Réflexions sur le fragment dans les pratiques scientifiques en ligne : entre matérialité documentaire et péricope

Gérald Kembellec, Thomas Bottini

Réflexions sur le fragment dans les pratiques scientifiques en ligne : *entre matérialité documentaire et péricope*

Thoughts about the concept of fragment in online scholar practices: between document material and pericope

Gérald KEMBELLEC (1), Thomas BOTTINI (2)

(1) Dicen-Idf, Cnam
gerald.kembellec@lecnam.net

(2) Dicen-Idf, Cnam
Thomas.bottini@lecnam.net

Résumé. Cette communication propose une réflexion pluridisciplinaire (SIC, ingénierie documentaire et théorie du document numérique, informatique, « humanités numériques », histoire des pratiques savantes) sur les usages du fragment dans les pratiques documentaires scientifiques en ligne. En prolongement de ces éléments théoriques sont proposés un modèle théorique de la segmentation des contenus en unités de sens (péricope) et des directions d'implémentation.

Mots clés. Web sémantique ; lecture savante ; primitives savantes ; fragment documentaire ; péricope ; *semantic publishing* ; citation ; rééditorialisation

Abstract. This paper proposes a multidisciplinary reflection (information sciences, document engineering and digital document theory, computer science, digital humanities, history of scholar practices) on the uses of the fragment in online scientific documentary practices. A segmentation of contents model into units of meaning (pericope) is proposed. Some directions of implementation as a continuation of these theoretical elements are also proposed.

Keywords. Linked Data ; scholars ; scholarly primitives ; document fragment ; periscope ; semantic publishing ; quotation; re-editorialisation

1 Introduction

La communauté du document numérique scientifique affiche une volonté de réappropriation des formats éditoriaux pour mieux correspondre aux nouveaux usages du numérique. Les formats traditionnels comme PDF, Word, LaTeX sont bousculés au profit de nouveaux formats comme le RASH (Piotrowski, 2016 ; Peroni et al, 2017) générant des documents Web hypertextes ré-éditorialisables, adressables et autodocumentés grâce aux techniques des *Linked Data* que ce soit sous forme de pages HTML ou de formats de publication nomades enrichis comme l'ePub.

Dans cette communication, nous proposons une vision théorique et technique pluridisciplinaire sur les dispositifs en ligne de médiation informationnelle scientifique. Pédauque (2006) a proposé une théorie du document numérique par des approches sémiotique, structurelle et communicationnelle. Plus récemment Broudoux (2015) a proposé de définir les « contours » du document numérique devenu connecté. Elle soulignait la convergence entre approches structurelle et communicationnelle, ce que nous proposons d'équiper par une méthodologie adaptée à l'écriture numérique de documents en ligne, ainsi qu'en apportant une réflexion sur son implémentation technique. L'enjeu est ici de penser la segmentation et l'inscription des contenus au sein d'un document Web afin d'en préparer la réception et la redocumentarisation (rééditorialisation). Nous pensons cette réception aussi bien par des destinataires logiciellement équipés, que par des machines relayant et traitant de l'information pour les usages du Web de données. Nous soulignons un impensé de la culture documentaire et du Web : la nécessité d'une fragmentation de la matière documentaire en unités de sens — c'est le concept de *péricope* — en complément de la vision plus abstraite du Web des données (ontologies et schémas).

Dans cette communication, nous contextualisons le rôle des fragments dans l'histoire des outils et pratiques savantes jusqu'au Web contemporain, proposons un regard théorique sur la fragmentation matérielle et sémantique en milieu numérique. Enfin, nous observons de manière critique les conséquences de cette mutation, et en dégageons des préconisations utiles dans les méthodes d'inscriptions documentaires dans les pratiques scientifiques en ligne.

2 Analyse et problématisation de la notion de fragment

Le *fragment*, dans sa double acception d'élément signifiant issu d'une production sémiotique principale et de bribe matérielle d'un support d'inscription, renvoie à des imaginaires sédimentés dans l'histoire des pratiques intellectuelles et à des faits techniques caractérisant les dispositifs matériels convoqués par ces pratiques. Nous en proposons ici une analyse pluridisciplinaire.

2.1 Fragments et fragmentation dans l'histoire des pratiques savantes

L'histoire de la fixation du contenu sur un support à des fins heuristiques et herméneutiques[1] est animée par une tendance d'augmentation des potentialités de fragmentation du sens et de son substrat matériel, débouchant sur un perfectionnement des techniques savantes et de *« repérage de l'information »* : structuration, adressage, « commentabilité » (Fayet-Scribe, 1997). Nous proposons des exemplifications de cette tendance dans le tableau chronologique n°1. À cette histoire matérielle s'ajoutent, sur le plan logique du contenu, des *techniques de spatialisation des fragments* permettant de percevoir des relations signifiantes invisibles, et des *traditions de segmentation* (péricope, analyse structurale, TEI, etc.)[2].

La polarisation globalisatrice des technologies du Web sémantique renvoie à l'époque scolastique dans cette volonté de qualifier et d'organiser la totalité de la

[1] Pour une synthèse historique plus complète, se référer à Bottini (2010) et Fayet-Scribe (1997).

[2] Par exemple, Lackner (2011) expose les techniques de fragmentation et d'analyse non linéaire des textes comme pratiques savantes de premier ordre en Chine entre le XII^e^ et le XIV^e^ siècle.

connaissance du monde, les standards W3C et les *data stores* dans le *cloud* s'étant substitués à la cathédrale[3]. Les possibilités de discrétisation, de qualification et de mise en relation des signes sont infinies, et les standards de description de contenu sont variés et riches. Mais le support numérique semble « en avance » sur les pratiques : si les développements visant une pratique interprétative spécifique des humanités numériques (HN) mobilisent des modèles communs « abstraits » de fragmentation et de description, ils réassument isolément des questions de base relatives à la création et la manipulation concrète des fragments dans l'environnement de travail. Les questions d'ergonomie cognitive et physique des supports savants (désormais, de leurs IHM) a toujours résulté historiquement d'une implication éditoriale, qui, étant moteur des améliorations techniques et typodispositionnelles du livre, a rendu possibles des travaux analytiques et synthétiques toujours plus fins.

