Et si on comprenait la structure de graphes de connaissances comme Wikidata ?

Hassan Abdallah, Béatrice Markhoff, Arnaud Soulet

Et si on comprenait la structure de graphes de connaissances comme Wikidata ?

Hassan Abdallah[1], Béatrice Markhoff[2], Arnaud Soulet[1]

[1] Université de Tours, EA 6300 LIFAT, Blois
[2] Université de Tours, UMR 7324 CITERES, Blois

prenom.nom@univ-tours.fr

Résumé

La production participative de graphes de connaissances comme Wikidata a permis l'émergence de vastes bases de connaissances agglomérant des expertises et des opinions variées. Fortement dépendant de sa communauté, ce processus décentralisé interroge sur sa capacité à faire émerger un graphe de connaissances représentatif et cohérent. Dans cet article, nous affirmons que la représentation des connaissances gagnerait à modéliser la manière dont les faits s'organisent dans la partie assertionnelle. L'objectif serait de définir et étudier des processus permettant de générer des données synthétiques qui ressemblent étroitement aux graphes de connaissances réels. Nous indiquons des retombées possibles de ces modèles en optimisation et en analyse de données. Enfin, nous envisageons les principaux verrous scientifiques à lever pour parvenir à modéliser la structure des graphes de connaissances.

Mots-clés

Graphe de connaissances, modèle génératif, Wikidata.

Abstract

Crowdsourcing of knowledge graphs (KGs) such as Wikidata has allowed the emergence of vast knowledge bases agglomerating various expertises and opinions. Strongly dependent on its community, this decentralized process questions its capacity to produce a representative and coherent KG. In this paper, we argue that knowledge representation would benefit from modeling the way facts are organized in the assertional part. The goal would be to define and study processes to generate synthetic graphs that closely resemble real KGs. We indicate possible implications of these models in optimization and KG analysis. Finally, we consider the main scientific obstacles to overcome in order to model KGs.

Keywords

Knowledge graph, generative model, Wikidata.

1 Introduction

La modélisation de la structure des réseaux et de leur dynamique est un domaine qui a pris de l'ampleur dans les années 2000. Elle s'intéresse aux relations et interconnexions entre des objets, dans des domaines aussi divers que les réseaux d'ordinateurs ou la biologie. En particulier, elle a apporté un ensemble important de connaissances sur l'émergence de structures et leurs dynamiques dans les artefacts du Web : réseau des pages HTML et réseaux sociaux [6], résultats de systèmes distribués d'annotations collaboratives (folksonomies) [19]. Parmi les caractéristiques communes à ces artefacts, il y a le fait d'être construits par des ensembles de personnes qui interagissent de façon distribuée, ne communiquant qu'à travers leurs actions sur des ressources du Web et, selon les cas, se coordonnant en communautés ou pas. L'étude de la structure des artefacts résultant de ces actions isolées, qui deviennent des interactions par le fait d'agir sur les mêmes ressources du Web (pointer vers une page Web depuis une autre, suivre ou répondre à un compte de réseau social, annoter la même ressource) apporte divers éclairages et permet des analyses riches (voir par exemple [23] sur la visualisation des structures et dynamiques de connaissances contenues dans des articles scientifiques). Notre intuition est que des modèles génératifs des graphes de connaissances du Web offriraient des perspectives de réflexion tout aussi riches, et auraient par ailleurs des applications concrètes pour améliorer l'exploitation de ces derniers. Ceci nous paraît particulièrement vrai pour les grands graphes construits de manière participative, de façon distribuée et peu contrainte, comme Wikidata. Un outil qui reproduirait une telle construction d'une connaissance commune, en d'autres termes une construction de consensus sur des connaissances, permettrait évidemment d'observer et analyser « in vitro » cette construction, de produire des benchmarks plus représentatifs, et d'élaborer des méthodes d'analyse statistique et d'apprentissage plus fiables.

