Évaluer une interface documentaire augmentée : étude de cas sur le projet ANR Archival

Éléonore Besnehard

Eléonore BESNEHARD

Master Information-Documentation / Première année

Mémoire de stage

Évaluer une interface documentaire augmentée : étude de cas sur le projet ANR Archival

Stage effectué du 17 avril au 20 juillet 2023 à la BnF (Site François-Mitterrand) et au laboratoire GERiiCO (Université de Lille, Campus Pont-de-Bois)

Sous la direction de :

M. Stéphane CHAUDIRON (tuteur pédagogique, laboratoire GERiiCO, Université de Lille)

M. Eric KERGOSIEN (tuteur professionnel, laboratoire GERiiCO, laboratoire Geriico)

M. Arnaud LABORDERIE (tuteur professionnel, service de la Coopération numérique et de Gallica, BnF)

Soutenu le 3 juillet 2023

Université de Lille (campus Pont-de-Bois)

BP 60 149, 59 653 Villeneuve d'Ascq Cedex

Année universitaire 2022/2023

Remerciements

Je tiens avant toute chose à exprimer ma profonde gratitude à mes trois tuteurs de stage : Stéphane CHAUDIRON, Éric KERGOSIEN et Arnaud LABORDERIE pour leur accompagnement quotidien, les bons conseils prodigués et la confiance qu'ils m'ont témoignée.

J'adresse également mes remerciements à Monique PUJOL, directrice du Département de la coopération, et toute l'équipe de la Coopération numérique et de Gallica à la BnF, notamment Sophie BERTRAND, cheffe de service, Mathieu GIOUX, chef de service adjoint, mais également Isabelle MANGOU et Fanny VERDIER pour le temps qu'elles m'ont accordé.

Je tiens également à témoigner ma gratitude à l'équipe du BnF DataLab, Marie CARLIN et Louise-Anne CHARLES, pour leur aide dans la préparation et leur soutien le jour de l'expérimentation du 15 mai 2023.

J'aimerais aussi témoigner ma profonde reconnaissance envers Irène BASTARD pour ses multiples conseils et son aide précieuse dans la réalisation de l'enquête ainsi qu'envers Lucie TERMIGNON pour l'organisation de la journée d'étude du 20 juin à la BnF.

Je remercie chaleureusement l'équipe Archival pour la collaboration fructueuse que nous avons fait naître ainsi que pour le très bon accueil qu'ils m'ont réservé. Je remercie en particulier Ghislaine AZEMARD, titulaire de la chaire UNESCO-ITEN, Michel AGNOLA, coordinateur du projet Archival et Samuel DA SILVA, développeur principal du démonstrateur.

Je remercie également les partenaires du projet Archival, notamment Frédéric BECHET[1], directeur du Lis-Lab, Géraldine DAMNATI[2], chercheuse chez Orange Labs, Guillaume GRAVIER, chercheur au CNRS et directeur de l'IRISA, et Pascale SEBILLOT[3], directrice adjointe de l'IRISA.

[1] Equipe TALEP (Traitement Automatique du Langage Ecrit et Parlé), Lis-Lab, Université Aix-Marseille (AMU)
[2] Orange Labs
[3] Équipe LINKMEDIA, Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), Université Rennes 1.

Je souhaite exprimer ma gratitude aux vingt testeurs interrogés dans le cadre de l'enquête menée au cours du stage pour leur disponibilité, leur bienveillance et leurs retours précieux : Alexandra BALLY, Laure BOLKA-TABARY, Jean-Michel BORDE, Nasreddine BOUHAI, Stéphane CHAUDIRON, Marie CROS, Jean-Pierre DALBERA, Axel ETO, Alexandre FAYE, Henri HUDRISIER, Olivier JACQUOT, Nadia KHELIFERI, Tanguy LAURENT, Laure LEROY, Xavier LEVOIN, Jean-Philippe MOREUX, John MOTTA, Mathieu QUINIOU, Lucie TERMIGNON, Luiz TORRES-YEPEZ.

J'adresse également mes sincères remerciements au laboratoire GERiiCO de l'Université de Lille pour avoir permis la réalisation de ce stage dans d'excellentes conditions, notamment Patrice DE LA BROISE, directeur du laboratoire.

Pour finir, j'aimerais remercier Joachim SCHÖPFEL, responsable du master Information-Documentation de l'Université de Lille, d'avoir accepté de faire partie du jury de la soutenance.

Résumé

La recherche d'information revêt une importance capitale dans nos sociétés actuelles et ses enjeux sont nombreux et divers. Dans ce cadre, les interfaces documentaires tentent de s'adapter sans cesse aux besoins renouvelés des usagers et aux technologies en constante évolution, en particulier dans le domaine de l'intelligence artificielle (IA). Ainsi, naissent depuis plusieurs années des interfaces documentaires dites « *intelligentes* » car augmentées par IA ; notre objectif étant d'en comprendre les possibilités, les apports actuels ainsi que les perspectives envisagées pour proposer une méthode d'évaluation de ce type de plateforme. Pour ce faire, nous avons travaillé sur le projet ANR Archival comme cas pratique afin d'éprouver notre protocole d'expérimentation du dispositif et d'initier notre évaluation de la plateforme.

Mots-clés

Interface documentaire, Lecture augmentée, Intelligence artificielle, Recherche d'information, Découvrabilité, Expérimentation et évaluation.

Abstract

Information retrieval is of paramount importance in today's society, and the issues involved are many and varied. In this context, document interfaces are constantly trying to adapt to the renewed needs of users and to constantly evolving technologies, particularly in the field of artificial intelligence (AI). For several years now, so-called "*intelligent*" documentary interfaces have been emerging, augmented by AI. Our aim is to understand the possibilities, current contributions and future prospects, and to propose a method for evaluating this type of platform. To this end, we used the ANR Archival project as a case study to test our experimental protocol and initiate our evaluation of the platform.

Key words

Documentary Interface, Augmented Reading, Artificial Intelligence, Information Retrieval, Discoverability, Experimentation and Evaluation.

Liste des acronymes utilisés

ANR : Agence nationale de la recherche

AUF : Agence Universitaire de la Francophonie

BnF : Bibliothèque nationale de France

CCFr : Catalogue collectif de France

CLEF : Cross Language Evaluation Forum

CollEx : Collection d'Excellence

DAI : Dispositif d'accès à l'information

DARPA : Defense Advanced Research Program Agency

DOD : Department of Defense

EAGLES : Evaluation of Natural Language Processing Systems

GERiiCO : Groupe d'Études et de Recherche Interdisciplinaire en Information et Communication, Université de Lille

IA : Intelligence artificielle

IHM : Interface Homme Machine

INEX : Initiative for the Evaluation of XML Retrieval

IREX : Information Retrieval and Extraction Exercise

IRISA : Institut de Recherche en Informatique et Systèmes Aléatoires, Université Rennes 1

ITEN : Innovation, Transmission, Edition Numérique, chaire de l'UNESCO

Lis-Lab : Laboratoire d'Informatique et Système, Université Aix-Marseille

MUC : Message Understanding Conference

MRIM : Modélisation et Recherche d'Information Multimédia

NIST : National Institute of Standards and Technology

RI : Recherche d'information

SCD : Service commun de la documentation

SHS : Sciences humaines et sociales

SIC : Sciences de l'information et de la communication

SID : Sciences de l'information et de la documentation

SRI : Système de recherche d'information

STIC : Sciences et Technologies de l'Information et de la Communication.

TALEP : Traitement Automatique du Langage Ecrit et Parlé, Université Aix-Marseille

TALN : Traitement automatique du langage naturel

TIDES : Translingual Information Detection, Extraction and Summarization

TREC : Text REtrieval Conference

WP : WorkPackage

Sommaire

Remerciements 1

Résumé 3

Mots-clés 3

Abstract 3

Key words 3

Liste des acronymes utilisés 4

Introduction 9

Chapitre 1 : Cadre théorique et méthodologique 15

I/ Genèse et évolution des systèmes de recherche d'information 15

1) Les fondements de la RI 15

2) La recherche d'information en contexte 20

3) Les interfaces de recherche documentaire 22

II/ Les pratiques évaluatives 24

1) L'approche techno-centrée 24

2) De l'approche orientée usager à l'approche croisée 27

3) Les protocoles d'évaluation 29

III/ Les outils et concepts pour l'évaluation 31

1) Mesures et métriques 31

2) Concepts mobilisés 35

3) L'évaluation : une approche pluridisciplinaire 39

Conclusion 41

Chapitre 2 : Étude de cas Archival 42

I/ Présentation du projet 42

1) Création du dispositif Archival 42

2) Les enjeux de l'explicabilité des algorithmes 46

3) Le dialogue entre informatique, SIC et SHS 48

II/ Mise en place de l'expérimentation 50
1) Réalisation de journées d'expérimentation 50
2) Définition des panels de testeurs 52
3) Les fonctionnalités du dispositif 55
III/ Élaboration d'une méthode d'évaluation 60
1) Usage de la littérature 60
2) Mise en place d'une méthode composite 62
3) Une démarche pluridisciplinaire et multi acteurs 64
Conclusion 66
Chapitre 3 : Résultats et recommandations 67
I/ Résultats obtenus sur l'interface 67
1) Sur l'interface graphique 67
2) Sur l'ergonomie et la navigation 71
3) Sur l'appropriation du dispositif 73
II/ Résultats obtenus sur le contenu 76
1) Sur le corpus 77
2) Sur les facettes de recherche classique 82
3) Sur les apports de l'intelligence artificielle 86
III/ Discussion et recommandations 91
1) Discussion des résultats 91
2) Recommandations pour les expérimentations futures 95
3) Portabilité de la méthode d'évaluation 99
Conclusion 101
Conclusion générale 102
Bibliographie 104
Sitographie 112
Résumé 114

Mots-clés 114

Introduction

Le 20 juin dernier, une journée d'études intitulée « *Penser la découvrabilité des contenus culturels* » a été organisée au petit auditorium de la Bibliothèque nationale de France (BnF), par l'institution en partenariat avec la chaire UNESCO-Innovation, Transmission, Edition Numérique (ITEN). Les sujets abordés étaient variés et les invités nombreux : deux projets ANR, le service numérique du ministère de la Culture, Radio France, le Pass Culture[4] ou encore l'entreprise Spidéo[5]. Il y a donc une réelle actualité au sein des institutions culturelles et patrimoniales françaises autour de la découvrabilité et de ses enjeux. Cela s'inscrit dans un contexte plus large où le ministère de la Culture français s'est associé depuis 2019 au ministère de la Culture et des communications du Québec pour promouvoir la découvrabilité en ligne des contenus culturels francophones au travers d'une mission et une stratégie communes[6]. La découvrabilité est définie par les ministères de la façon suivante :

> « *La découvrabilité d'un contenu dans l'environnement numérique se réfère à sa disponibilité en ligne et à sa capacité à être repéré parmi un vaste ensemble d'autres contenus, en particulier par une personne qui n'en faisait pas précisément la recherche.* »[7]

Ce sujet revêt une importance particulièrement actuelle car la masse documentaire mise à disposition du public en ligne est aujourd'hui constituée de plusieurs milliards de documents accessibles à tous gratuitement en quelques clics. Par ailleurs, l'offre culturelle en ligne ne fait que s'étoffer au travers d'acteurs, de produits et de collections qui se diversifient et proposent toujours plus de contenus. L'exemple de Gallica est révélateur de cette massification des contenus accessibles en ligne car la bibliothèque numérique a dépassé les dix millions de documents numérisés ces derniers mois, un phénomène qui ne va pas s'estomper -bien au contraire- puisque les chantiers de numérisation des collections de la BnF se poursuivent et que les documents accessibles en ligne continuent d'augmenter de jour en jour.

Penser la découvrabilité des contenus culturels et documentaires est donc une nécessité actuelle, créer des outils appropriés permettant aux usagers de découvrir les collections et de mettre en valeur les documents peu ou pas consultés est dès lors un défi de taille pour le présent

[4] Voir https://pass.culture.fr/ [en ligne] consulté le 22/06/2023.
[5] Voir https://spideo.com/ [en ligne] consulté le 22/06/2023.
[6] Voir https://www.culture.gouv.fr/Presse/Communiques-de-presse/Lancement-de-la-mission-franco-quebecoise-sur-la-decouvrabilite-des-contenus-culturels-francophones-en-ligne [en ligne] consulté le 22/06/23.
[7] Voir https://www.culture.gouv.fr/Thematiques/Europe-et-international/Publications/Decouvrabilite-en-ligne-des-contenus-culturels-francophones [en ligne] consulté le 22/06/23.

et le futur de la recherche documentaire, notamment pour les institutions culturelles. La problématique en jeu est alors non pas de savoir si ces outils seront utiles mais plutôt d'identifier comment les utiliser, qui peut les utiliser, quels sont les prérequis, quels sont les biais et les limites ? L'accroissement exponentiel des collections documentaires pose également la question de la trouvabilité d'un document ou d'une information parmi les milliards de ressources numérisées ou nativement numériques car s'il est nécessaire d'augmenter les collections en ligne afin de permettre à tous un accès facile et rapide, il est également primordial que chacun puisse trouver ce qu'il recherche.

La problématique de la découvrabilité et de la trouvabilité des documents et contenus culturels fait donc partie intégrante des sciences de l'information et de la communication (SIC) et plus précisément du domaine de la recherche d'information (RI). Aucune définition de la recherche d'information n'est établie de façon unanime cependant, dans son acception la plus large, elle peut être expliquée ainsi : « *la RI a pour thème central l'étude de modèles et systèmes d'interaction entre des utilisateurs humains et des corpus de documents numériques, en vue de la satisfaction de leurs besoins d'information* »[8]. Un domaine qui s'est constamment développé selon trois axes majeurs : la théorie et les modèles conceptuels, la réalisation et l'expérimentation de systèmes de recherche d'information (SRI), l'étude des comportements usagers au cours de ces expérimentations.

Définition des termes du sujet

Ces trois axes ont alors amené à questionner l'efficacité des SRI et à les évaluer, amenant à la création du domaine de l'évaluation :

> « *On peut définir l'évaluation comme une relation qui vise à déterminer un indice de satisfaction, c'est-à-dire à déterminer à quoi quelque chose est bon pour quelqu'un ; c'est une fonction qui lie le système à évaluer et ses utilisateurs à des finalités.* »[9]

L'évaluation est alors un domaine dont « *l'importance n'est plus à démontrer* »[10] mais qui est aussi et surtout sans cesse renouvelé par la création de nouveaux protocoles d'évaluation et de nouveaux dispositifs d'accès à l'information (DAI). En parallèle, les progrès récents de l'intelligence artificielle en font un objet d'étude et de recherche particulièrement foisonnant

[8] Chiaramella, Y. & Mulhem, P. (2007). La recherche d'information : De la documentation automatique à la recherche d'information en contexte. Document numérique, 10, p. 11-38.
[9] Chaudiron, S. (dir.) (2004). Évaluation des systèmes de traitement de l'information, Hermès.
[10] *Ibid.*

notamment en ce qui concerne la génération automatique de liens. Cependant, les chercheurs s'accordent pour rappeler que ces modèles de mise en relation des connaissances sont encore très largement imparfaits et nécessitent une attention toute particulière s'agissant de la pertinence des liens générés qu'il convient d'évaluer. Ce qui est d'autant plus important que l'intelligence artificielle intéresse de plus en plus les institutions (bibliothèques, musées, etc) principalement pour faciliter la recherche documentaire et la découvrabilité des contenus culturels. Renouvelant ainsi des questionnements de longue haleine sur la question de l'apport de l'intelligence artificielle et des modèles de langue pour les sciences humaines et sociales (SHS). Parmi ces nouveaux dispositifs d'accès à l'information, de récentes innovations méritent toute notre attention.

Objet de recherche : le projet Archival

Le projet ANR Archival (ANR-19-CE38-0011) est un projet de recherche interdisciplinaire dédié à la valorisation d'archives multimédia assistée par intelligence artificielle. Le postulat initial du projet est le suivant : bien que des travaux aient déjà été effectués sur les archives multimédia (textes, images, sons, vidéos), les interfaces de recherche actuelles ne permettent ni la navigation et l'exploration approfondies des contenus, ni une mise en relation efficace entre les archives et d'autres sources externes. Dans ce contexte, Archival vise à développer de nouvelles interfaces de lecture, de consultation des documents, de médiation et de transmission des savoirs grâce à l'intelligence artificielle. Les questions de départ de l'équipe projet sont les suivantes :

> « *quel rôle peuvent jouer les méthodes de compréhension par les machines dans la réinterprétation de fonds d'archives thématiques ? Selon quelles modalités des interfaces de médiation des contenus peuvent-elles exploiter des résultats générés par les méthodes actuelles d'Intelligence Artificielle ?* »[11]

La BnF et le laboratoire GERiiCO (Groupe d'Etudes et de recherche Interdisciplinaire en Information et COmmunication) ont été associés au projet en 2023 pour apporter un soutien méthodologique et pratique à l'expérimentation auprès des usagers et à l'évaluation finale du dispositif. Les apports de la BnF et du laboratoire GERiiCO fera l'objet de ce présent mémoire.

[11] Voir https://www.fmsh.fr/actualites/archival [en ligne] consulté le 31/05/2023.

Les missions de la BnF et de Gallica

La Bibliothèque nationale de France est associée au projet car en tant qu'établissement public sous tutelle du ministère de la Culture, ses missions sont multiples : « *collecter, cataloguer, conserver, enrichir et communiquer le patrimoine documentaire national. La BnF assure l'accès du plus grand nombre aux collections sur place, à distance et développe la coopération nationale et internationale* »[12]. Les missions de la Bibliothèque sont alors parfaitement en accord avec les problématiques énoncées précédemment à savoir la découvrabilité et la trouvabilité des documents et ce notamment en ligne, au travers de sa bibliothèque numérique : Gallica[13].

La BnF est régie par une organisation complexe constituée de plusieurs niveaux entre les délégations, les directions, les départements et les services. Parmi les cinq directions[14] se trouve celle des Services et des réseaux au sein de laquelle on compte six départements distincts qui travaillent en synergie[15] dont le Département de la Coopération. Ce dernier regroupe trois services pour une trentaine d'agents, le service du Catalogue collectif de France (CCFr), la mission Coopération régionale, communication, formation et enfin le service de la Coopération numérique et de Gallica, au sein duquel j'ai effectué mon stage. L'organigramme de la BnF est présenté en annexes de ce mémoire [voir annexe 1].

Le service de la Coopération numérique et de Gallica, situé sur le site François-Mitterrand à Paris, compte douze agents répartis sur trois pôles [voir annexe 2] : le pôle Services numériques aux partenaires, le pôle Usagers et médiation, le pôle Coordination scientifique et documentaire et le chef de produit Gallica.

Depuis 1997 jusqu'à aujourd'hui, les missions de la bibliothèque numérique sont de donner un accès libre et gratuit à plus de 10 millions de documents patrimoniaux à plusieurs dizaines de milliers de visiteurs par jour (tous types de publics) en leur permettant d'explorer les collections et de travailler et d'exporter les documents qui les intéressent par intérêt personnel ou pour leurs travaux[16]. Les collections de la bibliothèque ne cessent de s'étoffer

[12] Voir https://www.bnf.fr/fr/missions-et-organisation-de-la-bnf [en ligne] consulté le 23/06/2023.
[13] Voir https://gallica.bnf.fr/accueil/fr/ [en ligne] consulté le 22/06/2023.
[14] La Direction des Publics, la Direction des Collections, la Direction de l'Administration et du personnel, la Direction déléguée aux Ressources humaines ainsi que la Direction des Services et des réseaux.
[15] Le Département des Métadonnées, le Département de la Conservation, le Département du Dépôt légal, le Département des Images et prestations numériques, le Département des Systèmes d'information et le Département de la Coopération.
[16] Tous les chiffres mentionnés ici sont issus d'une présentation du service faite par Arnaud LABORDERIE le 18/04/2023 à la BnF.

d'année en année, ce qui démultiplie la masse documentaire à disposition des lecteurs et pose toujours plus vivement la question de la recherche d'information au sein de la plateforme. Il faut donc aujourd'hui continuer de compléter les collections tout en rendant les documents plus facilement trouvables et découvrables, il s'agit d'une des problématiques de recherche portée par Arnaud LABORDERIE, chef de projet Gallica et chargé de l'exploitation des données pour la recherche, et de ses collègues.

Un travail qui se fait en lien avec le BnF DataLab, le service aux chercheurs de la BnF créé en 2019 au sein de la Direction des Collections dans le Département de la Découverte des collections et de l'accompagnement à la recherche, dont nous aurons l'occasion de reparler.

Les travaux du laboratoire GERiiCO

Pour sa part, le laboratoire GERiiCO (laboratoire de SIC de l'Université de Lille) est « *centré sur la question des médiations des connaissances, des savoirs et des cultures dans la société contemporaine* »[17]. Structuré en quatre axes thématiques[18], le laboratoire explore :

> « *Comment informations, connaissances, savoirs et cultures circulent dans les groupes sociaux, sont organisés, capitalisés ? À travers quels dispositifs - matériels et humains - ils prennent forme et statut ? Quels sont les processus de médiation et de médiatisation qui les affectent à l'heure du numérique ? Quels sont les conditions, les contextes et les enjeux du développement et de l'insertion sociale des dispositifs technologiques d'information-communication ?* »

Dans ce cadre, le présent stage s'inscrit parfaitement dans les considérations de l'axe 3 (Innovation par l'usage et dispositifs numériques) qui mène « *des recherches pluridisciplinaires sur la conception, la mise en œuvre, l'appropriation, l'évaluation et l'analyse des usages des dispositifs numériques* » et dans les travaux sur l'évaluation des SRI réalisés par Stéphane CHAUDIRON et ceux d'Eric KERGOSIEN sur l'appropriation des dispositifs informationnels - notamment des systèmes d'information.

[17] Voir https://geriico.univ-lille.fr/lunite/presentation-de-lunite [en ligne] consulté le 22/06/2023.
[18] Axe 1 : Information et communication dans les organisations,
Axe 2 : Culture et médias dans l'espace public,
Axe 3 : Innovation par l'usage et dispositifs numériques,
Axe 4 : Circulation de l'information et organisation des connaissances
Voir https://geriico.univ-lille.fr/lunite/axes-thematiques [en ligne] consulté le 22/06/2023.

Enjeux et problématique

L'évaluation est de fait complexe et plurielle : que peut-on évaluer ? Sur quels critères baser l'évaluation ? Comment mettre en place un protocole scientifique d'évaluation ? Comment faire dialoguer les évaluations quantitatives et qualitatives ? L'évaluation d'un système de recherche d'information (SRI) s'appuie alors sur une méthodologie scientifique et peut prendre en compte une multitude de critères tels que la performance du système, la pertinence des résultats, l'adéquation avec les besoins informationnels et la satisfaction des usagers ou encore l'explicabilité du système. Notre travail est donc guidé par l'objectif suivant : élaborer, à partir des connaissances et de la littérature scientifique qui existent sur l'évaluation des SRI, une méthode permettant d'évaluer le dispositif Archival et voir quelles en sont les possibilités de réemplois pour d'autres interfaces documentaires, notamment Gallica.

Notre évaluation se base alors sur deux niveaux de résultats : d'abord, le prototype Archival en lui-même en tant que dispositif de lecture et de recherche d'information (interface, navigation, facettes) mais également le protocole mis en place pour évaluer les résultats collectés lors des phases d'expérimentation auprès des testeurs (observations, questionnaires et entretiens). La collecte et la mise en relation de ces résultats permettra dès lors d'étayer les analyses et réflexions issues de la littérature scientifique sur le sujet.

Ainsi, quelles recommandations formuler au regard de ces deux niveaux de résultats ? Les recommandations faites peuvent-elles constituer les fondations d'une boîte à outils, à la fois théorique et pratique, permettant d'étendre ce protocole à d'autres dispositifs et en particulier à la bibliothèque numérique Gallica ?

Plan

Le présent mémoire est structuré en trois chapitres, le premier est consacré à l'ancrage du sujet au sein du cadre théorique et méthodologique établi dans la littérature scientifique, le second s'attache à présenter le dispositif Archival ainsi que la méthode d'évaluation mise en place, le dernier explique les résultats obtenus par l'évaluation mais aussi les perspectives d'amélioration et de transposition de cette méthode à d'autres dispositifs.

Chapitre 1 : Cadre théorique et méthodologique

Le présent chapitre vise à établir le cadre théorique et méthodologique de notre réflexion sur l'évaluation des dispositifs d'accès à l'information. La littérature sur le sujet des DAI est très abondante, il convient donc d'en faire un rapide état des lieux pour en comprendre les enjeux, intérêts et évolutions dans une perspective diachronique et conceptuelle. Cela nous permettra de poser les premiers jalons de l'étude des DAI afin de nous intéresser à leur évaluation, domaine qui possède également une littérature foisonnante. Nous en dresserons un rapide aperçu en évoquant l'évolution des méthodes d'évaluation avec une attention portée aux protocoles existants. Dans la dernière partie, nous élaborerons une boîte à outils conceptuelle et pratique que nous pourrions employer pour mettre au point notre méthode d'évaluation.

I/ Genèse et évolution des systèmes de recherche d'information

La recherche d'information s'est structurée au fil des décennies, depuis les années 1940 à nos jours, nous évoquerons l'émergence du domaine avant d'en présenter les diverses évolutions pour nous pencher sur les DAI utilisés en RI par les chercheurs aujourd'hui.

1) Les fondements de la RI

La recherche d'information émerge sous sa première forme dans les années 1940 aux Etats-Unis grâce à l'avènement des ordinateurs, et aux travaux pionniers de Michael LESK[19] et Vannevar BUSH[20], ce dernier soutient l'idée de créer des grands dépôts de documents et les outils de recherche automatisés associés. Il a également développé le projet Memex soit un des premiers systèmes de gestion de base de données. Calvin MOOERS invente l'appellation *Information Retrieval* (IR) en 1948 dans son mémoire au MIT[21]. Ce terme s'impose alors et donne son nom à la discipline qui connaît en engouement fort pendant la Guerre froide du fait du caractère stratégique de l'information[22].

[19] Lesk, M. (1995). The Seven Ages of Information Retrieval. Conference for the 50th anniversary of « As We May Think ».
[20] Bush, V. (1945) « As We May Think », Atlantic Monthly, 176, 1, p. 101-108.
[21] Garfield, E. (1996). A tribute to Calvin N. Mooers, A pioneer of Information Retrieval. The Scientist, 11.
[22] Chiaramella, Y. & Mulhem, P. (2007). *Op. cit.*

Les premiers travaux sur la recherche bibliographique et les données booléennes voient le jour dans les années 1950 notamment avec le WRU (*Western Reserve University*) *Searching Selector* de James W. PERRY, mais la recherche d'information se développe et se structure surtout dans les décennies suivantes notamment avec des travaux portés sur l'indexation des données. Le domaine prend alors de l'ampleur grâce aux évolutions des ordinateurs plus puissants qui décuplent les performances et possibilités permettant d'indexer automatiquement des dizaines de milliers de documents. La RI a ensuite connu, dans les décennies suivantes, un développement constant selon les trois axes suivants[23] : « *la théorie et les modèles sous-jacents, la réalisation et l'expérimentation de systèmes de recherche d'information (SRI), l'étude des comportements usagers au travers de différentes applications et expérimentations.* » Les trois axes se sont développés conjointement et n'ont fait qu'accroitre les diverses avancées du domaine.

Les années 1960 sont foisonnantes et décisives pour la RI, notamment avec la création de l' « *approche probabiliste* » (*probabilistic retrieval*) et l'utilisation des mathématiques pour évaluer la pertinence des documents en s'appuyant sur un calcul de probabilité qu'un document retrouvé soit pertinent[24], idée reprise et développée par la suite par SPARCK, JONES[25] et ROBERTSON[26]. On peut également relever la création des premières notions de pondération (*term weighting*) et classement (*ranking*). On voit ainsi se développer une prise de conscience de l'importance de l'ordre des documents dans la visualisation et la consultation des résultats par les usagers. L'autre notion qui se développe à l'époque est celle de la classification des documents (*document clustering*) qui postule le fait que la similarité entre documents permet d'optimiser le processus de résolution des requêtes car « *si un document est très pertinent en réponse à une requête donnée, alors tout document fortement similaire à ce document est susceptible d'être également une bonne réponse à cette requête* »[27] ce qui est l'objet des travaux de W. B. CROFT en 1977[28].

C'est également l'époque où la recherche à partir de thésaurus documentaires est associée aux premiers moteurs de recherche spécialisés. On le voit aux Etats-Unis à partir de

[23] *Ibid.*
[24] Maron, M. E. (1960). On relevance, probabilistic indexing and information retrieval. Journal of the ACM, 66.
[25] Spärck Jones, K., & Willett, P. (1997). Readings. Dans Information Retrieval. Morgan and Kaufmann publishers.
[26] Robertson, S., & Spärck Jones, K. (1976). Relevance weighing of search terms. Journal of the American Society for Information Science, 27.
[27] Chiaramella, Y. & Mulhem, P. (2007). *Op. cit.*
[28] Croft W. B. (1977). Clustering large files of documents using a single link. Journal of the American Society for Information Science, 28.

1964 par l'association du thésaurus MeSH au système MEDLARS (*Medical Literature Analysis and Retrieval System*) créé par la *National Library of Medicine*, ce dernier sert à interroger une base de données bibliographiques médicale à l'aide de mots-clés[29]. Pour la France, on peut relever l'exemple du système DARC (Description, Acquisition, Recherche et Corrélation) dédié à la chimie et élaboré à partir de 1954 par Jean-Emile DUBOIS[30]. Ce type d'initiative permet également de révéler l'importance de l'information, de sa circulation et de la facilitation d'accès.

Par ailleurs, c'est vers la fin des années 1960 que les SIC commencent à s'intéresser à la recherche d'information, un domaine jusqu'alors très informatique, ce sont également les débuts de la RI en tant que domaine de recherche scientifique. Ce phénomène est dû entre autres aux travaux pionniers et retentissants de Gérard SALTON[31] qui a été un précurseur dans les approches « *modèles* » du domaine évoquées précédemment.

La recherche d'information en tant que sujet arrive alors peu à peu en France où l'*Information Retrieval* est traduite littéralement en *Recherche d'information*, terme qui peut sembler induit à première vue mais qui ne l'est pas tant que cela : *« Cette tentative de clarification conceptuelle est d'autant plus importante que de nombreux termes sont traduits de l'anglais, souvent maladroitement ce qui renforce la polysémie et la confusion qui en résulte.»*[32] Le terme anglais désigne, en effet, plutôt la notion de quête d'information ou processus d'information[33]. Le domaine est ensuite clarifié et redéfini par F. W. LANCASTER en 1979 ainsi : « *the process of searching a collection of documents with the goal of identifyng documents that are relate to a particular topic* »[34] distinguant alors deux types de recherche : la recherche documentaire (*information retrieval*) et la recherche de renseignements (*fact retrieval*)[35].

---

[29] Dalbin, S. (2007). Thésaurus et informatique documentaires : Des Noces d'Or. Documentaliste, 44(1), p. 76-80.
[30] *Ibid.*
[31]Salton, G. (1968). Automatic information organization and retrieval. Proceedings of the 11th coference on Computational linguistics ; Salton, G. (1989). Automatic text processing : the transformation, analysis, and retrieval of information by computer. Choice Reviews Online, 27(01). https://doi.org/10.5860/choice.27-0351 ; Salton, G., & McGill, M. J. (1983). Introduction to Modern Information Retrieval. Information storage and retrieval systems. New York, McGraw-Hill ; Salton, G. (1971). The SMART Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall, Inc eBooks.
[32] Ihadjadene, M., & Chaudiron, S. (2008). L'étude des dispositifs d'accès à l'information électronique. Dans HAL (Le Centre pour la Communication Scientifique Directe). French National Centre for Scientific Research.
[33] Chiaramella, Y. & Mulhem, P. (2007). *Op. cit.*
[34] Lancaster, F. W. (1979). Information Retrieval Systems : Characteristics, Testing, and Evaluation. New York, Toronto, Wiley.
[35] Ihadjadene, M., & Chaudiron, S. (2008). *Op. cit.*

Les années 1960 et surtout les années 1970, voient l'essor de la « *documentation automatique* » également appelée « *informatique documentaire* » notamment avec l'influence des travaux de Jean-Claude GARDIN[36] en France. Les premiers logiciels documentaires ouverts au grand public se développent en partie encouragés par l'intérêt grandissant des institutions à leur égard (bibliothèques, musées, archives…) : « *Au plan de la recherche française, la RI était alors surtout l'affaire des spécialistes des sciences de l'information, dont beaucoup ont rapidement perçu l'intérêt des outils informatiques dans leur domaine.* »[37] Car la recherche d'information n'était, auparavant, accessible qu'aux spécialistes et aux professionnels notamment de la documentation, une époque désormais révolue :

> « *où l'utilisateur final de l'information (le chercheur, l'ingénieur, le journaliste etc.) ne pouvait lui-même interagir avec les données, mais devait passer par les services de documentalistes ou de bibliothécaires, seuls habilités et formés au fonctionnement des systèmes de recherche d'information des années 60 et 70.* »

Au cours des deux décennies suivantes (les années 1970 et 1980), on constate une généralisation et ouverture de l'informatique à un plus large public encore ce qui permet aux utilisateurs d'avoir accès à l'information ou au document en ligne sans la médiation d'un documentaliste ou d'un centre documentaire. Les DAI ne sont alors plus uniquement destinés aux professionnels mais à tous types de publics qui s'en saisissent massivement.

