Dessine-moi un graphe de connaissances !

Fabien Gandon

binaire
L'informatique : science et technique au coeur du numérique
Avec la Société Informatique de France

05 OCTOBRE 2021 PAR BINAIRE

Dessine-moi un graphe de connaissances !

Comment représenter des connaissances de manière formelle pour que des logiciels puissent les utiliser ? Plein de trucs ont été essayés et ce qui marche bien c'est la structure de graphe. Les nœuds sont des entités et les liens des relations entre elles. Bon, on a un peu trop simplifié. Fabien Gandon nous parle des graphes de connaissance, une branche de l'IA avec des applications impressionnantes, peut-être moins connue que l'apprentissage automatique mais toute aussi passionnante. Fabien est informaticien, chercheur chez Inria. Il est Professeur au Data ScienceTech Institute, Titulaire d'une Chaire 3IA aux Instituts Interdisciplinaires d'Intelligence Artificielle de l'Université Côte d'Azur. C'est un des meilleurs spécialistes en représentation des connaissances et Web Sémantique. ***Serge Abiteboul, Ikram Chraibi Kaadoud, Thierry Viéville***

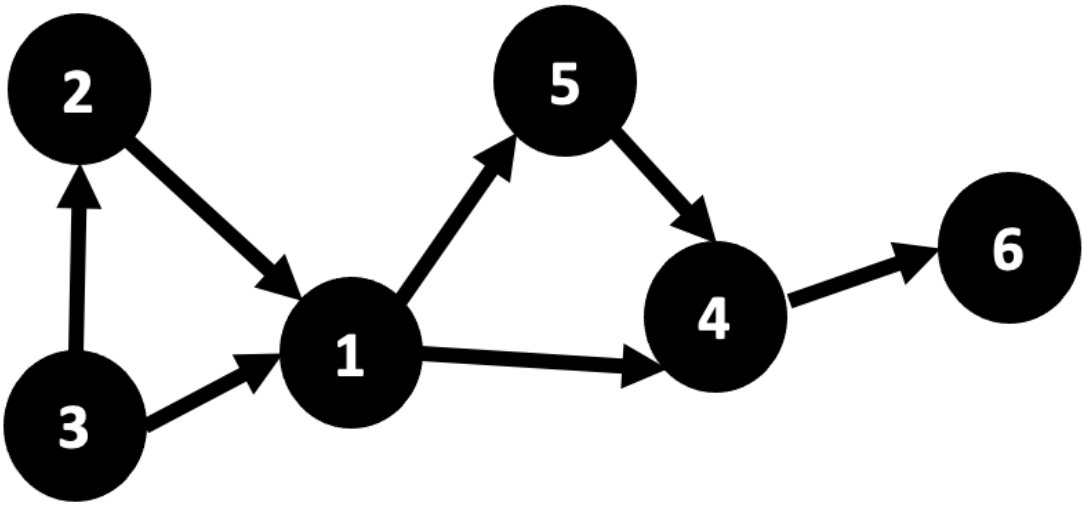
Page de Fabien Gandon, A partir de « Les défis de l'intelligence artificielle – Un reporter dans les labos de recherche », Jérémie Dres, 2021.

Le terme de « graphe de connaissance » existe depuis des décennies mais son utilisation par Google en 2012 pour un nouveau service, puis par un nombre grandissant d'autres entreprises, l'ont rendu extrêmement populaire dernièrement. De plus son couplage avec différentes techniques d'intelligence artificielle contribue à en faire un sujet d'intérêt d'actualité. Si, à l'instar de cette expression « intelligence artificielle », le terme « graphe de connaissance » ou *Knowledge Graph* est utilisé avec différentes acceptions et identifie actuellement une ressource numérique très différente d'un cas d'usage à un autre, le domaine de la représentation des connaissances à base de graphes existe depuis longtemps et étudie l'expressivité de ces modèles et la complexité de leurs traitements avec des interactions multidisciplinaires et des applications dans de nombreux domaines.

S'il vous plaît… dessine-moi un graphe de connaissances !

