École Nationale des Chartes – PSL
Master 1 – Humanités Numériques

Fondations numériques d’un imaginaire

L’écriture créative et communautaire en réseau dans l’exemple de *La Fondation SCP*

(Mini-)Mémoire préparé sous la direction de Carmen Brando et de Thierry Poibeau

Par Perrine MAUREL

Année universitaire 2021-2022

Sommaire

Introduction

« I don’t think people have lost the ability to tell stories as much as they have lost the expectation that it is their place to tell stories.[1] »[10] argumente Cynthia Kurtz[12], consultante en Participatory Narrative Inquiry. Or l’ère des plateformes digitales connectées offre une nouvelle alternative aux individus voulant s’essayer à la narration écrite, orale ou même multimédia, sans distinction de qualité, de compétence, de renommée ou d’exposition. Le rapport à l’objet, au processus de production de ce dernier, au public qui le reçoit s’en retrouvent modifiés : l’exercice d’une narration entièrement numérique diffusée en large public crée une singularité narrative qui, du fait de son accessibilité, a le potentiel de devenir une source d’inspiration. La sélection qui s’effectue ensuite de manière spontanée, entre les œuvres gagnant en notoriété et celles qui demeureront dans les zones d’ombre anonymes du web ainsi que toutes les nuances entre ces deux états, repose sur des phénomènes de diffusion de l’information et de critères d’appréciation subjectifs du plus grand nombre.

Mais les œuvres ne sont pas seules à bénéficier de l’ouverture offerte par la couverture mondiale d’Internet ; l’acte même d’écrire devient un motif de réunion facilité par l’inter-connectivité, où le fait d’écrire avec la contrainte (et l’aide) de l’autre devient le thème principal d’une œuvre, secondé par le thème littéraire de l’objet qui résultera de cet exercice. De manière analogue à la tradition du *renga*, genre de la poésie japonaise du XIII^{ème} siècle se caractérisant par une collaboration entre au moins deux auteurs qui écrivent en se répondant, plusieurs sites Internet constituent un appel libre à la création littéraire en groupe à plus ou moins grande échelle. Cet appel peut prendre plusieurs formes : certains empruntent à la tradition du jeu de rôle en demandant aux auteurs-joueurs d’assumer le rôle d’un personnage et de participer à l’histoire en réagissant aux évènements présentés par les narrateurs (*Ravel Pathfinder 2* [13], *Anarchy*[14]) ; d’autres reposent sur l’acte d’écrire uniquement en permettant aux auteurs de relayer leurs histoires plus librement (*L’Allée des Conteurs*[15], *Necromorial*[16]) ; d’autres encore investissent le procédé de création collaborative plus en détail en codifiant la création d’un univers commun que chacun peut enrichir à sa manière, y compris en reprenant les créations des autres auteurs pour ses propres productions (*Backrooms*[18]).

Dans cette dernière catégorie existe un objet littéraire singulier du fait de sa complexité, sa popularité et son expansion naturelle. Je fais ici référence au site de *La Fondation SCP*[19], une communauté d’écriture et de traduction en ligne. Les origines de sa création retracent le parcours

1 « Je ne pense pas que les gens aient perdu la capacité de raconter des histoires, plutôt, ils se sont convaincus que ce n’était pas leur rôle. » – Traduction par mes soins.

d’une initiative commune spontanée s’étant progressivement organisée et entretenue jusqu’à devenir à ce jour l’une des plus grandes communautés d’écriture au monde. L’existence même d’une communauté aussi large et aussi vivante dépend effectivement du média numérique interconnecté sur lequel elle se fonde, permettant une ouverture et une perméabilité du lieu d’écriture propice à l’inspiration comme à la sociabilité créative. L’initiative finit par dépasser l’auteur, car elle ne dépend plus d’une figure créative particulière, plutôt du processus créatif en lui-même : l’intertextualité devient une composante essentielle de l’existence de cet univers commun, construit pièce par pièce et enrichi constamment par l’ajout de différentes sensibilités, de différents genres littéraires et de nouveau concepts.

Il apparaît alors que la trajectoire des entités nommées à travers les textes permettrait de mettre en évidence l’évolution spontanée des concepts dans la communauté au sens large, puis d’identifier et de caractériser ceux dont la position au sein des textes se trouve consolidée par leur popularité et leur canonicité locale.

Il apparaît également que la figure d’auteur, dans un tel projet, ne peut se satisfaire d’une simple fonction : l’auteur se doit aussi d’être lecteur, brouillant ces deux statuts, et auxquels se rajoute la figure du traducteur.

Il apparaît enfin que la gestion collective d’un tel projet demande l’application de règles consensuelles codifiées et respectées, sujet sociétal que nous pourrons analyser à la lumière des différences entre les différentes sous-communautés de *La Fondation SCP*.

État de l’art

Serge Bouchardon, dans *La valeur heuristique de la littérature numérique*[1], interroge les perspectives ouvertes par l’étude de la littérature numérique. Il souligne en conclusion que la littérature numérique constitue un « laboratoire du numérique » : la littérature numérique redéfinit partiellement des objets que l’on pensait connus, tels que les textes, les auteurs, les lecteurs… et la narrativité elle-même. En outre, l’auteur relève que la littérature numérique fait appel à des compétences techniques indépendantes de la littérature, mises à profit par le format, et constitue donc un espace d’expérimentation littéraire et anthropologique foisonnant, d’où son intérêt pour la recherche. Serge Bouchardon observe que la littérature numérique interroge également les rapports entre littérature et communication, se plaçant parfois à la croisée de ces deux domaines, tant au niveau des technologies et méthodes employées qu’au niveau du contenu. C’est dans la continuité de cette pensée que j’effectue mes recherches, en me concentrant toutefois davantage sur une approche de critique génétique plutôt que de science de l’information. Je sélectionnerai toutefois les méthodes en science de l’information qui me permettront de quantifier les effets narratifs des textes étudiés.

L’écriture collective a été étudiée par Pauline Gosetti dans son mémoire *L’écriture collective : un* jeu *d’auteur(s)*[2]. Elle y interroge la place de l’auteur et du lecteur dans les œuvres collectives, cherchant à définir la figure autoriale dans leur contexte et à définir les modalités de réception qui environnent les textes issus d’une telle initiative. L’autrice observe que les dispositifs d’écriture propres à l’écriture collective demandent la définition d’un cadre précis et de règles permettant de véritablement concrétiser le lieu d’écriture ; elle insiste toutefois sur l’importance des interactions au sein de ce cadre, lesquelles en font l’essence. Les conclusions de Pauline Gosetti ont nourri mes propres questions de recherche : l’étude des forums et des interactions entre les auteurs pendant la production de leurs œuvres d’une part, et la part de lecture par les auteurs comme une modalité nécessaire pour la survie et l’évolution de l’univers commun, constituent deux pistes de recherche que j’aimerais examiner.

Concernant l’objet d’étude que constitue le site *La Fondation SCP*, ce dernier a été étudié par Eric Thomas Newsom dans sa thèse *Participatory Storytelling and the New Folklore of the Digital Age*[3]. L’auteur a choisi d’étudier l’émergence de ces projets de narration collaborative sous le prisme de la folkloristique, tout en accordant une attention toute particulière au rôle que joue le numérique dans la relation entre l’auteur, le lecteur et le texte. Eric Newsom prise l’idée d’une dépendance mutuelle entre la communauté d’écriture et le produit d’écriture : la communauté guide

la création du produit, lequel entérine alors l'existence de la communauté et la codifie petit à petit selon les éléments fondateurs qui ressortent du projet. Son texte ayant largement couvert la question de la formation de ces communautés, j'ai choisi de prendre une direction différente dans ce (mini) mémoire en me concentrant sur la formation de l'univers commun. Or, dans un livre de 2016 qu'Eric Thomas Newsom a écrit conjointement avec Shira Chess, *Folklore, Horror Stories and the Slender Man: the Development of an Internet Mythology*[4], la question de la malléabilité du personnage et de la mise en forme de son mythe à travers sa redéfinition créatrice aux mains de plusieurs narrateurs s'avère centrale et fournit plusieurs pistes de recherches sur le sujet. Les auteurs de l'ouvrage affirment ainsi que le mythe entourant la figure du *Slender Man* est le produit d'une négociation spontanée se réalisant à travers ses différentes apparitions jusqu'à l'acceptation d'une constante consensuelle faisant le mythe, bien que des alternatives marginales existent encore en dehors du canon général. Cette notion de négociation d'une ligne directrice n'est pas sans rappeler la confrontation des différents canons sur le site, et rejoint ma théorie première selon laquelle certains éléments de l'univers de la Fondation SCP deviennent spontanément des pierres emblématiques dans sa construction ; la question étant d'identifier les caractéristiques pouvant mener à ce statut. Les auteurs de l'ouvrage ayant conclu en annonçant que le phénomène de *Slender Man* annonçait la nature des narrations à venir, il serait intéressant d'étudier l'évolution de *La Fondation SCP* aujourd'hui en la comparant aux prévisions de l'ouvrage, cette dernière étant une œuvre contemporaine au *Slender Man* mais toujours en expansion.

Dans une moindre mesure, l'étude du format encyclopédique m'a semblé pertinente et j'ai ainsi étudié le livre *Wikipédia, au cœur de la plus grande encyclopédie du monde*[5] publié en 2021 par Rémy Mathis ; ceci afin de me familiariser avec les enjeux et les modalités des encyclopédies en ligne, modèle que reprend le site de *La Fondation SCP*. De même, je me suis intéressée au mémoire de Marion Lefebvre portant sur *La juridictionnalisation de la modération des sites internet*[6], qui étudie entre autres la modération effectuée sur le site de *La Fondation SCP*. La modération étant un aspect primordial du maintien d'une communauté et englobant des sujets créatifs autant que sociaux, j'ai voulu voir si cette lecture pouvait m'ouvrir certaines perspectives de recherche.

En ce qui concerne la méthodologie appliquée, je me suis référée à la thèse de Maud Ehrmann soutenue en 2008 à l'Université Paris Diderot, *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*[7]. Cette thèse expose en effet une définition extensive du concept d'entité nommée et présente les diverses méthodes employées pour les étudier.

Présentation des éléments de recherche

Cette partie du (mini) mémoire sert à présenter l'objet de recherche dans toutes ses spécificités ainsi que les méthodes et méthodologies employées pour cette étude.

