

Le démonstrateur des Archives nationales : sémantiser et explorer les métadonnées des archives notariales

Florence Clavaud, Pauline Charbonnier
Lab des Archives nationales
17 juin 2022

Restitution du projet Sparnatural : un outil pour explorer les graphes de métadonnées culturelles



MINISTÈRE
DE LA CULTURE

Liberté
Égalité
Fraternité

ARCHIVES
NATIONALES

SPARNA

Plan

- Objectifs du démonstrateur des Archives nationales
- Les contenus actuels du démonstrateur
 - Les métadonnées retenues
 - Historique rapide
 - Le processus de production du graphe de données
 - Éléments statistiques concernant le graphe RDF obtenu
- Conception d'une interface de recherche basée sur Sparnatural
 - Démonstrateur A
 - Démonstrateur B
- Bilan et perspectives

Objectifs du démonstrateur des Archives nationales

Le démonstrateur des Archives nationales (AN) devait servir :

- à **tester les évolutions de Sparnatural**, et par là-même contribuer au travail de développement ;
- à **mettre en place une toute première application web permettant d’explorer et d’interroger de façon intuitive une partie significative des métadonnées archivistiques de l’institution, converties en RDF/RiC-O.**

Ce faisant il s’agissait aussi :

- d’apporter la **preuve que la sémantisation à grande échelle de métadonnées archivistiques “classiques” est possible**, et **que cela ouvre de nouvelles possibilités de recherche aux utilisateurs** (ce après le prototype qualitatif [PIAAF](#) et le développement du logiciel [RiC-O Converter](#)) ;
- **de voir ce qu’une telle opération nous apprend sur nos métadonnées ;**
- **de construire une méthodologie et d’acquérir des compétences** sur la mise en œuvre de Sparnatural.

Tout cela dans le cadre d’une **stratégie globale d’évolution des métadonnées et du système d’information de l’institution vers des données de meilleure qualité, plus accessibles, et des graphes d’entités liées.**

Contenus actuels du démonstrateur

Les métadonnées des Archives nationales, ce sont (entre autres) :

- plus de 30 000 instruments de recherche ou inventaires (en XML/EAD 2002) ;
- plus de 15 400 notices d'autorité (en EAC-CPF) sur des producteurs ;
- des vocabulaires servant à l'indexation de ces inventaires et notices.

La [salle des inventaires virtuelle](#) permet aux utilisateurs de faire des recherches dans l'ensemble de ces métadonnées.

Pour alimenter le démonstrateur Sparnatural, dans le cadre de ce projet, il a fallu **sélectionner un sous-ensemble de ces métadonnées.**

→ Nous avons retenu **une partie des métadonnées décrivant les archives des notaires de Paris**, pour plusieurs raisons :

- elles forment un tout très cohérent ;
- leur description est riche, tout en présentant les caractéristiques classiques des répertoires numériques produits par les services publics d'archives français ;
- elles décrivent des documents très consultés, par divers types de publics ;
- des fonds de même nature sont également conservés en archives départementales.

Les métadonnées retenues

Les notices d'autorité sur les producteurs et les inventaires relatifs aux archives des 40 premières études sur les 122 études parisiennes, soit :

- 40 notices de collectivité décrivant les études (et une autre, celle de l'étude LIII) ;
- 1079 notices décrivant les notaires qui ont exercé au sein de ces études ;
- 1577 instruments de recherche archivistiques (32 % des 4902 instruments de recherche existants au Département du Minutier central des notaires de Paris - DMC - et validés en mars 2022) ;
- les notices (sur des personnes et organismes sujets des documents) et vocabulaires (types de documents, mots-matières, activités, lieux de Paris) utilisés dans ces notices et inventaires pour les indexer.

Les notices et vocabulaires sont publiés et disponibles dans un dépôt public des Archives nationales sur GitHub (<https://github.com/ArchivesNationalesFR/Referentiels>).

Quelques notices de producteurs et le vocabulaire des types de documents ont été enrichis dans le cadre du projet.