Un graphe est une structure mathématique contenant un ensemble d'objets dans lequel certaines paires

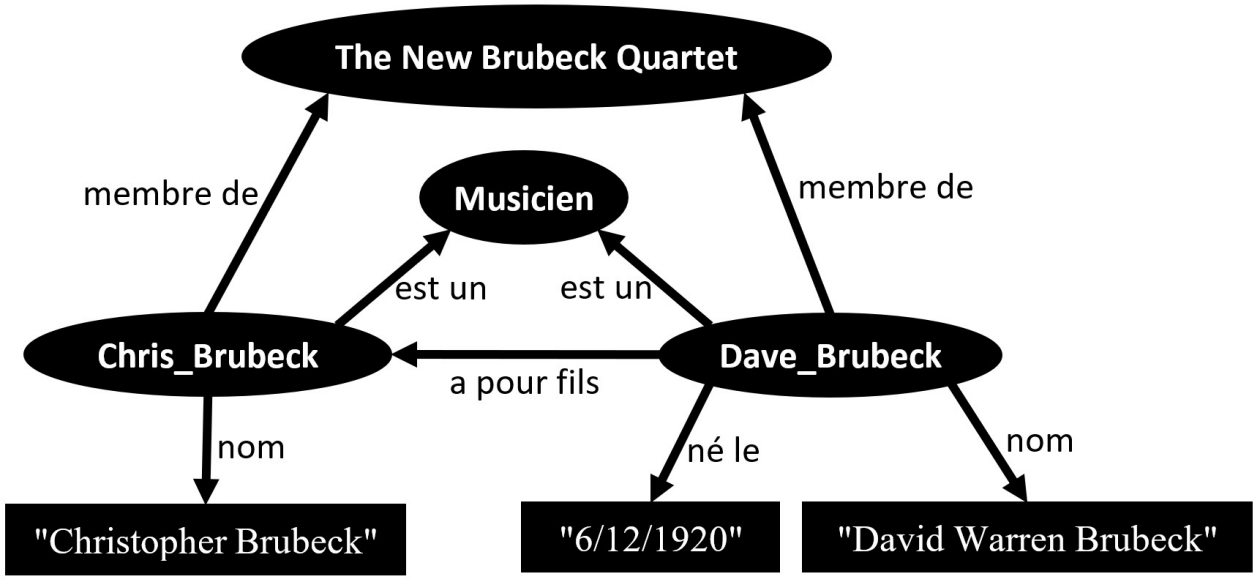
d'objets sont en relation. Les objets et les relations peuvent être très variés comme par exemple des villes reliées par des routes, des personnes reliées par des relations sociales ou des livres reliés par des citations. Un graphe est typiquement dessiné sous la forme de points représentant les objets (sommets du graphe) et de lignes entre eux représentant les relations (arêtes du graphe).



Un graphe avec six sommets et sept arêtes

Un graphe de connaissances représente des données très variées en les augmentant avec des connaissances explicites attachées aux sommets et aux arêtes du graphe pour donner des informations sur leur sens, leur structure et leur contexte. Il est explicitement utilisé pour représenter et formaliser nos connaissances dans des applications informatiques.

Prenons l'exemple d'un graphe de connaissances dans le domaine de la musique. Les sommets de ce graphe peuvent représenter des albums, des artistes, des concerts, des chansons, des labels, des langues, des genres, etc., et les arêtes peuvent capturer les relations d'auteur, compositeur, interprète, parolier, indiquer les influences artistiques, connecter les différentes versions d'un morceau ou grouper les morceaux d'un album, etc.



