Sémantisation et visualisation de métadonnées archivistiques : un projet de prototype

Florence Clavaud

Archives nationales

florence.clavaud@culture.gouv.fr

Journée d'études « Le Web sémantique pour les données culturelles : bilan et perspectives »
12 décembre 2016, INHA

Le projet en quelques mots

De quoi s'agit-il ?

- Objectifs : apporter ensemble la preuve qu'il est possible :

  - de **représenter de façon rigoureuse en RDF** des métadonnées archivistiques (notices d'autorité, descriptions de documents…) provenant de plusieurs institutions ;

  - d'**interconnecter** et d'enrichir les jeux de données obtenus, notamment en exprimant les relations qui existent entre les objets décrits ;

  - de **visualiser** utilement ces jeux de données de façon graphique, en permettant à l'utilisateur de parcourir le graphe et d'accéder aux jeux de métadonnées source

  > **une preuve de concept et d'utilité**

Les participants (Qui ?)

- **Trois partenaires institutionnels :**

  - les Archives nationales de France (ANF) (responsable du projet)
  - la Bibliothèque nationale de France (BnF)
  - le service interministériel des Archives de France (SIAF)

- **Un chercheur**, Emmanuel Château-Dutier (Centre de recherche interuniversitaire en humanités numériques – CRIHN, Université de Montréal, Canada)

- Gestion du projet : une petite équipe projet

Comment ?

- La préparation des métadonnées et leur conversion en RDF sont réalisées par l'équipe projet
- Le prototype sera réalisé par un prestataire dans le cadre d'un marché
- Projet co-financé par les trois institutions partenaires :
  - budget global : 45 000 euros (pour le prestataire)
  - temps de travail des membres de l'équipe projet apporté par les institutions
- Les librairies réalisées seront placées sous licence libre

Calendrier sommaire (Quand ?)

- Projet lancé en 2015

- Publication du prototype (sources logicielles dans un entrepôt public de sources, et instance sous la forme d'un démonstrateur en ligne) prévue en octobre 2017

Enjeux et contexte du projet

Des objectifs communs, des enjeux institutionnels spécifiques : pour le Service interministériel des Archives de France (SIAF)

- Tête du réseau des services publics d'archives français

  > le prototype, un outil à vocation pédagogique dans la diffusion des modèles et méthodes utilisés

- Ne produit pas lui-même de métadonnées, sauf dans le cadre du projet de création d'un **référentiel pour les notices d'autorité décrivant les administrations et collectivités territoriales** (en collaboration avec l'AAF) :
  http://www.archivistes.org/Notices-d-autorite-producteurs-1781
  Un « méta-référentiel » qui pourrait servir à organiser le graphe des collectivités territoriales en fédérant ces collectivités par catégorie, et à relier ce graphe avec celui des administrations et opérateurs de l'État, produit par les Archives nationales

  > le prototype, pour représenter en RDF les fichiers EAC-CPF produits, et disposer d'une interface de consultation dynamique et ergonomique pour ce graphe ; pour montrer la faisabilité et l'intérêt de la connexion de ces notices avec celles produites par les Archives nationales ;

- Pilote la conception et la réalisation du projet de portail national Francearchives.fr

  > les résultats du projet de prototype pourraient nourrir la réflexion pour les futures évolutions de ce portail

Des objectifs communs, des enjeux institutionnels spécifiques : pour les Archives nationales de France

- 1998-2013 : un important effort de normalisation et d'informatisation :
  - travaux pilotes ;
  - développement et mise en production en 2009-2013 d'un SI intégrant l'ensemble des tâches de la chaîne archivistique ;
  - mise en ligne de la salle des inventaires virtuelle (SIV) en septembre 2013 (https://www.siv.archives-nationales.culture.gouv.fr/siv/)

- Aujourd'hui :
  - plus de 25000 instruments de recherche en XML/EAD et plus de 14000 notices d'autorité sur les producteurs, en XML/EAC-CPF ;
  - de nouveaux instruments de recherche tous les jours ;
  - description systématique des producteurs