Ce corpus de métadonnées correspond au **tiers des fichiers de métadonnées décrivant les archives notariales aux Archives nationales** (qui forment un ensemble de plus de **26 km linéaires, 194 500 cartons**, dans lesquels on estime que sont conservés **20 millions d'actes** ; voir la [fiche d'aide à la recherche sur les chantiers d'inventaires au DMC](#)).

Historique rapide

- Mise en place par Sparna d'une **base RDF (instance du logiciel [GraphDB Free](#)) sur la grille de services de la TGIR Huma-Num du CNRS**, et d'un dépôt sur GitHub pour stocker le code du démonstrateur (été 2021) ; base remplie provisoirement avec des données RDF/RiC-O préexistantes.
- **Début des développements de la version 2 de Sparnatural (août 2021), en mode agile.**
- Ateliers Sparna-Lab de prise en main de la méthode de configuration de Sparnatural.
- Sélection du jeu de métadonnées, conversion en données RDF/RiC-O et import du jeu de métadonnées dans la base RDF ; premiers traitements complémentaires sur les données (Lab, automne 2021).
- **Premières configurations AN à l'automne 2021** (Lab avec l'aide de Sparna).
- **Ateliers utilisateurs** (équipe projet au complet, novembre 2021).
- **Intégration progressive par Sparna des développements Sparnatural dans le dépôt GitHub du démonstrateur ; tests et mise en œuvre** de ces nouveautés par le Lab dans l'interface de recherche du démonstrateur.
- **Deuxième conversion des métadonnées en RDF/RiC-O, et consolidation des configurations des interfaces de recherche** (mai-juin 2022).

Processus de production du graphe de données

Les métadonnées ont été converties en graphes de données liées, conformément au nouveau modèle conceptuel [Records in Contexts](#) et à sa transposition technique, l'ontologie [RiC-O](#) (version 0.2, publiée en février 2021).

Deux itérations ont été réalisées : une première en novembre 2021 et une autre en mai 2022, tirant les leçons des résultats de la première et des premiers travaux de configuration du démonstrateur.

Pour les fichiers EAD et EAC-CPF :

- utilisation du logiciel open source [RiC-O Converter](#), développé pour les Archives nationales en 2019, disponible sur GitHub depuis avril 2020 : <https://github.com/ArchivesNationalesFR/rico-converter> ;
- RiC-O Converter produit des fichiers RDF conformes à RiC-O 0.1. Les fichiers obtenus ont donc ensuite été rendus conformes à RiC-O 0.2 et un peu enrichis ou modifiés par le Lab, par des scripts spécifiques
- Parmi les enrichissements réalisés : propagation de la relation de provenance sur toutes les ressources archivistiques ; catégorisation (avec [rico:hasDocumentaryFormType](#) et le référentiel des types de documents) des instances identifiables comme étant des répertoires, listes ou mentions d'actes.

Pour les vocabulaires et notices utilisées pour indexer ces fichiers EAD et EAC-CPF : conversion grâce des scripts dédiés, produits et utilisés par le Lab.

Processus de production du graphe de données

Après import dans la base RDF, **le graphe de données a été à nouveau enrichi par inférence** : en particulier, **création et peuplement de classes spécifiques aux archives notariales** (exemple ci-contre pour la classe Répertoire)

