RAPPORT N°3

RECOMMANDATIONS

PROPOSITIONS DE FONCTIONNALITÉS POUR UN SERVICE DE TRADUCTION SCIENTIFIQUE OUTILLÉE

Ce rapport est publié dans le cadre du projet Traductions et science ouverte sous la coordination d'OPERAS.

—

Ce projet est financé par le Fonds national pour la science ouverte, qui rassemble des contributions du ministère de l'Enseignement supérieur et de la Recherche, d'universités et d'organismes de recherche français.

Pour ce projet, le Fonds national pour la science ouverte reçoit également une contribution spéciale du ministère de la culture français.

Auteurs : Xavier Auffret, Claire Lapassat, Philippe Lacour avec la participation de Claire Larsonneur et Thierry Poibeau

l'atelier universel

OPERAS

Disclaimer

The ideas and views expressed in the exploratory reports only reflect those of the experts involved in the studies and may not be representative of the opinions or policies promoted by any specific organization, institution, or government entity. The present report is therefore only intended for informational purposes.

Avertissement

Les idées et les perspectives exprimées dans les rapports exploratoires reflètent uniquement celles des spécialistes ayant contribué aux études et ne sont pas nécessairement représentatives des opinions ou des politiques promues par une organisation, une institution ou une entité gouvernementale spécifique. Le présent rapport est donc uniquement diffusé à des fins d'information.

LES AUTEURS

direction de l'étude

l'atelier universel

AGENCE DE DESIGN ET DE CONSEIL EN INNOVATION

- experte des démarches centrées utilisateurs,
- spécialisée dans les projets complexes, notamment dans les logiciels professionnels pour la recherche et la science ouverte,
- croisant design désirable et intrants scientifiques,
- habituée à travailler avec des chercheurs.

directeur scientifique

Philippe Lacour

DOCTEUR ET CHERCHEUR SPÉCIALISTE DE LA PHILOSOPHIE DU LANGAGE

- professeur titulaire au département de philosophie de l'Université de Brasilia (UnB, Brésil). Spécialiste de l'épistémologie des sciences de la culture (SHS, humanités),
- responsable du projet Traduxio (https://traduxio.org), outil de traduction collaborative multilingue,
- co-responsable du projet de recherche "L'IA et ses critiques" (UnB).

Site personnel : https ://philippelacour.net

comité d'experts

Claire Larsonneur

MAÎTRE DE CONFÉRENCE ET CHERCHEUSE SPÉCIALISÉE EN ÉCONOMIE DE LA TRADUCTION, TRADUCTION ET NUMÉRIQUE

Thierry Poibeau

DIRECTEUR DE RECHERCHE AU CNRS ET DIRECTEUR ADJOINT DU LABORATOIRE LATTICE

SOMMAIRE

CONTEXTE GÉNÉRAL DU PROJET

Le projet *Traductions et science ouverte*
L'étude n°2 *Use case study for a technology-based scientific translation service*

a_Le projet *Traductions et science ouverte*

Les technologies basées sur l'**intelligence artificielle** sont actuellement en plein essor dans de nombreux domaines et la **traduction** ne fait pas exception. Depuis l'avènement des moteurs neuronaux, la **traduction automatique**, en particulier, suscite de plus en plus d'enthousiasme auprès des spécialistes et du grand public. Cet enthousiasme est le résultat des indéniables progrès accomplis en termes de performances et d'accessibilité, au moins pour certaines paires linguistiques, de cette technologie. Les profils d'utilisateurs l'ayant adoptée et les possibilités d'application sont aujourd'hui extrêmement variés.

Les **enjeux du multilinguisme** étant de plus en plus importants dans la société contemporaine, **la traduction automatique fait l'objet d'une attention accrue et, par conséquent, d'investissements et de développements significatifs**. Comme toute technologie, cependant, cet outil incontestablement précieux n'est pas sans défaut. Au contraire, **son utilisation exige prudence et discernement**.

Comment **exploiter la traduction automatique de manière raisonnée pour favoriser le multilinguisme dans la communication scientifique**, un domaine historiquement marqué par une dominance de la langue anglaise et, en même temps, par une pénurie de ressources humaines et financières destinées à la traduction ? **Telle est la question que le Ministère français de l'Enseignement Supérieur et de la Recherche s'est posée en 2020**, un an après le lancement de **l'Initiative d'Helsinki sur le multilinguisme dans la communication savante**[1]. La réponse a été formulée par un groupe de travail, constitué d'experts en traduction et en traitement automatique des langues (TAL), ayant publié la même année **le rapport Traductions et science ouverte**[2].

Les recommandations du groupe de travail ont notamment permis d'élaborer une série de **quatre études exploratoires**, dont la coordination a été confiée en 2022 à **l'infrastructure de recherche européenne OPERAS**[3]. Les études ont été conçues afin de **préparer un déploiement des technologies de la traduction, structuré et mutualisé entre différents acteurs et utilisateurs de la communauté scientifique, notamment grâce à la création d'un service de traduction outillée, combinant technologies, ressources linguistiques numériques et expertises humaines.**

La **première étude Cartographie et collecte de corpus scientifiques bilingues** a porté sur l'identification des données textuelles bilingues existantes dans la paire linguistique anglais-français et dans trois disciplines pilotes[4]. Une fois les publications identifiées et les conditions d'utilisation analysées, environ 100 000 segments et 300 termes spécialisés ont été collectés pour chacune des disciplines considérées.

En parallèle, **une étude de cas d'usage a permis d'élaborer un état de l'art des technologies, pratiques et outils de la traduction et de formuler des propositions de fonctionnalités** pour le futur service de traduction scientifique outillée (cf. infra). Il s'agit de **l'étude 2** présentée en détail dans la prochaine section de cette publication.

Dans le cadre de la **troisième étude Évaluation de la traduction automatique dans la communication scientifique**, trois moteurs ont été entraînés avec les données textuelles collectées dans le cadre de l'étude 1 afin d'en évaluer les performances de traduction avec différents types de textes scientifiques.

La dernière étude **Feuille de route et projections budgétaires pour un service de traduction scientifique outillée** a hérité les informations et les questions des trois études précédentes, afin d'élaborer des feuilles de route, des projections budgétaires ainsi que des modèles opérationnels, juridiques et économiques viables pour le futur service de traduction scientifique outillée.

b_ L'étude n°2 *Use case study for a technology-based scientific translation service*

Ce rapport est le dernier de l'étude 2 qui se concentre donc sur :

- l'identification des **technologies, outils, pratiques et enjeux actuels de la traduction scientifique.**
- la **proposition de fonctionnalités** pour le futur service de traduction scientifique outillée.

Pour cela, d'un point de vue méthodologique, ont été réalisés lors de cette étude plusieurs interviews et ateliers détaillés page suivante.

1 https ://www.helsinki-initiative.org/fr
2 https ://hal-lara.archives-ouvertes.fr/OUVRIR-LA-SCIENCE/hal-03640511
3 https ://operas-eu.org/
4 Climatologie et changement climatique (Sciences physiques et de l'ingénierie) ; Neurosciences et troubles du système nerveux (Sciences de la vie) ; Mobilité humaine, environnement et espace (Sciences humaines et sociales).

Étude 2 - Approche méthodologique

l'Atelier Universel = Pilotage

ENTRETIENS FORMELS

Entretiens formels d'1h à 1h30 - dont une synthèse est disponible dans le présent rapport

- **3 entretiens avec 3 traducteurs** représentant chacun un champ disciplinaire différent (SHS, changement climatique, le médical).
- **1 entretien avec une équipe d'Erudit**, plateforme québécoise de diffusion dans le domaine des SHS.
- **1 entretien avec une personne de l'Académie des Sciences**, organisme français d'expertise scientifique et de soutien à la recherche qui publie 7 revues de recherche de niveau international dans différentes disciplines de STM (Science Technologie & Médecine).

ENTRETIENS INFORMELS

Entretiens informels auprès de chercheurs ou enseignants-chercheurs aux disciplines variées (psychologie comportementale, mathématiques appliquées à la physique, politiques publiques, traductologie), notamment :

- **Franck Barbin**
  Maître de conférences en traduction anglaise à l'université Rennes 2.
- **Stéphane Pouyllau**
  Ingénieur de recherche au CNRS, responsable de l'Huma-Num lab.