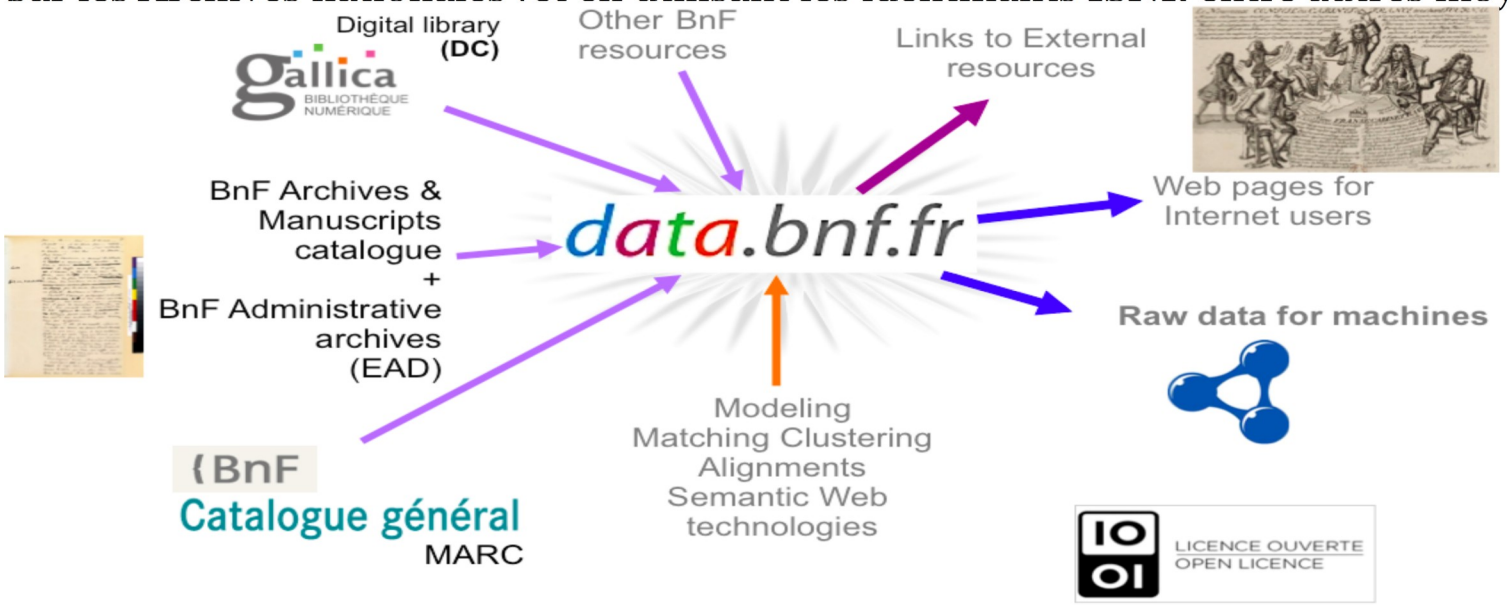
- Les métadonnées produites constituent d'ores et déjà un **réseau très complexe, articulant plusieurs types d'objets, et doté d'une dimension temporelle**

- <u>Problèmes</u> :
  - métadonnées non sémantisées, ce qui limite leur exploitation et leur réutilisation
  - les **moyens de recherche et de consultation de ce… graphe lui sont peu adaptés** ;
  - **les métadonnées** (en particulier les notices d'autorité) ne sont pas **techniquement liées à celles des institutions voisines**, alors que les AN s'inscrivent dans un écosystème administratif, patrimonial et scientifique vaste.

  De premiers jalons ont été posés (adhésion à la communauté ISNI et utilisation manuelle des identifiants ISNI ; tests d'alignement des notices avec celles de la BnF…) mais **il est temps de prouver l'intérêt de l'interconnexion**

Des objectifs communs, des enjeux institutionnels spécifiques : pour la Bibliothèque nationale de France

- Différentes applications web, différents entrepôts de données :
  - le catalogue général ;
  - Archives et manuscrits
  - …
- Expose ses données en Linked Open Data, à l'aide des technologies du web sémantique (http://data.bnf.fr), en utilisant les données d'autorité pour interconnecter les ressources des différents entrepôts
- Toutes les données de la BnF sont sous licence ouverte EtaLab
- <u>Problèmes</u> :
  - du côté du service des archives de l'institution : existence de notices d'autorité en EAC-CPF, hors du catalogue Archives et Manuscrits. Pas d'outil de publication ou de visualisation
  - souhait d'interconnecter ces données avec celles des autres services d'archives, en commençant par les Archives nationales (et en utilisant les identifiants ISNI, entre autres moyens)



Le projet du point de vue d'un historien

- E. Château-Dutier, professeur-adjoint au CRIHN à Montréal, travaille notamment sur l'histoire de l'administration de l'architecture publique au XIXe siècle.

