Harold : un système de requête itératif et interactif pour l'exploration de corpus de patrimoine culturel

Prunelle D. Treuil[1], Olivier Bruneau[2], Jean Lieber[1] , Emmanuel Nauer[1], Laurent Rollet[2]

[1] LORIA, Université de Lorraine, France
[2] AHP-PReST, Université de Lorraine, France

prunelle.daudre–treuil@loria.fr, olivier.bruneau@univ-lorraine.fr, jean.lieber@loria.fr, emmanuel.nauer@loria.fr, laurent.rollet@univ-lorraine.fr

Résumé

Cet article présente Harold, un système conversationnel permettant l'accès à des corpus de patrimoine culturel sans la nécessité de connaître de langage informatique comme SPARQL. Ce système organise les documents du corpus en utilisant l'analyse formelle de concepts (AFC) afin de présenter les résultats d'une recherche de façon synthétique. L'utilisateur peut interagir avec cette hiérarchie de concepts afin de sélectionner les documents intéressants ou retirer ceux qui ne le sont pas. De plus, un système de gestion d'ontologie est intégré à Harold pour structurer les concepts liés aux questions de recherche de ses utilisateurs.

Mots-clés

Gestion d'ontologie, Système conversationnel, Analyse formelle de concepts, Collections patrimoniales culturelles

Abstract

This paper presents Harold, a conversational system that enables access to cultural heritage corpora without requiring knowledge of computer languages such as SPARQL. The system organizes corpus documents using Formal Concept Analysis (FCA) to provide a synthetic view of search results. Users can interact with the resulting concept hierarchy to select relevant documents or discard irrelevant ones. Additionally, Harold integrates an ontology management component to structure concepts related to users' research questions.

Keywords

Ontology management, Conversational system, Formal Concept Analysis, Cultural Heritage Collection

1 Introduction

Cet article présente *Harold*, un système conçu pour explorer des corpus textuels, comme la correspondance d'Henri Poincaré. Développé avec les Archives Henri Poincaré, un laboratoire d'histoire et de philosophie des sciences [1], ce système permet une exploration interactive des plus de 2 000 lettres échangées par Henri Poincaré, numérisées, transcrites et annotées. Leur étude offre un aperçu unique de ses activités scientifiques. Ce corpus est accessible par un point d'accès SPARQL permettant des requêtes avancées mais dont l'usage reste complexe pour les non-spécialistes, justifiant le recours à une interface accessible comme Harold.

Pour simplifier l'accès au corpus sans recourir à des requêtes SPARQL, Harold propose une interface simple fondée sur trois principes. (1) Les questions de recherche sont souvent vagues et nécessitent une exploration itérative ; Harold utilise ainsi l'analyse formelle de concepts (AFC) pour grouper les lettres selon des propriétés partagées, permettant à l'utilisateur de découvrir des liens entre concepts dans les lettres et de s'en servir pour affiner progressivement sa recherche. (2) L'exploitation du contenu des lettres est essentielle et doit aller au-delà de l'exploitation de simples métadonnées : des techniques de traitement automatique du langage (TAL) sont employées pour extraire des termes (mots, groupes nominaux, entités) qui sont considérés comme de nouvelles propriétés sémantiques. (3) L'expertise du chercheur est valorisée via une ontologie construite à partir des termes extraits, permettant à l'historien d'organiser sa connaissance. L'interaction entre corpus, ontologie et utilisateurs facilite la réification de nouvelles connaissances.

La section 2 présente les travaux relatifs à ce travail. La section 3 introduit les préliminaires. La section 4 décrit le système Harold. La section 5 présente l'outil de gestion ontologique intégré.

