Assister l'édition manuelle de données RDF à l'aide du raisonnement à partir de cas

Journées IC 2021

Nicolas Lasolle, Olivier Bruneau, Jean Lieber,
Emmanuel Nauer, Siyana Pavlova

2 juillet 2021

Contexte

Web sémantique

- Données structurées selon le modèle RDF
- Connaissances du domaine représentées en RDFS
- Graphes RDF interrogeables et modifiables grâce à SPARQL

Application pour un corpus historique

Les travaux présentés sont appliqués pour enrichir et exploiter le corpus de la correspondance d'Henri Poincaré.

Problématique

Constat

L'édition manuelle de données RDF est une tâche fastidieuse avec un risque d'erreurs important :

- Erreur de duplication ;
- Erreur d'ambiguïté ;
- Erreur de frappe.

Problématique

Constat

L'édition manuelle de données RDF est une tâche fastidieuse avec un risque d'erreurs important :

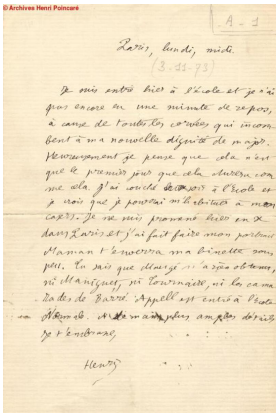
- Erreur de duplication ;
- Erreur d'ambiguïté ;
- Erreur de frappe.

Objectif

Proposer, implémenter, et évaluer des méthodes pour assister les contributeurs dans cette tâche.

Le corpus de la correspondance d'Henri Poincaré

Le corpus de la correspondance d'Henri Poincaré

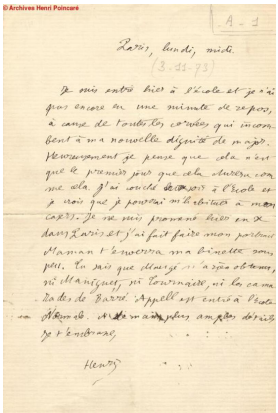


Henri Poincaré à Aline Poincaré, 1873

Un corpus historique

- Plus de 2000 lettres (échanges d'ordre scientifique, administratif et privé)

Le corpus de la correspondance d'Henri Poincaré

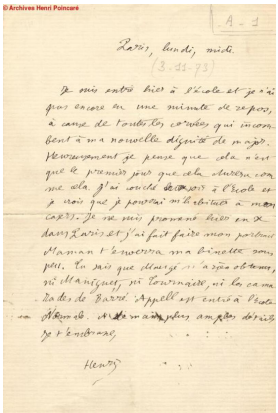


Henri Poincaré à Aline Poincaré, 1873

Un corpus historique

- Plus de 2000 lettres (échanges d'ordre scientifique, administratif et privé)
- Source d'informations importante pour les historiens

Le corpus de la correspondance d'Henri Poincaré

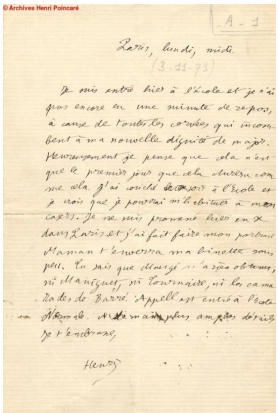


Henri Poincaré à Aline Poincaré, 1873

Un corpus historique

- Plus de 2000 lettres (échanges d'ordre scientifique, administratif et privé)
- Source d'informations importante pour les historiens
  - Théories scientifiques

Le corpus de la correspondance d'Henri Poincaré

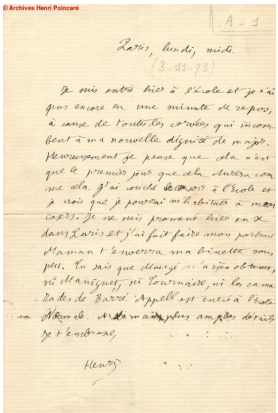


Henri Poincaré à Aline Poincaré, 1873

Un corpus historique

- Plus de 2000 lettres (échanges d'ordre scientifique, administratif et privé)
- Source d'informations importante pour les historiens
  - Théories scientifiques
  - Institutions et sociétés savantes

