L'hypertexte et les sciences (1991-2021) : des voies navigables pour les routes de connaissances

Renaud Fabre, Joachim Schöpfel

L'hypertexte et les sciences (1991-2021) : des voies navigables pour les routes de connaissances

Hypertext and Science (1991-2021): A Safer Navigation on Roads towards Knowledge

Renaud Fabre

Professeur émérite à l'Université de Paris 8, Sciences Économiques

Joachim Schöpfel

Univ. Lille, ULR 4073 - GERiiCO - Groupe d'Études et de Recherche Interdisciplinaire en Information et Communication, F-59000 Lille, France

Maître de conférences en sciences de l'information et de la communication à l'Université de Lille et membre du laboratoire GERiiCO.

Mots clés

Normes hypertextes, routes de connaissances, graphes de connaissances scientifiques, sélectivité, contenus générés par les utilisateurs

Keywords

HyperText standards, routes of knowledge, scientific knowledge graphs, selectivity, user-generated contents

Chapô

Sur Internet, depuis 1991, l'adoption du langage HTML a révolutionné la représentation dynamique des connaissances, en associant une architecture d'information unique et élémentaire, aux puissantes propriétés d'assemblage et d'analyse des Graphes: les hypertextes (« graphes dont les sommets sont des textes ») offrent à la science des possibilités illimitées d'exploration de l'IST pour les articuler aux connaissances scientifiques, que structurent aujourd'hui les « Scientific Knowledge Graphs » (SKG). Les réalisations sont innombrables mais de nombreux obstacles à une navigation fiable demeurent pour bâtir et certifier les parcours de connaissances : comment passe-t-on des corrélations automatisées aux savoirs validés par les pairs ? « Search is not Research » certes, nous dit la recherche en cours, mais comment va-t-on plus loin pour offrir une navigation plus sûre sur les réseaux hypertextes ?

Résumé

Faire évoluer la connaissance a toujours nécessité de représenter et de certifier l'information scientifique disponible. Dans ce sens, une révolution inattendue survient en 1991 avec le langage fondateur d'Internet, le protocole hypertexte (HTML) : il relie en effet à l'infini les informations scientifiques, en créant des cartes et routes de connaissances à partir d'une architecture modulaire d'arcs et de sommets (hyperliens, URL) couplée à la puissance analytique massive des Graphes. Candidate idéale à la navigation hypertexte, de par la structuration de ses vocabulaires et de ses communautés aux dimensions réduites, la science en a bénéficié de façon spectaculaire : sur ces nouveaux fondements en effet, en une génération (1991-2021), ont été construites pour la première fois d'immenses bases et bibliothèques globales, bibliothèques d'Alexandrie modernes (1991-2010).

Toutefois, le développement des systèmes hypertextes s'est heurté à divers obstacles (2010-2020): massivité de l'information, imprécisions des systèmes d'assemblage et d'identification des connaissances et de décompte de leurs usages, lacunes de la philologie et de l'épistémologie de ces nouveaux leviers du savoir. La représentation stable des connaissances peine ainsi à dépasser divers conflits systémiques : comment rendre mieux traçables les interactions entre les choix documentaires et les opinons de certification ? Comment renforcer l'intégrité du processus de validation des connaissances ?

Dans un hypertexte toutefois, le choix des liens de correspondance entre documents (Hyperlinks) n'est pas aléatoire : la « distance cognitive » entre les textes peut-elle être représentée avec intégrité ? Le programme d'une recherche en cours (2021), avec la topologie des graphes les sciences du vivant les SHS et les sciences de l'information, tend à fusionner la « recommandation » du document scientifique et la « contextualisation » de ses usages, en visant des cartes et des routes de connaissances plus lisibles et une navigation plus sure. Cette double exigence étend la Science ouverte à « l'Open Process », fluidifiant l'échange entre des architectures hypertextes mobilisées pour servir l'opinion de certification des connaissances. La recherche sur les hypertextes est confrontée à un défi, à travers les fonctions d'assemblage et leurs transitions, qui n'est rien moins que d'équilibrer les langages et le message véhiculés dans le partage des connaissances.

Le programme d'une recherche en cours (2021), avec la topologie des graphes, tend à fusionner la « recommandation » du document scientifique et la « contextualisation » de ses usages, en visant des cartes et des routes de connaissances plus lisibles et une navigation plus sure. Cette double exigence étend la Science ouverte à « l'Open Process », fluidifiant l'échange entre des architectures hypertextes mobilisées pour servir l'opinion de certification des connaissances.

Abstract

To advance knowledge, it has always been necessary to represent and certify the available scientific information. In this sense, an unexpected revolution occurred in 1991 with the founding language of the Internet, the hypertext protocol (HTML): it linked scientific information to infinity, creating maps and knowledge routes based on a modular architecture of arcs and vertices (hyperlinks, URLs) coupled with the massive analytical power of graphs. Science is an ideal candidate for hypertext navigation, due to the structuring of its vocabularies and its small-scale communities, and has benefited spectacularly from it: on these new foundations, in fact, in one generation (1991-2021), immense databases and global libraries were built for the first time, modern Alexandria libraries (1991-2010).

However, the development of hypertext systems has come up against various obstacles (2010-2020): the massiveness of information, the imprecision of systems for aggregating and identifying knowledge and counting its usage, and the lack of philology and epistemology of these new knowledge levers. The stable representation of knowledge thus struggles to overcome various systemic conflicts: as the choice of links is not random,how to make the interactions between documentary choices and certification opinions more traceable? How can the integrity of the knowledge validation process be strengthened?

The current research program (2021), with graph topology, tends to merge the "recommendation" of the scientific document and the "contextualization" of its uses, aiming at more readable maps and knowledge routes and safer navigation: knowledge is a path of multimodal data, not a piece of information. This double requirement extends Open Science to the "Open Process", fluidifying the exchange between hypertext architectures mobilized to serve the opinion of knowledge certification. As choice of links is not random, a research question is to know if structured representations of knowledge in hypertext could be achieved according to integrity.