L'étude de la structure des graphes de connaissances et de son évolution est donc impérative pour gagner en compréhension sur les bases de connaissances produites collaborativement. De manière immédiate, il est possible d'appliquer à ces graphes de connaissances des statistiques descriptives pour observer des phénomènes [13]. L'une des principales caractéristiques d'un graphe de connaissances est sa distribution de degrés, qui mesure le nombre de faits impliquant une entité avec les autres entités du graphe. De manière intéressante, cette information précieuse indique la répartition des faits au sein du graphe, mettant en lu-

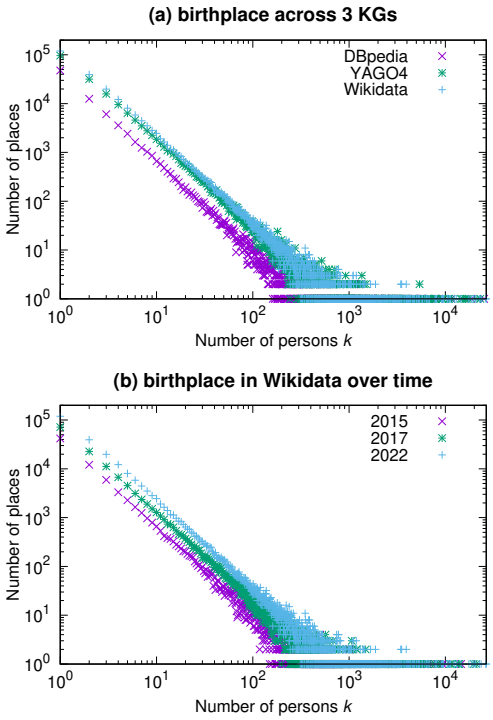

FIGURE 1 – Illustration de l'existence d'une structure pour la relation `place of birth` (dénotée par `wdt:P19`) (a) à travers 3 graphes de connaissances issus de productions collaboratives et (b) au fil du temps dans Wikidata.

mière des entités mal-renseignées et celles concentrant la majorité des connaissances. Prenons par exemple la relation `place of birth` (dénotée par `wdt:P19` dans Wikidata) présente dans DBpedia [5], YAGO4 [26] et Wikidata [40], la figure 1(a) représente le nombre de lieux en fonction du nombre de naissances pour ces 3 graphes. On y voit que dans ces trois ressources il y a de l'ordre de 10^5 lieux dans lesquels seulement une personne est déclarée être née (en haut à gauche). On voit également qu'il y a une concentration des déclarations de naissance sur relativement peu de lieux (en bas à droite, entre 100000 et plusieurs dizaines de millions pour une dizaine de villes). Étonnamment, bien que ces trois graphes ne portent pas sur les mêmes données, ils exhibent une structure similaire pour cette relation. De plus, la figure 1(b) représente la même distribution pour Wikidata en considérant 3 années différentes depuis sa création. Outre l'augmentation attendue du nombre de faits, nous constatons que la structure à travers le temps reste la même. Nous avons également observé cette même stabilité de la structure à travers le temps et les graphes pour d'autres relations : `instance of` (`wdt:P31`), `subclass of` (`wdt:P279`), `creator` (`wdt:P170`), etc. A la lumière de ces exemples, il parait nécessaire d'aller au-delà de simples observations statistiques pour comprendre comment s'accumulent les faits au sein de ces graphes de connaissances et les phénomènes sous-jacents conduisant à l'émergence de structures régulières.

