Développement d'une ontologie pour l'analyse de textes de décisions administratives d'Ancien Régime par des Grands Modèles de Langage

Gilles Falquet[1], Christophe Chazalon[1], Marco Sorbi[1], Stéphane Marchand-Maillet[1], Laurent Moccozet[1]

[1] Université de Genève, Centre Universitaire d'Informatique

Gilles.Falquet@unige.ch

Résumé

Cet article décrit la démarche qui a été mise en place pour le développement et l'implémentation d'une ontologie contextuelle dans le cadre du projet RCnum. Le but de cette ontologie est de faciliter l'analyse ultérieure des transcriptions des textes des Registres des Conseils de la ville de Genève par des Grands Modèles de Langages pour en permettre l'interrogation et l'exploration. Ces textes répertorient des décisions administratives concernant la ville de Genève à l'époque de Jean Calvin. La conception s'appuie sur les textes déjà collationnés et indexées pour une version papier pour guider la construction de l'ontologie et l'initialisation du graphe de connaissances. La démarche concilie le respect des contraintes historiques et techniques.

Mots-clés

Documents historiques, Entités nommées, Graphe de connaissance historique, Indexation sémantique, Ontologie contextuelle, RAG

Abstract

This article describes the approach that has been adopted for the development and implementation of a contextual ontology within the framework of the RCnum project. The aim of this ontology is to facilitate the subsequent analysis of the transcriptions of the texts of the Registers of the Council of the City of Geneva by Large Language Models to enable them to be queried and explored. These texts record administrative decisions concerning the city of Geneva at the time of John Calvin. The design is based on the texts already collated and indexed for a paper version to guide the construction of the ontology and the initialisation of the knowledge graph. The approach reconciles respect for historical and technical constraints.

Keywords

Historical documents, Named entities, Historical knowledge graph, Semantic indexing, Contextual ontology, RAG

1 Introduction

RCnum, projet basé sur les Registres des Conseils de Genève (RC) de 1545 à 1550, vise la création d'une plateforme informatique pérenne, ouverte et *open source*, proposant aux utilisatrices et utilisateurs, outre la transcription du texte, sa modernisation et sa traduction dans d'autres langues, ainsi qu'une série d'outils numériques et pédagogiques de visualisation et d'exploration avancée. Un point central du projet est la création d'un graphe de connaissance pour 1) le stockage des connaissances extraites des registres (transcription, traduction, entités et relations.) ; 2) l'accès au contenu des registres et aux connaissances connexes (navigation et recherche dans les textes, création de documents virtuels répondant à des questions, visualisation des connaissances, ...). Dans ce cadre, la création d'une ontologie orientée sur des décisions administratives d'Ancien Régime s'avère nécessaire car, à notre connaissance, une telle ontologie n'existe pas encore.

1.1 Cadre linguistique et terminologique

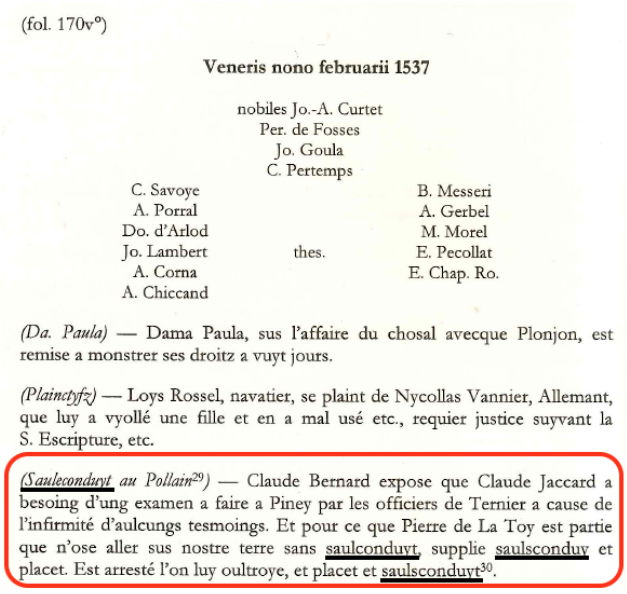
Les RC consistent en une série quasi complète de comptes rendus des séances et décisions des Conseils de la cité de Genève de 1409 à 1792 (c.f. Figure 1), date de la conquête napoléonienne. Deux séries ont été éditées couvrant les années 1409 à 1544. Le projet RCnum traite les RC des années 1545-1550 déjà transcrits. Malgré cet intervalle de temps court, plusieurs difficultés importantes restent à surmonter :