I. L'objet de recherche : un projet d'écriture communautaire unique

Avant toute chose, il convient d'étudier l'histoire et l'évolution du site d'écriture en tant que tel afin de contextualiser la façon dont il est devenu un ovni narratif. Ce sujet est traité de manière interne par des utilisateurs étant présents aux origines ou ayant retracé ces dernières, sur le document *Centre de l'Histoire de l'Univers*[20] du site, publié le 18 Février 2013.

En 2007, un utilisateur posta de manière anonyme sur le forum 'paranormal' de *4chan*[21] un document intradiégétique rédigé du point de vue d'une organisation secrète, nommée la Fondation SCP, cherchant à confiner et étudier une créature anormale, dont l'existence contredirait les lois naturelles de ce monde. Le document s'intitulait *SCP-173*[22]. Ce format particulier, dit du 'rapport', fut ensuite repris sur le même forum par d'autres utilisateurs pour exposer leurs propres créations, créant au fil du temps une collection conséquente de rapports partageant un même vocabulaire, un même style d'écriture et un même univers : en ce sens, l'univers de la Fondation SCP a bénéficié de la tradition des *creepypastas*, des œuvres horrifiques numériques ayant pour vocation d'être partagée en masse en tant que légendes urbaines. En 2008 naquit la première version d'une encyclopédie en ligne recensant les rapports de manière ordonnée, laquelle offrait également la possibilité de discuter dans les forums de l'univers et des écrits afin de les questionner et de les commenter. Le consensus de la communauté servait alors et sert encore à exclure les œuvres jugées inadaptées au projet ou trop pauvres en qualités littéraires. Cette communauté prit le nom de l'organisation désormais éponyme : *La Fondation SCP*[24].

Rapidement, ces discussions motivèrent l'organisation du forum en un lieu d'écriture où les auteurs commençaient à élaborer ensemble des idées et des principes à appliquer à leurs créations, offrant des retours critiques sur leurs œuvres en cours d'écriture. Le format du rapport étant assez limité bien qu'emblématique, les productions littéraires se sont étendues à d'autres formats : les 'contes', production littéraire en prose ou en vers bien plus libre et plus proche d'une narration traditionnelle ; les 'guides', production documentaire plutôt que littéraire voulant expliquer certains principes et néologismes propres à l'univers pour permettre leur réutilisation par les auteurs

néophytes ; les ‘centres’, production littéraire et documentaire présentant un aspect de l’univers de manière extensive.

Le site de *La Fondation SCP* étant anglophone d’origine, sa diffusion à travers le monde fut facilitée. Plusieurs initiatives de traduction virent alors le jour, visant à traduire et diffuser les écrits du site originel dans un espace linguistique particulier. Grâce à cette démocratisation du projet d’écriture, plusieurs auteurs non-anglophones rejoignirent ces sites de traduction pour produire, dans leur langue native, des écrits reprenant l’univers. La reconnaissance officielle de ces sous-communautés a donné lieu à l’appellation de ‘branche’, désignant un site officiel chargé de la gestion du projet dans une langue distincte. Il en existe aujourd’hui dix-sept, dans les locutions :

anglophone,
francophone,
italienne,
thaï,
japonaise,
lusophone,
germanophone,
tchèque,
russophone[2],
coréenne,
sinophone,
hispanophone,
polonaise,
vietnamienne,
ukrainienne[3],
du chinois traditionnel[4].

La dernière et dix-septième branche, SCP-INT[25], est une branche internationale utilisant l’anglais et servant de centre commun à toutes les branches non-anglophone pour y poster leurs œuvres natives traduites en anglais, permettant ainsi des traductions inter-branches et plus seulement de l’anglais aux autres langues. Quatre branches non-officielles attendent leur officialisation, étoffant pendant ce temps leur catalogue de traductions et de productions originales : une branche turque, une branche indonésienne, une branche grecque et une branche nordique. Chaque communauté dispose de ses propres règles d’écriture et d’accès au site. À travers les seize sites officiels[5], la communauté d’écriture et de traduction de la Fondation SCP compte[6] 14 766 utilisateurs actuellement actifs, 44 156 œuvres originales produites par 7729 auteurs distincts. Ce dernier

2 Actuellement inactive, les utilisateurs russes ayant été bannis par l’hébergeur du site le 24/05/2025 en réponse à une tentative de hack provenant de la Fédération Russe et en guise de protestation contre l’assaut militaire mené par la Russie contre l’Ukraine depuis le 24/02/2022. La communauté d’écriture en elle-même désapprouve dans l’ensemble cette mesure et demande qu’elle soit levée.

3 Actuellement inactive en raison de la situation militaire en Ukraine depuis le 24/02/2022.

4 Anciennement ‘branche taïwanaise’. L’appellation officielle a été changée afin d’éviter de placer la branche sinophone en infraction de la loi locale en Chine.

5 SCP-INT a été volontairement exclue du décompte : en effet, il s’agit de la réunion d’utilisateurs et de textes provenant à l’origine des autres branches, et ceux-ci seraient comptés deux fois.

6 Les chiffres proviennent du site de statistique scpper[9] et ont été compilés pour la dernière par mes soins le 25/05/2022 à 10h18. Les chiffres des branches russes (559 actifs, 2849 œuvres, 1179 auteurs) et ukrainiennes (52 actifs, 224 œuvres, 42 auteurs) ont été comptabilisés car leur inactivité est récente.

chiffre est à revoir légèrement à la baisse, puisque quelques rares auteurs publient sur plusieurs branches différentes et ont donc été comptés plusieurs fois.

La question de la cohérence est vite devenue essentielle à l'existence de cet univers collaboratif, sa taille massive impliquant également un contrôle moindre sur les éventuelles contradictions qui pourraient exister entre deux œuvres. Le parti pris de la communauté, pour survivre tout en se développant, fut donc d'abandonner le principe de 'canon fixe' au profit du principe de 'canon fluide'[7] : chaque auteur sélectionne les éléments de l'univers qui l'intéressent de manière à les agencer dans une vision plausible qui lui est propre. Cette vision est généralement exprimée à travers ses œuvres et celles des auteurs qui reprendraient ses idées ; s'il existe donc des éléments plus consensuels que d'autre, il n'existe pas de base narrative complètement inamovible à proprement parler, d'où la dénomination 'd'ovni narratif'. Le terme de 'canon' fut ensuite généralement appliqué à de larges projets d'écriture collaborative au sein même de la communauté, reposant sur une vision particulière de l'univers qui serait alternative aux interprétations usuelles. Par exemple, le canon *Dos au Mur*[26] créé par l'auteur DrTesla et étoffé par 8 autres auteurs sur la branche francophone expose une version de la Fondation SCP géopolitiquement fragile et menacée par les diverses influences étatiques du monde entier ; là où le consensus usuel veut plutôt que la Fondation SCP agisse en collaboration avec les états et soit même financée par eux. Un même auteur peut jouer entre les différents canons et les différentes interprétations à sa guise, ne se confinant pas à une vision unique.

De ce fait, la Fondation SCP est devenu une communauté particulièrement diverse tant au niveau des formats que des genres d'écriture. Bien qu'elle se rattache à l'origine aux genres horrifique, fantastique et scientifique, le projet compte désormais de nombreuses représentations des différents genres littéraire de contenu, de registre ou de forme variée. Pour entériner l'aspect collaboratif du projet, l'ensemble des œuvres liées à la Fondation SCP fut placé sous la licence Creative Commons Attribution-ShareAlike 3.0 License[27], laquelle permet la distribution et la modification libre du contenu tant que l'attribution de la parenté de l'œuvre est correctement effectuée. La communauté ne se limite ainsi donc pas aux sites d'écriture et de traduction : de nombreux agents extérieurs ont repris à leur compte l'univers pour créer des textes, des jeux vidéos, des séries filmées ou des jeux de société.

Au delà de l'aspect littéraire, cette œuvre constitue un objet de recherche particulièrement riche. En effet, il s'agit d'un exemple de premier choix pour le domaine de la critique génétique en raison de la documentation rigoureuse des procédés de création qui se déroulent au sein de la communauté, agissant sur l'auteur et son œuvre avant la rédaction, pendant la rédaction et souvent

7 Seule la branche italophone conserve volontairement le principe de canon fixe, demandant que les travaux des auteurs s'articulent de manière cohérente les uns avec les autres.

même après la publication. L'univers qui résulte de ces procédés est particulièrement complexe et remet en cause les principes de permanence des entités et des concepts, lesquels sont pourtant caractéristiques de l'œuvre canonique habituelle : écrite par un auteur unique qui, dans la conception classique de l'écriture, ne saurait se contredire s'il est compétent. L'incohérence assumée et inhérente à cette œuvre d'écriture complètement spontanée permet une étude comparative impossible dans un autre contexte : l'ensemble des textes constitue un écosystème naturel de néologismes et d'entités nommées en compétition virtuelle les uns avec les autres, certains devenant emblématiques jusqu'à dépasser même le cadre du site d'écriture et d'autres se cantonnant aux œuvres de l'auteur les ayant créés. L'étude de leur évolution à travers le temps, l'espace et les auteurs devrait permettre de mettre en valeur un certain nombre de caractéristiques propres aux entités nommées devenues 'fondamentales' dans l'univers et servant de manière effective à son expansion.

L'objectif principal de mes recherches est donc de créer un réseau d'entité nommées comprenant des lieux, des personnages, des événements, des organisations, des créatures, des classifications, des œuvres, des néologismes et des phénomènes avec une focalisation sur leurs origines extradiégétiques et leurs liaisons intradiégétiques, puis de l'interroger de manière ordonnée. Les données ainsi obtenues en matière d'existence littéraire et créative seront ensuite contextualisées dans le cadre de l'existence social et communautaire de l'objet de recherche, afin d'examiner tous les tenants et aboutissants propres à l'écriture créative et communautaire en réseau.

Dans un premier temps, je me limiterai au corpus de la branche francophone, en n'étudiant que les contes et les rapports : soit un corpus de base de 1272 textes.

II. Les méthodes de recherche : la chaîne de traitement et les données de terrain

1) Préparation du corpus

Différencier les textes pertinents des autres a été très facile en raison de la politique de 'tag' de la branche francophone, obligeant à mettre les étiquettes 'conte' sur les contes, 'scp' sur les rapports et 'fr' sur les productions francophones. C'est cet élément que j'ai donc utilisé pour effectuer une discrimination.