2 Travaux proches

De nombreux travaux portent sur l'interrogation de bases de connaissances, notamment représentées en RDF ou accessibles via un point d'accès SPARQL. Une classification des approches, allant des plus informelles (langage naturel) aux plus formelles (SPARQL), est proposée dans Kaufmann and Bernstein [9]. Diefenbach et al. [5] présentent une étude sur les systèmes de questions-réponses. Ces systèmes doivent transformer des requêtes en langage naturel en requêtes SPARQL, souvent à l'aide de techniques de TAL (reconnaissance d'entités nommées, analyse morpho-syntaxique, dépendances syntaxiques). Par exemple, le système SWIP [15] relie d'abord différents éléments de la re-

1. https://poincare.univ-lorraine.fr/

quête à des ressources et classes du graphe de connaissance interrogé, puis utilise des modèles de requêtes pour générer une requête SPARQL. Des travaux plus récents utilisent des techniques d'apprentissage profond appliquées aux systèmes de questions-réponses, notamment par l'utilisation de grands modèles de langues (LLM) pour obtenir des réponses directement en langage naturel. Biancofore et al. [2] présentent un état-de-l'art sur les systèmes de questions-réponses interactifs et Zaib et al. [18], une étude sur les systèmes questions-réponses conversationnels. Enfin, Lan et al. [10] présentent plusieurs techniques de questions-réponses appliquées aux bases de connaissances avec des requêtes en langage naturel. D'autres systèmes vont venir assister l'utilisateur via une interface. Dans Sparklis [8], une série de formulaires permet de construire progressivement les requêtes, cette requête étant traduite en langage naturel pour faciliter la compréhension de l'utilisateur. D'autres systèmes utilisent des langages visuels permettant de représenter les requêtes, comme Nitelight [16] qui les représente sous forme de graphes. Enfin, des interfaces spécifiques ont été développées, en particulier, pour interroger la correspondance de Henri Poincaré, incluant recherche par similarité [11] ou recherche approximative par transformation de requêtes SPARQL [3].

Un autre aspect abordé dans ce travail concerne l'accès synthétique aux documents, et la nécessité d'une interaction avec le système pour répondre à un problème de recherche en plusieurs étapes. Ce travail s'inspire du système CRECHAINDO [14], un système itératif et interactif de recherche d'information qui organise les résultats de recherche GOOGLE à l'aide de l'AFC, permettant d'organiser les réponses selon leurs propriétés dans une hiérarchie (voir section 3.2). Cette hiérarchie facilite l'exploration des résultats en fournissant un accès structuré, et l'utilisateur peut interagir avec celle-ci pour signaler les concepts pertinents ou non, ce qui modifie dynamiquement la hiérarchie.

Enfin, le processus de construction automatique d'ontologies s'inspire de la littérature existante (voir [1] pour un état de l'art). Dans Harold, la construction d'ontologies repose sur des potentiels concepts issus de techniques de traitement automatique des langues (TAL) et sur l'usage de l'AFC pour faire émerger les termes fréquents. Il couvre ainsi plusieurs niveaux du *Ontology Learning Layer Cake* (le gâteau à étage de l'apprentissage d'ontologie) proposé dans [4] : identification des termes, des synonymes (non détaillé ici), des concepts, et de leur organisation en hiérarchie. Le processus est également guidé par l'approche interactive pour la construction d'une ontologie orientée par l'expert du domaine, décrite dans [6], bien que notre objectif ne soit pas ici de formaliser des définitions de concepts.

3 Préliminaires

3.1 Le corpus et l'ontologie d'Henri Poincaré

Le corpus de la correspondance de Henri Poincaré comprend plus de 2000 lettres, envoyées ou reçues par Henri Poincaré entre 1873 et 1912 avec plus de 300 correspondants en cinq langues, principalement en français. Ces documents ont été numérisés, transcrits et annotés sémantiquement avec des métadonnées générales et des notes contextuelles, constituant l'*appareil critique* du corpus, fruit de plus de 30 ans de travaux historiques.