— la langue des RC (1545-1550) est en « ancien français », une notion mal définie oscillant entre « moyen français » et « français préclassique », en passant par le patois local issu du franco-provençal. De plus, à cette date, si le « francien » est officiellement la langue utilisée dans l'administration genevoise, on y trouve également des fragments en latin.

— l'absence de règles orthographiques, grammaticales et syntaxiques laisse une grande liberté aux scribes pour expérimenter ou proposer plusieurs graphies pour un même mot, souvent phonétiques.

— la diversité lexicale et d'écriture des 3 secrétaires des Conseils actifs entre 1545-1550 impose une adaptation des technologies numériques renforcée.

— le mélange entre « mise au propre » et « brouillon » des procès-verbaux rend la lecture difficile, tout autant que leur traitement informatique.



(fol. 170v°)

Veneris nono februarii 1537

nobiles Jo.-A. Curtet
Per. de Fosses
Jo. Goula
C. Pertemps

C. Savoye, A. Porral, Do. d'Arlod, Jo. Lambert, A. Corna, A. Chiccand — thes. — B. Messeri, A. Gerbel, M. Morel, E. Pecollat, E. Chap. Ro.

(Da. Paula) — Dama Paula, sus l'affaire du chosal avecque Plonjon, est remise a monstrer ses droitz a vuyt jours.

(Plainctyfz) — Loys Rossel, navatier, se plaint de Nycollas Vannier, Allemant, que luy a vyollé une fille et en a mal usé etc., requier justice suyvant la S. Escripture, etc.

(Sauleconduyt au Pollain[29]) — Claude Bernard expose que Claude Jaccard a besoing d'ung examen a faire a Piney par les officiers de Ternier a cause de l'infirmité d'aulcungs tesmoings. Et pour ce que Pierre de La Toy est partie que n'ose aller sus nostre terre sans saulconduyt, supplie saulsconduy et placet. Est arresté l'on luy oultroye, et placet et saulsconduyt[30].

FIGURE 1 – Extrait des Registres des Conseils de Genève au 9 Février 1537 : le 3ème paragraphe présente les graphies multiples de "sauf-conduit". *R.C. impr.*, t. II/1, p. 64

— la mise en place progressive d'une nouvelle administration se traduit par des tâtonnements, des hésitations et des décisions hétérogènes, parfois mêmes contradictoires, au fur et à mesure des années. Ainsi, au départ, en 1536, on y retrouve pêle-mêle politique intérieure, politique extérieure, finances publiques, législation, justice criminelle ou civile, affaires particulières, affaires religieuses, aumônes, etc. La spécialisation progressive influe sur le lexique utilisé et son volume.

Pour illustrer cette multiplicité de graphies dans un seul paragraphe, l'exemple de Claude Roset suffira. En février 1537, il propose 4 graphies différentes du mot *sauf-conduit* (c.f. Figure 1) : Sauleconduyt ; saulconduyt ; saulsconduy et saulsconduyt (*R.C. impr.*, n.s., t. II/1, p. 64 (fol. 170v), 1537-02-09). Ce cadre nous a amené à construire une ontologie qui possède une forte composante terminologique (représentée par des propriétés d'annotation).

1.2 Cadre spatio-temporel

La contextualisation temporelle et spatiale des connaissances est essentielle dans ce projet car :

— la terminologie ayant évolué depuis le XVIe, l'utilisateur actuel des RC doit pouvoir y accéder avec les termes modernes. Il est donc nécessaire pour chaque concept de savoir quels étaient les termes en vigueur à une époque donnée pour le désigner. Dans le projet actuel les deux spatio-contextes temporels à considérer sont (Genève, 1536-1550), (Europe, XXIe).

— les connaissances ont une validité spatio-temporelle. Par exemple l'assertion « Jean Favre est membre du Conseil des LX » est vraie de 1540 à 1544, ou bien « La danse est une infraction » est vraie à Genève au XVIe siècle (mais plus au XXIe).