- Le projet devrait lui permettre :
  - de mieux rendre compte des évolutions successives du service des Bâtiments civils et de tester une organisation plus pertinente des instruments de recherche à partir de la provenance ;

  - d'évaluer la pertinence et l'efficience des modèles documentaires archivistiques pour la description des fonctions et des collectivités ;

  - de mettre au point des visualisations diachroniques qui permettent d'explorer et de rendre compte des organigrammes administratifs historiques.

En résumé

- Des situations différentes
- Un **projet stratégique pour les trois partenaires institutionnels**
- Une ambition réelle
- Un intérêt très fort pour les méthodes et leçons du projet
- Des résultats qui pourraient aussi intéresser d'autres entités (lors des différentes actions de communication déjà menées, nombreuses manifestations d'intérêt)

Des visées qualitatives

- Le prototype sera développé et mis en production hors des SI des partenaires

- Jeux de métadonnées source de volumétrie réduite
  Il ne s'agit pas de traiter toutes les données disponibles, ce qui exigerait de prendre en compte une grande hétérogénéité et de choisir des outils très robustes

- **Les exigences concernent surtout :**

  - **la qualité des résultats :**
    - expressivité des données produites ;
    - pertinence et précision des relations ;
    - possibilités de raisonnement ;
    - clarté ;
    - ergonomie et réactivité de l'interface