Introduction : Information numérique et connaissances scientifiques : quelles interactions ?

L'empilement menace toujours la mise en réseau des informations, nous confie Umberto Eco, et nous risquons alors de vivre le « Vertige de la Liste » (Eco, 2009) : l'information finit par déborder et par compromettre toute possibilité d'analyse...écrit-il en étudiant toutes sortes de « listes » établies depuis le Moyen Age… Comment utiliser au mieux les « listes » d'informations scientifiques numériques et comment s'assurer qu'elles font bon ménage avec les connaissances véhiculées par Internet ?

La question générale est actuelle : « *Qu'est-ce que l'instrumentation numérique a fait à la science, et de la science ces dernières années ?* » s'interroge fin 2021, la 8e Journée Nationale d'Etude du Réseau des URFIST[1].

A première vue, on perçoit le protocole HTML, HyperText Markup Language, fondateur d'Internet, comme permettant d'écrire un « graphe dont les sommets sont des textes » ou encore un « livre de livres » mu par une sorte de « grammaire générative[2] », propre à *explorer* les informations… ainsi que les informations enregistrées sur l'usage de ces informations, les « contenus générés par les utilisateurs ». … La « distance cognitive » entre les textes peut-elle être représentée avec intégrité?

Cette exploration de l'information sur Internet, aujourd'hui, peut en effet s'enregistrer sur des « routes » identifiables, navigables, celles que pratiquent les communautés documentaires utilisatrices des hypertextes*. La démarche hypertexte permet en effet de représenter ces communautés d'information savante, d'en fournir la structuration complète sur la base d'une requête, et, à partir de là, de représenter les questions posées et les routes documentaires adoptées, par les communautés de la recherche* (Fabre 2019), pourvu qu'on dispose par ailleurs des *bons vocabulaires* et *d'une pratique savante des démarches suivies pour poser des questions à partir de la documentation*.

La question de base, posée de longue date, porte sur la démarche de production et d'assemblage des connaissances : cette démarche identifie classiquement une filiation entre la connaissance à venir et la connaissance passée, filiation que relie une information scientifique validée et certifiée, toujours diverse et autorisant la construction de multiples parcours alternatifs, afin de « chercher d'autres chemins pour démontrer un théorème » car « le progrès des sciences modernes n'a pas de fin » puisque toute question résolue en stimule une autre (Godelier, 2015) et invite à demeurer critique envers le monolithisme méthodologique, ce dont la relecture de Paul Feyerabend montre l'actualité (Brown & Kidd, 2016).

Une connaissance n'étant ainsi jamais « seule au monde », l'exploration des connaissances est un processus continu, évoqué par l'image d'Isaac Newton, selon laquelle on va plus loin en voyageant sur les « épaules de géants ». La vision s'est répandue d'une science « cumulative », reposant sur une *relation structurée* entre l'information scientifique, savoir en construction (nouveaux articles, données, commentaires…), et connaissance scientifique validée… jusqu'à la prochaine réfutation apportée par un nouveau résultat sous la forme de la validation d'un nouveau savoir en construction. Thomas S. Kuhn, dans son ouvrage célèbre sur « La structure des révolutions scientifiques » observe que ce processus de transition dans les connaissances, ne va pas de soi et que souvent, pour des raisons diverses, la connaissance nouvelle peine à s'imposer à la perception savante, en dépit des validations enregistrées par les connaissances nouvelles (Kuhn, 1962). Le repère stable des savoirs validés, depuis des siècles, a été donné par la hiérarchisation documentaire des connaissances. Cette hiérarchie a

[1] Unité Régionale de Formation à l'Information Scientifique et Technique https://urfistjne2021.univ-lyon1.fr/

[2] « Ensemble de règles finies permettant de construire un nombre infini de phrases »

reposé souvent sur l'image des forêts : ordonnées ou libres, les connaissances poussent toujours en symbiose en associant les futaies et les adventices : « la » connaissance, c'est la « Sylva Sylvarum », la Forêt des Forêts, écrira Francis Bacon en 1631.

À quelles connaissances « profite » l'information scientifique ? Ce n'est pas facile à décrire, et on peut ici ou là déceler des sortes d'« aubaines documentaires » : n'advient-il pas parfois, selon le mot de Montaigne, que « Le plus haut monté a souvent plus d'honneur que de mérite » ? Comment l'information nouvelle peut-elle être délimitée véritablement dans ce contexte très étiré aux structurations multiples, et comment peut-elle nourrir la connaissance ? Par ailleurs, sommes-nous toujours les auteurs des connaissances que nous « produisons » et dans quelle mesure ? Dans l'Article de présentation du mot « Encyclopédie »[3] Denis Diderot développe avec feu l'idée que la connaissance ne peut venir « d'un homme seul ». Cette parenté implicite et/ou ou explicite des connaissances était identifiée par Diderot comme une question délicate et parfois malencontreuse : cette question est au cœur d'un autre texte moins fréquenté[4], où il ne cache pas son embarras devant les ambiguïtés éditoriales de l'Encyclopédie, dont il vient d'achever la publication (1763). À travers l'information scientifique et ses usages, la généalogie des connaissances est donc un « fil rouge » ancien et bien connu. Dans l'univers studieux des classifications savantes, l'irruption inattendue de l'hypertexte, il y a tout juste une génération (1991-2021) va apporter un bouleversement qui, avec le recul du temps, apparait à la fois gigantesque, et tout juste amorcé…

En produisant à partir de la Toile un enregistrement structuré de toute l'information scientifique disponible et de toute information induite (donnée, débat, controverse… et données générées par et sur les chercheurs-utilisateurs) l'hypertexte a révélé la science en tant que candidate idéale à ses fonctionnalités informationnelles fondatrices, comme à sa neutralité intrinsèque[5] par principe et par construction. Beaucoup d'autres domaines d'usage des hypertextes en exploitent les possibilités sous des formes très diverses, peu souvent comparées entre elles ni nourries de leurs expériences réciproques…