[3] On doit à Panofsky une analyse du lien entre cet édifice particulier et les structures mentales de l'époque (Déotte, 2010).

Technologie	Effets pour la fragmentation
dès 4000 av. J.-C. (Mésopotamie) listes	premières structures d'organisation des mots, donc du sens, donc du monde
vers 1500 av. J.-C. (Ougarit, Syrie) l'alphabet	amorce la tradition de découpe-recomposition textuelle comme clef de voûte des mondes « lettrés », ainsi que la possibilité de classer et de renvoyer à d'autres textes
antiquité greco-romaine prise de notes fragmentaires relative aux expériences personnelles (idées, lecture…) (hupomnêmata)	externalisation de la mémoire en vue d'un usage futur
1er siècle codex en parchemin, qui supplante le volumen en papyrus	1) fragmentation du contenu en unités logiques et matérielles plus facilement manipulables, opposant la page et le rubricage à la séquentialité imposée du rouleau 2) espace articulé au texte, marginal ou interlinéaire, pour recevoir gloses et scholies, ce qui initie la lecture intensive et la tradition savante moderne
7ème siècle (Irlange) blanc entre les mots	point d'appui des stratégies de segmentation textuelle tant sémantiques que matérielles
12ème siècle *stratégies d'organisa*tion de la page	lisibilité du texte articulé à ses multiples strates de commentaires
période scolastique dispositifs d'identification, indexation et organisation des fragments-citations des autorités	1) *statim inveniri* (« accès immédiat »), notion consubstantielle à la lecture extensive qui connaîtra des développements contemporains (état de l'art, bibliométrie, surcharge cognitive, etc.) 2) prémices des primitives savantes avec les rôles de Bonaventure : scribe-copiste, compilateur, commentateur, auteur
Renaissance humaniste circulation des livres et utilisation des marges	formation d'un « réseau social savant contributif » européen
Renaissance standardisation de l'espace de la page amenée par l'imprimerie	disparition de la pratique du fragment-commentaire à-même le support divorce entre lecture et écriture
fin du 18ème siècle Évangiles synoptiques	fragmentation et spatialisation fine des textes pour leur étude structurelle précise
fin du 20ème siècle langages et méthodes de l'ingénierie documentaires	raison computationnelle appliquée au contenu

Tableau 1. *Synthèse chronologique de la fragmentation dans les pratiques savantes*

2.2 Les fragments à l'ère des documents Web

Les travaux théoriques présentés par les pionniers anglo-saxons de la documentation avaient un objectif d'accès distant à l'information (Otlet), puis de navigation entre idées au sein de documents distants (Bush, Engelbart, Nelson). Enfin, les principes du langage de balisage hypertexte mis en oeuvre initialement par T. Berners-Lee étaient de permettre l'accès à distance à une représentation-écran de documents structurés — même faiblement[4] — et hyperliés. Le Web évolua au tournant du 21e siècle vers la sémantique (ce terme n'est plus utilisé par son « créateur » qui lui préfère Web des « données liées »). Les questions de partage, d'adressage et de présentation du document étant réglées, il reste à mieux comprendre et améliorer la fragmentation ainsi que la description des contenus intradocumentaires.

Le formalisme du HTML tenait plus à l'origine d'aspects typodispositionnels que documentaires, mais a évolué pour répondre à des besoins plus pointus de formalisme (xHTML, HTML5) proposant *in fine* des possibilités de structuration sémantique riches formalisées à l'écran grâce aux « feuilles de style », mais pouvant aussi être utilisés à d'autres fins par la *navigation équipée*. L'outil capable d'analyser un hypertexte et d'en donner une représentation unifiée a été nommé *browser ou* « navigateur ». Nous entendons par navigation équipée l'usage de fonctionnalités complémentaires, ajoutées au navigateur par choix de l'usager pour des besoins spécifiques que nous cadrons ici uniquement dans une optique documentaire, qui peuvent être très spécifiques :

Ex. 1. Rétrécissement typologique

- Il peut s'agir de détection, signalisation et collecte automatisée de fragments informationnels ou de notices bibliographiques sur un sujet particulier dans les pages d'un vaste corpus.

Ex. 2. Ouverture

- Il peut également s'agir de sérendipité avec la possibilité de découvrir de nouveaux contenus en lien direct avec les entités nommées et concepts au sein de l'hypertexte en cours de lecture, méthode dite de *berrypicking* qui autorise le rebond informationnel entre documents (Bates, 1989).

En adéquation avec ces besoins, nous retrouvons des plateformes d'éditorialisation de documents, y compris scientifiques, permettant la structuration hypertextuelle sémantique dans une optique d'écriture et de partage de l'information[5] à destination des modes de lecture et de raisonnement de l'humain ET des machines désirantes[6]. Cette manière de repenser structurellement l'écriture numérique scientifique en portions liées, avec les données extérieures disponibles, est rendue possible par le *Linked Data* et a été nommée *semantic publishing* (Shotton, 2009).

[4] Initialement le HTML devait simplifier la structuration documentaire SGML et ajouter une couche de présentation en réseau.

[5] Cf. panorama de ce type de plateformes éditoriales scientifiques proposé par Broudoux et Kembellec (2017, p. 35 et 36).

[6] Terme emprunté à Deleuze et Guattari.