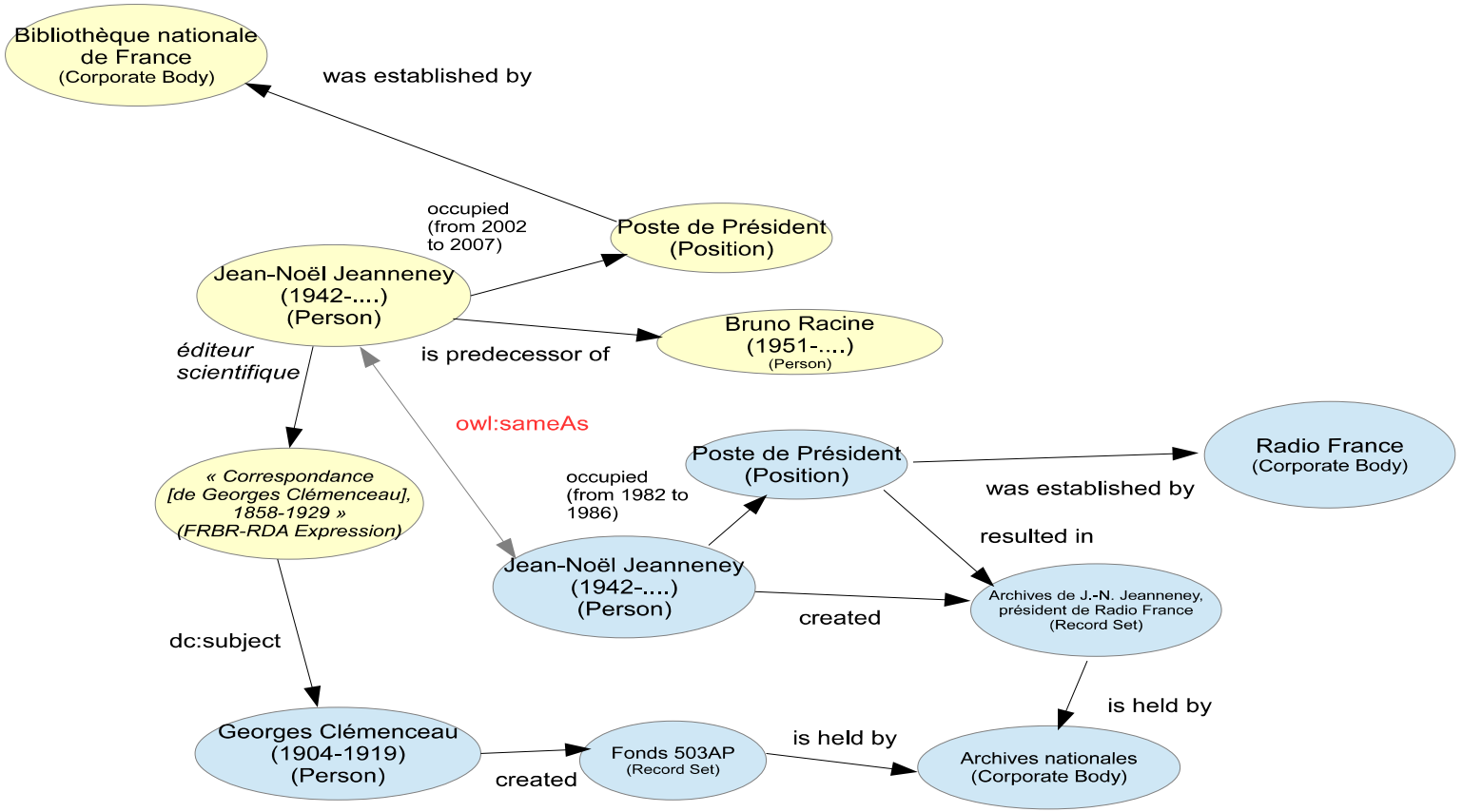
  - **la rigueur de la démarche**

Opportunités et défis technologiques

Une ontologie OWL pour la description des archives : RiC-O

- Par principe, pour le projet tel que défini, choix de ne pas élaborer d'ontologie spécifique
- Or jusqu'ici, pas d'ontologie générique du domaine des archives
  Seulement quatre normes conceptuelles publiées successivement par le CIA (Conseil International des Archives)
  Il existe quelques ontologies archivistiques élaborées dans le cadre de projets spécifiques, sans ambition de généricité
  Les ontologies génériques des domaines connexes du patrimoine culturel, quant à elles, ne sont pas adaptées pour exprimer pleinement les concepts et la complexité du domaine archivistique.
- Décision du CIA en 2013 : création d'un groupe d'experts, EGAD (Experts Group On Archival Description) pour élaborer un modèle conceptuel global (Records In Contexts-CM) et une ontologie formelle (Records In Contexts-O)
- RiC-CM : première version du modèle publiée en septembre 2016, avec appel à commentaires ouvert jusqu'au 31 décembre
  Voir <http://www.ica.org/fr/publication-de-records-contexts-par-legad>
- RiC-O : première version sera publiée probablement en février 2017

Un exemple très simple de graphe que RiC-O pourrait aider à produire



Mettre en œuvre pour la première fois RiC-O

- Florence Clavaud : pilote le projet de prototype, et coordonne le développement de RiC-O au sein du groupe EGAD
- Les métadonnées sélectionnées pour le projet décrivent des instances d'un sous-ensemble des classes définies par RiC-O
  Corporate Body, Person, Record Set, Record, Function (Abstract), Activity, Position…
- RiC-O devrait aussi être étendue pour le projet (ajout de propriétés et de classes)
- Le projet devrait donc :
  - servir de *use case* pour RiC-O
  - nourrir la réflexion du groupe EGAD sur RiC-O

La *data visualisation* sur le web

- Des technologies et librairies existent :
  - Langages de base : JavaScript, AJAX, JSON, HTML5
  - librairies web génériques et puissantes, pour la visualisation de données (comme D3), la génération de lignes de temps (ex. : Timeline du projet SIMILE)
  - autres développements web en cours dans des labos d'humanités numériques, souvent à partir de ces librairies, à suivre de près.
  Ex. : Palladio (http://palladio.designhumanities.org/#/), un projet du labo Humanities+Design (HDLab) de l'université de Stanford en Californie
  - Combinaison des deux (lignes de temps et graphe) sur le web plus rare
  Ex. : Kindred Britain (sur George Washington, http://kindred.stanford.edu/#/kin/half/half/none/I5457//)
- Mais rien de directement utilisable
- Globalement, notre problème est complexe, car le système à représenter est complexe :
  - différents types d'objets
  - différents types de relations
  - densité des relations
  - différentes perspectives institutionnelles à combiner
  - ...sans sacrifier la granularité informationnelle ni la lisibilité.