Le corpus de la correspondance d'Henri Poincaré

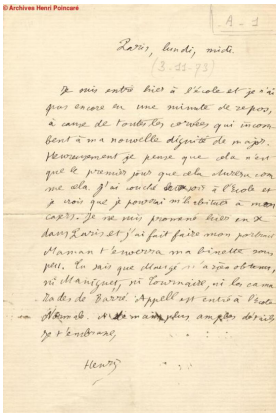


Henri Poincaré à Aline Poincaré, 1873

Un corpus historique

- Plus de 2000 lettres (échanges d'ordre scientifique, administratif et privé)
- Source d'informations importante pour les historiens
  - Théories scientifiques
  - Institutions et sociétés savantes
  - Contexte politique, social et culturel

Le corpus de la correspondance d'Henri Poincaré



Henri Poincaré à Aline Poincaré, 1873

Un corpus historique

- Plus de 2000 lettres (échanges d'ordre scientifique, administratif et privé)
- Source d'informations importante pour les historiens
  - Théories scientifiques
  - Institutions et sociétés savantes
  - Contexte politique, social et culturel
- L'édition de ce corpus est un projet collectif mené par les Archives Henri-Poincaré

Le corpus de la correspondance d'Henri Poincaré

L'édition numérique et le site Omeka S

LA CORRESPONDANCE D'HENRI POINCARÉ

Accueil | Lettres | Recherche | Index

LETTRE : Henri Poincaré à Aline Boutroux - 3 novembre 1873

Modifier l'item

Transcription | Métadonnées | Citer ce document

[3 novembre 1873[1]]

Paris, lundi, midi.

Je suis entré hier à l'école[2] et je n'ai pas encore eu une minute de repos, à cause de toutes les nouvelles corvées qui incombent à ma nouvelle dignité de major. Heureusement je pense que cela n'est que le premier jour que cela durera comme cela. J'ai couché hier soir à l'École et je crois que je pourrai m'habituer à mon casern. Je me suis promené hier en X[3] dans Paris et j'ai fait faire mon portrait[4]. Maman t'enverra ma binette sous peu. Tu sais que Maugé[5] n'a rien obtenu, ni Maniguet[6], ni Tournaire, ni les camarades de Barré[7]. Appell est entré à l'École Normale[8]. À demain plus amples détails.