2.3 Le fragment par les opérations savantes invariantes des humanités numériques

La fragmentation du sens et des supports étant la clef de voûte des pratiques

savantes, il est utile d'évoquer les travaux visant à dégager des invariants opératoires

dans la diversité de pratiques auxquelles renvoie l'expression « humanités

numériques ». Ce souci est initié par Unsworth (2000), qui identifie sept « primitives

savantes » (*« scholarly primitives »*) (cf. tab. 2). La systématicité et l'extensibilité de ce

concept en ont fait un cadre structurant pour plusieurs travaux analytiques et

prescriptifs. Ainsi, visant l'élucidation des pratiques informationnelles des

chercheurs pour penser des infrastructures informatiques plus efficaces en

bibliothèque, Palmer et al. (2009) étendent le nombre de primitives et les structurent en six activités essentielles (cf. fig. 2). Pour Blancke et Hedges (2013) le recours à des primitives savantes permet d'organiser la pluralité des analyses des pratiques, de spécifier des SI plus adaptés aux besoins des chercheurs et susceptibles de susciter de nouvelles pratiques et de nouveaux objets de recherche, et de garantir l'articulation des différents projets d'implémentation afférents. Hennicke et al. (2015) constatent que malgré les travaux théoriques et techniques dans le giron des HN, les *« scholars »* recourent principalement à des outils connus, communs, et donc peu adaptés à la singularité de leurs tâches, et postulent le manque d'un modèle unifiant les pratiques numériques et papier. À travers leur *Scholarly Domain Model*, les auteur.e.s poursuivent l'intention d'Unsworth de dégager une base commune aux humanités indépendante des disciplines et approches théoriques (cf. tab. 2). Ce modèle de l'activité de recherche s'ouvre sur la constitution du corpus, et attribue un rôle central à la contextualisation de ses éléments et fragments en vue de leur compréhension, par le prisme de la primitive *interpretive modelling*, appuyée par le recours à des données de référence externes (techniquement, des *données liées*) et à des structures de référence (ontologies, thésaurus, etc.) venant enrichir et qualifier le

[UNSWORTH 2000]	**annotating, comparing, discovering, illustrating, referring, representing, sampling**	
[PALMER et al. 2009] *extension & organisation des primitives savantes*	**searching**	direct searching, chaining, browsing, probing, accessing
	collecting	gathering, organizing
	reading	scanning, assessing, rereading
	writing	assembling, co-authoring, disseminating
	collaborating	coordinating, networking, consulting
	cross-cuting primitives	monitoring, notetaking, translating, data practices
[HENNICKE et al. 2015] *modèle plus large de l'activité*	**area**	*contexte général dans lequel l'activité de recherche s'insère*
	scholarly primitives	interpretative modeling, exploration, aggregation, augmentation, externalisation *grandes catégories abstraites de pratiques*
	scholarly activities	(direct) searching, discovering/foraging, browsing, probing, chaining, monitoring, reading, contextualising/conceptualising, translating, assessing, comparing, synthesising/filtering, sampling, organising, collecting/gathering, referring/linking, annotating, selecting, writing, assembling, notetaking, illustrating, sharing, publishing, disseminating
	scholarly operations	*concrétisation des scholarly activities dans le cadre d'un geste ou processus propre à une discipline identifiée*
[NAKAKOJI et al. 2005] *le travail savant en cinq phases*	**I**	collecte des fragments & données à partir des sources
	II	compréhension du matériel rassemblé (annotations subjectives & tissage de liens)
	III	adjonction de nouvelles idées
	IV	construction d'un « récit cohérent » articulant ces idées
	V	sélection des fragments en vue d'une structure documentaire publiable, et archivage des fragments non retenus pour usage plus tardif

sens.

En inventoriant des primitives, ces auteur.e.s sont amené.e.s à évoquer leur caractère interrelié, et tendent donc, en lisière de leur propos, vers l'idée d'une chaîne opératoire savante, telle que la *« chaîne lectoriale »* de Bottini (2010, 2017) dont les six maillons répondent à un objectif conceptuel similaire à celui des cinq primitives de Hennicke et al. (2015). Similairement, Nakakoji et al. (2005) donnent un modèle du travail savant en cinq phases dans lesquelles les fragments sont centraux (cf. tab. 2).

2.4 Le fragment comme clef de voûte de l'ingénierie documentaire

Souvent associé à une instabilité des formes documentaires et à la désorientation afférente des pratiques intellectuelles, le signe numérique rend également possible, du fait de sa nature « *autothétique* » (qui « ne pose que lui-même » (Bachimont, 2004)), une gamme d'usages et de réutilisations infinie en droit. Ceci fonde ce qui est nommé « rééditorialisation » en SIC, terme renvoyant aux pratiques socio-culturelles et non à la strate matérielle et technique des signes numériques. Voyant en la rééditorialisation la *« tendance »* de l'écriture numérique, Crozat (2012) identifie quatre techniques d'écritures spécifiques à celle-ci, constituant autant de manières d'impliquer un fragment dans un processus auctorial : le *polymorphisme* (adaptation de la forme du contenu au contexte de restitution), la *transclusion* (réutilisation de fragments par référence, sans production de copies divergentes), la *dérivation* (réutilisation des fragments dont certains éléments sont ajustables selon les besoins), et la *déclinaison* (variations *programmées* à même le contenu et sélectionnées selon des paramètres éditoriaux). Ces primitives documentaires sont *« destinées à favoriser la rééditorialisation tout en évitant ou contrôlant le clonage »*, le clonage conduisant à la redondance, à l'incohérence, à la perte du contexte éditorial originel, à la *« perte de valeur de vérité associée [aux] fragments »*. Toutefois, l'auteur ne donne du fragment qu'une définition technique — *« ressource numérique qui peut être intégrée par transclusion dans d'autres fragments »* —, et se tient donc en amont de toute tradition d'usage. En cela, son propos est complémentaire aux travaux sur les primitives savantes.