Ce travail, quoique non indispensable, a simplifié les travaux de configuration du démonstrateur et permis d'optimiser le fonctionnement du démonstrateur en réduisant le temps d'exécution de certaines requêtes.

```
PREFIX demoanonto: <https://sparnatural-demo-anf.huma-num.fr/ontology#>␣
PREFIX sparnademoan: <https://sparnatural-demo-anf.huma-num.fr/addedData/>␣
PREFIX owl: <http://www.w3.org/2002/07/owl#>␣
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>␣
PREFIX rico: <https://www.ica.org/standards/RiC/ontology#>␣
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>␣
PREFIX dc: <http://purl.org/dc/elements/1.1/>␣
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>␣
␣
# Déclaration de nouvelles classes␣
INSERT DATA { GRAPH sparnademoan: {␣
  demoanonto:Repertoire a owl:Class ;␣
  rdfs:subClassOf rico:Record ;␣
  rdfs:label "Répertoire"@fr ;␣
  rdfs:comment "Liste, en général chronologique,␣
  se présentant sous forme de registre ou d'ensemble de feuillets,␣
  que le notaire a obligation de tenir, aujourd'hui jour par jour,␣
  en y faisant figurer date, nature de l'acte [...].␣
  Une fois versé avec les minutes dans un service public d'archives, le répertoire sert␣
  principalement à repérer un acte donné [...], mais [...], comme la minute,␣
  il comporte intrinsèquement une valeur probatoire."@fr ; [...]"@fr .␣
} } ;␣
␣
# calcul des instances des nouvelles classes␣
␣
INSERT {␣
GRAPH sparnademoan: {␣
?rep a demoanonto:Repertoire }␣
}␣
WHERE {␣
?rep a rico:Record ;␣
  rico:hasDocumentaryFormType <http://data.archives-nationales.culture.gouv.fr/documentaryFormType/FRAN_RI_001-d-a04blhkn0-92bbw9jou9ed> .␣
}␣
}
```

Quelques informations sur le graphe obtenu

Environ **50 millions de triplets RDF**.

Donc la sémantisation, selon les mêmes modalités, de l'ensemble des métadonnées des archives notariales conservées aux Archives nationales produirait environ 150 millions de triplets.

Dans ce jeu de données, sont employées, parmi les composants définis dans RiC-O 0.2 :

- 34 classes (catégories d'entités) sur 106 ;
- 20 *datatype properties* (relations dont la cible est une chaîne de caractères) sur 62 ;
- 75 *object properties* (relations dont la cible est une entité) sur 423.

Première opération de sémantisation (conversion en RDF conformément à RiC-O) à grande échelle, par les Archives nationales (et en France sauf erreur), de métadonnées archivistiques « classiques ».

Un autre jeu de données moins massif et plus spécifique (relatif à des séries de photographies aériennes) a par ailleurs été publié en décembre 2021 dans le cadre du projet ALEGORIA (disponible à <https://github.com/ArchivesNationalesFR/ALEGORIA-datasets>, et également accessible via le [SPARQL endpoint](#) du projet).

Éléments statistiques sur le graphe

Quantités obtenues en interrogeant le graphe RDF :

- 72 661 groupes de documents (fonds, séries, liasses, dossiers de clients...)
- **400 570 actes notariés** (décrits à ce jour) au sein des liasses ;
- **104 383 mentions d'actes** relevées au sein des répertoires ;
- *3135 répertoires** ;
- **13 127 listes d'actes au sein des répertoires**, dont 11 670 numérisées ;
- *629 246 personnes** (dont 3 213 personnes dotées d'un IRI) ;
- 530 organismes dont 528 dotés d'un IRI ;
- 15 559 lieux dont 15 463 dotés d'un IRI.

Les quantités obtenues peuvent être considérées comme une bonne approximation pour les actes notariés, les mentions d'actes et les listes d'actes compte tenu des raisonnements assez fiables tenus pour catégoriser ces objets.

* Pour les répertoires et les personnes, les nombres obtenus constituent plutôt un ordre de grandeur.

Concevoir une interface de recherche basée sur Sparnatural

- **Se fait essentiellement en produisant une ontologie OWL**, qui s'appuie sur deux ontologies génériques incluses dans le code source du logiciel.
- On définit donc **un modèle ontologique pour la recherche et ses correspondances, simples ou plus élaborées, avec les classes et propriétés de l'ontologie métier.**
- Voir la documentation à : <https://docs.sparnatural.eu/OWL-based-configuration>

- Si on dispose d'un SPARQL endpoint, que l'on connaît HTML, le modèle métier employé, le ou les jeux de données à explorer et le langage OWL, **on peut donc construire son (ses) interface(s) de façon autonome.**
- L'enjeu est alors d'atteindre par ce biais un **compromis raisonnable entre la complexité du modèle métier et la nécessité de produire une interface compréhensible et performante pour les utilisateurs.**
- Comme il s'agit d'un fichier OWL, **une ontologie Sparnatural de recherche élaborée dans le cadre d'un projet donné peut être partagée pour être réutilisée dans d'autres projets utilisant le même outil.**
- Cela peut être intéressant dans le cas où l'ontologie utilisée pour produire les jeux de données RDF est un modèle générique (comme les modèles IFLA LRM, CIDOC CRM ou ICA RiC-O).