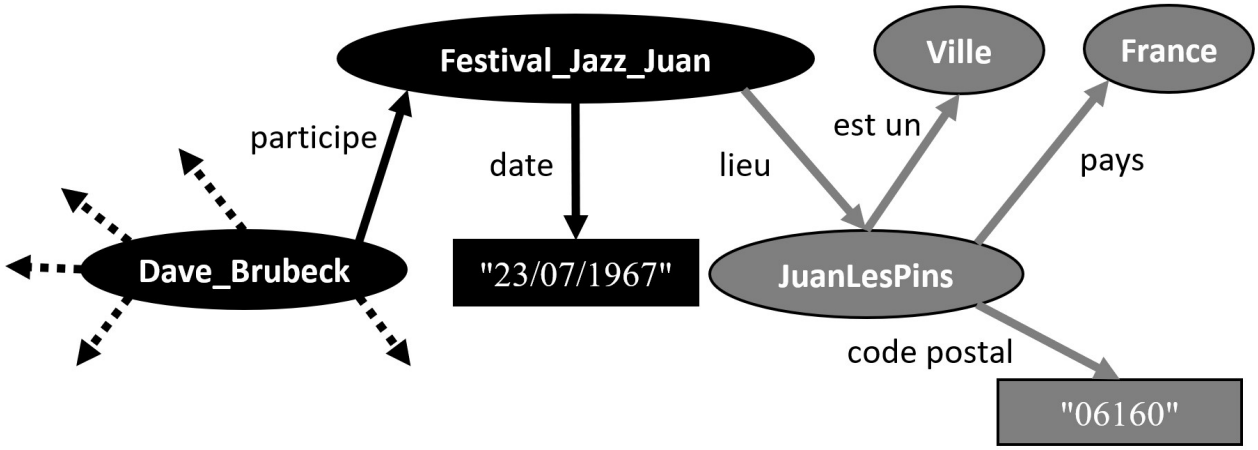
Un petit graphe de connaissance en musique

Dans un graphe de connaissance on trouvera typiquement deux types de sommets : ceux qui représentent des objets (ex. les musiciens) et ceux qui représentent des données (ex. une date, un texte). On trouvera donc aussi deux types d'arêtes : celles qui relient des objets (ex. un père et son fils) et celles qui indiquent

des attributs d'un objet (ex. la date de naissance d'une personne).

Des graphes à tout faire

Que ce soit au sein d'un même graphe ou entre des graphes différents, on trouve des connaissances de natures très variées dans ces graphes. Les connaissances peuvent être organisées dans des arbres pour une taxonomie d'espèces, ou plutôt en réseau pour un réseau social ou pour des liens entre sites web. On peut créer des ponts entre différents graphes de connaissances notamment en réutilisant des sommets de l'un dans l'autre. Par exemple, un graphe de connaissance géographique capturant des villes, des reliefs, des frontières, pourra en certains sommets rejoindre notre graphe sur la musique quand la description d'un concert indiquera le lieu de cet évènement.



Dans la pratique, une distinction peut se faire entre deux grandes familles de graphes de connaissances : les graphes de connaissance ouverts et les graphes de connaissance privés notamment les graphes d'entreprise.

Les graphes de connaissance ouverts sont publiés en ligne comme des biens publics. Certains sont publiés dans des domaines spécifiques, tels que les sciences naturelles (ex. le graphe UniProt décrivant les protéines), la géographie (ex. le graphe GeoNames) ou la musique (ex. le graphe de MusicBrainz). D'autres couvrent des connaissances générales comme DBpedia ou YAGO qui sont des graphes extraits de Wikipedia par des algorithmes, ou Wikidata qui est un graphe construit collaborativement par une communauté de volontaires.

Les graphes de connaissance d'entreprise sont généralement internes à celle-ci car ils font l'objet d'une utilisation commerciale ou sont au cœur de son système d'information. On en trouve dans tous les domaines, depuis l'industrie jusqu'aux différents acteurs de la finance en passant par les sites marchands, les services de relation client ou l'éducation.