En parallèle, la RI s'affirme davantage en tant que domaine de recherche pluridisciplinaire à la croisée de l'informatique et des SIC en particulier avec la conférence ACM-SIGIR[38] en 1971. De nouvelles réflexions naissent et élargissent à nouveau le domaine, c'est le cas de l'évaluation et des travaux de W. S. COOPER[39] sur la pertinence des résultats retournés par les systèmes.

---

[36] Gardin J-C. (1962). Documentation sur cartes perforées et travaux sur ordinateurs dans les sciences humaines. Revue internationale de documentation, vol. 29, p.83-92 ; Gardin J-C. (1967). Recherches sur l'indexation automatique des documents scientifiques. Revue d'informatique et de recherche opérationnelle, 1ere année, n°6, p. 27-46 ; Gardin J-C. (1974). Analyse documentaire et théorie linguistique. Les analyses de discours Neuchatel, Delchaux et Niestlé, (col Zéthos), p.120-128.

[37] Chiaramella, Y. & Mulhem, P. (2007). *Op. cit.*

[38] Special Interest Group for Information Retrieval.

[39] Cooper, W. E. (1971). A definition of relevance for information retrieval. Information Storage and Retrieval, 7(1), p. 19-37.

Ces travaux se poursuivent dans les années 1980 avec un intérêt grandissant de la part de la communauté scientifique française en particulier avec les travaux de FLUHR[40]. L'approche probabiliste continue de se développer en parallèle de nouveaux modèles comme les approches fondées sur la logique (*logic based IR*) en partie avec les travaux de C. J. van RIJSBERGEN[41] qui ont connu un succès retentissant en France et en Europe. Un groupe de recherche spécialisé dans la RI est créé en France, à Grenoble. Il s'agit du groupe MRIM (Modélisation et Recherche d'Information Multimédia) fondé en 1983 et encore actif aujourd'hui[42]. Par ailleurs, de nombreuses conférences sont organisées, permettant de structurer la discipline et d'en promouvoir les recherches. Les notions d'exhaustivité et de spécificité d'un document par rapport à la requête formulée voient le jour à cette époque, c'est également la période où la recherche multimédia connaît une réelle expansion et que les travaux sur l'hypertexte[43] et l'hyperdocument connaissent un grand engouement.

Il ne s'agissait pourtant jusqu'alors que des prémices de la RI qui connaît, à partir des années 1990, un développement encore plus important et rapide. Sans revenir sur l'évolution de tous les axes évoqués précédemment, les années 1990 marquent surtout un tournant dans l'histoire de la RI par l'arrivée et la popularisation d'Internet, la démocratisation de l'informatique dans les foyers et l'avènement des premiers moteurs de recherche en ligne : « *L'expansion des systèmes de recherche d'information grand public […] a entraîné à la fois une multiplication et une diversification des usagers et une hétérogénéité croissante des contenus* »[44].

Cette évolution multifactorielle permet aux recherches en RI de s'étendre à de nouveaux sous-domaines : les bases de données[45], les communications homme-machine[46], les débuts de

[40] Fluhr C. (1977). Algorithmes a apprentissage et traitement automatique des langues. Thèse de doctorat, Orsay, Université de Paris Sud ; Fluhr, C. (1981). Spirit : un système syntaxique et probabiliste d'indexation et de recherche d'informations textuelles. Dans Proceedings of ADBS-IDT81, Paris, ADBS, p. 113-116.
[41] Van Rijsbergen C. J. (1986). A New Theorical Framework for Information Retrieval. Proceedings of the ACM-SIGIR86 International Conference on Research and Development in Information Retrieval, Pisa, p. 194-200 ; Van Rijsbergen C. J. (1986) A non-classical logic for information retrieval. The Computer Journal 29, 6, p. 481-485, Van Rijsbergen C. J. (1989). Towards an information logic. Proceedings of ACM-SIGIR89 International Conference on Research and Developmtent in Information Retrieval Cambridge, Massachusetts USA, p. 77-86.
[42] Voir https://www.liglab.fr/fr/recherche/equipes-recherche/mrim [en ligne] consulté le 24/06/2023.
[43] Rouet, J. & Tricot, A. (1998). Chercher de l'information dans un hypertexte : vers un modèle des processus cognitifs. Les hypermédias : approches cognitives et ergonomiques, p. 57-74.
[44] Ihadjadene, M., & Chaudiron, S. (2008). *Op. cit.*
[45] Le Maitre J., Murisasco E. & Robert M. (1997). From Annotated Corpora to Databases : the SgmlQL Linguage. Dans J. Nerbonne (ed.), Linguistic Databases, CSLI Lecture Notes n°77, p. 37-58.
[46] Ingwersen, P. (1999). Cognitive Information Retrieval. Annual Review of Information Science & Technology, 34, p. 3-52.

la prise en compte des aspects cognitifs de la recherche[47], l'IA[48], l'axe multimédia[49], l'aspect multilingue (*cross language IR, multilingual IR*, TALN[50]), etc. C'est également l'époque des débuts des systèmes de questions-réponses (*question-answering systems*), dont nous aurons l'occasion de reparler. Un autre axe se développe également, celui du document structuré et dont l'objectif est « *de fournir aux usagers des réponses aussi focalisées que possible, en leur restituant les composants des documents les plus spécifiques à leur besoin* »[51]. Ce qui répond à une double problématique : le web a transformé la notion de document par l'aspect hypermédia de l'information contenue, le document devient alors un nouvel objet notamment théorisé par Roger T. PEDAUQUE qui en distingue trois caractéristiques majeures : le médium, la forme et le signe[52].

Les corpus s'étoffent considérablement d'année en année et ce de façon exponentielle, nécessitant d'améliorer les résultats des requêtes formulées. Cette prise de conscience s'accompagne de considérations nouvelles notamment le fait de ne retourner à l'utilisateur qu'un nombre limité de résultats pour éviter la surcharge cognitive de ce dernier[53].

2) La recherche d'information en contexte

Dans les années 2000, la massification et la diversification des publics qui pratiquent la RI mettent en lumière la variété de typologies et de situations de recherche soit les différentes conditions dans lesquelles l'usager effectue sa RI ce qui en modifie la méthode, l'objectif, le jugement de la pertinence, etc[54]. Cette prise en compte de la dimension utilisateur pour rendre la recherche la plus efficace possible se veut plus proche de la réalité, elle prend le nom de « *recherche d'information en contexte* » (*contextual IR*) et connaît, depuis sa création, un profond engouement dans le domaine : « *Parmi les limites des approches courantes, celle qui*

[47] Ingwersen P. & Järvelin K. (2005). The Turn : Integration of Information Seeking and Retrieval. Dans Contexte ; Kluver.
[48] Gallinari, P., Zaragoza H. & Amini M. (2002). Chapitre 11 : Apprentissage et Données Textuelles. Dans Bases de données et statistiques, Dunod.
[49] Baeza-Yates, R. & Ribeiro-Neto, B. (dir.). (1999). Modern Information Retrieval. Addison-Wesley.
[50] Traitement automatique du langage naturel.
[51] Chiaramella, Y. & Mulhem, P. (2007). *Op. cit.*
[52] Pédauque, Roger T. (2003). Document : forme, signe et médium, les re-formulations du numérique. STIC-CNRS.
[53] Luk, R. W. P., Leong, H. V., Dillon, T. S., Chan, A. T. S., Croft, W. B., & Allan, J. (2002). A survey in indexing and searching XML documents. Journal of the Association for Information Science and Technology, 53(6), p. 415-437. https://doi.org/10.1002/asi.10056
[54] Chiaramella, Y. & Mulhem, P. (2007). *Op. cit.*

a suscité un grand intérêt de la part de la communauté scientifique est la modélisation du contexte d'utilisation des systèmes et de leurs utilisateurs. »[55]

La RI devient alors encore plus vaste et multiforme[56] :

> « *il s'agit aussi bien de comprendre le fonctionnement intrinsèque des dispositifs (approche algorithmique et informatique), d'évaluer la performance du dispositif pour un usage donné, de comprendre les mécanismes d'appropriation par les usagers, d'identifier et de modéliser les pratiques informationnelles* ».

L'usager est alors désormais placé au centre de l'attention et des réflexions sur les DAI, le paradigme usager prend donc le pas sur le paradigme système qui dominait jusque-là sans l'évincer totalement toutefois.

Enfin, si les travaux ont d'abord porté sur le système en lui-même dans une perspective techno-centrée avant de s'intéresser à l'usager et la place qu'il occupe dans la RI (dans sa dimension cognitive, affective, etc), ils se sont rarement penchés sur les environnements socioéconomiques de l'usager lorsqu'il effectue sa ou ses recherche(s) : « *qui sont devenus déterminants avec l'émergence des moteurs de recherche internet. D'autres considérations interviennent désormais dans toute étude sur les usages des dispositifs d'accès à l'information : éthiques […], politiques […], juridiques […].* »[57] De nouvelles orientations qui ont nécessité d'avoir recours aux travaux d'autres disciplines que les SIC dont nous parlerons ultérieurement.

La recherche d'information en contexte consiste donc à ne plus s'intéresser uniquement au dispositif de recherche ni à l'utilisateur isolé au cours de sa session de recherche mais au contexte global de la RI (social, culturel, linguistique, etc). « *On cherche alors à intégrer dans l'analyse des pratiques informationnelles l'impact des dynamiques interpersonnelles et sociales.* »[58] Ces analyses rendent alors compte du « *comportement informationnel* » (*information behavior*) de l'usager, c'est le cas des travaux de P. INGWERSEN et A. TRICOT sur les modèles cognitifs, T. WILSON pour les modèles conceptuels et T. SARACEVIC pour le modèle stratifié.

Dans l'idée de prendre en compte l'environnement global impliqué dans la recherche, P. INGWERSEN a alors développé le modèle de « *polyreprésentation* » qui va bien au-delà du SRI

[55] Bellot, P., Cauvet, C., Pasi, G. & Vallès-Parlangeau, N. (2012). Introduction. Document numérique, 15, p. 7-8. https://www-cairn-info.ressources-electroniques.univ-lille.fr/revue--2012-1-page-7.htm.
[56] Ihadjadene, M., & Chaudiron, S. (2008). *Op. cit.*
[57] *Ibid.*
[58] *Ibid.*

en lui-même afin de démontrer le caractère dynamique de la RI. « *Il s'agit d'une nouvelle approche de la pertinence qui s'appuie sur le constat qu'une même source d'information est perçue différemment par les usagers, en fonction de leurs structures cognitives propres* »[59].

3) Les interfaces de recherche documentaire

La recherche d'information traduite de l'*Information Retrieval* comporte donc deux pans distincts : la recherche de renseignements (*fact retrieval*) et la recherche documentaire (*information retrieval*). C'est à la seconde que nous nous intéresserons désormais en lien avec les problématiques évoquées en introduction sur la découvrabilité des contenus au sein des interfaces documentaires car « *La conception des interfaces web est au cœur des architectures de l'information et constitue l'un des enjeux majeurs de nos sociétés de l'information* »[60] ce qui est particulièrement vrai dans le cas des bibliothèques : « *Dans le contexte des bibliothèques, le concept d'interface documentaire relève de l'architecture du système d'information documentaire* »[61].

Il convient de noter que les interfaces de recherche documentaire sont particulières, parmi les DAI, car elles ne permettent pas seulement l'accès à l'information, elles sont également des plateformes de travail et « *permettent l'accès à l'information électronique via des interfaces de recherche, de navigation, de lecture, de consultation, d'annotation…* »[62]. Les auteurs dressent ensuite une liste des critères permettant de distinguer et décrire les différents types de systèmes et mentionnent ainsi : la nature du contenu, le fonds documentaire, la façon de représenter l'information, le processus d'accès, les types de recherche et d'accès à l'information mais aussi la performance du système, les utilisateurs ou les possibilités de personnaliser certaines fonctions. Une grande diversité de critères qui engendre une multitude de pratiques et d'utilisation de ces interfaces.

Les interfaces de recherche documentaires connues et utilisées en France sont de natures assez diversifiées, on compte notamment les catalogues des bibliothèques (le Sudoc[63], le Catalogue de la BnF, WorldCat, etc), les catalogues des musées (JocondeLab), les répertoires

[59] *Ibid.*
[60] Muller, C. (2015). Interfaces documentaires innovantes. I2D - Information, données & documents, 52, p. 15-16. https://doi-org.ressources-electroniques.univ-lille.fr/10.3917/i2d.153.0015
[61] *Ibid.*
[62] Ihadjadene, M., & Chaudiron, S. (2008). *Op. cit.*
[63] Sudoc : Système Universitaire de Documentation

de revues (Cairn, DOAJ[64], project Muse, etc), les bases de données spécialisées (Scopus, Web of Science, PubMed, JSTOR, etc), les bibliothèques numériques (Gallica, Persée, Europeana, …), les archives ouvertes (HAL et ses déclinaisons, ArXiv). Par ailleurs, le champ de la recherche documentaire est également foisonnant car plusieurs projets ANR dédiés à l'innovation y travaillent actuellement, c'est le cas des projets Archival[65] et Philherit[66] présentés lors de la journée d'étude à la BnF le 20 juin.

C. MULLER propose alors une typologie des interfaces documentaires dites « *innovantes* » en distinguant trois groupes. Le premier est celui des interfaces les plus innovantes décrites ainsi :

> « *Les interfaces les plus innovantes ont intégré au moins cinq innovations ; elles ont développé des nouvelles fonctionnalités répondant aux pratiques actuelles des usagers et à l'évolution du Web. La force de ces interfaces tient à leur capacité de scénarisation des collections : en augmentant l'intelligence documentaire des contenus, elles proposent à l'usager un fil narratif à travers le dédale des ressources en réseaux.* » [67]

Le second groupe comporte les interfaces dites « *innovantes* » mais plus modestes car ne regroupant que deux à trois développements innovants. Le troisième groupe est constitué des interfaces les plus « *user friendly* », elles ont la particularité d'être très faciles à prendre en main, créatives au niveau du design etc mais de faire l'impasse sur l'innovation de quelque forme qu'elle soit.

Les interfaces de recherche documentaire ont d'ailleurs une place grandissante dans les pratiques des usagers pour plusieurs raisons : d'abord la numérisation massive des documents et collections des bibliothèques et centres d'archives permettant l'accès à distance, également les principes de la science ouverte (*open science*) qui valorisent la mise en ligne des contenus produits afin d'en permettre l'accès plus rapide, plus facile et au plus grand nombre mais

[64] DOAJ : Directory of Open Access Journals
[65] Porté par la Fondation Maison des Sciences de l'Homme (chaire UNESCO-ITEN) en partenariat avec Orange Lab, l'IRISA (Université Rennes 1) et LIS LAB (Université Aix Marseille), le programme Archival explore de nouvelles formes d'accès aux documents numériques grâce aux avancées dans le domaine de la compréhension automatique des contenus et de la génération de liens, à partir d'un corpus de recherche sur l'autogestion.
[66] Projet de recherche philosophique et interdisciplinaire portant sur l'héritage, PHILHERIT articule une Marque Blanche de Gallica avec des outils algorithmiques d'exploration de corpus issus du traitement automatique du langage (topic modeling, word embedding, etc.) pour proposer de nouvelles modalités de recherche et de navigation.
[67] Muller, C. (2015). *Op. cit.*

également car cela favorise la circulation des contenus. C'est ce que Sarra BEN LAGHA nomme le « *réseau d'information* » :

> « *La coopération et l'échange des données et des informations sont des concepts ancrés dans l'esprit et les pratiques des professionnels des bibliothèques. D'une part, parce que dans ce domaine on a toujours été conscient du fait qu'il est impossible de se procurer, par ses propres moyens, tous les documents dont on aurait besoin, et, d'autre part chaque ouvrage étant édité en centaines d'exemplaires, pour tout document qui arrive dans une bibliothèque, il y a de très fortes chances qu'un exemplaire ait déjà atterri dans une autre bibliothèque et qu'il y soit déjà catalogué. Pour éviter le travail en double du côté des professionnels et rendre de meilleurs services aux lecteurs ou chercheurs, les bibliothèques s'organisent généralement en réseaux d'information.* »[68]

II/ Les pratiques évaluatives

Après avoir présenté dans les grandes lignes les jalons de l'évolution des DAI et d'en avoir présenté quelques exemples, il est désormais temps de nous consacrer à l'étude de l'évaluation de ces dispositifs en trois points : les débuts de l'évaluation avec l'approche techno-centrée, le basculement vers l'approche orientée usager puis croisée et enfin les protocoles d'évaluation et campagnes existantes.

1) L'approche techno-centrée

Dès la création de la recherche d'information, il a été question d'évaluer les dispositifs mis en place permettant d'y accéder, d'abord puisque « *Toute activité de recherche qui veut revendiquer l'étiquette de scientifique se doit de disposer d'instruments de mesures* »[69] mais aussi car, dans le cas précis de la RI, il convient de déterminer l'efficacité des systèmes en vue de les améliorer :

[68] Ben Lagha, S. (2002). La numérisation des catalogues : une analyse rétrospective. Document numérique, 6, p. 81-97. https://doi-org.ressources-electroniques.univ-lille.fr/10.3166/dn.6.1-2.81-97
[69] Fluhr, C. (2004). Chapitre 1 : L'évaluation des systèmes de recherche d'informations textuelles. Dans Chaudiron, S. (dir.). Evaluation des systèmes de traitement de l'information, Hermès.

> « *Les premières études portant sur les dispositifs d'accès à l'information ont concerné l'analyse de leurs fonctionnalités (indexation, appariement, formulation et reformulation des requêtes, structuration et organisation des données et des informations, etc…) dans le but d'améliorer la performance globale des systèmes* »[70].

Les premiers travaux sur la RI traitaient déjà de cet aspect évaluatif et quantitatif. Les différentes évaluations de ce type ont été qualifiées de techno-centrées ou d'évaluations orientées système (*system oriented*) tandis que D. ELLIS lui donne en 1992 le nom de « *paradigme physique* » (*physical paradigm*)[71]. Cette approche est donc profondément tournée vers la technique, les fonctionnalités et la performance du système -dont une grande place est accordée à l'étude des algorithmes de recherche, aux modules d'appariement et d'ordonnancement-, elle a largement dominé les recherches en informatique documentaire de la création du domaine jusqu'aux années 2000.

Dans les systèmes de recherche documentaire, c'est la logique booléenne (issue des opérateurs booléens « *et* », « *ou* », et « *sauf* ») qui a été mise en place en premier lieu, c'est-à-dire interroger le système à l'aide de mots-clés articulés par les opérateurs. Pour faire simple, un index avec des mots-clés est associé à chaque document pour le caractériser, lorsqu'une requête est effectuée, le système cherche dans les index et doit rapporter les documents dont l'index correspond aux mots-clés recherchés[72]. Pour évaluer, il a donc été question de comparer la requête posée aux résultats associés.

Ce type de méthode est en particulier adaptée à l'étude des moteurs de recherche, domaine qui a suscité beaucoup d'intérêt et possède une littérature dédiée abondante -plus d'une centaine d'études spécialisées d'après Stéphane CHAUDIRON et Madjid IHADJADENE[73]. Des travaux de Craig SILVERSTEIN en 1998 sur le moteur Altavista[74] à ceux de Sherry KOSHMAN, Amanda SPINK et Bernard J. JANSEN en 2006 sur Vivisimo[75] (un métamoteur) ou plus

---

[70] Ihadjadene, M., & Chaudiron, S. (2008). *Op. cit.*
[71] Ellis D. (2002). The Physical and congitive paradigm in information retrieval research. Journal of Documentation, vol. 48, n°1, p. 45-64.
[72] Fluhr, C. (2004). *Op cit.*
[73] Ihadjadene, M., & Chaudiron, S. (2008 bis). Quelles analyses de l'usage des moteurs de recherche. Questions de communication, (14), p. 17-32. https://doi.org/10.4000/questionsdecommunication.604
[74] Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. Sigir Forum, 33(1), p. 6-12. https://doi.org/10.1145/331403.331405
[75] Jansen, B. J., Koshman, S., & Spink, A. (2006). Web searching on the Vivisimo search engine. Journal of the Association for Information Science and Technology, 57(14), p. 1875-1887. https://doi.org/10.1002/asi.20408

récemment aux travaux sur l'intelligence artificielle et son rôle dans les moteurs de recherche[76], les moteurs de recherche ont polarisé l'attention des chercheurs en particulier aux Etats-Unis et en France.

Dans les années 1960, naissent ainsi les premières esquisses de l'évaluation avec la création des mesures de « *rappel* » et de « *précision* » de Cyril CLEVERDON[77] : « *Le pionnier en la matière fut Cyril CLEVERDON, qui introduisit les mesures bien connues de rappel et de précision, et créa les toutes premières collections de test destinées à évaluer les performances qualitatives d'un SRI* »[78]. Dans la même décennie, les chercheurs en technologies de l'information ont proposé une nouvelle manière de juger les résultats non plus de manière binaire mais graduée selon un système de classement (*ranking*) par valeur de rang (*ranking value*), du plus pertinent au moins pertinent.

Il existe cependant plusieurs écueils auxquels se confrontent les chercheurs dans ce type de démarche : « *Pour aboutir à ce type de classement, il faut établir une métrique qui permet d'évaluer ce degré de pertinence.* »[79] Or, il existe une multitude de manières différentes de calculer cette gradation : modèles statistiques, linguistiques, neuronaux, etc. Dans ce cadre, les premiers résultats retournés par le système doivent être les plus pertinents et proches de la requête posée et les derniers les moins pertinents selon une courbe décroissante. Ce phénomène relève bien sûr de la théorie et d'un idéal à atteindre. Pour savoir si les documents pertinents ont été retournés par le système, il faut alors connaître l'intégralité du fonds interrogé ce qui n'est pas toujours possible (ce n'est pas le cas sur un moteur de recherche par exemple).

Il existe différentes méthodes d'évaluation en fonction des nombreux types de DAI, et elles sont répertoriés dans *Evaluation des systèmes de traitement de l'information*[80] où les différents SRI évoqués sont associés à leur méthode d'évaluation spécifique (la recherche textuelle, visuelle, par question-réponse, etc). Dans le cadre de notre étude, nous nous intéressons plus particulièrement à la recherche textuelle.

L'évaluation des DAI prend surtout de l'importance dans les années 1990 notamment avec les campagnes TREC (*Text REtrieval Conference*). Il s'agit d'un programme à l'origine

[76] Cazals, F. & Cazals, C. (2020). Intelligence artificielle : L'intelligence amplifiée par la technologie. De Boeck Supérieur.
[77] Cleverdon, C. W., Mills, J. & Keen, M. (1966). Factors determining the performance of indexing systems. Association of Special Libraries and Information Bureau, Cranfield (Angleterre) ; Cleverdon C. (1970). Progress in documentation, evaluation tests of information retrieval systems. Journal of Documentation, 26, p.55-67.
[78] Chiaramella, Y. & Mulhem, P. (2007). *Op. cit.*
[79] Fluhr, C. (2004). *Op cit.*
[80] Chaudiron, S. (dir.) (2004). *Op cit.*

très orienté système, laissant très peu de place à l'utilisateur, il mêle sciences dites « *dures* » et SHS car si les mesures de rappel et de précision sont utilisées quantitativement, elles se basent sur la notion floue de la pertinence qui ne peut être déterminée par un théorème, « *le domaine n'est pas aussi stricte que les mathématiques et la physique* »[81]. De plus, l'interprétation des résultats obtenus peut s'avérer complexe comme le rappelle l'auteur :

> « *l'évaluation donne un niveau de qualité qui prend comme référentiel l'évaluation par des professionnels de la documentation d'un nombre limité de documents (les meilleurs documents pour chaque système et chaque question). Cela indique qu'un grand nombre de documents ne sont pas évalués et qu'ils sont donc par défaut comptés comme non pertinents.* »

Il y a donc des biais conséquents et des limites importantes dans ce type d'évaluation et les critiques adressées sont diverses :

> « *les critiques portent sur le protocole d'évaluation lui-même […], l'absence de l'usager dans le processus d'évaluation […], le problème de la représentation trop mécanique du besoin informationnel […], la modélisation simplifiée à l'extrême des pratiques informationnelles […], et enfin une simplification également de la notion de pertinence.* »[82]

Ce à quoi les chercheurs ont souhaité remédier notamment en développant un nouveau paradigme de recherche dans l'évaluation : l'approche usager.

2) De l'approche orientée usager à l'approche croisée

Dans les années 1960 déjà, certains chercheurs nuançaient cette approche techno-centrée en alertant sur la complexité de la pertinence, sa gradation ou encore sur la diversité des besoins informationnels comme ce fut le cas de R. TAYLOR en 1968[83]. Dès lors, le besoin informationnel des usagers a amené à une sensibilisation sur l'importance de la prise en compte de l'usager et de ses besoins informationnels dans le domaine de la RI. On doit également à (SALTON, 1971, 1989) de nombreuses avancées sur l'évaluation des DAI, en particulier parce qu'il s'est quelque peu éloigné des approches booléennes pour proposer l'approche vectorielle.

[81] Fluhr, C. (2004). *Op cit.*
[82] Ihadjadene, M. & Chaudiron, S. (2008). *Op cit.*
[83] Taylor R. (1968). Question negociation and information seeking in libraries. College and research libraries, n°29, p. 178-194.

Cette dernière introduit l'idée de mesurer la similarité entre la requête et le document pour évaluer la pertinence d'un système[84].

L'approche orientée usager va cependant plus loin car elle « *considère que l'attention doit être portée sur les attentes réelles de l'usager vis-à-vis du système et sur son environnement* »[85]. Les auteurs identifient quatre courants majeurs sur lesquels travailler à partir de ce postulat : l'amélioration des protocoles d'évaluation avec une meilleure définition de la pertinence, une approche plus inspirée de la psychologie-cognitive pour mieux rendre compte des caractéristiques individuelles des expérimentateurs, la proposition de différents modèles et une approche dite « *holistique* » ou croisée pour passer outre l'opposition entre approches orientées système ou usager pour prendre en compte l'ensemble de la RI comme un tout où la session de recherche n'est qu'une étape.

L'approche orientée usager ne remplace pas l'approche techno-centrée, elle s'y ajoute et tente de la compléter. On compare alors la pertinence système avec la pertinence du côté usager pour les confronter et vérifier qu'elles correspondent sans quoi le système sera jugé moins bon :

> « *Cette modélisation entre réalité (le point de vue utilisateur) et modélisation (le point de vue du SRI) devient une dimension clé du domaine : elle place d'une part l'utilisateur au centre du processus (un système n'est pas acceptable si son modèle de pertinence n'est pas proche de celui des usagers), en même temps qu'elle impose à tout approche l'épreuve du feu, c'est-à-dire la confrontation à des situations réalistes.* »[86]

Enfin, les différents travaux prenant en compte les aspects cognitifs de la recherche ont permis de développer un modèle des processus cognitifs qui ont lieu lors d'une recherche d'information et se répètent de façon cyclique, il s'agit du modèle EST soit « *l'évaluation, la sélection et le traitement* », c'est un modèle de recherche cyclique qui postule que toute recherche d'information fait appel à ces trois processus.

[84] Chiaramella, Y. & Mulhem, P. (2007). *Op. cit.*
[85] Ihadjadene, M. & Chaudiron, S. (2008). *Op cit.*
[86] Chiaramella, Y. & Mulhem, P. (2007). *Op. cit.*

3) Les protocoles d'évaluation

Dans les années 1990, se développent des grands programmes ou campagnes d'évaluation des DAI qui sont aussi nombreux que variés. Sans chercher à dresser ici un état des lieux complets de tous les protocoles d'évaluation des DAI qui existent, il s'agit plutôt de faire mention des plus importants et d'en tirer des conclusions pour l'élaboration de notre méthode. Il convient également de rappeler que la RI proposait, dès 1960, des expérimentations afin d'évaluer les qualités, performances et limites des systèmes[87].

Le programme de Cranfield est celui qui a posé les premiers jalons de l'évaluation de manière assez complète, mais le plus connu d'entre eux est le programme TREC élaboré conjointement par le NIST (*National Institute of Standards and Technology*) et le DOD (*Department of Defense*) avec des crédits de la DARPA (*Defense Advanced Research Program Agency*) et dont la première conférence a eu lieu en 1992 et se déroule ensuite annuellement. Le programme était initialement dédié à deux tâches principales : le filtrage (*filtering*) et la recherche[88]. Il s'est ensuite étoffé pour comprendre aujourd'hui 7 axes différents d'expérimentation (nommés *Tracks*) dédiés à des sous-domaines spécifiques de la RI. Les *tracks* sont les suivantes : la recherche dans d'autres langues que l'anglais, l'interrogation interlingue (*cross-linguage interrogation*), la recherche précise dans les premières réponses, l'interrogation interactive qui comprend un utilisateur, la recherche sur des grands volumes et masses documentaires ou encore le système questions-réponses.

Le programme américain a été repris en France au travers de la campagne Amaryllis, initiée en 1996 par l'AUF (Agence Universitaire de la Francophonie) et le ministère français de la Recherche et de la Technologie et reconduit pour une deuxième campagne en 1998-1999. Basée sur le protocole TREC, Amaryllis correspond au premier *track* identifié à savoir la recherche dans d'autres langues que l'anglais. Plusieurs articles dédiés à la campagne présentent et expliquent le projet[89] : la première campagne adoptait une démarche exploratoire pour mettre en place la méthodologie d'évaluation et créer les premières collections pour les tests auprès de 8 participants, la seconde proposait trois axes (la recherche unilingue,

[87] Cabanac, G., Hubert, G., Boughanem, M. & Chrisment, C. (2011). Impact du « biais des ex aequo » dans les évaluations de recherche d'information. Document numérique, 14, p. 149-168.
[88] Fluhr, C. (2004). *Op cit.*
[89] Chaudiron, S. & Schmitt, L. (2000). Amaryllis : an evaluation-based program for Text Retrieval. Dans Jacquemin C., Mariani J. & Paroubek P. (dir.). Workshop Proceedings of LREC – Using Evaluation within HLT Programmes : Results and Trends. ELRA. Athens, p. 65-68 ; Coret A., Kremer P., Landi B. Schibler D., Schmitt L. & Viscogliosi N. (2000). Accès à l'information textuelle en français : le cycle exploratoire Amaryllis, Ressources et évaluation en ingénierie des langues, Bruxellles, De Boeck & Larcier, p.13-24.

multilingue et l'axe de test sur internet auprès de 11 participants). L'objectif annoncé de la démarche était divisé en trois axes : sensibiliser la communauté scientifique sur l'importance et la difficulté de l'évaluation, améliorer la qualité et l'efficacité technique des systèmes et services, constituer et mettre à disposition des collections de test pour l'évaluation des SRI[90].