Le prototype de visualisation en graphe du projet Social Networks And Archival Contexts (SNAC)

social networks and archival context PROTOTYPE HOME ABOUT PROTOTYPE

Lafayette, Marie Joseph Paul Yves Roch Gilbert Du Motier, 1757-1834

Alternative names

Dates:
birth 1757-09-06
death 1834-05-20
Gender:

Authority Source: VIAF, aps, colu, uchic, ual, afl, NLA, vah, oac, LC, WorldCat, umd, umi, crnlu, harvard, lc, nara, nypl, pacscl, yale
Nationality: France
Language: French, English Unknown

Biographical notes:

The Marquis de Lafayette (1757-1834) was a French patriot who enlisted in the Continental Army with the rank of Major-General. He was second-in-command to General Washington. He also participated in the French Revolution until 1790 when he was forced out of the country under the rule of the Jacobin party. He returned to France in 1800 and was active in French and European politics until he died in 1834.

From the description of Marie Joseph Paul Yves Roch Gilbert Du Motier Lafayette letter, 1786. (Georgia Historical Society). WorldCat record id: 44416715

Lafayette's estate near Paris was called LaGrange.

From the description of Catalog of the library at LaGrange, 1822. (Unknown). WorldCat record id: 122458915

French statesman and soldier, major general in the American Revolution.

From the description of Arthur H. and Mary Marden Dean Lafayette Collection, 1520-1849. (Cornell University Library). WorldCat record id: 63936469

Lieutenant en 1773, lié avec Franklin, Marie-Joseph-Paul-Yves-Roch-Gilbert Motier, marquis de Lafayette (1757-1834) partit en Amérique en 1777 pour aider les insurgés. De retour en France, il contribua à décider le gouvernement à apporter son aide officielle à la guerre d'Indépendance américaine et repartit en 1780 pour l'Amérique, où il fut nommé maréchal de camp en 1782. Franc-maçon, adepte des idées nouvelles, il fut favorable au doublement du tiers état lors de la réunion de l'Assemblée des notables de 1788. Il fut élu député de la noblesse aux états généraux de 1789. Nommé commandant de la Garde nationale après la prise de la Bastille, il voulut être l'instrument de la réconciliation du roi et de la Révolution, lors des journées des 5 et