Positionnement La représentation des connaissances [38] s'intéresse en profondeur à la formalisation et la modélisation de la partie terminologique avec notamment des travaux variés sur les logiques de description et les ontologies. Cette compréhension de la partie terminologique se répercute sur la forme des faits constituant la partie assertionnelle, mais elle ne décrit pas certaines singularités de la structuration des entités et des relations (comme la distribution des faits). De plus, il serait impossible d'étudier Wikidata à partir de son ontologie puisque ce graphe de connaissances ne repose pas sur une ontologie [41]. Nous pensons donc qu'il est nécessaire d'étendre le champs de la représentation des connaissances à l'étude de la manière dont s'organisent les faits pour constituer le graphe de connaissances. Il s'agit de développer des modèles pour capturer l'organisation des faits afin de pouvoir générer des données synthétiques qui ressemblent étroitement aux graphes de connaissances réels, à l'instar des travaux qui ont été menés pour expliquer la structuration du Web [6] ou celle des folksonomies [19]. En effet, l'objectif d'un modèle génératif est de reproduire les données réelles ; donc, si cette reproduction est fidèle (si le modèle est correct), alors l'étude de la structure générée par le modèle permet de découvrir des caractéristiques de la structure du graphe réel, et de son évolution. Pour étayer ce positionnement, nous rappelons en section 2 ce qui caractérise Wikidata d'une part, et d'autre part les travaux existants sur la compréhension des graphes de connaissances, puis nous nous intéressons aux retombées qu'aurait une modélisation de la structure des graphes de connaissances dans la section 3. Ensuite, nous envisageons les principaux verrous scientifiques à lever pour parvenir à modéliser les graphes de connaissances, dans la section 4, avec plusieurs directions de recherche.

2 Genèse de Wikidata et compréhension des graphes de connaissances

L'émergence des grands graphes de connaissances du Web, et en particulier de Wikidata, repose en grande partie sur la production collaborative, ce qui soulève la question de la légitimité des données construites en termes de représentativité et de cohérence. En effet, la production collaborative est un processus qui consiste à solliciter les contributions d'un grand groupe de personnes pour obtenir des données ou des informations [36]. Ce processus a gagné en popularité avec la croissance des technologies numériques facilitant la mise en relation de personnes du monde entier pour qu'elles puissent travailler ensemble à la réalisation d'un objectif commun. La production collaborative est particulièrement utile lorsque de grandes quantités de données ou un large éventail de perspectives sont nécessaires [32]. D'une part, les données ainsi produites peuvent être vastes et variées (e.g., textes ou images) avec la possibilité de rassembler de grands volumes en un temps relativement court. D'autre part, un autre avantage de la production collaborative est la potentielle diversité des perspectives exprimées. En sollicitant les contributions d'un grand nombre

de personnes, cela permet de s'assurer qu'un large éventail d'opinions, d'expériences et de points de vue sont représentés dans les données produites. Dans le cas de Wikidata [40], cet aspect s'avère particulièrement important. En s'appuyant sur les contributions d'une communauté diversifiée de contributeurs, Wikidata est en mesure d'élargir et d'affiner sa base de connaissances au fil du temps pour renforcer l'exactitude et l'exhaustivité des données. En revanche, la décentralisation de la génération des faits et leur maintenance par une large communauté d'individus soulève des incertitudes. Malgré la diversité désirée de ses contributeurs, la communauté peut être déséquilibrée introduisant des biais sur les données produites. Le volume important de données rassemblées peut négliger certains domaines, entraînant des biais de représentativité culturels ou sociaux [12]. De plus, pour un même domaine polémique, il est possible d'imaginer que sans organisation centralisée les divergences entre les points de vue des individus provoquent des inconsistances voire de l'instabilité. Par exemple, sans ontologie de référence, on peut s'interroger sur la convergence de la terminologie de Wikidata, co-construite par les contributeurs. De nombreux travaux se sont intéressés aux aspects sociaux et organisationnels de la communauté collaborant pour construire un tel graphe de connaissances [34, 28]. Mais, à notre connaissance, la répercussion de cette production collaborative sur l'émergence d'une structuration des données n'a pas reçu d'attention.