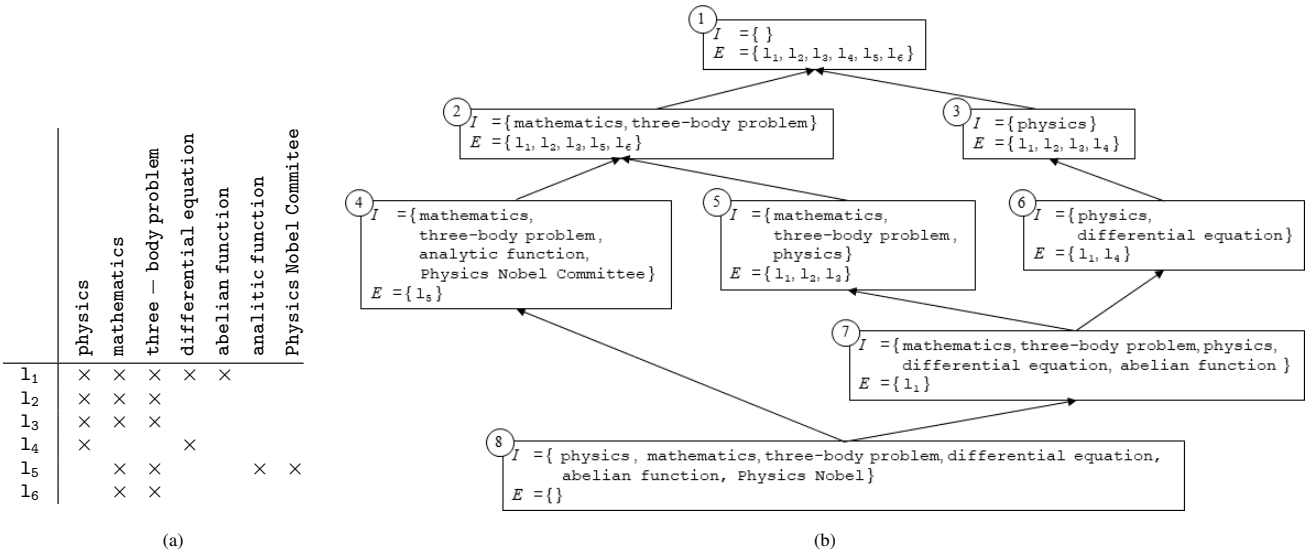
La correspondance de Henri Poincaré est accessible via des volumes publiés , p. ex. [13], un site web [2] reposant sur Omeka S [12], et un point d'accès SPARQL (accès restreint à un usage interne). Le site propose une recherche par mot-clé simple, tandis que le point d'accès SPARQL permet des requêtes précises mais nécessite de maîtriser SPARQL ainsi que le vocabulaire utilisé dans l'annotation en RDF pour les propriétés et les concepts. L'ontologie Henri Poincaré (`ahpo`), développée par les Archives, structure les métadonnées liées à la correspondance et à d'autres documents (articles, livres, thèses, etc.). Chaque lettre est décrite par des annotations sémantiques sous forme de triplets RDF, utilisant les propriétés de l'ontologie Henri Poincaré, telles que `ahpo:sentBy` (expéditeur), `ahpo:sentTo` (destinataire), `ahpo:language` (langue d'écriture), `ahpo:writingDate` (date de rédaction) et `ahpo:writtenAt` (lieu d'écriture).

3.2 L'analyse formelle de concepts

L'AFC est une approche de l'analyse de données fondée sur la théorie des treillis. Un *contexte formel* est un triplet $\mathcal{K} = (O, P, R)$, où O est un ensemble d'*objets*, P un ensemble de *propriétées* et R une relation sur $O \times P$ indiquant qu'un objet est décrit par une propriété [7]. Un exemple de contexte formel est donné en figure 1 (a), ce contexte représente 6 lettres ($\mathtt{l}_1$,...,$\mathtt{l}_6$) décrites par 6 termes et un destinataire (le comité du Nobel de Physique). Un *concept formel* est un couple (I, E), où E est un ensemble maximal d'objets (appelé *extension*) et I un ensemble maximal de propriétés (appelé *intension*) partagées par cette extension. Par exemple, $(\{\mathtt{physics}, \mathtt{differential\ equation}\}, \{\mathtt{l}_1, \mathtt{l}_4\})$ constitue un concept formel : il n'y a pas d'autres objets qui possèdent l'ensemble des propriétés de l'intension, et il n'y a pas d'autres propriétés communes aux objets de l'extension.