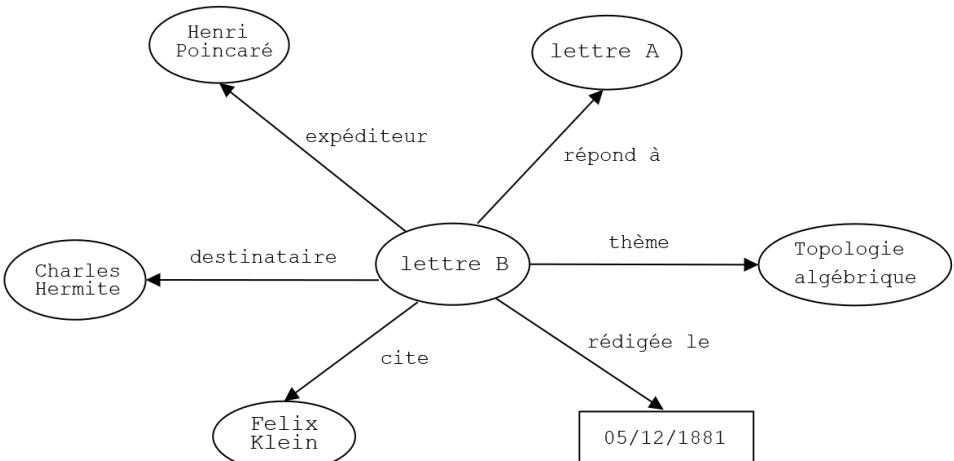
Je t'embrasse,

Henri

1. Comme de nombreuses autres, cette lettre n'avait pas de destinataire précis. La référence à Eugénie Poincaré laisse supposer qu'elle s'adressait plutôt à la sœur de Poincaré, Aline. On sait que toute la famille avait accompagné le jeune polytechnicien à Paris pour sa rentrée. Il est est probable qu'Eugénie Poincaré était restée quelques jours à Paris tandis qu'Aline et Émile-Léon Poincaré étaient retournés à Nancy.↩

Le corpus de la correspondance d'Henri Poincaré

Resource Description Framework (RDF)



Le corpus de la correspondance d'Henri Poincaré

SPARQL

Ce langage permet de formuler des requêtes afin d'explorer le graphe RDF du corpus

Exemple de requête (informelle)

$$\mathcal{Q} = \left| \begin{array}{l} \text{Donner les lettres envoyées par } \texttt{Henri Poincaré} \\ \text{à } \texttt{Félix Klein} \text{ entre 1890 et 1900 et ayant} \\ \text{pour thème la } \texttt{géométrie} \end{array} \right.$$

Des méthodes pour assister l'édition manuelle de données RDF

Comment assister l'édition manuelle de données RDF ?

Problématique

L'édition manuelle de données (i.e. saisir les triplets) est souvent une tâche longue et fastidieuse.

Comment assister l'édition manuelle de données RDF ?

Problématique

L'édition manuelle de données (i.e. saisir les triplets) est souvent une tâche longue et fastidieuse.

Objectif

Fournir une liste de suggestions ordonnées pour assister les contributeurs dans cette tâche.

Comment assister l'édition manuelle de données RDF ?

Problématique

L'édition manuelle de données (i.e. saisir les triplets) est souvent une tâche longue et fastidieuse.

Objectif

Fournir une liste de suggestions ordonnées pour assister les contributeurs dans cette tâche.

Questions d'édition

$\langle \boxed{?s} \; p \; o \rangle$ $\langle s \; \boxed{?p} \; o \rangle$ $\langle s \; ?p \; \boxed{?o} \rangle$ $\langle s \; \boxed{?p} \; ?o \rangle$ ...

Comment assister l'édition manuelle de données RDF ?

Problématique

L'édition manuelle de données (i.e. saisir les triplets) est souvent une tâche longue et fastidieuse.

Objectif

Fournir une liste de suggestions ordonnées pour assister les contributeurs dans cette tâche.

Questions d'édition

$\langle \boxed{?s} \;\; p \;\; o \rangle$ $\langle s \;\; \boxed{?p} \;\; o \rangle$ $\langle s \;\; ?p \;\; \boxed{?o} \rangle$ $\langle s \;\; \boxed{?p} \;\; ?o \rangle$ ...

Exemple

$\langle$ `lettreA destinataire` $\boxed{?o} \rangle$

Proposition d'un outil d'édition

4 versions du système de suggestions :

- **basique** classe les suggestions selon l'ordre alphabétique ;
- **déductif** bénéficie de l'utilisation des connaissances de l'ontologie ;
- **à base de cas** s'appuie sur des problèmes d'édition similaires au problème courant ;
- **combiné** combine les deux précédentes méthodes.

Proposition d'un outil d'édition

4 versions du système de suggestions :

- **basique** classe les suggestions selon l'ordre alphabétique ;
- **déductif** bénéficie de l'utilisation des connaissances de l'ontologie ;
- **à base de cas** s'appuie sur des problèmes d'édition similaires au problème courant ;
- **combiné** combine les deux précédentes méthodes.