Cette démarche est suivie en Europe dans les années 2000 avec les initiatives CLEF (*Cross Language Evaluation Forum*) dédiées spécifiquement au travail sur la RI multilingue, reprenant ainsi le second axe identifié par le TREC. Le protocole CLEF offre ainsi un cadre standardisé pour l'évaluation de la recherche multilingue aux différentes communautés scientifiques internationales. On pourrait également mentionner le programme INEX (*Initiative for the Evaluation of XML Retrieval*) cette fois consacré aux expérimentations concernant les documents XML. Il serait également possible de mentionner le programme IREX (*Information Retrieval and Extraction Exercise*), les campagnes NTCIR au Japon en collaboration avec Taïwan et la Corée en 1999 sur les langues asiatiques ou les nombreux programmes dédiés au TALN : le MUC (*Message Understanding Conference*), le TIDES (*Translingual Information Detection, Extraction and Summarization*) lancé par la DARPA aux Etats-Unis en 1999, le EAGLES (*Evaluation of Natural Language Processing Systems*), un programme européen de TALN entre 1991 et 1995.

Ces différents programmes apportent plusieurs évolutions : d'abord la taille des collections de test est très importante on pourrait même dire qu'elles sont massives avec parfois jusqu'à plusieurs millions de documents, ce qui correspond au souhait de les rendre réalistes par rapport aux conditions réelles de la RI[91] et tente d'éviter l'écueil du test en laboratoire ; ensuite, la réalisation de campagnes d'expérimentation d'ampleur ; enfin, l'aspect compétitif des expérimentations qui sont réalisées en parallèles et dont les résultats sont confrontés à la fin des expérimentations lors de conférences. Ainsi, ceux qui sont à l'origine d'un système ont l'occasion de le tester dans le cadre de la campagne et de le confronter à ceux des autres dans une perspective d'amélioration scientifique et commerciale. Les collections de documents sont également préparées à l'avance et spécialisées en vue de chaque expérimentation ce qui pousse sans cesse la création de nouvelles métriques adaptées à chaque évaluation.

Pour le programme TREC par exemple, les résultats discutés en fin d'expérimentation au NIST sont les suivants :

[90] Chaudiron, S. & Schmitt, L. (2000). *Op cit.*
[91] Chiaramella, Y. & Mulhem, P. (2007). *Op cit.*

> « *la liste des documents relevés pour chaque thème ; les actions de l'utilisateur et les principaux événements lors de l'interaction ainsi que le timing ; un questionnaire sur les connaissances antérieures de l'utilisateur avant le test et la satisfaction après le test ; une description complète du déroulement d'une des questions choisies par le NIST* »[92].

Ils permettent ainsi la comparaison entre les différents systèmes et la captation de leurs points faibles et avantages à chacun sur ces critères.

Les interfaces dites « *innovantes* » ont également fait l'objet de travaux dédiés à leur évaluation, c'est le cas de C. MULLER qui dresse les critères d'évaluation qui permettent d'identifier ce type d'interface à partir de l'évaluation de dix d'entre elles[93]. Elle explique ainsi qu'il existe trois types d'innovation : l'innovation technique, économique et politique et que chacun peut être divisée en deux branches : l'innovation sociale ou l'innovation graphique.

III/ Les outils et concepts pour l'évaluation

Cette partie tend à faire la synthèse de ce qui a été vu précédemment dans l'objectif de mettre au point une boîte à outils sur l'évaluation pouvant nous servir dans la réalisation de notre étude.

1) Mesures et métriques

Plusieurs mesures ont alors vu le jour au travers des différents protocoles et campagnes d'évaluation, des métriques qui -bien que contestées- demeurent importantes : « *Aujourd'hui, la volumétrie a évolué, les besoins se sont diversifiés et les problématiques se sont complexifiées, mais l'on observe encore les mêmes indicateurs de référence depuis des décennies.* »[94] Les métriques sous-entendues étant notamment le rappel, la précision et la F-mesure. Des métriques qui sont pourtant liées à des concepts bien plus difficiles à quantifier et à évaluer comme la pertinence, le besoin informationnel, le gain de temps ou encore la

[92] Fluhr, C. (2004). *Op cit.*
[93] Muller, C. (2015). *Op cit.*
[94] Timimi, I. (2020). Évaluation d'outils et outils d'évaluation. Revue COSSI, (9). https://doi.org/10.34745/numerev_1560

satisfaction de l'usager, comme le confirme l'auteur : « *ses concepts et paramètres sont associés à un réseau sémantique étendu (performance, pertinence, distance, référentiel, métrique, besoin informationnel)* ».

Rappel et précision

Les métriques les plus connues et courantes dans l'évaluation des SRI sont alors le rappel et la précision[95]. Le rappel consiste à quantifier le nombre de réponses correctes ou pertinentes rapportées par le système parmi l'ensemble de la collection, la précision s'intéresse au contraire au nombre de documents pertinents rapportés parmi les résultats retournés par le système. Les deux mesures se complètent et n'ont de sens qu'en synergie car l'une ou l'autre, prise isolément, ne donne pas de réelle indication sur la qualité des résultats proposés. La première traite l'exhaustivité des résultats rapportés tandis que l'autre reflète la qualité des réponses fournies.

Ces deux mesures permettent alors d'estimer le bruit et le silence parmi les résultats proposés par le DAI. Si un système rapporte tous les documents de sa collection, il aura donc le score maximal en termes de rappel car tous les documents pertinents seront retournés par le système mais il y aura aussi une masse conséquente de documents non pertinents rapportés, il s'agit du bruit. Les résultats non pertinents rapportés complexifient l'accès aux résultats pertinents et dégrade donc la qualité du système et de la satisfaction de l'usager.

À l'inverse, si aucun document non pertinent n'est rapporté par le système mais que cela implique d'avoir une liste de résultats très lacunaire voire vide, on parlera de silence car il manque des documents pertinents dans la liste de résultats. Pour qu'un système soit jugé satisfaisant, il faut trouver un équilibre entre les deux métriques et faire en sorte de les améliorer sans dégrader la qualité de l'autre. Il existe une mesure supplémentaire nommée F-mesure qui consiste à « *faire une synthèse entre rappel et précision, en favorisant les systèmes dont les mesures de rappel et de précision sont voisines.* »[96], cette mesure a pour objectif d'être une moyenne des deux métriques permettant de juger de la performance d'un système.

Or, si la précision est plutôt facile à quantifier, le rappel est -pour sa part- bien souvent problématique dans l'évaluation car il nécessite de connaître l'ensemble des documents pertinents pour la requête formulée ce qui s'avère souvent impossible :

[95] Cleverdon, C. W., Mills, J. & Keen, M. (1966). *Op cit* ; Cleverdon, C. (1970). *Op cit.*
[96] Nazarenko, A. & Poibeau, T. (2004). Chapitre 5 : L'évaluation des systèmes d'analyse et de compréhension de texte. Dans Chaudiron, S. (dir.). Evaluation des systèmes de traitement de l'information, Hermès.

> « *la précision est une notion facile à évaluer, car elle ne demande aucune information sur le nombre total de documents pertinents dans la collection, le rappel est une notion qui pose plusieurs problèmes. Il est d'ailleurs impossible de l'évaluer dans les cas réels où l'utilisateur ne connaît pas à l'avance le nombre de documents pertinents correspondant à sa requête.* »[97]

Cependant, pour calculer ces deux mesures, il faut déterminer de façon binaire et systématique si une réponse est correcte ou pertinente, ce qui engendre un certain nombre de complexités, notamment quant à la définition et la détermination de la pertinence d'un document ou d'un résultat.

La pertinence

Le rappel et la précision sont jugés en fonction de la notion de pertinence, une notion pourtant extrêmement difficile à définir car s'il existe pléthore de définitions, leur application concrète s'avère bien souvent très épineuse :

> « *L'évaluation dans le domaine de la recherche d'information est un problème compliqué. Cela vient du fait que la pertinence est une notion subjective : le même document peut être jugé diversement par deux utilisateurs, ou dans deux conditions différentes. Ainsi, rendre les machines capables d'évaluer l'efficacité des systèmes de recherche d'information n'est pas facile.* »[98]

Les premiers travaux concernant le concept de pertinence datent des années 1970 avec les recherche de T. SARACEVIC[99]. Plus récemment, la définition suivante a été proposée :

> « *La pertinence peut ainsi être définie selon au moins quatre dimensions : le besoin d'information, décomposé en besoin réel, besoin perçu par l'utilisateur, besoin exprimé, et besoin formalisé par un langage de requête ; les composants : l'information elle-même, la tâche et le contexte ; le temps relevé pour retrouver l'information ; la granularité de l'information recherchée : document complet, sujet du document, ou information précise à l'intérieur de ce document.* »[100]

---

[97] Audeh, B., Beaune, P. & Beigbeder, M. (2015). MOR : Mesure orientée rappel pour les systèmes de recherche d'information. Document numérique, 18, 37-54.
[98] *Ibid.*
[99] Saracevic T. (1970). The concept of Relevance. Introduction to Information Science, R.R. Bowker Compagny. Dans Chiaramella, Y. & Mulhem, P. (2007). *Op cit.*
[100] Sitbon, L., Bellot, P. & Blache, P. (2010). Vers une recherche d'information adaptée aux utilisateurs dyslexiques. Document numérique, 13, p. 161-185.

Il faut également prendre en compte le fait que la pertinence dépend des usagers, elle n'est pas la même selon chacun, d'une situation à l'autre, car elle peut être pertinente un jour et non pertinente le lendemain ou dans un autre contexte, et elle existe au travers d'une large gradation. Elle correspond alors à plusieurs critères comme l'expliquent B. MOULAHI, L. TAMINE et S. BEN YAHIA dans l'article qui y est consacré :

> « *la pertinence est estimée en globalité selon un ensemble de dimensions qui s'apparentent à des familles de critères ; parmi ces différentes dimensions, on cite les plus reconnues dont : la pertinence thématique (contenu et méta-contenu), la pertinence situationnelle (temps et géolocalisation) et la pertinence cognitive (expertise, centres d'intérêts).* »[101]

Les auteurs soulèvent donc le fait que la pertinence soit multifactorielle et que ces différentes dimensions s'articulent entre elles dans une relation d'interdépendance :

> « *Un autre résultat important est l'interdépendance de ces dimensions pour inférer la pertinence globale d'un document* […]. *En clair, un utilisateur juge de la pertinence d'un document en tenant compte conjointement de l'ensemble des critères de pertinence ; à titre d'exemple, un document est d'autant plus pertinent du point de vue du contenu que l'expertise de l'utilisateur est en lien avec ce contenu.* »[102]

Il est désormais aisé de comprendre la complexité de cette notion et la fragilité des mesures de rappel et de précision qui la prennent comme origine : « *Cette orientation [orientée usager] reprend finalement les hypothèses depuis longtemps formulées par les pionniers de la RI, quant à la complexité et au caractère multiforme de la notion de pertinence.* »[103]

Enfin, la fiabilité des informations retournées par le système est particulièrement importante dans l'évaluation car si un système soumet des résultats en accord avec le thème ou la requête formulée mais que ces résultats sont faux, ils risquent d'être pris en compte dans les calculs de performance alors qu'ils sont faux et de biaiser l'ensemble.

> « *Le principe de l'évaluation repose sur l'idée qu'une donnée ne peut être prise en compte si elle ne présente pas un degré de fiabilité suffisant au risque de*

[101] Moulahi, B., Tamine, L. & Ben Yahia, S. (2016). Estimation de la pertinence multidimensionnelle en recherche d'information : Évaluation de l'application d'un opérateur flou d'agrégation. Document numérique, 19, 59-82.
[102] *Ibid.*
[103] Chiaramella, Y. & Mulhem, P. (2007). *Op cit.*

polluer la connaissance produite. L'évaluation est donc supposée garantir la fiabilité des informations qui vont être exploitées. C'est pourquoi elle est donc considérée comme une étape essentielle nécessitant une attention particulière, puisque d'elle dépend la qualité de la connaissance produite. »[104]

La fiabilité et la véracité d'une information sont également des critères à prendre en compte bien que souvent difficile à estimer.

2) Concepts mobilisés

Comme l'explique I. TIMIMI, il existe des dichotomies et des oppositions entre les différentes méthodes « *quantitative/ qualitative, automatique/ manuelle, verticale/ horizontale, boîte transparente/ boîte noire, ex-ante/ex-post, intrants/extrants, interface dynamique/ à interface statique, orientée système/orientée usager…* »[105] dont nous allons désormais faire état.

Les différents types d'évaluation

Il existe donc plusieurs types d'évaluation selon l'objet observé, ce qui est attendu ou encore en fonction des personnes ou groupes qui la réalisent. Chaque type d'évaluation porte un nom précis et défini comme l'explique S. CHAUDIRON :

« *On peut définir l'évaluation comme une relation qui vise à déterminer un indice de satisfaction, c'est-à-dire à déterminer à quoi quelque chose est bon pour quelqu'un ; c'est une fonction qui lie le système à évaluer et ses utilisateurs à des finalités. Dans cette optique, une première distinction est effectuée entre trois types d'évaluation : l'évaluation de progression, l'évaluation de mise en adéquation, et l''évaluation de diagnostic.* »[106]

L'évaluation de progression, également appelée évaluation de performance, est elle-même subdivisée en deux branches : l'évaluation horizontale (d'appariement) ou l'évaluation verticale (de progression)[107]. La première consiste à comparer deux systèmes similaires tandis que la seconde s'attache à évaluer plusieurs versions successives d'un même système pour en déterminer la performance au regard d'indicateurs identiques définis en amont. Ce type

[104] Bulinge, F. (2022). Chapitre 6. Évaluer l'information. Dans : F. Bulinge, Maîtriser l'information stratégique: Méthodes et techniques d'analyse (pp. 105-119). Louvain-la-Neuve: De Boeck Supérieur.
[105] Timimi, I. (2020). Évaluation d'outils et outils d'évaluation. Revue COSSI, (9).
[106] Chaudiron, S. (dir.) (2004). *Op cit.*
[107] Timimi, I. (2020). *Op cit.*

d'évaluation est surtout utilisé par les concepteurs et développeurs des systèmes afin de faire ressortir qualités et défauts de chacun ou de mesurer l'évolution entre deux versions. C'est également le principe bien connu des méthodes dites « *A/B* » (test A/B ou *A/B testing*) qui permettent de proposer une interface, une fonctionnalité ou une présentation différente d'un même projet à deux groupes distincts pour choisir la plus efficace ou la plus plébiscitée par les usagers. Ce type de méthode « *aide entre autres à améliorer la présentation d'une offre d'un site* »[108].

La deuxième évaluation, dite de diagnostic « *permet de déterminer l'état d'un système afin d'en mesurer ses performances intrinsèques ou de découvrir l'origine et la cause des erreurs* »[109], il s'agit également d'une évaluation plutôt réalisée par les concepteurs d'un système qui permet de connaître la performance certes mais qui est également utilisée à des fins de corrections et d'amélioration du système. Ce type d'évaluation peut aussi être réalisée entre différentes technologies concurrentes tant que ces dernières sont suffisamment similaires pour être comparées. Dans ce type d'évaluation : « *l'usager cherche à déterminer à partir d'une série de tests les sources de performance ou d'imperfection d'un système par rapport à une tâche précise* »[110] assignée à l'avance et uniformément pour tous les testeurs.

Le troisième type est la mise en adéquation (*adequacy evaluation* ou *formative evaluation*), cette évaluation « *vise à déterminer l'adéquation d'un système au regard d'un quelconque usage souhaité de ce système.* »[111] avec une nuance toutefois entre les usages projetés par les concepteurs ou par les usagers du système :

> « *Si l'usage est "projeté" par les concepteurs du système, l'évaluation portera sur l'adéquation de l'offre technique par rapport à la demande ; l'analyse peut également porter sur les différentes logiques d'usage identifiées chez les utilisateurs prenant en compte notamment les détournements des objets techniques.* »[112]

[108] Medioni, S. & Benmoyal-Bouzaglo, S. (2018). Chapitre 4. L'évaluation du consommateur en situation d'hyperchoix. Dans : S. Medioni & S. Benmoyal-Bouzaglo (dir), Marketing digital (p. 141-184). Paris : Dunod.
[109] Chaudiron, S. (dir.) (2004). *Op cit.*
[110] *Ibid.*
[111] *Ibid.*
[112] *Ibid.*

Tandis que :

> « *D'un point de vue "utilisateur", l'analyse de l'adéquation d'une offre par rapport à un besoin est la plus importante. Il ne s'agit pas d'identifier le meilleur système dans l'absolu mais de procéder à une évaluation comparative permettant à l'utilisateur d'effectuer un choix.* »[113]

Boîte noire/ Boîte transparente

L'une des grandes distinctions opérées entre les différents types d'évaluation se situe aussi du côté de l'explication et la compréhension du fonctionnement du système évalué par le ou les évaluateurs. La démarche la plus courante est celle dite de la « *boîte noire* » (*black box*) car elle permet d'évaluer le résultat produit par le système sans prendre en compte le fonctionnement interne de ce dernier, elle est donc plus simple à mettre en œuvre et semble poser moins de problèmes d'interprétation. Une démarche « *"boîte noire", c'est-à-dire que l'on observe uniquement les données fournies au système (input) et les résultats que celui-ci produit (output), sans s'intéresser au fonctionnement interne du système* »[114] s'intéresse donc uniquement au résultat sans aucune forme de prise en compte du processus. Ainsi, l'évaluation de type boîte noire ne prend pas en compte la performance individuelle des différentes briques qui composent le système et ne s'attache qu'à l'ensemble produit, « *Les prétraitements des données effectués par les différents modules du système ne font l'objet d'aucune évaluation dans cette démarche.* »[115]

De l'autre côté, l'évaluation dite « *boîte transparente* » (*glass box*) est bien moins utilisée dans les différentes campagnes d'évaluation car plus complexe à mettre en place et plus coûteuse d'autant plus qu'elle est parfois considérée comme étant moins objective. Elle consiste à mettre en lumière la façon dont le système part d'une donnée, d'un corpus ou autre pour arriver au résultat produit et retourné à l'usager. Elle est plus complexe car il convient de faire évaluer le DAI par des évaluateurs capables de comprendre le fonctionnement du système et des différentes « *briques* » agencée ou de faire un système « *explicable* » qui soit en mesure de se rendre compréhensible par l'usager non spécialiste.

[113] *Ibid.*
[114] *Ibid.*
[115] Timimi, I. (2020). *Op cit.*

Evaluation statique ou dynamique

Il est également important de prendre en compte le fait que l'évaluation peut être statique ou dynamique, un choix qui est notamment dû à l'état du système évalué : s'il est achevé ou non. La première consiste à évaluer le système et ses performances à un moment donné et sans prise en compte de modifications ultérieures ou en cours, ni prise en compte d'ajouts de la part des testeurs « *cette approche s'adresse essentiellement aux utilisateurs des systèmes qui n'offrent pas la possibilité d'une modification ou d'un ajout de ressources extérieures telles que des connaissances linguistiques (thésaurus, lexique, dictionnaires, etc.)* »[116]

La seconde consiste à faire l'inverse, à savoir évaluer un système qui continue d'être modifié. Effectivement, il serait possible de penser -à première vue- qu'un système ne peut faire l'objet d'une évaluation qu'une fois terminé, or l'évaluation dynamique permet de repérer les problèmes avant la fin du projet pour les corriger, rectifier et apprendre des erreurs. C'est ce qu'explique I. TIMIMI : « *l'évaluation à interface dynamique permet d'étudier la valeur ajoutée et l'impact des ressources d'enrichissement introduites dans le système […] ou des choix d'orientation ordonnés par l'usager.* »[117] Ce type d'évaluation permet également d'adopter une démarche itérative d'essais, erreurs, progression permettant l'amélioration continue du système en continu. Cette démarche ne vise donc pas uniquement l'évaluation d'un résultat final mais l'évaluation d'un système à un moment en vue de l'améliorer. L'évaluation dynamique s'appuie donc en partie sur l'élaboration de recommandations afin d'améliorer le système proposé.

Autres dichotomies

Il serait également bon d'ajouter qu'il existe bien d'autres types de différenciations et ce notamment entre l'analyse quantitative et l'analyse qualitative, entre l'évaluation automatique ou manuelle, entre intrants/extrants... L'évaluation est également souvent associée (voire qualifiée) par sa finalité, son objectif ce qui permet sans cesse de créer de nouvelles catégorisations, c'est ce qu'avance I. TIMIMI :

> « *l'évaluation est souvent associée à ses finalités, elle peut avoir le sens de valorisation dans les activités économiques et marketing ; le sens de certification dans les activités d'assurance qualité ; le sens de notation, en DRH ou plus*

[116] Chaudiron, S. (dir.) (2004). *Op cit.*
[117] Timimi, I. (2020). *Op cit.*

généralement pour tout domaine ayant recours à une échelle ou un référentiel de cotation. »[118]

Une classification qui demeure toutefois assez peu citée et semble anecdotique par rapport à celles citées précédemment.

3) L'évaluation : une approche pluridisciplinaire

L'évaluation est une donc une discipline éminemment pluridisciplinaire dès ses origines entre informatique et SIC ce qui s'est renforcé avec les débuts de l'approche orientée usagers et au contexte dans lequel les recherches sont effectuées. Cette discipline se trouve aujourd'hui au croisement des SIC, et de nombreuses autres disciplines :

> « *L'étude des systèmes de recherche d'information (SRI) et plus généralement des dispositifs d'accès à l'information est un vaste domaine de recherche au carrefour de plusieurs disciplines, en particulier les sciences de l'information, de la psychologie cognitive, de la linguistique, de l'informatique et de l'intelligence artificielle.* »[119]

On pourrait ajouter les mathématiques, la sociologie des usages ou encore l'ergonomie et le design et bien d'autres. Une ouverture toujours plus importante :

> « *Ces domaines sont en pleine mutation et s'ouvrent en permanence à de nouvelles thématiques. Les travaux de recherche sont extrêmement variés et complémentaires, allant des modèles théoriques pour l'analyse de l'écrit et l'extraction d'information dans les images, jusqu'à l'analyse et la recherche d'information dans les documents structurés sur le web, en passant par l'usage de nouvelles interfaces homme-machine à base de stylets ou une recherche d'information intelligente.* »[120]

Les autres disciplines que les SIC utilisent également l'évaluation selon différentes formes, on peut penser aux sciences de l'éducation bien sûr mais également au domaine des

---

[118] *Ibid.*
[119] Ihadjadene, M. & Chaudiron, S. (2008). *Op cit.*
[120] Grau, B., Laleau, R. & Ramel, J. (2011). Introduction. Document numérique, 14, p. 7-10. https://www-cairn-info.ressources-electroniques.univ-lille.fr/revue--2011-2-page-7.htm.

ressources humaines, de la gestion, de la psychologie, etc qui peuvent ainsi l'enrichir de leurs approches et méthodes.

La psychologie cognitive a permis de nombreux apports : « *Plusieurs campagnes d'évaluation comme INEX ou InFile s'attachent à mieux prendre en compte l'usager et son contexte dans les métriques et les protocoles.* »[121] les campagnes d'évaluation s'inspirent alors progressivement de la psychologie cognitive pour l'appliquer à la RI (la démarche, les concepts, la méthode). Elle permet également de mettre en avant le fait que la RI possède une dimension affective importante qu'il convient de ne pas négliger, pensée qui se développe surtout dans les années 1980 avec la prise en compte de plus en plus marquée de la dimension cognitive des usagers : « *L'approche cognitive est perçue comme particulièrement féconde tant sur le plan conceptuel que méthodologique* » par une partie des chercheurs en SIC. Il pourrait être intéressant de noter qu'« *En psychologie cognitive, la recherche d'information est considérée comme une tâche secondaire au service d'une autre activité* »[122], un postulat qui ne peut qu'accroître les travaux sur la recherche d'information en contexte.

La sociologie des usages a également joué un rôle important dans l'évaluation des DAI à partir des années 1980 notamment avec les travaux de J. JOUET[123] ou plus récemment de D. BOULLIER[124].

Cependant, cette pluridisciplinarité ajoute également de la complexité à l'évaluation qui l'est déjà tant au niveau théorique que pratique comme le montre I. TIMIMI :

> « *l'évaluation dans le cas des systèmes de traitement d'information présente un travail de recherche multidisciplinaire assez préoccupant. Le manque de normalisation consensuelle additionné à la difficulté d'une modélisation rend l'évaluation comme objet de recherche complexe et très discutable dans les différents champs disciplinaires* »[125].

[121] Ihadjadene, M. & Chaudiron, S. (2008). *Op cit.*
[122] Dedencker C. & Kolmayer E. (2006) Eléments de la psychologie cognitive pour les sciences de l'information, Villeurbanne, Presses de l'ENSSIB.
[123] Jouët, J. (2000). Retour critique sur la sociologie des usages. Réseaux, 18(100), p. 487-521. https://doi.org/10.3406/reso.2000.2235
[124] Boullier, D. (2016). Chapitre 2. Sociologie des usages. Dans : , D. Boullier, Sociologie du numérique (p. 99-129). Paris: Armand Colin.
[125] Timimi, I. (2020). *Op cit*.

Il est donc important d'arriver à tirer profit de la littérature disponible sur le sujet car elle permet de faire avancer la recherche sur l'évaluation de la RI et en hybridant les approches et les modèles.

Conclusion

Au travers de ce premier chapitre, nous avons ainsi mis en lumière l'évolution de la recherche d'information des années 1940 à nos jours pour en comprendre les enjeux et problématiques, la question de l'évaluation s'est alors révélée à la fois nécessaire et complexe, tant dans la théorie que dans la pratique. L'étude des protocoles d'évaluation existants nous a permis de nous constituer une boîte à outils pratique et conceptuelle afin de construire notre protocole d'évaluation, ce qui s'avère nécessaire car aucun protocole actuel ne correspond au cas d'étude sur lequel nous avons travaillé. La boîte à outils qui pourra être remobilisée dans la construction de nouveaux protocoles d'évaluation. Nous ferons le choix d'élaborer une évaluation à la croisée de l'évaluation dite de diagnostic et celle de mise en adéquation dans une approche dynamique, qualitative et pluridisciplinaire présentée dans le second chapitre de ce mémoire.

Chapitre 2 : Étude de cas Archival

Après avoir présenté le cadre théorique et conceptuel sur la recherche d'information et les possibilités d'évaluation, il convient désormais de présenter notre cas d'étude, en détailler les objectifs et fonctionnalités pour présenter l'élaboration de notre méthode d'évaluation du dispositif.

I/ Présentation du projet

Présenter ce projet implique de nous pencher sur son contexte de création et ses objectifs originels dont l'explicabilité du système et le dialogue entre les disciplines tiennent une bonne place.

1) Création du dispositif Archival

Archival est un projet ANR dédié à la valorisation d'archives multimédias assistée par intelligence artificielle. Porté par la chaire UNESCO-ITEN et soutenu par la FMSH et l'Université Paris 8, le projet réunit de nombreux partenaires dont Orange Labs, l'IRISA (l'Institut de Recherche en Informatique et Systèmes Aléatoires de l'Université Rennes 1) via l'équipe LINKMEDIA et le Lis-Lab (Laboratoire d'Informatique et Système de l'Université Aix-Marseille) à travers l'équipe TALEP (Traitement Automatique du Langage Ecrit et Parlé).

Le projet est né de la volonté de développer une interface documentaire dite « *intelligente* » car, selon l'équipe projet, si de nombreux travaux ont déjà été effectués sur les archives multimédia (textes, images, vidéos), les interfaces de recherche actuelles ne permettent ni la navigation et l'exploration approfondies des contenus, ni une mise en relation efficace entre les archives et d'autres sources externes :

> « *Le numérique transforme l'accès aux savoirs, les sources éditoriales et fonds d'archives se démultiplient sur les réseaux dans de multiples formats, pourtant ces savoirs sont encore difficilement accessibles auprès des publics, les résultats*

affichés sous forme de liste, l'agrégation de ces données multimodales de sources différentes est difficile, les interfaces peu attractives. »[126]

La plupart des interfaces de recherche documentaire actuelles s'apparenteraient, selon l'équipe projet, à des listes plates ce à quoi ils souhaitent proposer une alternative : *« Très souvent présentées sous forme de listes exclusivement fonctionnelles, les interfaces de résultats de recherche offrent peu de possibilités de navigation et d'exploration et articulent mal les contenus entre eux.* »[127]

Dans ce contexte, le projet Archival vise à développer de nouvelles interfaces de lecture, de consultation des documents, de médiation et de transmission des savoirs grâce à un travail en intelligence artificielle (compréhension automatique multimodale du langage). Deux questions de départ ont été formulées par l'équipe projet : Quel rôle peuvent jouer les méthodes de compréhension par les machines dans la réinterprétation de fonds d'archives thématiques ? Selon quelles modalités des interfaces de médiation des contenus peuvent-elles exploiter des résultats générés par les méthodes actuelles d'Intelligence Artificielle ?

Le projet consiste alors proposer de nouvelles modalités d'augmentation des corpus documentaires par la création de nouveaux parcours grâce à la génération automatique de liens au travers de deux algorithmes : l'un qui propose des liens par la génération de questions sur le texte[128] et l'autre qui agit par rapprochement de similarités sémantiques entre les extraits textuels[129]. Le postulat de ce projet étant que l'exploration renouvelée des fonds d'archives par la création de ces nouveaux parcours générés par IA est pertinente pour l'utilisateur notamment en termes de découvrabilité.

Le projet Archival est donc une preuve de concept (*proof of concept*) ayant pour objectif d'apporter des réponses sur la modélisation des pratiques intellectuelles de découverte d'un

---

[126] Agnola, M., Azemard, G., & Da Silva, S. (2022). Conférences internationales EUTIC 2022 : « A l'intersection de l'art, de la science et de la technologie : dialogues entre les hommes et les machines » [Actes de conférence]. Académie Ionienne de Corfu, Grèce, Octobre 13-14-15.

[127] Voir https://anr.fr/Projet-ANR-19-CE38-0011 consulté le 15/06/2023.

[128] Béchet F., Antoine E., Auguste J., Damnati G. (2022). Question Generation and Answering for exploring Digital Humanities collections. 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille, France ; Antoine E., Kang H. J., Rousseau I., Azémard G., Béchet F., et al. (2023). Exploring Social Sciences Archives with Explainable Document Linkage through Question Generation. 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, ACL SIGHUM, Dubrovnik, Croatia.

[129] Nguyen D. H., Gravier G., Sébillot P. (2022). Filtrage et régularisation pour améliorer la plausibilité des poids d'attention dans la tâche d'inférence en langue naturelle. TALN 2022 - Traitement Automatique des Langues Naturelles, Avignon, France, p.95-103 ; Nguyen D. H., Mallart C., Gravier G., Sébillot P. (2023). Regularization, Semi-supervision, and Supervision for a Plausible Attention-Based Explanation. NLDB 2023 - 28th International Conference on Natural Language and Information Systems, Derby, United Kingdom, p.1-14.

corpus documentaire. Pour le rendre exploitable et permettre la mise en place de journées d'expérimentation auprès de testeurs, un démonstrateur a été créé sous forme de site web[130]. Ce démonstrateur, également appelé dispositif, assemble différentes fonctionnalités détaillées ci-après dans une partie dédiée, l'ensemble permettant d'obtenir des avis et retours de testeurs sur le concept mis en œuvre. L'objectif de ce type de plateforme n'est pas de développer une interface documentaire achevée mais de développer de manière concrète les idées et hypothèses émises afin de les soumettre à des testeurs et d'en démontrer la faisabilité. Il ne s'agit donc pas d'une interface documentaire mise en production, ni d'un logiciel ou d'une plateforme gouvernementale mais de la démonstration concrète d'un projet.

Le choix du corpus

Le choix du corpus s'est porté sur la revue *Autogestion* parue en France entre 1966 et 1986. Disponible en intégralité sur Persée, elle se compose de 46 numéros. Entièrement dédiée au concept éponyme, la revue *Autogestion* suit l'ambition encyclopédique de ses créateurs Georges GURVITCH, Jean BANCAL et Daniel GUERIN : elle fait ainsi état des recherches théoriques et scientifiques sur ce thème tout en explorant les expériences concrètes menées partout à travers le monde (en particulier en Yougoslavie). Il s'agit d'une revue notamment connue pour son engagement comme l'explique Claudie WEILL[131] dans son article dédié : « *Bien que non partisane et s'inscrivant dans la mouvance universitaire, c'était une revue engagée* »[132] où la philosophie et l'histoire tiennent une grande place parmi d'autres disciplines telles que l'économie, l'anthropologie, l'éducation, le droit, la psychologie, la politologie ou encore la sociologie. L'apport de ces nombreuses disciplines en fait un corpus riche et dense.