Concevoir une interface de recherche basée sur Sparnatural

Par exemple, définition à l'aide de l'éditeur open source d'ontologies [Protégé](#) de la classe **Archives** dans l'ontologie de la configuration A.

L'annotation *sparqlString* spécifie la correspondance (ici, très simple) entre la classe Sparnatural et la classe utilisée dans les données RDF (ici une classe spécifique au projet, Archive, qui réunit toutes les instances de rico:RecordResource sauf les instruments de recherche et les notices).

The screenshot displays the Protégé ontology editor interface. The top navigation bar shows the active ontology as 'ontology' with a search bar. Below the navigation bar, the 'Class hierarchy' tab is active, showing a tree view of classes. The 'Archives' class is selected and highlighted in blue. The 'Annotations' tab is also active, showing a list of annotations for the 'Archives' class. The annotations include:

- `rdfs:label` [language: fr] Archives
- `rdfs:label` [language: en] Records
- `defaultLabelProperty` `aPourIntitule`
- `fa:icon` `fad fa-archive`
- `order` [type: xsd:integer] 1
- `sparqlString` `<https://sparnatural-demo-anf.huma-num.fr/ontology#Archive>`
- `tooltip` [language: fr] Tout type de ressource archivistique, qu'il s'agisse d'un item (comme une minute ou un répertoire) ou d'un groupe de documents (série des minutes ou des répertoires d'un notaire, liasse de minutes, dossier de client...).
- `tooltip` [language: en] Any type of archival resource, whether it is an item (such as a minute or a registry) or a group of documents (series of minutes or registry of a notary, bundle of minutes, client file...).

Below the annotations, the 'Description: Archives' tab is active, showing the class description. The 'Equivalent To' section is highlighted in blue, and the 'SubClass Of' section shows that 'Archives' is a subclass of 'SparnaturalClass'. The 'General class axioms' section is also visible.

Concevoir une interface de recherche basée sur Sparnatural

Ce travail de configuration des interfaces de recherche a aussi permis de tester les fonctionnalités développées dans le cadre du projet d'évolution de Sparnatural :

- support de la clause OPTIONAL (on peut ainsi mieux prendre en compte les variations dans la complétude et la précision des données du graphe) ;
- support de FILTER NOT EXISTS ;
- sélection de valeurs dans une liste arborescente (de type thésaurus) ;
- ajout d'aides contextuelles (*tooltips*) ;
- choix des colonnes de résultats à afficher ;
- ajout de requêtes pré-enregistrées ;
- déclinaison de l'interface en plusieurs langues.

Démo

[Accueil](#)[Démonstrateur A](#)[Démonstrateur B](#)[Documentation](#)[Mentions légales](#)[Sparnatural](#)**ARCHIVES
NATIONALES**

Démonstrateur Sparnatural des Archives nationales

Français

Ce démonstrateur, construit avec l'outil [Sparnatural](#) vous permet d'interroger une partie des métadonnées décrivant les archives des notaires de Paris conservées aux Archives nationales. Ces métadonnées ont été sémantisées conformément à l'ontologie [Records In Contexts \(RIC-O\)](#). Le démonstrateur est composé de deux interfaces de recherche. Le [démonstrateur A](#) propose une interface simple et généraliste d'exploration dans ces métadonnées, qui pourrait être utilisée pour d'autres métadonnées archivistiques. Le [démonstrateur B](#) propose une interface d'exploration dans ces métadonnées qui s'appuie sur les spécificités des archives notariales. Par ailleurs les IRIs des ressources RDF ne sont actuellement pas déréférencables.