SESSIONS DE TRAVAIL AVEC DES EXPERTS SCIENTIFIQUES

Direction scientifique :

- **Philippe Lacour**
  Docteur et chercheur spécialiste de la philosophie du langage.

Expertise scientifique :

- **Claire Larsonneur**
  Maître de conférence et chercheuse, spécialisée en économie de la traduction, traduction et numérique.
- **Thierry Poibeau**
  Directeur de recherche au CNRS et directeur adjoint du laboratoire LATTICE.

PRÉSENTATION D'OUTILS

- Démonstration de Trados, commentée, par un traducteur.

1 ATELIER « PRATIQUES DE LA TRADUCTION SCIENTIFIQUE »

pour **préciser les pratiques** de la traduction scientifique, les grandes étapes, outils et identifier les besoins, les freins et motivations. Avec des **chercheurs** (4), **traducteurs** (4), **éditeurs-diffuseurs** (5) et l'**équipe projet complète**.

1 ATELIER « FONCTIONNALITÉS »

pour **identifier les fonctionnalités clefs** à proposer dans le futur service et faire de premières esquisses rapides. Avec des **chercheurs** (4), **traducteurs** (4), **éditeurs-diffuseurs** (3) et l'**équipe projet complète**.

1 ATELIER « ARBITRAGES »

pour **arbitrer parmi les propositions** faites lors de l'atelier précédent. Avec des **membres du Conseil Scientifique** et du **Comité de pilotage du projet *Traductions et science ouverte***, avec la participation de **spécialistes disciplinaires du Ministère de l'Enseignement Supérieur et de la Recherche** (MESR) et l'**équipe projet complète.**

l'Atelier Universel = Pilotage

INTRODUCTION DU RAPPORT

L’objectif stratégique du projet
Nos recommandations
Points d’attentions
Les bonnes pratiques à mettre en avant dans le service

*Ce rapport présente une **sélection de propositions** arbitrées de façon collégiale avec des membres du Conseil Scientifique et du Comité de pilotage du projet Traductions et science ouverte, avec la participation de spécialistes disciplinaires du Ministère de l'Enseignement Supérieur et de la Recherche (MESR).*

*Le projet a une vocation **opérationnelle** : l'enjeu est de construire un service concret, répondant de manière effective et efficace aux besoins et aux usages observés auprès des acteurs et utilisateurs cibles. Pour cela, il convient d'utiliser des technologies déjà existantes, ayant fait leurs preuves, et autour desquelles on puisse regrouper une communauté conséquente (utilisateurs).*

a_ L'objectif stratégique du projet

L'ambition est celle d'un **double rayonnement** de la production scientifique :

- de la recherche française à l'extérieur (diplomatie française et francophone)
- et de la recherche internationale à l'intérieur (faciliter son accès, science citoyenne).

Afin d'atteindre cet objectif, il apparaît comme prioritaire de concevoir une stratégie d'indexation multilingue des contenus scientifiques, afin d'en assurer la découvrabilité en plusieurs langues notamment via une traduction des métadonnées. Dans cette stratégie, la priorité est donnée à **l'automaticité** : les aspects collaboratifs ne sont pas exclus, mais ne sont pas prévus dans l'immédiat (cela dépendra des premiers résultats).

b_Nos recommandations

Le point de vue de **différents acteurs** doit être considéré : chercheurs, lecteurs grand public, institutions, éditeurs, traducteurs. Tous ces acteurs doivent trouver un intérêt au service, dont la **mutualisation** est le principal atout.

Concrètement, il s'agit de construire un **service** (para-public) auquel les utilisateurs puissent faire appel depuis des environnements **externes** (notamment par le biais de plug-ins, pour des plateformes de diffusion, pour les traducteurs, pour les individus, etc.) et depuis une interface, voire un environnement, **interne**.

La priorité donnée dans cette étude est bien celle de la **traduction** (automatique) ; néanmoins, il nous semble important de mentionner la **rédaction** (aide à la rédaction) voire la **génération** de textes (de synthèses par exemple) en langue cible. En effet, les développements en cours (modèles de langage) montrent une **tendance vers des moteurs capables de traduire autant que de proposer ou générer du texte**. Par ailleurs, les entretiens et ateliers **utilisateurs ont fait émerger ce type de besoins** pour certains profils ou situations ; par exemple, dans certaines disciplines la rédaction se fait usuellement directement dans une langue cible. Ces fonctionnalités d'aide à la rédaction ou de génération de texte, bien que moins centrales ou prioritaires, doivent être considérées, afin d'**anticiper leur correcte mise en œuvre** notamment en ce qui concerne les points de vue **techniques**, **juridiques** ou **éthiques**.

Points d'attention

L'attractivité du service est une question fondamentale, qui doit être abordée selon deux lignes principales :

- l'aspect **normatif** (éthique, juridique, politique) de la science ouverte et citoyenne : il s'agit de vaincre des appréhensions pour susciter la confiance des utilisateurs en un service d'intérêt général, par une construction graduelle de la réputation : protection des données, amélioration effective de l'indexation et de la découvrabilité, capacité à rendre des comptes, pérennité, etc.
- l'aspect **ergonomique** (aisance, intuitivité, facilité de la prise en main) : il s'agit de séduire, en faisant en sorte que le(s) plug-in(s) et l'interface incitent à l'utilisation par leur aspect plaisant, le sentiment d'une valeur ajoutée apportée par des fonctionnalités utiles, le gain de temps, la rapidité d'exécution d'une tâche, la fidélisation, etc. Exemple : au moment du dépôt d'un article dans HAL, possibilité de produire des métadonnées (titre, résumé, mots-clés) dans différentes langues, de façon automatique, avec une option de post-édition (éventuellement collaborative).

Il faudra en particulier veiller à **se distinguer** tout en tenant compte des habitudes des utilisateurs, et se mettre en valeur **par rapport aux concurrents actuels**, par exemple en proposant des points de **différenciation** forts, comme la mise à disposition de glossaires scientifiques.

L'un des aspects les plus incitatifs tient à la **taille critique du service** (en termes de données multilingues thésaurisées, préservées selon des garanties suscitant la confiance), qui peut permettre d'attirer des utilisateurs « isolés », convaincus de pouvoir gagner à cette mutualisation, ou rendus confiants par une alliance d'acteurs de référence de l'édition et de la communication scientifique (« puisque cet acteur de référence en fait partie je peux y aller aussi », ou « puisque telle institution l'utilise, ça vaut la peine d'essayer »). Une logique du **donnant-donnant** (ou mieux : du **gagnant-gagnant**) doit prévaloir. Les éditeurs doivent être convaincus de l'utilité du projet et prêts à s'associer, selon des garanties juridiques à définir soigneusement. Pour cette raison, s'appuyer sur un service public de diffusion déjà existant (type HAL, Open Edition, Érudit…) pourrait s'avérer précieux. Par ailleurs, les traducteurs doivent avoir suffisamment d'intérêt à la thésaurisation pour confier (sous conditions) leur travail à la base textuelle du service (qui doit en garder copie). L'intérêt pris à la collaboration est une possibilité subsidiaire, au cas où le service l'offrirait (en option). Enfin, le service doit garantir un certain accès aux utilisateurs sans que les données recueillies (textes originaux, traductions, métadonnées) ne puissent être récoltées de façon sauvage par les pratiques de « chalutage » des grands concurrents du Web.

Multilinguisme

Le système de stockage des données textuelles doit être d'emblée multilingue, que ce soit pour les items (les textes ou documents) ou les ensembles d'items (les corpus), afin de permettre un archivage qui puisse s'enrichir des éventuelles (diverses) traductions. Cela pose une **difficulté technique, que l'on peut résumer comme « le problème de l'ONU »** : dans cette institution internationale, une résolution peut entrer dans une des langues officielles, mais doit être traduite immédiatement dans les autres ; or la langue d'entrée peut être différente à chaque fois. Si l'on opte pour un archivage par « couple » de langues (ou par corpus bilingue), on est alors vite confronté à une prolifération exponentielle, très coûteuse (mémoire, énergie, procédures).

Concernant les items (documents), mieux vaut donc opter dès le départ pour un **archivage multilingue** : chaque texte déposé dans le service doit ainsi se voir attribuer un nombre indéfini de traductions possibles, ce qui permet au système de toujours pouvoir rajouter, à l'avenir, de nouvelles langues. Par exemple, un texte original français voit son résumé traduit en anglais ; mais, face à la demande internationale, il est décidé de rajouter une version chinoise, espagnole, et portugaise ; et quelques années plus tard, un spécialiste d'outre-Rhin veut rajouter une version allemande, etc.