L'autogestion est, en effet, « *une figure centrale du mouvement d'émancipation tout au long du XX^e^ siècle.* »[133] Elle pourrait être définie très simplement comme la « *gestion d'une collectivité par elle-même* »[134], comme un « *ensemble de pratiques alternatives au capitalisme* »[135] mais nous préférons employer la définition avancée par Valentin SCHAEPELYNCK, Engin SUSTAM dans leur article dédié à la question :

---

[130] Voir https://archival.msh-paris.fr/ consulté le 15/06/2023.
[131] Claudie WEILL a écrit de nombreux articles (une trentaine environ) pour la revue de 1977 à 1985.
[132] Weill, C. (1999). La revue Autogestion comme observatoire des mouvements d'émancipation. L'Homme et la société, 132(2), p. 29-36. https://doi.org/10.3406/homso.1999.3008
[133] *Ibid.*
[134] Article « Autogestion », dictionnaire Larousse [en ligne] consulté le 15/06/2023. URL : https://www.larousse.fr/dictionnaires/francais/autogestion/6714
[135] Christophe LE DIGOL, « Notre histoire intellectuelle et politique 1968-2018 (P. Rosanvallon) - Fiche de lecture », Encyclopædia Universalis [en ligne], consulté le 15 juin 2023. URL : http://www.universalis-edu.com.ressources-electroniques.univ-lille.fr/encyclopedie/notre-histoire-intellectuelle-et-politique-1968-2018/

> « *l'autogestion renvoie en droit à tout espace social, de travail ou d'activité, qui est gouverné directement par ses acteurs ou ses producteurs, qui en établissent et en instituent collectivement et directement les règles, les normes et les institutions, refusant toute hiérarchie verticale, toute division entre gouvernants et gouvernés, patrons et salariés, éducateurs et éduqués. Problématique transversale, autorisant une grande diversité de sens et d'appropriations, l'autogestion renvoie à un spectre d'expériences, passées ou contemporaines, très large et pluriel* »[136]

La pluridisciplinarité et la transversalité de la revue sont des critères qui ont joué dans le choix du corpus utilisé pour la création du dispositif Archival. Puisque le thème de l'autogestion couvre de nombreux domaines des SHS, l'utiliser dans la création d'un dispositif de lecture augmentée et une interface « *intelligente* » permet d'interroger des panels diversifiés, d'explorer différentes pratiques de lecture et des cas d'usage très disparates en fonction des disciplines. Un fonds d'autant plus intéressant qu'il a été labellisé CollEx (collection d'Excellence) dans le cadre de l'AMI CollEx-Persée.

Les ressources documentaires externes

Le dispositif Archival est donc basé, en grande majorité, sur la revue *Autogestion* qui constitue un corpus homogène et clos, mais il ne s'y cantonne pas grâce à l'ouverture vers de nombreuses ressources documentaires externes. Ces ressources sont diversifiées (textes, vidéos, sons, images, …) et proviennent des fonds documentaires de la BnF (Gallica et dataBnF), de la FMSH et de Wikidata mais également des ressources audiovisuelles de CanalU et de l'INA. Bien qu'une partie des liens n'aient pas encore été implémentés à ce jour, l'équipe Archival a réellement pensé le démonstrateur comme un ensemble d'interfaces documentaires multimodale permettant le rebond et le dialogue entre les ressources internes et externes, justifiant sa qualification d'interface « *augmentée* » par l'équipe projet.

L'intelligence artificielle tient alors une place de choix dans le projet car il s'agit de l'articulation principale permettant les liens et les rebonds entre les fonds. Le dispositif propose ainsi une lecture « *augmentée* » des corpus avec pour objectif d'apporter une valeur ajoutée à

[136] Schaepelynck, V. & Sustam, E. (2018). Autogestion. Le Télémaque, 54, p. 27-36. https://doi.org/10.3917/tele.054.0027

l'usager : « *Archival a pour objectif global de placer les ressources de l'IA au service de l'utilisateur* »[137].

Lancé en 2019 et débuté en 2020, le projet Archival est prévu pour une durée de 42 mois, le travail a donc été réparti selon un calendrier structuré en 5 WorkPackages (WP) de la coordination du projet et suivi scientifique jusqu'à l'évaluation finale. Le projet ANR touche désormais à sa fin et entre dans cette dernière phase, raison pour laquelle nous y avons été associés et que nous allons détailler dans la suite de ce mémoire.

2) Les enjeux de l'explicabilité des algorithmes

Les algorithmes sont donc au cœur du projet car ils permettent la mise en lien des contenus et constituent un type d'entrée « *intelligente* » dans ces derniers. Puisque le corpus est tourné vers les SHS, les usagers potentiels du dispositif sont des chercheurs des différentes disciplines, des documentalistes, des étudiants, etc. Le dispositif n'est donc pas destiné à des informaticiens ni à des spécialistes de l'intelligence artificielle. Le consortium Archival a donc à cœur d'expliquer le fonctionnement des algorithmes aux lecteurs qui seront amenés à les utiliser comme expliqué sur la page du projet sur le site de l'ANR :

> « *Les équipes d'ARCHIVAL s'interrogeront sur les manières de structurer une collection de documents hétérogènes en faisant apparaître de manière explicite les liens implicites, de révéler la nature de ces liens et de les valoriser de manière intelligible par la médiation d'interfaces ergonomiques qui garantissent une appropriation réussie des contenus.* »[138]

Cette démarche dite « *d'explicabilité* » vise donc à faire comprendre à l'usager comment l'algorithme procède au résultat, c'est-à-dire les logiques employées afin de permettre la mise en lien des différents contenus, qu'il s'agisse des similarités ou des questions générées. Il est vrai que la démarche et la façon de procéder d'un algorithme peuvent paraître très obscures pour un néophyte, rien que le terme « *algorithme* » peut être difficile à appréhender voire effrayer certains par manque de connaissances sur la question. En suivant la définition avancée par le Conseil national du numérique, on comprend qu'un algorithme est une séquence

[137] Voir https://archival.msh-paris.fr [en ligne] consulté le 26/06/2023.
[138] Voir https://anr.fr/Projet-ANR-19-CE38-0011 [en ligne] consulté le 15/06/2023.

d'instructions appliquées à un ensemble de données en vue de produire un résultat[139]. Et c'est la réalisation du résultat qui guide l'ensemble du processus comme l'explique Gilles ROUET :

> « *Les programmes, les implémentations, la mobilisation de matériel et dispositifs sont des instruments au service de l'objectif : obtenir un résultat, qui peut être de différentes natures, correspondre à des besoins ou à des intentions particulières.* »[140]

Bien qu'assez simples au regard des définitions proposées, les algorithmes sont en grande majorité développés pour être des boîtes noires (*black boxs*) qui permettent d'arriver à un résultat sans que le processus ne soit compris par l'usager. L'idée qui guide cette démarche est le fait que l'usager utilise le système pour les résultats qu'il produit et non pour en connaître le fonctionnement. Ils ont ainsi « *suscité de nombreuses critiques, en particulier par rapport à la transparence nécessaire de ce type de dispositif : quelles étaient les données utilisées et comment étaient-elles traitées pour l'obtention des résultats ?* »[141] ce qui a progressivement permis de renouveler le regard porté sur les algorithmes en s'intéressant non plus uniquement au résultat mais également au processus qui l'a engendré. Le projet Archival tente alors d'adopter une démarche dite de « *boîte transparente* » (*glass box*), à l'encontre de la complexité apparente du fonctionnement des DAI.

L'intérêt grandissant pour le fonctionnement des algorithmes et la façon dont les résultats sont générés amène donc à des réflexions sur la notion d'explicabilité, au cœur de la démarche *glass box*. Le but de l'équipe Archival est de rendre explicite la présence des algorithmes et d'en expliquer les possibilités, pour permettre à l'usager de l'utiliser en comprenant son fonctionnement. C'est donc une démarche qui vise à rationaliser le travail fait par l'algorithme, il n'y a pas un résultat isolé mais un processus compréhensible par le lecteur -du moins en théorie. Cela permet de comprendre les rapprochements qui sont opérés et juger de leur pertinence par exemple.

Il devient alors possible de repérer et comprendre les rapprochements proposés et éventuellement d'identifier les erreurs commises par le système et les biais de ce dernier : « *Les algorithmes naissent d'objectifs et doivent fournir des résultats. Constructions humaines, y compris quand il s'agit de dispositif susceptible d'auto-apprentissage, les algorithmes ne*

---

[139] Rouet, G. (2019). Démystifier les algorithmes. Hermès, La Revue, 85, p. 21-31. https://doi-org.ressources-electroniques.univ-lille.fr/10.3917/herm.085.0021
[140] *Ibid.*
[141] *Ibid.*

peuvent pas être neutres »[142] ce qui est particulièrement important pour l'usager surtout s'il est expert ou s'il a un objectif de recherche scientifique car ses exigences sont plus grandes que celles des autres usagers. La confiance placée par l'usager dans le système est primordiale, s'il en comprend le fonctionnement, il sera alors plus à même de l'utiliser :

> « *De plus, pour certaines de ces technologies, notamment celles qui mobilisent l'apprentissage profond, il existe ce qu'on appelle un phénomène de « boîte noire » : il est parfois difficile de comprendre et d'expliquer comment le système est arrivé à une suggestion ou à une décision. Or, ce manque d'explicabilité est problématique, car moins un système est transparent, moins les utilisateurs lui feront confiance et voudront l'adopter.* »[143]

L'explicabilité correspond alors à des enjeux éthiques majeurs, il est donc très intéressant de constater que le consortium Archival met cet aspect en lumière dans le projet.

3) Le dialogue entre informatique, SIC et SHS

Il existe aujourd'hui une nécessité : celle de faire dialoguer les progrès récents de l'informatique et notamment de l'intelligence artificielle avec les besoins documentaires et de recherche en SHS, c'est notamment ce qu'explique Nicolas SAURET :

> « *Depuis trente ans, les musées, les archives et les bibliothèques ont massivement numérisé leurs fonds, constituant d'immenses corpus numériques que les institutions cherchent progressivement à ouvrir aux chercheur·e·s et à leur public. Mais la numérisation n'est pas tout, car le plus gros du travail pour ces institutions consistent surtout à produire, pour chaque objet numérisé, les métadonnées, les descriptions et les transcriptions qui faciliteront leur véritable exploitation. Sur ce chantier, les méthodes de machine et de deep learning apparaissent particulièrement appropriées.* » [144]

Le croisement entre les disciplines et les domaines semble alors évident mais s'avère en réalité être bien plus complexe qu'il n'y paraît. La pluridisciplinarité pose, en effet, de

---

[142] *Ibid.*
[143] Martineau, J. (2023). Transition numérique et intelligence artificielle : d'importants enjeux éthiques à surveiller. Gestion, 48, p. 60-64. https://doi-org.ressources-electroniques.univ-lille.fr/10.3917/riges.481.0060
[144] Sauret, N. (2022). Intelligence artificielle & Sciences humaines et sociales (SHS) : opportunités, défis et perspectives. I2D - Information, données & documents, 1, p. 97-103. https://doi-org.ressources-electroniques.univ-lille.fr/10.3917/i2d.221.0097

nombreuses questions à la fois théoriques et pratiques en particulier sur la place et le rôle de l'intelligence artificielle proposée par les informaticiens vis-à-vis de l'intelligence humaine des chercheurs en SHS. C'est notamment ce qu'avancent Annette CASAGRANDE et Laurent VUILLON dans leur article « *Sciences humaines et sociales et méthodes du numérique, un mariage heureux ?* », ils écrivent :

> « *Les méthodes du numérique ne doivent pas s'imposer aux SHS mais se doivent d'être à l'écoute de leur problématique afin de leur proposer des solutions adaptées. Chaque champ disciplinaire a ses codes, son univers sémantique, ses méthodes...* »[145]

L'équipe Archival apporte une double réponse à cet impératif. D'abord, elle annonce subordonner l'intelligence artificielle à l'intelligence humaine et à l'expérience de l'utilisateur « *Le programme Archival a pour objectif global de placer les ressources de l'IA au service de l'utilisateur* […] *Cette approche permet "d'augmenter" l'expérience utilisateur* » cantonnant alors l'IA à un rôle d'auxiliaire dans la recherche et la découverte des corpus. Ensuite, l'équipe met en avant sa diversité grâce à la pluridisciplinarité de ses membres :

> « *Pluridisciplinaire et multi-acteurs, le projet vise à faire collaborer des chercheurs issus des Sciences de l'Information et de la Communication et de l'Informatique autour de la valorisation des archives et du partage des savoirs pour les arts, la culture et le patrimoine.* »[146]

Cela permet de contrer l'argument avancé par Thierry FOUCART :

> « *il ne suffit pas d'être mathématicien, informaticien, médecin, gestionnaire, sociologue ou psychologue pour mettre en œuvre une méthode statistique et en analyser correctement les résultats. Il ne suffit pas non plus d'être statisticien : il faut posséder des compétences multiples, ou travailler en équipe pluridisciplinaire.* »[147]

Le consortium Archival est pourtant confronté à des difficultés qui découlent de sa pluridisciplinarité. Le vocabulaire, les centres d'intérêt et les méthodes de travail, tout semble

---

[145] Casagrande, A. & Vuillon, L. (2017). Sciences humaines et sociales et méthodes du numérique, un mariage heureux ?. Les Cahiers du numérique, 13, p. 115-136. https://www-cairn-info.ressources-electroniques.univ-lille.fr/revue--2017-3-page-115.htm.

[146] Voir https://www.fmsh.fr/actualites/archival consulté le 15/06/2023.

[147] Foucart T. (2001). L'interprétation des résultats statistiques. Mathématiques et sciences humaines, 153. DOI : https://doi.org/10.4000/msh.2840

opposer les membres de l'équipe ce qui n'est pas étonnant au regard des différences majeures qui existent entre les domaines : « *La grande difficulté d'un travail scientifique pluridisciplinaire réside dans la compréhension mutuelle des chercheurs de disciplines différentes.* »[148] Leurs attentes et objectifs divergents souvent et il peut s'avérer difficile de trouver un juste milieu comme l'explique un des membres de l'équipe :

> « *On est hybride non pas par choix, on l'a été un peu par défaut puisqu'il y avait quand même l'ambition de faire une interface interactive et intéressante qui intègre des fonctionnalités traditionnelles et en même temps de valoriser le travail pour lequel vraiment l'ANR nous avait mandaté surtout les collègues informaticiens* » [voir annexe n°49].

Un autre confie d'ailleurs en entretien se sentir pris entre deux feux et ajoute : « *il y a des dialogues qui ne sont pas faciles à établir donc on a réussi à communiquer mais peut-être pas jusqu'au bout…* ». Les membres perçoivent alors ce phénomène à la fois comme une contrainte quotidienne pour s'entendre et se comprendre mais comme une force sur le long terme par la pluridisciplinarité et le croisement des méthodes et paradigmes. Un phénomène qui n'est donc pas anecdotique car il entraîne des répercussions sur le développement du projet et sur l'évaluation que nous menons.

II/ Mise en place de l'expérimentation

Il convient désormais d'entrer dans le cœur de notre travail, à savoir la réalisation de journées d'expérimentation afin de faire tester le dispositif à des usagers potentiels de ce type de dispositif et d'en retirer des matériaux pour notre évaluation.

1) Réalisation de journées d'expérimentation

Deux journées d'expérimentation ont été réalisées le 15 mai 2023 et le 09 juin 2023 permettant la mise en place de trois sessions de test distinctes. Nous allons présenter séparément bien qu'il y ait de nombreuses similitudes car il y a eu quelques ajustements entre la première et la deuxième journée.

[148] Casagrande, A. & Vuillon, L. (2017). *Op cit.*

L’atelier du 15 mai 2023

La première journée de test s’est tenue au BnF DataLab, en salle X de la bibliothèque de recherche de la BnF (site François-Mitterrand, Paris). Seize testeurs ont été réunis pour cet atelier de découverte et de prise en main de la plateforme Archival. La journée s'est déroulée en deux temps : une première session de 2h30 le matin avec dix testeurs, principalement des chercheurs en sciences de l'information et de la communication, suivie d'une seconde session de 2h30 l'après-midi, regroupant six professionnels des bibliothèques et de la documentation.

Les testeurs ont eu l'occasion de prendre en main la plateforme individuellement dans les boxs situés en mezzanine du BnF DataLab le matin et dans une salle de formation du service l’après-midi. Deux parcours leur ont été proposés : le premier était entièrement libre, tandis que le second était « *guidé* » à partir d'articles d'entrée suggérés. À la fin de chaque parcours, les testeurs ont été invités à remplir un questionnaire en ligne et au terme de chaque session, les participants ont été réunis pour discuter et échanger collectivement sur le dispositif, poser des questions, émettre des remarques et exprimer leur ressenti général. Tout au long de l’expérimentation, le parcours des testeurs est enregistré grâce aux identifiants anonymisés distribués à chacun, l’enregistrement de leur activité dans le démonstrateur permet de collecter les traces ou *logs* de connexion de chacun afin de les analyser.

L’atelier du 09 juin 2023

La seconde journée s’est déroulée à l’Université de Lille, dans une salle informatique du département Sciences de l’information et de la documentation (SID). Quatre testeurs ont été invités : deux enseignants-chercheurs en SIC, une doctorante et professeure documentaliste et une documentaliste du Service Commun de la Documentation (SCD) de l’Université. La durée de l’expérimentation a été rallongée à 3h30 pour ce deuxième atelier au regard du manque de temps constaté lors de la première journée d’expérimentation.

Les testeurs ont suivi le même parcours que ceux du 15 mai 2023 à ceci près qu’ils ont eu plus de temps pour chaque étape.

Objectifs des sessions de test

Les journées d’expérimentation s’inscrivent dans le WP5, annoncé par le calendrier du projet Archival, dédié à l’expérimentation et l’évaluation du travail réalisé. Ces journées font donc partie du travail d’évaluation réalisé par l’équipe projet, leurs objectifs sont multiples mais

convergents : évaluer le système et la génération automatique de liens et évaluer l'interface et le parcours utilisateur.

La chaire UNESCO-ITEN et la FMSH souhaitent participer « *selon des méthodes agiles, à l'évaluation de l'appropriation des outils par les usagers* »[149] et se placent donc du côté de la réception du dispositif par l'usager. Pour les informaticiens de l'équipe, les objectifs se situent plus au niveau de la compréhension du dispositif et de son fonctionnement par l'usager, donc au niveau de l'explicabilité. L'équipe a alors annoncé en réunion vouloir obtenir deux feedbacks pendant les phases de test : un sur le jugement de pertinence et l'intérêt des liens générés par les usagers, un autre sur la compréhension et la réception des liens par les usagers.

Poursuites

L'équipe Archival -en particulier la chaire UNESCO-ITEN et la FMSH- envisagent de poursuivre les journées de test par des entretiens individuels avec les expérimentateurs des différentes sessions. Ces entretiens individuels seraient bien plus poussés que ce qui a pu être dit lors des entretiens collectifs et, d'autre part, plus qualitatifs permettant à chaque interrogé d'aborder en détails les éléments les plus importants selon lui ou qui l'ont le plus touché tout en offrant la possibilité de questionner chacun sur son domaine d'expertise au regard du dispositif proposé. Par ailleurs, les laboratoires d'informatique s'attachent à traiter et à analyser les logs de connexion des différents testeurs afin de déterminer leurs parcours, les interfaces consultées et celles qui le sont moins, l'usage des algorithmes fait par les différents testeurs, etc.

2) Définition des panels de testeurs

Sans prendre en compte les différentes sessions d'expérimentation, deux panels de testeurs distincts ont été invités lors des ateliers. On distingue alors d'une part des chercheurs en SIC (maîtres de conférences, professeurs et professeurs émérites) et d'autre part des professionnels des bibliothèques et de la documentation (conservateurs, conservateurs en chef, chefs de projets et un étudiant).

Les panels invités à participer aux sessions d'expérimentation ont alors été qualifiés d'« *experts* », cependant, la question de l'expertise est complexe et mérite d'être plus

[149] Voir https://anr.fr/Projet-ANR-19-CE38-0011 consulté le 15/06/2023.

amplement expliquée. Rien que la définition du terme semble créer de nombreux débats ; du latin *expertus*, l'expert peut être défini comme celui qui a éprouvé, qui est exercé et connaît par la pratique. C'est notamment ce qu'explique Corinne DELMAS dans *Sociologie de l'expertise* : « *À l'origine, l'expert est le détenteur d'un savoir particulier, lié à la pratique de son métier ; il devient un spécialiste reconnu dans son domaine, sollicité pour émettre un avis.* » [150] elle ajoute « *La notion d' "expertise" est souvent perçue comme synonyme de "compétence", notamment professionnelle* ». L'expert ne l'est donc que dans un domaine, au regard d'une ou plusieurs compétences, on est expert de quelque chose et non expert tout simplement. Un phénomène d'autant plus important que la notion d'expert est « *utilisée avec une banalité déconcertante* »[151].

Pourtant, lors des journées d'expérimentation et en particulier de la première (le 15 mai 2023), les testeurs invités ont été qualifiés d'« *experts* » sans plus de précision ni sur la définition du terme, ni sur le domaine et les compétences précises de chacun. Cela engendre alors un flou relatif vis-à-vis du panel interrogé pris comme un tout alors que la logique voudrait que chacun soit expert de son domaine et que ce soit l'expertise précise de chacun qui soit interrogée -et non une expertise générale et informe. L'expertise, pouvant être qualifiée d'« *examen de quelque chose en vue de son estimation, de son évaluation, etc.* »[152], devrait alors être sollicitée au regard de l'individualité et des compétences de chacun. Un point important qui se retrouve dans l'étymologie et l'évolution du terme :

> « *Le substantif "expert", qui serait apparu en 1580 chez Montaigne[153], est surtout un terme juridique qui renvoie dès le XVIII^e^ siècle à la notion de 'personne choisie pour ses connaissances techniques et chargée de faire, en vue de la solution d'un procès, des examens, constatations ou appréciations de faits*'. *Dès 1807, le verbe apparaît avec le sens de "soumettre à une expertise", "apprécier, estimer, évaluer".* »[154]

De plus, il existe des degrés dans l'expertise, il n'y a pas de rupture brutale entre l'expert et le novice mais plutôt une gradation complexe qui mériterait également qu'on y porte une attention toute particulière lors de la création des panels de testeurs. C'est notamment ce qui est

[150] Delmas, C. (2011). Introduction. Dans : Corinne Delmas éd., Sociologie politique de l'expertise (p. 3-8). Paris: La Découverte.
[151] Chaudat, P. & Pierre Mirralès, P. (2009). L'évaluation des experts dans les organisations. Revue Interventions économiques, 39. DOI : https://doi.org/10.4000/interventionseconomiques.195
[152] Article « Expertise », dictionnaire Larousse [en ligne] consulté le 16/06/2023.
[153] Montaigne. (1980) Trésor de la langue française, vol. 8, p. 472.
[154] Delmas, C. (2011). *Op cit.*

démontré par dans l'article « *Expertise dans le domaine et expertise dans Internet : leurs effets sur la recherche d'informations* » :

> « *Plusieurs études empiriques mirent en évidence que les stratégies de recherche et de navigation sont soumises à des variabilités inter-individuelles. Ainsi beaucoup d'auteurs ont constaté que l'expertise domaine est un facteur important pour la performance de recherche d'information*. »[155]

Nous avons identifié trois types d'expertise au sein des différents panels : une expertise du domaine (sur la recherche d'information et les interfaces documentaires), une expertise fonctionnelle (sur les interfaces web, l'UX, le développement web) et une expertise du sujet (sur le fond et le corpus proposé).

Enfin, l'équipe Archival considérait qu'il n'était pas opportun d'observer les testeurs au cours de l'expérimentation pour ne pas les influencer dans leur navigation, les testeurs de la première session étant qualifiés d' « *experts* » -à la fois du domaine et des aspects fonctionnels, il a été suggéré de nous en remettre à leur expertise. Nous n'avons donc pas effectué d'observation ciblée le matin mais avons observé de façon distante et discrète pour ne pas les influencer. La posture à adopter face à l'expert n'est pas simple à trouver et le débat sur ce sujet reste entier comme l'explique : « *Faut-il expertiser les experts ? Cette question en forme de boutade pourrait résumer la place paradoxale de l'expertise dans les sociétés contemporaines*. »[156] Au cours de la matinée, nous avons pu observer qu'un accompagnement dans la découverte et la prise en main du démonstrateur serait bienvenu pour une partie des testeurs, nous avons donc fait évoluer la méthode entre les deux premières sessions pour ajouter l'observation ciblée l'après-midi du 15 mai 2023. Les observations ciblées menées auprès des testeurs au cours de la deuxième et de la troisième session ont obtenu des résultats probants et ont révélé l'intérêt de cette méthode.

Au regard de ces explications, nous pouvons affirmer que « *Le problème de l'expertise est durable* »[157] et qu'il mériterait un travail de fond plus poussé sur la relation entre expert et évaluation. Un travail qui pourrait s'appuyer entre autres sur les travaux de Y. BERARD et R.

[155] Ihadjadene M. & Martins D. (2004). Expertise dans le domaine et expertise dans Internet : leurs effets sur la recherche d'informations. Paris, Hermès, 39, p. 133-142.
[156] Delmas, C. (2011). *Op cit.*
[157] Demortain, D. (2021). Experts scientifiques et action publique : paradoxe et perspectives de recherche pour la sociologie politique de l'expertise: Commentaire, 39, p. 33-41.

CRESPIN[158], P. B. JOLY[159], L. MAXIM et G. ARNOLD[160], A. ROGER et P. ROGER[161], ou encore J. Y. TREPOS[162].

3) Les fonctionnalités du dispositif

Pour mieux comprendre le dispositif et la suite des explications, il convient désormais de présenter en détails les différentes interfaces, leurs fonctionnalités et les liens entre elles. Pour permettre à nos lecteurs de suivre cette présentation, nous les invitons à consulter les annexes de ce mémoire en parallèle de leur lecture car le démonstrateur n'est accessible qu'avec des identifiants de connexion[163].

Page d'accueil

La page d'accueil du site présente le projet ANR, ses différents aspects et enjeux [voir illustrations n°3 et 3 bis], il s'agit d'une page informative. Dans le coin supérieur droit, l'encart « *Se connecter* » permet d'entrer ses identifiants personnels et d'accéder au démonstrateur. La consultation des fonds et l'usage du dispositif n'est pas permis sans identification préalable, elle n'est donc pas ouverte au grand public actuellement.

Une fois connecté, une nouvelle page d'accueil apparaît à l'écran, beaucoup plus simple et vide que la première, elle compte : le logo Archival, le titre de la revue, ses dates, une petite synthèse à côté de laquelle se trouve un bref texte explicatif sur le projet [voir illustration n°4]. Dans le coin supérieur droit, un moteur de recherche apparaît à côté des identifiants du testeur. Dans le coin supérieur opposé se trouve un menu latéral composé de trois sections et huit entrées : Découvrir les corpus (La revue autogestion, Vidéothèque), Explorer les contenus (Notions, Personnages, À travers le monde, Co-citations), À propos (Projet ANR Archival, Algorithmes utilisés) [voir illustration n°5].

[158] Bérard, Y. & Crespin, R. (dir.) (2015). Aux frontières de l'expertise : dialogues entre savoirs et pouvoirs. Rennes, Presses universitaires de Rennes.
[159] Joly, P.B. (2012). La fabrique de l'expertise scientifique. Hermès, 64, p. 22-28.
[160] Maxim, L. & Arnold, G. (2012). Entre recherche académique et expertise scientifique : des mondes de chercheurs. Hermès, La Revue, 64, p. 9-13.
[161] Roger, A. & Roger, P. (2001). Rôle et place des experts dans une société de l'information. Actes du XIIe Congrès de l'AGRH.
[162] Trepos J.Y. (1996), La sociologie de l'expertise, Presses Universitaires de France.
[163] Le lien du dispositif : https://archival.msh-paris.fr/ [en ligne].

Les différentes fonctionnalités

Le dispositif propose deux grands types d'entrée dans le corpus : des entrées dites « *classiques* » en recherche documentaire (le moteur de recherche par mots-clés, l'indexation des personnes et des notions présentes la revue), ainsi que des entrées « *intelligentes* » (permises par l'IA grâce à deux algorithmes de génération automatique de liens, proposant des questions sur le texte et des rapprochements d'extraits basés sur la similarité des contenus).

Le moteur de recherche est accessible directement depuis la page d'accueil et le reste tout au long de la navigation. Assez basique, il permet d'entrer des mots-clés mais ne permet pas de recherche avancée. Les résultats de recherche apparaissent sur la page d'accueil et sont présentés sous forme de longues listes où chaque lien est cliquable [voir illustration n°6]. Les résultats sont décrits par le titre et l'auteur de l'article ainsi que par la phrase ou les quelques mots où apparaissent les termes recherchés. Juste au-dessous, en grisé très clair, on peut lire l'identifiant donné par la FMSH et à droite, souligné « voir l'article » pour accéder directement à la page où le(s) mot(s)-clé(s) a/ont été trouvé(s). Les liens issus d'un même article sont présentés à la suite les uns des autres par ordre d'apparition dans le texte, les liens issus de différents articles sont séparés par des barres typographiques noires de la largeur de l'écran.

Dans le menu latéral, la première entrée proposée et celle de la revue : elle ouvre une page composée d'une brève présentation de la revue *Autogestion* suivie d'une frise chronologique de l'ensemble des numéros parus [voir illustrations n°7 et 7 bis]. La frise est cliquable et affiche au-dessous d'elle le(s) numéro(s) de l'année correspondant(s) à la date sélectionnée avec la photo de couverture du numéro, le sommaire et les articles qui le composent également cliquables (parfois, il y a également une bibliographie). Il n'est pas possible de consulter l'ensemble d'un numéro en un clic, il faut sélectionner article par article. Il est également possible d'accéder aux différents numéros directement grâce à la barre de défilement au-dessous de la frise.

L'entrée suivante nommée « *Vidéothèque* » permet de visionner diverses vidéos en lien avec l'autogestion [voir illustration n°8] allant de la propagande boulangiste à la grève des LIP en passant par des reportages sur l'économie sociale. Chaque vignette est cliquable et ouvre une fiche composée du titre de la vidéo, du lecteur vidéo et des mots-clés qui lui sont associés cliquables également [voir illustration n°9].

L'usager a ensuite la possibilité d'entrer dans le corpus via la page notions dédiée aux termes de la revue indexés dans le thésaurus du projet par les chercheurs de la chaire UNESCO-

ITEN et de la FMSH [voir illustration n°10]. Près de 350 termes y sont recensés et classés par nombre d'occurrences, le terme le plus cité étant « *société* » avec 136 occurrences, le moins cité est « *taylorisme* » avec une seule apparition dans la revue (à égalité avec d'autres termes cités une seule fois comme « *républicains* »). Les termes sont affichés via deux graphiques : un histogramme circulaire puis un diagramme linéaire en bâtons, les notions sont cliquables sur le premier mais pas sur le second.

Les notions indexées dans le thésaurus ont été réparties en quatre catégories (« *Organisations sociales* », « *Dynamiques collectives* », « *Modèle économique* » et « *Modèle politique* ») par les chercheurs de la chaire UNESCO-ITEN et de la FMSH et une teinte de vert leur a été associée à chacune. Sur la droite du diagramme circulaire, deux fonctions avancées permettent de modifier les paramètres d'affichage du premier graphique. D'abord, les quatre catégories qui constituent la légende sont cliquables grâce au carré de couleur permettant en sélectionnant ou désélectionnant les catégories de les afficher ou non. La seconde fonction avancée se situe juste au-dessous de la légende et permet de faire apparaître les termes en fonction de leur nombre d'occurrences. Il est alors possible de changer les termes qui apparaissent à l'écran en faisant glisser les crochets sur le petit histogramme. Pour générer le nouveau graphique, il faut cliquer sur « *actualiser* » [voir illustrations n°10 et 11]. Au-dessous du diagramme circulaire, se trouve un diagramme en bâtons qui recense également tous les termes du thésaurus par ordre décroissant. Il interagit avec la légende comme le premier graphique mais pas avec l'histogramme à crochets permettant de sélectionner les termes par nombre d'occurrences.