Le langage hypertexte offre en effet en temps réel, la possibilité d'actualiser toute controverse et de fournir une base générale et massive de validation des informations scientifiques au profit de la construction des connaissances nouvelles. Les graphes hypertexte, comme les rayons de miel des ruches, sont en effet directement exploitables et interprétables comme structures de stockage des connaissance construites : on trie désormais des « grappes » et des « grains » en analysant directement les URL pertinentes d'un graphe hypertexte, qui correspondent chacune à un document unique parmi des trillions. De plus les hypertextes sont des « rayons de miel façonnables et modulables » : on écrit et produit du graphe hypertexte, pour réagencer des contenus (articles) et leur « représentation » (Liu et al., 2021) par des architectures de documents, par des routes. Voilà donc une révolution numérique en marche pour donner au savant un outil de tri des savoirs, et d'exploration des nouveaux mondes possibles. Certes l'hypertexte ne produit à lui seul aucune science, il est neutre au résultat, mais ouvre à une « mise à plat » de l'infinité de voies parcourues et de leurs voisinages...

Alors réjouissons-nous, mais jusqu'où et de quoi exactement ?

Depuis la naissance d'Internet jusqu'à nos jours, l'histoire des mutations introduites par les hypertextes peut être ponctuée en trois phases approximatives :

[3] http://enccre.academie-sciences.fr/encyclopedie/article/v5-1249-0/ Article « Encyclopédie »
[4] https://gallica.bnf.fr/ark:/12148/bpt6k6443411c.textelmage
[5] « The intention in the design of the web was that normal links should simply be references, with no implied meaning. » Note interne W3C Tim Berners-Lee https://www.w3.org/DesignIssues/LinkLaw

« **La découverte des sommets** » (1991-2010) survient avec l'entrée de l'information scientifique et technique dans l'ère des hypertextes et de leur Graphes, au service de gigantesques réarrangements des connaissances scientifiques ;

« **Les conflits de logiques** » (2010-2020) avec l'irruption massive des hypertextes dans le travail de la science et l'apparition de conflits fonctionnels et éthiques entre les processus de représentation de l'information scientifique, de la science en train de se faire[6], et des connaissances validées qui y sont associées : où sont les frontières, comment les gérer ?

« **Les convergences possibles** » **: un hypertexte stabilisé** ? (2020-2021) Les solutions possibles des conflits, se dessinent à travers les architectures des « Graphes de Connaissances Scientifiques » (Scientific Knowledge Graphs, SKG). Ces objets hypertextes dédiés visent en effet à mettre en cohérence les contextes d'usage et les contenus de l'information (C2C) : « l'information contextualisée » tend ainsi à fournir une base intègre à toute validation savante… C'est le chemin que poursuit intensément la recherche, largement pluridisciplinaire, qui est encore loin de sa maturité[7].

En conclusion, un principe d'« Open Process » sera proposé comme une exigence pour la Science ouverte, afin, d'individualiser et partager les architectures de « routes navigables » de connaissances, pratiquées dans l'intégrité sur les hypertextes (codes, algorithmes, processus, éthique…).

A. « A la découverte des sommets » (1991-2010)

1. Information scientifique et connaissances avant Internet : une relation structurée mais discontinue

Comment avons-nous fait pour trouver l'information et y associer des connaissances, « avant » ? Les premiers dispositifs sont les bibliothèques et fonds documentaires, avec leurs outils bien à eux – les catalogues avec leurs trois index titre, auteur et sujet ; les listes des nouveaux arrivages ; puis tout simplement les collections physiques avec leurs rayonnages et travées en libre accès ou en magasin, avec leurs systèmes de classement qui permettent aussi bien une recherche précise qu'une exploration plus ouverte, sans but précis.

Une relation construite entre information scientifique et connaissances

Ces dispositifs créent des liens entre les documents, des liens physiques (proximité dans les rayonnages) aussi bien que des liens virtuels (tous les livres d'un auteur ou sur un sujet), et la navigation se fait physiquement (entre les travées, salles ou magasins) et/ou virtuellement (entre les catalogues, listes et fiches), aidée parfois et orientée par les bibliothécaires, ces « cartographes et pilotes de l'archipel des savoirs » (Baltz, 2003). En fait, « la question de l'accès à des documentations hétérogènes en évolution constante, destinées au plus grand nombre, alimente des problématiques auxquelles les bibliothèques traditionnelles sont confrontées depuis bien longtemps » (Papy, 2018).

L'étude des « systematic reviews » montre que l'appel aux experts d'un domaine est une autre option pour trouver de l'information pertinente, en particulier en ce qui concerne des ressources difficiles à identifier et à trouver ; le contact avec des « experts-clés », avec les auteurs d'actes de conférences potentiellement pertinents, avec les chercheurs les plus productifs dans le domaine ou avec les chercheurs des organisations-clés dans ce domaine est un autre moyen d'identifier les publications

[6] On y reviendra plus loin, la moitié des articles publiés, des connaissances certifiées, ne sont pas lus ni téléchargés aujourd'hui…Pourquoi ? Entropie de l'information ? Surcharge de la mémoire et des machines de tri ?

[7] La revue *Quantitative Science Studies* vient tout juste de consacrer un Special Issue aux SKG, qui sera publié fin 2021. Cf. aussi le projet DRIM - Approches Interdisciplinaires Des Systèmes Complexes https://hackmd.iscpif.fr/s/ryWU8VV4K

potentiellement pertinentes et aussi d'obtenir des informations sur toute recherche en cours ou non publiée (Schöpfel & Prost, 2021).