Links to collections

- Archival Collections 1214
- Related Resources 101
- Related External Links 9

Related names in SNAC

- People 552
- Families 52
- Organizations 92

Visual

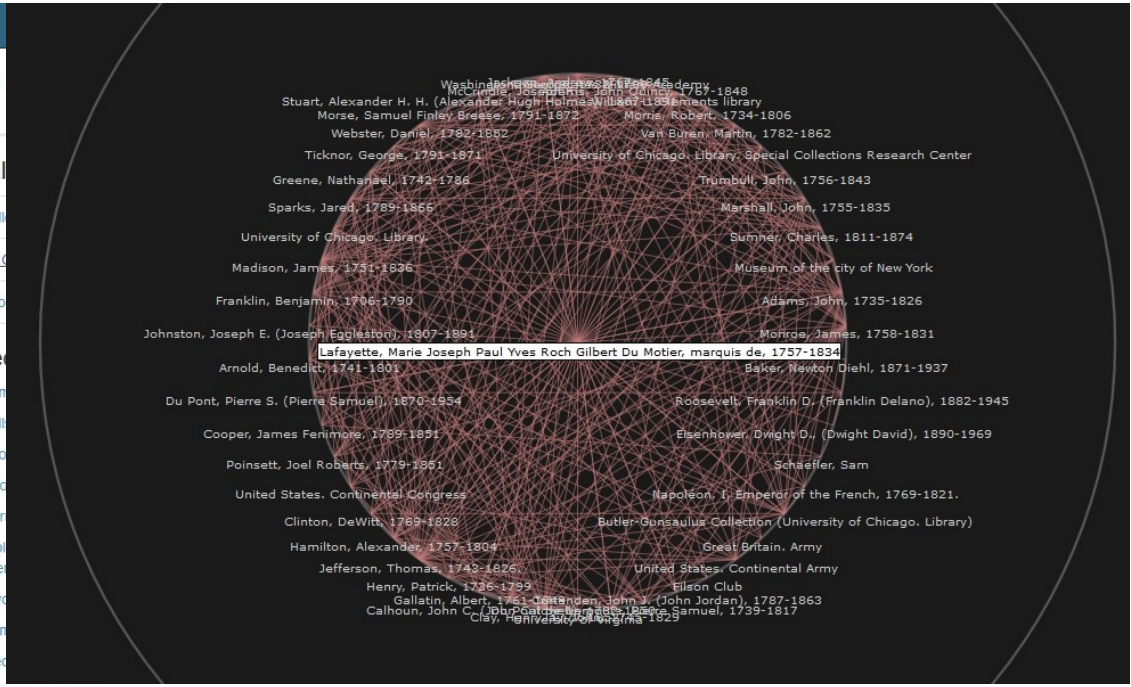
List coll

Radial

