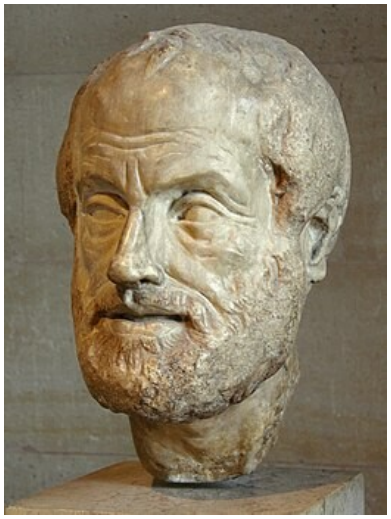
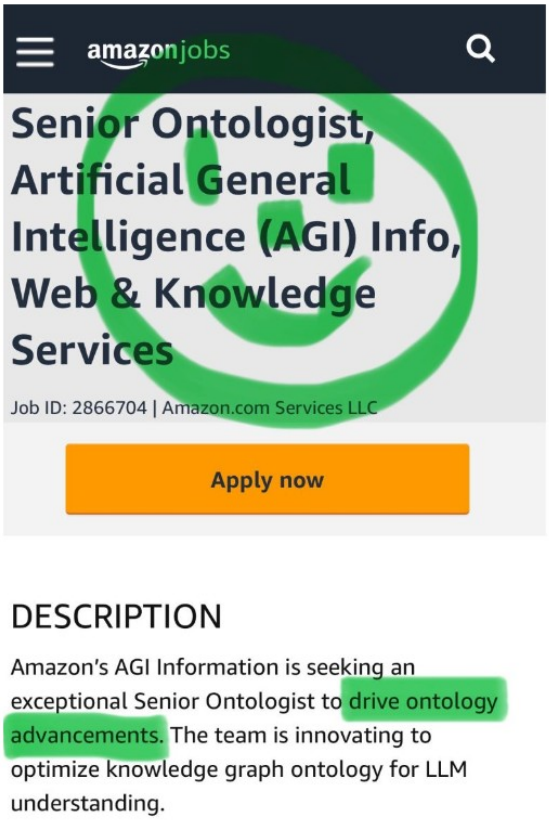
Des Idées aux Données :

Les ontologies et leurs applications, d’Aristote à Palantir





Cours M2 Traitement Automatique des Langues INALCO 2025
Louis Jourdain

L’ingénierie des connaissances, un terme à l’allure pompeusevoire technocratique, représente un champ d’application historique du traitement automatique des langues (NLP). Elle a connu son apogée entre les années 1990 et 2000 avant de sombrer dans un relatif oubli. Aujourd’hui, ce domaine est à peine mentionné dans les discussions contemporaines, sauf dans certains cercles académiques et industriels isolés. La meilleure illustration de ce désintérêt réside dans les programmes d’enseignement des grandes universités américaines – Stanford, Carnegie Mellon, NYU, MIT – qui dominent l’offre en informatique et en NLP. Aucun cours n’est spécifiquement consacré à l’ingénierie des connaissances ou aux ontologies. À Stanford, par exemple, le cours de Jure Leskovec, une figure emblématique des réseaux de neurones de graphes (GNN), n’aborde les graphes de connaissances qu’en surface, et uniquement pour montrer comment les convertir en représentations vectorielles (embeddings). Pire encore, l’histoire des ontologies et des graphes de connaissances dans le NLP, de leur émergence dans les années 1980 jusqu’à leur rôle dans les technologies modernes, n’est quasiment jamais évoquée, même dans les institutions qui ont été des pionnières dans ce domaine. Stanford, créateur du célèbre outil Protégé, illustre ce paradoxe : le dernier cours intitulé « Knowledge Representation » remonte à 2011, un signe clair du déclin de l'intérêt pour ce sujet.

Cependant, depuis 2021, un regain d'intérêt subtil mais significatif se manifeste, alimenté par les progrès spectaculaires de l’intelligence artificielle, et notamment par les modèles de langage de grande taille (LLMs). Ces derniers ont permis de dépasser l’un des principaux obstacles historiques de l’ingénierie des connaissances : le « bottleneck » de l’acquisition des connaissances. Ce goulot d’étranglement, qui rendait l’élaboration de graphes de connaissances laborieuse et coûteuse, est désormais contourné grâce à la capacité des LLMs à extraire, structurer et relier de grandes quantités d’informations issues de corpus textuels massifs.

Paradoxalement, cette redécouverte de l’ingénierie des connaissances est aussi motivée par les faiblesses des LLMs eux-mêmes. Ces modèles, bien qu'impressionnants dans leur capacité à générer du texte, souffrent de limitations majeures : leur propension aux hallucinations (fabrication de faits erronés), leur incapacité à vérifier leurs propres affirmations et leur tendance à manquer de fondement factuel. Les graphes de connaissances et les ontologies apparaissent alors comme des solutions complémentaires précieuses pour ancrer les affirmations des LLMs dans des bases factuelles solides. Ils permettent de fournir un cadre structuré et vérifiable, essentiel pour des applications critiques telles que le fact-checking, la lutte contre les biais informationnels et la validation des résultats.

Ainsi, l’ingénierie des connaissances, longtemps reléguée au rang de discipline dépassée, retrouve progressivement une pertinence dans l’écosystème technologique contemporain. À la croisée des chemins entre l’héritage des systèmes à base de règles et les promesses de l’apprentissage automatique, elle illustre comment les méthodes traditionnelles peuvent se réinventer pour répondre aux défis modernes. Ce renouveau, bien que timide, pourrait s’avérer crucial pour surmonter les limites fondamentales des représentations vectorielles (embeddings) et assurer une interaction plus fiable entre les LLMs et le monde réel.

On étudiera dans ce cours l’évolution historique des techniques de construction des bases de connaissances, des années 1980 à nos jours et étudiera leur cas d’application contemporains à l’heure où la GenAI les remet sous le feu des projecteurs.

Chapitre 1) : aux sources de l'ontologie

1) une brève histoire de la transmission de l'information

La connaissance humaine (au sens d'ensemble des savoirs constitués par l'homme dans divers domaine à partir de son expérience) a pris dans son histoire différentes formes. Cette connaissance est celle de tout une communauté, et l'enjeu principal a toujours été celui de l'accessibilité de cette connaissance de la communauté par l'individu, c'est à dire sa transmission.
La connaissance pour pouvoir être partagée doit pouvoir être :
* consignée, c'est à dire encodée et décodée.
* transmise d'un individu à un autre.
* stockée de façon sécurisée et pérenne.

a) L'encodage de la connaissance

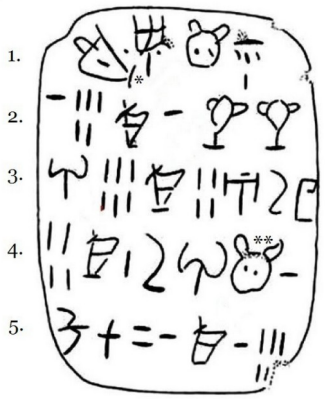
La connaissance humaine, qu'elle soit scientifique ou non, a connu de nombreux médias et supports. Avant toute chose, elle est encodée et informée par le langage. A priori on ne peut pas vraiment parler de connaissance avant son invention. La connaissance se transmettait forcément d'individu à individu (formation selon un rapport maître élève étroit dans l'antiquité, envoi de messagers…). La parole ne transmettait pas forcément uniquement des connaissances individuelles mais pouvait incarner le savoir de tout un peuple.

C'est ainsi que l'on peut interpréter la transmission des récits épiques dans l'antiquité dans différentes aires géographiques. L'exemple le plus connu sont bien sûr les deux épopées d'Homère que l'on conserve, l'Illiade et l'Odyssée. Avant d'être consignées à un moment sous forme écrite par plusieurs plumes différentes, les épopées dites homériques désignaient surtout un ensemble de pratiques de récitations orales, notamment lors d'événements importants, comme l'ont démontré Parry et Lord. Les aèdes (bardes) suivaient une longue formation incluant mémorisation des trames narratives essentielles et d'un formulaire poétique et entraînement à l'improvisation.
Or ces récits épiques illustrent non seulement des points d'« histoire » (rapport à l'Orient, importance de la civilisation minoenne…) mais également les valeurs de tout une société. Que ce soit au niveau des pratiques guerrières (descriptions des actions et du matériel très précises) des concepts sociaux (valeurs aristocratique, comme par exemple l'*arété*, la vertu héroïque, *kleos*, la gloire acquise par le fait d'arme, *timé*, la réputation sociale, *xenia*, l'hospitalité …) ou des concepts plus existentiels tel que la *tuché*, la variabilité du sort et du destin, sont illustrés et développés par les récits. Il ne serait pas, ce me semble, exagéré d'affirmer que ces épopées contiennent une ontologie du monde grec archaïque : les concepts, leur relation… sont enchâssées dans les vers de l'épopée qui était un vecteur important de la transmission des valeurs.