L'ensemble $\mathcal{C}_\mathcal{K}$ de tous les concepts formels du contexte $\mathcal{K} = (O, P, R)$ est partiellement ordonné par inclusion d'extensions, également appelée *spécialisation* (notée $\leq_\mathcal{K}$) entre concepts. $\mathcal{L} = \langle \mathcal{C}_\mathcal{K}, \leq_\mathcal{K} \rangle$ forme un treillis complet, appelé *treillis de concepts*. Ce treillis $\mathcal{L}$ peut être représenté sous la forme d'un diagramme de Hasse. La figure 1 (b) présente le treillis de concepts issu du contexte formel de la figure 1 (a). Le concept supérieur contient toutes les lettres ; son intension est vide car aucune propriété n'est partagée par l'ensemble des lettres. À l'inverse, le concept inférieur est défini par l'ensemble de toutes les propriétés ; son extension est vide car aucune lettre n'est décrite par toutes les propriétés. Plusieurs algorithmes ont été proposés pour la construction des treillis de concepts [7]. Dans le cadre de Harold, CORON [3] est utilisée. CORON est une plateforme logicielle qui implémente plusieurs méthodes pour l'ex-

2. https://henripoincare.fr/
3. http://coron.loria.fr



(a) (b)

FIGURE 1 – Un context formel (a) et le treillis de concepts associés (b)

traction symbolique de connaissances, y compris des programmes de construction de treillis de concepts [17].

4 Le système Harold

Le système Harold aide les historiens à explorer la correspondance de Henri Poincaré en récupérant et analysant les lettres. Une étape de TAL identifie les termes dans les textes complets pour exploiter à la fois le contenu des lettres et leur appareil critique. La section 4.1 décrit l'annotation et la section 4.2 présente, sur un exemple concret de recherche historique, les fonctionnalités et les possibilités de Harold.

4.1 Exploitation du plein texte par les outils de TAL

Les historiens analysant la correspondance de Henri Poincaré doivent examiner à la fois le texte complet et l'appareil critique. Dans ce but, des termes en ont été extraits à la fois grâce à une liste préexistante et à une extraction automatique de groupes nominaux. En effet, les termes scientifiques sont souvent identifiés par des groupes nominaux, tandis que les termes issus de l'environnement du mathématicien, comme le jargon polytechnicien, se limitent souvent à un mot. Cette liste, validée par les historiens, est cruciale, car de nombreux termes ont des significations contextuelles. Une liste des œuvres citées dans les lettres est également maintenue.

4.2 Exemple de cas d'usage

Prenons l'exemple d'un historien souhaitant étudier l'impact du travail de Henri Poincaré sur la physique. Cet historien utilise Harold pour rechercher des lettres échangées avec des physiciens ou contenant des mots-clés en physique mathématique (*mathématiques*, *physique*, *équation différentielle*). Plusieurs recherches et analyses sont nécessaires, avec des affinements en fonction des résultats avec de nouveaux termes ou en excluant des résultats. L'organisation des concepts dans une ontologie permet d'affiner le processus de recherche.

La section 4.3 présente le formulaire de recherche pour explorer le corpus. La section 4.4 explique l'affichage des résultats. La section 4.5 illustre les types d'interactions avec Harold.

4.3 Rechercher les lettres par propriétés

Harold permet de rechercher dans la correspondance de Henri Poincaré des lettres avec des propriétés spécifiques via un formulaire (`Sender`, `Recipient`, `Containing`, etc.). Ce formulaire permet également de choisir les propriétés à considérer par l'AFC. Le formulaire est transformé en requête SPARQL pour trouver les lettres correspondantes. Trouver l'ensemble des lettres répondant à une problématique dès l'exécution de la première requête étant rare, la recherche est itérative, faisant évoluer l'ensemble des lettres pertinentes pour une recherche à chaque nouvelle requête. Le formulaire peut être utilisé plusieurs fois pour retrouver de nouvelles lettres ou de nouveaux critères.