La plupart des approches de construction de graphes de connaissances consistent en des workflows de production partant de diverses sources de données (textes, pages web, tableurs, bases de données sous toutes les formes) et utilisant une ou plusieurs ontologies, ainsi qu'un ou plusieurs thésauri de référence. Les initiateurs et pilotes de Wikidata ont délibérément choisi de ne pas construire une ontologie au préalable et de ne pas imposer de thesaurus non plus, sachant que Wikipedia en a fourni le noyau de départ [41]. Pour autant, cela reste aux ontologies, ou schémas, que l'on pense quand il est question de la structure des graphes de connaissances [20]. Au-delà de la nécessité pour les utilisateurs de découvrir et comprendre les ontologies utilisées dans les graphes de connaissances, comme il est toujours difficile de saisir rapidement le contenu de ces graphes, de nombreux travaux s'attachent à en extraire des informations statistiques [7], des résumés [10, 17], des schémas [21], des profils [35, 14], des contraintes de forme [30], ou bien à les munir d'interfaces d'interrogation aussi intuitives que possible, par exemple [24, 39] pour Wikidata. En parallèle à ces efforts pour exhiber la sémantique des données contenues dans les graphes de connaissances, il y a également de très nombreuses propositions consistant à aborder les graphes de connaissances avec des outils de deep learning [25, 31]. Globalement ces modèles génératifs profonds permettent une reproduction fine mais nécessitent des données d'apprentissage et apportent peu de compréhension des graphes car ils requièrent trop de paramètres. Malgré la diversité des approches que nous venons de résumer, nous n'avons pas trouvé de proposition visant à comprendre et modéliser la topologie propre aux graphes de connaissances.

3 Intérêt d'une modélisation

Il ne fait aucun doute que la proposition de modèles précis pour les graphes de connaissances apporterait énormément au domaine tant d'un point de vue pratique que théorique. Au-delà de la création de données synthétiques, l'analyse théorique du modèle génératif peut en effet apporter une loi de probabilité sur les données. Par exemple, la distribution du nombre de naissances de la figure 1 suit une loi de puissance d'exposant 1.91 dans Wikidata. La connaissance d'une telle loi est évidemment utile à diverses fins notamment en ingénierie des données, en analyse de données et en qualification des données.

3.1 Optimisation et benchmarking

Comme la taille des graphes de connaissances les plus importants comme Wikidata ne cesse de croître, il est important d'améliorer la performance des systèmes d'interrogation comme les moteurs SPARQL [1, 3]. A l'instar des bases de données, il est nécessaire d'exploiter les propriétés des données pour optimiser l'exécution des requêtes et de s'appuyer sur des benchmarks pour comparer les systèmes. D'une part, les propriétés mathématiques dérivées des modèles peuvent être injectées directement dans le système d'interrogation à la place des statistiques sur les données. En effet, le stockage de données et l'optimisation du plan d'exécution nécessitent des informations sur la répartition des données pour améliorer l'exécution des requêtes. Ces statistiques sont coûteuses à obtenir et à maintenir voire sont remplacées par des heuristiques imprécises [37]. A l'inverse, les modèles reposent sur des paramètres d'entrée qui résument précisément les principales caractéristiques des données réelles à simuler (e.g., l'exposant 1.91 indique la répartition globale des faits de la relation `place of birth`). D'autre part, le développement et le test des systèmes d'interrogation nécessitent aussi des benchmarks variés pour analyser les différents contextes d'interrogation. La génération de données synthétiques est une approche qui permet de relever ce défi. Elle consiste à créer de nouvelles données ayant des caractéristiques statistiques similaires aux données réelles, ce qui permet aux chercheurs de tester et d'optimiser leurs propositions sans avoir recours à des données réelles. Cette approche est particulièrement utile pour pouvoir observer la répercussion d'un paramètre précis de la génération de données sur le système d'interrogation pour mieux comprendre ses forces et ses faiblesses. A notre connaissance, les principaux générateurs de données synthétiques pour les graphes de connaissances utilisent uniquement des lois normale ou uniforme pour générer la distribution des faits comme BSBM [8] ou LUBM [18]. Il est clair que ces benchmarks ne sont pas adaptés pour simuler des relations comme `place of birth` se rapprochant davantage d'une loi de puissance. Par conséquent, la proposition de modèles fins pour générer des données synthétiques est cruciale pour construire des benchmarks plus réalistes.