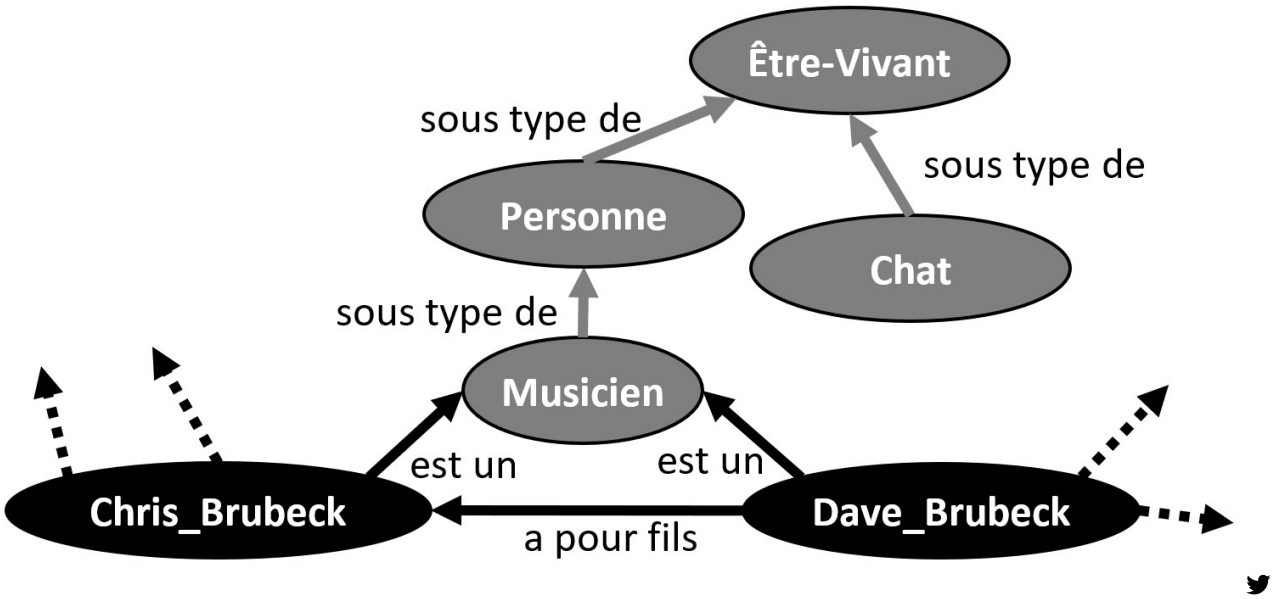
Mais la variété des graphes de connaissance concerne bien d'autres aspects de ces structures. Ils peuvent être petits comme ceux qui capturent quelques données personnelles d'un individu ou très gros comme ceux qui forment les bases de connaissances biologiques. Ils peuvent être assez statiques comme un graphe de connaissances linguistiques du Latin ou très dynamiques comme ceux produits par le réseau des capteurs d'une ville.

Les connaissances communes d'un domaine : les schémas des graphes

En tant qu'êtres humains, nous pouvons déduire de l'exemple du graphe sur la musique que deux artistes se connaissent car ils jouent dans le même groupe. Nous pouvons déduire plus de choses que ce que les arêtes du graphe indiquent explicitement parce que nous faisons appel à des connaissances générales que nous

partageons avec de nombreuses personnes. Pour un graphe plus spécialisé, ce phénomène se reproduit avec des connaissances partagées par les experts du domaine, les « connaissances de domaine ». Ces connaissances lorsqu'elles sont explicitement représentées en informatique sont appelées des « schémas » ou encore des vocabulaires ou des ontologies en fonction notamment du type de connaissances qu'ils capturent (ex. des connaissances pour valider la qualité des données vs. des connaissances pour déduire de nouvelles choses ; ou encore un lexique vs. une théorie formelle des catégories d'un domaine).

Ces schémas sont eux aussi des graphes de connaissances qui se relient aux autres, mais ils se concentrent sur des connaissances générales partagées, par exemple en indiquant que la catégorie « Musicien » est une sous-catégorie de « Personne » par une arête entre ces deux sommets, sans s'intéresser à un musicien ou une personne en particulier.



Graphe de connaissances et schéma

Les graphes de connaissances et leurs schémas sont alors utiles à diverses méthodes, notamment d'apprentissage et de raisonnement et permettent d'améliorer les réponses à nos requêtes, la classification automatique, la recherche d'incohérences, la suggestion de nouvelles connaissances, etc.