Un tournant majeur dans l'histoire de la connaissance humaine intervient lorsqu'elle n'a plus eu besoin d'un médiateur humain pour être transmise et que son encodage a été rendu possible dans des objets. Sans surprise, l'invention de l'écriture (environ 5000 avant J.-C.) qui s'est faite de façon parallèle à plusieurs endroits du globe est en est le tournant majeur. Dans les différentes aires géographiques (Égypte et croissant fertile, Chine), l'évolution du système d'écriture a souvent été parallèle.
Au départ, des dessins représentaient des mots (idéogrammes). Par exemple une tête de bœuf désignait un bœuf, ce qui permettait l'écriture de documents comptables simples, des registres de commerce. Si ce système à base d'idéogramme à l'avantage d'être universel et permet de s'exprimer sur les sujets quotidiens, l'expressivité de ce système de signes était très limité : la taille du vocabulaire ne peut pas être infinie (souci de mémoire) et il y a des concepts que l'on ne peut pas représenter.

cows? on a Bronze Age fresco, at Akrotiri on Santorini Island

Linear A tablet HT 118 (Haghia Triada)



Linear A Latinized:
1. ideogram for "pig" + I + madi
2. 15 + KI 10 + qaqa-
3. ru 6 + KI 4 + arisu
4. 4 + KI 1 + riruma 10
5. kuro 30 + KI 15+

Decipherment:
1. ideogram for "pig" + supersyllabogram I = ima? ἱμάς = leather strap/thong, i.e. leash? + madi = pig?
2. 15 + KI supersyllabogram for "plot of land"? kitina κτοίνα Cf. Linear B *kotona* 10 + qaqaru = a livestock animal, possibly a cow or bull or ox
3. 6 + arisu, another kind of livestock, any one of the above
4. 4 + KI = "a plot of land" + riruma = another kind of livestock, permutation of 1 of the 3 above + 10
5. total = 50 on plots of land 15

Free translation:
a leather strap to restrain a pig, 15 pigs on a plot of land, 6 cows or bulls or oxen, 4 cows or bulls or oxen (permuted from previous livestock) and 10 of the same livestock, permuted again, for a total of 50 livestock on 15 plots of land. *Et voilà. C'est assez simple.*

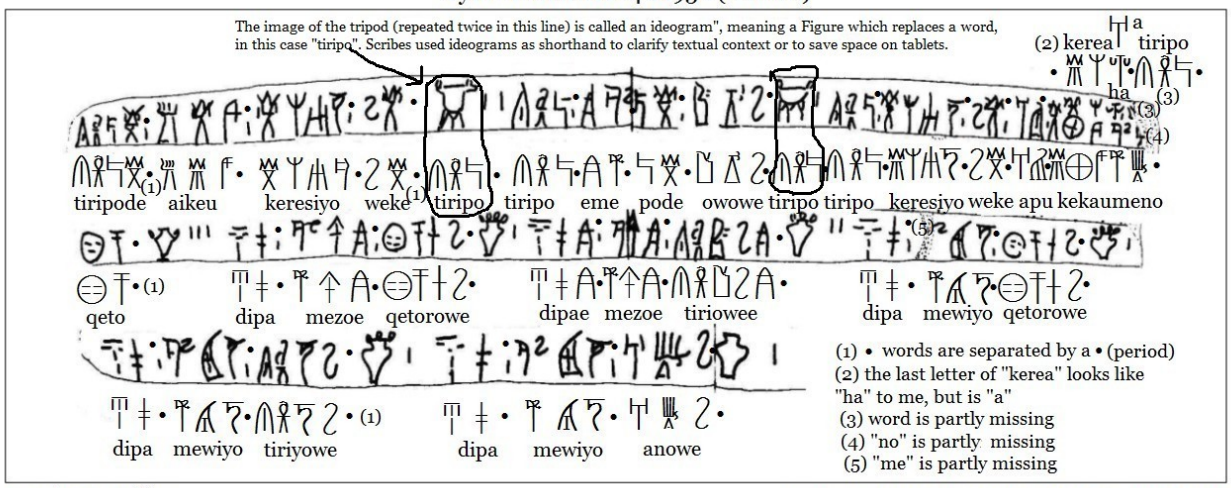
Pour pallier ces limites, certaines civilisations ont commencé à associer les symboles non plus aux idées, mais aux sons de leur langue parlée. C'est le passage aux phonogrammes, où un symbole représente un son ou une syllabe plutôt qu'un concept. Certains idéogrammes ont commencé à être utilisés pour représenter des sons associés à leur nom. Par exemple, un dessin représentant "abeille" pourrait être utilisé pour le son "a" ou "ab". Cela a permis un principe de rébus : Ce mécanisme consiste à combiner des symboles pour écrire des mots en utilisant leur valeur phonétique plutôt que leur signification initiale.

Ainsi par exemple le mot en égyptien hyéroglyphqiue nfr est écrit avec un hiéroglyphe représentant un cœur et une trachée. Ce symbole était à l'origine un idéogramme représentant l'anatomie, mais il est également utilisé comme phonogramme pour les sons **nfr**, qui composent le mot "beauté"

Cette transition de l'idéogramme au phonogramme a été successive et les deux systèmes ont souvent coexisté. Un bon exemple est les tablettes mycéniennes, découvertes dans les palais de Crète et de Grèce continentale. À l'origine tablettes d'argile dédiées à un inventaire temporaire des possessions des rois et des transactions marchandes, des incendies ont fixé leur contenu. Certaines tablettes contiennent essentiellement des idéogrammes comme celle à droite qui propose un inventaire de bétail.

Mais les mêmes signes et d'autres sont utilisés comme syllabaire pour noter d'autres mots. Regardez les mots entourés dans la tablette suivante. L'idéogramme représentant un tripode (sorte de vase à trois pieds) est retranscrit par des syllabes. Mais il est suivi dans le texte de la translittération phonétique du mot (ti ri po : τρίπος) signe que les deux systèmes coexistaient.

Pylos Tablet TA 641-1952 (Ventris)



Translation to follow

Au fil du temps, les phonogrammes se sont éloignés de leurs origines graphiques et ont évolué vers des symboles purement abstraits désignant des syllabes (akkadien et langues notées par le cunéiforme, mycénien) puis des lettres (alphabets). On peut retrouver des traces des idéogrammes dans la graphie des lettres (alpha, a, aleph et alif dérivent tous de l'idéogramme représentant une tête de bœuf). L'invention de l'alphabet par les phéniciens s'est rapidement répandue dans tout le bassin méditerranéen et a été adapté pour noter les différentes langues qui y étaient parlées.

Cette révolution alphabétique a donc réduit considérablement l'inventaire des signes à un ensemble de taille mémorisable (nécessite toutefois une formation à la littératie) mais rend possible l'encodage de n'importe quel discours, de n'importe quelle connaissance sur un format matériel avec un système d'une trentaine de signes. Un concept et sa définition est donc exprimable en quelques centaines de signes et la connaissance peut commencer à circuler (cf section suivante).

Si cette écriture (modulo la nécessité de traduction entre les langues) a permis les échanges même au long cours pendant des siècles, sa limite a été atteinte à la fin du dix neuvième siècle et l'avènement de l'époque industrielle. La vitesse de transmission était insuffisante et il a fallu trouver des moyens pour encoder l'information autrement et la faire transiter par des moyens techniques. En 1837, Samuel Morse parvient à un système de télégraphe électrique en arrivant à discrétiser les textes. Par un système d'alternance entre envoi d'un signal et pause (on/off) il met à jour un code permettant d'encoder tout l'alphabet en seulement deux signes.

L'idée de Morse trouve un prolongement conceptuel avec l'algèbre de Boole, développée par George Boole en 1854. Ce système mathématique repose sur la manipulation de variables prenant uniquement deux valeurs possibles, souvent interprétées comme 0 et 1. Lorsque l'électricité est introduite dans ce cadre conceptuel, cela mène l'électrification de la logique : Les circuits électriques peuvent être conçus pour exécuter des opérations booléennes comme ET, OU...La discrétisation des textes amorcée par le Code Morse a initié une révolution dans le traitement de l'information en introduisant le principe d'un **encodage universe**l, c'est à dire le principe de réduire l'information complexe à des unités élémentaires est le fondement des systèmes numériques actuels, comme l'encodage binaire (ASCII, Unicode, etc.).

Ainsi, chaque caractère de l'alphabet peut être noté par une série de 0 et de 1 et stockée sur 1 'bit' (0 ou 1). Il faudrait *a minima* 2^5=32 (suffisant) pour encoder l'alphabet latin, 2^6 avec les majuscule (doubler le nombre de caractères). Le système Ascii encode les caractères usuels sur 7bit alors que unicode, bien plus complet va utiliser jusqu'à 32 bits (4 octets) ce qui est normalement suffisant pour encode n'importe quel texte.

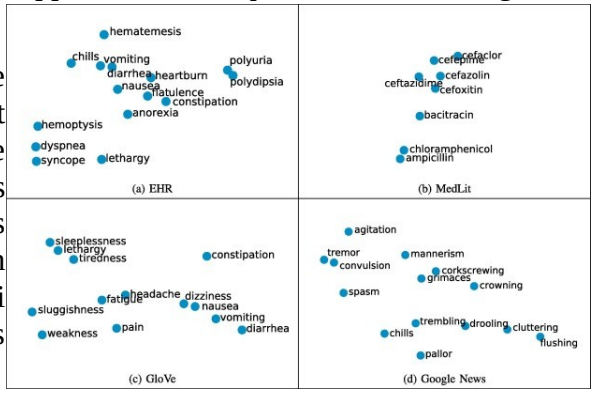
Ce qui est intéressant de constater, est que même si le moyen de stockage de l'information (sa vitesse de transmission et le pouvoir de compression) a été démultiplié, le moyen dont la connaissance humaine est stockée n'a pas changé : elle est toujours présentée sous forme de texte ! C'est une des raisons pour lesquelles les machines, qui manipulent (lisent, écrivent, transforment) des bits sont incapables (pour l'instant) de traiter le langage comme un être humain le fait : elles peuvent fournir un accès à la connaissance (indexation du texte, requêtage) mais pas l'utiliser. Elles peuvent regrouper des contenus qui « se ressemblent » (techniques de clustering, voir chapitre 6) mais pas comprendre conceptuellement pourquoi ces contenus sont liés (si ce n'est repérer la présence des mêmes mots clés, sans savoir les définir autrement que par la requête d'une définition dans une base de données).

Pour remédier à cette limitation fondamentale, des chercheurs ont cherché à dépasser cette approche strictement symbolique du langage en transformant le sens des mots en une représentation numérique. Cette transformation repose sur l'idée que les mots, au-delà de leur forme écrite, peuvent être représentés dans un espace continu où leurs significations sont traduites en vecteurs mathématiques. Ces vecteurs, calculés à partir des contextes dans lesquels les mots apparaissent, permettent de capturer les relations sémantiques entre eux. Ainsi, deux mots ayant des significations

proches, comme "roi" et "reine", seront représentés par des vecteurs également proches dans cet espace. Cette avancée a ouvert la voie à des systèmes capables de manipuler le langage avec une finesse accrue, comme les modèles de langage pré-entraînés. Ces derniers utilisent des millions, voire des milliards de paramètres pour analyser des textes à grande échelle et établir des corrélations complexes entre les mots, les phrases et les concepts qu'ils expriment. Pourtant, bien que ces systèmes semblent parfois "comprendre" le langage, ils se limitent en réalité à manipuler des statistiques et des corrélations, sans accéder à une véritable compréhension conceptuelle. Ils excellent à simuler une forme de sens, mais ils ne possèdent pas la capacité humaine de relier des informations abstraites à une expérience vécue ou à une intuition.