3.2 Exploration de données

Nous pensons que la connaissance de modèles pourrait bénéficier à la fois aux méthodes d'exploration de données descriptive et prédictive. De manière évidente, les statistiques descriptives s'appuient sur les lois statistiques connues pour les observations. A la lumière de la figure 1, il semble judicieux d'utiliser une loi de pareto plutôt qu'une loi normale pour analyser le nombre de naissances par ville. De la même manière, si l'on souhaite classifier les villes en différents groupes, la distribution suggère d'éviter d'utiliser K-means [22], plutôt adapté à un mélange de gaussiennes. La difficulté est que la distribution dépend de chaque relation, renforçant l'intérêt de disposer d'un modèle qui permettrait de les distinguer. En outre, l'objectif d'un modèle génératif pour les graphes de connaissances est aussi de simuler la croissance du graphe dans le temps. Par analogie avec d'autres travaux en science des réseaux, il serait alors possible d'estimer la probabilité pour une entité de recevoir un nouveau fait en se basant uniquement sur la structure du graphe. Cette connaissance s'avère évidemment précieuse pour analyser des données. La prédiction de lien est une tâche prédictive très populaire pour compléter et augmenter les graphes de connaissances [31]. Les approches actuelles basées sur des indices locaux pourraient alors prendre en compte la structure globale du graphe. Inversement, de nombreuses méthodes en détection d'anomalie [2] exploitent une connaissance structurelle attendue sur les graphes pour identifier comme anomalies les arcs ou les noeuds qui dévient de cette structure. Schématiquement, pour la relation `place of birth`, il est possible de modéliser le nombre de naissances nouvelles rattachées annuellement à chaque ville. Dès lors, une ville qui recevrait subitement de nombreux faits au quotidien contre une estimation attendue de quelques faits sur l'année, serait identifiée comme une anomalie.

3.3 Qualification des données

Une question critique est de déterminer si la structure du graphe de connaissances est stable. Il est essentiel de comprendre si la distribution de probabilité des degrés dans le graphe de connaissances converge ou non [19]. En analysant cette caractéristique du graphe, il est possible d'avoir une idée de la stabilité générale du graphe et de la probabilité qu'il évolue au fil du temps [16]. La représentativité de la connaissance est une autre caractéristique critique qui doit être prise en considération lors de l'exploitation de graphes de connaissances. Il s'agit de savoir dans quelle mesure le graphe reliant des entités est complet et non biaisé. Un modèle adéquat de la structure du graphe et de son évolution permettrait de s'assurer de sa représentativité en identifiant le nombre minimum d'entités requis pour la garantir [33]. Par ailleurs, un modèle génératif peut aider à déterminer les entités vulnérables ou au contraire robustes dans le graphe de connaissances, en déterminant si l'ajout ou la suppression d'un fait peut remettre en cause la connaissance courante. Un autre aspect encore, qui peut être analysé à l'aide d'un modèle de la structure du graphe, est la découverte de nouvelles relations entre des entités, qui peut émerger d'interactions entre les contributeurs, ce qui apporterait un éclairage complémentaire sur la dynamique sociale de l'ingénierie collaborative des connaissances [28, 27]. De plus, un modèle peut permettre de détecter des erreurs ou incohérences dans les données. Par exemple, dans la figure1(a), il est clair que les trois sources sont concordantes, aussi un modèle de cette relation pourrait montrer des contradictions dans les informations provenant d'une autre source, concernant cette même relation. Cela aiderait à améliorer la qualité des données du graphe [29] et le rendrait ainsi plus utile à différentes applications. Par conséquent, un modèle bien conçu permettrait de montrer des caractéristiques non apparentes et des relations implicites, menant à de nouvelles idées et découvertes.

4 Défis de la modélisation

Pour modéliser la structuration de la production collaborative de connaissances, il serait possible d'utiliser des techniques d'analyse de réseaux comme initié par [6]. Ces techniques fournissent un moyen de décrire les relations entre les entités, et de mettre en évidence des motifs d'interactions et de collaboration qui apparaissent au sein de communautés de contributeurs et contributrices. De nombreux travaux ont suivi ceux de Albert et Barabási [6], encore tout récemment un nouveau modèle générique a été proposé dans [16] pour reproduire la distribution de différents réseaux complexes [11] en focalisant sur la fonction d'attachement pour la connexion des noeuds. Cette fonction gouverne la manière dont les nouveaux noeuds sont connectés au graphe existant. Un autre modèle de croissance a été introduit dans [9] pour décrire les graphes *dirigés* sans-échelle dont la taille augmente grâce à une fonction d'attachement préférentiel.