L'ensemble du corpus se trouvant librement disponible en ligne, la première étape de ma méthode fut de créer un outil de *web scrapping* afin de récupérer les textes ainsi que leurs

métadonnées. La première version de cet outil fut créée dans le langage R : chaque page effectuant des appels vers des serveurs pour récupérer les métadonnées stockées autre part que dans le code HTML, il m'a fallu reproduire ces appels pour obtenir les données en question. J'ai bénéficié de l'aide de Corentin Poupry, qui m'a aidé à adapter son code[28] du Python au R à cet effet. J'ai ensuite adapté mon code en Python via la bibliothèque `BeautifulSoup`[29], puis la bibliothèque `lxml`[30].

Les données ainsi récupérées ont pour vocation d'être organisées dans deux variables *`dataframe`*[8] différentes : l'une récupère les informations du module de crédit introduit manuellement par les utilisateurs du site, l'autre récupère les métadonnées automatiquement générées lors de la création de la page. L'importance de cette distinction permet d'obtenir des données généralement plus précises de la part des utilisateurs, et de vérifier l'efficacité du système de crédit mis en place sur la branche francophone. Le texte à étudier est également inclus dans ce tableau, après avoir été élagué des balises HTML[9].

Un module de crédit manuel désigne un élément HTML remplit lors de la publication d'une page par son auteur ou traducteur, qui peut contenir plusieurs informations différentes dont les plus importantes sont le nom de l'auteur originel de l'œuvre, la date de publication de l'original et les droits d'auteur si des éléments extérieurs ont été utilisés ou mentionnés dans le texte, par exemple des images.

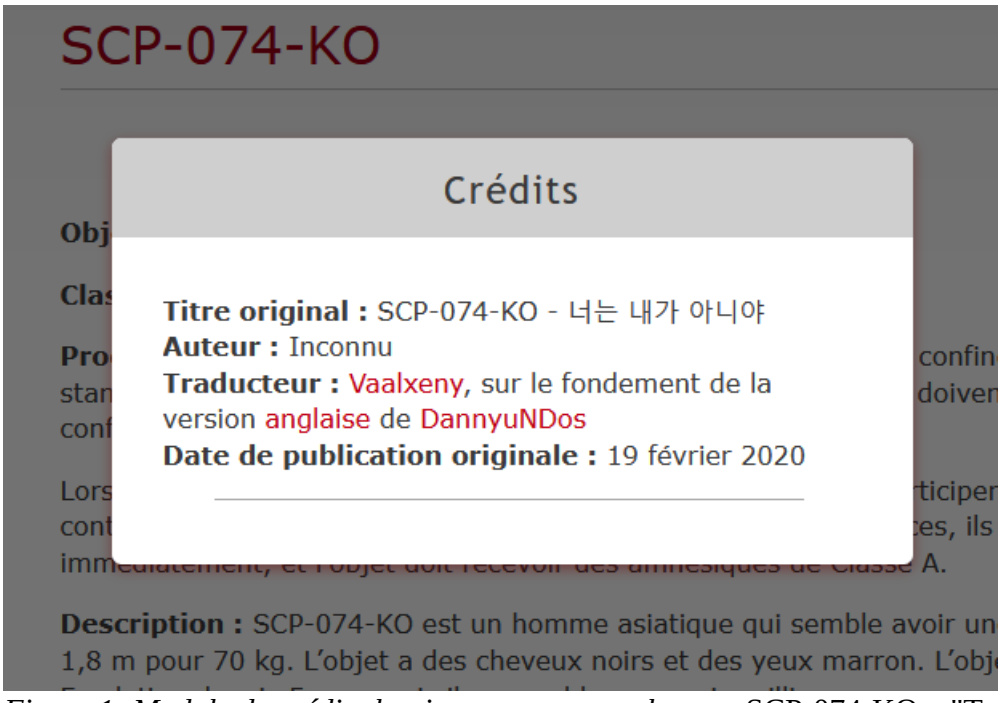

Figure 1: Module de crédit classique, concernant le texte SCP-074-KO – "Tu n'es pas moi"[31] *de la branche coréenne.*

8 Classe issue de la bibliothèque Python *numpy*.
9 À l'exception des balises URL qui sont importantes pour établir un réseau.

Les dataframes suivent le schéma suivant :

Tableau 1 : `user_born_data`

Désigne les données renseignées par l'utilisateur, en plus de quelques métadonnées natives renseignées également pour plus d'exhaustivité. Une ligne par texte, treize colonnes.

Informations issues du module de crédit

Certaines colonnes peuvent être renseignées ou laissées vides selon si le module de crédit comporte ou non un champ équivalent. En cas de module de crédit absent, la mention « Module de crédit absent » sera ajoutée en lieu de donnée.

- `titre` *(char)* : titre original de l'œuvre
- `auteurices` (*list[char]*) : liste des auteurs de l'œuvre
- `traducteurices` (*list[char]*) : liste des traducteurs de l'œuvre, si pertinent[10]
- `date_creation` (*char*) : date de création de l'œuvre
- `date_traduction` (c*har*) : date de traduction de l'œuvre, si pertinent
- `informations_images` (*char*) : informations de copyright des images
- `remerciements` (*char*) : remerciements de l'auteur
- `commentaires` (*char*) : texte additionnel désigné comme étant une remarque particulière sur le texte
- `inclassable` (*char*) : tout segment de texte qui n'aurait pas été identifié par le programme comme appartenant aux catégories précédentes

Informations issues des métadonnées de la page

- `tags` (*list[char]*) : liste des étiquettes données à la page par les utilisateurs
- `texte` (*char*) : le texte de la page
- `lien` (*char*) : le lien menant à la page étudiée, sert d'identifiant unique

Autres informations

- `temps_requete` (*float*) : le temps pris par le programme pour récupérer et ordonner dans une ligne du dataframe les métadonnées

10 Peut parfois englober par erreur des notes de traduction. Il sera possible de corriger cela dans le futur en ne se référant qu'aux éléments disposant d'une URL menant vers un profil d'utilisateur.

Tableau 2 : `digital_born_data`
Désigne les données obtenues dans les métadonnées de la page. Une ligne par texte, sept colonnes.
Informations issues des métadonnées de la page • `titre` *(char)* : titre original de la page • `auteurice` (*char*) : créateur de la page, peut être le traducteur ou l'auteur si l'œuvre est originaire de la branche francophone • `date` (*timestamp*) : date de création de la page, peut être la date de publication de la traduction ou la date de publication du texte si l'œuvre est originaire de la branche francophone • `tags` (*list[char]*) : liste des étiquettes données à la page par les utilisateurs • `texte` (*char*) : le texte de la page • `lien` (*char*) : le lien menant à la page étudiée, sert d'identifiant unique
Autres informations • `temps_requete` (*float*) : le temps pris par le programme pour récupérer et ordonner dans une ligne du dataframe les métadonnées

2) Étiquetage XML

Une fois ces données ordonnées, un travail de profondeur doit être effectué sur les textes, trouvés dans la colonne 11 du dataframe `user_born_data`. Ce travail est un effort d'étiquetage XML des entités nommées reconnaissables, ayant pour objectif d'en constituer un index exploitable. Une première application entièrement manuelle de cette approche a été effectuée sur l'ensemble des textes constituant l'œuvre *Une Tragédie du Roi Pendu*[32] – que je désignerai ensuite sous le terme de 'texte-exemple' –, une pièce de théâtre francophone sélectionnée en raison de la large richesse de ses références, tant intertextuelles que réelles.

Les entités nommées classiques sur lesquelles je me suis focalisée sont : les personnages, les lieux, les créatures, les classifications, les organisations et les objets d'art titrés, sous toutes leurs

formes. J'ai également choisi de m'intéresser aux évènements et aux néologismes, deux types d'entités nommées particuliers : les évènements sont difficiles à définir car toute action ou acte au sein d'un texte peut potentiellement être évènement. J'ai donc décidé de limiter la définition de « l'évènement » à un incident ayant une valeur historique, pouvant se constituer comme un point chronologique, et ayant reçu au sein des textes une dénomination propre et particulière ayant pour nature celle des noms propres. Quant aux néologismes, ils sont, à ma connaissance, un genre d'entité nommée assez peu étudié, appartenant davantage au domaine de la linguistique que de la reconnaissance d'entités nommés dans la conception traditionnelle du TAL. Toutefois, en me calquant sur la définition au sens large de « nom propre et autres expressions »[7] des entités nommées, il est concevable de traiter les néologismes en tant que tel. En effet, certains de ces néologismes ont pour nature de mot celles des noms propres, induisant une typification du texte allant dans mon sens ; et les questions que j'ai été amenée à me poser sur leur sujet rejoignaient systématiquement celle des autres entités nommées, facilitant donc une étude liée. Il convient également de mentionner que ce que j'entends par 'néologisme' désigne à la fois les mots nouveaux issus d'une transformation grammaticale, et les mots dont la forme morphologique existe déjà mais dont le fond sémantique s'est vu être redéfini dans le cadre de la communauté d'écriture.

J'ai donc élaboré en guise d'exemple une documentation ODD titrée *oddTragédieRoiPendu* qui suffisait dans le cadre de mon texte-exemple, et que je serai amenée à étoffer en avançant dans mes recherches. Celle-ci suit généralement le format de balisage TEI, adapté selon mes besoins. J'ai donc créé ou repris une balise XML par entité nommée, puis j'ai identifié en priorité pour chaque entité nommée :

Tableau 3 : étiquetage XML

Cas général		si elle fait référence ou non à un équivalent réel :	
		Personnages	grâce à l'utilisation du caractère spécial '`#`' en début de `xml:id` ;
		Général	grâce à l'attribut `reel`.
		l'auteur à l'origine de l'entité nommée grâce à l'attribut `author` :	
		Réelles	c'est-à-dire la personne à laquelle est attribuée la paternité dans le monde réel ;
		Imaginaires	c'est-à-dire le personnage auquel est

		attribuée la paternité dans l'univers imaginaire.
		l'affiliation de l'entité nommée à une organisation réelle ou imaginaire, grâce à l'attribut `affiliation`.
réelles		si elle désigne une variante imaginaire de cet équivalent réel, ayant subi des altérations.
imaginaires		la source de l'entité nommée comme étant le lieu de sa première mention, c'est-à-dire un titre d'œuvre ou une URL, grâce à l'attribut `source` ;
		la branche d'origine de l'entité nommée si celle-ci provient de la Fondation SCP, grâce à l'attribut `branch` ;
		le sous-type d'entité nommée, grâce aux attributs `type` et `sous-type` ;

Au cours de cet étiquetage manuel, deux problématiques insoupçonnées au sein de ma méthode de travail se sont présentées à moi :

La première concernait le statut des néologismes. Si leur assimilation au groupe des entités nommées permet leur étude suivant la même méthodologie, il s'est avéré que certains de ces néologismes correspondaient également à un autre type particulier d'une autre entité nommée. Cette catégorisation croisée concerne surtout les entités nommées de 'classification' : en effet, nombre de ces classifications sont dérivées de références à des mots préexistants dont le sens a été modifié au sein de l'univers littéraire étudié. Le problème ne s'est pas présenté pour les autres catégories d'entités nommées car selon la définition établie plus haut d'un 'néologisme' dans le cadre de notre étude, les mots inventés sans fond sémantique notoire, ne servant qu'à référer à un lieu, un personnage… imaginaire, ne sont pas considérés comme des néologismes. Ainsi, « Classe Euclide » est un néologisme, là où le nom « Hématite » ne l'est pas.