Ces techniques d'*embeddings* ont des limites techniques et pratiques : au départ les vecteurs représentant un mots étaient figés, un mot avait toujours la même représentation. Or un mot n'a pas toujours le même sens car le langage est par nature polysémique. Le sens d'un mot est aussi défini par son contexte d'utilisation voire son usage pratique (cf Wittgenstein, les jeux de langage). L'avènement des modèles de langues (2019) et leur capacité à créer des représentations contextuelles des mots a en partie résolu cette limitation importante. À l'heure actuelle, ces modèles peinent à offrir une couverture réellement multilingue, notamment pour les langues rares ou peu documentées. Cela s'explique en partie par le manque de ressources textuelles pour entraîner les modèles dans ces langues, ce qui limite leur capacité à comprendre et à représenter correctement leur structure et leurs nuances. De plus, ces modèles éprouvent souvent des difficultés à saisir le sens de concepts clés dans des domaines spécialisés comme la médecine, le droit, ou encore les sciences fondamentales. Dans ces disciplines, le langage est fortement codifié, et les mots ou expressions peuvent revêtir des significations précises qui diffèrent du langage courant. Par exemple, le terme "force" a une connotation bien différente en physique qu'en sociologie ou en littérature. Les modèles de langue, même entraînés sur des corpus volumineux, risquent d'interpréter ces termes de manière erronée ou simpliste s'ils n'ont pas été spécifiquement exposés à des données de haute qualité issues de ces champs disciplinaires.

Cette lacune souligne une limite structurelle : bien que les modèles actuels soient des outils d'une remarquable polyvalence, ils restent dépendants de la qualité et de la diversité des données qui leur sont fournies. Sans corpus adaptés, ils génèrent des représentations imprécises ou biaisées, notamment lorsqu'ils doivent traiter des textes hautement techniques ou dans des langues sous-représentées. Il en résulte des erreurs de compréhension ou d'interprétation qui peuvent avoir des conséquences problématiques, par exemple dans des applications critiques comme le diagnostic médical automatisé ou l'analyse juridique.



Un autre défi réside dans l'adaptation continue de ces modèles. Les connaissances humaines évoluent constamment, et le langage, lui aussi, change avec le temps. De nouvelles terminologies émergent, tandis que d'autres deviennent obsolètes. Les modèles doivent donc être mis à jour régulièrement, non seulement pour intégrer ces évolutions, mais aussi pour corriger les biais hérités de corpus d'entraînement datés ou incomplets.

Ainsi, si on veut d'un système qui soit efficace sur notre domaine de spécialité, il faut nécessairement fine-tuner le modèle ce qui implique un coût et un travail supplémentaire. Mais surtout il faut renoncer fermement à l'idée qu'un seul modèle puisse encoder avec précision tous les textes produits. La révolution des embeddings est donc un échec cuisant dans la tentative de création d'un pont entre connaissances humaine et systèmes informatiques.

Cette frontière entre manipulation et compréhension demeure l'un des plus grands défis de l'intelligence artificielle. Alors que les machines deviennent de plus en plus performantes dans leur capacité à traiter, classer et même générer du texte, la question reste ouverte : pourront-elles un jour

accéder à la signification profonde des mots et, par extension, à la connaissance humaine qu'ils véhiculent ?

b) La transmission de la connaissance

Un autre enjeu majeur a été la vitesse de transmission de la connaissance humaine. Avant l'invention de l'écriture, la transmission d'une information reposait essentiellement sur la mémoire humaine et la capacité des messagers à transporter cette information d'un point à un autre. Cette méthode était intrinsèquement lente et peu fiable, car la vitesse de transmission dépendait de plusieurs facteurs : la distance à parcourir, les moyens de locomotion disponibles (à pied, à cheval ou par bateau), et le temps nécessaire pour localiser le destinataire. Une information pouvait ainsi prendre des jours, voire des mois, pour atteindre sa destination, et le risque d'altération ou de perte du message en chemin n'était pas négligeable.

Pour remédier à ces limites, des systèmes de communication alternatifs ont été développés pour permettre une transmission plus rapide des informations sur de longues distances. Parmi les premières innovations, on peut citer l'utilisation de signaux visuels, tels que les feux de signalisation ou les miroirs pour refléter la lumière du soleil. Les civilisations de l'Antiquité, comme les Grecs et les Romains, ont également conçu des réseaux de tours de guet où des signaux de fumée ou des flammes pouvaient transmettre des messages simples, mais rapides, entre des points éloignés. Ainsi selon la légende, Clytemnestre à Mycénes aurait été informée de la chute de Troie le soir même (des centaines de kilomètres à parcourir!). Ces systèmes étaient toutefois limités en portée et dans la complexité des messages qu'ils pouvaient véhiculer, en l'occurrence un signal binaire ici. Un bit a traversé la méditerranée, et combien de bûches brûlées… Certains systèmes avaient une plus grande expressivité. Dans plusieurs régions d'Afrique, les tambours ont été utilisés comme un système de communication codée, souvent appelé *langage tambouriné*. Ce système repose sur la capacité des tambours à imiter les tonalités et les rythmes du langage parlé, en particulier dans les langues tonales. Par exemple, dans des régions comme l'Afrique de l'Ouest, où des langues comme le yoruba ou le ewe sont parlées, les tambours peuvent reproduire les inflexions de la voix humaine pour transmettre des messages complexes. Les tambours parlants, tels que les *dunun* ou les *atumpan* au Ghana, sont capables d'envoyer des messages à plusieurs kilomètres. Les tambourinaires expérimentés utilisent un langage codé pour annoncer des événements importants, avertir d'un danger imminent ou coordonner des activités communautaires. Ces messages sont souvent répétitifs pour s'assurer qu'ils sont bien compris par ceux qui les reçoivent et peuvent être transmis de villages en villages parfois sur des dizaine de kilomètres.

Le nombre d'information pouvant être transmis à la minute reste limité : on ne peut pas tambouriner plus d'une information à la fois, soit quelques phrases à la minute. L'invention de l'écriture permet à un messager de dépasser la taille de sa simple mémoire et à transmettre des informations bien plus complexes, parfois qu'il ne comprend pas du tout, comme par exemple des traités de médecine ou des plans de bataille ennemis. L'amélioration des moyens de transport a accru la circulation de l'information, mais les systèmes de transmission de texte ont montré leurs limites sur la quantité de données qu'ils pouvaient permettre de véhiculer en un temps limité.

Le passage à l'époque industrielle a radicalement transformé ces dynamiques. Avec l'invention du télégraphe au XIXe siècle, une nouvelle étape a été franchie dans la transmission de l'information. Désormais, il devenait possible de transmettre des messages presque instantanément, sur de très longues distances, en codant les informations en impulsions électriques. Le code Morse, qui alternait des signaux courts et longs (points et traits) pour représenter des lettres et des chiffres, permettait d'envoyer des messages d'une grande précision. Mais même cette technologie, révolutionnaire pour l'époque, restait limitée dans la quantité d'informations qu'elle pouvait transmettre en un temps donné (environ 50 mots par minute, nécessitait au départ un opérateur humain pour opérer la traduction, puis 200 avec l'amélioration du système).

C’est dans ce contexte que l’idée d’automatiser et de compresser davantage les informations a émergé. Avec l’avènement du téléphone, puis des systèmes numériques et des réseaux de données au XXe siècle, le volume d’informations transmissibles en un laps de temps donné a explosé. Le passage de supports physiques (comme les lettres et les livres) à des formats immatériels (signaux électriques, puis numériques) a permis de dépasser les contraintes des supports traditionnels. Aujourd’hui, nous pouvons transmettre en quelques secondes des quantités astronomiques de données, allant de simples messages texte à des images, des vidéos, et des bases de données complexes. L’émergence des ordinateurs et des réseaux numériques a marqué une accélération drastique de la transmission de données. Les premiers modems des années 1960-1970, comme ceux utilisés dans ARPANET (le prédécesseur d’Internet), avaient des vitesses limitées :

* Modem 300 bauds (années 1970) : Transmettant environ 30 caractères par seconde, cela équivalait à environ 6 mots par seconde, soit 360 mots par minute.

* Modem 56k (années 1990) : Avec une vitesse de 56 kilobits par seconde, un texte non compressé de 7 bits par caractère permettait d’envoyer environ 1000 mots par seconde, soit 60 000 mots par minute.

Les technologies modernes comme la fibre optique permettent des vitesses de transmission stupéfiantes.

* Connexion fibre 1 Gbps : Avec une vitesse de 1 gigabit par seconde, un texte compressé (environ 5 bits par caractère) peut transmettre 25 millions de caractères par seconde, soit environ 5 millions de mots par seconde. Cela représente 300 millions de mots par minute, suffisant pour transmettre l’intégralité de certains livres en quelques secondes. Les réseaux 5G, introduits récemment, offrent des vitesses comparables à la fibre optique avec des temps de latence extrêmement faibles. En théorie, la 5G : Avec des vitesses atteignant 10 Gbps, pourrait envoyer environ 3 milliards de mots par minute, rendant possible la transmission quasi instantanée de bibliothèques entière

Malgré cette révolution technologique, l’écriture demeure au cœur de la transmission de la connaissance humaine. Les progrès dans les moyens de communication n’ont fait qu’accroître l’importance de structurer et d’organiser l’information, afin qu’elle puisse être interprétée correctement par ses destinataires. Toutefois, ce processus soulève de nouvelles questions : si la vitesse de transmission est désormais quasi instantanée, comment s’assurer que les informations restent compréhensibles et pertinentes dans un flot de données toujours plus massif ?

c) Le stockage de l’information.

Au fur et à mesure que la civilisation humaine développait de nouveaux moyens d'écriture et de diffusion du savoir, la quantité d'informations disponibles a cru de manière exponentielle. Au début, le stockage de la connaissance se faisait sur des supports lourds et durables, tels que la pierre ou les tablettes d’argile, des matériaux qui, bien qu’offrant une certaine pérennité, limitaient l'accessibilité et la transmission des savoirs. L’apparition du papyrus en Égypte ancienne, puis du parchemin et du papier, a permis une réduction significative du poids et de la taille des supports, facilitant ainsi la production de livres et leur transport.