Tableau 2. *Synthèse des primitives savantes par auteurs*

L'ingénierie documentaire a également pour enjeu la maîtrise de la variabilité des contenus dans le temps. Il a été souligné en SIC comment le milieu calculatoire conférant au document informatique une nature dynamique le prive en même temps de sa valeur probatoire sur le passé. Ainsi, Crozat (2012) pose que *« le document numérique n'existe pas, la locution est oxymorique. Il ne peut exister que des constructions numériques dont le traitement calculatoire permet de simuler un ordre documentaire »*. Toutefois, les « effets » du numérique ne doivent pas être évalués uniquement sur la base de ses caractéristiques (dynamicité, évolutivité permanente), mais en considérant la totalité des dispositifs informatiques d'instrumentation des fragments[7]. Ainsi, si la stabilité du contenu d'un fragment numérique ne lui est pas immanente, elle peut être assurée par un dispositif logiciel. L'évolutivité des fragments peut être prise en charge par un système de gestion des versions

[7] Nul ne considérerait les effets socio-cognitifs du livre par le prisme exclusif des propriétés d'absorption de l'encre par le papier.

successives et concurrentes, ces systèmes permettant de situer toute modification dans l'espace du document et dans le temps des contributions. Deux niveaux de fragmentation sont à considérer : le niveau technique de la forme sémiotique d'expression ou de la forme d'enregistrement (Bachimont, 2004), qui rend possible l'établissement d'un différentiel entre deux états d'un même contenu[8], et le niveau socio-coopératif qui découpe le document en unités signifiantes autonomes - au sens de Zacklad (2015) - dans le cadre du Document pour l'Action) afin d'organiser l'effort commun. Nous pensons que des extension des systèmes de *versioning* distribués tels que Git (en dépassant le niveau strictement technique des fragments nécessaire au calcul des différentiels), instrumentés par des plateformes contributives comme Github (pour le travail collectif autour de fragments comme unités de sens) et Stackoverflow (pour le suivi des controverses), le tout adossé à des systèmes de « signature » tels que la *blockchain*, permettraient de garantir une traçabilité des signes numériques et d'arriver à une finesse de suivi philologique sans précédent.

3 Deux directions pour la fragmentation « savante » à l'heure du Web

Les primitives savantes s'appliquent à l'analyse des pratiques, sans se prononcer sur la nature des objets documentaires convoqués, et ont de ce fait un intérêt heuristique limité pour la conception de SI savants. Des efforts comme ceux de Crozat (2012) s'attachent à l'inverse à formaliser des primitives documentaires établissant les critères fondamentaux de l'auctorialité numérique — lesquels reviennent à s'intéresser au cycle de vie des fragments — indépendamment de toute pratique. À notre connaissance, aucun travaux visant les pratiques savantes ne semblent avoir concilié ces dimensions sans sacrifier la généricité. Nous proposons ici deux directions pour contribuer à ce programme, l'une résultant d'une leçon tirée d'un projet de recherche, l'autre étant plus spéculative.

3.1 Fragmenter la matière pour ouvrir les possibles, le cas des herbiers numérisés

L'infrastructure e-ReColNat[9] fédère les efforts d'informatisation des collections végétales sèches institutionnelles. Les diverses pratiques potentielles autour de celles-ci offrent à la conception de SI contributifs savants un « cas » où la réflexion sur la nature du fragment est déterminante.

Les métadonnées calligraphiées des étiquettes sont au cœur de toute démarche botanique. Leur transcription fait l'objet d'un processus de redocumentarisation participative dans le site *Les Herbonautes*[10], induisant des formes d'écriture fragmentée (Chupin, 2017). La fragmentation est d'abord matérielle (localisation de l'étiquette au sein de la planche, puis identification des « champs »), avant de prescrire une organisation du travail : les informations relatives au lieu de collecte sollicitent les compétences historico-géographiques des participant.e.s, et celles relatives à la dénomination savante ou vernaculaire de la plante, leurs connaissances botaniques. Les interprétations divergentes sont discutées et argumentées dans des forums attachés aux différents niveaux de fragmentation structurelle

[8] Dumas (2016) propose une synthèse des différentes stratégies techniques de différentiel entre les évolutions d'un fragment.

[9] Infrastructure Nationale en Biologie et Santé e-ReColNat (ANR-11-INBS-0004)

[10] http://lesherbonautes.mnhn.fr

(planche/étiquette/champ), dans lesquels des figures d'autorité locale émergent. Les processus de fragmentation à l'œuvre génèrent donc une « triple indexation » de l'information : matérielle, sémantique (taxons et lieux géographiques étant adossés à des référentiels) et sociale.

L'instrumentarium logiciel e-ReColNat compte également le Collaboratoire[11], environnement scientifique permettant d'effectuer mesures & annotations sur les planches. La fragmentation s'y opère également à plusieurs niveaux : matérialité des médias (localiser des zones d'intérêt dans une continuité spatiale : feuille, tige, bourgeon, étiquette, etc.), production des données scientifiques (chacune étant rattachée à un aspect scientifique ou à un élément matériel spécifique d'une planche), et organisation de l'espace de travail (répartition des planches en *sets*, spatialisation des planches pour élaborer des catégorisations implicites à des fins de systématique).