Proposition d'un outil d'édition

4 versions du système de suggestions :

- **basique** classe les suggestions selon l'ordre alphabétique ;
- **déductif** bénéficie de l'utilisation des connaissances de l'ontologie ;
- **à base de cas** s'appuie sur des problèmes d'édition similaires au problème courant ;
- **combiné** combine les deux précédentes méthodes.

Proposition d'un outil d'édition

4 versions du système de suggestions :

- **basique** classe les suggestions selon l'ordre alphabétique ;
- **déductif** bénéficie de l'utilisation des connaissances de l'ontologie ;
- **à base de cas** s'appuie sur des problèmes d'édition similaires au problème courant ;
- **combiné** combine les deux précédentes méthodes.

Méthodologie

Le système s'appuie sur les domaines (`rdfs:domain`) et co-domaines (`rdfs:range`) des propriétés de l'ontologie.

Question d'édition $\langle$`lettreA destinataire` `?o`$\rangle$

Solution pour le classement des valeurs potentielles

Comme les classes `Personne` et `Institution` font partie du co-domaine de la propriété `destinataire`, cette connaissance est utilisée pour favoriser les instances de ces classes.

Méthodologie

Le système s'appuie sur les domaines (`rdfs:domain`) et co-domaines (`rdfs:range`) des propriétés de l'ontologie.

Question d'édition $\langle$`lettreA destinataire` `?o`$\rangle$

Solution pour le classement des valeurs potentielles

Comme les classes `Personne` et `Institution` font partie du co-domaine de la propriété `destinataire`, cette connaissance est utilisée pour favoriser les instances de ces classes.

Limite

Plus de 1000 personnes dans le corpus !

Raisonnement à partir du cas

Une méthodologie qui s'appuie sur des situations précédentes pour résoudre des problèmes nouveaux.

Raisonnement à partir du cas

Une méthodologie qui s'appuie sur des situations précédentes pour résoudre des problèmes nouveaux.

Définitions

Cas : $(\mathtt{x}, \mathtt{y})$ où $\mathtt{x}$: problème et $\mathtt{y}$: solution
Problème cible : le problème à résoudre $\mathtt{x}^{\mathtt{cible}}$
Cas source : $(\mathtt{x}^{\mathtt{s}}, \mathtt{y}^{\mathtt{s}}) \in \mathit{BaseCas}$ utilisé pour résoudre $\mathtt{x}^{\mathtt{cible}}$

Assister l'édition manuelle de données RDF

Système à base de cas

Le problème cible

$x^{\texttt{cible}} =$

question : ⟨lettreA destinataire ?o⟩

contexte :
⟨lettreA expéditeur henriPoincaré⟩
⟨lettreA thème topologieAlgébrique⟩
⟨lettreA cite charlesHermite⟩
⟨lettreA rédigéeEn 1880⟩

Assister l'édition manuelle de données RDF

Système à base de cas

Le problème cible

$x^{\texttt{cible}} =$

question : ⟨lettreA destinataire ?o⟩

contexte :
⟨lettreA expéditeur henriPoincaré⟩
⟨lettreA thème topologieAlgébrique⟩
⟨lettreA cite charlesHermite⟩
⟨lettreA rédigéeEn 1880⟩

$y^{\texttt{cible}}$ = valeur pour ?o

Assister l'édition manuelle de données RDF

Système à base de cas

Le problème cible

$x^{\texttt{cible}} =$

question : ⟨lettreA destinataire ?o⟩

contexte :
⟨lettreA expéditeur henriPoincaré⟩
⟨lettreA thème topologieAlgébrique⟩
⟨lettreA cite charlesHermite⟩
⟨lettreA rédigéeEn 1880⟩