La seconde problématique concernait la présence d'un même élément sur divers niveaux d'abstraction. L'exemple le plus explicite concerne *La Tragédie du Roi Pendu*[33], une œuvre anglophone – à ne pas confondre avec *Une Tragédie du Roi Pendu*, notre texte-exemple francophone qui reprend et étoffe la narration anglophone. Cette appellation désigne ainsi dans le même temps :

- *SCP-701 – La Tragédie du Roi Pendu*[33], un rapport fictif de la branche anglophone ;
- *La Tragédie du Roi Pendu*, un ensemble de phénomènes d'altérations littéraires portant sur une pièce de théâtre fictive ;
- *La Tragédie du Roi Pendu*, une pièce de théâtre fictive.

On observe ainsi divers degrés d'abstraction et de compréhension sur une même entité nommée : le degré du réel portant sur le texte appartenant au corpus étudié ; puis, un degré d'abstraction en dessous, sur un ensemble de variations d'une même pièce de théâtre ; puis, encore en dessous, la pièce de théâtre originelle, sans variation.

J'ai jugé pouvoir résoudre ces deux problématiques par l'utilisation de balises imbriquées. Puisque j'ai pris le parti d'utiliser des `xml:id` uniques pour chaque entité, il m'est possible de faire des références entre mes différentes balises et d'associer ainsi plusieurs classes. Voici donc la documentation de deux éléments XML personnalisés, `neologisme` et `coucheInterne`, directement tirées de mon fichier *`oddTragédieRoiPendu`* :

1.7. <coucheInterne>

<coucheInterne> Lorsqu'une entité référencée dans le texte dispose à la fois d'une présence intradiégétique et extradiégétique, cet élément est utilisé dans une balise

Module	derived-module-oddbyexample		
	Attributes		
	type	Indique le type de l'entité.	
		Status	Optional
		Datatype	teidata.word
	title	Indique l'œuvre référencée, si c'est une œuvre.	
		Status	Optional
		Datatype	teidata.word
Attributes	author	Indique l'auteur de l'entité, si elle est fictive.	
		Status	Optional
		Datatype	teidata.word
	date	Indique la date de publication de l'entité, si c'est une œuvre.	
		Status	Optional
		Datatype	teidata.word
	publisher	Indique l'organisme ou la personne à l'origine de cette édition de l'entité, si c'est une œuvre.	
		Status	Optional
		Datatype	teidata.word

1.21. <neologisme>

<neologisme> Désigne un néologisme (au sens large) propre à l'univers étudié : soit un mot entièrement nouveau dont la morphologie est inventée, soit un mot existant dont la signification a été modifiée.

Module derived-module-oddbyexample

Attributes

type	Indique si le néologisme est un mot emprunté et redéfini, ou un néologisme au sens classique du terme. Status Optional Datatype teidata.word
author	L'auteurice qui a créé ce néologisme. Status Optional Datatype teidata.word
branch	Désigne la branche d'origine du néologisme en référant son identifiant (EN, FR...) Status Optional Datatype teidata.word
definition	Donne une définition du néologisme. Status Optional Datatype teidata.word
nature	Si le néologisme est un mot nouveau, établit la catégorie spécifique de sa création (mot-valise...). Status Optional Datatype teidata.word

Attributes

origine	Si le néologisme est un mot nouveau, établit le ou les concepts/mots parents. Status Optional Datatype teidata.word
source	Désigne la source originelle dont vient le néologisme. Peut prendre la forme d'un identifiant renvoyant à un objet spécifique, ou d'un lien URL si la première mention s'est faite en ligne. Le lien URL indique en priorité le texte traduit en Français s'il existe, puis le texte original si la traduction n'existe pas. Status Optional Datatype teidata.word
traducteur	Désigne la personne à l'origine de la traduction du mot et qui a donc créé sa version française. Status Optional Datatype teidata.word
traduction	Désigne la traduction du néologisme dans sa langue originelle. Status Optional Datatype teidata.word

3) Chaîne de traitement actuelle et projections de Topic modeling

Une fois l'étiquetage terminé, il suffit d'utiliser des formules de statistiques classiques afin d'obtenir de premiers résultats et en retirer de premières conclusions. Certains outils propres aux langages utilisés (les librairies `ggplot2`[34] pour R, `matplotlib`[35] pour Python et le chemin Xpath pour XML) permettent également d'effectuer une première modélisation numérique et graphique. En conclusion, la chaîne de traitement de mes données est la suivante :

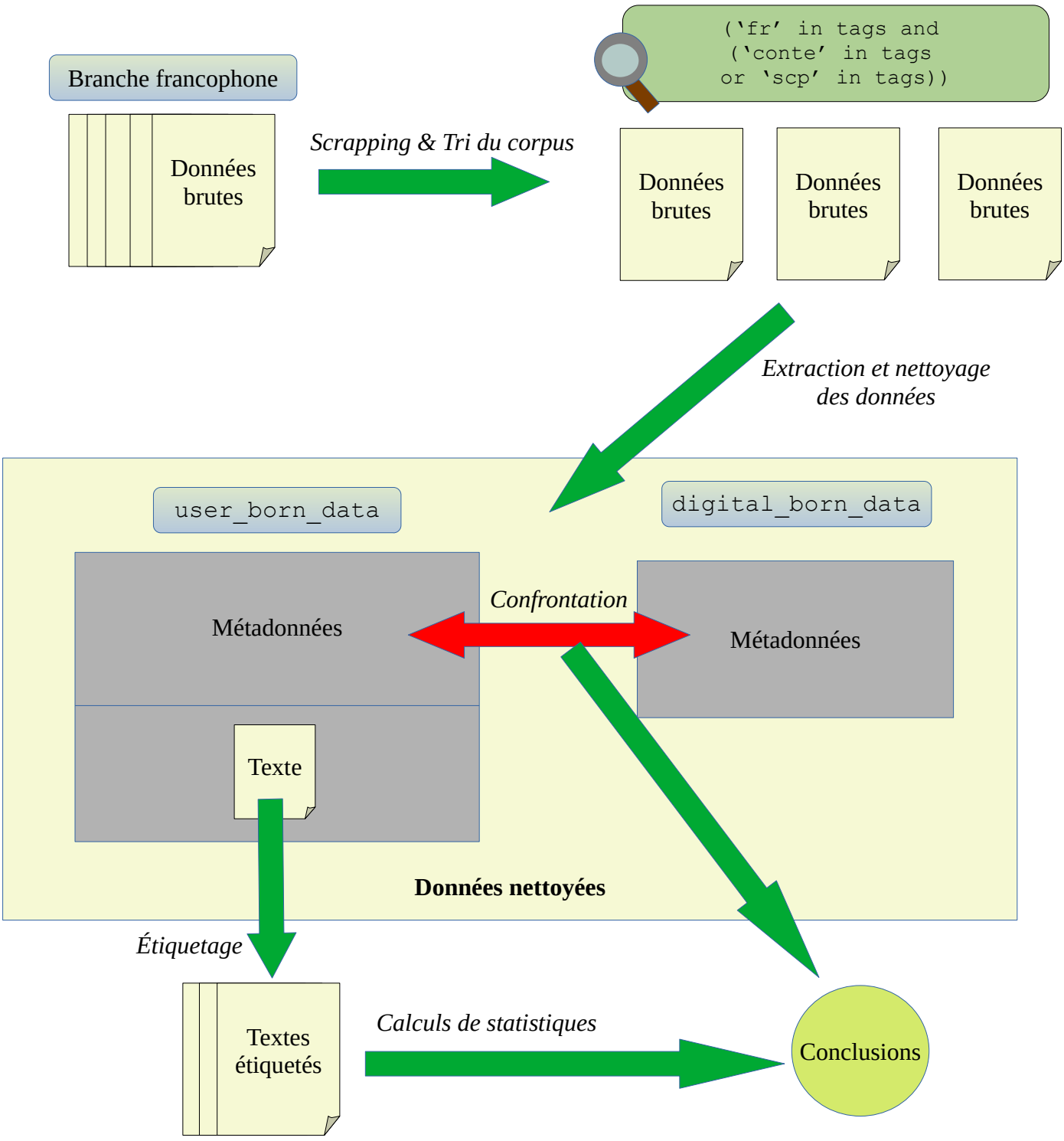

Figure 2: Chaîne de traitement des données

En outre, voici un exemple possible de modélisation des réseaux d'entités nommées grâce à la sérialisation RDF/XML. Cette modélisation se limite encore au stade théorique et conceptuel pour le moment :