Ce sont de telles connaissances qui permettent à un moteur de recherche de capturer et de répondre, à la question « quelle est la date de naissance de Dave Brubeck ? » directement « le 6 décembre 1920 », plutôt que de vous proposer comme réponses une liste de pages du web

L'adoption d'un même schéma par plusieurs acteurs d'un domaine ou par plusieurs graphes de connaissances permet aussi à ces derniers d'être des éléments clefs dans l'intégration de données et l'intégration d'applications dans ce domaine.

La flexibilité des graphes et de leurs schémas est particulièrement importante lorsque l'on s'intéresse à découvrir des données dans un processus continuel par exemple lorsque ces données sont obtenues en parcourant le web en permanence ou lorsqu'elles sont issues de nouvelles expériences et analyses biologiques arrivant quotidiennement.

La vie rêvée d'un graphe

Les méthodes et outils de création et enrichissement de graphes de connaissances se basent sur des sources de données diverses qui peuvent aller du texte ou de la donnée brute, aux données très structurées. De plus, la flexibilité et l'extensibilité naturelle des graphes de connaissance se prête à une approche incrémentale et agile partant d'un petit graphe initial qui est progressivement enrichi à partir de sources multiples.

Ces extractions qui viennent nourrir les graphes seront généralement incomplètes ou en doublons, avec des contradictions ou même des erreurs. Un second ensemble de méthodes et outils s'intéresse à évaluer et raffiner les graphes de connaissances pour en assurer la qualité et, par répercussion, la fiabilité des applications construites au-dessus.

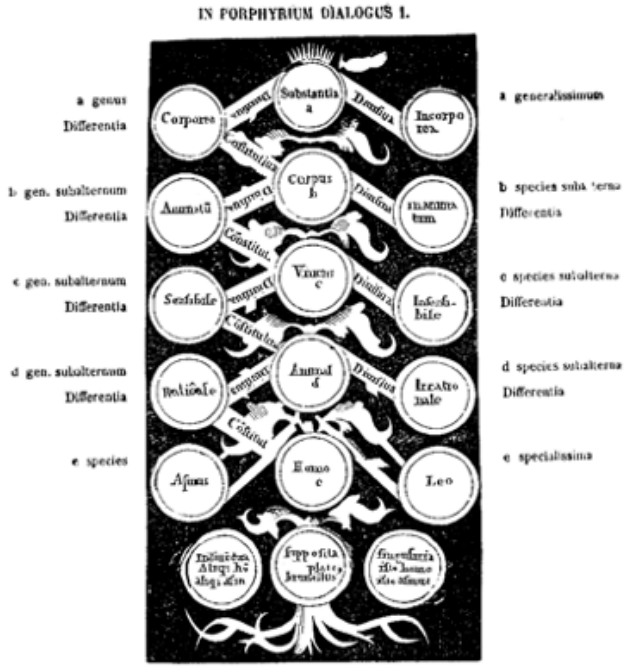
La variété des graphes de connaissances implique aussi une variété d'outils plus ou moins adaptés aux différents usages. Un outil performant pour un graphe de connaissances pourra se révéler inadapté pour un autre s'ils ont différentes caractéristiques en termes de dynamicité, de traitement ou de taille par exemple.

Outre l'extraction de connaissances qui les nourrit, les graphes de connaissance ont un autre lien particulier avec l'intelligence artificielle : ils font en effet partie des modèles de données de choix quand il s'agit de fournir les entrées ou de capturer les sorties des algorithmes que ce soit pour simuler un raisonnement ou un apprentissage. Le graphe de connaissance peut donc aussi jouer un rôle important dans l'intégration de différentes méthodes d'intelligence artificielle.

Ce double couplage de l'intelligence artificielle et des graphes de connaissance permet d'envisager un cercle vertueux ou le graphe de connaissances en entrée est suffisamment riche pour permettre des traitements intelligents et, en retour, les traitements intelligents augmentent et améliorent la qualité et l'accès au graphe. Dans l'exemple sur la musique, le graphe peut ainsi permettre en entrée d'améliorer un moteur de recherche avec des raisonnements ou de fournir des exemples pour entrainer une méthode d'apprentissage à reconnaitre un genre musical et, en retour, ces mêmes algorithmes d'intelligence artificielle peuvent nous permettre de détecter des manques ou des oublis dans le graphe et de l'améliorer par exemple en suggérant le genre d'un morceau qui manquait dans le graphe.