Des discontinuités identifiées

S'agissant de l'information scientifique, cette notion devenue « caméléonesque » (Morin, 1997) : comment distinguer information et connaissance ? Quand une « information » devient-elle « connaissance » ? Là où la « pyramide de la sagesse » fait une distinction claire entre information (déduite des données, répondant à des questions qui, quoi, comment…) et connaissance (savoir-faire, c'est ce qui permet de transformer une information en instruction) (Rowley, 2007), d'autres y voient plutôt un ensemble, un continuum : « l'information est une connaissance inscrite (enregistrée) sous forme écrite (…), orale ou audiovisuelle sur un support spatio-temporel » (Le Coadic, 1994). L'information comporte un « élément de sens », c'est une « signification transmise » (idem) qui correspond à une réduction d'incertitude (Dion, 1997).

La transition vers Internet a été analysée. Dans un essai sur les mutations du document, Jean-Michel Salaün (2012) constate que « pour le document, le changement (induit par le Web) est radical.

2. A partir d'Internet : l'apparition des hypertextes et leur adéquation aux besoins propres à la Science

Le langage et les applications hypertextes sont déjà en phase de maturité au moment où leur transposition à Internet intervient : depuis 1987, premier colloque global consacré aux hypertextes, annuellement, de nombreux articles de chercheurs et d'industriels décrivent déjà, avec souvent des accents visionnaires, cette nouvelle architecture d'information. Ci-joint, on trouvera les articles de ces colloques du passé, fort opportunément conservés et utiles à la recherche[8].

L'HyperText Markup Language (HTML), publié en 1991 au CERN par le physicien britannique Tim Berners-Lee, avait l'objectif, faut-il le rappeler, d'échanger et de publier des documents scientifiques, dans un langage (XML) et selon un format ouvert, non « propriétaire », en connectant des « adresses » de ressources (URL : Uniform Ressource Locator) à partir d'« hyperliens » (hyperlinks[9]). D'emblée on peut donc assimiler un hypertexte à un « graphe dont les sommets sont des textes », puisque les URL sont continument reliées par des hyperliens sur toute la toile, et ceci avant même que la recherche, beaucoup plus tard, autour de 2010, développe comme un objet spécifique le « Graphe de Connaissances Scientifique » (Scientific Knowledge Graph, SKG), qui va se développer en s'appuyant sur le système de connexion des URL par l'arc des hyperliens.

La Science, candidate privilégiée aux usages hypertextes

D'emblée, la science apparait comme l'une meilleures candidates, parmi les communautés d'information potentiellement utilisatrices des hypertextes avec une intégrité et une fiabilité de l'information partagée : l'article est désormais une « donnée » (Egret, 2021 [article 6 du Dossier]) et les connaissances sont des chemins de données hypertextes, comparables pour une même requête. Rappelons-le en effet, la production de la science à l'échelle mondiale émane d'une « très petite » communauté d'information (environ 4,5 millions de chercheurs publiants réguliers dans 45 000 journaux), aux hiérarchies d'acteurs transparentes réduites et structurées (grades, qualifications…), aux vocabulaires et aux supports d'information (articles) très normalisés, aux standards

[8] Memex and Beyond Web Site http://cs.brown.edu/memex/bibliography.html#104
[9] Creating Hyperlinks https://www.ironspider.ca/format_text/hyperlinks.htm

d'homologation évolutifs et précis (peer-review), et acceptant la controverse selon des procédures formalisées (colloques, thèses…).

L'explosion des usages liant URL et hyperliens : bases globales, architectures dédiées, outils.

Des bases globales sur l'essentiel de la production scientifique validée et citée (WoS, Scopus, PubMed Central, Semantic Scholar, Google Scholar etc.) sont développées et évaluées : elles font l'objet d'innombrables publications comparatives évaluant les périmètres et leurs limitations (Stahlschmidt & Stephen, 2020), cependant que les logiques d'assemblage des articles et données sur les bases, sont optimisées et « unifiées » par de multiples démarches (Van Eck & Waltman, 2010 ; Waltman et al., 2010) : sur une simple requête, l'URL de l'objet recherché va pouvoir donner accès, via les hyperliens correspondants, à l'univers des requêtes associées. L'architecture hypertexte va également s'appliquer à un très large éventail de sous-produits des compilations globales, menées sur les 45 000 journaux répertoriés dans l'édition scientifique privée et associative, mais également sur les 16000 journaux qui, à un titre ou à un autre, relèvent de l'Accès ouvert à travers le « Directory of Open Access Journals » et donnent ainsi d'ores et déjà la possibilité de consulter directement le texte intégral des publications[10]. ISTEX en France, à partir de l'INIST et selon un modèle original au plan international[11], offre accès au texte intégral et à l'exploration de l'essentiel de la littérature scientifique.

Des hypertextes d'analyse et d'interprétation des bases : l'aide à la construction des connaissances

Ces hypertextes associés aux contenus des informations scientifiques forment une constellation de bases d'analyse de la science, sur lesquelles divers traitements peuvent être pratiqués à partir d'outils dédiés (hypertools[12]). Un aperçu, qui ne peut être développé ici, fait ressortir un très large éventail de fonctionnalités complémentaires de traitement d'URL corrélées: -représentation des connaissances par les hyperliens des influences et collaborations dans un domaine (voir l'impressionnante base Paperscape mise à jour continument par l'Université Cornell pour la physique des hautes énergies[13]) ; -représentation des opinions savantes en biologie-médecine à l'échelle globale (voir la base PubFacts[14] répertoriant les liens de commentaires sur les articles exposés sur Pub Med Central) ; -la publication des informations scientifiques, qu'elles aient été ou non acceptées par les pairs et la représentation de connaissances en libre accès avec l'archive ouverte HAL[15].

B. « Les conflits de logiques » (2010-2020)

Avec l'irruption massive des hypertextes dans le travail de la science et l'apparition de courses-poursuites entre l'information scientifique disponible[16] et la connaissance validée, sont apparues diverses formes de conflits systémiques. En voici les principales formes, à défaut de produire ici une analyse structurée qui nécessite une large recherche dédiée (2010-2020).