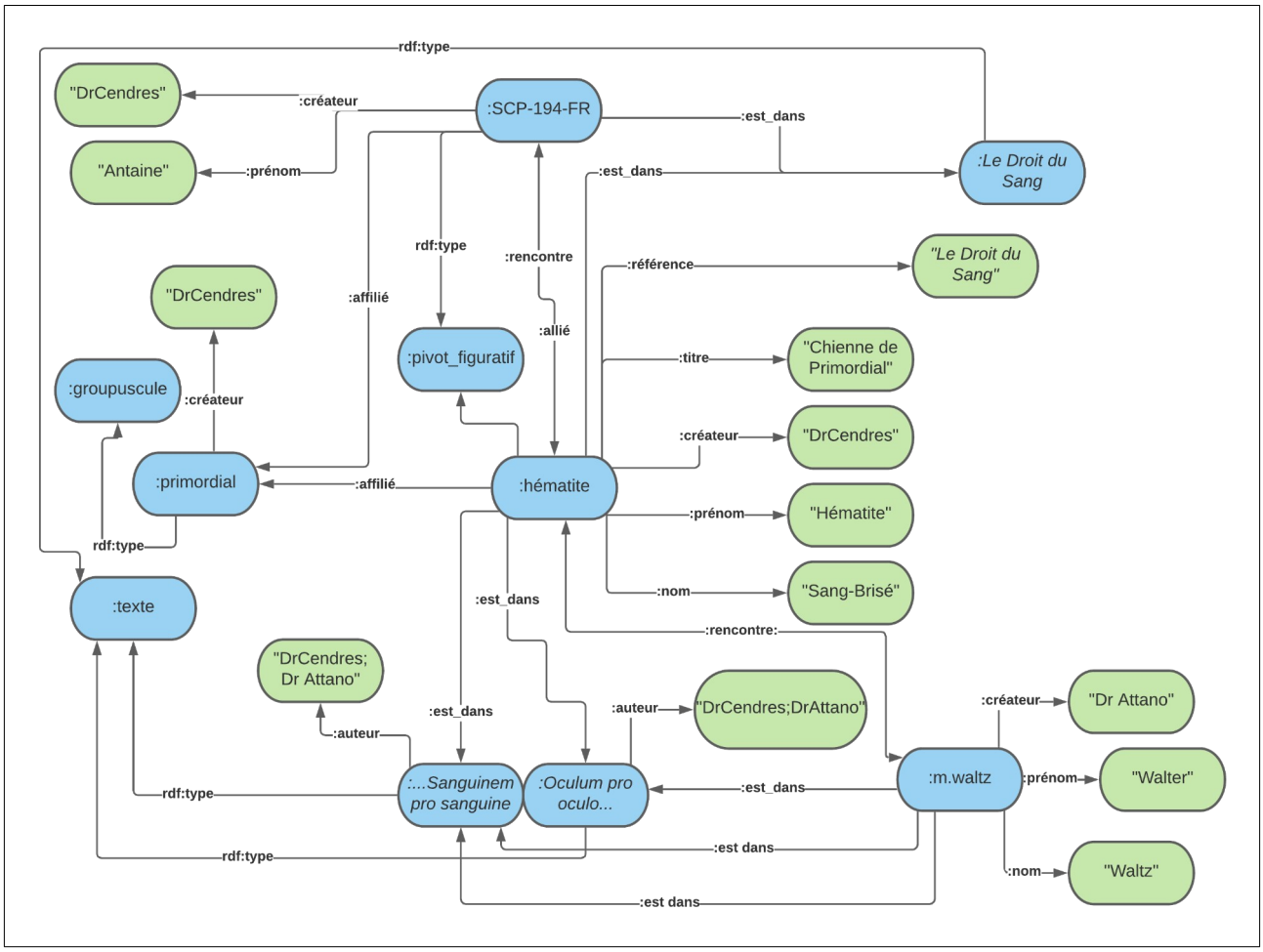

Figure 3: Graphe RDF contenant des informations choisies qui concernent l'entité nommée Hématite[36] de la branche francophone.

Comme l'illustre ce schéma, l'objectif d'un tel réseau serait d'identifier les différents croisements et composantes entre entité nommées, tout particulièrement au niveau des auteurs et des branches. De cette manière, il serait possible d'établir des zones de concentration et d'identifier les entités nommées se trouvant en leur centre, autour desquelles gravitent les procédés créatifs.

Il sera alors possible de quantifier leur importance en établissant un indice de canonicité, concept directement inspiré des travaux de Jean Barré[8] et redéfini dans le cadre de cette étude. Cet indice caractériserait la présence d'un même élément au sein d'un échantillon de textes diversifié tant au niveau des auteurs que celui de la branche d'origine.

III. Résultats de la recherche, conclusions et ouvertures

1) Étude comparée des néologismes

En établissant un ensemble de néologismes choisis sur la base de ceux existant[37], j'ai pu rechercher leur nombre d'occurrence dans les textes du corpus et estimer leur fréquence, en cherchant à étudier les prédominants.

La liste de néologismes choisis était la suivante :

Tableau 4 : néologismes étudiés

Morphologie francophone	Définition	Branche d'origine du concept
amnésiant	Médicament effaçant les souvenirs.	EN
amnésique	Variation orthographique de 'amnésiant'.	EN
anart	Art anormal.	EN
cinquiste	A trait au cinquisme.	EN
cinquisme	Dogme de la Cinquième Église.	EN
daevite	Peuple barbare de conquérants non-humain du Sud de la Russie.	EN
diacrinochrome	État de présenter des motifs différenciant le sujet de ses pairs.	FR
gendastre	Membre de la Gendastrerie.	FR
Gendastrerie	Corps de la gendarmerie chargé des infractions anormales.	FR
humes	Unité de mesure de la réalité.	EN
mekhanite	A trait à l'Église du Dieu Brisé.	EN
mémétique	Information culturelle se diffusant de manière anormale.	EN
orthothan	Peuple de réfugiés.	FR
sarkiste	A trait à la magie du corps et des offrandes.	EN
SCP	Acronyme 'Sécuriser, Contenir, Protéger' devenu par abus de langage la manière de se référer à une anomalie confinée par la Fondation SCP.	EN
singularme	Arme de destruction massive.	EN
télékill	Alliage de métaux absorbant les informations mémétiques dans ses environs immédiats.	FR

J'ai sélectionné des néologismes provenant de la branche anglophone et de la branche francophone afin de diversifier ma sélection et d'étudier l'impact des néologismes natifs ou traduits

dans les productions francophones originales. J'ai également choisi de regrouper ensemble les mots de la même famille lexicale en recherchant prioritairement leur racine commune. 'Cinquisme' et 'cinquiste' sont donc étudiés ensemble sous la dénomination 'cinquis', de même pour 'gendastre' et 'gendastrerie' sous la dénomination 'gendastre'. En revanche, j'ai fait le choix d'étudier 'amnésiant' et 'amnésique' de manière séparée afin d'obtenir quelques premières données sur les variations orthographiques.

La modélisation visuelle des fréquences de ces néologismes est la suivante :

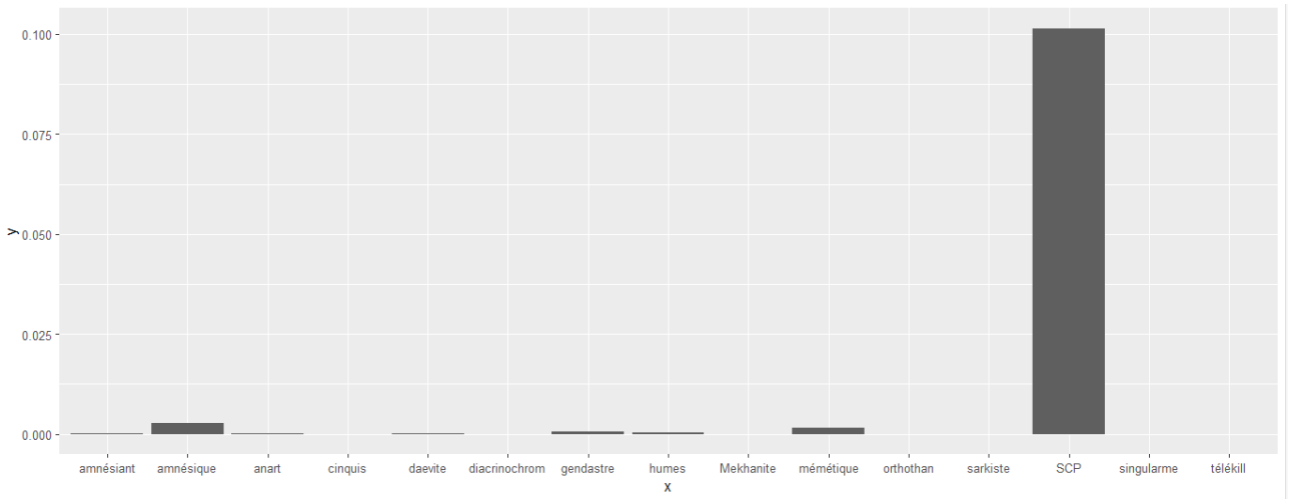

Figure 4: Graphe `df`*, néologismes en abscisses et fréquence*10 en ordonnées*

On constate une variance particulièrement élevée en raison de la forte disparité entre l'élément de tête 'SCP' et les autres néologismes. Je n'ai donc pas jugé utile de calculer la moyenne arithmétique de leur fréquence, celle-ci étant dès lors peu représentative et donc peu pertinente.

Pour expliquer ce phénomène, j'ai émis l'hypothèse que le néologisme principal 'SCP' se trouvait être très fortement utilisé dans le format des rapports ; en effet, la description des anomalies confinées étant effectuée systématiquement sous la dénomination 'SCP-Numéro', ce néologisme est largement employé. Les autres néologismes seraient utilisés de manière plus égale et ponctuelle à travers les différents formats. Pour tester cette hypothèse, j'ai décidé de calculer la fréquence des néologismes dans les contes et celle dans les rapports, puis de compiler ces informations dans un graphique avec l'application de la méthode de réduction directionnelle ACP.