En cliquant sur un terme, une nouvelle fenêtre apparaît à l'écran avec une définition, des liens externes au site (vers d'autres interfaces documentaires) et la liste des articles de la revue dans lesquels il est cité, tous cliquables également [voir illustration n°12]. Le bandeau d'en-tête de la notion est de la couleur de la catégorie à laquelle elle est rattachée dans le thésaurus (vert très sombre : organisations sociales, vert foncé : dynamiques collectives, vert pomme : modèle économique et vert clair : modèle politique).

Le dispositif propose ensuite d'explorer le corpus par personnages. Cette page recense l'ensemble de noms propres cités dans la revue, qu'il s'agisse de personnes citées dans le corps du texte ou des auteurs des différents articles [voir illustration n°13]. Ils sont classés par ordre alphabétique par nom de famille ou surnom (ex : Montesquieu), une liste alphabétique verticale sur la droite de l'écran permet d'accéder directement à la lettre souhaitée. Chaque personnage dispose d'une vignette de présentation avec son nom en bandeau, une photo ou un portrait issu

de Wikipédia lorsque c'est possible, ses identifiants JSON (Persée, idref, BnF, isni, viaf, Wikidata, orcid, bdpedia, etc), une brève mention de ce pourquoi il est connu -indépendamment de la revue- et le nombre de fois où il est auteur et/ou cité dans l'ensemble du corpus. En cliquant sur la vignette du personnage souhaité, une fiche qui répertorie les informations à son sujet s'ouvre [voir illustration n°14]. Elle contient : la photo ou le portrait, le nom, la fonction ou les hauts-faits connus, la liste des articles où il est cité puis celle des articles dont il est auteur et enfin la bibliographie BnF associée à la personne. La liste des liens est cliquable et renvoie directement vers lesdits articles et vers le site de la BnF [voir illustration n°15].

Le menu latéral proposera par la suite deux autres types d'entrées dans le corpus, à savoir « *À travers le monde* » et « *Co-citations* » mais il nous est actuellement impossible d'en parler, ces entrées devraient cependant arriver dans les prochains mois (d'ici décembre 2023).

Dans la section suivante nommée « *À propos* », il ne s'agit plus de fonctionnalités d'entrées mais de pages informatives à destination du lecteur. La page intitulée « *Projet ANR Archival* » renvoie vers la première page d'accueil du site, avant la connexion grâce aux identifiants personnels, présentée précédemment, page sur laquelle il peut revenir à tout moment en cliquant sur le logo du projet en haut de chaque page. La seconde nommée « *Algorithmes utilisées* » est une page d'explications sur les algorithmes présents dans Archival et leur fonctionnement [voir illustration n°16]. Deux longs textes explicatifs rédigés par les informaticiens du projet présentent donc l'algorithme de génération de questions et celui qui opère des rapprochements par similarités.

Lire un article

Une fois un article sélectionné, il s'ouvre en pleine page dans un lecteur et sera, par défaut, en mode image [voir illustration n°17]. Pour activer le mode texte et ainsi pouvoir sélectionner du texte, effectuer un copier-coller ou autre, il faut cliquer sur le texte. Cette modification ne s'applique qu'à la page sélectionnée et il n'est possible de le faire que page par page. Des changements de mise en forme sont à prévoir lorsque la page est affichée en mode texte (texte brut sans mise en forme) surtout pour les pages de titres, de bibliographie, etc. Il est possible d'annuler cette action et retrouver le format image en cliquant en bas de la page affichée sur « *voir l'image* ». Une fois en mode texte, il est possible de sélectionner une portion de texte (une phrase minimum, la page au maximum, impossible de sélectionner du texte sur plusieurs pages en même temps) [voir illustration n°18]. Ce texte apparaît alors surligné en jaune et une nouvelle fenêtre nommée « *Algorithmes de génération de liens* » s'ouvre sur le

côté droit du lecteur. Elle propose l'exploration du texte surligné par le biais des deux algorithmes : l'exploration par questions et l'algorithme de similarités. Chaque algorithme est accompagné d'un bref descriptif, pour en savoir plus, le lecteur peut se reporter à la page « *Algorithmes utilisés* ».

Le bandeau d'en-tête bleu indique le mois et l'année de parution du numéro consulté (mais pas son numéro), le titre de l'article sélectionné ainsi que son auteur, il permet d'accéder à trois fonctionnalités : lire l'article, personnages cités et notions liées. Lire l'article est la fonctionnalité qui s'ouvre automatiquement lorsque l'on clique sur l'article, personnages cités permet d'accéder à la galerie de personnages cités dans l'article (auteurs et personnages historiques) [voir illustration n°19]. Chacun possède une vignette si possible avec une photo ou un portrait, son nom et en cliquant dessus, l'usager ouvre la fiche personnage correspondante. La fonctionnalité suivante « *Notions liées* » ouvre deux listes de mots-clés cités dans l'article ; celle de gauche est constituée des termes entrés dans le thésaurus du projet, ils sont tous cliquables et ouvrent chacun la vignette liée à la notion correspondante. La seconde liste, à droite, recense les mots-clés identifiés par l'algorithme dans le texte, il s'agit d'une liste ouverte et non contrôlée à l'inverse de la précédente. Tous les termes sont suivis d'un nombre qui correspond au score algorithmique attribué à chacun par la machine permettant d'estimer l'importance des termes les uns par rapport aux autres.

Le système de bureau et de fiches

Le dispositif Archival est basé sur un système de bureau et de navigation par fiches, système qui nécessite d'être expliqué car il recèle quelques difficultés. Lorsque l'usager navigue dans le démonstrateur, il peut le faire à deux niveaux : dans les pages d'entrée présentées successivement ci-dessus ou dans l'espace de bureau grâce à la navigation par fiches. Les fiches peuvent être de quatre natures différentes : les articles, les personnages, les notions et les fiches audiovisuelles. Une fois entré dans le corpus, l'interface première passe au second plan, le fond de la page s'assombrit pour devenir gris foncé et les fiches s'ouvrent successivement sur ce bureau au premier plan [voir illustration n°20]. Les fiches s'ouvrent alors successivement de droite à gauche (les plus récentes sont à gauches et les plus anciennes à droite). Il est possible de les réduire, les fermer ou les agrandir. Pour retourner dans l'interface de départ et quitter le bureau, il faut cliquer sur « *masquer les fiches* » dans le coin supérieur droit de l'écran, pour rouvrir le bureau (et toutes les fiches qu'il contient) il faut faire l'inverse et cliquer sur « *voir les fiches* ». Les fiches se positionnent automatiquement mais il est possible de les déplacer manuellement [voir illustration n°21].

III/ Élaboration d'une méthode d'évaluation

Nous pouvons désormais présenter l'élaboration de notre méthode d'évaluation tirée en partie de la littérature, de la nécessité de mettre en place une démarche composite et pluridisciplinaire.

1) Usage de la littérature

Afin de mettre en place notre méthode pour évaluer Archival, nous avons donc puisé dans la littérature scientifique des différents domaines cités précédemment en adoptant une démarche agile et pluridisciplinaire.

Le type d'évaluation choisi

Nous avons ainsi eu le choix entre plusieurs possibilités dans la construction de cette méthode notamment entre évaluation de progression, de diagnostic ou de mise en adéquation. C'est la dernière évaluation qui a été retenue pour ce projet car l'évaluation de mise en adéquation est celle qui répond le mieux aux attentes et questions formulées initialement par l'équipe Archival. Le consortium a, en effet, sollicité notre aide pour évaluer les usages du dispositif, son appropriation par les usagers et comprendre la mise en adéquation entre le dispositif proposé et les attentes et habitudes des utilisateurs.

L'évaluation dite de « *diagnostic* » a également été réalisée mais par l'équipe Archival directement, comme cela a été mentionné plus tôt, ce type d'évaluation est souvent réalisée par les concepteurs afin de mesurer les performances d'un système (ce qu'ils ont fait avec les deux algorithmes) et ainsi en pallier les erreurs. Ce sont donc les informaticiens du consortium qui ont réalisé ce travail, nous n'avons pas pris part à ce pan de l'évaluation.

Il aurait également été possible et très certainement intéressant de réaliser une évaluation de performance de type horizontale (ou d'appariement) en comparant Archival avec un autre système similaire. Dans ce cas, nous avions envisagé de réaliser cette étude comparée entre le dispositif Archival d'un côté et Persée de l'autre, puisque la revue y est intégralement numérisée. Cette évaluation comparative aurait permis de mettre en lumière les apports permis par Archival en comparaison d'un DAI plus traditionnel et quotidiennement utilisé par en SHS. Nous avons évincé cette piste nous éloignant trop des objectifs fixés par l'équipe projet pour privilégier une autre approche mais elle semble pertinente également.

Archival : une « *boîte transparente* » ?

Par ailleurs, l'explicabilité souhaitée par l'équipe projet au sujet du fonctionnement des algorithmes inscrit l'évaluation dans une démarche dite de « *boîte transparente* » (*glass box*), les algorithmes expliquent leur fonctionnement et le processus qui leur a permis d'arriver au résultat. Une démarche majeure au sein du projet et qui a fait l'objet de questions nombreuses et variées auprès des publics interrogés.

Une évaluation *de facto* dynamique

Archival étant encore en développement lors de l'évaluation, nous avons de fait eu recours à une évaluation dynamique. Nous aurions pu faire le choix de réaliser une évaluation statique à date de la première session d'expérimentation sans prise en compte des améliorations et modifications apportées après, cela nous aurait cependant empêché de réaliser une seconde journée de test et d'affiner notre méthode d'enquête et d'évaluation. Nous avons alors adopté une démarche agile permettant de faire évoluer la méthode au cours de nos expérimentations en vue de l'améliorer et d'en combler les lacunes. Cette évaluation dynamique a donc permis de mettre en lumière les apports et améliorations effectuées entre les deux journées de test tout en étudiant problèmes restants.

L'évaluation quali-quantitative

L'évaluation peut ainsi être quantitative, qualitative ou hybridée. Dans le cas d'Archival, deux évaluations ont lieu en parallèle : une évaluation quantitative réalisée par les informaticiens du projet pour mesurer la performance des algorithmes et une évaluation qualitative de notre côté pour déterminer la satisfaction des usagers, leur compréhension du système et l'appropriation qu'ils s'en font. De notre côté, nous réalisons ainsi une démarche qualitative et humaine. L'évaluation que nous avons mise en place est donc uniquement qualitative mais elle s'inscrit dans une démarche plus globale initiée par l'équipe Archival dans le cadre du WP5 du projet.

Nous espérons cependant pouvoir hybrider les résultats obtenus par ces deux démarches notamment en ce qui concerne l'analyse des *logs* des différents testeurs par exemple, et il serait intéressant de pouvoir confronter cette analyse des logs avec l'enquête. L'analyse des traces ou *logs* sur les SRI web permet d'adapter le système aux besoins réels des usagers, à leurs pratiques. Une hybridation qui serait tout à fait pertinente puisque, d'un côté les laboratoires d'informatique s'occupent du web analytique, à savoir analyser quantitativement les parcours

des testeurs, leurs traces (les clics, le temps passé sur les pages, le nombre de pages consultées, etc) tandis que nous avons les traces plus humaines grâce à l'observation mise en place lors des sessions de test pouvant ainsi apporter du contexte et des précisions aux résultats quantitatifs obtenus.

2) Mise en place d'une méthode composite

Pour tirer profit des journées d'expérimentation et permettre de baser notre évaluation sur les retours réels des testeurs, une enquête a été réalisée auprès d'eux. Réalisée en collaboration entre la chaire UNESCO-ITEN, la FMSH, la BnF et le laboratoire GERiiCO, l'enquête est constituée de trois types de matériaux : les observations réalisées par l'équipe Archival et associés, les questionnaires complétés par les testeurs et les transcriptions des entretiens collectifs à partir des enregistrements. Tous les matériaux bruts sont présentés en annexes de ce mémoire.

L'observation

Nous avons distingué et pratiqué deux types d'observation différents au cours de la journée : l'observation ciblée et l'observation générale. La première consiste à placer un observateur à proximité d'un testeur dans le box ou assis à côté de lui. Le testeur expérimente le dispositif pendant que l'observateur note le parcours emprunté, les remarques et commentaires éventuels ainsi que les émotions et blocages ressentis au cours de la navigation. Noter le parcours du testeur permet de comprendre le cheminement suivi par ce dernier, ce qui complète et affine l'analyse des logs réalisée par l'équipe Archival. Le point central de ce type observation est alors la captation des émotions ressenties et des blocages rencontrés car cela permet d'évaluer qualitativement l'expérimentation du testeur et d'expliquer d'un point de vue humain le parcours enregistré automatiquement dans les logs. Pour faciliter ce travail, nous avons réalisé en amont une grille d'observation uniforme distribuée à tous les observateurs [voir illustration n°22], uniquement utilisée le 15 mai 2023 car pour le 9 juin nous avons simplement utilisé des feuilles blanches en format A4, jugées plus pratiques pour noter rapidement. Toutes les fiches d'observation récoltées sont intégrées aux annexes [de l'annexe n°23 à 25].

La seconde observation, dite « *observation générale* » ou « *macro* », consistait plutôt à adopter un point de vue global sur l'expérimentation : ce qui fonctionne ou qui fonctionne moins, les réussites et les points à améliorer, etc. Ce type d'observation ne se concentre donc pas sur un ou plusieurs testeurs mais sur le dispositif en général : le lieu, la durée, la situation,

la cohésion des expérimentateurs, etc. L'objectif étant de relever un maximum d'informations sur le déroulement de la journée, tant sur la plateforme à analyser que sur le dispositif d'évaluation mis en place.

Il convient de noter que ces deux types d'observation ont donné des résultats différents et très complémentaires. L'observation ciblée permet notamment d'éclairer les résultats quantitatifs et précis obtenus par les logs de chaque testeur tandis que l'observation générale permet d'évaluer l'expérimentation générale en vue de l'améliorer pour les sessions futures.

Le questionnaire

Pour compléter les observations, nous avons soumis un questionnaire aux testeurs afin de récupérer leurs avis individuellement pendant ou après les phases de test. Il s'agit d'un questionnaire élaboré par la chaire UNESCO-ITEN et l'équipe Archival mis en ligne et créé sur Framaforms dans un souci de respect du RGPD[164]. Le questionnaire est composé de quatre sections : la première dédiée au profil du testeur, la deuxième axée sur le parcours libre, la troisième consacrée au parcours guidé et la dernière permettant de donner un avis global sur le démonstrateur et le test. L'ensemble des questions soumises aux testeurs se trouve en annexes [de l'annexe n°26 à 28], les deux questionnaires sont différents car nous avons procédé à quelques ajustements entre les deux journées de test. Les réponses individuelles des testeurs anonymisés se trouvent également en annexes [de l'annexe n°29 à 48].

Enfin, les testeurs du matin étaient libres dans leur rédaction des questionnaires et ont pu les remplir à la fin de chaque session de test ou au cours de la session. Sur 10 testeurs, 2 d'entre eux ont répondu aux questionnaires au fur et à mesure tandis que les autres ont répondu à la fin. L'après-midi, en revanche, les testeurs étaient accompagnés d'un observateur et ont donc tous expérimenté le dispositif puis complété les questionnaires en même temps, à la fin. Les deux méthodes comportent avantages et inconvénients : en remplissant le questionnaire parallèlement à l'expérimentation, les testeurs sont certains d'avoir le temps de répondre à tous les champs mais se coupent leur expérience et peuvent orienter leur navigation avec les questions posées tandis que remplir à la fin leur assure de rester neutres mais les met en situation d'urgence au moment de compléter le questionnaire ce qui s'est avéré être le cas pour beaucoup d'entre eux.

[164] Le lien vers le questionnaire : https://framaforms.org/questionnaire-dexperimentation-1683730735

L'entretien collectif

À la fin de l'expérimentation, et une fois que les testeurs ont eu le temps de faire les deux parcours et de remplir les questionnaires correspondants, nous les avons réunis pour discuter collectivement de leur expérience. Il leur était alors demandé en quelques mots de donner leur ressenti sur la prise en main, la navigation, la pertinence du dispositif Archival. Chacun a alors pu prendre la parole pour donner son ressenti et exprimer ses remarques, commentaires et questions. Il ne s'agissait alors pas de faire un retour complet et détaillé -de toute façon le temps ne le permettait pas- mais d'exposer un point de vue global en s'appuyant sur des exemples. Le temps de parole de chacun était estimé de 3 à 5 minutes pour la première session, environ 5 minutes pour la seconde et de 5 à 15 minutes pour la dernière. Les transcriptions de ces retours sont également présentées en annexes de ce mémoire [de l'annexe n°49 à 51].

3) Une démarche pluridisciplinaire et multi acteurs

L'évaluation du dispositif Archival s'inscrit donc dans une démarche pluridisciplinaire tant dans l'évaluation du dispositif en lui-même que dans la construction de la méthode d'enquête. Pour ce faire, nous avons été puiser dans la littérature existante sur le sujet et avons croisé les approches en co-construisant la méthode avec les différents membres de l'équipe Archival.

Au sein de l'équipe, l'évaluation est divisée en deux branches majeures : l'évaluation orientée système confiée aux trois laboratoires d'informatique ainsi que l'évaluation orientée usagers prise en charge par la Chaire UNESCO-ITEN et la FMSH. Cependant, ce travail est moins binaire qu'il n'y paraît puisque les journées d'expérimentation au cours desquelles nous avons fait tester le démonstrateur à des groupes de testeurs est un croisement de ces deux branches liant les objectifs de chacune qu'ils soient individuels ou partagés. L'enquête mise en œuvre au cours de ces journées avait pour objectif de collecter les retours des testeurs sur les algorithmes et leur pertinence pour l'usager, sur la pertinence des liens générés par IA mais aussi sur les pratiques et usages documentaires des usagers et sur leur appréciation et appropriation du dispositif.

Ghislaine AZEMARD (chaire UNESCO-ITEN) et Michel AGNOLA (FMSH) étaient en chargés de réaliser l'enquête pour les journées d'expérimentation notamment par la création d'un questionnaire faisant la synthèse des éléments énoncés ci-dessus. Dans l'enquête et en

particulier dans le questionnaire, il a donc été nécessaire de croiser les attentes et les demandes de chaque groupe, nous avons épaulé l'équipe projet dans cette tâche.

Par ailleurs, l'enquête est en elle-même une démarche pluridisciplinaire à la croisée des SIC, de la sociologie, la psychologie, la statistique, l'anthropologie, etc, où chaque discipline enrichit les autres. Pour l'enquête Archival, nous avons mêlé les approches en associant Irène BASTARD, sociologue de la BnF, au projet afin qu'elle nous oriente, nous prodigue des conseils et assiste aux journées d'expérimentations pour nous faire des retours et nous permettre d'améliorer notre méthode. Spécialisée dans l'enquête auprès des publics, elle nous a surtout épaulé dans l'élaboration de notre méthode d'observation pour savoir quoi observer, comment le faire et comment exploiter les résultats obtenus. Nous avons également eu l'occasion de travailler avec elle sur la réalisation de questionnaires et d'entretiens collectifs, aspects que nous n'avons pour l'instant pas pu exploiter réellement car le questionnaire et l'entretien collectif dans l'enquête Archival étaient prévus par l'équipe projet pour la première journée d'expérimentation, trame que nous avons suivie pour la seconde journée afin d'obtenir des résultats comparables et de ne pas dénaturer le protocole mis en place lors de la première journée.

La démarche globale est également pluridisciplinaire puisque les aspects quantitatifs sont évalués par les informaticiens du projet tandis que nous nous occupons des aspects qualitatifs. Ainsi, la performance des algorithmes (le nombre de questions et de liens générés, taux de questions écrites dans un français correct) a été mesurée par les trois laboratoires d'informatique du consortium Archival. Il en est de même pour l'analyse des *logs* de connexion des vingt expérimentateurs qui seront analysés dans le courant du mois de juillet, analyses devraient ensuite être confrontées aux observations que nous avons réalisées au cours des différentes sessions. Une hybridation des méthodes et compétences qui devrait permettre de tirer pleinement profit des traces laissées par les testeurs.

Nous avons ainsi essayé d'intégrer à notre démarche des conseils et pratiques d'autres disciplines que les SIC bien que cela soit encore marginal dans la création du protocole. Nous aimerions pouvoir étayer ces apports pour les expérimentations futures. Il serait également possible d'ouvrir la démarche à d'autres disciplines telles que l'ergonomie cognitive et le design web, la BnF travaille notamment en étroite collaboration avec des spécialistes en ergonomie et interfaces web de chez TALAN (un cabinet de conseil en innovation et transformation par la technologie qui travaille régulièrement pour la BnF notamment à des fins d'évaluation).

Conclusion

Au travers de ce second chapitre, nous avons ainsi pu présenter le cas d'étude permettant de nous exercer, sur un projet réel, à la mise en place d'un protocole d'expérimentation dédié à une interface et la réalisation de celui-ci. Pour évaluer la pertinence, l'apport et la compréhension des liens générés par algorithmes, des testeurs issus des SIC et de la documentation ont été interrogés par le biais d'une triple enquête (observation, questionnaire, entretien) construite à partir de la boîte à outils élaborée dans le chapitre 1.

Chapitre 3 : Résultats et recommandations

Au regard de ce qui a été énoncé précédemment, il est désormais possible de présenter et discuter les résultats obtenus grâce à la méthode d'évaluation élaborée en vue d'en comprendre les intérêts, limites et la portabilité à d'autres dispositifs. Pour ce faire, les résultats seront traités séparément entre interface et contenu avant d'être discutés ce qui permettra l'élaboration de recommandations à la fois théoriques et pratiques tant sur le dispositif que sur les expérimentations.

I/ Résultats obtenus sur l'interface

Grâce à la mise en place de la méthode d'enquête agile et composite, nous avons réussi à collecter un grand nombre de retours auprès des testeurs à la fois sur l'interface graphique, l'ergonomie et la navigation mais également sur leur appropriation du dispositif.

1) Sur l'interface graphique

L'interface graphique pensée par le consortium Archival est une interface homme machine (IHM) basée sur l'expérience de lecture concrète et physique connue par le lecteur. Pour l'expliquer plus concrètement, le démonstrateur a été conçu comme une abstraction de la réalité : le lecteur y consulte une revue prenant la forme d'une liste ou pile de numéros, il peut les ouvrir dans un bureau virtuel et les lire un a un ou en parallèle les uns des autres, page après page, la lecture est appuyée par l'apport de fiches amovibles et de ressources externes. Le démonstrateur comporte alors deux niveaux : la découverte générale des collections et du corpus par la navigation dans la revue, dans les personnages cités ou dans les notions et d'autre part, le travail sur le corpus dans l'interface dite de « *bureau* » au sein de laquelle le lecteur peut ouvrir les articles, les lire page à page, travailler sur le texte à l'aide des deux algorithmes et étendre son travail avec les ressources et rebond proposés.

En explorant le dispositif Archival, les testeurs découvrent des interfaces graphiques colorées, ludiques et interactives. Au regard des retours collectés, elles semblent plutôt appréciées par les testeurs et ce indépendamment des panels interrogés. C'est d'ailleurs un des éléments mis en avant par les expérimentateurs lors des entretiens collectifs « *visuellement, je trouve que c'est très réussi* » (testeur 3), « *l'interface est chouette* » (testeur 11) ou bien «

j'adore l'interface » (testeur 16). Puisqu'aucune question ne permettait d'aborder ce point dans le questionnaire soumis aux premiers testeurs, nous y avons remédié dans le second questionnaire afin de systématiser les réponses. Les réponses à la question « *Que pensez-vous visuellement de l'interface ?* » sont alors majoritairement élogieuses, comme nous l'avions constaté à l'oral (« *design plutôt agréable* » testeur 17, « *attractif* » testeur 20, « *agréable visuellement* » testeur 19).

Bien qu'il s'agisse de goûts personnels et donc d'un point de vue tout à fait subjectif, le résultat nous semble significatif. Un aspect d'autant plus intéressant qu'une partie des testeurs interrogés sont spécialisés dans le design d'interfaces, l'un d'entre eux qui déclare alors :

> « *Je suis graphiste donc tout ce qui est graphique ça me parle énormément, donc le fait que l'interface ait énormément de champs visuels et qu'on puisse organiser l'information de façon très visuelle c'était un atout pour moi* » (testeur 7).

Cependant, si le premier ressenti est en très grande majorité positif, il est souvent nuancé par le répondant qui évoque par la suite des problèmes et pistes d'amélioration. Nombreux sont ceux qui ont déploré le fait que les différentes interfaces soient trop chargées et ont jugé qu'il faudrait les alléger pour les rendre plus lisibles et agréables. Un problème de taille bien résumé par le testeur 13 qualifiant le démonstrateur comme « *un environnement qui est assez complexe et déroutant, on aurait besoin d'un peu de simplicité en fait dans l'interface* ». Elle semble alors pouvoir plaire aux usagers mais demeure encore trop complexe à appréhender pour qu'ils en soient pleinement satisfaits : « *je me suis senti perdu un certain moment sur la complexité de l'interface* » (testeur 7), « *ce n'est pas évident* » (testeur 6) ou plus précisément :

> « *L'interface est agréable visuellement au départ de l'utilisation. Une fois que la navigation est lancée, certains affichages ne sont visuellement pas "agréables" : Vignettes, listes des articles dans la largeur de fenêtre, affichage des articles entre mode image et mode texte »* (testeur 19).

Puisque le dispositif actuel est un prototype, il est tout à fait normal que le graphisme des différentes interfaces ne soit pas la priorité dans le développement du démonstrateur. Cela a toutefois constitué un frein pour certains testeurs qui auraient préféré travailler sur des interfaces plus travaillées, un travail plus poussé sur le design de certaines interfaces leur semble alors nécessaire et ils citent de nombreux « *détails* » pour les améliorer. Car le design - bien trop souvent relayé à une fonction esthétique- est en réalité primordial dans la conception

d'un produit y compris pour une interface documentaire. En design, il y a des grands principes à respecter applicables à toute conception. Don NORMAN est celui qui a érigé la liste la plus communément admise de ces principes, il en répertorie six incontournables et applicables à toute forme de design[165] : la visibilité (*visibility*), les retours utilisateur (*feedback*), situer la progression (*mapping*), designer par la contrainte (*design through constraint*), l'homogénéité (*consistency*) et l'affordance (*affordance*).

En reprenant ces principes point par point, on en comprend que de nombreuses petites modifications pourraient améliorer les interfaces et le démonstrateur dans son ensemble. La visibilité c'est permettre à l'utilisateur de trouver la fonctionnalité qu'il cherche facilement. Dans le cas d'Archival, le dispositif a les défauts de ses qualités, comme les interfaces sont riches et recèlent de nombreuses informations et fonctionnalités, la visibilité de ces dernières n'est pas optimale. L'écran est lourd et chargé, certains onglets ne sont pas réellement mis en valeur (notamment masquer les fiches qui est difficile à trouver) et d'autres s'avèrent même très difficiles à trouver -comme les algorithmes par exemple. Cette difficulté d'accès à la fonctionnalité cherchée est un frein important dans la navigation du testeur, ce qui a été soulevé maintes fois au cours des différentes sessions.

Les retours utilisateurs servent à communiquer une information sur l'action qui vient d'être faite, ils sont pourtant parfois inexistants dans les différentes interfaces. Il est très perturbant pour un utilisateur de cliquer quelque part sans avoir de retour, ce phénomène a notamment été relevé pour les pouces d'évaluation des algorithmes, le manque d'une roue de chargement (pour la première journée), pas de question générée sans aucune explication ni retour, etc. Un point qu'il convient de nuancer toutefois car depuis les expérimentations, quelques feedbacks ont été ajoutés aux différentes interfaces.

Le mapping sert à situer l'utilisateur dans le dispositif qu'il utilise (comme l'exemple connu du « *plan du site* »), il pourrait également être amélioré car nombreux sont les testeurs qui avaient du mal à se situer dans le démonstrateur et ce peu importe les profils, domaines ou spécialités. Ce problème n'a, certes, pas été partagé par tous mais s'est avéré très handicapant pour ceux qui y étaient confrontés soit près de la moitié des testeurs.

Le design par la contrainte consiste à limiter les actions possibles au sein des interfaces comme cacher les fonctions qui ne sont pas accessibles momentanément, cela permet de faciliter la navigation des utilisateurs et leur évite les erreurs 404 ou autres. Dans le cas

[165] Norman, D. (2013). The Design of Everyday Things : Revised and Expanded Edition. Constellation.

d'Archival, on note que les facettes de recherche qui devraient être développées mais ne le sont pas encore à l'heure actuelle ne sont pas cliquables, évitant ainsi au testeur de perdre du temps à aller sur une interface vide ou qui lui retournerait une erreur ce qui est un point positif. Dans Archival, le design par la contrainte pourrait être développé mais n'est pas le plus préoccupant.

L'homogénéité du dispositif est plus contrariante que l'aspect précédent. En design, l'homogénéité consiste à faire en sorte que tous les éléments soient similaires pour des tâches similaires, ainsi, tous les textes cliquables doivent être indiqués de la même manière, toutes les polices doivent être similaires ou du moins s'accorder entre-elles. Or, le dispositif fait état d'une disparité telle qu'il est difficile de trouver plusieurs fois la même police [voir illustration n°53]. Le testeur 19 a particulièrement été sensible à cette question (bien qu'évoquée par d'autres précédemment) et explique dans le questionnaire « *Un utilisateur "Dys" aura sans doute beaucoup de mal à utiliser l'interface.* » et développe à l'oral que ce problème est notamment dû au grand nombre de polices diverses et variées. Sans même parler de rendre le site conforme aux normes d'accessibilité[166], il est pourtant nécessaire d'en faciliter l'accès en particulier en uniformisant le design des interfaces et les polices d'écriture[167].

Enfin, l'affordance est le dernier principe mis en avant par Don NORMAN, il s'agit de la capacité d'une fonction à renvoyer à son utilisation sans explication ou tutoriel, de sa bonne usabilité. Ce point fait également défaut dans le dispositif et pourrait être optimisé, certains testeurs ont par exemple eu du mal à comprendre que la frise chronologique de l'entrée revue était cliquable pour afficher les dates rapidement, un testeur nous a d'ailleurs avoué à la fin du parcours guidé qu'il venait de comprendre que les fiches étaient déplaçables, s'exclamant : *« Déjà je viens de découvrir qu'on peut glisser les fiches* » après plus d'une heure de navigation. Les interfaces étant chargées, il est d'autant plus difficile de se concentrer sur les fonctionnalités et d'en comprendre le fonctionnement.

Nous aimerions ajouter qu'il manque un élément dans cette liste à savoir le confort de l'utilisateur, paramètre qui nous semble être primordial -d'autant plus dans une approche orientée usager-, il est notamment mis en avant dans un article dédié à l'analyse de la qualité et de la performance des sites web : « *Pour qu'un site Web puisse assurer une expérience utilisateur optimale, il est indispensable de s'assurer qu'il répond toujours aux exigences des*

[166] Voir https://accessibilite.numerique.gouv.fr/ [en ligne] consulté le 16/06/2023.
[167] Sitbon, L., Bellot, P. & Blache, P. (2010). *Op cit.*

internautes et des évolutions technologiques »[168]. Ce point a été soulevé concernant différents aspects des interfaces, le fond gris foncé du bureau dénote trop avec la clarté de la page et a provoqué un inconfort visuel important à plusieurs testeurs, il en est de même pour le menu alphabétique des personnages qui défile très rapidement sous les yeux lorsqu'une lettre est sélectionnée ce qui est très désagréable visuellement (« *ça donne mal au cœur un peu* », testeur 18, « *ça m'a dérangé visuellement, je vous le dit parce que c'était pas agréable en fait* », testeur 19).

2) Sur l'ergonomie et la navigation

Les résultats obtenus sur la navigation sont plus nuancés et critiques que ceux sur l'interface graphique. Les testeurs sont en désaccord quant à la facilité de navigation au sein des interfaces car, là où certains ont trouvé qu'il était facile de se repérer et de naviguer pour les expérimenter, d'autres ont eu beaucoup de difficultés à se situer au cours de leur test : « *il m'a fallu demander un peu d'aide pour aller d'un espace à l'autre* » (testeur 4) ou « *Pas toute suite, j'ai eu besoin de naviguer pour me situer.* » (testeur 10), un problème partagé par la majorité des expérimentateurs bien que les avis soient plus ou moins tranchés.