Dans le Moyen Âge, les manuscrits étaient des objets précieux, souvent réalisés à la main par des scribes dans les monastères ou dans des ateliers de copistes. Un manuscrit médiéval typique, souvent contenu dans un volume d’environ un litre, pouvait contenir environ 30 000 mots. Cette quantité pouvait varier selon la taille de la police, la mise en page, et d'autres facteurs, mais elle donne une idée de la densité d'information qu'un livre pouvait contenir à l'époque. Bien que cette forme de stockage ait permis une diffusion accrue du savoir par l'ouverture de bibliothèques et la multiplication des centres de copie, cette expansion restait limitée par la rareté des livres et la lourdeur de leur production. Chaque exemplaire devait être copié à la main, un processus long et coûteux qui restreignait l'accès au savoir à une élite éduquée et fortunée. Cet aspect de rareté et de concentration de la connaissance rappelle d’ailleurs l’histoire de la bibliothèque d’Alexandrie, un

exemple frappant de l’accumulation du savoir ancien et des pertes irréparables que peuvent causer la destruction des supports de transmission. Fondée au IIIe siècle avant J.-C., la bibliothèque d’Alexandrie était l'un des centres de connaissance les plus vastes et les plus prestigieux du monde antique, abritant des milliers de manuscrits, grattés sur des rouleaux de papyrus. Ces écrits couvraient une multitude de domaines, de la philosophie à la science, de l'histoire à la littérature. Toutefois, la bibliothèque n’a pas survécu aux aléas du temps et aux attaques, en particulier lors de l'incendie qui l’a ravagée à plusieurs reprises, et les écrits qu’elle contenait ont été pour une large part perdus à jamais. Bien que l’on ne sache pas exactement combien de textes ont été détruits, on estime que la bibliothèque pouvait abriter entre 40 000 et 400 000 rouleaux, dont une partie irremplaçable de la pensée et du savoir de l'Antiquité. Cet événement n’est pas le seul. Sur la centaine de pièces écrites par Sophocle et dont on a trouvé les titre, on en conserve que… 7.

La bibliothèque a donc longtemps été le lieu par excellence de la conservation de la connaissance humaine. La limitation matérielle et économique du livre a été une contrainte majeure à l'essaimage de la connaissance. Même si des bibliothèques prestigieuses telles que celle de Charlemagne ou celle de l'abbaye de Saint-Gall ont joué un rôle clé dans la diffusion des idées, la portée de ces connaissances restait confinée à des cercles restreints. La connaissance, bien qu'en augmentation constante, ne pouvait se répandre de manière véritablement massive, à cause de la lenteur du processus de copie et de la rareté des ressources pour produire des livres.

C'est seulement avec l'invention de l'imprimerie au XVe siècle, et surtout avec l'avènement de la révolution industrielle et de l'édition de masse, que la diffusion de la connaissance a pu prendre une véritable ampleur, facilitée par des moyens de production plus rapides et moins coûteux. Avec la multiplication des sources d’écriture, le savoir s’est démocratisé, mais s’est également retrouvé dilué dans un plus grand nombre de supports. D’où la tentative de certains de proposer des sommes structurées et hiérarchisées des connaissances humaines. Reprenant les projets antique de catalogages de la réalité (Pline) l'"Encyclopédie" de Diderot et d'Alembert au XVIIIe siècle a non seulement compilé les savoirs scientifiques et philosophiques de son époque, mais a aussi joué un rôle clé dans la propagation des idées des Lumières. Cette œuvre ambitieuse visait à démocratiser l'accès à la connaissance et à mettre en lumière la raison humaine comme source de progrès. Ce projet encyclopédique s’est étendu à d’autres pays, puis s’est spécialisé avec la rédaction d’ouvrages de référence sur certains domaines, qui se voulaient être une somme totalisante du savoir dans une discipline.

Ce projet encyclopédique reste encore d’actualité aujourd’hui à l’heure du numérique. Des projets comme Wikipedia, lancée en 2001, ont complètement redéfini le paysage de l'encodage et du partage des savoirs. Ces nouvelles formes d'encyclopédies sont en constante mise à jour et peuvent être éditées par n'importe qui, ce qui a permis une démocratisation encore plus large de l'accès à l'information. Wikipédia est devenu une sorte de base de connaissances mondiale, actualisée même si partielle.

Les révolutions de l’imprimerie, de l’industrie puis du numérique ont conduit à l’explosion des contenus textuels produits. Avec le web 2.0, l’utilisateur du web n’a plus uniquement le rôle d’utilisateur et d’accès au savoir stocké sur internet, mais a le rôle de producteur. Chaque tweet, post, mail, appel téléphonique, vidéo contient potentiellement des connaissances humaines précieuses. Par exemple, imaginez que j’ai une idée originale sur l’emploi d’ontologies pour améliorer les agents de code automatique et que, tout enthousiaste j’envoie un sms sur le sujet à un ami, cette échange verbal privé contient des éléments contribuant à ce que l’on peut appeler connaissance humaine. Mais que représente vraiment cette masse de données textuelles ?

En 2023, il y a environ 1,9 milliard de sites web actifs, mais ce chiffre ne comprend que les sites accessibles au public et indexés par les moteurs de recherche. Sur chaque site, le nombre de mots peut varier énormément. Cependant, une estimation moyenne du nombre de mots par page web pourrait se situer entre 1 000 et 2 000 mots par page (en prenant en compte la longueur des articles, blogs, descriptions de produits, etc.). Si on suppose qu'il y a 1,9 milliard de sites web, et

que chaque site a environ 100 pages, cela donne environ 190 milliards de pages web. Si chaque page contient environ 1 000 à 2 000 mots, cela signifie qu'il pourrait y avoir entre 190 billions de mots (1,9 x 10^{14} mots) et 380 billions de mots (3,8 x 10^{14} mots) stockés sur le web.

Cela représente des péta octets de données (juste pour le texte du web!). Selon google 175 zettabytes en 2022 avec les pages html et les médias (photos, vidéos). Cela représente 75 millions de machines puissantes (1 péta de mémoire, c'est énorme!) et une telle infrastructure coûterait environ 20 Milliards de dollars, plus 2 milliards d'entretien par an… Évidemment ces chiffres astronomiques contiennent de nombreuses redondances (de pages, de contenus, de savoirs… En effet un article de Britannica et de Wikipedia sur Napoléon vont contenir quasiment le même contenu… qui sera bien inférieur à une monographie spécialisée qui existe quelque part sur le web en format digital.

Comment donc peut on désormais parvenir à structurer une telle quantités d'information ? De rare et difficile à transmettre et à stocker, l'information est devenue omniprésente et inévitable. D'où le besoin de systèmes de traitement de la donnée hétérogène en masse pour organiser le magma informe de la donnée en un savoir cohérent et exploitable. C'est là le tout le rôle et le défi de l'ingénierie de la connaissance à l'époque moderne.

2) L'ingénierie de la connaissance : définition et tâches

L'ingénierie de la connaissance (ou Knowledge Engineering) est née simplement d'un besoin de communication.Elle désigne l'ensemble des pratiques, techniques et méthodologies utilisées pour construire des systèmes capables de simuler l'expertise humaine dans des domaines spécialisés Le développement de l'informatique depuis les années 70-80 a amené les informaticiens à vouloir automatiser des tâches très complexe nécessitant une importante connaissance du métier, ce que l'on a par la suite appelé des systèmes experts. Pour s'assurer que les systèmes codés correspondent bien au cas d'usage appliqué, les informaticiens ont été mis en contact avec des experts du domaine pour être guidés dans leur travail d'automatisation. Néanmoins, les informaticiens n'avaient aucune connaissance du domaine sur lesquels ils devaient travailler.Ce manque de compréhension a conduit à des erreurs de conception, des incompréhensions entre les informaticiens et les experts, et à des systèmes qui ne correspondaient pas parfaitement aux besoins réels des utilisateurs finaux. D'autre part, les experts dans leur domaine ne comprenaient pas non plus vraiment ce que faisaient les informaticiens et comment exactement cela pourra leur être utile. Pour surmonter ces obstacles, l'ingénierie de la connaissance a évolué pour intégrer les experts du domaine directement dans le processus de développement des systèmes, en leur demandant de formaliser leur savoir et de le traduire en un format compréhensible et manipulable par les machines. Cela a permis de créer des passerelles entre les deux mondes, celui des informaticiens et celui des spécialistes du domaine, et de concevoir des systèmes experts mieux adaptés aux réalités du terrain.

a) définir et clarifier

Derrière ce titre un peu pompeux, et surtout avec l'explosion de la quantité de données numérique, il est nécessaire de définir plus précisément ce que l'on entend par « connaissance » qui est censée être l'apanage de ce mystérieux ingénieur que l'on décrit. Ce concept doit d'être distingués d'autre assez proches.

Le dictionnaire de Cambridge donne la définition suivante [1]:

1 https://dictionary.cambridge.org/dictionary/english/knowledge

Knowlege : understanding of or information about a subject that you get by experience or study, either known by one person or by people generally. the state of knowing about or being familiar with something.
Her knowledge of English grammar is very extensive.
He has a limited knowledge of French.
The details of the scandal are now common knowledge (= familiar to most people).
In this town there are only a couple of restaurants that to my knowledge (= judging from my personal experience and information) serve good food.
C'est un nom déverbal du moyen anglais knowlechen ("to find out, acknowledge").
Même schéma étymologique pour le mot « connaissance » qui est la meilleure traduction française possible de ce terme.

Là où la « connaissance » peut en français désigner aussi bien le processus d'acquisition (« faire connaissance », la démarche y menant, et aussi la faculté de connaître de façon plus générale, le terme anglais désigne plutôt le résultat, à savoir l'état interne qui se produit quand on sait quelque chose («j'en ai connaissance »), voir substantivé « la connaissance » « une connaissance » « some knowledge » désigne un savoir individualisé. La source de cette connaissance peut être très diverse, être issue de l'expérience subjective, la transmission depuis autrui, la familiarité avec un fait, une connaissance technique. « la connaissance » désigne un fait prouvé, justifiable par opposition à l'opinion. On peut parler de « connaissance propositionnelle ».
Ce mot qui est pourtant à la base de notre vocabulaire philosophique et technique est fortement polysémique. Par exemple, en grec ancien il ne recouvre pas moins de 4 mots différents :
epistēmē (connaissance théorique, des choses qui ne changent jamais, sciences dures), technē (savoir techniques), mētis (connaissance stratégique, habilité décisionnelle), et gnōsis (connaissance personnelle, lié à la capacité de penser).Mais de laquelle de ces formes de connaissance notre ingénierie doit-elle se soucier ? Il convient de définir deux autres termes pour faire émerger par contraste le sens de ce mot.