Le schéma qui en a résulté est le suivant :

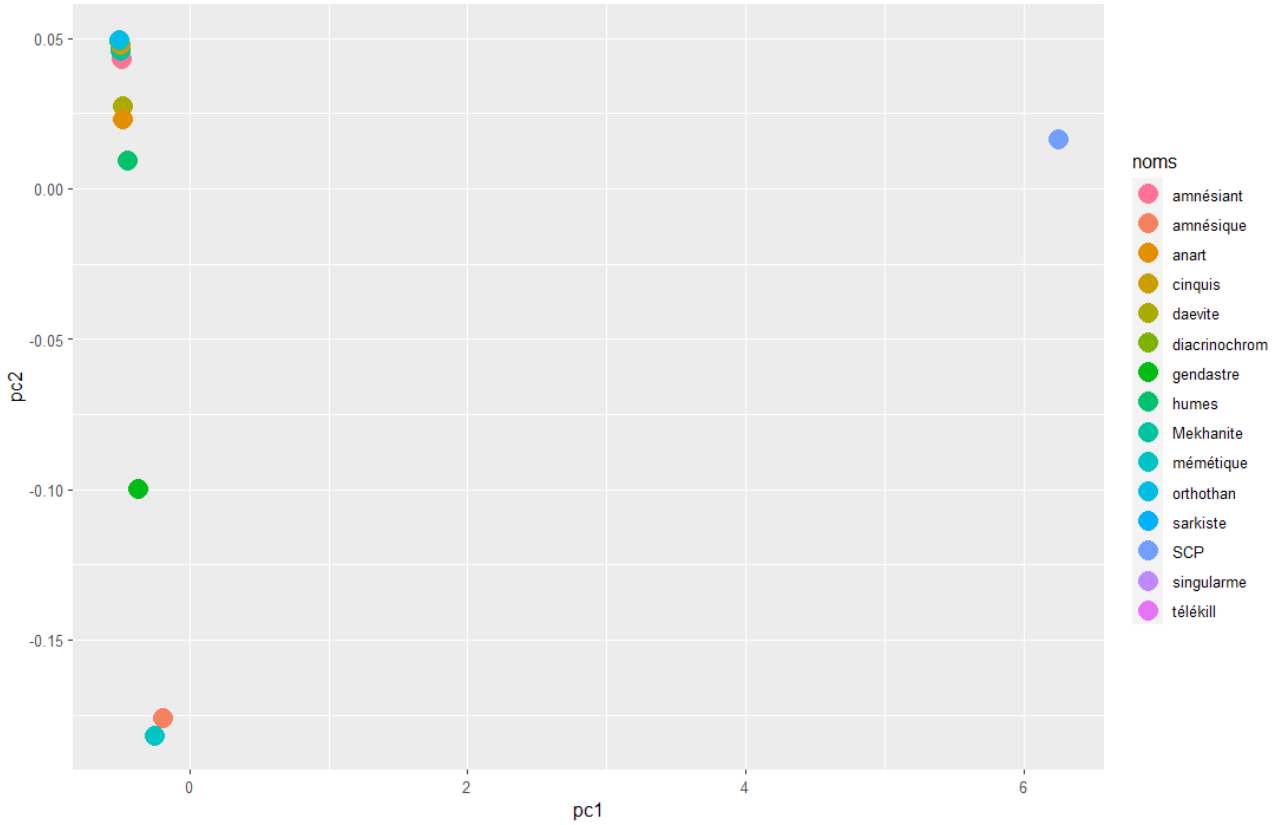

Figure 5: Graphe `df_stat`*, compare la fréquence des néologismes dans le corpus entier, dans les rapports (PC1) et dans les contes (PC2)*

Ce graphe confirme mon hypothèse de travail. Il serait dès lors intéressant d'exclure le néologisme 'SCP' de nos recherches afin d'obtenir une meilleure modélisation des autres néologismes.

2) Étude de cas du texte *Une Tragédie du Roi Pendu*

En utilisant des requêtes Xpath, j'ai pu produire plusieurs calculs comparant la présence des entités nommées francophones et anglophones – repérées grâce à l'attribut `@branch` qui leur est exclusif. Les entités nommées ne disposant pas d'un attribut `@branch` et ne venant donc pas du site de *La Fondation SCP* n'ont pas été comptées lors de ces calculs. On effectue les requêtes suivantes :

```
//*[@branch]
```

et

```
count(distinct-values(//*[@branch]))
```

Notamment, on dénombre en tout 1153 mentions d'entités nommées pour 321 entités distinctes.

Ainsi, les requêtes :

```
//*[@branch='FR']
```

et

```
count(distinct-values(//*[@branch='FR']))
```

révèlent que le texte comporte 625 mentions différentes d'une entité nommée imaginaire venant de la branche francophone, réparties entre 148 entités distinctes. Comparativement aux entités anglophones :

```
//*[@branch='EN']
```

et

```
count(distinct-values(//*[@branch='EN']))
```

qui sont mentionnées 520 fois pour 167 entités distinctes.

La requête :

```
//*[@branch and not(@branch='FR') and not(@branch='EN')]
```

révèle que 8 autres mentions ne sont pas affiliées à la branche anglophone ou francophone : on compte une entité nommée liée à la branche internationale, une autre à la branche italienne, une autre constitue un concept commun dont l'attribut `@branch` prend pour valeur `'all'`. Les 5 autres mentions sont d'origine inconnue et concernent 3 entités distinctes. Ces dernières n'ont pas été indiquées dans le Tableau 5 en raison de leur faible proportion.

En considérant maintenant que le texte-exemple fait un total de 205 195 caractères et 35 877 mots, on obtient donc les statistiques suivantes :

Tableau 5 : statistiques portant sur les entités nommées francophones et anglophones

Nature du calcul	***Entités de la Branche FR***	***Entités de la Branche EN***	***Taux de variation***[11]
Pourcentage des mentions représentées	54,20 %	45,09 %	- 20 %

11 En prenant la moyenne FR en valeur finale et la moyenne EN en valeur initiale. En arrondissant.

Pourcentage des entités nommées représentées	46,11 %	52,02 %	11 %
Fréquence des mentions	$1,7*10^{-2}$ mention par mot	$1,4*10^{-2}$ mention par mot	- 18 %
Fréquence des entités distinctes	$4,1*10^{-3}$ entité distincte par mot	$4,7*10^{-3}$ entité distincte par mot	13 %

J'ai choisi de modéliser la composition générale du texte en matière d'entités nommées grâce à deux diagrammes en secteur :

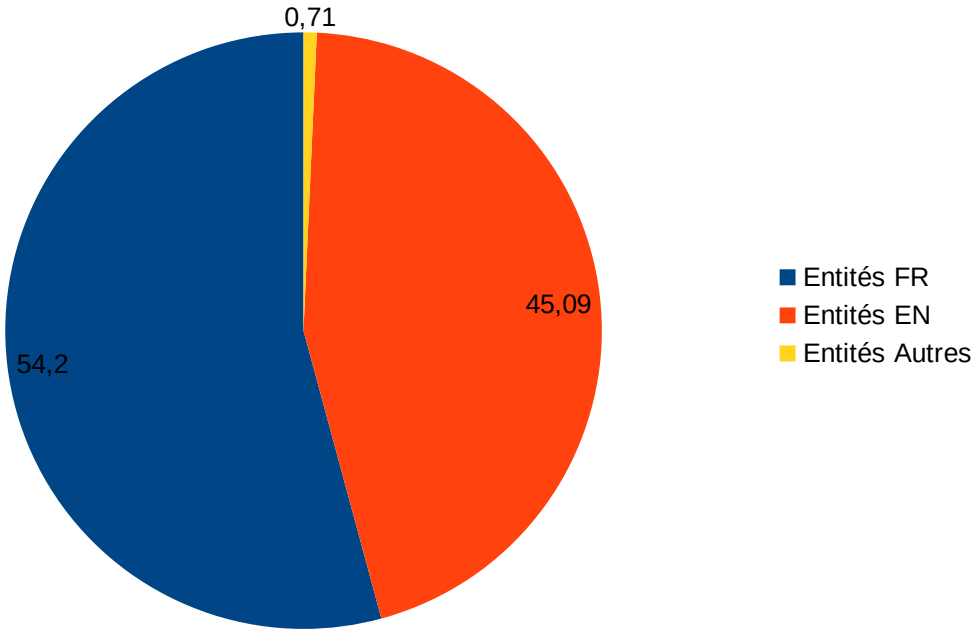

Figure 6 : Répartition en pourcentage des mentions d'entités nommées dans le texte

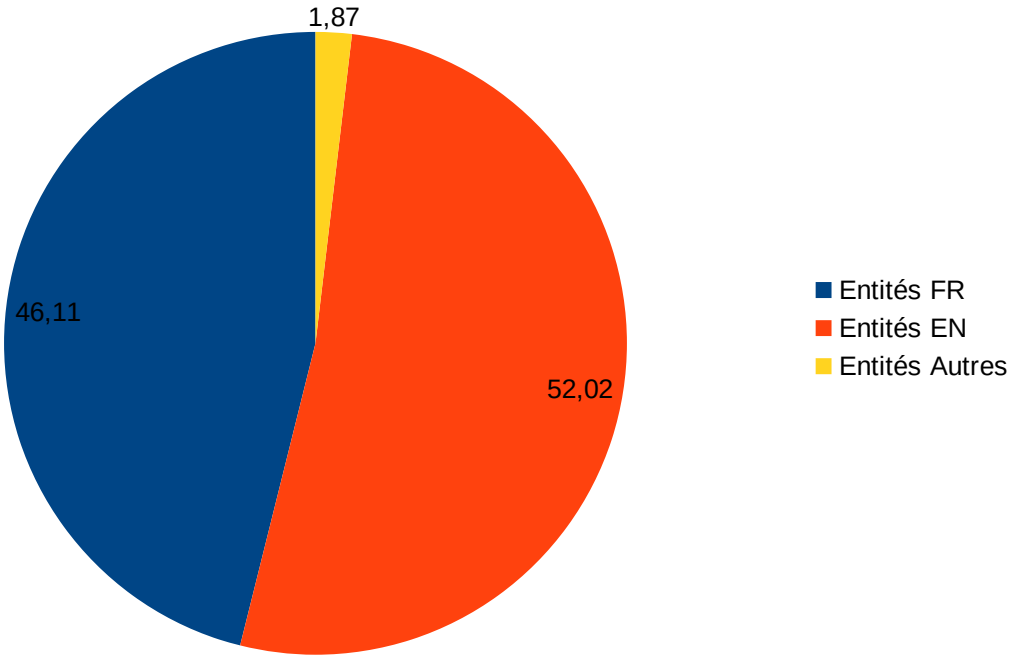

Figure 7 : Répartition en pourcentage des entités nommées distinctes dans le texte

On observe ainsi un équilibre statistique généralisé dans le texte entre les entités nommées anglophones et francophones : en effet, la moyenne des valeurs absolus des taux de variation pour chacune de mes statistiques est de 15,5 %, avec un écart-type de 4,2 %.