Si le regard porté par les botanistes sur les différentes informations produites dans ces outils dépendent de leur discipline (systématiciens et écologues ne s'intéressent pas aux mêmes aspects du réel et ne produisent pas les mêmes savoirs), le SI e-ReColNat dans son ensemble nous permet de dégager la leçon que leur coopération est néanmoins rendue possible par la dualité matériel/sémantique de la fragmentation, et l'indépendance de deux étapes de l'indexation que sont la « localisation » et la « qualification » (Bachimont, 2004) En effet, la localisation d'une portion « physique » (feuille, tige, étiquette…) au sein d'un continuum génère une ressource non encore polarisée pouvant ainsi être convoquée par une multiplicité de points de vue et d'usages. Pour exemples : disposer d'une indexation/localisation matérielle d'un grand nombre de feuilles permet de disposer d'une définition par extension sous forme d'exemplification du concept de feuille dans sa variabilité ; constituer une série d'étiquettes peut appuyer une recherche historique où la source est centrale ; mettre en série les traces et annotations spatiales aide à comprendre où le regard et la main des scientifiques se portent sur les herbiers ; etc. Ce qui fait dire à Bachimont (2004) que *« l'indexation n'est ni définitive ni universelle »*, c'est que son étape de qualification porte la marque du contexte culturel et d'usage de sa production. Mais la localisation matérielle des fragments d'intérêt peut trouver des usages plus larges que ce que les annotations dont elle constitue l'ancrage ne le laissent supposer[12].

3.2 Fragmenter pour donner une autonomie de sens aux inscriptions, vers une « péricope moderne »

Nous n'avons pas trouvé dans l'état de l'art d'éléments structurels pour la segmentation d'un document en unités de sens pour aller vers la vision de Conklin (1987) d'un hyperdocument préindexé autorisant une lecture non linéaire.

La tradition exégétique des grandes religions révélées a développé des méthodes de segmentation du sens allant au-delà de la simple structuration visant l'accès à l'information et fondant sa citabilité (livres, chapitres, versets/sourates). Pour exemple, dans *« (Luc 15:11-32) »*, l'unité de sens qu'est la parabole biblique du fils

[11] https://lab.recolnat.org

[12] Dans le champ artistique, on peut ainsi remarquer que le Hip hop s'est constitué comme genre musical autonome par la fragmentation d'œuvres enregistrées dont les passages instrumentaux sémantisés comme simples breaks anecdotiques par leurs auteurs ont endossé une fonction de structuration rythmique fondamentale. La négation du sens premier du fragment est donc ici la condition de son émancipation.

prodigue est matériellement localisée, mais peut être lue de manière *autonome*, comme une simple histoire. La tradition catholique propose également des *péricopes*, ou *« passages »* pour Meynet (2013), petits fragments parfois titrés dans les traductions, et ne se recouvrant pas nécessairement selon les choix éditoriaux (ibid). De son côté, la tradition juive offre une organisation du texte sacré basée sur l'étude annuelle de la Torah qui suit une logique similaire dite de la *parasha*.

Définition : Une péricope est un fragment textuel formant une unité ou une pensée cohérente qui a du sens dans son cadre d'écriture, mais aussi indépendamment de son contexte.

Ainsi, un découpage du texte lors de sa transition au numérique peut être réalisé à un niveau lexical par la sémantique. Afin d'étendre l'usage de la sémantique, nous proposons un niveau de structuration documentaire intermédiaire, un niveau macroscopique qui serait consacré au sens plus qu'à la terminologie. Cette architecture s'inspirerait du fragment (tel que la péricope) et permettrait de proposer des fonctions de traitement identiques à la sémantique des concepts, aux hyperliens auto-décrits, à la désambiguïsation des entités nommées, mais répondant réellement à l'usage de la notion de fragment, fondamentale à la description de la littérature scientifique. Le découpage documentaire et la typologie des fragments offriraient alors un accès simple et sans ambiguïté à des portions de documents possiblement autoporteuses de sens.

Pour aller dans le sens d'une rhétorique de l'hyperlien (Saemmer, 2015), si le lien pointé par le document *« géniteur »* est un fragment (*a fortiori* s'il est une péricope), l'hyperlien sera d'autant plus fin rhétoriquement qu'il ne pointera plus sur l'ensemble d'un document en ligne, mais uniquement sur la partie souhaitée qui serait typée. Il serait ainsi possible de déduire la valeur argumentative de l'hyperlien au sens auctorial. Ainsi, l'incertitude — supposée — du lecteur face au sens argumentaire à donner à l'hyperlien (ibid, p.23) serait réduite par une lecture équipée du texte dans un navigateur.

Beretta et Letricot (2017) proposent dans le cadre projet SyMoGIH des textes historiques dont les « unités de connaissances » sont systématiquement balisées selon des schémas prédéfinis. Ils partent du postulat que *« dans une perspective de partage et de remédiation des données scientifiques, il paraît judicieux de mettre à disposition de la communauté disciplinaire et du public le texte de la source étudiée, enrichi de l'apparat critique constitué par le travail de recherche qui va en faciliter la compréhension et l'appropriation »*. Il s'agit de permettre de lier sémantiquement des événements, des personnes, des lieux, des faits historiques entre eux et avec le reste des ressources disponibles dans la sphère *Linked Data*. Cette activité de mise en exergue de concepts au sein du texte est fondamentale, mais pas assez macroscopique pour être réutilisée dans le cadre du balisage, de l'annotation et de l'interconnexion des fragments de littérature scientifique.

Strictement appliqué à l'écriture scientifique, il pourrait être envisagé de segmenter un article selon le modèle IMRaD (Introduction, Méthodes, Résultats et Discussion) si l'on souhaite typer les fragments. Cependant, ce modèle bien que très répandu en sciences « dures » s'applique difficilement aux SHS où la liberté auctoriale est privilégiée par rapport aux normes éditoriales. Il serait donc intéressant d'ajouter des types de fragments plus ou moins longs comme une définition éventuellement accompagnée d'exemples. Au sein des fragments, le typage des péricopes

numériques — en écriture scientifique des humanités — sera le sujet d'un débat (ultérieur) pour rester en prise avec les besoins auctoriaux et éditoriaux.