Par conséquent nous visons à créer un graphe de connaissances de modèle similaire à Wikidata. où les énoncés peuvent être qualifiés par un domaine de validité spatio-temporel.

1.3 Connexion avec les outils d'IA

Le graphe de connaissances (GdC) agit en étroite relation avec l'IA, l'un l'autre se soutenant mutuellement :

— des outils d'extraction de connaissances sont et seront utilisés pour peupler le GdC, par exemple pour associer des personnes, des lieux, des actions aux objets (décisions) des RC ou pour établir des liens entre personnes (parenté, économique,...). Dans cette situation, il est important de conserver la provenance exacte des connaissances (sources et méthodes utilisées) ainsi que des taux de confiance.

— le GdC sera utilisé pour augmenter le système de recherche d'informations dans les RC (technique RAG). Il devra permettre l'expérimentation et la comparaison de différentes techniques d'augmentation basées sur les textes bruts, sur les textes traduits, sur les énoncés du graphe, etc.

— l'un des buts du GdC est de permettre d'inférer des connaissances latentes, soit à l'aide de moteurs d'inférence utilisant des règles ou des axiomes de l'ontologie, soit par des techniques de génération d'arêtes par apprentissage automatique.

2 Élaboration d'une méthodologie pour la création d'une ontologie orientée sur des décisions administratives d'Ancien Régime

Le projet RCnum se concentre sur les années 1545-1550, mais le développement de l'ontologie se base sur le contenu des RC de 1536 à 1544. Ce choix s'explique par la qualité et la complétude des données pour ces années, qui ont bénéficié d'une édition papier avec transcription, annotation et indexation. En revanche, les années 1545-1550, bien que transcrites, ne sont pas collationées. Elles sont donc moins fiables, ce qui pourrait compromettre la création d'une ontologie précise sans un important travail de nettoyage ultérieur.
Plus encore, les index offrent un point de départ utile et efficace pour l'ontologie car ils contiennent déjà en substance l'essentiel des principaux concepts à la base des classes et sous-classes que l'on retrouvera pour les RC (1545-1550).
Enfin, il existe plusieurs dictionnaires historiques qui répertorient graphies ou définitions des concepts en usage au XVIe siècle, permettant un référencement scientifique.

2.1 Identification et définition des concepts

Dès lors, deux approches possibles existent :

— extraire les concepts via des TAL (Traitement automatique du langage naturel) pour créer automatiquement l'ontologie, puis en faire remodeler et améliorer le résultat par des expert-e-s.

— créer manuellement l'ontologie avec les expert-e-s, puis utiliser des outils numériques pour l'améliorer et la compléter, en particulier par inférence.

Nous avons opté pour la seconde approche en raison de la complexité du sujet (lexiques et graphies), qui aurait nécessité trop de corrections manuelles avec la première [9, 6]. Ce choix a des conséquences importantes. Déjà, le temps requis pour cette tâche croît fortement car les expert-e-s, souvent non compétents en informatique, doivent apprendre sur le tas. La collaboration interdisciplinaire est essentielle, l'informaticien-ne devant former l'expert-e. Plus encore, la création d'une ontologie formelle exige que chaque concept ait une définition unique, ce qui est contraire à l'approche des historien-ne-s, qui tendent à regrouper les concepts. Ainsi, pour un index, ils réunissent des termes comme « dot », « accroît » et « contrat de mariage », tandis que l'ontologie les distingue. Dès lors la création de l'ontologie RCnum les oblige à revoir et préciser le sens des termes utilisés dans les documents d'archives. Prenons trois exemples pour illustrer cette nécessité :

— « religeux » : l'adjectif concerne la religion en général (ex. : les objets religieux). Cet adjectif, dans une ontologie, ne désigne pas un concept mais il pourra conduire à la définition de sous-concepts : *Objet Religieux* $\sqsubseteq$ *Objet*, *Cérémonie Religieuse* $\sqsubseteq$ *Cérémonie*. Le substantif, lui, désigne spécifiquement un moine ou un membre d'un ordre régulier. Ainsi, si un franciscain est un « religieux », un prêtre catholique ou un évêque anglican ne le sont pas. Ce sont des « ecclésiastiques ».