[10] DOAJ https://doaj.org/
[11] ISTEX https://www.istex.fr/
[12] https://hypertools.readthedocs.io/en/latest/auto_examples/
[13] https://paperscape.org/
[14] https://www.pubfacts.com/
[15] https://hal.archives-ouvertes.fr/
[16] On y reviendra plus loin, la moitié des articles publiés, des connaissances certifiées, ne sont pas lus ni téléchargés aujourd'hui… Pourquoi ? Entropie de l'information ? Surcharge de la mémoire et des machines de tri ?

1. Connaissance produite et Information exploitée : une cohérence inégale

L'information scientifique apparait inégalement mobilisée, ce dont s'était inquiété l'Académie des Sciences qui avait mené une Enquête nationale assortie de nombreuses auditions (2013-2014) sur les mutations de l'Edition scientifique et sur « l'hétérogénéité » croissante des sources de la publication scientifique, en préambule au rapport commandé par la ministre de l'Enseignement supérieure et de la recherche à Jean Yves Mérindol sur ce thème (Mérindol, 2020), qui, en constatant un déclin général, préconisait éloquemment la reprise de l'investissement éditorial, aujourd'hui à l'étiage dans tous les champs de la publication (voir son article dans ce numéro).

Par ailleurs, plusieurs études récentes portent sur les usages par les chercheurs et personnels de recherche des grands outils de consultation de l'information scientifique. Les données qui s'en dégagent donnent des images souvent floues ou elliptiques, généralement négatives, de l'usage des grandes bases d'information scientifiques.

2. Connaissances produites et connaissances mobilisées : les « contenus générés par les utilisateurs » d'hypertextes

Les cartes hypertextes de connaissances et les routes documentaires qu'y empruntent les chercheurs, sont sujettes à des évolutions d'usages qu'enregistrent les moteurs de recherche : en compilant nos requêtes, ces systèmes produisent une information sur l'information que sont les « contenus générés par les utilisateurs » et dont l'analyse sémantique produit l'état des interrogations en cours de la recherche.

La science est une utilisatrice assidue de ces données, et les Éditeurs tirent de l'analyse des URL consultées et téléchargées, de nombreux sous-produits d'information scientifique pour la gestion des abonnements, comme pour l'identification des pratiques documentaires des chercheurs (Base REAXYS d'Elsevier en Chimie, ou encore l'identification massive des « Prominent Topics » dans la Base SciVal à partir des téléchargements des articles, récents ou anciens). Au CNRS, à l'INIST, des applications innovantes et fréquentées de haute qualité ont été développées pour analyser les usages de l'information scientifique[17].

La réutilisation structurée de ces données au profit de la construction d'hypertextes sur les « routes de connaissances » et leurs voies alternatives reste toutefois peu répandue, sauf exceptions (Chimie notamment) dans la recherche publique (Fabre, 2019). Ces développements nécessiteraient, il est vrai, en parallèle, des dispositifs de sécurisation des voies de réutilisation de l'information, qui, à notre connaissance, ne sont pas projetés pour l'instant.

3. Parcours de connaissances empruntées et navigation traçable dans les connaissances : vers des standards de « libre parcours » (Open Route) ?

Les « **libres parcours**[18]» à travers l'architecture des concepts et des idées publiées sont ainsi nommés tels par le droit, car le droit protège et sécurise le libre partage sans condition des idées : ce sont donc des objets de Science ouverte par excellence. Or les « libres parcours », de façon apparemment paradoxale pour des activités construites en « démarches », sont inégalement revendiqués par les

[17] https://www.inist.fr/services/acceder/ezmesure/

[18] Arrêt du 22 juin 2017 de la Cour de cassation, l'arrêt, casse un arrêt d'appel ayant condamné pour parasitisme un exploitant de propriété intellectuelle

sciences à l'heure hypertexte : la traçabilité des chemins documentaires numériques demeure en effet trop souvent incertaine et mal protégée, alors même que le droit est ouvert au libre parcours.

Des « routes navigables », ouvertes et/ou protégées seraient donc utiles, à l'heure du développement annoncé de longue date du TDM (Bellot & Grau, 2014), pour calibrer ou comparer les voies d'exploration massive des hypertextes de connaissances indispensables à la recherche (Nedellec, 2018).

4. Des « routes en construction » : exemples et témoignages

Des usages de pointe de la publication scientifique se développent partout, selon l'exemple « ancien » de la bio-informatique[19] à l'EMBL (Laboratoire Européen de Biologie Moléculaire). Les segments de « routes » d'information construites ici pour les besoins de projets de recherche, deviennent ensuite, une fois validés et partagés, les bases de « cartes » de connaissances conduisant à la construction de nouvelles routes. Ces projets prennent tous en charge une étape de construction hypertexte, mettant en présence des résultats scientifiques et des publications sur des voies de découverte en cours d'exploration (Kahsay et al., 2020 ; Ferrarotti et al., 2019 ; Griss et al., 2020).

5. La solution des conflits et les conventions d'articulation actuelles des informations et connaissances à l'heure des hypertextes : la délimitation des connaissances

L'organisation de l'information en « routes » répond à la nécessité de fournir un support documentaire solide et traçable, pour partager une vue d'ensemble des articulations jugées scientifiquement nécessaires entre des méthodologies retenues et des résultats attendus. Cette démarche a pour but final d'exposer les documents permettant au scientifique de tester la cohérence et la validité de ses hypothèses, à l'amont de l'expérimentation ou parallèlement à celle-ci, et d'offrir ainsi une vision d'ensemble à ses pairs, sur l'environnement de ses choix.

Force est de constater que ces démarches appellent des clarifications, auxquelles la recherche la plus récente (bibliométrie et recherche d'information) vient encore de réaffirmer son attachement et sa large disponibilité[20] en s'intéressant également aux contenus générés par les utilisateurs des hypertextes concernés (commentaires, associations documentaires, annotations)[21] et (Mustafa El Hadi & Timimi 2021 [article 8 du Dossier]).