Par ailleurs, **la même réflexion vaut pour les corpus** : le stockage doit être fait de telle sorte qu'il permette de constituer des **collections sur mesure** (multilingues, selon certains groupements disciplinaires ou d'époque, de tradition scientifique, etc.), et pas simplement des ensembles « rigides » (par exemple : bilingue avec un alignement par phrases). L'étiquetage des corpus devrait aussi pouvoir être réalisé de façon souple, afin de permettre plusieurs modes de classement (sur le type des classifications collaboratives spontanées (*folksonomies*), mais de façon plus ordonnée). Ainsi, la **thésaurisation souple** sera la meilleure garantie d'une **pérennité** et d'une **interopérabilité** démultipliée, susceptible de s'adapter à différentes exigences d'exportation, de traitement et d'interprétation des documents et des données stockées.

Les bonnes pratiques à mettre en avant dans le service

Traçabilité & valorisation des contributeurs
À chaque fois qu'une traduction (ou éventuellement une génération) automatique de texte sera lancée, il faut que l'on puisse savoir à partir de quel moteur elle a été créée, à quelle date, et quelles modifications elle a subies (quand et par qui), ceci afin d'assurer une meilleure compréhension du produit textuel qu'on a sous les yeux. Par ailleurs, en cas de post-édition, voire de collaboration pour la traduction, le nom des traducteurs doit pouvoir apparaître et être cité dans la référence (Dony et al. 2023).

Confiance et usage des données
Les métadonnées (notamment : titre, résumé, mot-clés) doivent pouvoir être stockées en toute sécurité, dans un espace d'emblée multilingue, afin de permettre un archivage qui puisse s'enrichir des éventuelles (diverses) traductions demandées par les utilisateurs.

Pérennité et interopérabilité
Une grande partie de l'intérêt d'un tel service tient à ses possibilités de thésaurisation. Les textes déposés, générés et traduits, et l'ensemble des métadonnées stockées, doivent pouvoir être consultés, retrouvés, exportés et traités facilement. Il convient notamment de penser à la possibilité, de niveau supérieur, de constituer des corpus de textes de façon souple et multilingue.

Qualité scientifique
Elle doit être garantie par plusieurs critères. D'abord, la qualité du moteur (son pré-entraînement, son entraînement, son suivi et son guidage humain) et de son corpus d'apprentissage. À noter que, dans cette perspective semi-automatique, **la qualité a vocation à s'améliorer par incrémentation**, dès lors que le corpus d'entraînement est abondé en nouveaux textes générés ou traduits, voire améliorés par la post-édition. Ensuite, la qualité provient du contrôle de l'accès au service (éventuellement par identification pour un certain niveau de service) qui permet d'éviter les interventions malignes de robots ou d'internautes. Enfin la qualité dépend de la précision des traductions et de la rigueur du travail de post-édition, laquelle est essentielle dans une optique de stockage et de thésaurisation (Dony et al. 2023). La traduction et éventuellement la génération automatique de textes doivent être bien encadrées et supervisées afin de repérer et corriger les possibles « hallucinations » de la machine, problème bien connu dans la recherche en traitement automatique des langues (TAL).

Différentes directions ont été envisagées dans les précédents rapports, allant de la traduction automatique à des outils d'aide à la rédaction dans une langue cible, en passant par des fonctionnalités collaboratives autour des glossaires par exemple. Lors de l'atelier, il a été décidé de se concentrer en priorité sur la ***mise en place d'un moteur entraîné avec des données textuelles spécialisées*** *et sur* ***son utilisation double : directe et autonome*** *(via une interface de traduction automatique), mais aussi* ***indirecte et intégrée à un écosystème d'outils existants*** *(entre autres sous la forme de plug-ins). Cette priorisation passe également par une* ***thésaurisation de données ou de connaissances de qualité****, faite par des collectes significatives organisées et progressivement (de façon incrémentale)* ***par les utilisateurs et devant leur demeurer accessible****.*

PRÉSENTATION DES PISTES RETENUES

Introduction : Quelles pistes retenues et quelle grille de lecture ?
Un moteur spécialisé accessible via différentes modalités
Une thésaurisation des données
Feuille de route macro

01

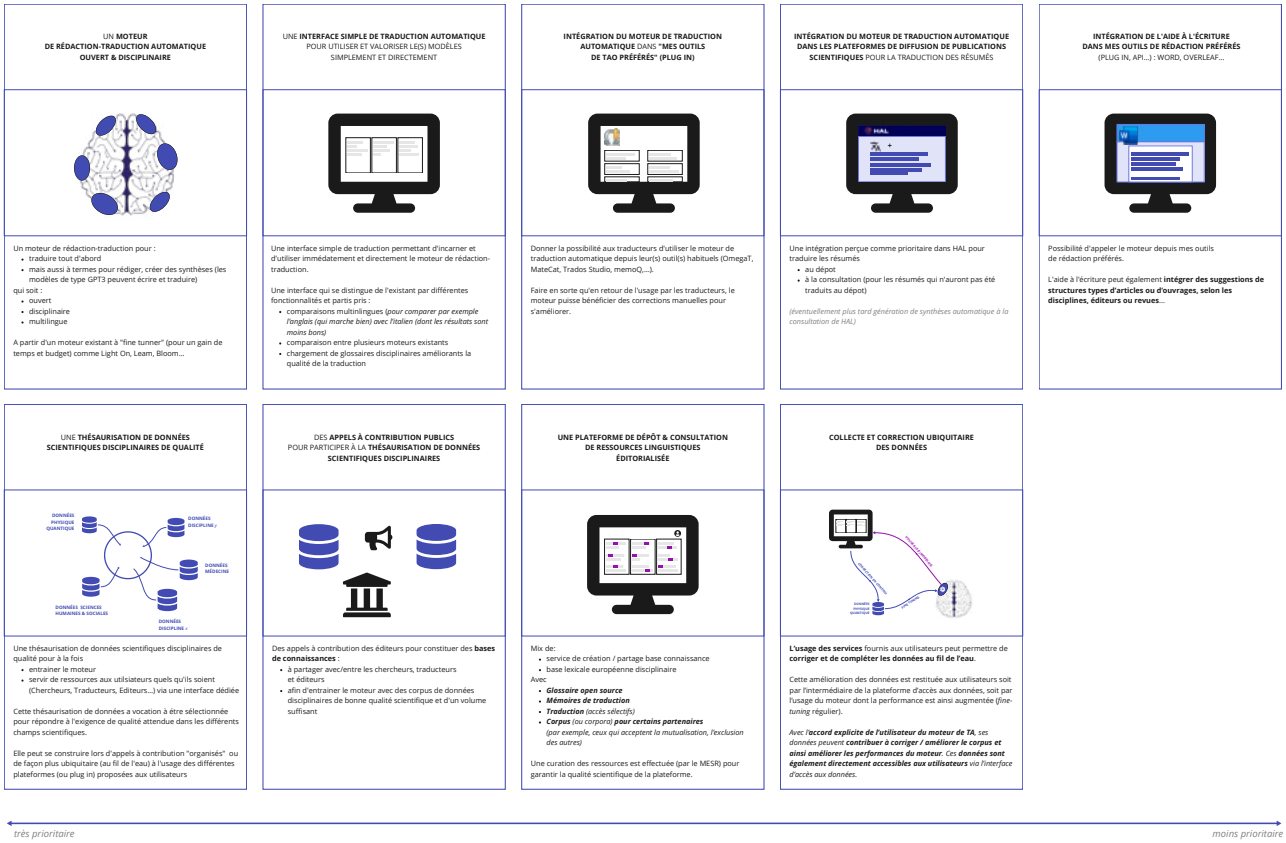
INTRODUCTION : QUELLES PISTES RETENUES ET QUELLE GRILLE DE LECTURE ?