— « faber » : mot latin, que l'on retrouve dans les RC sous la forme « fabri », généralement traduit par « forgeron » ou « maréchal ». Dans les faits, le sens est plus large. Il s'agit « d'un artisan ou un ouvrier travaillant le métal », appelé « fèvre » en français. Dans l'ontologie, « fèvre » sera donc une classe avec pour sous-classes « forgeron », « maréchal-ferrant », « ferratier », « taillandier », etc.

— « avocat » : ce mot désigne soit l'avocat en tant que défenseur, soit un jurisconsulte. Les deux sens étant clairement distincts, l'ontologie contiendra deux concepts, soit Avocat (Défenseur) et Avocat (Jurisconsulte), associés aux mêmes graphies.

La création de l'ontologie et la précision sémantique qu'elle implique, enrichissent la compréhension du domaine d'expertise. D'autre part, l'unicité de la définition d'un concept est confrontée à la variabilité et à l'ambiguïté de la langue (« fèvre » peut désigner un artisan ou un simple ouvrier travaillant le fer). Cependant, les documents sources ne permettent pas toujours de distinguer les nuances, et les dictionnaires et glossaires ne fournissent pas non plus de réponses claires. Dès lors, une classe peut regrouper des sens difficiles à distinguer, laissant à l'utilisateur le soin de trouver, voire de proposer des solutions de désambiguïsation.
Notons encore que l'extraction automatisée, dans l'intégralité du corpus des RC (1537-1544 : collationnés et 1545-1550 : non collationnés), des mots uniques a permis de mettre en évidence de nouveaux concepts jusque-là écartés ou volontairement omis, voire ignorés par les expert-e-s dans le cadre de leur indexation. C'est le cas, par exemple, de « manipule » (ornement ecclésiastique), « blot » (pincement) ou de « cascavel » (grelot). La création de l'ontologie permet donc d'enrichir la connaissance du contenu des RC par rapport aux éditions traditionnelles papier.
En conclusion, les index des RC répertorient déjà la majeure partie des concepts et des individus, mais une vérification du sens de chaque mot imposée par l'ontologie permet d'en préciser, voire d'en corriger le contenu. En s'appuyant sur les index des RC (1536-1544) existants pour les classes et les individus, deux approches distinctes sont adoptées :

traitement des individus : un index manuel détaillé (patronymes, filiations, métiers) sert de base au marquage automatisé des données 1545-1550, combinant expertise historique et extraction informatique.

définition des classes : une ontologie est crée manuellement par l'historien-ne en partant des termes contextuels des textes (approche *sémasiologique*) [2], utilisant dictionnaires et thésaurus pour affronter polysémie et variations graphiques. Cette méthode, bien que chronophage, garantit une précision supérieure aux traitements automatiques dans ce contexte archivistique complexe.

Cette dualité méthodologique permet de concilier rigueur conceptuelle (définitions ancrées dans l'usage historique des mots) et efficacité technique (réutilisation des index validés), tout en s'inscrivant dans les pratiques modernes d'interopérabilité des ontologies.

De l'usage des dictionnaires et autres glossaires.

Comme on l'a vu, la langue des RC est ancienne et mal définie. Certains termes utilisés par les scribes sont ambigus. Pour établir une méthodologie efficace, nous avons d'abord travaillé sur une partie de l'ontologie : les éléments physiques comme les « personnes », les « animaux » et les « objets ». L'ontologie fournit pour chaque classe une définition unique, mais pas univoque, basée sur plusieurs dictionnaires et glossaires, à savoir Ahokas, Covelle, DMF, FEW, Godefroy, GPSR et TLFi. Divers ouvrages sont utilisés en complément, de manière ponctuelle, comme le *Dictionnaire historique de la Suisse*, le *Littré*, ou les éditions du *Dictionnaire* de l'Académie française. Un constat cependant s'impose : tous les dictionnaires et glossaires diffèrent par leur structure et les mots qu'ils traitent, chacun apportant une réponse spécifique. Or, l'ontologie exige une définition unique par classe. C'est pourquoi il a été décidé de proposer une pluralité de définitions similaires mais légèrement discordantes, évitant ainsi de figer un sens qui pourrait ne pas correspondre à l'intention du scribe. Cette diversité permet une vision globale des mots et de leur évolution à travers le temps. Chaque définition est référencée par un lien web lorsque l'ouvrage est accessible en ligne, permettant de vérifier ou d'explorer d'autres définitions proposées par ses auteur(e)s.
D'un point de vue technique, la définition d'un concept de l'ontologie RCnum est la disjonction de définitions trou-