C. « Les convergences possibles » : vers un hypertexte stabilisé ? (2020-2021)

Les solutions possibles des conflits précédemment évoqués (2021) se dessinent dans les évolutions actuelles des hypertextes dans les évolutions en cours des « Graphes de Connaissances Scientifiques » (Scientific Knowledge Graphs, SKG).

Plusieurs piliers d'amélioration de la fiabilité des hypertextes sont évoqués ci-après.

---

[19] https://www.ebi.ac.uk/research

[20] 10th Workshop on Bibliometric-enhanced Information Retrieval https://sites.google.com/view/bir-ws/bir-2020

[21] https://www.connectedpapers.com/main/5a00ab293237c4038b9e902adb3fce11ca9e801d/A-Searchable-Space-with-Routes-for-Querying-Scientific-Information/graph

1. Une recherche interdisciplinaire et une épistémologie de la « valeur » des textes : vers des cartes de « distance cognitive » pour orienter les routes de connaissances

Les parcours d'informations scientifiques doivent proposer des parcours de données pertinentes et interprétables. C'est un premier domaine d'approfondissements en cours, à partir des sciences de l'information. On mentionnera le travail en cours de l'équipe de Open Edition, d'Hypothèses.org [22] (l'annotation automatique des références bibliographiques) ainsi que les approches fondamentales en plein développement visent la valeur des données comme Valda[23] (« Valeur à partir de données ») et Tyrex[24] à INRIA, ou encore celle de Pablo Jensen (cf. Bouiller, 2019). Mais le lien entre textes (Hyperlien) reliant les sommets (URL) n'est pas aléatoire, c'est un choix : la recherche peut-elle nous renseigner sur la « distance cognitive » entre ces choix ? Diverses étapes ont été franchies dans ce sens : toutefois, mesurer les distances entre les textes, et le relief qui les assemble (cartes) ne saurait suffire. Encore faut-il naviguer avec ces cartes pour explorer, connaitre et faire ainsi des choix sur des « routes » de connaissances, sûres ou pas. Il faut donc pour cela représenter l'itinéraire de toutes les routes pratiquées, et les offrir toutes ensembles au regard du chercheur pour qu'il choisisse la sienne ou qu'il en invente une autre... Seul un hypertexte d'hypertextes à géométrie « pertinente et neutre » peut représenter les routes de la connaissance (Fabre, Azeroual, Bellot, Schöpfel, Egret, 2021).

La qualité épistémologique, la protection de l'intégrité, font l'objet d'approfondissements continus au W3C, sur l'hypertexte et la loi ; d'autres travaux, complémentaires, relèvent de la philologie des hypertextes et leur analyse structurale, sur laquelle se remarquent les approches peer-to-peer (Lip6). Les approches pluridisciplinaires empruntant aux sciences du vivant sont particulièrement stimulantes (dès G. Deleuze avec le rhizome, stigmergie, intelligence en essaim…). Il manque encore une extension à l'hypertexte de théories de la connaissance ouvertes aux influences indirectes issues de l'environnement (à partir de la stigmergie notamment, cf. Marsh & Onof, 2008).

La philosophie des sciences, évidement décisive pour approcher l'intégrité des hypertextes, est également en chantier sur une très large gamme de travaux stimulés notamment par la récente pandémie, mais surtout bien antérieurs. Tout près, avec le recul de toute l'œuvre anthropologique de Maurice Godelier pour saisir structures, concepts et transitions, commence à se dessiner une « Anthropologie des connaissances », à partir des savoirs synthétisés sur l'art humain « d'inventer des mondes » (Godelier 2016).

2. Le droit des parcours hypertexte : une démarche engagée

Dans la « Lettre ouverte sur le commerce de la librairie », déjà citée, Diderot[25] se pose une question qui pourrait se résumer ainsi : « Une fois que nous sommes des auteurs de connaissances, sommes-nous pour autant des « propriétaires » ? L'article 30 et l'article 38 de la Loi pour une République numérique ont apporté des réponses en cours d'application, à partir de riches travaux préparatoires, dont un Livre blanc préfigurait les bases (DIST CNRS, 2016).

Les réponses multiples apportées par les statuts éditoriaux sont-elles même partielles si l'on a l'ambition de rendre compte de la complexité des modèles de la propriété intellectuelle numérique,

[22] https://lab.hypotheses.org/
[23] https://team.inria.fr/valda/fr/
[24] https://tyrex.inria.fr/
[25] https://gallica.bnf.fr/ark:/12148/bpt6k6443411c.texteImage

modèles souvent plus audacieux et complexes que ceux de la publication proprement dite (Fabre & Bensoussan, 2017).

Le droit de propriété numérique lui-même, propriété individuelle et propriété commune « commons » définis par Elinor Ostrom (cf. Holland & Sene, 2010) ne sont-ils pas en train de s'étendre bien au-delà des catégories courantes encore au XXe siècle, vers ce que le droit romain qualifiait de « choses sans maitre » (« res nullius » dont l'humanité entière est redevable à défaut d'en être détentrice) ? Ce dernier domaine du droit s'étend en effet à un domaine où réside aujourd'hui une bonne part de l'héritage et du patrimoine des connaissances, mais aussi de l'environnement naturel. Dans une heureuse synthèse « hypertexte », avec une évidente actualité, le droit médical y place désormais le statut juridique… des microbes (Lucas-Baloup, 1999).

Une évolution, semble-t-il décisive est intervenue pratiquement sans commentaires au niveau global, avec l'adoption par le W3C d'une nouvelle norme de commentaire (Open Annotation) permettant d'assembler les hypertextes : on y reviendra en conclusion.