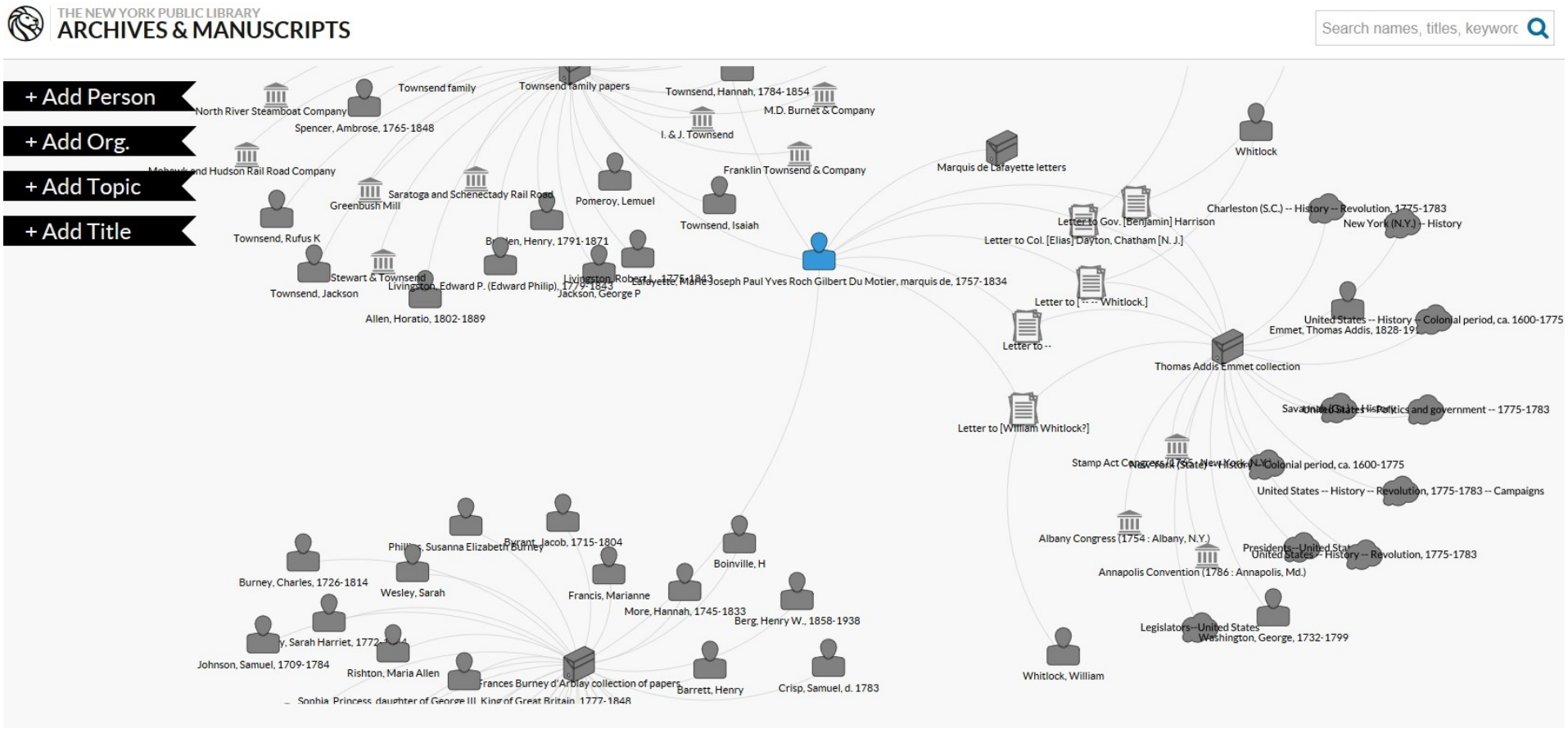
View so

Subje

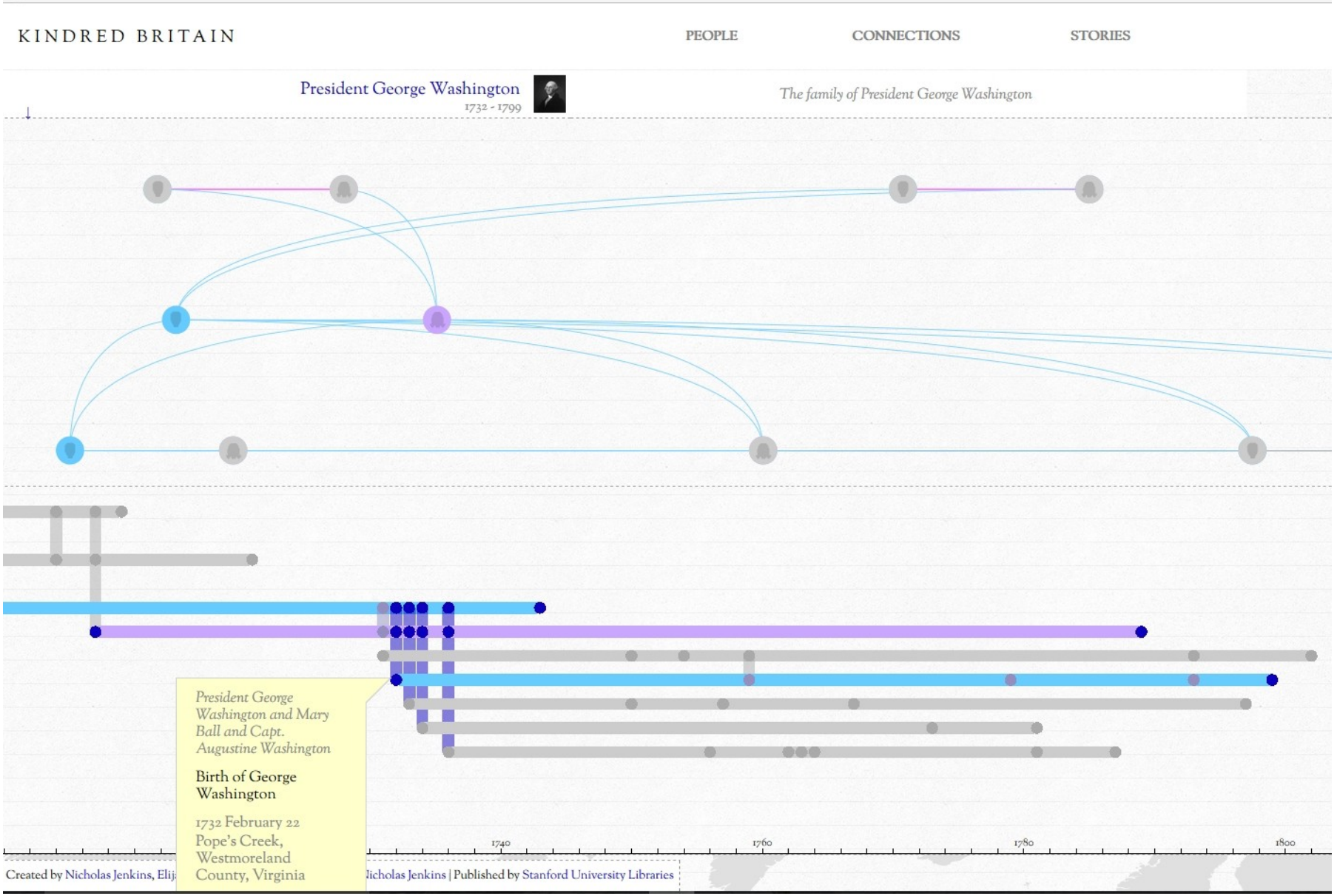
- Adm
- Ball
- Boo
- Cho
- Cor
- Dipl Fre
- Divo
- Don
- Elec
- Engraving--History
- Gifts
- Imprisonment
- Land grants
- Land tenure--History
- Land tenure--Law and legislation