vées dans les dictionnaires. Une entité appartient à l'extension du concept si elle satisfait à au moins une des définitions sélectionnées.

Graphies variables et multiples, un challenge pour la désambiguïsation

Il est apparu utile pour le traitement informatique et la désambiguïsation, tout autant que pour l'aide à la précision de la définition, d'ajouter d'autres annotations à l'ontologie, au nombre de trois :

— graphies multiples : l'ontologie des RC intègre les variantes graphiques identifiées en deux étapes. Tout d'abord, une recherche manuelle a été effectuée dans les RC (1536-1544), en utilisant les graphies mentionnées dans l'index et l'expérience de l'historien-ne. Le résultat montre que « syndic » apparaît sous 17 graphies en français (dont « contre-indicque », « consyndicque » ou « santique »). De même, « châtelain » connaît 13 graphies en français. Dans un deuxième temps, une extraction systématique a été réalisée. Tous les mots uniques des RC (1536-1544) ont été extraits, classés alphabétiquement, puis intégrés ou rejetés méthodiquement dans l'ontologie initiale. Ce processus permet de consolider la structure et la richesse sémantique de l'ontologie, préparant ainsi le terrain pour le traitement automatique des RC (1545-1550) et potentiellement pour ceux jusqu'à 1792. Cette approche est donc particulièrement utile compte tenu de l'absence d'orthographe officielle figée à l'époque.

— exemple pour chaque graphie : pour clarifier le choix du sens par l'expert-e qui crée l'ontologie, un extrait des RC comprenant au moins une occurrence du terme est ajouté pour chaque graphie, ceci dans deux buts : aider à la désambiguïsation et permettre à l'usager d'accéder directement au mot dans son contexte pour confirmer ou infirmer le choix de définition de l'expert-e. Pour des raisons de vérification et d'utilisation, ces extraits ou exemples sont dûment référencés.

— verbes et adjectifs : bien souvent le substantif n'est pas directement utilisé pour désigner un concept. Le scribe lui préfère une tournure verbale. Par exemple, « l'amodiation du Pré-l'Évêque a été attribuée à Pierre Dupond », est transcrite par « le Pré-l'Évêque a été admodié à Pierre Dupond ». Dès lors, à l'image des variantes graphiques, des annotations spécifiques aux verbes et aux adjectifs sont créées, toujours dans le but de faciliter la désambiguïsation et de permettre une meilleure réponse aux recherches des usagers effectuées sur la plateforme web de consultation.

3 Implémentation de l'ontologie

3.1 Aspects spatio-temporels

Les langages ontologiques formels communément utilisés, en particulier RDFS et OWL permettent d'annoter les axiomes et par conséquent d'y ajouter leur contexte de validité. Cependant ces annotations ne sont pas prises en compte par les raisonneurs usuels. Par conséquent, il n'est pas possible de créer deux axiomes annotés

$$\begin{array}{ll} \mathsf{Danse} \sqsubseteq \mathsf{Infraction} & (\mathrm{XVI^e}, \mathsf{Genève}) \\ \mathsf{Danse}\ \mathbf{disjoint}\ \mathsf{Infraction} & (\mathrm{XXI^e}, \mathsf{Europe}) \end{array}$$

sans engendrer une incohérence dans l'ontologie. Bien que des propositions d'extension contextuelle des logiques de descriptions existent [3][1], elles ne font actuellement l'objet d'aucune standardisation ni d'un outillage de raisonnement facilement utilisable. Par contre, dans les modèles de graphes de connaissances tels Wikibase [1], Neo4J [2] ou RDF* [3] les annotations (représentées de diverses manières : attributs, qualificateurs, triplets ayant un triplet comme sujet,...) font partie intégrante du modèle. Mais ces modèles ne sont pas associés à une logique formelle permettant de définir les notions de consistance et d'inférence logique. Par conséquent nous avons décidé d'appliquer le processus suivant :

