégative.
Exemples :
– Le texte anglais de « The fall of the House of Usher » d'Edgar Allan Poe a été agrégé dans une expression agrégative, produisant la manifestation agrégative « The Oxford book of short stories » ; sélectionné par V.S. Pritchett.
– L'expression agrégée qui produit la série monographique « IFLA series on bibliographic control » a agrégé le texte anglais de « ISBD: International bibliographic description standard », édition consolidée 2011.
– L'expression agrégée qui produit la série monographique « Povremena izdanja Hrvatskoga knjižnicarskog društva. Novi niz » a agrégé le texte croate 2014 de « ISBD: International bibliographic description standard », édition consolidée 2011.

Plus d'information (en anglais) :
https://repository.ifla.org/bitstream/123456789/40/1/ifla-lrm-august-2017_rev201712.pdf