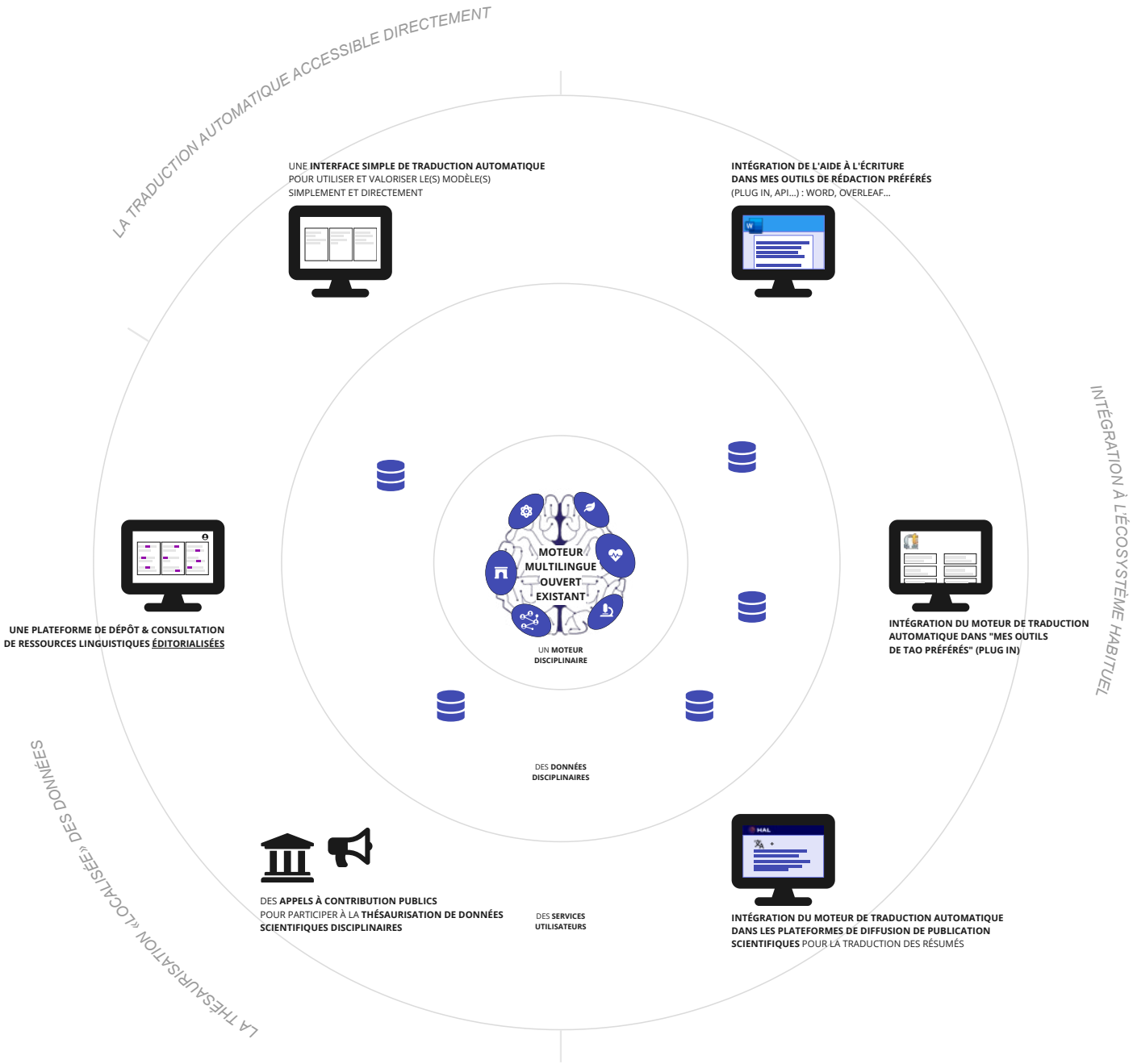
a_ Les pistes retenues (vue d'ensemble)

Les pistes retenues peuvent être présentées de façon synthétique comme suit :

- **Le moteur spécialisé de traduction automatiques de textes scientifiques** (voire d'aide à la rédaction et génération de textes)
- **Les données textuelles multilingues, spécialisées et de qualité** (utiles pour l'entraînement du moteur disciplinaire mais aussi en l'état par les utilisateurs)
- **Les services utilisateurs**, c'est-à-dire les modalités d'**accès en lecture et écriture au moteur et aux données** ; il s'agit des interfaces utilisateurs, par opposition au cœur technique, artefacts rendant des services aux utilisateurs et permettant d'alimenter le thésaurus.

Cartographie des pistes retenues





▲

Architecture fonctionnelle de principe des pistes retenues (simplifiée).

b_ Les critères clefs

Afin d'aider à **lire les propositions retenues** collégialement, nous proposons les critères suivants : *précision scientifique, bénéfice utilisateurs, souveraineté, pérennité, multilinguisme*.

Précision et qualité scientifique des productions
Les propositions qui seront développées doivent contribuer à la production de textes scientifiques de qualité. Pour cela, il est essentiel d'utiliser un (ou des) moteur(s) de traduction, et éventuellement de rédaction et génération, automatique(s) éprouvé(s), entrainé(s) sur des corpus pertinents (par champ disciplinaire, par thème ou par langue, par exemple), et dont la précision devra s'améliorer dans le temps, au fur et à mesure des usages (chaque génération / traduction augmentant arithmétiquement le nombre de documents dans la base de données que le moteur moissonnera). Afin de garantir de la manière la plus objective possible cette qualité, certains critères formels pourront être mis en avant, tels que le réquisit de traduction de l'intégralité du résumé depuis la langue originale, l'identification de l'auteur, la traçabilité de la traduction (moteur utilisé, date, nom du traducteur en cas de post-édition), l'utilisation de glossaires apportant un certain label de qualité pour les termes techniques, etc.

Bénéfice utilisateur
Le bénéfice utilisateur désigne la valeur ajoutée que les propositions développées doivent apporter aux divers utilisateurs, qu'ils soient institutionnels ou simples particuliers : traducteurs, chercheurs, éditeurs… Le gain concerne les **fonctionnalités proposées** qui doivent **répondre à des besoins identifiés** (qu'ils soient formulés explicitement par les utilisateurs ou issus de l'observation de leurs pratiques) et **améliorer les gestes et pratiques des utilisateurs** (en améliorant des fonctions existantes encore mal réalisées ou en en proposant de nouvelles pour des besoins non pourvus). Voici quelques exemples de besoins identifiés : génération ou traduction automatique de résumés, amélioration de la découvrabilité (traduction des métadonnées, suivi de l'impact d'une publication, etc.) mais aussi systèmes permettant une reconnaissance professionnelle (pour les traducteurs), une valorisation réputationnelle (pour les chercheurs donnant un peu de leur temps pour effectuer la post-édition et améliorer les traductions - les leurs ou celles des autres - par exemple des badges, etc.), de mise en réseau (entre les utilisateurs qui offrent et ceux qui cherchent une aide à la traduction), de travail collaboratif (pour une post-édition collective, par exemple, voire une traduction humaine coopérative du contenu d'un article entier).

Souveraineté
C'est un des aspects clefs auquel doivent contribuer les pistes qui seront développées : la souveraineté, nationale ou européenne, soit la maîtrise des technologies et des données. Il s'agit notamment d'assurer un contrôle sur la thésaurisation et l'usage des textes et corpus créés par incrémentation. Une telle base de données documentaire ne doit pas pouvoir être moissonnée sans autorisation ou impunément. Il convient donc d'en réguler l'accès : les GAFAM ou autres concurrents, par exemple, ne doivent pas pouvoir aspirer ce savoir accumulé. C'est à cette condition que pourra être garantie une réelle autonomie et indépendance par rapport aux rivaux privés ou extra-européens, ainsi qu'une protection et une portabilité des données. Il est impératif que les questions liées à l'accès, l'autonomie et la souveraineté des données soient abordées en détail dans le cadre de la dernière étude du projet *Traductions et science ouverte*, portant sur les modèles opérationnels et économiques du service.