— créer une ontologie OWL pour chaque contexte spatio-temporel

— interconnecter ces ontologies pour former un réseau d'ontologies

— utiliser un raisonneur OWL (standard) pour tester la cohérence du réseau

— injecter chaque ontologie dans le GdC en annotant ses axiomes avec le domaine de validité correspondant

3.2 Réseau d'ontologie

Pour l'instant le réseau d'ontologies est composé de :

— l'ontologie **rc-xvi-ge** qui comprend 282 classes dont chacune est annotée avec en moyenne 5 références aux définitions des dictionnaires historique ; 3,5 exemples provenant des registres 1536-1544 ; 3,6 graphies trouvées dans ces mêmes registres et 2,7 verbes et adjectifs pouvant désigner des objets ou propriétés de cette classe. Les axiomes de l'ontologie servent essentiellement à définir la hiérarchie des concepts.

— l'ontologie rc-xxi qui contient les graphies actuelles des concepts des registres et les axiomes de hiérarchisation.

— l'ontologie **rc-univ** dont les axiomes sont considérés comme universellement valides. Par exemple :

$$\begin{array}{l} \mathsf{Personne} \sqsubseteq \mathsf{mère}\ \mathbf{only}\ \mathsf{Personne} \\ \mathsf{Condamnation} \sqsubseteq \mathsf{peine}\ \mathbf{some}\ \mathsf{Peine} \\ \qquad \mathbf{and}\ \mathsf{coupable}\ \mathbf{some}\ \mathsf{Acteur} \\ \mathsf{Registre} \sqsubseteq \mathsf{Document} \end{array}$$

Les concepts de haut-niveau de cette ontologie (Action, Acteur, Lieu, Document,...) sont alignés avec ceux

1. https://www.mediawiki.org/wiki/Wikibase/DataModel
2. https://neo4j.com/docs/getting-started/appendix/graphdb-concepts/
3. https://www.ontotext.com/knowledgehub/fundamentals/what-is-rdf-star/

de CIDOC-CRM (E7_Activity, E39_Actor, E53_Place, E31_Document,...)
L'interconnexion des ontologies est assurée par la relation d'importation entre ontologies et par le choix d'un même URI pour les concepts équivalents. Il aurait été possible de définir des concepts différents dans chaque ontologie et de les lier avec le prédicat owl:SameAs. Cependant cette solution conduit à une multiplication des URI désignant le même concept. En effet, de très nombreux concepts restent en usage au cours du temps. Par contre les annotations lexicales et les axiomes qui s'appliquent à ceux-ci peuvent être spécifiques à un contexte. On aura par exemple :

dans **rc-xvi-ge**

Danse graphie "dances" (axiome d'annotation)
Danse $\sqsubseteq$ Infraction

dans **rc-xxi-ge**

Danse graphie "danse" (axiome d'annotation)
Danse **disjoint** Infraction

dans **rc-univ**

Danse $\sqsubseteq$ Action

Pris tous ensemble ces axiomes sont incohérents. Pour un tel réseau d'ontologies, il faut définir une notion de cohérence plus faible que la cohérence globale. Nous adoptons la définition suivante qui s'applique à un réseau d'ontologies tel que pour toute paire d'ontologies, soit leurs contextes sont disjoints soit ils sont inclus l'un dans l'autre :

Definition. Un réseau d'ontologies contextuelles est *cohérent* si pour chaque ontologie O_i du réseau, l'ontologie $O_i \cup O'_1 \cup ... \cup O'_n$, où les O'_j sont les ontologies telles que $\text{contexte}(O_i) \subseteq \text{contexte}(O'_j)$, est cohérente.

Dans l'exemple précédent, cette condition est satisfaite car **rc-xvi-ge** $\cup$ **rc-univ** et **rc-xxi-ge** $\cup$ **rc-univ** sont toutes deux cohérentes.