La connaissance, globale s'oppose à l'information qui est plus unitaire.L'information désigne la représentation de faits ou de valeurs organisées d'une manière qui donne sens et utilité. Par exemple, dans une base de données, l'information se compose de données structurées et bien définies qui peuvent être utilisées pour effectuer des calculs, des analyses ou des prises de décision.
Dans certains contextes, information peut être synonyme de « connaissance » et savoir et désigner la compréhension ou à la conscience qu'un individu peut avoir à propos d'un sujet, acquise par l'apprentissage, l'expérience ou la recherche. Par exemple, connaître les causes et les effets du réchauffement climatique constitue une forme d'information au sens de savoir.

L''information désigne un message ou un contenu qui est transmis d'un émetteur à un récepteur dans le cadre d'une communication. Cela peut inclure des messages verbaux, écrits, audio-visuels, ou même des signaux non verbaux. Le but est de transmettre un sens, une instruction ou une donnée entre les parties. On peut aussi comprendre l'information comme réduction de l'incertitude : Cela renvoie à la capacité d'une donnée ou d'un message à réduire les options possibles ou à éclairer un point précis. Plus un message ou une donnée réduit l'incertitude sur un sujet, plus l'information qu'il apporte est considérée comme précieuse.

C'est surtout Shannon qui a déduit ce dernier sens du troisième. Dans le cadre de la théorie de l'information développée par Claude Shannon dans les années 1940, l'information est vue sous un angle beaucoup plus formel et mathématique. Selon Shannon, l'information est la réduction de l'incertitude ou l'élimination des possibilités inconnues dans un système donné. Son objectif n'était pas de comprendre l'information en tant que connaissance ou contenu significatif, mais plutôt de quantifier et de mesurer le flux d'information entre un émetteur et un récepteur dans un canal de communication. Par exemple, lorsqu'une question a plusieurs réponses possibles, recevoir la bonne réponse réduit l'incertitude à propos de cette question. Shannon l'exprime par la probabilité de l'occurrence d'un événement : plus l'événement est probable, moins il apporte d'information. À

l'inverse, plus un événement est improbable, plus il apporte de l'information lorsqu'il se produit. Shannon a introduit le concept d'entropie, qui mesure le degré d'incertitude ou de surprise associé à un ensemble de messages possibles. Plus l'entropie est élevée, plus il y a d'incertitude dans les choix disponibles. Par exemple, un tirage au sort avec des résultats totalement aléatoires (tous les résultats étant également probables) génère plus d'incertitude (et donc plus d'entropie) qu'un tirage avec des résultats prévisibles. Formellement, l'entropie H(X) d'une variable aléatoire X avec n événements possibles est donnée par la formule :

$H(X)=-i=1\sum npilog2pi$ où pi est la probabilité d'occurrence de chaque événement.

On peut ainsi définir la quantité maximale d'information qui peut transiter dans un canal en un temps donné. Pour Shannon, l'information n'est pas définie par sa signification ou sa pertinence, mais par sa capacité à réduire l'incertitude dans un système donné. C'est une définition dont on voit directement l'application dans des domaines techniques.

Le contenu de ce qui transite dans les canaux de communication est souvent appelé « donnée ». Un**e donnée** peut être définie comme une représentation brute d'un fait, d'une observation ou d'une mesure, sans contexte, interprétation ou signification ajoutée. Les données sont souvent des éléments isolés, comme des chiffres, des textes ou des symboles, qui par eux-mêmes ne produisent pas de sens immédiat. Elles constituent la matière première à partir de laquelle l'information peut être extraite lorsque des relations ou des structures sont identifiées, permettant ainsi leur interprétation.

Le passage de la **donnée** à l'**information** se fait lorsqu'une certaine organisation ou structuration des données permet d'en extraire un sens. L'information, par définition, est une donnée qui a été traitée et contextuellement enrichie. Cela signifie que les données, une fois traitées, sont organisées de manière à ce qu'elles aient une utilité ou une pertinence particulière pour un individu ou un système. Par exemple, une température mesurée à un instant donné est une donnée brute, mais si cette température est interprétée comme étant supérieure ou inférieure à une norme saisonnière, elle devient une information utile. L'information permet de réduire l'incertitude ou de répondre à une question précise, alors que les données seules, sans contexte, restent incompréhensibles.

Le passage de l'**information** à l'**intelligence** repose sur l'aptitude à utiliser cette information pour prendre des décisions, résoudre des problèmes ou générer de nouvelles connaissances. L'intelligence, dans ce cadre, n'est pas simplement une accumulation d'informations, mais la capacité à **analyser**, **comprendre** et **agir** en fonction de celles-ci. Elle implique des processus cognitifs complexes, souvent automatisés dans le cas des systèmes informatiques, où l'information est traitée à travers des algorithmes pour produire des résultats utiles, comme des prédictions, des recommandations ou des solutions.

La différence entre **information** et **connaissance** réside principalement dans leur niveau de complexité et leur rôle dans le processus cognitif et décisionnel.

L'**information** est généralement définie comme des données traitées, organisées et contextualisées de manière à avoir une signification ou une pertinence. Elle réduit l'incertitude sur un sujet spécifique et permet de répondre à des questions ou de remplir des fonctions précises. Par exemple, le rapport entre la prise d'un médicament et l'apparition d'une pathologie secondaire. Elles sont structurées, mais elles ne contiennent pas nécessairement de savoirs ou d'expériences personnelles qui expliquent leur signification ou leur application.

La **connaissance**, en revanche, va plus loin. Elle représente une **compréhension approfondie** de l'information acquise et son **intégration dans un cadre plus large**. La connaissance est le résultat d'un processus cognitif où l'information est non seulement comprise, mais également analysée, interprétée et mise en relation avec d'autres informations et expériences antérieures. Elle permet de faire des connexions, d'expliquer des phénomènes et de prendre des

décisions éclairées. Par exemple, comprendre pourquoi un certain produit rencontre des difficultés à se vendre sur le marché requiert une connaissance des tendances économiques, du comportement des consommateurs et des stratégies de marketing. La connaissance implique un niveau plus élevé de réflexion, de discernement et de jugement par rapport à l'information brute.

En résumé, l'information est un élément de base, structuré et souvent factuel, tandis que la connaissance est le **produit de la réflexion** et de l'expérience qui permet d'agir, de comprendre et d'expliquer. L'information peut être acquise de manière passive, mais la connaissance nécessite une **activité cognitive**, un processus de compréhension qui va au-delà de la simple accumulation d'informations.[2]

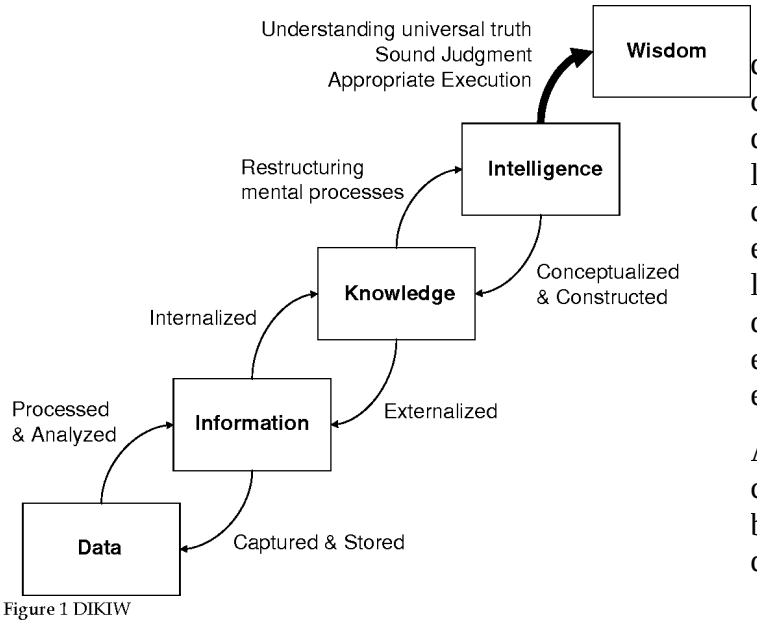


Figure 1 DIKIW

Le travail de l'ingénieur de la connaissance, est donc de transformer l'information en connaissance exploitable pour la prise de décision. À l'heure du big data, cela inclut toute la chaîne allant de la réception du flux de données, de leur connexion pertinente pour extraire l'information (voir chapitre 3 sur la RI), l'intégration de cette information dans une base de connaissance, puis rendre cette base exploitable pour la prise de décision humaine et/ou machine.

Au XXI ième siècle, l'ingénieur de la connaissance ne se contente plus de construire de belles ontologies, mais il doit transformer la donnée en intelligence.

b) l'ingénieur de la connaissance : son rôle et son évolution

Les tâches de l'ingénieur de la connaissance suivent donc la lente conversion de la donnée en intelligence.

1. Collecte de la connaissance : Cela implique d'interroger les experts du domaine pour recueillir leur savoir tacite et explicite. Les informaticiens doivent poser les bonnes questions et avoir les bons outils pour capturer toutes les informations pertinentes, qu'il s'agisse de règles, de procédures, d'exemples ou de bonnes pratiques.
2. Formalisation de la connaissance : Une fois la connaissance collectée, elle doit être transformée en une forme qui peut être utilisée par un système informatique. Cela peut inclure la modélisation de la connaissance sous forme de règles, de réseaux sémantiques ou d'ontologies. L'objectif est de rendre la connaissance manipulable, compréhensible et exploitée par un programme informatique.
3. Représentation de la connaissance : Cette tâche consiste à choisir la meilleure structure pour représenter la connaissance dans le système. Il peut s'agir de bases de données, de graphes, de systèmes de règles ou de modèles probabilistes, selon le type de connaissance et l'application.
4. Acquisition et mise à jour continue de la connaissance : Les domaines d'expertise sont en constante évolution. Il est donc crucial que les systèmes informatiques puissent intégrer de

2 https://www.semanticscholar.org/paper/DIKIW%3A-Data%2C-Information%2C-Knowledge%2C-Intelligence%2C-Liew/695f73fef84353bcec7cb66c0683f582522e18e2

nouvelles informations, ajuster leurs modèles de représentation et mettre à jour les règles ou données qui les gouvernent.