Pérennité
La pérennité désigne le caractère **durable** des documents produits, stockés, thésaurisés et manipulés au sein de la base de données choisie. Elle est décisive, tant pour les utilisateurs particuliers que pour les éditeurs, les institutions de recherche et les traducteurs, et implique plusieurs aspects :

- **l'interopérabilité** : les utilisateurs peuvent vouloir insérer ou récupérer un texte dans des formats très différents (.txt, .word, .pdf, .ppt, .html, etc). Il est essentiel de garantir une compatibilité afin que chacun puisse construire librement son propre environnement de travail (avec OpenOffice, Overleaf, etc). Nous pouvons également ajouter la possibilité de pouvoir à tout moment partir avec ses données (dans le format de son choix). Il est également fondamental d'assurer l'interopérabilité des données textuelles multilingues, en permettant des échanges dans les principaux format d'importation et exportation : tmx, tbx, xliff, csv ou tsv, etc.
- **l'incrémentation** : quelle que soit l'opération choisie (génération, traduction automatique, post-édition, etc.), il faut qu'une copie soit légalement déposée dans la base de données documentaire du système, dûment étiquetée (avec mention de la date, de l'auteur, du traducteur, du moteur, etc.). C'est en effet à cette condition qu'une logique vertueuse de thésaurisation pourra être mise en place, en donnant ainsi toute sa fécondité au projet. Cette croissance progressive permettra peu à peu des suggestions de navigation sur corpus, selon différents points de vue (comme en littérature comparée), en permettant des rapprochements inédits : par exemple, croiser les articles de géographie humaine et d'épidémiologie concernant le concept de « population », ou d'histoire et de science de la terre concernant le « climat », etc.
- **l'entraînement** : le moteur utilisé pour la traduction automatique et éventuellement l'aide à la rédaction et la génération de textes aura fait l'objet d'un pré-entraînement. Mais au fur et à mesure de l'incrémentation de la base de données textuelles, il devra être perfectionné de manière pertinente et raisonnée, par exemple par domaine disciplinaire, langue, et pourquoi pas selon d'autres critères (époque, registre de discours, etc.).
- **le stockage multilingue** (cf. infra) souple, permettant différents points de vue tant sur les items (les textes ou documents) que sur les ensembles d'items (corpus). La question de la durée du stockage mérite d'être prise en considération, notamment dans les disciplines où certaines informations sont susceptibles d'être considérées comme un peu dépassées, voire franchement obsolètes. Le cas des sciences sociales et humanités, qui entretiennent un rapport structurel à leur passé, est un peu différent, car la thésaurisation n'en est que plus importante. Le point le plus délicat concerne la norme juridique permettant le stockage: il convient de pouvoir en fixer la détermination de façon souple et évolutive, en tenant également compte des contraintes et des besoins techniques.

Multilinguisme
On pourrait envisager d'utiliser une technologie comme **CouchDB** (commode), même si une autre infrastructure du même type serait également légitime. Il s'agit d'un système de gestion de **base de données orientée documents**[1], par opposition à un système de gestion de base de données relationnelles (où l'information est organisée dans des tableaux à deux dimensions appelés des relations ou tables) ou de base de données de graphiques (où le stockage fournit une adjacence entre éléments voisins). En particulier, alors que, dans les bases de données relationnelles, un même objet peut apparaître dans différents tableaux, les bases de données orientées documents placent l'information concernant un certain objet à un seul endroit, de sorte que tout objet emmagasiné peut être différent des autres : cela est très utile pour traiter des textes indépendants les uns des autres.

Mais **le plus important n'est pas tant l'infrastructure que la manière dont elle sera conçue et utilisée**. Il faut éviter une manipulation rigide et **privilégier la souplesse**. En effet, le risque est que, en utilisant une base de données relationnelles, on voit tout comme des tableaux ; ou que, avec une base de données à base de graphes, on voit tout comme un graphe ; ou enfin que, avec une base de données à base de documents, on conçoive tout comme un document. Alors que, en réalité, rien n'empêche de « faire entrer » un graphe ou un document dans un tableau et inversement.

Mieux vaudrait donc **prêter une sorte de valeur heuristique aux documents de CouchDB** pour inviter les informaticiens à considérer les documents du traducteur. Cela implique de ne pas concevoir les segments stockés (les données) indépendamment des textes dont ils sont issus (qui constituent leur contexte), et de pouvoir toujours leur associer la marque de leur auteur (traçabilité). Autrement dit, les documents doivent pouvoir être **interprétés** par les différents utilisateurs qui doivent pouvoir comparer leurs points de vue (les différentes traductions). D'où l'idée d'associer à chaque item un nombre indéfini (n) de traductions possibles, et non de procéder par couple de langues. De même, les documents doivent pouvoir être **tracés**, afin de savoir quel(s) texte(s) précis ont servi de source à telle traduction. Dans cette perspective, le document est conçu comme le résultat d'un **geste de documentation**, **assumé** par un (ou plusieurs) auteur(s), et non comme une simple donnée (dont on ne sait en fait jamais qui la donne, ni comment et pourquoi).

1 https://fr.wikipedia.org/wiki/CouchDB

02

UN MOTEUR SPÉCIALISÉ ACCESSIBLE VIA DIFFÉRENTES MODALITÉS

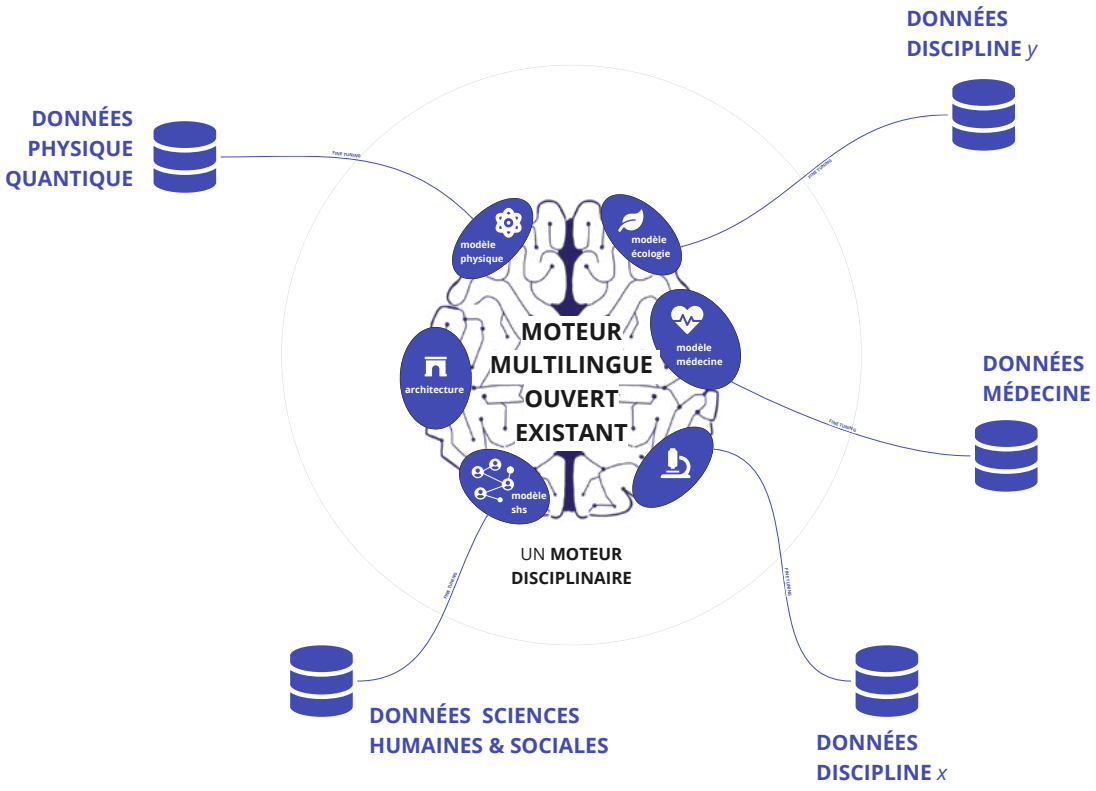
a_ Un moteur de traduction et rédaction automatique ouvert & spécialisé

Une piste qui semblait essentielle à tous les acteurs impliqués est celle d'un moteur de traduction et rédaction automatique spécialisé, une des principales raisons étant d'être en mesure de proposer **une alternative à la concurrence, plus transparente et respectueuse des données utilisateurs.**

Ce moteur est proposé pour :

- **en priorité : traduire**,
- mais aussi à terme : pour **rédiger**, voire **créer des synthèses** ; les modèles de type GPT3, pour citer le plus connu, peuvent en effet écrire et traduire. Il pourrait ainsi constituer une aide à l'écriture de résumé, « en français facile » ou en anglais standard (« machine »). L'aide à la rédaction ainsi que la génération automatique ont été remontées lors des entretiens et ateliers utilisateurs ; par exemple dans certaines disciplines, la rédaction directement dans une langue cible est la pratique ; cette dernière pourrait être aidée par un moteur et permettre d'augmenter la qualité de la rédaction.