3.3 Injection dans le GdC

Le GdC « RC » est un graphe RDF* (RDF 1.2 à l'avenir)) [4] dans lequel un triplet peut être sujet ou objet d'un autre triplet. Pour représenter un triplet muni d'un contexte, mais également d'une provenance et d'un niveau de confiance, nous utilisons la structure RDF* [5] :

$<<$ *Triplet* $>>$ `:annotation` [
`:context` *Contexte*;
`:prov` *Provenance*;
`:conf` *Confiance*]

Les axiomes de l'exemple précédents seront donc représentés par les triplets :

```
<< :Danse} :graphie "dances" >>
 :annotation [:context :ge_xvi ].
<< :Danse rdfs:subClassOf :Infraction >>
  :annotation [:context :ge_xvi].
<< :Danse :graphie "danse" >>
  :annotation [:context :ge_xxi ].
<< :Danse owl:disjointWith :Infraction>>
 :annotation [:context :ge_xxi ].
<< :Danse rdfs:subClassOf :Action >>
  :annotation [:context :univ].
:ge_xvi :time :XVIe ; :space :Genève.
:univ :time :all ; :space :Europe
```

3.4 Indexation sémantique

À partir des index sont construits des GdC contenant des informations sur les personnes et les lieux, représentées sous forme d'entités. Ces GdC sont utilisés avec les ontologies pour trouver et identifier les entités dans le texte des documents, soit les RC. Cette procédure s'appelle le traitement des Entités Nommées, qui se compose de plusieurs étapes [4] :

- Entity Recognition : reconnaître et délimiter les occurrences des entités dans le texte.
- Entity Classification : déterminer le type de l'entité trouvée.
- Entity Discrimination, ou Linking : spécifier à quelle entité du GdC le texte fait référence, et désambiguïser parmi plusieurs entités candidates.
- Relation Extraction : découvrir de nouvelles relations entre les entités.

Il existe deux approches principales pour ces tâches : l'approche basée sur des règles et l'approche basée sur des données (avec l'IA). Nous avons appliqué la première approche pour reconnaître et désambiguïser les entités dans le texte, en recherchant des correspondances dans l'index et les ontologies, et en utilisant des expressions régulières pour prendre en compte les variations orthographiques. Cependant, des difficultés sont apparues dans les deux tâches en raison des caractéristiques linguistiques des RC :

- Entity Recognition : difficulté à identifier toutes les orthographes multiples d'une même entité.
- Entity Discrimination : difficulté à désambiguïser personnes et lieux, pour les cas de patronymes issus de noms de lieux (ou inversement), ainsi que les personnes ayant le même nom/prénoms ou des noms/prénoms similaires.

En suivant les approches de CLEF-HIPE-2020 [5], ces défis pourront être relevés en utilisant des méthodes d'IA comme les Conditional Random Fields, la Bidirectional LSTM, et les Pretrained Language Models, ou en intégrant l'index en tant que Gazetteer et les connaissances d'expert-e-s pour la validation du système.
L'établissement de liens entre les entités contenues dans les documents et le GdC améliorera grandement la navigation

4. Il aurait été possible de représenter les contextes des axiomes par des graphes nommés, mais l'ajout d'autres annotations nécessaires (confiance, provenance) conduirait à une explosion combinatoire du nombre de graphes
5. Par souci de concision, nous omettons les définitions de préfixes.

dans les documents et permettra d'effectuer des tâches supplémentaires telles que la recherche avancée ou la génération augmentée de récupération basée sur les graphes (GraphRAG), en exploitant les entités liées.

3.5 Génération Augmentée de Récupération

La RAG [6] est un moyen d'améliorer la réponse d'un modèle de langage en fournissant des sources pertinentes comme input. Elle peut être implémentée dans ce contexte pour répondre à des questions liées aux RC. Classiquement, cette méthode consiste à repérer des morceaux de documents similaires à la requête et à les ajouter à la requête elle-même, mais elle présente des difficultés pour les requêtes qui nécessitent un contexte d'informations pertinentes et un raisonnement multi-sauts [7]. Pour surmonter ces limitations, la GraphRAG est un type de RAG qui utilise des documents organisés sous forme de graphe, comme dans notre cas, pour exploiter les relations précalculées et trouver des informations plus abstraites [8]. L'idée est ici d'utiliser les GdC mis à disposition et les entités liées pour construire une GraphRAG et interroger les documents en langage naturel (en la comparant à une RAG classique pour évaluer l'impact des relations). Ce système permettra aux professionnels et au grand public d'interroger les documents et d'extraire des informations variées, mais pertinentes et ce, malgré les difficultés imposées par les RC.