Le découpage du document scientifique peut se faire a posteriori de son écriture de manière automatisée selon des modèles prédéfinis, souvent grâce aux outils de TAL[13]. SciAnnotDoc (de Ribaupierre et Falquet, 2017), pour prendre un exemple récent, permet de répondre au besoin de désagrégation et réagrégation de fragments de littérature scientifique (Bishop, 1999).
La démarche inverse, celle qui consisterait à offrir a priori une écriture nativement fragmentaire et sémantique a été pensée par les « chaînes éditoriales » avancées comme le montrent les travaux de Crozat (2016) sur l'écriture avec et pour des « machines à calculer ». En effet, la notion de calculabilité numérique — qui amène au travail intellectuel de nouvelles possibilités manipulatoires et représentationnelles (Bottini et al., 2008) — appliquée ici permettrait un filtrage des fragments informationnels scientifiques, selon une typologie argumentaire proche de la rhétorique. Un usager de documents encodés sur ce modèle, pourra par exemple sélectionner tous les arguments relatifs à une notion dans un corpus, ou encore relever tous les états de l'art sur une même thématique. Cette vision, bien que coûteuse en temps lors de sa mise en œuvre auctoriale, présente l'avantage de se forcer à une structuration mentale forte du document fini et a pour corollaire d'écrire nativement pour tous supports de diffusion, y compris pour la lecture hypertextuelle[14].

4 Préconisations et perspectives techniques pour la matérialisation des unités de sens sur le Web

4.1 HTML5 et les balises sémantiques : l'évidence trompeuse ?

Le W3C propose avec le HTML5 des balises dites sémantiques permettant de penser à la fois les aspects typo-dispositionnels, mais aussi de segmenter les contenus d'un document au sens de Crozat et d'en typer les fragments, permettant à ce langage de présentation d'être également structurant. Avec des balises comme <article>, <dfn>[15], <summary>, <section> ou encore <aside>[16] le HTML5 semble apporter un début de réponse. Ces balises peuvent-elles réellement donner du sens à des fragments, permettre à des liens hypertextes y aboutissant d'être porteurs d'un type structurel : ce lien pointe vers une définition ou un résumé dans un document scientifique en ligne ? Il semble que non : typer un fragment documentaire, une péricope par exemple, il faudra user d'une spécification commune à l'auteur et au lecteur et que le fragment dispose d'une URI. Cela se fait habituellement au moyen de langages avancés de segmentation comme le TEI. Voyons de manière empirique comment envisager cela dans une optique de lecture Web et ce sans le coût logiciel d'un moteur de transformation.

[13] Traitement Automatisé du Langage.
[14] À l'aide de feuilles de style adaptées selon les préconisations CSS3 du W3C qui permettent un large panel de mise en forme du même document, du document imprimé au smartphone en passant par l'écran d'ordinateur.
[15] <dfn> est utilisée pour encadrer une définition, elle peut servir pour appliquer un style à une définition et/ou la signaler comme étant un fragment. De plus, étant dite « sémantique » cette balise fixe la réception du fragment à une fonction de définition (à moteur de recherche par exemple).
[16] Intéressant pour toute forme de paratexte éditorial ou auctorial

4.2 Attributs et ancres : affiner la typologie des fragments

Les balises en HTML5 permettent donc de segmenter et typer des éléments simples au sein d'un document : état de l'art, définitions ou autres fragments. Si elles respectaient strictement une règle éditoriale universelle d'écriture scientifique, elles deviendraient automatiquement porteuses intrinsèquement de sens et la question de l'adéquation la matérialité documentaire des fragments avec celle de leur inscription et réception serait réglée. Mais en l'état, le *semantic publishing* ne peut pas se fondre dans le HTML5, qui n'offre en réalité nativement que peu de sémantique. Il faut donc faire appel à des techniques permettant de typer scientifiquement les balises structurelles souhaitées et encadrer les fragments selon des schémas partagés avec la communauté et ancrer ces mêmes balises-fragments au sein des documents avec des identifiants uniques pour leur donner une possible vie propre hors du document. Ainsi, cette pratique permettrait de ne citer que les hypothèses d'une thèse en ligne et en autoriserait éventuellement la transclusion dans une perspective de rééditorialisation.

Une implémentation de ce modèle convoquerait en priorité les modèles techniques basés sur les triplets sémantiques (RDFa, microdonnées…), s'adjoignant l'usage des identifiants d'autorités (LoC, BnF…) et des schémas structurels sémantiques (schema.org). Il n'existe malheureusement pas (encore) de chaînes éditoriales capables de proposer un cadre d'écriture convivial pour un chercheur non spécialiste des technologies sémantiques avec un éditeur *What You See Is What You Mean* scientifique. En attendant cette avancée, nous proposons une méthode transitoire très simple proche des microformats[17] : le HTML offre des outils s'ajustant aux besoins évoqués, ce qui permettra en théorie d'utiliser un simple éditeur WYSIWYG pour répondre aux besoins basiques par réappropriation de certaines fonctionnalités :

1) Pilier des fonctions de navigation hypertexte, l'ancre va autoriser directement l'accès à une partie d'un document grâce à l'usage de l'attribut « *id* » d'une balise par la création d'une URI. Cette fonctionnalité est utilisée dans l'index hypertexte de Wikipédia pour accéder à un fragment d'article.
2) L'attribut «*class*[18] »va permettre de typer une balise, s'il est traditionnellement utilisé pour la mise en page, il peut également caractériser un fragment documentaire.
3) L'attribut « *title* » pourrait également servir à contextualiser le fragment en le nommant.

Précisons ici que nous entendons détourner l'usage d'une feuille de style en ligne afin de servir d'autorité pour la typologie des classes utilisables. Si cette méthode détourne les usages traditionnels des classes, le couple ancre/classe n'en est pas moins efficace pour typer et adresser des fragments documentaires.