Nous n'avons trouvé pour les graphes de connaissances aucune proposition de modélisation de la structure et de la dynamique du graphe en lui-même. Ceci s'explique sans doute par un ensemble de caractéristiques propres aux graphes de connaissance, qui font que les modèles existants ne peuvent pas s'y appliquer, ni même s'y adapter simplement. La principale tient sans doute au fait qu'ils représentent des connaissances, et qui plus est, pour un graphe comme celui de Wikidata, des connaissances dans des domaines très divers. Ces connaissances sont donc décrites par *des relations très diverses*, dont le nombre augmente avec la croissance du graphe. Même si leur création est nettement plus encadrée que la création d'entités, il n'en demeure pas moins qu'en 2015 il y avait de l'ordre de 500 relations directes (nous ne considérons pas celles qui permettent de caractériser les assertions) et en 2022 il y en a pratiquement trois fois plus. Alors que les modèles existant pour les réseaux considèrent tous les arcs du graphe de façon indistincte, pour un graphe de connaissances il est crucial que le modèle représente précisément le comportement des ensembles d'arcs qui ont la même étiquette, correspondant à une relation décrivant un ensemble d'entités du monde réel.

Par conséquent concevoir un modèle génératif fiable pour des graphes de connaissances est un vrai défi. Pour s'y atta-

quer il faut prendre en compte plusieurs dimensions, parmi lesquelles la véracité, le volume et la variété des données de ces graphes.

Véracité : Concernant la véracité des données, la conception d'un modèle robuste d'un graphe de connaissances doit prendre en compte le fait que les connaissances présentes sont fréquemment incomplètes ou, plus rarement, erronées [33], ce qui rend nécessaire de développer un modèle capable de traiter des éléments manquants ou bruités. Le modèle doit aussi pouvoir rendre compte de la nature dynamique du monde représenté, par exemple l'objet de la relation `wdt:P39` (`position held`) change régulièrement pour une personne donnée. Ces évolutions du graphe, différentes de simples ajouts d'entités, doivent être modélisées aussi. De plus, tandis que la plupart des modèles existants s'attachent à ajouter des entités, pour les graphes de connaissances il faudrait également modéliser l'évolution inverse, la suppression d'entités (par exemple des entités devenues obsolètes).

Volume de données : L'une des principales difficultés réside dans le très grand volume de données des graphes comme Wikidata. Or pour mener les expérimentations nécessaires au paramétrage d'un modèle génératif, il est nécessaire de recueillir les données du graphe par des requêtes analytiques. Cela nécessite d'utiliser des algorithmes capables de traiter ces très grands volumes de données, sans compromettre leur performance. De plus, le modèle en lui-même doit être efficace dans la génération de gros graphes synthétiques, ce qui souleve des défis algorithmiques [4]. Pour la mise au point du modèle comme pour son exécution, la gestion de la mémoire est un autre problème crucial à considérer.

Variété : La modélisation de graphes de connaissances qui contiennent des connaissances multidisciplinaires, à l'image de Wikidata, est une tâche compliquée qui nécessiterait une approche interdisciplinaire. En d'autres termes, lors de la génération des données synthétiques, il serait nécessaire de vérifier la capacité du modèle à fournir une représentation pertinente de ces connaissances du monde réel. Dans les différents domaines représentés par les graphes de connaissances, certains bénéficient déjà de modèles bien connus, par exemple en bibliométrie [15], tandis que d'autres pas. Mettre au point un modèle pour un domaine requiert une connaissance de ce domaine et des méthodes statistiques pour générer des valeurs fiables. Cependant, créer un modèle unique et spécifique à chaque domaine n'est pas faisable. Aussi, il faudrait un modèle capable de fonctionner pour les diverses disciplines et domaines.