English

Through this prototype, built with [Sparnatural](#), you can query a part of the metadata describing the archives of Paris notaries kept at the French national Archives. These metadata have been semantized according to the ontology [Records In Contexts \(RIC-O\)](#). The prototype is composed of two research interfaces. [Interface A](#) offers a simple and general exploration interface in this metadata, which could be used for other archival metadata. [Interface B](#) offers an exploration interface in these metadata based on the specificities of notarial archives. Furthermore, the IRIs of RDF resources are currently not dereferenceable.



中華人民共和國
國家圖書館

SPARNA

<https://sparna-git.github.io/sparnatural-demonstrateur-an/>

La première configuration, destinée à la découverte

The image shows a query builder interface with a light green background. It features four horizontal query clauses connected by logical operators. The first clause is 'Archives' (with a folder icon) 'ont pour type de document' (with a document icon) 'Type de document' 'inventaire après décès (...)' (with a document icon and a plus sign). The second clause is 'Archives' 'concernent' 'Personne'. The third clause is 'Personne' 'a pour activité' 'Activité' 'orfèvre'. The fourth clause is 'Archives' 'ont pour date de création' 'Date' 'de 1600 à 1700'. Logical operators 'et' (and) connect the first and second clauses, and the second and third clauses. An 'où' (where) operator connects the second and third clauses. Each clause has a refresh icon (circular arrow) and a delete icon (circle with an X). At the bottom, there is a search bar with a play button, a dropdown menu showing 'this', a 'Tri' (Sort) section with icons for ascending, descending, and random, and a toggle for 'Montrer le nom des colonnes' (Show column names).

Simple et généraliste ; elle peut quasiment s'appliquer à tout jeu de données archivistiques conformes à RiC-O, dans la mesure où s'y trouvent indexées des entités de contexte comme les producteurs, des agents sujets des documents, des lieux.

<https://sparna-git.github.io/sparnatural-demonstrateur-an/index-A.html>

Configuration A

Catégories d'objets (classes) utilisables :

- Archives
- Personne, Organisme, Lieu (correspondent aux classes RiC-O [Person](#), [CorporateBody](#) et [Place](#))
- Type de document, Mot-matière, Activité (classes des concepts SKOS dans les 3 vocabulaires de même nom aux Archives nationales)
- Reproduction numérique
- et d'autres 'entités' associées, correspondant à des valeurs littérales de propriétés RiC-O : date, intitulé, nom.

Deux requêtes préenregistrées :

- les [archives relatives à Jean-Baptiste Poquelin](#)
- les [inventaires après décès d'orfèvres au XVIIIe siècle](#)

La deuxième configuration permet d'explorer le graphe plus en profondeur

Le formulaire de recherche disponible sur la présente page propose **une interface d'exploration dans ces métadonnées qui s'appuie sur les spécificités des archives notariales.**

Groupes de documents produits par des notaires ayant exercé dans le quartier S v

Service SPARQL: <https://sparnatural-demo-anf.huma-num.fr/graphdb/repositories/sparnatural-demo-anf>

où

Notaire > a pour provenance

Notaire

et

Notaire > A exercé dans le quartier

Quartier, paroisse ou arrondissement de Paris

Saint-Germain-l'Auxerrois...

et

Notaire > Date de début d'exercice

Date

A partir de 1550

et

Notaire > Date de fin d'exercice

Date

Jusqu'à 1700

Elle exploite les spécificités des archives notariales et de leur description.

Plutôt adaptée aux besoins d'une personne familière de cet univers, archiviste ou usager habitué.

<https://sparna-git.github.io/sparnatural-demonstrateur-an/index-B.html>

Configuration B

De **nouvelles catégories d'objets** sont définies dans le démonstrateur :

- GroupeDeDocuments (qui correspond à la classe [rico:RecordSet](#)), Repertoire, ListeActes, ActeNotarie, MentionActe, Cote
- Etude, NumeroEtude, Notaire
- QuartierOuParoisse, Voie, Edifice, Adresse

La plupart de ces entités sont **utilisables dès le début de l'écriture d'une requête**.