L'âge de graphe

Comme pour d'autres sujets en intelligence artificielle, si l'on regarde l'histoire des graphes de connaissances, plutôt que de dire qu'il s'agit d'une nouveauté on pourrait dire qu'il s'agit d'un regain d'intérêt dû à un certain nombre de progrès et d'évolutions du contexte scientifique, technique et économique.

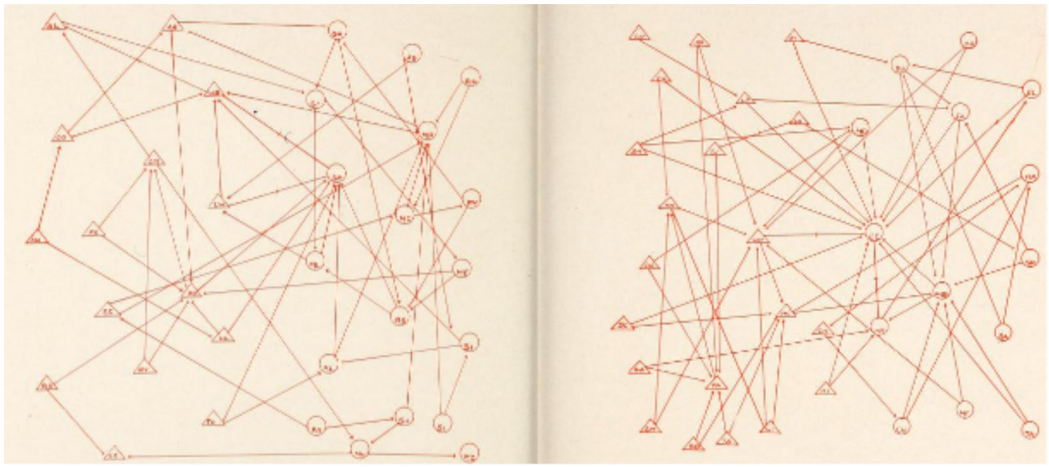


Arbre de Porphyre de Tyr pour son Introduction aux Catégories d'Aristote (vers 268) et représenté par Boèce au 6e siècle

On trouve des diagrammes de représentations de connaissances et raisonnements dès l'antiquité et, en mathématique, les graphes sont introduits et utilisés pour représenter une variété de réseaux plus ou moins complexes. Au 19e siècle, on représente des connaissances linguistiques sous forme de graphes. Au début

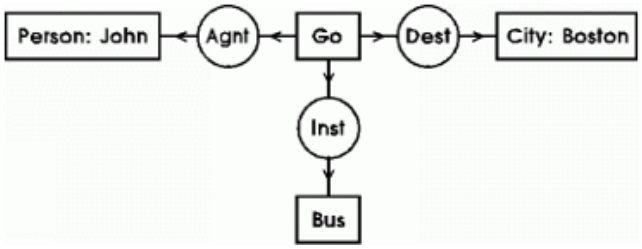
du 20e siècle, les sociogrammes capturent les connaissances sociales. Au début de la deuxième moitié du 20e siècle, les réseaux sémantiques font le lien entre modèles de mémoire humaine et représentation informatique.

Le besoin de langages de haut niveau pour gérer



Sociogrammes de J. L. Moreno dans son livre "Who Shall Survive: A New Approach to the Problem of Human Interrelations" 1934

automatiquement des données numériques indépendamment de leurs traitements et la recherche de l'indépendance aux représentations en machine vont encourager les progrès en matière de modèles de données en général et de graphes de données en particulier. Les années suivantes verront la proposition du modèle relationnel et l'émergence des bases de données, du modèle de graphe Entité-Relation, la formalisation logique des réseaux sémantiques, les modèles de *frames* et les graphes conceptuels, la programmation logique, les systèmes à base de règles et leur application aux systèmes experts et systèmes à base de connaissances, notamment sur des bases de graphes.