4 Conclusion

Nous avons décrit la démarche mise en place pour concevoir et implémenter une ontologie contextuelle adaptée aux transcriptions de textes de décisions administratives. Nous avons illustré les particularités des textes à différents niveaux et expliqué comment nous avons pris ces particularités en compte dans l'ontologie. Dans cette démarche, l'interaction historien-ne-s/informaticien-ne-s permet de définir des spécifications pour la représentation des connaissances qui prennent en compte les spécificités des documents et des modalités de leur rédaction (en particulier le cadre linguistique et terminologique) tout en répondant aux critères historiques et techniques. Cette ontologie, constituée à partir de transcriptions de textes déjà collationnés et indexés manuellement pour les années précédentes, permet d'initialiser un graphe de connaissance composé de concepts d'intérêt historique, de personnes et de lieux. L'ensemble fournit un cadre pour indexer sémantiquement les textes de la période concernée par le projet avec des approches récentes d'IA et de développer des interfaces d'interrogations en langage naturel avec la RAG pour permettre d'interroger le contenu des textes. Le projet RCnum est *open source* et *open access*, et l'ontologie sera mise librement à diposition pour d'autres projets similaires concernant des documents d'archives administratives.

Remerciements

Ces travaux sont développés dans le cadre du projet "Une édition sémantique et multilingue en ligne des registres du Conseil de Genève (1545-1550)" [7] soutenu par le FNS (Fonds National Suisse, grant 215733).

Références

[1] Sahar ALJALBOUT, Didier BUCHS et Gilles FALQUET : OWLˆ C : A Contextual Two-Dimensional Web Ontology Language. *In 2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[2] Charles BODON et Jean CHARLET : Les ontologies informatiques au service de la communication interdisciplinaire : l'interopérabilité sémantique. *Revue Intelligibilité du Numérique*, (1|2020), 2020.

[3] Loris BOZZATO et Luciano SERAFINI : Materialization calculus for contexts in the Semantic Web. *CEUR Workshop Proceedings*, 1014:552–572, 2013.

[4] Maud EHRMANN, Ahmed HAMDI, Elvys Linhares PONTES, Matteo ROMANELLO et Antoine DOUCET : Named entity recognition and classification in historical documents : A survey. *ACM Comput. Surv.*, 56(2), septembre 2023.

[5] Maud EHRMANN, Matteo ROMANELLO, Alex FLUCKIGER et Simon CLEMATIDE : Extended Overview of CLEF HIPE 2020 : Named Entity Processing on Historical Newspapers.

[6] Anna Sofia LIPPOLIS, Miguel CERIANI, Sara ZUPPIROLI et Andrea Giovanni NUZZOLESE : Ontogenia : Ontology Generation with Metacognitive Prompting in Large Language Models. *In* Meroño Peñuela et AL., éditeur : *The Semantic Web : ESWC 2024 Satellite Events*, pages 259–265, Cham, 2025. Springer Nature Switzerland.

[7] Balazs MOSOLYGO, Bahareh FATEMI, Fazle RABBI et Andreas OPDAHL : Evaluating graphrag's role in improving contextual understanding of news in newsrooms. *Norsk IKT-konferanse for forskning og utdanning*, 2024.

[8] Boci PENG, Yun ZHU, Yongchao LIU, Xiaohe BO, Haizhou SHI, Chuntao HONG, Yan ZHANG et Siliang TANG : Graph Retrieval-Augmented Generation : A Survey, septembre 2024.

[9] Mohammad Javad SAEEDIZADE et Eva BLOMQVIST : Navigating Ontology Development with Large Language Models. *In* Meroño Peñuela et AL., éditeur : *The Semantic Web*, pages 143–161, Cham, 2024. Springer Nature Switzerland.

6. Nous utilisons l'acronyme anglais, plus commun dans la littérature.

7. https ://www.unige.ch/registresconseilge/