Naviguer dans le démonstrateur s'est avéré inconfortable pour certains, difficile pour d'autres et très laborieux pour les derniers. Ainsi, ils sont plusieurs à avoir mis en avant le fait qu'une présentation du dispositif ou la mise en place d'un didacticiel avant le début de la navigation était nécessaire, l'un d'entre eux explique : « *Pour comprendre le fonctionnement du site et son mode de consultation, la présentation initiale s'est avérée primordiale. L'outil ne paraît immédiatement/intuitivement accessible sans cela.* » (testeur 12) ce qui est corroboré par d'autres. Ce phénomène constitue une réelle entrave dans la prise en main du dispositif par les expérimentateurs d'autant plus que -bien qu'il s'agisse actuellement d'une preuve de concept (*proof of concept*)- ce genre d'interface documentaire est pensé pour que l'usager puisse y accéder à distance et cela, sans barrière technologique à l'entrée.

Or, la présentation préalable de 10 à 15 minutes faite par l'un des membres de l'équipe Archival ne peut être reproduite pour chaque personne qui voudrait se connecter au dispositif, il faudrait donc certainement envisager la création d'un didacticiel accessible lors de la première consultation du site et retrouvable à tout moment dans le dispositif en cas de besoin ou bien la

[168] Truphème, S. & Gastaud, P. (2023). Outil 22. L'analyse de la qualité et de la performance d'un site Web. Dans : S. Truphème & P. Gastaud (dir), La boîte à outils du Marketing digital (p. 74-75). Paris : Dunod.

création d'un assistant de navigation. Y ayant quelque peu réfléchi, nous pensons qu'il pourrait prendre la forme du robot Archival présenté dans la vidéo sur la page d'accueil [voir illustration n°53], se situer en bas à droite de l'écran tout au long de la navigation et serait un agent informationnel et non conversationnel [voir illustration n°54]. Il serait chargé de faire un court didacticiel dès que l'usager arrive pour la première fois sur une page (didacticiel qui puisse être passé si l'usager juge ne pas en avoir besoin ou revisionné s'il en ressent de nouveau besoin) et s'occuperait de l'explicabilité simplifiée des algorithmes et fonctionnalités en revoyant directement dans le texte vers les pages plus étayées (explicabilité des algorithmes et algorithmes utilisés). Projet qui pourrait être plus amplement étayé si l'idée semble prometteuse même s'il ne verra *a priori* pas le jour.

Concernant l'efficacité du dispositif, les avis sont également partagés puisque si quelques-uns semblent trouver la démonstration « *productive* » (testeur 1) car permettant d'entrer rapidement dans les contenus, nombreux sont ceux qui ont éprouvé des difficultés avant de commencer à naviguer sans article d'entrée pendant le parcours libre. Cela est notamment dû au fait que les articles et le thésaurus sont centraux dans le projet et que naviguer hors de ce cadre n'est pas aisé à appréhender surtout pour ceux qui ne connaissaient ni le thème, ni la revue. La plupart des testeurs étaient tout de même globalement satisfaits et ont réussi à naviguer au sein du dispositif grâce aux différentes fonctionnalités : « *Les menus de base sont très intéressants et permettent d'aborder la recherche des éléments complexes de façon simplifie notamment avec les graphiques et la chronologie* » (testeur 7).

Les testeurs issus du monde des bibliothèques et de la documentation ont trouvé très particulière voire peu lisible la façon d'agencer les listes de résultats. Ils déplorent que ces listes manquent de clarté et conseillent de les revoir et les réorganiser plus distinctement. L'un d'entre eux s'exprime sur ce point en disant : « *Après on est des professionnels des listes de résultats, on est un peu maniaques peut-être. On ne sait pas où on en est, quand on a 3 ou 4 résultats, on comprend que c'est dans le même article mais ce n'est pas d'une clarté absolue.* » (testeur 14), prise de parole qui est acquiescée par les autres testeurs de la session. Les résultats sont, en effet, présentés dans des listes avec très peu de catégorisations. Il serait possible de mieux les différencier en fonction des numéros, des articles, etc [voir illustration n°56].

Nombreux sont ceux qui ont également trouvé que renseigner deux listes de mots-clés par page était trop lourd en termes de lecture et de surcharge cognitive, d'autant plus que le démonstrateur est déjà complexe. Ils auraient alors préféré avoir un résumé ou une synthèse des mots-clés par article car ils estiment qu'il y a une « *perte d'une vision d'ensemble sur le numéro*

ou l'article d'autant plus que les mots clés sont mis à la page. […] ça brouille énormément l'information, on perd la vue d'ensemble » (testeur 12) et que les notions, mots-clés et personnages sont moins liés à la page consultée qu'à l'article en lui-même. Un autre testeur confirme ce point de vue : « *les mots-clés à la page dans ce démonstrateur ce n'est pas forcément obligatoire de les afficher* » (testeur 16).

Les expérimentateurs ont aussi globalement trouvé que l'ergonomie générale pâtie du manque de micro-interactions habituelles sur les interfaces web *(« il n'y a pas de "retour utilisateur"* », testeur 4), comme la roue de chargement, le changement de couleur des liens au survol, les feedbacks sur les pouces, … Ces manques de clarté n'empêchent pas les testeurs de naviguer dans le démonstrateur mais bloquent certains usages et les freinent dans leur expérimentation : « *Il y a des petites améliorations d'ergonomie qui peuvent vraiment faciliter la navigation* » (testeur 14). Ils ont cependant compris que le dispositif était encore en développement et que ces ajouts seraient effectués sous peu.

Enfin, l'un des testeurs évoque la question épineuse du responsive, comprenant qu'elle est impossible à résoudre pour ce genre de projet « *On a beaucoup de texte à l'écran, s'il fallait chercher une interface responsive je ne sais pas trop comment on ferait* » (testeur 13), ce qui est confirmé par un autre testeur dans ses réponses au questionnaire « *Le site nécessite de grands écrans pour bénéficier de l'affichage horizontal des fenêtres qui s'ajoutent. Quid d'une consultation sur un device mobile ?* » (testeur 12). Une question qui semble pourtant insoluble au regard de la complexité du dispositif proposé, il serait intéressant de creuser cette question du *responsive* dans le cas de démonstrateurs tels que celui proposé par l'équipe Archival.

3) Sur l'appropriation du dispositif

La combinaison des deux parties précédentes nous pousse alors à questionner l'appropriation du dispositif de recherche et d'exploration permise par le dispositif actuel. Une question difficile à synthétiser de façon uniforme au regard des nombreux changements qui ont eu lieu et sont encore en cours mais qui nous permettent de déceler quelques axes majeurs.

Notion qu'il convient toutefois de resituer dans le contexte scientifique puisque le terme -assez galvaudé- risquerait d'engendrer des incompréhensions. Sa définition fait d'ailleurs

l'objet d'une partie dans l'article « *L'étude des dispositifs d'accès à l'information électronique* »[169] où les auteurs expliquent :

> « *Ce terme d'"appropriation" nécessite en effet une attention particulière. Le thème de l'appropriation est récurrent dans la théorie de l'innovation, qu'elle soit technique ou sociale qui met en évidence que, selon les cas, le processus d'appropriation peut déboucher sur l'aliénation ou sur la libération.* »[170]

Ils différencient ainsi les deux acceptions du terme :

> « *La première fait référence au concept d'adaptation, il s'agit de "rendre propre à une destination précise", de "se conformer" à quelque chose ou à une situation tandis que la seconde définition renvoie au fait de "s'attribuer", le plus souvent "indûment", quelque chose, d'en faire sa "propriété".* »[171]

Il ne s'agit, dans le cas de cette expérimentation, non pas de s'attribuer le dispositif ou de se l'accaparer mais bien de « *négocier le protocole d'utilisation tel qu'il a été rédigé par les concepteurs du dispositif* »[172] car c'est ce type d'appropriation qui aide le plus à déterminer les usages réels des testeurs en dehors de l'expérimentation.

Il a été demandé aux testeurs d'exprimer leur ressenti par rapport à l'appropriation du dispositif dans le questionnaire : « *Est-ce que ce type d'interface facilite la découverte et l'appropriation des données documentaires ?* » (en une seule question pour le premier questionnaire et divisé en deux questions pour le second). Les réponses sont variées puisque les résultats sont les suivants : 6 réponses sont positives, 4 réponses plutôt positives, 5 réponses mitigées, 3 réponses négatives et 2 réponses vides. La plupart des réponses sont positives comme « *C'est très intéressant* » (testeur 13), « *plutôt satisfaisant* » (testeur 8) ou « *effectivement elle facilitera l'appropriation des données* » (testeur 19).

Les chercheurs semblent notamment s'être projetés à utiliser ce type de dispositif pour leurs recherches comme l'explique l'un d'eux : « *j'aimerais avoir la même chose pour les articles de mon corpus de recherche* » (testeur 5), une telle projection est tout à fait encourageante du point de vue de l'appropriation du dispositif par les usagers. Un professionnel des bibliothèques confirme ce point en mettant l'accent sur l'avantage de l'interface bureau qui

[169] Ihadjadene, M. & Chaudiron, S. (2008). *Op cit.*
[170] *Ibid.*
[171] *Ibid.*
[172] *Ibid.*

lui semble prometteuse : « *On aimerait bien tous avoir ça pour les portails de recherches documentaires en général : un espace de travail et puis des fonctionnalités avancées* » (testeur 16).

Il semble donc que l'appropriation soit permise par le dispositif mais perfectible notamment comme l'expliquent plusieurs testeurs : « *pour l'appropriation, il me manque la possibilité d'exporter les résultats de mes recherches pour l'intégrer à mes propres outils de travail.* » (testeur 13) ou encore « *Cela facilite la découvrabilité. Pour l'appropriation, cela dépendra des possibilités d'export ou de mise au panier et globalement la manière dont on garde trace de son travail de lecture* » (testeur 14). En revanche, les derniers testeurs ne semblent pas conquis : « *Pour le moment, je ne suis pas très convaincu* » (testeur 3) ou encore « *non* » (testeur 17).

Ce manque d'adhésion pourrait être lié à plusieurs facteurs. D'abord, les testeurs semblent d'accord pour dire qu'il manque des fonctionnalités de base au démonstrateur - notamment en comparaison avec les outils de recherche documentaire habituels (Cairn, Persée, Gallica ont été mentionnés). Ils déplorent ainsi l'impossibilité de télécharger le contenu (les articles, notions ou personnages) en PDF, d'extraire les données, de les imprimer, de copier-coller du texte dans un logiciel de traitement de texte, d'exporter la bibliographie dans Zotero, etc. Toutes ces fonctionnalités habituelles leur ont manqué et n'ont pas permis de réelle appropriation faute de possibilité de travailler réellement dans Archival et non uniquement lire. Un phénomène renforcé par le fait que la lecture seule n'est pas non plus agréable du fait de la complexité de l'interface, du fonctionnement par page et de la surcharge cognitive qu'il engendre.

De plus, dans leurs réponses aux questionnaires, les testeurs ont également mis en lumière l'importance des habitudes et de l'expérience personnelle/ professionnelle de chacun dans l'utilisation d'un outil. C'est notamment visible dans leurs réponses à la question « *Quel point d'entrée dans le corpus avez-vous privilégié (moteur de recherche, revue, personnages, ...) ? Pourquoi ? En quoi cela vous a-t-il été utile ?* » car beaucoup affirment avoir préféré une fonctionnalité plutôt d'une autre par habitude : « *on est habitué* » (testeur 8), « *par tropisme* » testeur 12), « *par habitude* » (testeur 13). La familiarité avec un outil joue pour beaucoup dans la capacité de l'usager à se l'approprier, il n'est donc pas étonnant que l'approche sensible et le parcours individuel de chacun ait une grande part dans ce genre d'expérimentation. L'appropriation et le rôle des habitudes sont très importants dans les facteurs à prendre en compte lors de notre évaluation :

> « *L'expérience utilisateur est passée au premier plan, elle considère la satisfaction de l'usager, son ressenti par rapport à un service ; l'expérience utilisateur renvoie à des entrées variées que sont la performance fonctionnelle, l'ergonomie de l'interface, le ressenti émotionnel, la relation avec le producteur… Le design interactif est au cœur de cette dimension.* »[173]

Il convient enfin d'ajouter que la démarche d'explicabilité semble aider l'appropriation du dispositif par les usagers grâce à la confiance qu'il peuvent accorder au système. Un usager n'utilisera pas un dispositif dans lequel il n'a pas confiance surtout pour des recherches professionnelles ou scientifiques en particulier lorsque l'on parle d'intelligence artificielle. C'est notamment ce que démontre J. MARTINEAU :

> « *Autre préoccupation majeure : la confiance envers ces technologies et leur acceptabilité sociale. Ainsi, la confiance envers les technologies d'IA est indispensable à leur acceptation par les différents acteurs dans nos sociétés […]. Si le public est méfiant ou ne comprend pas le fonctionnement d'une technologie, il pourrait la rejeter, ce qui nuira nécessairement à son utilisation et à son adoption.* »[174]

Ainsi, l'amélioration de l'explicabilité des algorithmes et du système en général devrait permettre une meilleure confiance des usagers en le dispositif par la compréhension de son fonctionnement, de ses forces et ses faiblesses pour à terme en permettre une plus grande appropriation.

II/ Résultats obtenus sur le contenu

Les retours des testeurs ont été tout aussi nombreux et diversifiés sur le contenu du dispositif que sur l'interface graphique, nous avons classé les résultats obtenus en trois parties :

[173] Boustany, J., Broudoux, É. & Chartron, G. (2013). Introduction. Diversification des médiations informationnelles. Dans : Joumana Boustany éd., La médiation numérique : renouvellement et diversification des pratiques: Actes du colloque Document numérique et société, Zagreb 2013 (p. 7-10). Louvain-la-Neuve: De Boeck Supérieur. https://doi-org.ressources-electroniques.univ-lille.fr/10.3917/dbu.chron.2013.01.0007
[174] Martineau, J. (2023). *Op cit.*

le corpus, les facettes de recherche « *classiques* » et les facettes dites « *intelligentes* » à savoir les deux algorithmes.

1) Sur le corpus

Les résultats mis en avant par les testeurs sur le corpus révèlent une grande variété de réactions à la fois positives et négatives et dont le résultat est très contrasté. Ils s'axent sur des aspects différents qui peuvent être synthétisés en trois points majeurs : l'importance voire la prépondérance du thésaurus dans le projet permettant difficilement d'effectuer des recherches en dehors du cadre établi, l'étendue du corpus difficile à cerner et à comprendre et la granularité du dispositif parfois bloquante pour accéder aux résultats escomptés.

Le rôle prépondérant du thésaurus

Le démonstrateur proposé repose sur les facettes énoncées précédemment, à savoir les personnages, les notions, la revue, le moteur de recherche, etc. La page « *notions* » contient alors les termes classés dans le thésaurus du projet par les chercheurs de la chaire UNESCO-ITEN et de la FMSH, les quelques centaines de mots recensés étant divisés en quatre catégories. Mais en lisant un article, l'usager voit au bas de chaque page deux types de notions liées au texte : les notions entrées dans le thésaurus et des termes générés automatiquement en lien avec la page en cours de consultation. Ces deux listes de termes sont donc différentes car l'une est close, a été pensée en amont par l'intelligence humaine tandis que l'autre est ouverte et générée automatiquement par algorithme soit par intelligence artificielle. Cette dualité n'est cependant pas clairement expliquée ce qui a pu poser problème à quelques testeurs lors de leur consultation.

Le dispositif est fait de telle manière que le thésaurus réalisé par la chaire UNSECO-ITEN et la FMSH est central dans la recherche documentaire, la recherche et l'exploration par mots-clés hors du thésaurus s'avère donc presque impossible comme l'ont révélé les testeurs qui souhaitaient travailler sur le féminisme et l'autogestion par exemple : « *le féminisme et l'autoG qui n'est pas un thème central de la revue et qui n'était pas dans le thésaurus (pas de notion féminisme ou femme ou associée)* » (testeur 5). Ce qui est renforcé par le manque de connaissances sur le thème et la revue d'un certain nombre de testeurs interrogés.

Par ailleurs, les testeurs sont plusieurs à avoir demandé comment le thésaurus a été créé, comment et par qui les catégories ont été générées. C'est la demande formulée par le testeur

15 : « *les notions ça fonctionnait pas mal mais comment les 4 catégories ont-elles été réalisées ? Éventuellement donner une légende pour expliquer la classification* ». Le travail réalisé par la FMSH et la chaire UNESCO-ITEN est alors intéressant mais pas suffisamment clair, les usagers ont besoin de savoir d'où viennent les mots-clés, pourquoi ils ont été sélectionnés et comment ils ont été classés d'autant plus que la distinction entre ceux du thésaurus et ceux générés par IA n'est pas limpide pour tous :

> « *Par contre, j'aurais besoin d'un peu plus d'explicitation entre ce qui vient du thésaurus et ce qui vient des mots-clés, vous l'avez expliqué rapidement mais ça peut être sous forme d'outil, de quelque chose en tout cas qui me rendra facilement la différence entre les deux listes de concepts* » (testeur 13).

Les testeurs ont d'ailleurs été nombreux à vouloir ajouter leurs propres notions au thésaurus pour faciliter les recherches et permettre une navigation personnalisée : « *je me suis dit que ce serait sympa quand même de pouvoir ajouter ses propres notions au thésaurus.* […] *c'est frustrant quand même de pas avoir la main sur le thésaurus car on sent quand même qu'il est très central* » (testeur 14).

Enfin, ce type de travail pose la question de la possibilité de reproduire ou adapter le dispositif et l'étendre à d'autres corpus, d'autres revues sur l'autogestion ou des thèmes connexes. Ce travail de choix des termes et de classification est long, fastidieux et doit être fait à la main, puisque le thésaurus est aussi central dans le dispositif, ce dernier est-il réalisable sur d'autres corpus plus vastes ? Est-il adaptable à plus grande échelle ?

> « *On a un vrai problème de découvrabilité, je ne connais pas du tout la revue Autogestion, donc on voit un peu à partir des mots-clés etc, on commence à comprendre de quoi ça parle. Mais admettons que j'ai la même interface pour 10 revues, comment on fait ?* » (testeur 12).

Un rôle prépondérant d'autant plus problématique -surtout mentionné par les professionnels de la documentation- que le travail sur les entités nommées dans le texte est actuellement très léger et peu opérant alors même qu'il s'agit d'une fonctionnalité de recherche basique et nécessaire. Le manque de recherche par entités nommées a perturbé et déçu les testeurs des différentes sessions, un désarroi que l'on retrouve chez le testeur 12 : « *pareil pour les personnages, on travaille en bibliothèques, la recherche par entités nommées c'est quelque chose que tout le monde met en avant, on fait plein de programmes pour rechercher les entités nommées…* »

L'étendue du corpus

D'abord, il est important de noter que le périmètre du dispositif semble clair pour certains usagers mais pas pour les autres, une incertitude gênant quelques testeurs au cours de leur navigation comme ils le mentionnent au cours des entretiens collectifs : « *Je n'ai pas compris quel était le périmètre documentaire interrogé en dehors de la revue Autogestion.* » (testeur 15), « *J'avais du mal à distinguer quand je suis dans le corpus et hors corpus* » (testeur 9) mais aussi « *des fois on ne sait pas où on va ni quel est le périmètre, est-ce qu'il y a une revue ou plusieurs ? On ne sait pas trop ce qu'on interroge* » (testeur 14). Il faudrait donc s'assurer de clarifier ce point pour permettre aux testeurs de mieux se repérer dans le dispositif dans lequel ils naviguent. Un autre point intéressant est soulevé par cette question du repérage au sein du démonstrateur car ce sont surtout les professionnels des bibliothèques qui ont noté cet aspect comme étant négatif tandis que les chercheurs ont moins relevé et mentionné ce point qui semble donc les avoir moins touché ou marqué.

Enfin, « *L'outil implique de connaître le corpus initial et ne permet a priori pas d'appréhender un corpus inconnu. On est incité à aller rapidement sur le contenu, à lire en détail page/page sans avoir de vue d'ensemble du corpus* » ce qui est encore plus handicapant pour les usagers n'ayant aucune connaissance du corpus.

Il convient également de noter que le dispositif n'est basé que sur une seule revue qui constitue un corpus homogène et clos sur un thème unique -bien que vaste et pluridisciplinaire. Même s'il renvoie vers des ressources documentaires externes de natures variées, il ne constitue pas un corpus suffisant pour mener une recherche sur le thème de l'autogestion. Sur ce point, nous en sommes en désaccord avec les testeurs qui ont répondu majoritairement positivement à la question « *Est-ce que les fonds mis à disposition vous paraissent suffisants pour une investigation sur l'autogestion ?* ». Le corpus ne peut pas être suffisant pour une recherche sur un thème car il est basé sur une source augmentée de quelques apports extérieurs alors qu'une recherche implique de réellement croiser différentes sources et de les faire dialoguer. Tout au mieux, le corpus permet actuellement une découverte et une exploration du thème et de la revue. De plus, aucun lien n'est généré vers des articles traitant de la revue en elle-même comme C. WEILL[175] ou V. SCHAEPELYNCK et E. SUSTAM[176] ce qui serait pourtant un apport intéressant afin

[175] Weill, C. (1999). *Op cit*.
[176] Schaepelynck, V. & Sustam, E. (2018). *Op cit.*

de remettre la revue en perspectives comme l'explique le testeur 16 : « *on ne sait pas situer a priori cette revue dans le champ de réflexion plus large portant sur l'autogestion* ».

La granularité du dispositif

L'un des problèmes majeurs soulevés par les testeurs est celui de la granularité de la recherche d'information proposée par Archival. Le degré de granularité d'un dispositif d'accès à l'information est particulièrement important car il permet un accès plus ou moins fin à l'ensemble ou une partie du corpus, un point important soulevé par Y. CHIARAMELLA et P. MULHEM dans une partie dédiée aux perspectives pour les SRI :

> « *La granularité des documents. Si un objectif futur des SRI est de fournir les parties de documents les plus adaptées au besoin de l'utilisateur (focalisation des réponses), il faut alors être en mesure d'extraire les différents grains d'information et de les manipuler de manière efficace et satisfaisante pour les utilisateurs.* »[177]

Il y a, en effet, deux échelles au sein du dispositif Archival : la vue d'ensemble très générale avec les notions et personnages mais hors de la revue à proprement parler et le focus très rapproché sur un paragraphe ou quelques phrases au sein d'un article. Quelques testeurs qui ont beaucoup apprécié la dualité de la plateforme surtout pour la première session ce qui est résumé par le testeur 11 expliquant qu'il aime naviguer sur une interface en effectuant des « *zooms* » et « *dézooms* » dans le corpus pour découvrir le corpus et les fonctionnalités, en avoir une vision globale et pouvoir s'arrêter et approfondir ce qui l'intéresse. Cette dualité entre survol et attention profonde est un réel avantage dans le cadre d'une démarche exploratoire selon lui, ce qui est confirmé par d'autres.

Une grande partie des testeurs ont, en revanche, été beaucoup plus critiques avec ce fonctionnement (dont la plupart sont des professionnels de la documentation et de la RI), déplorant le manque d'usabilité ou d'intérêt d'un tel système pour la recherche documentaire. Ils ne déplorent pas la dualité de l'interface en elle-même mais l'articulation complexe entre exploration distante et lecture rapprochée. Le problème étant qu'Archival ne propose pas réellement de gradation ni dans la lecture, ni dans la recherche d'information et que s'il est possible de passer d'un niveau à l'autre, cela se fait toutefois sans gradation. Il n'est pas possible

[177] Chiaramella, Y. & Mulhem, P. (2007). *Op cit.*

de les faire dialoguer les deux niveaux entre eux ou de trouver une gradation intermédiaire et c'est ce point précis qui bloque les testeurs dans leur expérimentation :

> « *Ce qui est perturbant dans l'interface, c'est la dissociation entre les premières pages de dataviz sur l'ensemble du corpus (les graphes sur les notions) et l'interface de recherche après par numéro et par article. En fait, du coup, on perd cette vision d'ensemble du corpus et on est incité à plonger directement avec un degré très fin finalement.* » (testeur 12)

Or, cette perte de vision d'ensemble est gênante pour l'utilisateur, il se retrouve à utiliser un système qu'il comprend mal et qui peut constituer un frein pour son utilisation et les possibilités d'appropriation du dispositif. Mais aussi, et surtout, cela l'empêche de mener à bien ses recherches, ce qui est confirmé par l'expérimentateur suivant : « *On ne sait pas trop ce qu'on interroge* » (testeur 14) ou comme en témoigne le testeur 17 qui s'est posé les deux questions suivantes pendant le parcours guidé : « *Quelles sont les différentes définitions de l'autogestion fournies par les différentes sphères politiques et de logiques (l'anarcho-syndicalisme etc) ? et puis, par ailleurs : Quels étaient dans le corpus tous les documents qui relevaient de l'association "communisme et autogestion" ?* » et avoue ne pas avoir réussi à répondre ni à l'une ni à l'autre de ces deux questions alors qu'il est spécialisé en recherche d'information.

Cela est notamment dû au fait que « *La page n'a pas de sens conceptuellement* » (testeur 17), elle ne constitue pas un ensemble sémantique complet et clos. Un exemple très simple permettant de l'expliquer est le fait qu'un paragraphe ou qu'une phrase puisse commencer sur une page et se terminer sur la suivante. Travailler page par page n'a alors pas de sens ce qui explique en partie les difficultés de certains testeurs à en trouver, comme le disaient déjà Y. CHIARAMELLA et P. MULHEM en 2007 : « *Il serait urgent de développer des modèles intégrant une relation formelle entre structure du document et contenu sémantique.* »[178] Les auteurs ajoutent que l'« *améliorer les performances de la RI ne peuvent qu'être entreprises via un travail préalable de modélisation formelle* » du contenu d'un document mais cela n'est pas permis actuellement par le démonstrateur.

Les testeurs sont alors plusieurs à avoir conclu que le dispositif permettait la lecture de texte mais pas de réelle recherche documentaire ciblée mettant en avant la découvrabilité

[178] Chiaramella, Y. & Mulhem, P. (2007). *Op cit.*

(relative) mais pas la trouvabilité au sein du dispositif. Lecture qui n'est cependant pas optimale à cause des problèmes d'ergonomie, de surcharge cognitive etc que nous avons déjà évoqué.

2) Sur les facettes de recherche classique

Le dispositif propose deux types de facettes de recherche différentes : les facettes habituelles de recherche dont l'entrée notions, personnages, vidéothèque et le moteur de recherche mais également d'entrées dites « *intelligentes* » par le biais des deux algorithmes qui mettent en lien les contenus. Nous avons donc séparé les retours utilisateurs en fonction de cette typologie et commençons par les facettes dites « *classiques* ». Par ailleurs, puisque l'entrée « *vidéothèque* » a été ajoutée au dispositif après l'expérimentation du 9 juin, nous ne disposons d'aucun retour d'utilisateur sur ce point, nous ne l'avons pas intégré aux présentes analyses.

Les notions et mots-clés

L'interface dédiée aux notions a été plutôt appréciée par les testeurs, c'est notamment ce qui est mis en avant par les réponses à la question suivante : « *Quelles interfaces et/ou fonctionnalités vous semblent être les plus abouties ?* » car sur vingt réponses, dix mentionnent les notions. À titre de comparaison, les personnages ne sont cités que deux fois, les fiches trois fois, la revue trois fois, les algorithmes une fois et quatre testeurs n'ont pas répondu ou ont déclaré n'avoir pas d'avis sur la question. L'interface notions est donc celle qui semble connaître le plus grand engouement de la part des expérimentateurs, bien que quelques améliorations soient à apporter. Il a notamment été proposé d'étayer l'interface « *notions* » pour y ajouter une liste alphabétique facilitant la recherche ou encore des nuages de points faisant des liens entre les concepts testeur 15 et testeur 19.

Un des testeurs alerte l'équipe projet en disant : « *Pour un bibliothécaire, un thésaurus sert à indexer par sujet, non par mot-clé (génère du bruit)* » (testeur 16). À la question « *Avez-vous consulté les notions ? Si oui, à partir de quelles interfaces ? Ces informations vous ont-elles été utiles ?* » il explique « *Perturbant : les articles sont rattachés à une entrée de thésaurus par un degré de pertinence parfois très faible (par ex. "utopie" : une occurrence du mot dans un article suffit).* » Le système d'indexation des mots-clés pose ainsi des problèmes de bruit dans les notions mentionnées en bas de chaque page ou dans l'onglet « *notions liées* » de l'article.

En ce qui concerne les listes de mots-clés situées au-dessous des textes, les testeurs étaient en désaccord quant à leur pertinence et à leur utilisation ce qui correspond sûrement aux goûts et habitudes personnelles. Certains n'ont en revanche pas compris la différence entre les deux listes et auraient aimé avoir plus d'explications sur leur élaboration. Il a également été dit que les scores attribués aux mots-clés ne sont pas clairs : *« Je n'ai pas très bien compris le nombre de mots clés et le nombre de mots du thésaurus dans chaque page et les notions liées car on en avait si peu finalement qui étaient liées.* » (testeur 12) ou encore « *l'indice que je n'ai pas compris […] moi j'ai cru que c'était le numéro de page* » (testeur 17). Un manque de clarification qui peut créer de la frustration lors de la consultation c'est pourquoi, le testeur 20 tire la conclusion suivante « *À la limite, il ne faudrait pas les scores* », ce qui rejoint les avis évoqués plus haut sur la simplification de l'interface notamment au niveau de la page.

Les personnages

Dans l'entrée par personnages, les testeurs s'accordent pour dire que la navigation n'est pas simple ce qui ne les aide pas à donner des retours autres que ceux portés sur la complexité de l'interface (surtout pour la première journée de test). Ce phénomène nous permet également de mettre en lumière la difficulté rencontrée par bon nombre de testeurs à évaluer un dispositif alors que les interfaces ne sont pas terminées.

Il ressort de la majeure partie des retours utilisateurs que l'interface personnage est intéressante mais devrait être augmentée pour être réellement efficace, l'ajout d'un moteur de recherche pour la page personnage a été plébiscité par certains afin de retrouver plus aisément les personnes recherchées. D'autres proposent plutôt de diversifier les entrées de cette page pour permettre à chacun d'utiliser la plus efficace pour ses recherches : une entrée par liste alphabétique, une entrée chronologique, une entrée par nombre d'apparitions dans le corpus (à l'image de ce qui a été fait pour les notions) ou encore une entrée par nuage de points permettant de refléter les groupes de pensée. Il manque également une fonctionnalité intéressante dans l'interface « *personnages* » : la possibilité de filtrer les personnages affichés s'ils sont cités dans le corpus, s'ils sont auteurs d'articles ou les deux. Un filtre qui s'avère être important puisqu'il y a plusieurs centaines de personnages mentionnés et qu'il est difficile d'en avoir une vue d'ensemble (« *Il serait bien d'avoir des moyens de tri, pour les personnages, les notions, par importance, par qualité (auteur ou non)* », testeur 6).

Par ailleurs, dans les fiches « *personnages* », les concepts qui leur sont habituellement associés ne ressortent pas ou pas suffisamment, il faudrait retravailler ce point pour permettre une meilleure exploration du corpus à partir de concepts, ce que déplore le testeur 2 :

> « *Les références bibliographiques et liens avec d'autres auteurs sont riches, les informations biographiques sont comparativement assez limitées, les notions associées aux auteurs sont déconcertantes et/ou plates et ne correspondent pas aux notions qui sont habituellement associées à l'auteur et son œuvre* ».

C'est d'ailleurs ce que révèle le questionnaire avec le duo de questions « *Avez-vous consulté les fiches personnages ? Ces informations vous ont-elles été utiles ?* » puisque les avis sont très partagés pour la deuxième partie de la question : 7 personnes ont trouvé les informations utiles, 7 étaient mitigées et 3 n'ont pas jugé les informations utiles (sachant que 3 testeurs ont déclaré ne pas avoir consulté les fiches personnages).

De plus, quand le testeur ouvre un article à partir d'une fiche personnage, il n'y a pas la possibilité de surligner l'entité nommée recherchée dans le texte ce qui manque réellement aux usagers sachant qu'il s'agit d'une fonctionnalité habituelle en recherche documentaire comme l'explique un conservateur des bibliothèques, il est donc déceptif pour beaucoup de testeurs de ne pas la retrouver dans Archival de même que les autres fonctionnalités classiques en recherche documentaire citées précédemment.