De nouvelles relations sont également présentes, par exemple :

- entre étude et notaires (Notaire *appartient* à Étude notariale).
- entre notaires (Notaire *a pour prédécesseur* Notaire),
- entre ressources archivistiques (comme Répertoire *inclut ou mentionne* Mention d'acte ou Acte notarié).

Requêtes préenregistrées :

- [Groupes de documents produits par des notaires ayant exercé dans le quartier de Saint-Germain l'Auxerrois entre 1550 et 1700](#)
- [Actes notariés relatifs au Louvre, ayant ou non une copie numérique, avec leur cote et le nom du notaire](#)
- [Types de documents des actes notariés datant de l'année 1848, avec l'indication de leurs cotes](#)
- [Voies parisiennes dans lesquelles les notaires ont exercé du XVe au XXe siècle](#)

Bilan

- **Bilan très positif :**

- en ce qui concerne la **preuve de la faisabilité de la conversion de grandes quantités de métadonnées archivistiques de nature “classique” en données RDF/RiC-O** ;
- en ce qui concerne les **possibilités d’exploration de graphe** offertes par l’outil ;
- en ce qui concerne les **nouvelles possibilités de recherche** qui sont offertes (cf. démo).

Bien sûr nous espérons recueillir d’autres réactions, remarques, suggestions et retours d’expérience que ceux déjà reçus lors des ateliers utilisateurs

- **Autres résultats concrets du projet :**

- Travail préalable à la mise à jour de RiC-O Converter réalisé ;
- Travail de traduction en français des libellés anglais des classes et propriétés RiC-O, va être intégré dans les sources de RiC-O.

- **Limites du démonstrateur :**

- Sont essentiellement celles de Sparnatural, qui permet de construire une interface de recherche, pas d’afficher les données listées dans les tableaux de résultats.
- Pour le démonstrateur AN, il fallait pouvoir combler la lacune (faute pour les AN de proposer déjà le déréréférencement des URIS). Cela a été fait en utilisant [LodView](#).

Perspectives

- **Les deux configurations pourraient être fusionnées en une seule dans une possible suite du projet** (en hiérarchisant classes et propriétés, pour laisser le choix à l'utilisateur d'une exploration rapide ou plus précise). Cela impliquera de faire évoluer les possibilités de configuration offertes par Sparnatural.
- **Travaux d'amélioration des métadonnées source, pour en particulier réduire les redondances liées à la longue histoire de la description des archives aux Archives nationales**, qui affectent le graphe et rendent aussi en SIV certains résultats de recherche difficiles à comprendre et à exploiter.

Ainsi, par exemple, actuellement un même répertoire de notaire est le plus souvent décrit au moins deux fois dans deux instruments de recherche différents, sans qu'un lien ait été établi entre les deux unités de description (éléments EAD <c>), ce qui donne pour l'instant dans le graphe 2 instances de la classe Répertoire pour en fait un seul objet culturel conservé.

- **Production et publication d'une ontologie Sparnatural générique** de recherche pour les données RDF conformes à RiC-O en prenant pour point de départ la configuration A.
- Le Lab des Archives nationales est désormais en mesure de concevoir seul une interface de recherche Sparnatural pour d'autres jeux de données. La réutilisation de Sparnatural est donc prévue dans le cadre d'autres projets de recherche, par exemple pour explorer les données du projet [ORESME](#) (Œuvres et Référentiels des Etudiants, Suppôts et Maîtres de l'université de Paris au Moyen Âge). Voir à ce sujet la présentation, faite en novembre 2021, de la [preuve de concept](#) déjà réalisée.

Pour conclure

La [documentation du démonstrateur](#) donne des précisions complémentaires à notre présentation. Elle sera bientôt traduite en anglais.

Nous envisageons un atelier plus détaillé, au moins pour nos collègues des Archives nationales.

Le jeu de données RDF produit est susceptible d'être rendu public ; en attendant, nous contacter si vous êtes intéressé(e) !

Et enfin, n'hésitez pas à contacter le Lab (le-lab.archives-nationales@culture.gouv.fr) pour toute question, appréciation, remarque ou suggestion.

Merci de votre attention !