The New York Public Library Archives and Manuscripts. Beta Tools & Experiments



Kindred Britain



Sélection et préparation des jeux de métadonnées source

Définition des jeux de métadonnées

- Objectif : identifier, au sein des métadonnées provenant des quatre partenaires, des jeux à la fois **pertinents**, **maîtrisables** et **articulables**
- Choix de **deux sous-domaines fonctionnels relevant actuellement, directement ou non, du ministère de la Culture et de la Communication** :
  - la gestion des monuments historiques et des « bâtiments civils » (édifices publics)
  - la lecture publique et les bibliothèques
- Ces domaines :
  - sont les champs de compétence d'entités qui sont des subdivisions des administrations centrales de l'État, des opérateurs nationaux, et aussi de services déconcentrés et de collectivités territoriales ;
  - s'inscrivent dans une large amplitude temporelle : la période contemporaine ;
  - sont des centres d'intérêt pour tous les partenaires, et l'objet de certains de leurs travaux actuels

Méthode

- Point de départ : les notices d'autorité de chaque institution sur les acteurs-producteurs (des collectivités, et aussi des personnes)
- Prise en compte de quelques instruments de recherche archivistiques existants décrivant les archives de ces entités
- Chaque partenaire utilise ses outils et conserve sa propre perspective (pas de règles communes précises de rédaction)
- Cependant, consensus sur :

  - les formats d'encodage (XML/EAC-CPF et XML/EAD 2002)

  - la forme des noms des entités (respect des normes françaises en vigueur)

  - les informations obligatoires

  - la typologie des relations (cf. ontologie RiC-O)

- Décision d'élaborer ensemble deux vocabulaires communs pour l'indexation des domaines fonctionnels et des activités des collectivités.
  Utilisation de l'instance MCC du logiciel Ginco pour la saisie de ces vocabulaires et leur export au format RDF/SKOS

Le point à ce jour

- Les jeux de métadonnées sont prêts
  - environ 300 notices d'autorité
  - environ 50 instruments de recherche synthétiques ou analytiques
  - deux vocabulaires d'indexation des notices d'autorité

- Il y a bel et bien chevauchement et complémentarités entre ces jeux de métadonnées, au moins via les notices d'autorité :
  - chevauchements significatifs entre notices de la BnF et des AN (personnes liées aux deux institutions, écosystèmes administratifs qui se recoupent) ;
  - notices complémentaires : le réseau constitué par chacun des quatre jeux de notices étend ceux constitués par les autres jeux

  > Pertinence de l'objectif d'interconnexion

Quelques problématiques à prendre en compte

- La dimension historique, et les divergences dans la représentation du temps

  Pour les collectivités, les AN identifient le plus souvent plus d'entités successives que la BnF ou le SIAF

- Nécessité de conserver la trace de la provenance, voire des informations de gestion, des jeux de métadonnées
- Trouver la meilleure solution pour représenter dans l'ontologie de référence (donc, une extension de RiC-O) les catégories de collectivités territoriales décrites par le SIAF

Le cahier des charges du prototype

Quelques éléments significatifs

- Forte interaction souhaitée avec le titulaire du marché (approche agile)

- Première tâche : analyse de la qualité des jeux de données RDF fournis

- En ce qui concerne le prototype logiciel proprement dit :
  - deux types d'utilisateurs finaux : le professionnel du patrimoine culturel, le public (utilisateurs des métadonnées archivistiques, amateurs ou chercheurs) ;
  - montée en charge industrielle à prendre en compte dans l'évaluation des solutions techniques d'implémentation ;
  - fonctionnalités les plus importantes : alignement de données, inteerogation experte des données et construction de nouveaux triplets par inférence ; exploration des données via des dispositifs de visualisation interactifs.

28

Conclusion (très provisoire)

En guise de conclusion provisoire

- Nécessité de travailler ensemble sur un projet pendant plusieurs années, sans oublier les dynamiques, intérêts et perspectives de chacune des institutions

  Assez nouveau pour les partenaires !

- Beaucoup de temps sera consacré fin 2017 à un bilan critique collectif, qui comme les sources logicielles du prototype, devrait servir plus longtemps que le démonstrateur proprement dit, et ouvrir des perspectives

Merci !