Le moteur de recherche

Le moteur de recherche est une fonctionnalité ajoutée tardivement au démonstrateur par l'équipe Archival. À l'origine du projet, l'équipe ne pensait pas implémenter de moteur de recherche dans le démonstrateur pour inciter les testeurs à naviguer au travers des liens et rebonds permis par les algorithmes. Il a cependant été décidé d'ajouter un moteur de recherche au démonstrateur en amont des journées d'expérimentation afin de ne pas frustrer les testeurs et leur permettre d'utiliser les fonctionnalités de recherche dont ils ont l'habitude, au sein desquelles les moteurs de recherche tiennent une place importante. Ce choix tardif s'est donc fait non pas dans l'optique d'offrir une expérience de recherche complète et aboutie grâce au moteur de recherche mais plus pour montrer les possibilités de recherche permises par la combinaison de facettes habituelles de recherche et de facettes « *intelligentes* », pour montrer que l'un ne se substitue pas à l'autre mais que les deux peuvent fonctionner de concert.

Lors des journées d'expérimentation, les testeurs ont toutefois été presque unanimes pour dire que le moteur de recherche doit être amélioré car les résultats qu'il propose ne sont, pour l'instant, pas satisfaisants pour les usagers : « *il faut un moteur de recherche puissant avec une indexation puissante derrière pour que ça puisse donner des résultats pertinents, motivés* » (testeur 9) ce qui est corroboré par un autre testeur : « *La chose qu'on fait quand on va consulter une base de données en général c'est qu'on a déjà une référence et là n'on a pas la possibilité de la mettre et d'être sûr que c'est la première qui sort, là le moteur de recherche ne permet pas ça* ». Ce phénomène est lié à une multitude de facteurs : d'abord la granularité du dispositif qui ne permet pas de filtrer la recherche sur une partie ou l'autre du corpus ; ensuite, le manque de travail sur le moteur de recherche (développé rapidement pour l'expérimentation, il n'a pas fait l'objet d'un travail de fond et ne permet aucune recherche avancée ni utilisation des opérateurs booléens) ; enfin, seuls les articles sont retournés par le moteur ce qui implique que chercher une notion ou un personnage ne permet pas de trouver la fiche correspondante mais un article ou il/elle est cité(e).

Ces problèmes sont particulièrement mis en lumière car le moteur de recherche fait partie des premières entrées choisies par les testeurs comme le rappelle le testeur 9 : « *quand on arrive sur un outil de recherche documentaire et d'aide à la recherche documentaire on attend un moteur de recherche on commence généralement par ça quoi, n'importe qui.* » Il s'agit donc pour beaucoup de la ou d'une des première(s) fonctionnalité(s) essayée(s). Etant relativement peu travaillé en comparaison des autres fonctionnalités du démonstrateur, le moteur de recherche s'est donc avéré être quelque peu déceptif pour une partie des expérimentateurs. Un phénomène qui nous permet de questionner le principe même de faire tester un démonstrateur encore en cours de développement à des expérimentateurs sachant que leur expérience ne sera pas optimale et pourrait être discuté plus amplement dans de prochains travaux.

Il serait également intéressant de lier les résultats à des petits résumés pour justifier leur pertinence par rapport à la requête posée : « *c'est délicat de ne pas avoir des petits résumés ou des petits extraits avec la raison pour laquelle on voit le texte comme pertinent au début qui est une fonctionnalité de base d'accès à une interface documentaire* » (testeur 2), cela aiderait les usagers à comprendre le choix des résultats (pourquoi ils ont été retournés par le système, leur classement, etc).

En résumé, les facettes de recherche sont utiles et plaisent en majorité aux testeurs mais elles leur semblent parfois inabouties. Ce phénomène peut s'avérer problématique pour

quelques testeurs car ils sont habitués à trouver ce type de fonctionnalités dans les autres interfaces documentaires qu'ils les utilisent habituellement, ils sont alors critiques au regard de ce qu'ils connaissent et des manques et limites qu'ils perçoivent. La satisfaction des testeurs sur ces facettes varie alors beaucoup en fonction des individus car si certains ont admis qu'il s'agissait d'un test et d'un démonstrateur en cours de construction (« *j'estime que le système est productif dans le sens où il m'a permis d'entrer très vite dans le corpus et d'en exploiter des facettes que je n'aurai pas forcément appréhendées ou ressenties à priori* », testeur 1 ou **«** *pour une méthode exploratoire c'est très pratique.* », testeur 10), d'autres sont plus critiques et mettent en avant les lacunes de ce type de dispositif sur des fonction jugées habituelles et ont eu des difficultés à dépasser cela. Cela renouvèle les questionnements au sujet de la méthode d'évaluation proposée auprès de testeurs alors que le dispositif proposé est encore en cours de développement.

3) Sur les apports de l'intelligence artificielle

La question du public cible est notamment très importante pour juger de la pertinence des résultats proposés par la machine. Dans le questionnaire, les testeurs ont été sondé sur ces liens générés par algorithmes au travers de nombreuses questions dont « *Quelle est votre appréciation sur la pertinence des liens générés automatiquement ?* » qui a obtenu 2 réponses positives, 7 plutôt positives, 4 mitigés, 2 plutôt négative, 1 négative, 4 sans avis ou sans réponse. Les testeurs ont donc jugé la pertinence des liens plutôt correcte mais les avis sont très disparates.

Certains sont favorables à l'introduction de liens générés par intelligence artificielle dans le corpus documentaire (« *plutôt favorable* » testeur 1, « *extra* » testeur 4, « *globalement intéressant* » testeur 6), d'autres sont plus mitigés (« *intérêt modéré* » testeur 2, « *très variables selon les cas* » testeur 13) et aimeraient voir des améliorations pour se prononcer davantage tandis que les derniers n'ont pas particulièrement apprécié (« *faible pour les questions, moyen pour la similarité* » testeur 16, « *décevant pour un travail plus fin sur le contenu du texte* » testeur 18). On distingue cependant majoritairement des réponses positives parmi le panel de chercheurs et des réponses plus modérées voire négatives pour les professionnels de la documentation bien que ce ne soit pas une règle absolue.

Il est important de noter que le problème de granularité soulevé précédemment entrave grandement la possibilité de générer des questions pertinentes et des liens qui peuvent intéresser

le lecteur sans trop l'éloigner de sa recherche. Le testeur 17 soulève ce problème dans la question « *Ces 2 méthodes algorithmiques vous paraissent-elles complémentaires et/ou adaptées à des usages différents ?* » en répondant :

> « *En théorie, oui, mais il serait intéressant de pouvoir définir nous-mêmes le niveau de granularité (une section, une page ou l'article entier). Pour ma part, je regrette d'être limité à la page car celle-ci n'est pas toujours pertinente du point de vue du sens (elle l'est même rarement).* »

Ce témoignage permet de comprendre l'un des problèmes principaux des algorithmes : ils fonctionnement sur une portion de texte sélectionnée sans prise en compte aucune du contexte. Ainsi, les questions et lien générés sont sur quelques mots, une page tout au plus (dû à l'impossibilité d'en sélectionner plusieurs) sans prise en compte de l'article du numéro ou de la partie dans lequel le lecteur se situe.

Sur l'algorithme de questions/ réponses

Nous avons expérimenté ce problème avec un cas pratique, prenons l'article « *L'autogestion industrielle en Algérie* » par Damien HELIE dans le numéro de septembre-décembre 1969. En sélectionnant le premier paragraphe de l'article de « *Pendant toute la période* » à « *tous ces problèmes* », onze questions sont générées avec plusieurs liens et renvois chacune, la majorité des liens poussent le lecteur à sortir totalement du contexte dans lequel il effectue ses recherches, à savoir l'autogestion et l'Algérie.

L'autre problème majeur des questions est celui de la pertinence pour l'usager. Nombreux sont les testeurs à avoir trouvé les questions incompréhensibles, chronophages voire totalement inutiles, en témoigne le florilège de remarques à ce sujet : « *ça reste un peu superficiel. "Je vais bien. Comment il va ? Et bah il va bien !"* » (testeur 9), « *Il y a des trucs que je ne comprends vraiment pas, les questions sont débiles* » (testeur 19) ou encore :

> « *Les questions étaient souvent extraordinairement artificielles.* […] *Les questions ça m'a rappelé les questionnaires militaires avec des questions absurdes : de quoi sont les pieds ? La réponse était : "les pieds sont l'objet de soins constants de la part des soldats"* » (testeur 3)

Le testeur 3 aurait préféré que l'algorithme lui propose une succession de concepts plutôt que des mots-clés, il continue : « *Je n'ai pas vu l'intérêt, parfois j'ai même vu des contresens, c'est-à-dire qu'il y avait des questions qui étaient liées à des passages dans le texte*

et c'était purement un contresens. » C'est également ce que nous avons constaté au cours de nos recherches sur la plateforme, en témoigne la question « *Qu'apporte l'idéologie nationaliste ?* » générée à partir de la phrase « *L'idéologie nationaliste n'apporte pas de réponse à tous ces problèmes* » et dont seule la fin de la phrase a été utilisée par l'algorithme à savoir « *de réponse à tous ces problèmes* » [voir illustrations n° 57 et 58].

Le testeur 13 a également mis en avant le fait que les questions générées semblent meilleures au début et à la fin du texte comme il l'explique : « *J'aurais tendance à penser qu'on aurait une question plus pertinente si on va taper dans le début de l'article, dans l'introduction ou dans la conclusion parce que c'est l'endroit où les concepts traités sont plus ramassés.* » ce que nous pouvons confirmer au regard de nos test personnels mais qui devrait être soumis à plus de tests.

Sans compter que les questions ne sont pas toutes écrites dans français correct *« pb de syntaxe des questions* » (testeur 16), « *Le module de génération de questions n'a rien à voir sur le plan syntaxique… J'ai l'impression de lire mes étudiants* » (testeur 17). Pour l'article cité précédemment, l'algorithme pose entre autres les questions suivantes : « *Sur quoi a été mis l'accent ?* », « *De quoi s'affranchit tout pouvoir sorti du peuple ?* » ou encore « *Qui veulent placer leur pays dans la voie du développement ?* ». Le problème étant que ces incorrections de langue ont bloqué certains testeurs et les ont poussé à rejeter ce premier algorithme alors même que l'intérêt du programme n'est pas de générer des questions parfaites mais de permettre la mise en relation de contenus par le biais de questions générées automatiquement. Les testeurs ont cependant raison en mettant en avant l'importance d'améliorer la qualité syntaxique et la formulation des questions mais cela ne joue pas sur la qualité du lien et de la mise en relation de contenus proposés.

Ainsi, le module questions/réponses a obtenu des avis mitigés dont une partie sont très négatifs, justifiés par les propositions actuelles de l'algorithme qui sont très inégales et souvent mauvaises. Les testeurs ont alors largement remis en question ce module, comme l'exprime le testeur 9 : « *on se pose après la question, il est où le plus à part que ça aide vraiment à faire une synthèse. Qu'est-ce que ça apporte pour la recherche documentaire en interne et en externe ?* »

Il semble pourtant que les testeurs soient en grande majorité passés à côté de l'objectif réel du premier algorithme, ce dernier génère des questions certes mais il ne s'agit là que d'un moyen permettant la mise en lien de contenus. Les testeurs ont ainsi critiqué la forme que prend

le lien à savoir la question plus ou moins bien formulée mais très peu d'entre eux ont donné un avis sur le lien proposé. C'est le testeur 6 qui a le mieux compris et exprimé cette dualité entre la question générée et le lien proposé, il dit : « *Sur les questions, je trouvé que le fond de ce que je lisais était intéressant, cad que en cliquant sur les questions je trouvais des informations qui parlaient vraiment du sujet et apportaient vraiment des informations complémentaires* » mais ajoute :

> « *par contre c'est pas forcément évident que c'est ça qu'on va trouver derrière. La première question elle est toujours un peu bizarre, et on a l'impression que c'est genre un truc presque enfantin qui répond au texte comme si on était en guide de lecture en cours de français. Et donc du coup, c'est pas évident de se dire qu'on cliquant dessus pour le coup on va arriver à une réelle information.*»

Un travail de nettoyage de la syntaxe des questions générées est en cours et devrait permettre une amélioration des questions retournées par le système et pourrait modifier positivement les jugements émis par les futurs testeurs quant à la pertinence de cet algorithme. Il pourrait également être intéressant de plus mettre en avant l'objectif de cet algorithme soit la mise en relation de contenus ou de documents et non la génération de questions en elle-même.

Sur l'algorithme par similarités

L'algorithme 2 fait également l'objet de critiques positives comme négatives, les problèmes étant notamment liés à la question de la granularité du dispositif sur laquelle nous ne revenons pas.

Les points positifs soulevés par les testeurs sont : la démarche d'explicabilité du système (« *Je suis très fan de l'idée de rendre explicites les mécanismes de rapprochement par similarité* » testeur 13, « *mention spéciale pour "explicabilité par inférence" qui permet d'ouvrir et découvrir cette blackbox* » testeur 11, « *La transparence de l'algo et du périmètre documentaire sont effectivement deux éléments clés pour avoir la sensation de garder le contrôle sur les résultats de sa recherche.* » testeur 14) ou encore :

> « *oui, c'est hyper important que l'algo m'explique comment il a réfléchi, notamment pour savoir dès le début si les questions liées seront pertinentes. Beaucoup d'algo tentent de se montrer plus intelligents qu'ils ne le sont en n'expliquant pas leur fonctionnement, mais du coup, leur réponse a moins de valeur, de crédibilité* » (testeur 5).

Mais aussi l'aspect de découvrabilité du corpus permis par les similarités : « *Après sur les articles, c'était très intéressant cette possibilité de sélectionner et identifier une partie du texte pour voir ce que l'algorithme nous montre comme suggestion, ça c'est vraiment intéressant.* » (testeur 10). L'un des testeurs trouve pourtant que « *au-delà d'un rapprochement par similarité, le lien n'est pas éclairant sur le contenu.* » (testeur 12) remettant ainsi en cause le modèle d'algorithme proposé. Les limites sont notamment liées à l'élaboration d'un corpus homogène et clos car le système retourne beaucoup de bruit comme l'explique un des experts de la documentation « *La similarité est utile avec des limites aussi (échelle réduite, bruit)* » (testeur 16).

En le comparant à l'algorithme précédent, on se rend également compte du fait que le principe des similarités semble moins faciles à comprendre que celui des questions pour les usagers. Si l'on met en regard la question « *Est-ce que les questions associées aux fragments de texte mis en relation par l'algorithme 1 (génération de questions) vous ont aidé à appréhender le lien établi ?* » qui obtient 8 oui, 1 plutôt oui, 2 mitigés, 4 plutôt non, 4 non, 1 sans réponse et la question *« Est-ce que les mots surlignés dans les fragments de texte mis en relation par l' algo 2 vous ont aidé à appréhender le lien établi ?* » qui comptabilise 5 oui, 3 plutôt oui, 4 mitigés, 1 plutôt non, 4 non, 3 sans avis ou sans réponse ; on en comprend que - bien que les résultats soient très mitigés, la plupart des testeurs trouvent les questions générées par l'algorithme 1 plus faciles à appréhender que les liens générés par l'algorithme 2.

Il est également important de mettre en lumière le fait que les articles proposés dans le module similarités sont très souvent les mêmes y compris quand le testeur change d'article, de thème, d'année, d'auteur, etc. Ce phénomène s'est vraiment beaucoup reproduit lors de l'exploration du testeur 18, il a donc commencé à ouvrir des articles très différents pour systématiquement aller vérifier les similarités proposées et tirer la conclusion suivante : « *Ce n'est pas très fin comme travail algorithmique* », il ajoute « *l'algorithme a des promesses qui ne sont pas vraiment tenues* » révélant sa déception et son incompréhension vis-à-vis de ce second algorithme.

Les algorithmes ont des biais, qui peuvent être liés au corpus qu'ils traitent, à leurs données d'entraînement, à des décisions humaines incorrectes etc. C'est notamment ce que mentionne J. MARTINEAU :

> « *Par ailleurs, puisque leur apprentissage repose sur des données et des décisions humaines parfois imparfaites, ces systèmes peuvent reproduire et*

> *amplifier les biais et les performances au détriment de certains groupes marginalisés. C'est ce qu'on appelle la "discrimination algorithmique".* »[179]

Dans le cas d'Archival, un des biais notoires est le fait que le corpus soit très homogène, trop pour permettre un réel travail de similarités tant les liens entre tous les documents sont étroits. G. ROUET mentionne également ce problème dans son article « *Démystifier les algorithmes* » : « *Des corrélations, même illusoires, peuvent ainsi être utilisées alors même que les événements mesurés sont indépendants*. »[180] Ce système de similarités fonctionnerait nettement mieux sur des collections plus vastes et variées car elles seraient moins abusives et plus pertinentes. Cependant, repenser la granularité du dispositif permettrait de modifier les similarités entre documents, un travail intéressant mais de longue haleine : « *Les documents d'un espace ne sont pas autonomes […]. En fait, le problème est de déterminer pour les documents, ou les parties de documents, les relations qu'ils entretiennent pour les utiliser lors de l'indexation et de la recherche.* »[181] ; leurs auteurs déploraient alors :

> « *Il manque des efforts de fond pour déterminer comment des documents peuvent être liés (de manière explicite à la construction ou non), comment prendre en compte ces éléments lors de l'indexation, et comment les prendre en compte au niveau du traitement des requêtes.* »[182]

III/ Discussion et recommandations

Maintenant que les résultats obtenus auprès des testeurs ont été présentés et expliqués, il temps de les discuter afin d'en tirer des recommandations pour les expérimentations futures et identifier la portabilité de la méthode pour d'autres protocoles.

1) Discussion des résultats

Au regard de notre problématique de départ, à savoir élaborer une méthode d'évaluation d'Archival en tant qu'interface documentaire « *intelligente* » permettant d'évaluer

[179] Martineau, J. (2023). *Op cit.*
[180] Rouet, G. (2019). *Op cit.*
[181] Chiaramella, Y. & Mulhem, P. (2007). *Op cit.*
[182] *Ibid.*

l'appropriation du dispositif par les usagers, leur compréhension et leur jugement de la pertinence des liens générés automatiquement par algorithmes, il semble que le contrat soit rempli. Les résultats sont, en effet, nombreux, plutôt diversifiés et très éclairant sur la façon dont les testeurs ont perçu le démonstrateur et le projet.

L'interface graphique en elle-même plaît en très grande majorité mais étant encore en cours de conception, de nombreux problèmes d'ergonomie et de design ont dérangé voire bloqué les testeurs dans leur expérimentation révélant ainsi la difficulté pour un certain nombre d'entre eux d'expérimenter une plateforme encore en version *beta*. L'ergonomie et la navigation pâtissent des mêmes problèmes, il y avait donc une barrière à l'entrée dans la réalisation de ces expérimentations, à savoir être en capacité de suffisamment maîtriser les outils informatiques ou faire preuve d'abstraction et d'imagination pour dépasser le cadre de l'interface encore très limité.

Pour l'appropriation du dispositif, les résultats sont bien différents car bien que les testeurs soient nombreux à déclarer ne pas s'être approprié le système ou ne pas avoir réussi à le faire en avançant le fait qu'il s'agit d'un manque de connaissances sur le corpus ou à cause de l'état des différentes interfaces. Ils ont pourtant, dans la majorité des cas, négocié les utilisations du dispositif pensées par l'équipe Archival. Un phénomène qui pourrait être dû à la spécialisation des testeurs, ils ont l'habitude de travailler sur des interfaces documentaires -qu'il s'agisse des chercheurs en SIC ou des professionnels des bibliothèques- et sont familiers des différentes facettes et outils de recherche documentaire proposés (hormis les algorithmes). Si certains testeurs ne se sont pas approprié le dispositif en lui-même, ils se sont tous ou presque approprié une ou plusieurs briques qui le composent, la plupart du temps sans s'en rendre compte. Puisque tous les testeurs n'ont pas été observés, il est possible que certains détournements d'usage n'aient pas été repérés par l'observation.

Les résultats sur le contenu sont tout aussi éclairants notamment car les testeurs ont eu beaucoup de difficultés à comprendre le corpus documentaire interrogé, perdus entre la revue et les liens internes et externes -ce qui n'a certainement pas été arrangé par le fait que la revue porte trois noms différents successifs-. Ce manque de compréhension du contenu du démonstrateur est notamment révélateur d'un manque de rigueur dans la présentation documentaire : les numéros des articles ne sont pas cités, la plupart du temps ils sont affichés sans sources, les renseignements fournis à l'utilisateur ne sont généralement pas suffisants (pas de date ou pas d'auteur), tout cela place donc le lecteur face à un texte sans qu'il dispose de suffisamment d'éléments pour le situer. Ce problème est également lié à la granularité du

corpus, déroutant pour les testeurs qui ne peuvent naviguer que de manière très large dans la revue, les notions, les personnages et vidéos ou lire précisément un article page à page. Enfin, les liens sont nombreux (dans le numéro, hors du numéro, dans la revue ou hors de la revue) et perdent l'usager, il faudrait améliorer les listes de résultats et la description de ces derniers pour mieux mettre en avant leur provenance.

Les facettes de recherche ont été jugé intéressantes mais pas novatrices par les testeurs qui sont habitués à ce type d'entrées et qui les ont jugés -pour beaucoup- trop peu abouties ou incomplètes. L'entrée par notions est la plus appréciée, celles par personnages et par la chronologie ont été jugées plus sévèrement. Puisqu'il ne s'agissait pas du cœur de la démonstration du prototype Archival, ces facettes ont été développées à titre d'indication de ce qui pourrait être fait, ce qui est rappelé par l'équipe au cours des entretiens. Chaque interface se concentre alors sur une fonctionnalité particulière : la liste pour les personnages, l'histogramme pour les notions et la chronologie pour la revue. Si cela a perturbé certains testeurs qui auraient aimé retrouver les différentes fonctionnalités dans chaque entrée, il s'agit tout de même d'une façon pertinente de dévoiler les possibilités du dispositif.

Enfin, il est intéressant de noter que si les facettes « *intelligentes* », soit les liens et questions générés par les deux algorithmes, constituent le cœur du démonstrateur pour l'équipe Archival, cela n'a pas été le cas pour les testeurs qui les ont considérées comme les autres fonctionnalités, certains ne les ont d'ailleurs presque pas utilisées et ont préféré les autres modules du dispositif. Cela s'est avéré amplifié par le fait que les algorithmes n'ont pas d'entrée dans le menu latéral, il faut donc ouvrir un article puis sélectionner du texte pour les afficher, une démarche bien trop complexe pour que l'usager arrive à trouver le module facilement et l'utilise réellement comme une fonctionnalité du dispositif. Enfin, le jugement de la pertinence des questions et liens générés par les testeurs est plutôt déceptif car les plus enthousiastes disent que les algorithmes sont prometteurs, les testeurs plus critiques expliquent qu'ils sont aléatoires et les derniers les jugent mauvais voire inutiles. Un constat plutôt négatif mais qui est en partie lié aux problèmes de granularité du dispositif ne permettant pas un travail efficace de la part des algorithmes. Permettre à l'usager de changer la granularité de la recherche de façon plus ou moins précise en fonction de sa demande et de ses besoins dévoilerait de bien meilleurs résultats, probablement pas parfaits toutefois.

Il serait probablement pertinent de creuser l'analyse des résultats des campagnes TREC dédiées aux systèmes de questions-réponses en vue d'évaluer le module Q/A (l'algorithme question-réponse) d'Archival.

Le prototype a donc pu être évalué grâce au protocole d'expérimentation mis en place auprès des vingt testeurs. Ainsi, le dispositif de lecture fonctionne indéniablement mais il n'est pas efficace pour autant à cause des problèmes de surcharge cognitive évoqués, de l'impossibilité de mettre le texte en pleine page ou la présence de « *pollution* » dans la lecture notamment à cause des mots-clés. La lecture numérique mériterait donc que l'on y porte plus attention, comme le préconise les ouvrages dédiés[183]. La problématique de la lecture « *augmentée* » ou « *enrichie* » pourrait également être éclairée au regard des ouvrages et articles sur le sujet[184].

En tant que dispositif de recherche d'information en revanche, les résultats sont plus nuancés et il convient de rappeler que la RI peut concerner une recherche ciblée, guidée par un but (la recherche d'information, *information retrieval*) ou être guidée par l'exploration (la recherche de renseignements, *fact retrieval*)[185]. Il s'agit donc de deux démarches bien différentes. Archival met alors l'accent sur la découvrabilité des contenus soit la revue *Autogestion* et tous les liens qui y sont associés depuis CanalU, l'Ina, la BnF, Wikipédia, etc. Découvrabilité qui a été jugée très relative par les testeurs, d'abord parce que certains ont rencontré des difficultés à naviguer, d'autres ont peiné à juger la découvrabilité sans connaître le corpus, d'autres encore ont manqué de motivation car aucun but ni objectif ne guidait leur recherche, etc. Ceux qui connaissaient la revue ou du moins le thème ont jugé ce point de façon plus clémente dans la majorité des cas trouvant qu'effectivement, le dispositif peut permettre la découvrabilité des contenus.

La trouvabilité de l'information et du document sont, en revanche, bien plus problématique car il difficile de trouver un article du corpus avec le moteur de recherche, il faut alors le chercher manuellement dans la liste des résultats année par année par exemple. La trouvabilité n'est certes pas le point sur lequel l'équipe a souhaité mettre l'accent, mais une interface documentaire doit permettre d'accéder au document ce qui n'est pas le cas dans

[183] Baccino, T., Drai-Zerbib, V. (2015). La lecture numérique. Presses universitaires de Grenoble. https://doi-org.ressources-electroniques.univ-lille.fr/10.3917/pug.bacci.2015.01 ; Barbagelata, P., Inaudi, A. & Pelissier, M. (2014). Le numérique vecteur d'un renouveau des pratiques de lecture : leurre ou opportunité ?. Études de communication, 43, p. 17-38. https://doi-org.ressources-electroniques.univ-lille.fr/10.4000/edc.5965 ; Micheau, B. (2016). Saemmer Alexandra : Rhétorique du texte numérique : figures de la lecture, anticipations de pratiques. Études de communication, 46, p. 201-206. https://doi-org.ressources-electroniques.univ-lille.fr/10.4000/edc.6556

[184] Laborderie, A. (2020). Le livre augmenté : un nouveau paradigme du livre ?. Revue de la BNF, 60, p. 148-159. https://doi-org.ressources-electroniques.univ-lille.fr/10.3917/rbnf.060.0148 ; Francart, T. (2016). Des textes augmentés avec les données du Web. I2D - Information, données & documents, 53, p. 45-45. https://doi-org.ressources-electroniques.univ-lille.fr/10.3917/i2d.162.0045

[185] Lancaster, F. W. (1979). *Op cit.*

Archival qui s'avère être extrêmement complexe. Il faudrait donc retravailler cette notion de trouvabilité du document et de l'information[186]. Malgré l'évaluation, il reste donc de nombreux questionnements et problématiques en suspens.

2) Recommandations pour les expérimentations futures

Les deux journées d'expérimentation divisées en trois sessions distinctes de tests nous ont permis d'éprouver notre méthode et de la faire évoluer en vue de l'améliorer. Cela nous a également permis de mettre en lumière les points à perfectionner ou à revoir. Cependant, si cette démarche agile nous a aidé à mieux cibler les attentes de l'équipe projet et les possibilités d'expérimentation auprès des testeurs, les expérimentations futurs devraient suivre un cadre plus homogène et mieux défini en amont pour permettre une meilleure comparaison des résultats entre les sessions. Un point également lié au fait que le démonstrateur était encore en cours d'évolution entre les journées d'expérimentation et que ces modifications nous ont poussé à nous adapter.

Sur les panels

La première chose qu'il nous semble important de mentionner est qu'actuellement, les testeurs étaient des chercheurs en SIC et des professionnels des bibliothèques, aucun chercheur dans d'autres sections des SHS n'a été mobilisé, qu'il s'agisse d'historiens, de sociologues, de philosophes, de politologues, etc. Il y a donc un biais dans la création des panels de testeurs actuels car le dispositif a été pensé et créé par les SIC et l'informatique pour les chercheurs en SHS et dont la représentativité du panel est actuellement limitée. Les expérimentations futures devraient donc s'attacher à élargir le panel aux différentes disciplines concernées par le corpus. Cela devrait également permettre de faire expérimenter le démonstrateur par des publics plus vastes et hétérogènes mais également avec des appétences en informatique également plus variables. Ce genre de test devrait également permettre de faire émerger des cas d'usages et pratiques différentes de ce qui a été constaté le 15 mai et le 9 juin 2023.

Par ailleurs, il est important de noter que, si les panels étaient composés d'experts de différents horizons, quelques-uns connaissaient le thème de l'autogestion mais seul un d'entre eux était spécialiste du sujet. Le thème de la revue est pourtant central dans le dispositif comme nous l'avons démontré précédemment. Les réponses aux questionnaires dévoilent d'ailleurs une

[186] Lipsyc, C. & Ihadjadene, M. (2013). Architecture de l'information et éditorialisation. Études de communication, 41, p. 103-118. https://doi-org.ressources-electroniques.univ-lille.fr/10.4000/edc.5406

connaissance très limitée ou inexistante du sujet pour la plupart d'entre eux : 12 testeurs sur 20 avouent ne pas connaître l'autogestion, 4 affirment ne connaître qu'un peu et les 4 derniers connaissent le thème. Il serait donc intéressant, pour les expérimentations futures, d'interroger des experts du domaine qui seraient alors plus portés sur le contenu de la revue pour obtenir des retours complémentaires à ceux du 15 mai et du 9 juin 2023.

Sur l'observation

Quelques ajustements sont ainsi nécessaires pour améliorer ces procédés tant dans la réalisation que dans l'exploitation. Nous avons pratiqué deux observations différentes entre les trois sessions de test car nous n'avions observé que de façon générale pendant la première session et avons hybridé les deux types d'observation pour les deux autres sessions. Puisque l'hybridation des observations s'est avérée fructueuse les deux dernières sessions, il convient donc de noter que cette double observation devrait être mise en place au cours des futurs tests. De plus, lors de la première journée de test, l'observation aurait mérité d'être plus préparée afin que chaque observateur soit familier avec les attentes, les objectifs, les grilles et que les pratiques soient plus harmonisées. Il faudrait alors former tous les observateurs à cette pratique de manière uniforme en amont du test, un temps qui permettrait également d'en expliquer les enjeux car il s'agit d'une démarche qualitative qui nécessite de s'attacher au ressenti des testeurs, chose que nous avons pu faire pour la deuxième journée de test et qui s'est avéré concluante.

Enfin, nous avons eu un nombre de testeurs très diversifié entre les trois sessions puisqu'ils étaient 10 pour la première, 6 pour la seconde et 4 pour la dernière. Or, il se trouve qu'il est plus facile de se coordonner pour le déroulé de l'expérimentation et pour le remplissage des grilles d'observation en petit comité, les deux dernières expérimentations étaient donc plus fructueuses et agréables, le nombre de testeurs idéal se situe alors entre 4 et 6. Nous aimerions également pouvoir placer un observateur par testeur, lui demander de raconter son parcours, ce qu'il souhaite faire, s'il met en place une stratégie de recherche et ses éventuelles difficultés en enregistrant le tout pendant sa navigation. Cela pourrait être un réel gain pour l'enquête car les retours instantanés des testeurs sont bien plus proches de ce qu'ils pensent vraiment que leurs réponses aux questionnaires ou leurs prises de parole à l'oral à la fin de session qui sont reformulées et souvent édulcorées. Les testeurs se permettent plus de critiques lors de la navigation mais ces retours sont quelque peu difficiles à capter pendant l'observation suivant la méthode actuelle car il faut arriver à suivre deux parcours et les testeurs n'osent pas forcément parler de peur de couper l'autre. Il peut également y avoir des déséquilibres si l'un est bavard

et l'autre plus introverti ce qui complexifie la méthode d'observation. Cette possibilité se ferait, en revanche, à condition de doubler le nombre d'observateurs actuels pour en avoir un par testeur, solution qui est difficilement envisageable car il faut bien connaître la plateforme pour observer des testeurs et il est préférable d'avoir pratiqué l'observation ou du moins d'en connaître les objectifs ce qui nécessite un temps d'adaptation ou de formation préalable.