4.4 Visualiser les résultats

Au lieu de juste énumérer les lettres récupérées, Harold applique l'AFC pour les regrouper selon les propriétés qu'elles partagent, créant une vue hiérarchique soulignant leurs caractéristiques communes. Les résultats sont affichés sous la forme d'une hiérarchie. Chaque ligne dans la figure 2 représente un concept formel issu de la recherche de toutes les lettres contenant le terme *physique mathématique*. L'indentation indique que les propriétés de niveaux supérieurs font partie de l'extension du concept. Par exemple, les 18 lettres envoyées au *Physics Nobel Committee* contiennent aussi *physique mathématique*. D'autres termes intéressants pour la recherche apparaissent également : *dérivées partielles de la physique mathématique*, *propagation de la chaleur* et *problème de Dirichlet*.

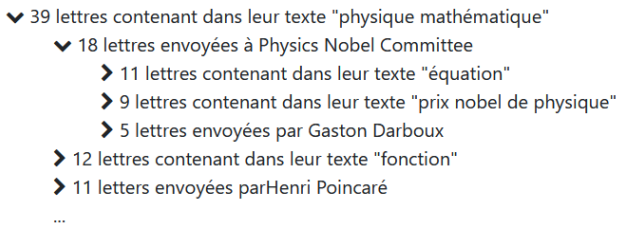


FIGURE 2 – Hiérarchie des résultats après avoir commencé une recherche avec le terme *physique mathématique*.

4.5 Intéragir avec les résultats

Pour trouver toutes les lettres liées au sujet *Mathématiques pour la physique*, une recherche uniquement sur *physique mathématique* est insuffisante, car de nombreuses lettres ne contiennent pas cette expression exacte. Les fonctionnalités interactives de Harold permettent à l'utilisateur d'affiner les recherches, d'ajouter des lettres pertinentes ou d'exclure celles qui sont non pertinentes, améliorant ainsi les résultats de l'analyse formelle des concepts. Plusieurs actions sont disponibles via un menu contextuel pour chaque groupe de lettres ; trois pour l'analyse des lettres, une pour la gestion de l'ontologie (voir section 5), et une permettant d'accéder aux lettres elles-mêmes. Par exemple, pour le concept *envoyé au Physics Nobel Committee*, ces actions seront :

- *Ajouter les lettres envoyées au Physics Nobel Committee* : provoque une nouvelle interaction avec le SPARQL endpoint pour ajouter les lettres répondant à ce critére à l'ensemble des lettres analysées.
- *Ne pas utiliser "envoyé au Physics Nobel Committee" dans le résultat* : indique à l'AFC de ne pas utiliser cette propriété dans la construction de la hiérarchie ce qui permet de simplifier les résultats.
- *Supprimer les lettres envoyées au Physics Nobel Committee* : supprime du context formel les lettres ayant cette propriété, ce qui permet d'éliminer des lettres jugées non pertinentes par l'historien pour son problème de recherche.
- *Ajouter le concept "Physics Nobel Committee" à l'ontologie* : aide l'utilisateur à acquérir des connaissances sur le contenu de la lettre. Le terme *Physics Nobel Committee* est ajouté à l'ensemble des concepts non étiquetés de l'interface de gestion de l'ontologie.