La différence notable entre le nombre d'entités distinctes, favorisant la branche anglophone, et le nombre de mentions de ces entités, favorisant la branche francophone, peut relever d'une opposition entre les principes de narration et de construction de la narration : en effet, si le texte-exemple reprend effectivement des concepts propres à la branche anglophone et plus précisément à la mythologie du Roi Pendu[38], il demeure que la narration proposée est une production originale de la branche francophone. Une partie des entités nommées, tout particulièrement les protagonistes, sont donc endémiques au texte-exemple, du moins jusqu'à ce qu'un autre auteur les reprenne dans son œuvre. Ainsi, les mentions d'entité nommées indiqueraient une place effective dans la narration et donc une part essentielle du texte ; là où le nombre d'entités nommées mentionnées et leur diversité permettrait de définir la complexité de la construction d'univers mise en place par l'auteur, du canon dans lequel l'œuvre s'inscrit, en quantifiant ce concept.

On veut alors définir un concept de 'localité' désignant les entités nommées natives ou même endémiques au texte étudié. Ce concept ne s'applique qu'au texte dont provient l'entité nommée et non à l'auteur l'ayant créée. Une entité nommée créée dans un autre texte du même auteur n'est pas considérée comme locale.

On examine par exemple les entités locales du texte-exemple afin d'étudier leur volume au sein des entités francophones qui sont si régulièrement mentionnées :

```
//*[@source='http://fondationscp.wikidot.com/une-tragedie-centre']
```

Ainsi, sur les 625 mentions d'entités francophones, 583 concernent des entités ayant pour source *Une Tragédie Du Roi Pendu*[32] : elles font donc leur première apparition dans le texte-exemple et constituent ainsi la partie purement créatrice du texte, l'ajout de nouveaux éléments innovants à l'univers. Les autres entités francophones non locales et les entités anglophones sont largement moins mentionnées. Ces premiers résultats semblent confirmer ma théorie initiale d'une différenciation statistique entre les entités nommées intégrées à la narration, et les entités nommées qui construisent la narration.

Tableau 6 : statistiques portant sur les entités locales du texte

Nature du calcul	***Entités FR locales***	***Entités FR non locales***	***Entités EN***
Pourcentage des mentions représentées	50,56 %	3,64 %	45,09 %

Une étude littéraire du texte-exemple permet de vérifier ces statistiques et de les contextualiser : en effet, *Une Tragédie du Roi Pendu*[32] étant à la fois une production francophone et une exploitation du mythe du Roi Pendu[38] anglophone, les entités nommées les plus actives au sein de la narration sont soit des entités locales, soit des entités anglophone issues du mythe que l'auteur a choisi d'exploiter.

En observant la liste des étiquettes affectées au texte-exemple, disponibles dans la colonne `tags` du dataframe `digital_born_data` [Tableau 2], je note que la liste contient les étiquettes : `alagadda`, `roi_pendu`, `fr`. Les deux premières étiquettes sont des étiquettes anglophones désignant deux éléments importants du mythe du Roi pendu[38]. Sur la base de cette observation, j'émets l'hypothèse que l'étude des étiquettes d'une page pourrait permettre de créer un modèle prédictif de la présence des entités nommées (tant au niveau des mentions que de la diversité et des rapports de force entre les entités venant de branches différentes) au sein d'un texte.

Ces considérations mettent finalement en valeur l'équilibre délicat que représente l'écriture communautaire en réseau, entre procédé d'invention classique et connaissance littéraire de l'univers déjà existant. Ces deux éléments ne sont pas mutuellement exclusifs toutefois : l'utilisation d'éléments provenant d'autres textes implique une part de réinterprétation et de réappropriation qui ne peut pas être constatée empiriquement par le biais des statistiques précédemment citées.

En appliquant le principe de localité et de proximité à un graphe de réseau, il pourrait peut-être possible de mettre en évidence la part de réinterprétation propre à chaque œuvre en retraçant le chemin chronologique et créatif des entités nommées étudiées jusque dans l'œuvre d'un auteur précis.

3) Autres pistes de recherche

Lors de mon étude préliminaire du sujet, j'ai décidé de m'intéresser à la figure de l'auteur au sein de la communauté, et à son assimilation à un rôle de lecteur et d'érudit en matière de connaissances portant sur l'univers commun. Je n'ai finalement pas eu le temps de m'y intéresser cette année et compte donc prolonger mon étude du sujet afin de pouvoir aborder cet aspect.

En effet, plusieurs outils auxiliaires ont été mis en place par la communauté afin de faciliter l'écriture de ses membres, tout particulièrement l'écriture à plusieurs. La mise à disposition d'un 'bac-à-sable' propre à chaque membre pour y faire ses tests, les forums de critique ou de correction des traductions et l'organisation d'équipes de critiqueurs, de vétérans, de correcteurs et d'experts font partie de ces mesures. Il serait intéressant d'étudier chez les auteurs la notion de 'vote' également mesurée par le site : un vote constitue une action indiquant avec certitude qu'un texte a été lu par tel ou tel auteur, et qu'il a été apprécié ou non. Il serait alors aisé d'étudier l'influence de ces lectures sur les textes des auteurs, en comparant la similarité ou la différence entre les éléments repris dans leurs textes et les éléments découverts dans d'autres. De même, le rôle de traducteur qui recoupe parfois celui d'auteur pourrait expliquer certaines influences, un utilisateur ayant tendance à traduire les textes qui lui plaisent ou qu'il juge essentiels à l'univers.

Un paradoxe demeure néanmoins au centre de cette double fonction de l'auteur-lecteur, ou 'lectauteur' comme le nomme Pauline Gosetti [2] en traduisant un concept anglophone de George P. Landow, « *wreader* ». Il s'agit d'une opposition ponctuelle entre la notion de productivité, chère à l'auteur, et la notion de plausibilité, chère au lecteur. En effet, les éléments esthétiques du site cherchant à établir la plausibilité de l'univers pour le lecteur se heurtent parfois à des éléments

fonctionnels brisant l'immersion amorcée : notamment les mentions légales et les modules de crédit. La liberté de format s'applique également à la forme du texte, justifiant l'apparition de nombreux thèmes CSS personnalisés, censés favoriser l'immersion et renforcer la thématique abordée dans un texte. Beaucoup de ces thèmes sont d'ailleurs associés à des entités nommées particulières, notamment celle des groupes d'intérêt, afin d'affirmer visuellement le point de vue adopté par un texte.

Ils font l'objet de fréquentes controverses au sein de la communauté, certains thèmes étant accusés de briser l'immersion et la plausibilité, d'autres de rendre la lecture trop complexe et l'étude des textes trop laborieuses en raison des éléments externes de décoration qui distraient l'œil, bien qu'ils constituent une partie intégrante de l'immersion. C'est ma conviction personnelle que ce sujet particulier de débat – observable dans les forums et commentaires des sites officiels – est symptomatique d'une communauté particulière de 'lectauteurs' tiraillée entre les différentes fonctions actives ou passives qui incombent à ses membres. Il pourrait donc nous apporter de nouveaux éléments statistiques sur la manière dont chaque utilisateur se positionne au sein du groupe social et créatif.

Un autre sujet me paraissant intéressant serait celui de la perméabilité de la communauté aux concepts extérieurs, ainsi qu'à l'étude du phénomène inverse. En effet, comme établi précédemment, la Fondation SCP est une entité existant au-delà des limites de la communauté d'écriture et de traduction, reprise sur plusieurs médias différents et particulièrement appréciée de la communauté horrifique du web. Sa grande souplesse lui permet assez aisément de se greffer à d'autres éléments culturels pré-existants : on peut noter l'existence du *Projet Crossover* [42] consistant à effectuer des croisements entre la Fondation SCP et d'autres œuvres variées et diverses ; ou l'existence du canon francophone *Sous le Monde* [43] qui reprend le principe du comics britanniques *La Ligue des gentlemen extraordinaires* [44] considérant un univers où toutes les fictions locales ont véritablement existé. Certaines entités nommées au sein du texte-exemple prenaient ainsi pour source des œuvres extérieures au site, classique ou obscures, ou au minimum s'en inspirant : je parle par exemple de *SCP-4028 – La Historia de Don Quixote de la Mancha*[45], inspirée de l'œuvre *L'Ingénieux Hidalgo Don Quichotte de la Manche* [46] de Miguel de Cervantes.

Plus largement, le caractère international de la communauté et les délimitations linguistiques qui séparent les branches les unes des autres les exposent également à subir les évènements du monde réel. La récente inactivité de la branche russe et ukrainienne [Voir Notes de bas de page 2 & 3] et le changement de nom de la branche de chinois traditionnel [Voir Note de bas de page 4] sont autant d'exemples des vulnérabilités d'un projet de cette nature aux changements mondiaux. Il

serait intéressant de les étudier dans une perspective sociologique ; mais ce n'est pas le sujet de mon mémoire ni ma spécialité et les données obtenues seraient trop récentes pour être utilisables.

J'ai parlé plus haut d'un indice de canonicité sur lequel je n'ai pu m'étendre jusqu'alors. C'est parce que j'ai pas encore défini précisément la formule qui pourrait permettre d'établir cet indice. Je pense qu'il me faudra mobiliser pour cela la répartition des entités nommées à travers les différents textes, travaux d'auteurs distincts et branches différentes.

En plus de ces pistes théoriques, la suite de mes recherches consistera à appliquer l'étiquetage à un plus grand nombre de textes afin de pouvoir créer un réseau d'entités nommées exploitable et d'en obtenir de plus amples statistiques.

Il faut noter également qu'avant de m'engager dans cette voie, il me faudra tout d'abord corriger mon code. En effet, l'hébergeur du site de *La Fondation SCP* ayant été sous maintenance d'urgence du 19 Mai 2022 au 24 Mai 2022 suite à une tentative de hack, plusieurs modifications de fond ont été apportées à la plateforme durant cette maintenance, rendant mon code actuel obsolète. Il me faudra donc le recalibrer afin de pouvoir récupérer les informations du corpus.