3. L'alimentation en données des hypertextes : données sur les documents et données sur les usages et les contextes d'usage

Les systèmes de recommandation dans les sciences ont fait l'objet de nombreuses revues montrant leurs différences d'approches et de fonctionnalités, et en définitive d'efficacité, laissant subsister globalement un certain flou des réponses. D'où le développement de démarches visant à améliorer l'information sur le contexte de la requête. L'Association internationale des Editeurs scientifiques STM multiplie désormais les informations de contexte en précisant davantage les usages de la publication scientifique[26] ; l'Association STM publie désormais un Usage Factor for Journals (UFJ1)[27] en développant un commentaire détaillé de contenus générés par les utilisateurs.

Plus généralement, avec le développement d'une approche « multimodale » (multimodal knowledge acquisition) sur les graphes de connaissances scientifiques de l'information scientifique (Jaradeh et al., 2019), s'étend l'association de matériaux de recherche variés (articles, données, commentaires principalement), en s'écartant des fragilités des indicateurs classiques de métrique de publication (Egret, [article 6 du Dossier]) au profit d'approches de l'évaluation, quantitative et qualitative, beaucoup plus larges.

4. Hypertextes et captation géométrique de l'information : nouvelles approches de la géométrie des graphes

Le dessein général fourni par C. S. Shannon (1948) prévoyait que « l'information mutuelle » repose sur une architecture dessinée pour répondre à « toute question possible »[28] au moment de son usage. Comment répondre avec l'hypertexte à ce besoin « d'information mutuelle » ? La modélisation des choix de requêtes sur le Web est développée dans ce sens, tout comme la construction de systèmes et de normes d'assemblage et de traitement de l'information dans ce cadre.

De nouvelles fonctionnalités de structuration et de circulation dans les graphes se multiplient : compas, historique, détection de communautés et routes alternatives… (Fabre et al., 2021).

---

[26] https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf (pp. 57-8, 72)

[27] This standard is « *the Median Value in a set of ordered full-text article usage data (i.e., the number of successful full text article requests) for a specified Usage Period of articles published in a journal during a specified Publication Period* ».

[28] « *The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design* ».

La géométrie analytique a par ailleurs depuis dix ans apporté aux graphes bon nombre de nouvelles propriétés descriptives et interprétatives : les « agnostic neural networks » modélisent des informations sans « entrainement » préalable[29], après les algorithmes génétiques (Stanley & Miikkulainen, 2002) et les « graph transformer networks » (LeCun et al., 1998), associés au développement de la « automated deduction in geometry » appliquée aux graphes (Schreck et al., 2011). Depuis peu, ces démarches débouchent sur un champ entièrement vierge de « geometric deep learning » qui vient de donner lieu (avril 2021) à une première revue de littérature très panoramique (Bronstein et al., 2021).

Si elle se confirme, comme semble le montrer l'inventaire des nombreuses applications en développement actuellement, cette démarche apportera aux hypertextes, des transformations significatives du traitement et de l'analyse de l'information.

Conclusion : L'« Open Process » : partager les architectures hypertextes

Héritière directe de la démarche encyclopédique avec ses « renvois » entre connaissance, la démarche hypertexte est à la croisée de plusieurs chemins.

Une condition pour cela essentielle est que soient garanties les conditions d'un partage des informations et surtout des modèles et architectures qui les sous-tendent sur une base nouvelle de Science ouverte.

En affirmant un principe d'Open Process, d'accès ouvert aux systèmes et à leurs interactions, à leur co-construction, la recherche peut se doter de nouveaux moyens étendus non seulement d'échanger des résultats mais aussi et surtout, à l'amont, d'échanger des démarches et des outils d'assemblage et renforcer ainsi le potentiel et l'autorité des choix savants. Si elle allait dans ce sens, la recherche pourrait bénéficier d'un nouvel atout, qu'est la création d'une nouvelle norme globale applicable aux hypertextes.

Depuis fin 2017 en effet, le W3C, le Consortium Internet, vient de diffuser une nouvelle norme d'Open Annotation[30], exploitée désormais par les Scientific Knowledge Graphs. Cette norme observe de façon liminaire : « Annotating, the act of creating associations between distinct pieces of information, is a pervasive activity online in many guises », et elle développe un ensemble précis de standards de construction et d 'assemblage directement tournés vers la valorisation des hypertextes (construction automatique d'hypertextes et surtout, normes d'assemblage de contenus - texte, image, données - appartenant à plusieurs hypertextes distincts).

De son côté, le Rapport déjà cité de l'Association des Editeurs STM pour 2018, observe[31]: « Open annotation shares some features with simpler forms of annotation (e.g., social bookmarking services) but supports multiple annotation types, including bookmarking, highlighting, tagging and commenting. Annotation does not require either the permission from the content annotated website or that it installations of any new software on its part ».

En s'annonçant ainsi par *la voie d'un standard global* (W3C) d'accès à la construction et au partage du contenu des hypertextes sur tout l'Internet, une nouvelle faculté globale de liaison et de commentaire

[29] La question de recherche est en effet de savoir « to what extent neural network architectures alone, without learning any weight parameters, can encode solutions for a given task ».
[30] https://www.w3.org/TR/annotation-model/
[31] https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf (p. 171)

s'offre à tous les usages, mais en particulier à la représentation des résultats scientifiques : elle pourra en effet être développée avec des services comme ceux de OpenEdition[32] et permettre ainsi à une association libre de choix, une construction commune de voies navigables pour les routes de connaissances.

Bibliographie

Bacon, F. (1631). *Sylva Sylvarum or A natural history in ten centuries*. London : Rawley.