De plus, notons que les graphes de connaissances sont constitués de différents types d'entités et de relations, en particulier des entités correspondant à l'observation du réel et des entités plus conceptuelles, décrivant des connaissances dérivées des données d'observation. Prenons l'exemple d'un graphe de connaissances en biologie médicale. Ce graphe peut inclure à la fois des entités réelles, telles que les maladies et les symptômes, et des entités conceptuelles, telles que les mécanismes biologiques sous-jacents à l'origine de ces maladies. La diversité des entités et des relations, ainsi que la variété des domaines et disciplines, rendent difficile la création d'un modèle de croissance aléatoire qui refléterait avec précision la mosaïque structurelle des graphes de connaissances.

5 Conclusion

Cet article propose de s'intéresser à la structure des graphes de connaissances en concevant et exploitant des modèles génératifs inspirés de la science des réseaux. Leur mise au point soulève de nombreux défis liés à la forme des graphes de connaissances bien plus complexe, avec la superposition des sémantiques portées par chacune des relations. Pourtant, de tels modèles singeant les données réelles seraient idéals pour mieux comprendre la structure sous-jacente des graphes et leur dynamique. Cette compréhension serait utile pour de nombreux travaux sur les graphes de connaissances allant de l'optimisation à la qualification des données. Au-delà, comme le graphe est un miroir des connaissances d'un domaine, elle permettrait aussi d'apporter un éclairage sur la constitution des connaissances au sein de ce domaine.

Références

[1] Ibrahim Abdelaziz, Razen Harbi, Zuhair Khayyat, and Panos Kalnis. A survey and experimental comparison of distributed sparql engines for very large rdf data. *VLDB*, 10(13) :2049–2060, 2017.

[2] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description : a survey. *Data mining and knowledge discovery*, 29 :626–688, 2015.

[3] Waqas Ali, Muhammad Saleem, Bin Yao, Aidan Hogan, and Axel-Cyrille Ngonga Ngomo. A survey of rdf stores & sparql engines for querying knowledge graphs. *The VLDB Journal*, pages 1–26, 2022.

[4] James Atwood, Bruno Ribeiro, and Don Towsley. Efficient network generation under general preferential attachment. In *WWW*, pages 695–700, 2014.

[5] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia : A nucleus for a web of open data. In *ISWC*, pages 722–735, 2007.

[6] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439) :509–512, 1999.

[7] Mohamed Ben Ellefi, Zohra Bellahsene, John G Breslin, Elena Demidova, Stefan Dietze, Julian Szymański, and Konstantin Todorov. Rdf dataset profiling–a survey of features, methods, vocabularies and applications. *Semantic Web*, 9(5) :677–705, 2018.

[8] Christian Bizer and Andreas Schultz. The berlin sparql benchmark. *International Journal on Semantic Web and Information Systems*, 5(2) :1–24, 2009.

[9] Béla Bollobás, Christian Borgs, Jennifer T Chayes, and Oliver Riordan. Directed scale-free graphs. In *SODA*, volume 3, pages 132–139, 2003.

[10] S. Cebiric, F. Goasdoue, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, and M. Zneika. Summarizing Semantic Graphs : A survey. *The VLDB Journal*, 28 :295–327, 2018.

[11] Fan Chung, Fan RK Chung, Fan Chung Graham, Linyuan Lu, et al. *Complex graphs and networks*. Number 107. American Mathematical Soc., 2006.

[12] Gianluca Demartini. Implicit bias in crowdsourced knowledge graphs. In *WWW*, pages 624–630, 2019.

[13] Li Ding and Tim Finin. Characterizing the semantic web on the web. In *ISWC*, pages 242–257, 2006.

[14] Lamine Diop, Béatrice Markhoff, and Arnaud Soulet. TTProfiler : types and terms profile building for online cultural heritage knowledge graphs. *JOCCH*, 2023.

[15] Leo Egghe. *Power laws in the information production process : Lotkaian informetrics*. 2005.

[16] Frédéric Giroire, Stéphane Pérennes, and Thibaud Trolliet. A random growth model with any real or theoretical degree distribution. *Theoretical Computer Science*, 940 :36–51, 2023.

[17] F. Goasdoue, P. Guzewicz, and I. Manolescu. RDF graph summarization for first-sight structure discovery. *The VLDB Journal*, 29(5) :1191–1218, 2020.