On note ici que la 2e partie d'introduction (cf. Fig. 1) « Aspects historiques du concept » est présentée comme une péricope et prétend à ce titre subir une

[17] cf. modèles d'extraction des microformats : https://www.w3.org/2007/07/grddl-pressrelease

[18] https://www.w3schools.com/tags/att_global_class.asp

```html
<!doctype html>
<html>
 <head><meta charset="utf-8"></head>
<body>
<article>
    <!-- fragments IMRAD -->
    <section id="introduction" class="IMRAD-introduction" title="Introduction">
        <p id="intro1">
        L'objet de cet article...
        </p>
        <div id="intro2" class="pericope" title="Aspects historiques du concept">
        Historiquement, ...
        </div>
    </section>
    <section id="methodes" class="IMRAD-methods" title="Matériel et méthodes">
        <p id="material" class="material" title="Matériel utilisé">
            Nous avons utilisé un <span id="material1">Scanner Xerox</span> et un
            <span id="material2">logiciel d'OCR</span> ...
        </p>
        <div id="methods" class="methods" title="Méthodes suivies">
            <div id="method1" class="method" title="La méthode du collègue"></div>
            <div id="method2" class="method" title="La méthode de l'autre collègue"></div>
            <div id="method3" class="method" title="Ma méthode"></div>
        </div>
    </section>
    <section id="results" class="IMRAD-results" title="Résultats">
        <p id="result1">Bla-bla... 27 % de réussite, Bla-bla...</p>
        <p id="result2">Bla-bla... 46 % de réussite, Bla-bla...</p>
        <p id="result3">Bla-bla... 75 % de réussite, ça marche très bien</p>
    </section>
    <section id="discussion" class="IMRAD-discussion" title="Discussion">Bla-bla...</section>
    <section id="conclusion" class="conclusion" title="Conclusion">En conclusion, ...</section>
    <section id="bibliography" class="bibliography" title="Bibliographie">Bla-bla...</section>
</article>
</body>
</html>
```

transclusion sans perte de sens. Les autres fragments sont également adressables et donc citables, mais sans offrir une autonomie de sens.

4.3 Aspects d'une IHM centrée fragments et citation sur le Web

Cette trame de méthode rudimentaire d'implémentation HTML de fragments documentaires scientifiques présente une simplicité certaine et l'avantage de pouvoir s'allier à une feuille de style qui offrirait de tendre vers le WYSIWYM avec un éditeur mixte code/aperçu du résultat. Un développement *ad hoc* pourrait même proposer une aide à la rédaction de documents scientifiques WYSIWYM sur le modèle de Scenari.

Figure 1. Exemple rudimentaire de l'adaptation d'un plan d'article IMRAD au modèle

Les augmentations présentées dans le paragraphe précédent ne sont pas toutes tangibles en lecture-écran traditionnelle, dans la continuité de notre proposition. Il reste à équiper le navigateur d'un *plug-in* de lecture augmentée capable de mettre en exergue les fragments et de les différencier selon leur typologie sur le modèle de l'extension *OpenLink Structured Data Sniffer*, qui s'active et se désactive à la demande, permettant de comprendre la structure documentaire, puis de filtrer et rebondir vers d'autres fragments liés.

5 Conclusion

Dans cette communication, nous avons mené une réflexion sur la contextualisation de la notion de fragment dans la littérature scientifique hypertexte en ligne. Après une rétrospective historique de l'usage de la liaison hypertextuelle des fragments informationnels numériques, avons proposé une symétrie entre la notion historique philologique de péricope et celle que l'on pourrait appliquer à l'exégèse scientifique : la péricope comme fragment ayant une autonomie de sens. Nous avons poursuivi par la possibilité de penser un modèle théorique et les modalités potentielles de son implantation technique. Les potentialités d'analyse de tels liens entre les fragments scientifiques typés et sémantisés présentés dans cet article sont exploitables tant sur le plan argumentatif que pour la scientométrie, comme outil d'analyse critique de la citation.

Depuis février 2017, le *Web Annotation Working Group* du W3C propose une recommandation de modèle d'annotation de fragments documentaires en ligne le

Web Annotation Data Model[19]. Il est raisonnable de penser que le futur de la qualification de fragments ou des péricopes en ligne peut passer par l'implémentation de cette recommandation.

Avec le vif intérêt des éditeurs scientifiques pour les plateformes éditoriales du *semantic publishing*, il reste également à la communauté et aux éditeurs à continuer la réflexion sur les interfaces qui permettront aux auteurs une rédaction WYSIWYM des fragments scientifiques.

6 Références bibliographiques

Bachimont, B. (2004). *Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle*. Mémoire d'habilitation à diriger des recherches, Université de Technologie de Compiègne.

Bates, M-J. (1989) The design of browsing and berrypicking techniques for the online search interface. *Online review*, vol. 13, num. 5, p. 407-424.

Beretta, F. et Letricot, R. (2017). Le portail XML du projet symogih.org : un projet d'édition numérique collaborative de sources et d'informations historiques. In *Écrilecture augmentée dans les communautés scientifiques. Humanités numériques et construction des savoirs,* Kembellec G. et Broudoux É. (Eds.), pp. 125-145. ISTE Editions.

Bishop, A. P. (1999). Document structure and digital libraries: how researchers mobilize information in journal articles. In *Information Processing & Management*, vol. 35, num. 3, p. 255-279.

Blancke, T. et Hedges, M. (2013) Scholarly primitives: Building institutional infrastructure for humanities e-Science, pp. 654-661, In. *FUTURE GENERATION COMPUTER SYSTEMS*, vol. 29, num. 2.