5. Validation et évaluation des systèmes : Un aspect fondamental de l'ingénierie de la connaissance est de s'assurer que les systèmes produits sont conformes aux attentes des experts et qu'ils sont capables de résoudre correctement les problèmes qui leur sont soumis. Cela passe par des tests réguliers, des ajustements et des validations en conditions réelles.
6. Gestion de la connaissance : Au-delà de la création de systèmes experts, l'ingénierie de la connaissance englobe également la gestion de la connaissance organisationnelle. Cela comprend la mise en place de processus pour stocker, organiser et retrouver efficacement la connaissance au sein d'une organisation.

Néanmoins, l'explosion du big data et l'arrivée de l'IA ont complètement chamboulé la discipline[3]. Et le chamboulement est encore en cours.

L'ingénierie des connaissances va bientôt avoir 50 ans, et la discipline, jamais vraiment bien définie, n'a cessé de se transformer.Une grande distinction est déjà faite entre le travail nécessaire pour représenter la connaissance et la stocker, et celui qui est nécessaire pour l'employer à des fins concrètes (optimisation du temps, automatisation de tâches…). Une division fondamentale était placée entre la description et l'action. Cette frontière s'est fragilisée avec l'évolution des rôles : l'ingénierie des connaissances devient de plus en plus technique. Des titres aux fonctions floues comme le « data scientists » peuvent la prendre en charge.

L'ingénierie des connaissances a également reçu quelques critiques : les systèmes experts conçus dans les premières années étaient… trop experts et pas généralisables (par exemple d'une châine de production de voiture à une autre). D'où une remise en cause de la discipline dans les années 2000 qui a conduit à l'émergence de nouveaux formalismes. Ce formalisme et cette mise en règle ont apporté encore plus de technicité au domaine qui, au départ à mi chemin entre sciences informatiques et sciences humaines s'est fait lentement happer par l'ingénierie[4].

L'arrivée de l'IA modifie encore la donne au moment ou l'utilité de l'ingénierie des connaissances, qui n'a jamais eu de fondations théoriques claires, est remise en questions au milieu de l'émergence de nouvelles technologies aux applications plus directes. La définition de l'Ingénierie des connaissances ressemble au premier des trois stades de l'évolution de l'IA prévus par la DARPA : "handcrafted knowledge" la connaissance humaine est transmise à la machine sous forme de règles directement applicables à un cas pratique. La seconde correspond au "statistical learning" : on convertit des objets en des vecteurs par feature engineering avant d'appliquer des modèles simples. Ces modèles efficaces manquent toutefois de pouvoir de généralisation et sont très sensibles au bruit à cause du manque de compréhension contextuelle. La troisième vague de l'IA correspond est donc "contextual adaption". On a eu BERT puis les LLMs.
Mais force est de constater que cette adaptation au contexte pose encore d'énormes difficultés sur les domaines de spécialité. Est-ce que l'ingénierie des connaissances pourrait faire un come-back et se proposer comme solution pour adapter les systèmes de GenAI aux domaines de spécialité ?

c) Application actuelles de l'ingénierie des connaissances.

L'ingénierie des connaissances joue un rôle crucial dans divers domaines, allant de la médecine à la justice en passant par le divertissement et la sécurité. Voici quelques exemples domaine d'applications marquantes de cette discipline :

3 https://www.sciencedirect.com/science/article/abs/pii/S1474034623000204

4*« Studer et al. describe the shift from a transfer process of human knowledge into a knowledge base to a more systematic modelling approach and thus turning the construction of KBS from an art into an engineering discipline – the so-called knowledge engineering »*

Le domaine médical ou la médecine assistée par ordinateur

La médecine, avec ces mots valises de 40 caractères, le charme de ses molécules et agents pathogènes divers est le domaine de spécialité par excellence, et un champs bien investi par l'IA. C'est sur ce domaine particulier que portent beaucoup de travaux sur les ontologies, comme par exemple SNOMED.

Dans les années 1970, le système MYCIN a marqué l'histoire en tant que l'un des premiers systèmes experts médicaux. Développé à l'Université Stanford, il avait pour objectif d'aider les médecins à diagnostiquer les infections bactériennes et à recommander des traitements antibiotiques adaptés. MYCIN fonctionnait grâce à une base de règles, dans laquelle des connaissances médicales étaient formalisées sous forme de "si... alors". Par exemple, il pouvait déduire qu'un patient souffrant de fièvre, d'éruptions cutanées et d'un certain profil d'antibiogramme avait probablement une infection streptococcique. L'un des aspects novateurs de MYCIN était son mécanisme d'explication, permettant au médecin de comprendre pourquoi une recommandation était formulée. Bien que MYCIN n'ait jamais été utilisé en pratique clinique en raison de limitations réglementaires, il a ouvert la voie à une nouvelle ère de systèmes d'aide à la décision médicale.

Aujourd'hui « boosté par l'IA », IBM Watson Health représente une avancée significative. Exploitant des bases de connaissances médicales combinées à des capacités d'apprentissage automatique, Watson Health est conçu pour analyser des millions de données cliniques, des essais cliniques aux publications scientifiques, et fournir des recommandations personnalisées pour les patients. Par exemple, dans le traitement du cancer, Watson Health peut suggérer des protocoles thérapeutiques adaptés à chaque patient en fonction de leurs données génétiques, des caractéristiques de leur maladie et des options de traitement disponibles. L'objectif est de compléter l'expertise humaine en permettant aux médecins d'accéder à une vue d'ensemble des connaissances médicales actualisées, réduisant ainsi le risque d'erreurs et augmentant les chances de succès des traitements.

le soutien juridique par les bases de connaissances

Le domaine juridique contient aussi son nombre de concepts et de règles bien particulier. Dans le domaine juridique, Lexis Nexis est une plateforme incontournable, utilisée par les avocats et les chercheurs juridiques du monde entier. Grâce à une gigantesque base de données juridiques, Lexis Nexis permet d'accéder rapidement à des décisions de justice, des législations, des contrats et d'autres documents clés. Mais ce qui distingue réellement cet outil, c'est son utilisation de l'ingénierie des connaissances pour faciliter la recherche et l'analyse. Par exemple, en utilisant des algorithmes sophistiqués de traitement du langage naturel, Lexis Nexis est capable de suggérer des précédents juridiques pertinents pour un cas donné, même lorsque les termes de recherche ne correspondent pas directement. De plus, la plateforme aide les utilisateurs à identifier des relations entre différents documents juridiques, offrant ainsi un aperçu stratégique des cas complexes.

Les systèmes de décision dans les jeux vidéo

L'ingénierie des connaissances joue également un rôle majeur dans les jeux vidéo, en rendant les personnages non-joueurs (PNJ) plus intelligents et interactifs. Les jeux modernes utilisent des systèmes basés sur des règles ou des réseaux de connaissances pour modéliser les comportements des PNJ. Par exemple, dans un jeu de rôle, un PNJ pourrait décider d'attaquer, de se défendre ou de fuir en fonction de son environnement, de sa santé et des actions du joueur. Des titres comme *The Elder Scrolls* ou *The Witcher* exploitent des systèmes de décision avancés pour créer des mondes immersifs où chaque personnage agit de manière cohérente avec ses objectifs et son rôle. Cela enrichit l'expérience du joueur, en introduisant des défis dynamiques et en renforçant l'impression que le monde virtuel est vivant.

Le renseignement : la connaissance au service de la sécurité

Dans le domaine du renseignement, l'ingénierie des connaissances est un levier stratégique pour analyser et interpréter des masses colossales de données. Les agences de renseignement utilisent des bases de connaissances pour structurer des informations provenant de multiples sources, comme les réseaux sociaux, les communications interceptées et les bases de données ouvertes. Par exemple, un outil d'analyse basé sur des connaissances peut identifier des liens entre des individus, des lieux et des événements, facilitant ainsi la détection de réseaux criminels ou terroristes. Ces systèmes sont souvent couplés à des algorithmes d'apprentissage automatique pour détecter des schémas cachés ou anticiper des menaces. Le défi réside dans la capacité à intégrer et interpréter des informations souvent incomplètes ou contradictoires, tout en garantissant la confidentialité et l'éthique des opérations.

L'entreprise américaine Palantir, basée à Denver Colorado est le parfait exemple de l'intégration des techniques d'ingénierie des connaissances au service du renseignement. Palantir Technologies intègre les ontologies comme une composante essentielle de sa plateforme Foundry, où elles jouent un rôle central dans la transformation de données brutes en connaissances exploitables. L'ontologie dans Foundry n'est pas qu'une simple base de données relationnelle ou une cartographie statique ; elle constitue une couche sémantique qui relie les données, les modèles et les concepts du monde réel en une représentation dynamique et contextuelle.Dans cette plateforme, les données brutes, souvent issues de sources hétérogènes, sont organisées en entités conceptuelles telles que des équipements industriels, des commandes clients ou des transactions financières. Ces entités sont enrichies par leurs propriétés, qui décrivent leurs caractéristiques spécifiques, et par les relations qui les lient à d'autres entités. Par exemple, une commande peut être associée à un client, et un client peut être lié à plusieurs sites d'opérations. Grâce à cette modélisation, Foundry permet non seulement de structurer les données, mais aussi de représenter les interactions complexes qui définissent des processus opérationnels. Ce cadre ontologique est également conçu pour intégrer des actions et des fonctions opérationnelles, ce qui signifie qu'il ne s'agit pas seulement de représenter des états statiques, mais aussi de modéliser des dynamiques comme des flux de travail, des décisions ou des processus automatisés. Par exemple, dans un contexte industriel, l'ontologie peut refléter le cycle complet d'une chaîne d'approvisionnement, en montrant comment un retard dans une livraison pourrait impacter un site de production ou les niveaux d'inventaire d'un entrepôt. Ce type de modélisation permet aux organisations d'identifier rapidement les points faibles, de simuler des scénarios, et de réagir en temps réel à des événements imprévus.

3) L’obsession de saisir le monde.

Une ontologie représente une représentation structurée du savoir humain. En ingénierie des connaissances, les ontologies sont notamment créées pour pouvoir être manipulées par la machine et servir à l’automatisation de processus. Il s’agit de trouver la meilleure manière de décrire les êtres, les choses et leurs interactions, ce qui a été aussi le sujet de toute une branche de la philosophie.