5 Gestion de l'ontologie

L'interface de gestion de l'ontologie comprend actuellement des concepts atomiques et leurs relations hiérarchiques. Elle est séparée en deux parties : un *ensemble de concepts non classés*, contenant les concepts candidats (ajoutés manuellement ou via la fonctionnalité *Ajouter à l'ontologie*) et un *ensemble de concepts classés*, structuré selon une relation spécifique/générique. Cela signifie que si *fonction* est l'enfant de *mathématiques* dans la hiérarchie, cela doit être interprété comme « Une lettre annotée par *fonction* va être aussi annotée par *mathématiques*. » Formellement, en utilisant la notation des logiques de descriptions, la formule suivante sera ajoutée à l'ontologie :

$$\texttt{Letter} \sqcap \exists \texttt{isAnnotatedBy}.\{\texttt{"fonction"}\} \sqsubseteq \texttt{Letter} \sqcap \exists \texttt{isAnnotatedBy}.\{\texttt{"mathématiques"}\}$$

La structuration hiérarchique s'effectue simplement en déplaçant dans l'interface un terme sous un autre.

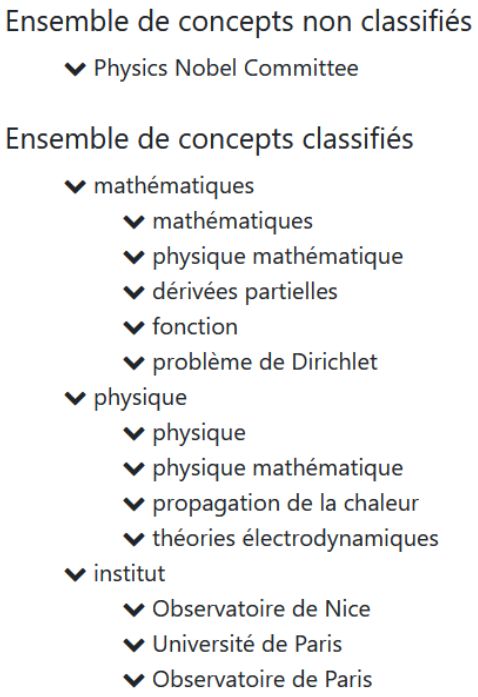


FIGURE 3 – L'interface de gestion de l'ontologie

L'intégration de l'ontologie dans l'analyse par AFC permet d'obtenir une structuration des résultats reflétant la hiérarchie. Ainsi, des lettres contenant *propagation de la chaleur* ou *théories électrodynamiques* sont regroupées sous le concept *physique*. Un groupe de 40 lettres associées à la fois à *mathématiques* et *physique* peut être identifié, même sans présence explicite de ces termes, mais par ajout de propriétés plus générales issues de l'ontologie dans le contexte formel pour étendre la description des lettres. Par ailleurs, lors des recherches, l'utilisation de l'ontologie assure que la saisie de *physique* inclut toutes les lettres liées aux concepts relevant de ce domaine.

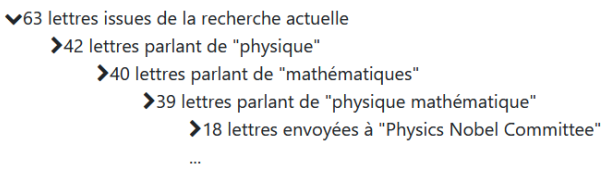


FIGURE 4 – Les résultats après la prise en compte de l'ontologie

6 Discussion et conclusion

Cet article a présenté Harold, un système itératif et interactif d'exploration de la correspondance de Henri Poincaré, ne nécessitant pas de connaissance préalable des technologies du web sémantique. Il permet l'analyse conjointe des métadonnées et du texte intégral, en s'appuyant sur l'AFC pour regrouper les lettres selon des propriétés partagées. L'ajout d'un système de gestion d'ontologie facilite la découverte de connaissances en améliorant l'organisation, par les historiens, des concepts relatifs à une question. Cette ontologie permet de généraliser les résultats synthétiques, faisant émerger des liens de plus hauts niveaux à partir de la présence dans les textes de concepts parfois peu représentés. Harold peut être appliqué à d'autres corpus. Une application en cours cherche à analyser les publications en informatique (web sémantique, ontologies, humanités numériques). D'autres perspectives incluent une évaluation comparative entre deux historiens, l'un utilisant Harold et l'autre non, afin de mesurer la couverture, la pertinence des lettres retrouvées et les connaissances ainsi acquises. Enfin, des travaux vont porter sur l'enrichissement de la gestion ontologique : intégration de relations variées entre concepts, meilleure prise en charge des propriétés (p. ex. *échangé avec* plus général que *envoyé à*), et définition précise de certains concepts (p. ex. années d'étude de Henri Poincaré entre 1873 et 1879).