Illustration de principe d'un moteur multilingue, entrainé avec des données spécialisées.



Parmis les principales caractéristiques retenues pour ce moteur nous recommandons qu'il soit :

- **ouvert** (le moteur et ses données d'entraînement),
- **spécialisé** (entraîné sur des données scientifiques disciplinaires de qualité),
- **multilingue**.

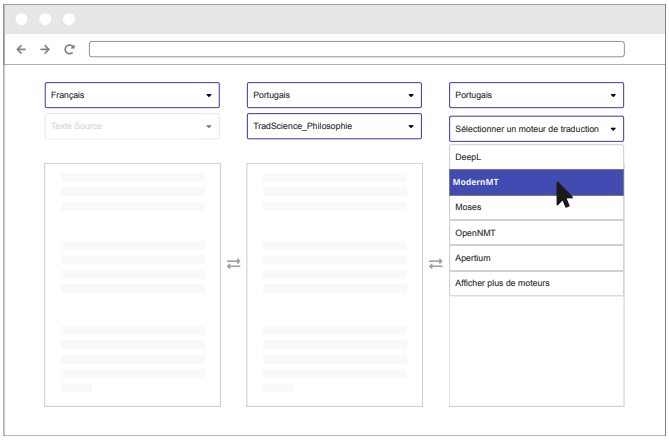
Dans un souci de **gain de temps et de budget**, il nous semble pertinent de **s'appuyer sur un moteur existant**, à spécialiser à partir de données issues de publications scientifiques et disciplinaires. Ce processus de spécialisation est appelé *fine-tuning* en anglais. Les données peuvent inclure des corpus monolingues, bilingues, multilingues, des glossaires, des mémoires de traductions, etc. Parmi les moteurs existants actuels, nous pouvons noter entre autres **LightOn**, **Leam**, ou encore **Bloom**. Il est souhaitable de privilégier des moteurs *open source*.

Ce moteur, comme nous le détaillons dans la suite, peut **être accessible** :

- **directement** par une interface de traduction automatique simple (que nous suggérons de créer en même temps),
- **indirectement**, en s'intégrant dans un écosystème d'outils de traduction assistée (OmegaT, Matecat, Trados, etc.), de plateformes de diffusion (HAL, OpenEdition, Thèses.fr, etc.), voire d'éditeurs de texte (Word, OverLeaf, etc.).

Autres recommandations :

- Les produits doivent être **pérennes** (il ne faut pas dépendre d'un moteur qui perdra son financement dans un an ou deux, ou qui abandonnera son statut open source brusquement).
- Les productions doivent être **interopérables** (le texte traité ou traduit doit pouvoir être importé depuis et exporté vers les formats de fichier métier et grand public les plus communs).
- Un ou plusieurs **corpus** doivent pouvoir être construits par incrémentation.
- Les productions doivent bénéficier d'une **traçabilité** impeccable (« traduit / généré avec LSPDIDR », LeServicePublicDiffusionInternationaleDeLaRecherche).
- Dans la construction de ce commun numérique, la **transparence** sur l'usage qui est fait des données doit être le mot d'ordre.
- Les résultats doivent se caractériser par la **précision** et la **pertinence**.
- **Définir en amont le périmètre du moteur** pour **concentrer les efforts de mise en œuvre** tout en étant en mesure de garantir la qualité de ses propositions ; il s'agit de faire très bien sur un contexte précis plutôt que de faire moyennement sur un contexte large. Nous recommandons donc dans un premier temps de limiter les disciplines et les langues.
- Faire **tester le moteur** par des acteurs différents.



Esquisse de l'interface de traduction automatique que nous proposons de développer en même temps que le moteur (piste détaillée juste après).

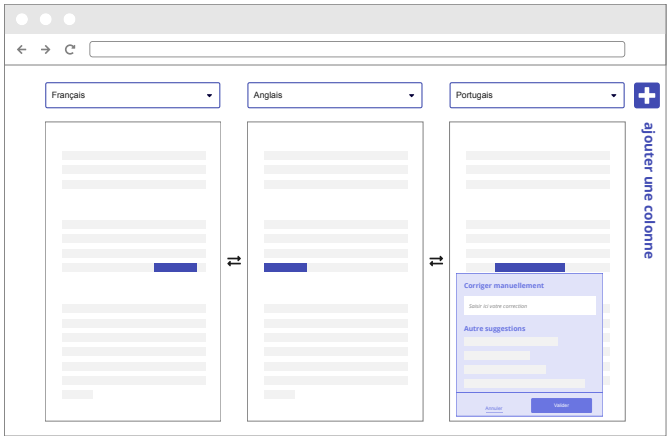
b_ Une interface simple de traduction automatique pour utiliser et valoriser le(s) moteur(s) simplement et directement

L'objectif de cette interface simple de traduction est de permettre une **utilisation intuitive et directe du moteur spécialisé**.

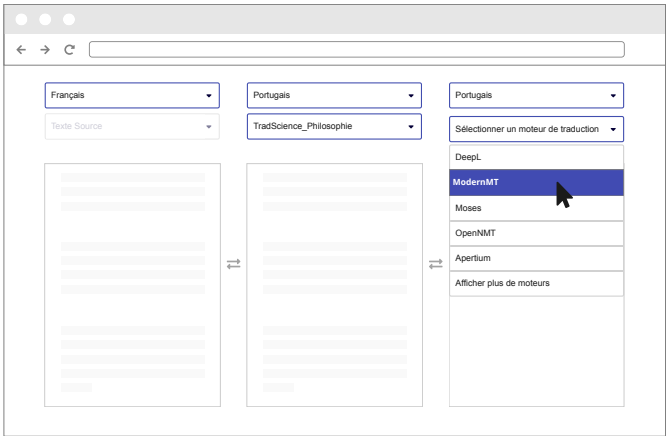
Nous proposons que cette interface **se distingue de l'existant par différentes fonctionnalités et partis pris** :

- **comparaisons multilingues** (pour comparer, par exemple, des langues pour lesquelles on obtient de bons résultats avec d'autres langues dont les résultats sont moins bons),
- **comparaison** entre **plusieurs moteurs** existants,
- **chargement de glossaires disciplinaires** pour permettre à la fois d'améliorer la performance de la traduction automatique mais aussi le geste (humain) de traduction (par un traducteur, chercheur…) ; en effet ces glossaires vont également permettre aux personnes traduisant d'avoir une idée plus claire de la qualité et, éventuellement, mieux interpréter la traduction générée automatiquement.

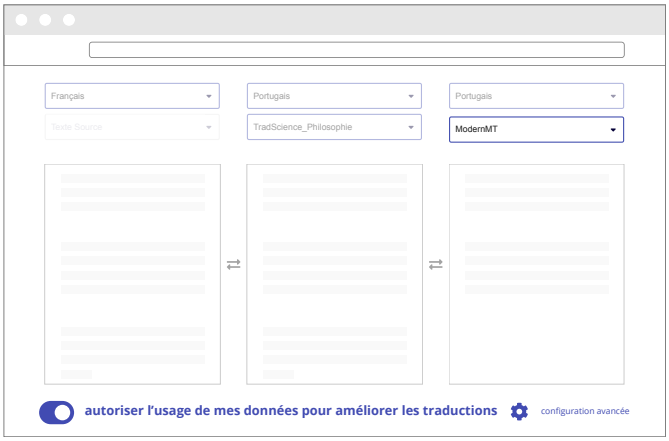
Ci-dessous quelques **illustrations mettant en avant les fonctionnalités qui nous semblent clefs** pour cette interface simple de traduction (les points d'attention sont représentés en bleu).