$y^{\texttt{cible}}$ = valeur pour ?o

Base de cas = la base RDF $\mathcal{D}$

Assister l'édition manuelle de données RDF

Système à base de cas

Le problème cible

$x^{\texttt{cible}} =$

question : ⟨lettreA destinataire ?o⟩

contexte :
⟨lettreA expéditeur henriPoincaré⟩
⟨lettreA thème topologieAlgébrique⟩
⟨lettreA cite charlesHermite⟩
⟨lettreA rédigéeEn 1880⟩

$y^{\texttt{cible}}$ = valeur pour ?o

Base de cas = la base RDF $\mathcal{D}$

x^s = une ressource de $\mathcal{D}$

y^s = une valeur candidate pour ?o
$\langle x^s \ \texttt{destinataire} \ y^s \rangle$

Assister l'édition manuelle de données RDF

Système à base de cas

$x^{\texttt{cible}} =$

question : ⟨lettreA destinataire ?o⟩

contexte :
⟨lettreA expéditeur henriPoincaré⟩
⟨lettreA thème topologieAlgébrique⟩
⟨lettreA cite charlesHermite⟩
⟨lettreA rédigéeEn 1880⟩

Requête SPARQL initiale

Retrouver les ressources pour lesquelles le contexte est le même que `lettreA`.

Requête initiale

$$\mathcal{Q} = \left| \begin{array}{l} \text{Donner les lettres envoyées par Poincaré en 1880} \\ \text{ayant pour thème la } \texttt{topologie algébrique} \\ \text{et citant } \texttt{Charles Hermite}. \end{array} \right.$$

Requête initiale

$$\mathcal{Q} = \left| \begin{array}{l} \text{Donner les lettres envoyées par Poincaré en 1880} \\ \text{ayant pour thème la } \texttt{topologie algébrique} \\ \text{et citant } \texttt{Charles Hermite}. \end{array} \right.$$

Problème

Ce n'est pas courant d'avoir deux ressources avec le même contexte. Comment retrouver les cas les plus similaires dans la base ?

Requête initiale

$$\mathcal{Q} = \left| \begin{array}{l} \text{Donner les lettres envoyées par Poincaré en 1880} \\ \text{ayant pour thème la } \texttt{topologie algébrique} \\ \text{et citant } \texttt{Charles Hermite}. \end{array} \right.$$

Problème

Ce n'est pas courant d'avoir deux ressources avec le même contexte. Comment retrouver les cas les plus similaires dans la base ?

Proposition

Utiliser d'un mécanisme de recherche flexible s'appuyant sur des règles de transformation de requêtes SPARQL.

Assister l'édition manuelle de données RDF

Système à base de cas

Requête initiale

$$\mathcal{Q} = \left| \begin{array}{l} \text{Donne moi les lettres envoyées par Poincaré en 1880} \\ \text{ayant pour thème la } \texttt{topologie algébrique} \\ \text{et citant } \texttt{Charles Hermite}. \end{array} \right.$$

Requête générée (en généralisant le thème)

$$\mathcal{Q}_1 = \left| \begin{array}{l} \text{Donner les lettres envoyées par Poincaré en 1880} \\ \text{ayant pour thème les } \texttt{mathématiques} \\ \text{et citant } \texttt{Charles Hermite}. \end{array} \right.$$

Assister l'édition manuelle de données RDF

Système à base de cas

Requête initiale

$$\mathcal{Q} = \left| \begin{array}{l} \text{Donner les lettres envoyées par Poincaré en 1880} \\ \text{ayant pour thème la } \texttt{topologie algébrique} \\ \text{et citant } \texttt{Charles Hermite}. \end{array} \right.$$

Requête générée (en remplaçant une personne citée par une personne liée)

$$\mathcal{Q}_2 = \left| \begin{array}{l} \text{Donner les lettres envoyées par Poincaré en 1880} \\ \text{ayant pour thème la } \texttt{topologie algébrique} \\ \text{et citant } \texttt{Paul Appell}. \end{array} \right.$$