Sur le questionnaire

Le questionnaire est composé de plus d'une trentaine de questions, il s'agit donc d'un questionnaire assez long et conséquent, d'autant plus que les champs nécessitent réflexion et rédaction de la part du répondant. Le temps proposé pour y répondre a alors été jugé trop court par bon nombre de testeurs surtout lors de la première journée de test, correct selon les autres. Le rallonger pour la deuxième journée de test s'est avéré bien plus confortable pour les testeurs qui ont eu la possibilité de développer leurs réponses, le temps pourrait encore être allongé mais cela risquerait de rendre l'expérimentation très longue voire trop longue.

La lecture du tableau Excel généré comportant les réponses des testeurs permet également de prendre conscience de la difficulté de répondre à autant de questions ouvertes pour les testeurs à cause du manque de temps (surtout pour la première journée de test). Un problème renforcé par le fait que plusieurs testeurs expriment leur incapacité à donner un avis par manque de temps d'expérimentation, un phénomène qui se reproduit sous de nombreuses questions : « *Difficile à éprouver en si peu de temps* » (testeur 1), « *Il faut plus de temps pour être pertinent* » (testeur 8), « *c'est un peu tôt pour avoir un avis* » (testeur 13) ou encore « *il me manque un peu plus de temps pour bien explorer* » (testeur 10).

Sur l'entretien collectif

Ces retours ont été très fructueux au cours de chaque session mais, là où les testeurs de la première session avaient environ 3 minutes de parole, les testeurs de la deuxième expérimentation ont eu près du double et les testeurs de la troisième session ont eu entre 5 et 15 minutes de parole chacun ce qui leur a réellement permis de développer leur point de vue de façon plus exhaustive et précise. Ce temps de parole en fin d'expérimentation est nécessaire à l'enquête car les testeurs parlent de ce qui les a marqués, ce qu'ils ont aimé ou qui les a déplu et leur permet de répondre librement alors que le questionnaire est contraint. Cela a également permis à l'équipe Archival et associés d'intervenir, de répondre aux questions, d'en poser et d'alimenter le débat, etc. Les testeurs ont d'ailleurs eu l'occasion d'intervenir plusieurs fois pour certains afin d'apporter des compléments à ce qui avait été dit ce qui a nourri davantage

les échanges. Leur donner à chacun un vrai temps de parole s'est avéré nécessaire et très intéressant pour construire les présentes analyses, il convient donc de ne pas le rogner.

Il semble important, au regard des transcriptions des échanges, de noter que les échanges de la deuxième et de la troisième expérimentation nous paraissent plus profitables à la fois pour les testeurs qui n'ont pas été brusqués et ont pu exprimer tout ce qu'ils avaient à dire mais également pour l'équipe projet car les membres ont pu poser des questions sur les améliorations à venir (« *Peut-être que c'est pas un bon choix au départ de faire un site web ?* », équipe Archival), les explications sur la façon dont le projet a été pensé (« *C'était un peu alors donc la manière dont nous on a travaillé* », équipe Archival), les défauts actuels (« *la problématique de lecture ce n'est pas évident* », « *Après moi j'ai peur -et je ne sais pas ce que vous en pensez- parce que c'est déjà très dense, très difficile à lire* », équipe Archival) mais également de demander un avis précis (« *sur la génération des liens, tu as pu tester un peu L. ?* », associés, « *Vous aurez souhaité retrouver plus d'infos ?* », équipe Archival). La deuxième et la troisième session ont permis plus de dialogue et se rapprochaient en cela plus d'un réel entretien collectif que la première.

Sur les parcours proposés

Il faudrait également revoir le déroulement et en particulier les parcours proposés aux testeurs. Nombreux sont ceux qui étaient perdus, manquaient de motivation pour la recherche ou trouvaient simplement que le parcours dit « *guidé* » était en réalité un second parcours libre mais avec des articles d'entrée suggérés. Il serait ainsi possible de proposer préalablement un parcours libre comme ça a été le cas et d'ensuite proposer un parcours qui s'apparenterait à un scénario de recherche plus poussé voire très détaillé. Cela permettrait également à tous les testeurs de parcourir l'ensemble des interfaces et des fonctionnalités car les deux parcours proposés actuellement les laissent tellement libres qu'il est possible pour un testeur de passer à côté des algorithmes.

Avant chaque session de test, Samuel DA SILVA (développeur du projet) a pris le temps de présenter le démonstrateur et d'en expliquer les fonctionnalités plus ou moins en détails en fonction du temps qui lui était accordé dans le déroulement de l'expérimentation. Lors de la première journée, les testeurs ont mis en avant les aspects positifs de cette démarche, leur permettant de comprendre le projet, le démonstrateur avant de s'y plonger eux-mêmes. Lors de la seconde journée de test en revanche, deux testeurs sur quatre ont mis en avant le fait que ces explications étaient trop précises et biaisaient l'expérimentation des testeurs avant même qu'ils

n'arrivent sur le dispositif et n'explorent par eux-mêmes. Cela se ressent dans les réponses au questionnaire du testeur 18 :

> « *L'interface vous a-t- elle aidée à comprendre les fonctionnalités algorithmiques ? Expliqué oralement avant l'usage donc difficile de juger.*
>
> *L'explication qui vous a été donnée sur le fonctionnement de chacun des algorithmes vous semble-t-elle suffisante ?*
>
> *Expliqué oralement avant l'usage donc difficile de juger.* »

Cela est notamment dû au fait que nous ayons allongé le temps alloué à cette présentation initiale de 7 à 15 minutes, nous savons donc désormais que si ce temps est nécessaire, il ne faut pas qu'il soit trop long ou détaillé au risque de couper l'herbe sous le pied des testeurs avant le début de leur découverte.

Sur la méthode

La méthode d'enquête a été voulue et pensée par l'équipe Archival et réalisée avec la collaboration des associés, la diversification des acteurs a amené à de nombreux compromis qui ne nous ont pas permis de mettre en place tout ce que nous aurions souhaité (comme l'observation ciblée sur la première session par exemple). Pour pallier ce problème, l'évaluation devrait être réalisée par un groupe d'acteurs moins nombreux -tout en restant pluridisciplinaire- et avec une partie de testeurs plus spécialisés sur la question. De plus, l'évaluation devrait être laissée à des acteurs extérieurs au projet de l'élaboration de la méthode à la réalisation d'expérimentations pour plus de neutralité notamment en termes de retours, ce qui a été observé dans la différence entre les retours faits par les testeurs invités par l'équipe Archival et ceux invités par les associés. Enfin, il faudrait prendre plus de temps et de recul en amont pour puiser dans la littérature sur les protocoles existants notamment en termes d'évaluation UX des interfaces, en ce qui concerne les enquêtes et les études sur les publics.

3) Portabilité de la méthode d'évaluation

Quelles perspectives pour Gallica ?

Gallica, en tant que bibliothèque numérique, est un dispositif d'accès à l'information. La refonte de sa page d'accueil et de ses modalités de recherche va avoir lieu en 2024 et se prépare dès maintenant. Le laboratoire GERiiCO sera alors de nouveau associé à la BnF dans

ce projet afin d'évaluer les possibilités de recherche sur les interfaces, la performance du système, les avancées permises ainsi que les points à améliorer. La méthode d'évaluation, telle que nous l'avons pensée actuellement, est un modèle *ad hoc* conçu pour Archival en collaboration avec l'équipe projet. Cela implique que nous n'avons pas eu la main sur l'ensemble du processus d'évaluation – nous n'avons pas évalué la performance du système, ni les logs des testeurs et n'avons pas choisi et rédigé les questions soumises lors des questionnaires - et avons surtout joué un rôle de conseil et d'observation. Pour l'évaluation de Gallica, il en sera autrement puisque la BnF et le laboratoire GERiiCO seront en charge de mener le protocole dans son intégralité.

La méthode d'expérimentation composite telle que pratiquée pour Archival avec une enquête hybride nous semble prometteuse. Il est envisageable de réutiliser ce principe d'expérimentation en l'adaptant à Gallica et ce auprès de publics diversifiés. Les parcours devront, en revanche, être revus en particulier le parcours dit « *guidé* » qui ne l'est pas réellement dans le protocole d'expérimentation Archival. Il faudrait, pour Gallica, mettre sur pied un scénario de recherche adapté à différents publics cibles à l'image de ce qui est fait dans les vidéos explicatives sur le fonctionnement de la plateforme disponibles en libre accès sur Youtube[187]. Plusieurs scénarios pourraient être imaginés en fonction des publics interrogés : chercheurs, Gallicanautes, grand public, enseignants, historiens, politologues… Il faudra également s'intéresser à l'autre versant de l'évaluation, à savoir l'évaluation quantitative et performative de la recherche en nous basant sur des protocoles et métriques tout en la combinant à l'évaluation qualitative.

Un travail qui pourrait également être facilité par rapport à Archival car Gallica dispose d'ores et déjà d'une bibliographie dédiée[188].

---

[187] Tutoriel n° 1 « *Partir à la recherche des trois Mousquetaires* » [en ligne] consultée le 18/06/2023. URL : https://www.youtube.com/watch?v=7p6TuD2K7tw
Tutoriel n° 2 « Percer les mystères du Titanic » [en ligne] consultée le 18/06/2023. URL : https://www.youtube.com/watch?v=Zi7csqvifHs
Tutoriel n° 3 « Créer votre corpus n'a jamais été si facile » [en ligne] consultée le 18/06/2023. URL : https://www.youtube.com/watch?v=rD2fsann5mQ
Tutoriel n° 4 : *à venir*.

[188] Chevallier, P. (2018). Les données au service de la connaissance des usages en ligne : l'exemple de l'analyse des logs de Gallica. Les Enjeux de l'information et de la communication, 19(2), p. 57-67. https://doi-org.ressources-electroniques.univ-lille.fr/10.3917/enic.025.0057

Conclusion

Le protocole d'enquête a donc permis de collecter les retours des testeurs et de les utiliser pour notre évaluation du dispositif Archival mettant en valeur les apports ainsi que les limites du projet et du démonstrateur. Les retours sont très diversifiés et permettent ainsi d'avoir, sur vingt expérimentateurs, un large éventail d'avis, de remarques et d'idées notamment en vue d'améliorer ce prototype et de nourrir la recherche et le débat sur les interfaces de recherche - notamment les interfaces dites « *intelligentes* » ou « *augmentées* ».

Conclusion générale

Notre objectif initial était d'élaborer une méthode d'évaluation du dispositif Archival tant dans la théorie du projet qu'au niveau du démonstrateur proposé, évaluation qui est donc en cours actuellement car si les deux sessions d'expérimentation sont passées, les projets avec l'équipe Archival devraient se poursuivre et peut-être amener à une troisième journée d'expérimentation et une mise en commun de l'analyse des parcours/*logs*.

La mise en place du cadre théorique et conceptuel a permis de nous situer dans le domaine et de poser les premiers jalons d'une boîte à outils conceptuelle et pratique qui devra être étoffée par la suite pour servir à de prochaines évaluations. Elle nous a permis d'initier les premières lectures sur la recherche d'information et les problématiques liées à l'évaluation de plateformes de recherche documentaire. La mise en place de l'enquête nous a donné la possibilité de nous confronter à la réalité du terrain et à la différence qu'il peut y avoir entre le protocole imaginé et l'expérience vécue en elle-même de même qu'il existe un écart entre les usages projetés par les concepteurs d'un système et les usages réels de celui-ci. L'évaluation qualitative, agile et hybride mise en place de notre côté a porté ses fruits, elle nous a permis de collecter des retours très intéressants et complémentaires à nos constats et analyses. Des retours riches de la part des testeurs qui ont permis de mieux saisir les attentes des usagers en matière de recherche d'information pour les comparer avec ce qui a été présenté dans le dispositif Archival.

Une évaluation qui demeure toutefois perfectible sur de nombreux points. Une multitude de champs et domaines n'ont pas été (suffisamment) explorés au travers de la littérature et donc exploités dans la création du protocole pour que l'on puisse parler d'une réelle évaluation pluridisciplinaire. L'évaluation est partielle car elle ne prend en compte que les aspects qualitatifs. Le quantitatif étant traité par les informaticiens du projet, nous n'y avons pas accès au moment de la rédaction de ce mémoire, ce qui aurait pourtant pu être éclairant sur les résultats obtenus grâce à l'enquête (notamment en ce qui concerne la performance des algorithmes).

Le protocole proposé nous a ainsi permis d'atteindre des résultats satisfaisants si ce n'est prometteurs pour la suite. Un travail qui ne fait que commencer et devrait donc être poursuivi, amélioré et augmenté en vue d'évaluer certaines interfaces et fonctionnalités de Gallica au cours du stage de deuxième année de master. L'évaluation de certaines interfaces et fonctionnalités de Gallica constitue un travail tout autre mais qui prend ses racines dans celui que nous avons

amorcé cette année et permettant de traiter d'autres problématiques telles que la diversité des publics, la masse documentaire (plus de 10 millions de documents sont accessibles en ligne sur la bibliothèque numérique) ou encore les modalités de recherche notamment de recherche avancée.

Bibliographie

Agnola, M., Azemard, G. & Da Silva, S. (2022). Conférences internationales EUTIC 2022 : « A l'intersection de l'art, de la science et de la technologie : dialogues entre les hommes et les machines » [Actes de conférence]. Académie Ionienne de Corfu, Grèce.

Antoine E., Kang H. J., Rousseau I., Azémard G., Béchet F., et al. (2023). Exploring Social Sciences Archives with Explainable Document Linkage through Question Generation. 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, ACL SIGHUM, Dubrovnik, Croatia.

Audeh, B., Beaune, P. & Beigbeder, M. (2015). MOR : Mesure orientée rappel pour les systèmes de recherche d'information. Document numérique, 18, p. 37-54.

Baccino, T. & Drai-Zerbib, V. (2015). La lecture numérique. Presses universitaires de Grenoble.

Baeza-Yates, R. & Ribeiro-Neto, B. (dir.). (1999). Modern Information Retrieval. Addison-Wesley.

Barbagelata, P., Inaudi, A. & Pelissier, M. (2014). Le numérique vecteur d'un renouveau des pratiques de lecture : leurre ou opportunité ?. Études de communication, 43, p. 17-38.

Béchet F., Antoine E., Auguste J. & Damnati G. (2022). Question Generation and Answering for exploring Digital Humanities collections. 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.

Bellot, P., Cauvet, C., Pasi, G. & Vallès-Parlangeau, N. (2012). Introduction. Document numérique, 15, p. 7-8.

Ben Lagha, S. (2002). La numérisation des catalogues : une analyse rétrospective. Document numérique, 6, p. 81-97.

Bérard, Y. & Crespin, R. (dir.). (2015). Aux frontières de l'expertise : dialogues entre savoirs et pouvoirs. Rennes, Presses universitaires de Rennes.

Boullier, D. (2016). Chapitre 2. Sociologie des usages. Dans : D. Boullier, Sociologie du numérique (p. 99-129). Paris : Armand Colin.

Boustany, J., Broudoux, É. & Chartron, G. (2013). Introduction. Diversification des médiations informationnelles. Dans : Joumana Boustany éd., La médiation numérique : renouvellement et

diversification des pratiques : Actes du colloque Document numérique et société, Zagreb (p. 7-10). Louvain-la-Neuve : De Boeck Supérieur.

Bulinge, F. (2022). Chapitre 6. Évaluer l'information. Dans : F. Bulinge, Maîtriser l'information stratégique : Méthodes et techniques d'analyse (p. 105-119). Louvain-la-Neuve : De Boeck Supérieur.

Bush, V. (1945) « As We May Think », Atlantic Monthly, 176, 1, p. 101-108.

Cabanac, G., Hubert, G., Boughanem, M. & Chrisment, C. (2011). Impact du « biais des ex aequo » dans les évaluations de recherche d'information. Document numérique, 14, p. 149-168.

Casagrande, A. & Vuillon, L. (2017). Sciences humaines et sociales et méthodes du numérique, un mariage heureux ?. Les Cahiers du numérique, 13, p. 115-136.

Cazals, F. & Cazals, C. (2020). Intelligence artificielle : L'intelligence amplifiée par la technologie. De Boeck Supérieur.

Chaudat, P. & Pierre Mirralès, P. (2009). L'évaluation des experts dans les organisations. Revue Interventions économiques, 39.

Chaudiron, S. (dir.) (2004). *Evaluation des systèmes de traitement de l'information*, Hermès.

Chaudiron, S. & Schmitt, L. (2000). Amaryllis : an evaluation-based program for Text Retrieval. Dans Jacquemin C., Mariani J. & Paroubek P. (dir.). Workshop Proceedings of LREC – Using Evaluation within HLT Programmes : Results and Trends. ELRA. Athens, p. 65-68.

Chevallier, P. (2018). Les données au service de la connaissance des usages en ligne : l'exemple de l'analyse des logs de Gallica. Les Enjeux de l'information et de la communication, 19(2), p. 57-67.

Chiaramella, Y. & Mulhem, P. (2007). La recherche d'information : De la documentation automatique à la recherche d'information en contexte. Document numérique, 10, p. 11-38.

Cleverdon, C. W., Mills, J. & Keen, M. (1966). Factors determining the performance of indexing systems. Association of Special Libraries and Information Bureau, Cranfield (Angleterre).

Cleverdon, C. (1970). Progress in documentation, evaluation tests of information retrieval systems. Journal of Documentation, 26, p. 55-67.

Cooper, W. E. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, *7*(1), p. 19-37.

Coret A., Kremer P., Landi B. Schibler D., Schmitt L. & Viscogliosi N. (2000). Accès à l'information textuelle en français : le cycle exploratoire Amaryllis. Ressources et évaluation en ingénierie des langues, Bruxelles, De Boeck & Larcier, p. 13-24.

Croft, W. B. (1977). Clustering large files of documents using a single link. Journal of the American Society for Information Science, 28.

Dalbin, S. (2007). Thésaurus et informatique documentaires : Des Noces d'Or. Documentaliste, 44(1).

Dedencker, C. & Kolmayer, E. (2006) Eléments de la psychologie cognitive pour les sciences de l'information, Villeurbanne, Presses de l'ENSSIB.

Delmas, C. (2011). Introduction. Dans : Corinne Delmas éd., Sociologie politique de l'expertise (p. 3-8). Paris : La Découverte.

Demortain, D. (2021). Experts scientifiques et action publique : paradoxe et perspectives de recherche pour la sociologie politique de l'expertise : Commentaire, 39, p. 33-41.

Ellis, D. (2002). The Physical and congitive paradigm in information retrieval research. Journal of Documentation, vol. 48, n°1, p. 45-64.

Fluhr, C. (1977). Algorithmes a apprentissage et traitement automatique des langues. Thèse de doctorat, Orsay, Université de Paris Sud.

Fluhr, C. (1981). Spirit : un système syntaxique et probabiliste d'indexation et de recherche d'informations textuelles. Dans Proceedings of ADBS-IDT81, Paris, ADBS, p. 113-116.

Fluhr, C. (2004). Chapitre 1 : L'évaluation des systèmes de recherche d'informations textuelles. Dans Chaudiron, S. (dir.). Evaluation des systèmes de traitement de l'information, Hermès.

Foucart, T. (2001). L'interprétation des résultats statistiques. Mathématiques et sciences humaines, 153.

Francart, T. (2016). Des textes augmentés avec les données du Web. I2D - Information, données & documents, 53, p. 45-45.

Gallinari, P., Zaragoza H. & Amini M. (2002). Chapitre 11 : Apprentissage et Données Textuelles. Dans Bases de données et statistiques, Dunod.

Gardin, J-C. (1962). Documentation sur cartes perforées et travaux sur ordinateurs dans les sciences humaines. Revue internationale de documentation, vol. 29, p.83-92.

Gardin, J-C. (1967). Recherches sur l'indexation automatique des documents scientifiques. Revue d'informatique et de recherche opérationnelle, 1ere année, n°6, p. 27-46.

Gardin, J-C. (1974). Analyse documentaire et théorie linguistique. Les analyses de discours Neuchatel, Delchaux et Niestlé, (col Zéthos), p.120-128.

Garfield, E. (1996). A tribute to Calvin N. Mooers, A pioneer of Information Retrieval. The Scientist, 11.

Grau, B., Laleau, R. & Ramel, J. (2011). Introduction. Document numérique, 14, p. 7-10.

Ihadjadene M. & Martins D. (2004). Expertise dans le domaine et expertise dans Internet : leurs effets sur la recherche d'informations. Paris, Hermès, 39, p. 133-142.

Ihadjadene, M. & Chaudiron, S. (2008). L'étude des dispositifs d'accès à l'information électronique. Dans HAL (Le Centre pour la Communication Scientifique Directe). French National Centre for Scientific Research.

Ihadjadene, M. & Chaudiron, S. (2008 bis). Quelles analyses de l'usage des moteurs de recherche. Questions de communication, (14), p. 17-32.

Ingwersen, P. (1999). Cognitive Information Retrieval. Annual Review of Information Science & Technology, 34, p. 3-52.

Ingwersen, P. & Järvelin, K. (2005). The Turn : Integration of Information Seeking and Retrieval. Dans Contexte, Kluver.

Jansen, B. J., Koshman, S. & Spink, A. (2006). Web searching on the Vivisimo search engine. Journal of the Association for Information Science and Technology, 57(14), p. 1875-1887.

Jouët, J. (2000). Retour critique sur la sociologie des usages. Réseaux, 18(100), p. 487-521.

Joly, P.B. (2012). La fabrique de l'expertise scientifique. Hermès, 64, p. 22-28.

Laborderie, A. (2020). Le livre augmenté : un nouveau paradigme du livre ?. Revue de la BNF, 60, p. 148-159.

Lancaster, F. W. (1979). Information Retrieval Systems : Characteristics, Testing, and Evaluation. New York, Toronto, Wiley.

Le Maitre J., Murisasco E. & Robert M. (1997). From Annotated Corpora to Databases : the SgmlQL Linguage. Dans J. Nerbonne (ed.), Linguistic Databases, CSLI Lecture Notes n°77, p. 37-58.

Lesk, M. (1995). The Seven Ages of Information Retrieval. *Conference for the 50th anniversary of « As We May Think »*.

Lipsyc, C. & Ihadjadene, M. (2013). Architecture de l'information et éditorialisation. Études de communication, 41, p. 103-118.

Luk, R. W. P., Leong, H. V., Dillon, T. S., Chan, A. T. S., Croft, W. B. & Allan, J. (2002). A survey in indexing and searching XML documents. Journal of the Association for Information Science and Technology, 53(6), p. 415-437.

Maron, M. E. (1960). On relevance, probabilistic indexing and information retrieval. Journal of the ACM, 66.

Martineau, J. (2023). Transition numérique et intelligence artificielle : d'importants enjeux éthiques à surveiller. Gestion, 48, p. 60-64.

Maxim, L. & Arnold, G. (2012). Entre recherche académique et expertise scientifique : des mondes de chercheurs. Hermès, La Revue, 64, p. 9-13.

Medioni, S. & Benmoyal-Bouzaglo, S. (2018). Chapitre 4. L'évaluation du consommateur en situation d'hyperchoix. Dans : S. Medioni & S. Benmoyal-Bouzaglo (dir), Marketing digital (p. 141-184). Paris : Dunod.

Micheau, B. (2016). Saemmer Alexandra : Rhétorique du texte numérique : figures de la lecture, anticipations de pratiques. Études de communication, 46, p. 201-206.

Montaigne. (1980) Trésor de la langue française, vol. 8, p. 472.

Moulahi, B., Tamine, L. & Ben Yahia, S. (2016). Estimation de la pertinence multidimensionnelle en recherche d'information : Évaluation de l'application d'un opérateur flou d'agrégation. Document numérique, 19, p. 59-82.

Muller, C. (2015). Interfaces documentaires innovantes. I2D - Information, données & documents, 52, p. 15-16.

Nazarenko, A. & Poibeau, T. (2004). Chapitre 5 : L'évaluation des systèmes d'analyse et de compréhension de texte. Dans Chaudiron, S. (dir.). Evaluation des systèmes de traitement de l'information, Hermès.

Nguyen D. H., Gravier G. & Sébillot P. (2022). Filtrage et régularisation pour améliorer la plausibilité des poids d'attention dans la tâche d'inférence en langue naturelle. TALN 2022 - Traitement Automatique des Langues Naturelles, Avignon, France, p.95-103.

Nguyen D. H., Mallart C., Gravier G. & Sébillot P. (2023). Regularization, Semi-supervision, and Supervision for a Plausible Attention-Based Explanation. NLDB 2023 - 28th International Conference on Natural Language and Information Systems, Derby, United Kingdom, p.1-14.

Norman, D. (2013). The Design of Everyday Things : Revised and Expanded Edition. Constellation.

Pédauque, Roger T. (2003). Document : forme, signe et médium, les re-formulations du numérique. STIC-CNRS.

Robertson, S. & Spärck Jones, K. (1976). Relevance weighing of search terms. Journal of the American Society for Information Science, 27.

Roger, A. & Roger, P. (2001). Rôle et place des experts dans une société de l'information. Actes du XIIe Congrès de l'AGRH.

Rouet, J. & Tricot, A. (1998). Chercher de l'information dans un hypertexte : vers un modèle des processus cognitifs. Les hypermédias : approches cognitives et ergonomiques, p. 57-74.

Rouet, G. (2019). Démystifier les algorithmes. Hermès, La Revue, 85, p. 21-31.

Salton, G. (1968). Automatic information organization and retrieval. Proceedings of the 11th coference on Computational linguistics.

Salton, G. (1971). The SMART Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall, Inc eBooks.

Salton, G. (1989). Automatic text processing : the transformation, analysis, and retrieval of information by computer. Choice Reviews Online, 27(01).

Salton, G. & McGill, M. J. (1983). Introduction to Modern Information Retrieval. Information storage and retrieval systems. New York, McGraw-Hill.

Saracevic, T. (1970). The concept of Relevance. Introduction to Information Science, R.R. Bowker Compagny.

Sauret, N. (2022). Intelligence artificielle & Sciences humaines et sociales (SHS) : opportunités, défis et perspectives. I2D - Information, données & documents, 1, p. 97-103.

Schaepelynck, V. & Sustam, E. (2018). Autogestion. Le Télémaque, 54, p. 27-36.

Silverstein, C., Marais, H., Henzinger, M. & Moricz, M. (1999). Analysis of a very large web search engine query log. Sigir Forum, 33(1), p. 6-12.

Sitbon, L., Bellot, P. & Blache, P. (2010). Vers une recherche d'information adaptée aux utilisateurs dyslexiques. Document numérique, 13, p. 161-185.

Spärck Jones, K. & Willett, P. (1997). Readings. Dans Information Retrieval. Morgan and Kaufmann publishers.

Taylor, R. (1968). Question negociation and information seeking in libraries, College and research libraries, n°29, p. 178-194.

Timimi, I. (2020). Évaluation d'outils et outils d'évaluation. Revue COSSI, (9).

Trepos J.Y. (1996), La sociologie de l'expertise, Presses Universitaires de France.

Truphème, S. & Gastaud, P. (2023). Outil 22. L'analyse de la qualité et de la performance d'un site Web. Dans : S. Truphème & P. Gastaud (dir), La boîte à outils du Marketing digital (p. 74-75). Paris : Dunod.

Van Rijsbergen, C. J. (1986). A New Theorical Framework for Information Retrieval. Proceedings of the ACM-SIGIR86 International Conference on Research and Development in Information Retrieval, Pisa, p. 194-200.

Van Rijsbergen, C. J. (1986) A non-classical logic for information retrieval. The Computer Journal 29, 6, p. 481-485.

Van Rijsbergen, C. J. (1989). Towards an information logic. Proceedings of ACM-SIGIR89 International Conference on Research and Developmtent in Information Retrieval Cambridge, Massachusetts USA, p. 77-86.

Weill, C. (1999). La revue Autogestion comme observatoire des mouvements d'émancipation. *L'Homme et la société*, *132*(2), p. 29-36.

Sitographie

Dans l'ordre d'apparition dans le texte

- Culture, P. (2023, 6 février). Accueil - Pass culture. Consulté à l'adresse https://pass.culture.fr/
- Spideo. (2023, 5 juillet). Spideo - Humanized Recommendations. built to scale. Consulté à l'adresse https://spideo.com/
- Lancement de la mission franco-québécoise sur la découvrabilité* des contenus culturels francophones en ligne. (s. d.). Consulté à l'adresse https://www.culture.gouv.fr/Presse/Communiques-de-presse/Lancement-de-la-mission-franco-quebecoise-sur-la-decouvrabilite-des-contenus-culturels-francophones-en-ligne
- Découvrabilité en ligne des contenus culturels francophones. (s. d.). Consulté à l'adresse https://www.culture.gouv.fr/Thematiques/Europe-et-international/Publications/Decouvrabilite-en-ligne-des-contenus-culturels-francophones
- ARCHIVAL. (s. d.). Consulté à l'adresse https://www.fmsh.fr/actualites/archival
- Missions et organisation de la BnF. (s. d.). Consulté à l'adresse https://www.bnf.fr/fr/missions-et-organisation-de-la-bnf
- Gallica. (s. d.). Consulté à l'adresse https://gallica.bnf.fr/accueil/fr/
- De Lille, U. (s. d.). Présentation de l'unité : Recherche Interdisciplinaire en Information et Communication - EA 4073. Consulté à l'adresse https://geriico.univ-lille.fr/lunite/presentation-de-lunite
- De Lille, U. (s. d.-a). Axes thématiques : Recherche Interdisciplinaire en Information et Communication - EA 4073. Consulté à l'adresse https://geriico.univ-lille.fr/lunite/axes-thematiques
- LIG - Université Grenoble Alpes - MRIM - Modélisation et Recherche d'Information Multimédia. (s. d.). Consulté à l'adresse https://www.liglab.fr/fr/recherche/equipes-recherche/mrim
- Archival - Autogestion. (s. d.). Consulté à l'adresse https://archival.msh-paris.fr/
- Valorisation d'archives multimédia : Compréhension automatique multimodale du langage pour de nouvelles interfaces intelligentes de médiation et de transmission des savoirs | ANR. (s. d.). Consulté à l'adresse https://anr.fr/Projet-ANR-19-CE38-0011

- Questionnaire d’expérimentation | Framaforms.org. (s. d.). Consulté à l’adresse https://framaforms.org/questionnaire-dexperimentation-1683730735
- Rendre les sites et services numériques accessibles à toutes et à tous - Référentiel général d’amélioration de l’accessibilité. (s. d.). Consulté à l’adresse https://accessibilite.numerique.gouv.fr/
- Bibliothèque nationale de France BnF. (2022, 15 février). *Partir à la recherche des trois Mousquetaires | Chercher, trouver dans Gallica Tuto # 1* [Fichier vidéo]. Consulté à l’adresse https://www.youtube.com/watch?v=7p6TuD2K7tw
- Bibliothèque nationale de France BnF. (2022b, février 15). Percer les mystères du Titanic | Chercher, trouver dans Gallica Tuto # 2 [Fichier vidéo]. Consulté à l’adresse https://www.youtube.com/watch?v=Zi7csqvifHs
- Bibliothèque nationale de France BnF. (2022c, juin 30). Créer votre corpus n’a jamais été si facile ! | Chercher, trouver dans Gallica Tuto # 3 [Fichier vidéo]. Consulté à l’adresse https://www.youtube.com/watch?v=rD2fsann5mQ

Résumé

La recherche d'information revêt une importance capitale dans nos sociétés actuelles et ses enjeux sont nombreux et divers. Dans ce cadre, les interfaces documentaires tentent de s'adapter sans cesse aux besoins renouvelés des usagers et aux technologies en constante évolution, en particulier dans le domaine de l'intelligence artificielle (IA). Ainsi, naissent depuis plusieurs années des interfaces documentaires dites « *intelligentes* » car augmentées par IA ; notre objectif étant d'en comprendre les possibilités, les apports actuels ainsi que les perspectives envisagées pour proposer une méthode d'évaluation de ce type de plateforme. Pour ce faire, nous avons travaillé sur le projet ANR Archival comme cas pratique afin d'éprouver notre protocole d'expérimentation du dispositif et d'initier notre évaluation de la plateforme.

Mots-clés

Interface documentaire, Lecture augmentée, Intelligence artificielle, Recherche d'information, Découvrabilité, Expérimentation et évaluation.