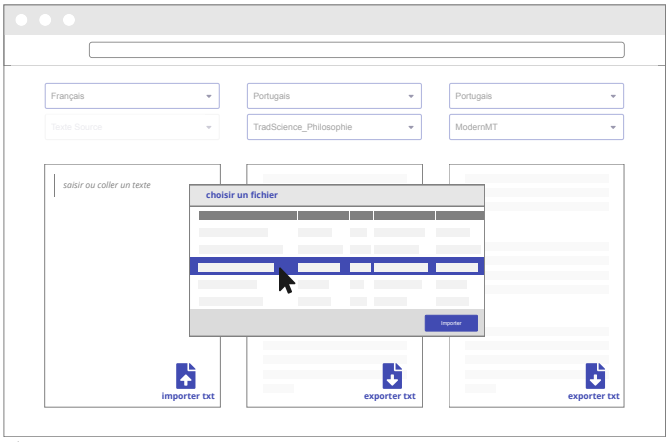
*Une interface de traduction automatique **multilingue** permettant de **comparer la qualité de la traduction selon les langues** (par exemple : meilleure en anglais qu'en portugais). Il est alors possible de **corriger / post-éditer les textes manuellement** ou à partir de suggestions.*



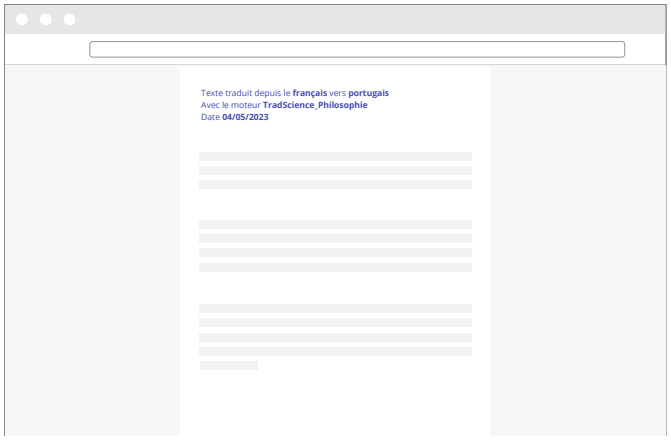
*Une interface de traduction automatique permettant de **comparer plusieurs modèles** (par exemple : TradScience, ModernMT, DeepL, Apertium, etc.). Pour pouvoir utiliser un moteur payant directement depuis l'interface, il est nécessaire de donner à l'utilisateur la possibilité de « connecter son compte » depuis une page dédiée.*



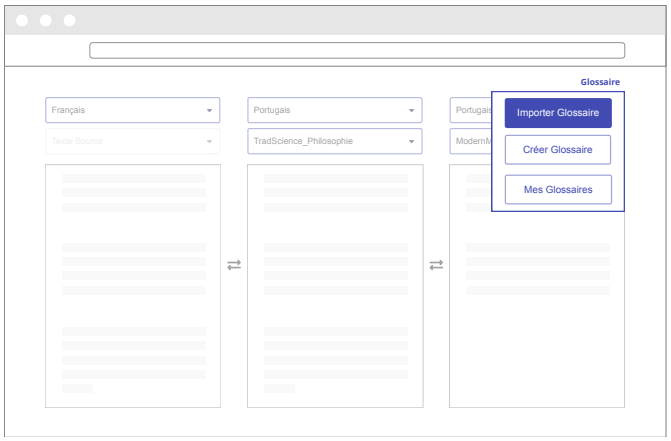
*Garder à tout moment **la maîtrise de l'usage qui est fait de vos données !***



***Importer et exporter des textes** directement dans l'interface sous différents formats (pdf, word, LaTeX, etc.).*



*À l'export, les fichiers textes sont marqués comme étant **traduits avec le moteur TradScience !***



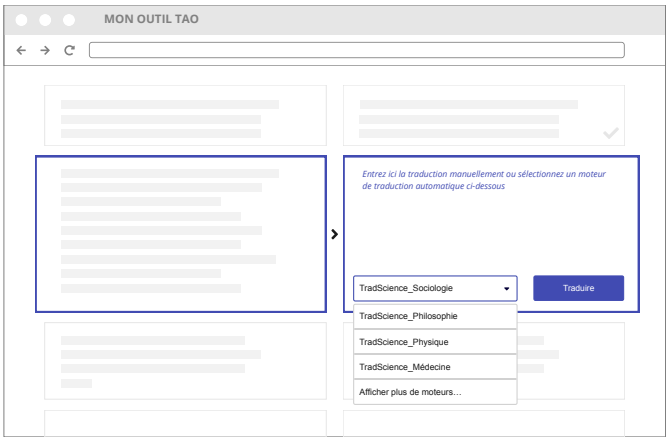
*Une interface de traduction automatique permettant **d'importer des glossaires disciplinaires, de les créer ou de gérer les siens** (en mode connecté), ceci afin d'améliorer la traduction. Ces glossaires sont **également accessibles depuis l'interface de dépôt et consultation de ressources linguistiques** suggérée et précisée dans la partie « thésaurisation de données ».*

c_ Intégration du moteur de traduction automatique dans « mes outils de TAO préférés » (plug-in)

Il s'agit ici de donner la possibilité aux traducteurs d'utiliser le moteur de traduction automatique depuis leur(s) outil(s) habituel(s) (OmegaT, MateCat, Trados Studio, memoQ, etc.).

Un point d'attention est alors de faire en sorte qu'en retour de l'usage par les traducteurs, le moteur puisse bénéficier des corrections manuelles pour s'améliorer : c'est la formule « gagnant-gagnant » ou « donnant-donnant » employée plus haut dans ce rapport.

Voici ci-dessous quelques **illustrations mettant en avant les fonctionnalités qui nous semblent clefs** (les points d'attention sont représentés en bleu).



*Depuis **mon outil de TAO préféré**, **sélection du moteur de traduction automatique** (parmi l'ensemble des champs disciplinaires existants) pour traduire **segment par segment ou la totalité du texte.***

d_ Intégration du moteur de traduction automatique dans les plateformes de diffusion de publications scientifiques pour la traduction des résumés

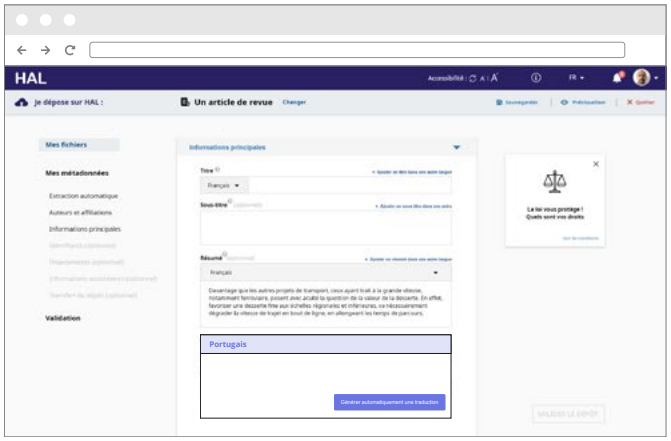
Afin de répondre à l'enjeu stratégique de rayonnement de la recherche dans plusieurs langues, mentionné plus haut, la **traduction des résumés** semble essentielle pour **améliorer la découvrabilité** ; cette traduction va en effet améliorer **l'indexation**, mais aussi aider à **parcourir et mieux comprendre** dans les grandes lignes les publications.

Cette traduction facilitée de résumés, nous suggérons de l'intégrer aux plateformes de diffusion de publications scientifiques. Une intégration en particulier est perçue comme prioritaire, celle dans HAL pour traduire les résumés :

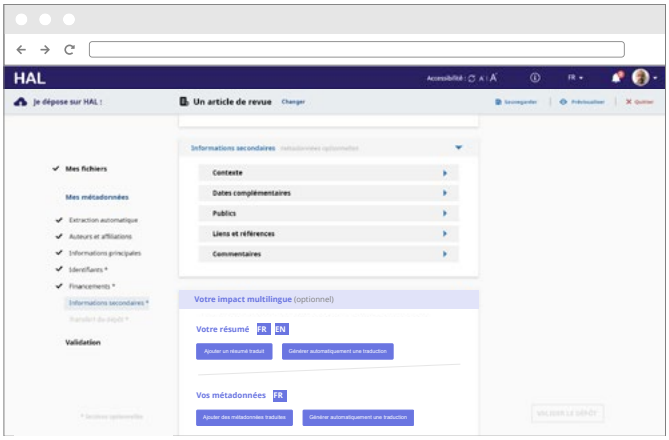
- **au dépôt** (aidant le chercheur à améliorer la découvrabilité de son texte en d'autres langues sans ajouter un coût pour l'auteur),
- **à la consultation** (pour les résumés qui n'auront pas été traduits au dépôt).

Il est également possible d'envisager **plus tard la génération de synthèses automatique** à la consultation de HAL.