Assister l'édition manuelle de données RDF

Système à base de cas

Requête initiale

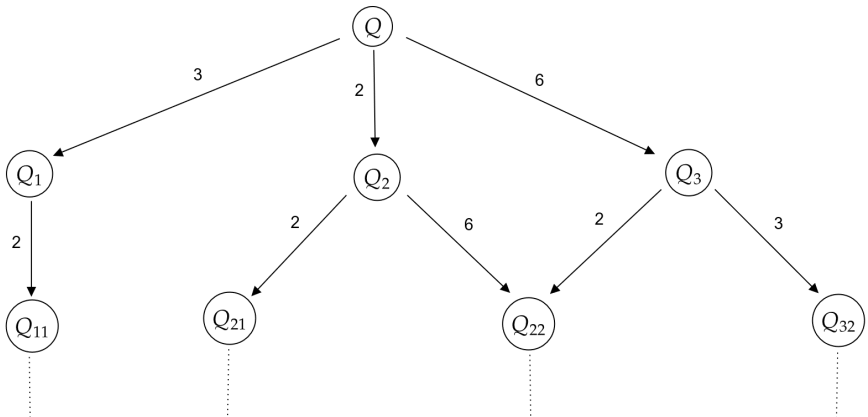
$$\mathcal{Q} = \left| \begin{array}{l} \text{Donner les lettres envoyées par Poincaré en 1880} \\ \text{ayant pour thème la } \texttt{topologie algébrique} \\ \text{et citant } \texttt{Charles Hermite}. \end{array} \right.$$

Requête générée (en étendant les bornes temporelles liées à la date de rédaction)

$$\mathcal{Q}_3 = \left| \begin{array}{l} \text{Donner les lettres envoyées par Poincaré entre 1875} \\ \text{et 1885 ayant pour thème la } \texttt{topologie algébrique} \\ \text{et citant } \texttt{Charles Hermite}. \end{array} \right.$$

Assister l'édition manuelle de données RDF

Coûts de transformation



Implémentation

Évaluation des méthodes

Évaluation humaine

Méthodologie

- Édition manuelle d'un ensemble de 100 lettres inédites complétant l'actuel corpus
- La base de cas est la base actuelle du corpus
- Pour chaque version du système de suggestions présentées dans un ordre aléatoire et inconnu, l'évaluateur :
  - édite l'ensemble des lettres ;
  - complète un questionnaire à propos du système ;
  - Attribue un score relatif à la pertinence des suggestions pour chaque propriété du système.

Évaluation humaine

Méthodologie

- Édition manuelle d'un ensemble de 100 lettres inédites complétant l'actuel corpus
- La base de cas est la base actuelle du corpus
- Pour chaque version du système de suggestions présentées dans un ordre aléatoire et inconnu, l'évaluateur :
  - édite l'ensemble des lettres ;
  - complète un questionnaire à propos du système ;
  - Attribue un score relatif à la pertinence des suggestions pour chaque propriété du système.

Résultats

	Basique	Déductif	À base de cas	Combiné
Score moyen (De 1 à 7)	3,4	5,7	5,3	7

Évaluation automatique

Méthodologie

- 200 lettres aléatoirement extraites du corpus
- Les triplets correspondants sont utilisés pour simuler des questions d'édition
- Le rang de la valeur attendue est sauvegardé pour chaque version du système de suggestions

Évaluation automatique

Méthodologie

- 200 lettres aléatoirement extraites du corpus
- Les triplets correspondants sont utilisés pour simuler des questions d'édition
- Le rang de la valeur attendue est sauvegardé pour chaque version du système de suggestions

Résultats

	Basique	Déductif	À base de cas	Combiné
$rang \leqslant 5$	2,7%	19,23%	33,6%	34,7%
$rang \leqslant 10$	7,1%	21,15%	43,2%	43,2%
$rang \leqslant 15$	11,3%	22,11%	49,0%	49,5%

Merci de votre attention !

nicolas.lasolle@univ-lorraine.fr