1) de l’ontologie aux ontologies

L’ontologie désigne la branche de la philosophie qui tente de répondre à la question « qu’est ce que l’être ». Si le terme en lui même n’appraît que bien plus tard, au XVII ième siècle, les préoccupations autour de l’être et ses propriétés ont structuré la pensée classique, et sont présents déjà dans les aphorismes pré socratiques.

C’est Aristote, le premier, dans ses livres sur la métaphysique[5] que ces questionnements sur l’être prennent un cadre formel plus rigoureux, et très intéressant. Dans le passage connu Γ, 1 de la Métaphysique. Aristote définit l**a « science de l’être en tant qu’être » :**
Il y a une science qui étudie l’être en tant qu’être et les attributs qui lui appartiennent essentiellement. Elle ne se confond avec aucune des science dites particulières, car aucune de ces autres sciences ne considère en général l’être en tant qu’être, mais découpant une certaine partie de l’être, c’est seulement de cette partie qu’elles étudient l’attribut essentiel ; tel est le cas des sciences mathématiques. Mais puisque nous recherchons les principes premiers et les causes les plus élevées, il est évident qu’il existe nécessairement quelque réalité à laquelle ces principes et ces causes appartiennent en vertu de sa nature propre. Si donc les philosophes qui recherchaient les éléments des êtres recherchaient ces mêmes principes, il en résulte nécessairement que les éléments de l’être sont éléments de l’être non pas en tant qu’accident, mais en tant qu’être. C’est pourquoi nous devons aussi appréhender les causes premières de l’être en tant qu’être.
Aristote tente de définir l’être et ses propriétés : en quoi une chose se sépare t’elle d’une autre ? Quelles sont les propriétés ? Ses attributs essentiel et non ses accidents ?
C’est dans un autre traité, les catégories (cf κατηγορία (katêgoria) « qualité attribuée à un objet » par un jugement cognitif), qu’il développe des aspects particulièrement intéressants de sa philosophie sur l’ontologie, la logique et le langage :
* Le traité s’ouvre sur une réflexion sur les rapports linguistiques entre les mots (synonymie, paronymie) et dresse une distinction entre les « êtres » (entités, étants) et les prédicats, c’est à dire les affirmations que l’on peut avoir sur ces êtres (relations!).
* Il procède à la distinction importante entre genre et espèce.[6] Aristote introduit une hiérarchie conceptuelle entre genre (ce qui regroupe plusieurs êtres par des caractéristiques communes) et espèce (ce qui spécifie un sous-ensemble d’êtres partageant des propriétés spécifiques). Ce modèle d’inclusion rappelle directement les taxonomies modernes utilisées dans les ontologies computationnelles, où le "genre" correspond à des classes abstraites et l’"espèce" à des instances ou sous-classes plus spécifiques.

5 Littéralement les livres qui ont été publiés / classés après ceux qu’il a écrits sur les sciences physiques.
6 pour les genres distincts (non subordonnés entre eux), les différences également sont d'espèce distincte ; des genres rangés les uns sous les autres peuvent avoir les mêmes différences ;les genres supérieurs sont prédicats des genres inférieurs.

* Dans la suite du traité, il définit les catégories qui sont les modalités de l'être. Les catégories sont avant tout linguistiques et désignent tout ce à quoi le verbe être peut associer à un individu en grec : la substance (ou essence), la quantité, la qualité, la relation, le lieu, le temps, la position, la possession, l'action, la passion.
Les quatre premières catégories sont fondamentales et servent à décrire l'entité et la relier ou la comparer à d'autres.

Catégories	Terme grec	Latin	Question associée
Lieu, où	*pou* / πού	*ubi*	Où est-ce ?
Temps, quand	*pote* / πότε	*quando*	Quand est-ce ?
Position, état	*keisthai* / κεῖσθαι	*situs*	Dans quelle position est-il ?
Possession, avoir	*echein* / ἔχειν	*habitus*	Qu'a la chose ou la personne ?
Action, faire	*poiein* / ποιεῖν	*actio*	Que fait cette chose ?
Passion (au sens de subir)	*paschein* / πάσχειν	*passio*	Que subit la chose ?

Les six suivantes[7] désignent son état (plus accidentel). Chacune est liée à un rôle syntaxique ou à un verbe fondamental de la langue grecque.
Il n'est peut être pas exagéré d'affirmer que dans sa description des catégories, Aristote propose une première liste de relations taxonomiques fondamentales qui permettent de décrire la réalité, c'est à dire les informations essentielles concernant une entité et les liens qu'elle entretient avec d'autres.

Aristote propose dans sa pensée métaphysique, qui est solidement amarrée à sa pensée linguistique et logique ce qui est peut être le premier système formel de description de la réalité. Cette description formelle de la réalité prend sa source dans le langage : Aristote établit un pont entre la réalité ontologique et sa représentation cognitive et linguistique. Cette idée est fondamentale pour la création d'ontologies modernes[8], qui doivent décrire non seulement les entités mais aussi leurs relations et les affirmations que l'on peut formuler à leur sujet.

Durant les siècles suivants, la métaphysique a erré sur les prémices d'Aristote pour statuer de l'existence de Dieu avant de revenir proposer de nouveaux formalismes à l'époque moderne. L'époque scolastique aura toutefois vu le développement de la logique comme langage de description des choses et de leurs relations. Leibniz travailla par exemple sur une langue universelle (characteristica universalis) et une méthode logique pour représenter les concepts et leurs relations. Le siècle des lumières vient remettre ensuite l'humain et le monde physique au centre des préoccupations. On voit émerger des systèmes de classification dans les sciences dures comme par exemple le système de Linné. Si les approches phénoménologistes du XIX et du début du XX ième ont replacé l'expérience subjective au premier plan, la philosophie du cercle de Vienne et notamment les travaux de Carnap sur le développement d'une langue philosophique comme outil de représentation des concepts ont peu à peu posé les fondations de la pensée ontologique moderne.
La métaphysique, branche la plus abstraite de la philosophie a traité dans son histoire un nombre de questions fondamentales qui ont informé le développement actuel des ontologies et de l'ingénirie des connaissanes :

7 https://fr.wikipedia.org/wiki/Cat%C3%A9gories_(Aristote)
8 Après recherches, je ne suis pas le premier à proposer ce rapprochement : https://www.jfsowa.com/talks/aristo.pdf
Voir John Sowa Knowledge Representation: Logical, Philosophical, and Computational Foundations
Franz Brentano : On the Several Senses of Being in Aristotle
+ Jonathan Barnes : Aristotle (daté mais référence)

* Qu’est ce qu’un objet ?
* Qu’est ce qu’une représentation ?
* Comment distinguer l’essentiel de l’accidentel ?
* Que peut-on connaître ?

Tous ces questionnements traduisent ce désir humain de donner au monde une représentation structurée pour mieux l’appréhender (paradigme de la technique).

b) l’échec des formalismes

Mais comment est-on passé de l’ontologie, néologisme néo latin du XVIIième siècle (1606, Ogdoas Scholastica, Jacob Lorhard (Lorhardus) formé par calque sur le nom d’autres disciplines telles que la philologie et la biologie, et censé désigner la branche de la métaphysique destinée à s’ocupper de l’être et de ses propriétés, aux ontologies utilisées en Traitement Automatique des Langues ? Le glissement est double ; à la fois glissement sémantique, et chute des aspirations.

Par un passage de l’abstrait au concret, de théorie de l’être, l’ontologie a fini par pouvoir désigner « une vision du monde », l’ensemble des classes conceptuelles et matérielles des choses qui sont et leurs relations. L’ontologie est une partie d’une théorie ou représentation du monde.
Substantivée, *une* ontologie a finit par désigner une vision du monde particulière, ou la théorie d’un penseur particulier. L’usage en ce sens dans la logique pour désigner une théorie des classes (Lesniewski, début du Xxième siècle) a fini par faire rentrer ce terme dans le monde des sciences fondamentales et des sciences de l’information, jusqu’à pouvoir désigner le schéma d’organisation structuré d’un domaine particulier. La définition est restée floue depuis. En 1998, Rudi Studer, Richard Benjamins et Dieter Fensel (en) définissent une ontologie comme une « spécification formelle et explicite d’une conceptualisation partagée ». On en déduit que :
* Une ontologie reste une construction abstraite, un fait de la pensée
* Elle traite surtout de concepts plus que de choses
* Elle a pour but d’être partagée.

Le concept d’ontologie est définitivement introduit en IA par le projet ARPA (Neches et al 1991).
1 ère définition Gruber (1993) [9]:Leur but est de mettre des connaissances à disposition des utilisateurs.
Les ontologies ont donc perdu de leur pureté abstraite et sont passé de théorie de l’être à « simple » système de modélisation.

Le développement des ontologies a été de pair avec celui du web sémantique et du soit disant « web des objets ». Ce dernier a connu l’émergence de divers formalismes aujourd’hui loin des projecteurs.Avec l’émergence du **Web sémantique**, une nouvelle tentative est apparue : représenter les objets du monde sous forme de triplets RDF (Resource Description Framework). Chaque triplet se compose d’un **sujet**, d’un **prédicat**, et d’un **objet** : une manière simple et élégante de modéliser des relations. Par exemple, dans l’univers du *Seigneur des Anneaux*, la relation entre Frodo et l’Anneau Unique pourrait être exprimée ainsi :

- **Sujet** : "Frodo"
- **Prédicat** : "porte"
- **Objet** : "Anneau Unique". Encore noté (Frodo, porte, Anneau Unique)

9 An ontology is a formal, explicit specification of a shared conceptualisation. Conceptualisation refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine-readable Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group.

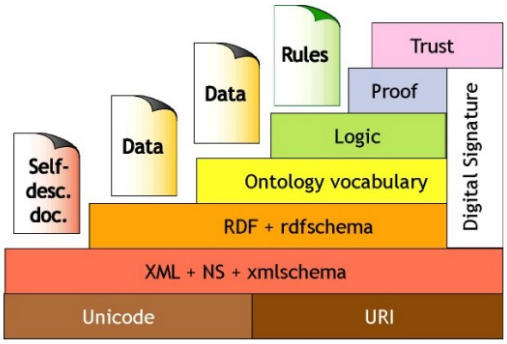
Cette structure flexible permet de modéliser une grande variété de relations. Cependant, RDF repose sur une faiblesse fondamentale : **l'interprétation des triplets est laissée à l'utilisateur final ou au système applicatif**, sans garantie d'uniformité. De plus les triplets sont un format limité pour décrire la réalité (voir chapitre 3).