Baltz, C. (2003). Quand la documentation s'éveillera... *Documentaliste-Sciences de l'Information*, *40*(2), 148–153. https://doi.org/10.3917/docsi.402.0148

Bellot, P., & Grau, B. (2014). Recherche d'information et fouille de textes. *Information Grammaticale*, 37–45. https://doi.org/10.2143/IG.141.0.3023303

Boullier, D. (2019). Pablo Jensen, *Pourquoi la société ne se laisse pas mettre en équations*, Paris, Seuil, 2018, 336 p. *Réseaux*, *n° 214-215*(2), 357–367. https://doi.org/10.3917/res.214.0357

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.13478*. https://arxiv.org/abs/2104.13478

Brown, M. J., & Kidd, I. J. (2016). Special Issue: Reappraising Paul Feyerabend. *Studies in History and Philosophy of Science Part A*, 57(June), 1–154. https://www.sciencedirect.com/journal/studies-in-history-and-philosophy-of-science-part-a/vol/57/suppl/C

Dion, E. (1997). *Invitation à la théorie de l'information*. Paris : Seuil.

Direction de l'Information Scientifique et Technique CNRS. (2016). *Livre blanc — Une Science ouverte dans une République numérique*. Marseille : OpenEdition Press. https://doi.org/10.4000/books.oep.1548

Eco, U. (2009). *Vertige de la liste*. Paris : Flammarion.

Fabre, R. (2019). A "Searchable" Space with Routes for Querying Scientific Information. *8th BIR@ECIR 2019*, Cologne, Germany: 112-124. https://dblp.org/db/conf/ecir/bir2019.html#Fabre19

Fabre, R., & Bensoussan, A. (2017). *La fabrique numérique des connaissances*. *Production et valorisation des résultats scientifiques*. London : ISTE Editions.

Fabre, R., Azeroual, O., Bellot, P., Schöpfel, J., & Egret, D. (2021). A Scientific Knowledge Graph with Community Detection and Routes of Search. Testing "GRAPHYP" as a Toolkit for Resilient Upgrade of Scholarly Content. Preprint. https://hal.archives-ouvertes.fr/hal-03365118

Ferrarotti, M. J., Rocchia, W., & Decherchi, S. (2019). Finding Principal Paths in Data Space. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(8), 2449–2462. https://doi.org/10.1109/TNNLS.2018.2884792

Godelier, M. (2015. *L'imaginé, l'imaginaire et le symbolique*. Paris : CNRS Editions.

[32] https://www.openedition.org/6438

Griss, J., Viteri, G., Sidiropoulos, K., Nguyen, V., Fabregat, A., & Hermjakob, H. (2020). ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. *BioRxiv*. https://doi.org/10.1101/2020.04.16.044958

Holland, G., & Sene, O. (2010). Elinor Ostrom et la Gouvernance Economique. *Revue d'économie Politique*, *120*(3), 441–452. https://doi.org/10.3917/redp.203.0441

Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., Kismihók, G., … Auer, S. (2019). Open Research Knowledge Graph. *Proceedings of the 10th International Conference on Knowledge Capture*, 243–246. https://doi.org/10.1145/3360901.3364435

Kahsay, R., Vora, J., Navelkar, R., Mousavi, R., Fochtman, B. C., Holmes, X., … Mazumder, R. (2020). GlyGen data model and processing workflow. *Bioinformatics*, *36*(12), 3941–3943. https://doi.org/10.1093/bioinformatics/btaa238

Kuhn, T. (1962). *La structure des révolutions scientifiques*. Paris : Flammarion.

Le Coadic, Y.-F. (1994). *La science de l'information*. Paris : Presses universitaires de France.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, S. (1998). Gradientbased learning applied to document recognition. *Proc. IEEE*, *86*(11), 2278–2324.

Liu, Y., Wei, X., Chen, W., Hu, L., & He, Z. (2021). A graph-traversal approach to identify influential nodes in a network. *Patterns*, 2(9), 100321. https://doi.org/10.1016/j.patter.2021.100321

Lucas-Baloup, I. (1999). Le microbe : une res nullius cause étrangère ? *Revue Générale de Droit Médical*, (102), 91–110.

Marsh, L., & Onof, C. (2008). Stigmergic epistemology, stigmergic cognition. *Cognitive Systems Research*, *9*(1–2), 136–149. https://doi.org/10.1016/j.cogsys.2007.06.009

Mérindol, J.-Y. (2020). *L'avenir de l'édition scientifique en France et la science ouverte*. Paris : Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. https://www.enseignementsup-recherche.gouv.fr/cid148896/www.enseignementsup-recherche.gouv.fr/cid148896/les-pouvoirs-publics-et-l-edition-scientifique-en-france.html

Morin, E. (1977). *La méthode*. Paris : Le Seuil.

Nedellec, C., Bossy, R., Chaix, E., & Deléger, L. (2018). Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity. arXiv preprint: https://arxiv.org/abs/1805.04107

Papy, F. (2018). Mundaneum numérique et internet augmenté : visions et intuitions de Paul Otlet. In W. Mustafa el Hadi (coord.), *Fondements épistémologiques et théoriques de la science de l'information-documentation*. London : ISTE Editions, pp. 217–227.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, *33*(2), 163–180. https://doi.org/10.1177/0165551506070706

Salaün, J.-M. (2012). *Vu, lu, su*. Paris : La Découverte.

Schöpfel, J., & Prost, H. (2021). How scientific papers mention grey literature: a scientometric study based on Scopus data. *Collection and Curation*, *40*(3), 77–82. https://doi.org/10.1108/CC-12-2019-0044

Schreck, P., Narboux, J., & Richter-Gebert, J. (2011). Automated Deduction in Geometry. *8th International Workshop, ADG 2010, Munich, Germany, July 22-24, 2010*. Cham : Springer.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Stahlschmidt, S., & Stephen, D. (2020). *Comparison of Web of Science, Scopus and Dimensions databases. KB Forschungspoolprojekt 2020*. Hannover : DZHW. https://bibliometrie.info/downloads/DZHW-Comparison-DIM-SCP-WOS.PDF

Stanley, K. O., & Miikkulainen, R. (2002). Efficient Evolution of Neural Network Topologies. *Proceedings of the 2002 Congress on Evolutionary Computation (CEC '02).* Retrieved from http://nn.cs.utexas.edu/downloads/papers/stanley.cec02.pdf

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635. https://doi.org/10.1016/j.joi.2010.07.002