Bottini, T. (2017). Les espaces de la critique. Une étude des conditions de possibilité d'une lecture savante multimédia. In *Écrilecture augmentée dans les communautés scientifiques. Humanités numériques et construction des savoirs.* Kembellec G. et Broudoux É. (Eds.), pp. 57–69. ISTE Editions.

Bottini, T. (2010) *Instrumenter la lecture critique personnelle multimédia*. Thèse de l'Université de Technologie de Compiègne, 2010.

Bottini, T., Morizet-Mahoudeaux, P., et Bachimont, B. (2008). Instrumenter la lecture savante de documents multimédia temporels. In *Actes du onzième Colloque International sur le Document Electronique* (CIDE 11), Rouen, France.

Broudoux, É. (2015). Contours du document numérique connecté. In *Documents et dispositifs à l'ère post-numérique, actes du 18e Colloque international sur le document Electronique* (*CIDE 18*). p. 7-15. Montpellier, France

Broudoux, É. et Kembellec, G. (2017). Introduction à l'écrilecture scientifique et aux modalités techniques de son augmentation, In *Écrilecture augmentée dans les communautés scientifiques. Humanités numériques et construction des savoirs* Kembellec G. et Broudoux É. (Eds.), ISTE Editions, mai.

Chupin, Lisa. (2017). La construction de normes d'écrilecture pour la transcription collaborative du patrimoine numérisé : entre algorithme, transmission et élaboration communautaire ». In *Écrilecture augmentée dans les communautés scientifiques. Humanités numériques et construction des savoirs* G. Kembellec et É. Broudoux (Eds.), pp. 89–106. ISTE Editions, mai.

[19] https://www.w3.org/TR/annotation-model/

Conklin, J. (1987). HyperText : An Introduction and Survey. In *IEEE Computer*, septembre 1987, vol. 18, num. 9, p. 17-4.

Crozat, S. (2016). Écrire avec une machine à calculer, écrire pour une machine à calculer. In *I2D–Information, données & documents*, ADBS, vol. 53, num. 2, p. 62-64.

Crozat, S. (2012). Chaînes éditoriales et rééditorialisation de contenus numériques. In *Le document numérique à l'heure du web*, Calderan, L., Laurent, P., Lowinger, H. et Millet J. (Eds.), ADBS, Sciences et techniques de l'information ; Le document numérique à l'heure du web de données. pp.179–220

Déotte, J-L. (2010). Bourdieu et Panofsky : l'appareil de l'habitus scolastique, In *Revue Appareil*, novembre. Disponible à : http://appareil.revues.org/1136

Dumas, L. (2016). *Conception de formes de relecture dans les chaînes éditoriales numériques*, thèse de doctorat en informatique de l'Université de Technologie de Compiègne.

Fayet-Scribe, S. (1997). Chronologie des supports, des dispositifs spatiaux, des outils de repérage, de l'information, In *Le savoir et ses outils d'accès : repères historiques*. SOLARIS, les cahiers du Groupe Interuniversitaire de Recherche en Science de l'Information (GIRSIC), num. 4.

Hennicke, S., Gradmann, S., Dill, K., Tschumpel, G., Thoden, K., Morbidoni, C., Pichler, A. (2015). *D3.4 – Research Report on DH Scholarly Primitives*. D2ME Project.

Lackner, M. (2011). « Les diagrammes d'analyse textuelle : une pratique savante de la tradition chinoise », *Lieux de savoir, tome II, Les mains de l'intellect*, pp. 824–844, Christian Jacob, Albin Michel.

Meynet, R. (2013). La rhétorique biblique et sémitique. In *Catholic Theology and Thought*, 2013, num. 72, p. 44-77.

Nakakoji, K., Yamamoto, Y., Akaishi, M., Hori, K. (2015). Interaction design for scholarly writing: hypertext representations as a means for creative knowledge work. In *The New Review of Hypermedia and Multimedia, Special issue: Scholarly hypermedia*, vol. 11, num. 1, Taylor & Francis.

Palmer, C. L., Teffeau, L. C. & Pirmann, C. M. (2009). *Scholarly Information Practices in the Online Environment – Themes from the Literature and Implications for Library Service Development*, rapport, OCLC Research.

Pédauque, R. T. (2006). *Le document à la lumière du numérique. Forme, texte, médium: comprendre le rôle du document numérique dans l'émergence d'une nouvelle modernité.*

Peroni S., Osborne F., Di Iorio A., Nuzzolese A.G., Poggi F., Vitali F., Motta E. (2017). Research Articles in Simplified HTML: a Web-first format for HTML-based scholarly articles. In *PeerJ*. Disponible à : https://doi.org/10.7717/peerj-cs.132

Piotrowki, M. (2016). Future Publishing Formats. In Proceedings of the 2016 ACM Symposium on Document Engineering. ACM, p. 7-8.

de Ribaupierre, H. et Falquet, G. (2017). Extracting discourse elements and annotating scientific documents using the SciAnnotDoc model: a use case in gender documents. In *International Journal on Digital Libraries*, p. 1-16.

Saemmer, A. (2015). *Rhétorique du texte numérique : figures de la lecture, anticipations de pratiques : essai*. Presses de l'Enssib.

Shotton, D. (2009). Semantic publishing: the coming revolution. In *scientific journal publishing*. Learned Publishing, vol. 22, num. 2, p. 85-94.

Unsworth, J. (2000). « Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? » Part of a

symposium on *"Humanities Computing: formal methods, experimental practice"* sponsored by King's College, London, May 13.

Zacklad, M. (2015). Genre de dispositifs de médiation numérique et régimes de documentalité, In *Les genres de documents dans les organisations, Analyse théorique et pratique*, Gagnon-Arguin, L., Mas S. et Maurel, D. (Eds.), pp. 145–183, PUQ, Québec.