Pour pallier les lacunes de RDF, le Web sémantique a introduit OWL (Web Ontology Language), basé sur les logiques de description. OWL permet de représenter des concepts, des classes, et leurs relations avec une sémantique formelle, permettant des inférences automatiques :

1. Déclaration : "Un hobbit est une créature".
2. Déclaration : "Frodo est un hobbit".
3. Inférence : Le système déduit automatiquement que "Frodo est une créature".

Cela permet de déduire des nouvelles connexions. Cependant, OWL souffre de plusieurs limites Les **logiques de description** imposent des définitions rigides : chaque concept doit être précisément défini, ce qui rend difficile la modélisation concepts polysémiques. OWL est conçu comme une **surcouche XML**, un langage technique conçu pour structurer les données mais dépourvu d'une capacité descriptive naturelle.

Le ver mis dans le fruit a été de lier trop étroitement la représentation conceptuelle choisie à l'outil technique choisie pour le noter. L'outil technique a contraint l'expressivité de la représentation et l'a condamnée. Les deux sont censés rester indépendants (le schéma vs l'implémentation).



En parallèle des efforts centrés sur le Web sémantique, l'ingénierie logicielle a développé des formalismes comme UML (Unified Modeling Language) pour modéliser des systèmes complexes. UML excelle dans la représentation des relations structurelles entre objets, notamment dans le cadre des logiciels, mais il s'avère inadéquat pour représenter des connaissances ouvertes et non déterministes. La rigidité des diagrammes UML rend difficile l'intégration des exceptions, des nuances, et des variations propres au monde réel.

Le web sémantique a donc péché par excès de formes. Sa structuration multi langages (conceptuels ET informatiques) interdépendants, qui forment une sorte de mille feuille théorique ont fini par le rendre incompréhensible pour l'homme comme pour la machine.

Ce grand échec des formalismes à représenter les objets du monde peut être attribué à plusieurs faiblesses intrinsèques :

1. **Rigidité et manque de flexibilité** : Les formalismes nécessitent des définitions précises et rigides, ce qui les rend mal adaptés aux concepts flous ou ambigus.
2. **Problème d'échelle** : Les systèmes formels peinent à capturer l'immense diversité des relations et des contextes qui caractérisent le savoir humain.
3. **Conflits d'interprétation** : Les formalismes comme RDF supposent une interprétation uniforme des relations, mais en pratique, les utilisateurs peuvent les comprendre différemment.
4. **Complexité et adoption limitée** : Des langages comme OWL et UML sont trop complexes pour une adoption large, ce qui limite leur utilisation en dehors de niches spécialisées.

Les ontologies ont toutefois su survivre et même se rendre indispensables dans des cas applicatifs précis ou des domaines spécialisés. Elles ont gagné en efficacité quand les ambitions du projet (décrire le monde) se sont réduit en terme d'ampleur comme de précision.

c) le fantasme de capturer le monde dans la machine

Quel avenir dès lors pour l'ingénierie des connaissance alors que le fantasme des formalismes a subi un échec criant et que les ontologies ne prospèrent que dans des niches de spécialité ? La complexification du formalisme pour gagner en expressivité et en nuance entraîne un coût technique d'appropriation de l'outil plus important. Si la représentation de l'objet devient moins intuitive que l'objet lui-même, il y a un problème. L'objectif est de trouver un langage commun pour faire comprendre à la machine la structure du monde, mais si l'objet décrit n'est plus perceptible ou appréhensible par l'humain, on a totalement perdu l'objectif de départ. À l'inverse, un affaiblissement du formalisme conduit à une perte d'interopérabilité et d'universalité.

Cela pourrait mener à l'abandon du projet ambitieux d'une seule structure pour décrire tous les objets du monde. La fin de toute l'entreprise ? Peut-être. Mais cela offre également l'opportunité de prendre un pas de recul par rapport au formalisme et de s'interroger sur ce qu'est véritablement un concept, ce qu'est un objet.

Il semblerait pourtant que le rêve totalisant de décrire le monde par et pour la machine n'ait pas disparu. Au contraire, ce fantasme semble prendre une nouvelle vigueur à l'ère du Big Data, où l'immense quantité d'informations collectées alimente l'idée qu'il serait possible de modéliser et de comprendre toute chose grâce à des systèmes intelligents. Ce désir s'accompagne d'une évolution majeure dans la manière dont nous communiquons avec les machines. Autrefois, les systèmes recevaient principalement des variables ou des données brutes, isolées et sans structure complexe. Désormais, ce que nous leur transmettons tend de plus en plus à ressembler à des connaissances, à des ensembles organisés et riches de relations sémantiques. Cette fusion entre donnée et connaissance redéfinit les interactions homme-machine, tout en posant de nouvelles questions sur la manière de transmettre des informations toujours plus riches, plus nuancées, et dotées d'une signification explicite.

Ce mouvement n'est pas sans lien avec la quête d'une Intelligence Artificielle Générale (AGI), capable de comprendre et de raisonner à un niveau comparable à celui de l'humain. La transmission de connaissances devient ainsi un enjeu central : comment fournir à ces systèmes non seulement des faits, mais aussi une représentation du monde suffisamment complexe pour leur permettre de naviguer dans des réalités imprévisibles ? Cette ambition évoque les bouleversements provoqués par d'autres révolutions technologiques, comme celle du cinéma. Lors des premières projections, le public fut ébahi de voir la réalité capturée et projetée sur un écran, même si cette réalité n'était qu'un point de vue fixe, figé et statique, dépouillé de sa profondeur tridimensionnelle. Ce qui paraissait alors comme une prouesse était en réalité une réduction invisible : une perte de dimensionnalité, un aplatissement de la réalité.

Aujourd'hui, une volonté englobante similaire se manifeste dans le domaine des technologies de l'information. Mais, contrairement au cinéma, ce n'est plus seulement une question de représentation : il s'agit de contrôle. Des initiatives comme Palantir, qui repose sur l'utilisation d'ontologies sophistiquées, illustrent comment la modélisation des objets du monde devient un outil stratégique, voire le nerf de la guerre du renseignement. En structurant les données sous forme de réseaux sémantiques interconnectés, ces systèmes prétendent capturer une vision totale et omniprésente, un miroir du monde utilisable pour prévoir et influencer les événements. Mais cette volonté de maîtrise soulève des questions profondes : jusqu'où peut-on aller dans cette entreprise sans perdre la complexité intrinsèque de la réalité, et qu'advient-il lorsque cette quête pour comprendre devient une arme pour contrôler ?

La définition traditionnelle de la connaissance faisait avant tout référence à un savoir humain, enraciné dans l'expérience et les compétences des individus, qui devait être transposé dans un format numérique. Ce savoir comprenait des savoir-faire de haut niveau, souvent implicites et rarement verbalisés. La construction de systèmes informatiques capables de représenter ces savoirs impliquait ainsi un travail d'explicitation : interroger l'expert, analyser ses productions et ses documents de travail, ou encore observer son activité pour comprendre comment il structure et mobilise ses connaissances. Ce n'est qu'ultérieurement que le document textuel est devenu une source privilégiée pour l'acquisition de connaissances, au point de s'imposer comme un pilier central des approches contemporaines. Les textes, en tant que « sources » de connaissance au sens premier du terme, permettent de capturer des éléments jugés stables, consensuels, et partagés dans un domaine donné. L'analyse textuelle repose ainsi sur une hypothèse fondamentale : les textes, en organisant et en structurant l'information, fournissent un point d'accès privilégié à des « contextes riches en connaissances ». La sélection des documents pertinents et leur analyse deviennent alors des étapes clés pour extraire, organiser, et formaliser des savoirs.

À l'ère du Big Data, cette ambition de créer une ontologie généralisée semble trouver un nouvel élan. Les données massives, et en particulier les données textuelles, offrent une opportunité sans précédent d'exploiter une variété infinie de contenus pour construire des modèles complexes du monde. Contrairement aux approches classiques centrées sur l'expertise humaine, le Big Data promet une extraction automatisée des connaissances à partir de corpus gigantesques, rassemblant des textes de tous genres : documents techniques, articles scientifiques, contenus web, et bien d'autres. L'idée est de dépasser les limites de la cognition humaine en s'appuyant sur des algorithmes capables de détecter des relations implicites, de repérer des modèles cachés, et d'établir des structures ontologiques à partir de ce flux ininterrompu de données.

Cependant, cette ambition repose sur un défi fondamental : comment garantir que les données textuelles, au cœur de cette entreprise, sont effectivement représentatives, fiables, et utiles pour bâtir une ontologie universelle ? Les textes ne sont pas neutres : ils reflètent des biais culturels, des interprétations contextuelles, et des perspectives limitées. Leur analyse nécessite des outils capables de décoder non seulement les significations explicites, mais aussi les nuances implicites qui échappent parfois à une simple lecture algorithmique. Si le Big Data peut offrir une quantité sans précédent d'informations, il reste à savoir si cette abondance permettra réellement de construire des ontologies pertinentes, ou si elle renforcera les limites déjà inhérentes à la donnée textuelle comme source de connaissance. Le défi est d'ordre technique : les données du Big Data sont-elles suffisantes pour entraîner des modèles au capacités de raisonnement suffisants ? (problème du deep learning) Sont-elles suffisantes pour construire un système de connaissances exploitable et fiable (problème de l'ingénieur de la connaissance).

Dans ce cours, on étudiera les différentes techniques de TAL et de data science nécessaires à la gestion du document et de son contenu (le rendre lisible, exploitable, exploité) sans omettre le problème de la captation de la donnée qui ne constitue pas a priori un fond documentaire (flux du monde). On proposera de prendre du recul par rapport à la définition classique de document et à prendre comme unité essentielle de connaissance la donnée. On verra comment les derniers outils de NLP sont devenus indispensables pour créer ces ontologies fantasmées et comment elles peuvent en retour décupler la puissance des outils de TAL. NLP, RI et ingénierie des connaissances sont en effet en relation symbiotique. Les ontologies viennent booster des systèmes de RI (indexation des documents) et de NLP (extraction d'entités), mais leur création automatique repose sur ces mêmes outils. On verra tout ça.