Bibliographie

- **1 –** Serge Bouchardon. *La valeur heuristique de la littérature numérique*. Sciences de l'information et de la communication. Université de Technologie de Compiègne, 2012. Français. Mémoire d'habilitation à diriger des recherches.
- **2 –** Pauline Gosetti. *L'écriture collective : un* jeu *d'auteur(s)*. Littérature européenne. Université Paris 3 Sorbonne Nouvelle, 2016. Français. Mémoire de Master en Lettres Modernes.
- **3 –** Eric Thomas Newsom. *Participatory Storytelling and the New Folklore of the Digital Age*. Philosophie de la Communication et de la Rhétorique. Graduate Faculty of Rensselaer Polytechnic Institute, 2013. Anglais.
- **4 –** Eric Thomas Newsom et Shira Chess. *Folklore, Horror Stories and the Slender Man: the Development of an Internet Mythology*. Littérature, Folklore, Psychologie. PALGRAVE MACMILLAN®, 2015. Anglais.
- **5 –** Rémi Mathis. *Wikipédia, au cœur de la plus grande encyclopédie du monde*. Communication, Littérature. First Éditions, 2021. Français.
- **6 –** Marion Lefebvre. *La juridictionnalisation de la modération des sites internet*. Droit. Université de Lilles, 2021. Français. Mémoire de Master en Études Judiciaires et Processuelles.
- **7 –** Maud Ehrmann. *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Informatique et langage [cs.CL]. Paris Diderot University, 2008. Français. tel-01639190
- **8 –** Jean Barré. *Entre Canon et Archive, étude des dynamiques textuelles : La valeur littéraire au révélateur des méthodes quantitatives*. Littérature, Humanités numériques. Université Paris, Sciences & Lettres, 2021. Français. Mémoire de Master 1 en Humanités Numériques, dir. [Thierry Poibeau, Jean-Baptiste Camps].

…

- **44 –** Alan Moore. *La Ligue des gentlemen extraordinaires*. WildStorm Production, DC Comic, 1999.
- **46 –** Miguel de Cervantes. *L'Ingénieux Hidalgo Don Quichotte de la Manche*. 1605.

Sitographie

Toutes les ressources en ligne ont vu leur accès vérifié pour la dernière fois le 29/05/2022.

- **9 –** Alexander ‘FiftyNine’ Krivopalov. *Scpper*, *https://scpper.com/*.
- **10 –** KatHansen. Sur *A Storied career*[11], « Q&A with a Story Guru: Cynthia Kurtz, Part 3 », http://astoriedcareer.com/qa_with_a_story_guru_cynthia_k_2. 17 Décembre 2008.
- **11 –** KatHansen. *A Storied career*, http://astoriedcareer.com/. Mai 2005.
- **12 –** Cynthia F. Kurtz. https://www.cfkurtz.com/.
- **13 –** Kirk. *Ravel Pathfinder 2*, https://www.ravel.pathfinder2.fr/.
- **14 –** France 4. *Anarchy*, https://www.france.tv/france-4/anarchy/. Automne 2014. Le lieu d’écriture a été depuis supprimée, mais une archive est accessible dans le mémoire de Pauline Gosetti [2].
- **15 –** *L’Allées des Conteurs*, https://www.alleedesconteurs.fr/.
- **16 –** Creepypasta From The Crypt[17]. *Necromorial*, https://necronomorial.blogspot.com/.
- **17 –** *Creepypasta From The Crypt*, https://creepypastafromthecrypt.blogspot.com/.
- **18 –** *Backrooms*, http://backrooms-wiki.wikidot.com/. Mars 2020.
- **19 –** *La Fondation SCP*, http://fondationscp.wikidot.com/. 2012.
- **20 –** RJB_BR. *Centre de l’Histoire de l’Univers*, http://fondationscp.wikidot.com/history-of-the-universe-hub. 18 Février 2013. Version originale ici : https://scp-wiki.wikidot.com/history-of-the-universe-hub. Traduction française actuelle de DrLekter, première version de la traduction le 23 Juillet 2018 par un compte d’utilisateur supprimé depuis.
- **21 –** Christopher Poole. *4Chan*, https://www.4chan.org/. 1er Octobre 2003.
- **22 –** Moto42, ou The U.S.S. Walrus. *SCP-173 – La Sculpture*, http://fondationscp.wikidot.com/scp-173. 25 Juillet 2008. Traduction française de DrMarcus, publiée le 12 Juin 2012. La version originelle du texte a été publiée le 22 Juin 2007 par l’auteur Moto42, sous le pseudo ‘The U.S.S. Walrus’ et sur le site *4chan*[21]. Elle a depuis été supprimée puis retrouvée et archivée sur le site *LostMediaWiki*[23] ici : https://lostmediawiki.com/SCP-173_(found_4chan_post;_2007).

- **23 –** *LostMediaWiki*, https://lostmediawiki.com/Home.
- **24 –** FritzWillie, ou The Administrator. *SCP Foundation*, https://scp-wiki.wikidot.com/. 25 Juillet 2008.
- **25 –** *SCP-INT*, http://scp-int.wikidot.com/. 30 Janvier 2017.
- **26 –** *Dos au Mur*, http://fondationscp.wikidot.com/dos-au-mur. 13 Octobre 2017.
- **27 –** Creative Commons. *Creative Commons Attribution-ShareAlike 3.0 License*, https://creativecommons.org/licenses/by-sa/3.0/.
- **28 –** Corentin Poupry. Dépôt : https://github.com/foundation-int-tech-team/sherlock/blob/e1d44d115cb6263c229b16ccfc6625ebe846563e51/sherlock/utils/wikidot.py#L24.
- **29 –** Leonard Richson. Documentation BeautifulSoup4 : https://beautiful-soup-4.readthedocs.io/en/latest/.
- **30 –** Stefan, ou scoder. Documentation LXML : https://lxml.de/. Dépôt : https://github.com/scoder.
- **31 –** Utilisateur inconnu. *SCP-74-KO – Tu n'es pas moi*, http://fondationscp.wikidot.com/scp-074-ko. 19 Février 2020. Version originale ici : http://scpko.wikidot.com/scp-074-ko. Traduction française de Vaalxeny, publiée le 25 Mai 2022 ; réalisée à partir de la traduction anglaise de DannyuNDos disponible depuis le 13 Octobre 2018 sur SCP-INT[25] : http://scp-int.wikidot.com/scp-074-ko.
- **32 –** Felter Finalis. *Une Tragédie du Roi Pendu*. 16 Juin 2021. Se compose des textes suivants :
  - *Une Tragédie du Roi Pendu – Centre*, http://fondationscp.wikidot.com/une-tragedie-centre.
  - *Acte 1 – Sur Les Planches*, http://fondationscp.wikidot.com/une-tragedie-acte-1.
  - *Acte 2 – Derrière Les Pendrillons*, http://fondationscp.wikidot.com/une-tragedie-acte-2.
  - *Acte 3 – Dans Les Coulisses*, http://fondationscp.wikidot.com/une-tragedie-acte-3.
  - *Acte 4 – Sous Les Projecteurs*, http://fondationscp.wikidot.com/une-tragedie-acte-4.
  - *Acte 5 – À la Régie*, http://fondationscp.wikidot.com/une-tragedie-acte-5.
  - *Acte 6 – Au Fond des Sièges*, http://fondationscp.wikidot.com/une-tragedie-acte-6.
- **33 –** tinwatchman. *SCP-701 – La Tragédie du Roi Pendu*, http://fondationscp.wikidot.com/scp-701. 27 Mars 2009. Version originale ici : https://scp-wiki.wikidot.com/scp-701. Traduction française de DrJohannes, publiée le 10 Septembre 2013.

- **34 –** Hadley Wickham. Documentation Ggplot2 : https://ggplot2.tidyverse.org/. Dépôt : https://github.com/tidyverse/ggplot2.
- **35 –** John D. Hunter, Michael Droettboom et al. Documentation Matplotlib : https://matplotlib.org/. Dépôt : https://github.com/matplotlib/matplotlib.
- **36 –** Agent Koop. *Rencontre au sommet*, http://fondationscp.wikidot.com/rencontre-au-sommet. 12 Novembre 2017. S'il s'agit bien là de la première apparition du personnage « Hématite » auquel la note fait référence, il a en réalité été créé par DrCendres et fait sa première apparition sous la plume de l'auteurice le 10 Septembre 2020, dans *Oculum pro oculo…*, http://fondationscp.wikidot.com/oculum-pro-oculo, co-écrit avec DrAttano.
- **37 –** Dr Goupil, DrGemini et al. *Glossaire Uniformisé Dédié à l'Univers et au Lore Étendu*, http://fondationscp.wikidot.com/gudule. 19 Mai 2019.
- **38 –** La mythologie du Roi Pendu se constitue de quatre œuvres majoritaires :
  - *SCP-701 – La Tragédie du Roi Pendu* [33]
  - *SCP-2264 - La cour d'Alagadda* [39]
  - *Un Arpenteur à la cour du Roi Pendu* [40]
  - *Et ainsi rirent les Corbeaux* [41]
- **39 –** Metaphysician. *SCP-2264 - La cour d'Alagadda*, http://fondationscp.wikidot.com/scp-2264. 14 Janvier 2015. Version originale ici : http://scp-wiki.wikidot.com/scp-2264. Traduction française de Vassago310, publiée le 1er Février 2016.
- **40 –** Metaphysician. *Un Arpenteur à la cour du Roi Pendu*, http://fondationscp.wikidot.com/a-wandsman-in-the-court-of-the-hanged-king. 26 Mai 2015. Version originale ici : https://scp-wiki.wikidot.com/a-wandsman-in-the-court-of-the-hanged-king. Traduction française de Vassago310, publiée le 8 Février 2016.
- **41 –** SunnyClockwork. *Et ainsi rirent les corbeaux*, http://fondationscp.wikidot.com/and-so-the-crows-laughed. 30 Mai 2015. Version originale ici : https://scp-wiki.wikidot.com/and-so-the-crows-laughed. Traduction française de Charles Magne, publiée le 17 Janvier 2020.
- **42 –** DrClef. *Projet Crossover*, http://fondationscp.wikidot.com/crossoverprojectindex. 12 Janvier 2012. Version originale ici : https://scp-wiki.wikidot.com/crossoverprojectindex. Traduction française de Dr Goupil, publiée le 24 Avril 2016.
- **43 –** DrGemini. *Sous le monde*, http://fondationscp.wikidot.com/sous-le-monde-centre. 9 Juin 2021.
- **45 –** The Great Hippo. *SCP-4028 – La Historia de Don Quixote de la Mancha*, http://fondationscp.wikidot.com/scp-4028. 21 Août 2018. Version originale ici : https://scp-

wiki.wikidot.com/scp-4028. Traduction française de Felter Finalis, publiée le 23 Novembre 2019.

Figures et tableaux

Index des figures

Index des tableaux