Exemple de Graphe Conceptuel (« John va à Boston en bus ») de John Sowa conçu dès les années 70

Dans les années 80 et 90, les langages orientés objets suivis par les représentations graphiques comme UML, mais aussi le développement des notions de schéma et d'ontologies en base de données et en représentation des connaissances renforcent encore l'indépendance des représentations et enrichissent les modèles de graphes de connaissances devenant plus modulaires et réutilisables. Le compromis entre le pouvoir expressif des modèles de représentation des connaissances et la complexité informatique de leur traitement est alors systématiquement étudié.

Le terme de *Knowledge Graph* (graphe de connaissance) apparait dans des titres de publications académiques à la fin des années 80 et au début des années 90 mais ne se répandra pas vraiment avant la deuxième décennie du siècle suivant. Internet puis le Web vont aussi augmenter à la fois le besoin et les solutions pour représenter, traiter et échanger des données. En particulier, la fin des années 90 voit le lancement au W3C (consortium de standardisation du Web) des langages standards du Web qui nous permettent maintenant de représenter, publier, interroger valider et raisonner sur des graphes de connaissances sur la toile.

Des années 2000 à nos jours, on assiste avant tout au déluge des données, notamment en termes de volume et d'hétérogénéité, suivi par le renouveau de l'intelligence artificielle nourrie par ces données. Dans ce

contexte, les graphes de connaissances apparaissent comme un moyen de relier et d'intégrer ces données et leurs métadonnées. Sur le Web, les graphes de connaissances publics apparaissent sous le terme de *Linked Data* (Données Liées). Facebook annonce son *Open Graph Protocol* en 2010 et en 2012, Google annonce un produit appelé *Knowledge Graph* après son rachat de l'entreprise Freebase quelques années avant. A ce stade, beaucoup de vieilles idées atteignent une popularité mondiale et commence alors une adoption massive des graphes de connaissances par de grandes entreprises dans tous les domaines.

On lie… un peu… beaucoup… à l'infini

Les graphes de connaissances sont donc des ressources numériques en pleine ascension, des graphes de données destinés à accumuler et à transmettre des connaissances, dont les sommets représentent des entités d'intérêt et dont les arêtes représentent leurs relations. Ils deviennent le substrat commun à beaucoup d'activités humaines et informatiques, la mémoire collective de communautés hybrides d'intelligences artificielles et naturelles. Ils ne cessent de grandir, de s'enrichir et de se relier entre eux sur virtuellement tous les sujets. Il y a donc de fortes chances que les défis et résultats des travaux sur les graphes de connaissances soient encore pour longtemps au croisement de multiples disciplines et domaines d'activité, avec un fort potentiel de retombées sociétales.

Fabien Gandon, Inria

Pour en savoir plus… vous aussi suivez les liens :

Trois références sur les différentes facettes et activités autour des graphes de connaissances :

– Hogan et al., Knowledge Graphs, 24 Jan 2021, arXiv:2003.02320
– Claudio Gutierrez and Juan F. Sequeda. 2021. Knowledge graphs. Commun. ACM 64, 3 (March 2021), 96–104. DOI: https://doi.org/10.1145/3418294
– Michel Chein et Marie-Laure Mugnier, Graph-based Knowledge Representation, 2009, Springer, ISBN 978-1-84800-286-9

Quatre références sur les graphes de connaissances sur le Web et les données liées :

– Fabien Gandon. A Survey of the First 20 Years of Research on Semantic Web and Linked Data. *Revue des Sciences et Technologies de l'Information – Série ISI : Ingénierie des Systèmes d'Information*, Lavoisier, 2018, ⟨3166/ISI.23.3-4.11-56⟩. ⟨hal-01935898⟩
– Allemang, D., Hendler, J., and Gandon, F. (2020). SemanticWeb for the Working Ontologist. ACM Books, ISBN-13: 978-1450376143
– Michael Uschold, Demystifying OWL for the Enterprise, ISBN: 9781681731278
– Fabien Gandon, Catherine Faron, Olivier Corby, Le web sémantique – Comment lier les données et les schémas sur le web ? Dunod, 2012, ISBN-13 : 978-2100572946