Références

[1] Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood, and Hafiza Mahnoor Abbasi. A survey of ontology learning techniques and applications. *Database : The Journal of Biological Databases and Curation*, 2018, 2018.

[2] Giovanni Maria Biancofiore, Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Fedelucio Narducci. Interactive question answering systems : Literature review. *ACM Comput. Surv.*, 56(9) :239 :1–239 :38, 2024.

[3] Olivier Bruneau, Nicolas Lasolle, Jean Lieber, Emmanuel Nauer, Siyana Pavlova, and Laurent Rollet. Applying and developing semantic web technologies for exploiting a corpus in history of science : The case study of the Henri Poincaré correspondence. *Semantic Web*, 12(2) :359–378, 2021.

[4] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. *Ontology Learning from Text : An Overview*, volume 123, pages 3–12. IOS Press, Amsterdam, 01 2005.

[5] Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. Core techniques of question answering systems over knowledge bases : a survey. *Knowledge and Information Systems*, 55(3) :529–569, 2018.

[6] Mathieu D'aquin. TaBIIC : Taxonomy Building through Iterative and Interactive Clustering. In *Formal Ontology in Information Systems*, pages 155–168. IOS Press, 2023.

[7] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis : Mathematical Foundations*. Springer, Berlin, 1999.

[8] Eero Hyvönen and Sébastien Ferré. Sparklis : An expressive query builder for SPARQL endpoints with guidance in natural language. *Semantic Web*, 8(3) :405–418, 2016.

[9] Esther Kaufmann and Abraham Bernstein. Evaluating the usability of natural language query languages and interfaces to semantic web knowledge bases. *Journal of Web Semantics*, 8(4) :377–393, 2010.

[10] Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. Complex knowledge base question answering : A survey, 2022.

[11] Nicolas Lasolle. A Navigation Tool for Exploring Semantic Web Corpora. In *Proceedings of the ISWC 2021 Posters, Demos and Industry Tracks*, Virtual conference, France, October 2021.

[12] Jean-Marc Meunier, Samuel Szoniecky, and Daniel Berthereau. Utilisation d'Omeka-S pour la conception et le partage de ressources pédagogiques. In *Zotero & Omeka-des outils pour les humanités numériques*, Poitiers, France, 2019.

[13] Philipe Nabonnand. *La Correspondance entre Henri Poincaré et Gösta Mittag-Leffler*. Publication des Archives Henri Poincaré. Birkhäuser Basel, 1998.

[14] Emmanuel Nauer and Yannick Toussaint. CreChainDo : an iterative and interactive web information retrieval system based on lattices. *International Journal of General Systems*, 38 :363–378, 2009.

[15] Camille Pradel, Ollivier Haemmerlé, and Nathalie Hernandez. Swip : A natural language to SPARQL interface implemented with SPARQL. In Nathalie Hernandez, Robert Jäschke, and Madalina Croitoru, editors, *Graph-Based Representation and Reasoning*, pages 260–274, Cham, 2014. Springer International Publishing.

[16] Alistair Russell and Paul Smart. NITELIGHT : A Graphical Editor for SPARQL Queries. In *7th International Semantic Web Conference (ISWC 2008), Poster*, October 2008.

[17] Laszlo Szathmary and Amadeo Napoli. CORON : A framework for levelwise itemset mining algorithms. *Supplementary Proceedings of The Third International Conference on Formal Concept Analysis (ICFCA '05), Lens, France*, pages 110–113, 2005.

[18] Munazza Zaib, Wei Emma Zhang, Quan Z. Sheng, Adnan Mahmood, and Yang Zhang. Conversational question answering : a survey. *Knowledge and Information Systems*, 64(12) :3151–3195, 2022.