[18] Yuanbo Guo, Zhengxiang Pan, and Jeff Heflin. Lubm : A benchmark for owl knowledge base systems. *Journal of Web Semantics*, 3(2-3) :158–182, 2005.

[19] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *WWW*, pages 211–220, 2007.

[20] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4) :1–37, 2021.

[21] Kenza Kellou-Menouer, Nikolaos Kardoulakis, Georgia Troullinou, Zoubida Kedad, Dimitris Plexousakis, and Haridimos Kondylakis. A survey on semantic schema discovery. *The VLDB Journal*, pages 1–36, 2021.

[22] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2) :129–137, 1982.

[23] Quentin Lobbé, Alexandre Delanoë, and David Chavalarias. Exploring, browsing and interacting with multi-level and multi-scale dynamics of knowledge. *Information Visualization*, 21(1) :17–37, 2022.

[24] Stanislav Malyshev, Markus Krötzsch, Larry González, Julius Gonsior, and Adrian Bielefeldt. Getting the most out of Wikidata : semantic technology usage in wikipedia's knowledge graph. In *ISWC*, pages 376–394, 2018.

[25] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proc. of the IEEE*, 104(1) :11–33, 2015.

[26] Thomas Pellissier Tanon, Gerhard Weikum, and Fabian Suchanek. Yago 4 : A reason-able knowledge base. In *ESWC*, pages 583–596, 2020.

[27] Guangyuan Piao and Weipeng Huang. Learning to predict the departure dynamics of Wikidata editors. In *ISWC*, pages 39–55, 2021.

[28] Alessandro Piscopo and Elena Simperl. Who models the world ? collaborative ontology creation and user roles in Wikidata. *HCI*, 2 :1–18, 2018.

[29] Alessandro Piscopo and Elena Simperl. What we talk about when we talk about Wikidata quality : a literature survey. In *the 15th International Symposium on Open Collaboration*, pages 1–11, 2019.

[30] Kashif Rabbani, Matteo Lissandrini, and Katja Hose. Extraction of validating shapes from very large knowledge graphs. *PVLDB*, 16(5) :1023–1032, 2023.

[31] Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. Knowledge graph embedding for link prediction : A comparative analysis. *ACM Transactions on Knowledge Discovery from Data*, 15(2) :1–49, 2021.

[32] Cristina Sarasua, Elena Simperl, Natasha F Noy, Abraham Bernstein, and Jan Marco Leimeister. Crowdsourcing and the semantic web : A research manifesto. *Human Computation*, 2(1), 2015.

[33] Suhas Shrinivasan and Simon Razniewski. How stable is knowledge base knowledge ? *arXiv preprint arXiv :2211.00989*, 2022.

[34] Elena Simperl and Markus Luczak-Rösch. Collaborative ontology engineering : a survey. *The Knowledge Engineering Review*, 29(1) :101–131, 2014.

[35] Blerina Spahiu, Riccardo Porrini, Matteo Palmonari, Anisa Rula, and Andrea Maurino. ABSTAT : Ontology-Driven Linked Data Summaries with Pattern Minimalization. In *ESWC*, pages 381–395, 2016.

[36] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[37] Petros Tsialiamanis, Lefteris Sidirourgos, Irini Fundulaki, Vassilis Christophides, and Peter Boncz. Heuristics-based query optimisation for sparql. In *ICDT/EBDT*, pages 324–335, 2012.

[38] Frank Van Harmelen, Vladimir Lifschitz, and Bruce Porter. *Handbook of knowledge representation*. Elsevier, 2008.

[39] Hernán Vargas, Carlos Buil-Aranda, Aidan Hogan, and Claudia López. A user interface for exploring and querying knowledge graphs (extended abstract). In *IJCAI*, pages 4785–4789, 2020.

[40] Denny Vrandečić and Markus Krötzsch. Wikidata : a free collaborative knowledgebase. *Communications of the ACM*, 57(10) :78–85, 2014.

[41] Denny Vrandečić, Lydia Pintscher, and Markus Krötzsch. Wikidata : The making of. In *the ACM Web Conference 2023*, pages 615–624, 2023.