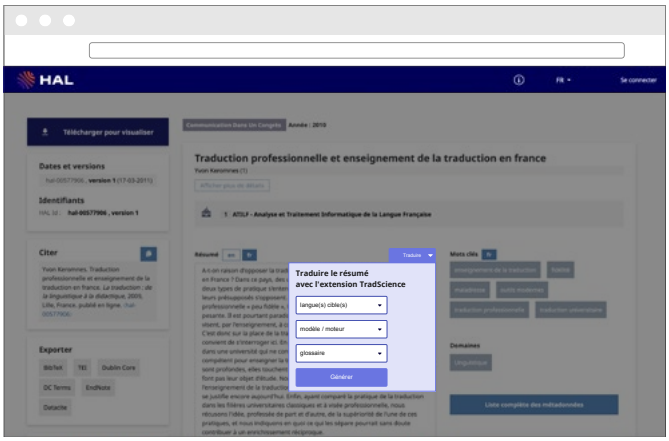
Voici ci-dessous **quelques illustrations mettant en avant les fonctionnalités qui nous semblent clefs** (les points d'attention sont représentés en bleu).



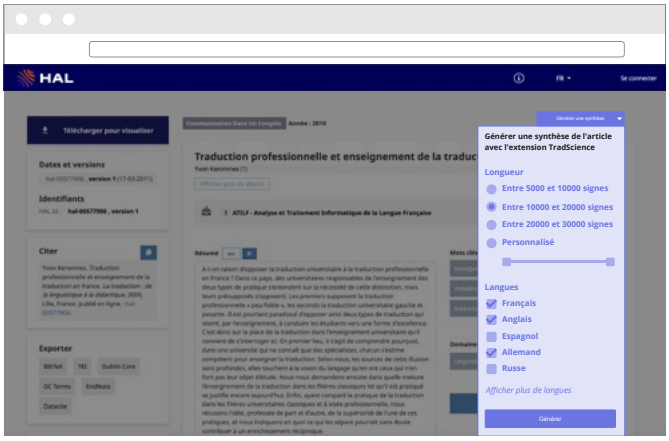
Traduire automatiquement *à l'aide du moteur* ***le résumé*** *dans d'autres langues* ***au moment du dépôt*** *dans HAL. Laisser la possibilité de relire et corriger la traduction proposée. Il est fondamental de garder et d'afficher l'information sur le statut de la traduction : a-t-elle été produite par le moteur ? Dans ce cas, un travail de post-édition a-t-il été fait ?*



Augmenter son impact multilingue *en ajoutant* ***des résumés dans d'autres langues*** *ou en* ***traduisant les métadonnées****.*



Traduire automatiquement le résumé *et les* ***métadonnées*** *d'une publication* ***à la consultation*** *de sa notice sur HAL, pour* ***mieux cerner le sujet****.*



Générer automatiquement la synthèse *d'une publication* ***lors de la consultation*** *de sa notice sur HAL.*

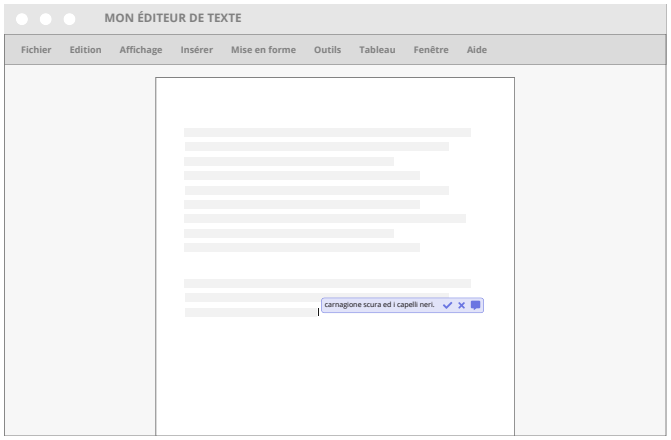
e_ Intégration de l'aide à l'écriture dans mes outils de rédaction préférés (plug-in, API, etc.) : Word, Overleaf…

Cette fonctionnalité d'aide à l'écriture a été retenue mais identifiée comme **moins prioritaire**.

Cette piste vise à **exploiter les possibilités de génération de texte des modèles** (de langage) **pour faire des suggestions de rédaction** aux auteurs. À noter que la pertinence de ces suggestions va dépendre des disciplines scientifiques et sera a priori meilleure pour celles plus « techniques ». Le plus simple pour intégrer cette aide à l'écriture, mais aussi le meilleur parti pour garantir son succès auprès des utilisateurs est de proposer cette **fonctionnalité de façon intégrée aux outils de rédaction habituels** (Word, OverLeaf, etc.).

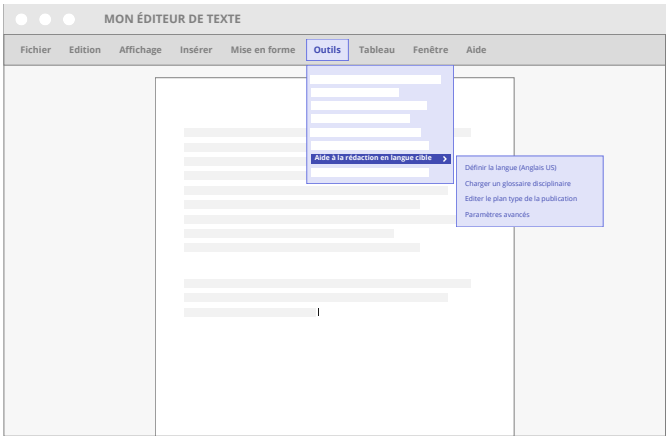
À noter également qu'à terme l'aide à l'écriture peut **intégrer des suggestions de structures types d'articles ou d'ouvrages, selon les disciplines, éditeurs ou revues**… La plateforme *TooWrite* peut servir d'inspiration sur ce point : elle aide les auteurs à structurer leur rédaction en fonction du contexte de publication.

Voici ci-dessous **quelques illustrations mettant en avant les fonctionnalités qui nous semblent clefs** (les points d'attention sont représentés en bleu).



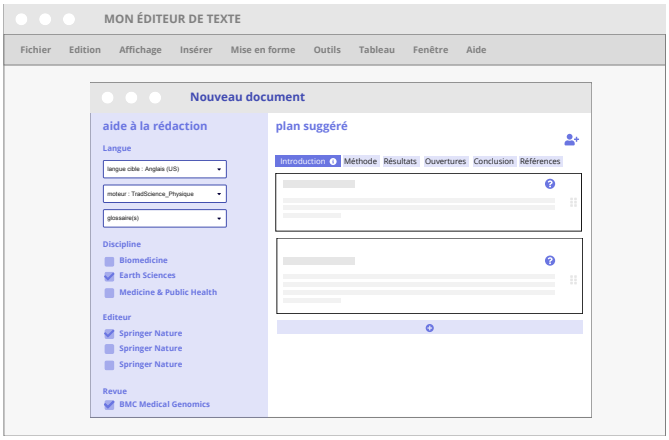
▲

*Depuis **mon éditeur de texte préféré** (Word, OverLeaf, etc.), **suggestions de formulations habituelles dans ma discipline**, dans la langue cible. Les **suggestions peuvent être acceptées ou refusées** — ces choix peuvent **contribuer à l'amélioration des nouvelles suggestions** faites dans le document.*



▲

*Les **paramètres d'aide à la rédaction peuvent être ajustés à tout moment** ; il est possible notamment de **charger un nouveau glossaire disciplinaire** à tout moment.*



▲

*Depuis **mon éditeur de texte préféré** (Word, OverLeaf, etc.), possibilité de **créer des trames habituelles adaptées aux différents contextes** (éditeur, discipline, etc.). Les trames **suggérées automatiquement** en fonction des paramètres d'entrée **peuvent être ajustées à la main à tout moment**.*

03

UNE THÉSAURISATION DES DONNÉES

a_ Des données / ressources scientifiques disciplinaires à agréger

Les **données** sont à la **base de tout apprentissage d'intelligence artificielle** et leur qualité constitue un facteur déterminant pour obtenir des résultats optimaux. Mais ces données n'ont pas qu'une valeur à travers le moteur : elles sont **aussi précieuses pour différents acteurs, par exemple les traducteurs qui vont les consulter et s'en servir pour améliorer la cohérence de leurs traductions**.

Dans tous les cas, **la thésaurisation de données scientifiques disciplinaires est identifiée comme une priorité** par l'ensemble des acteurs ayant été impliqués dans le projet.

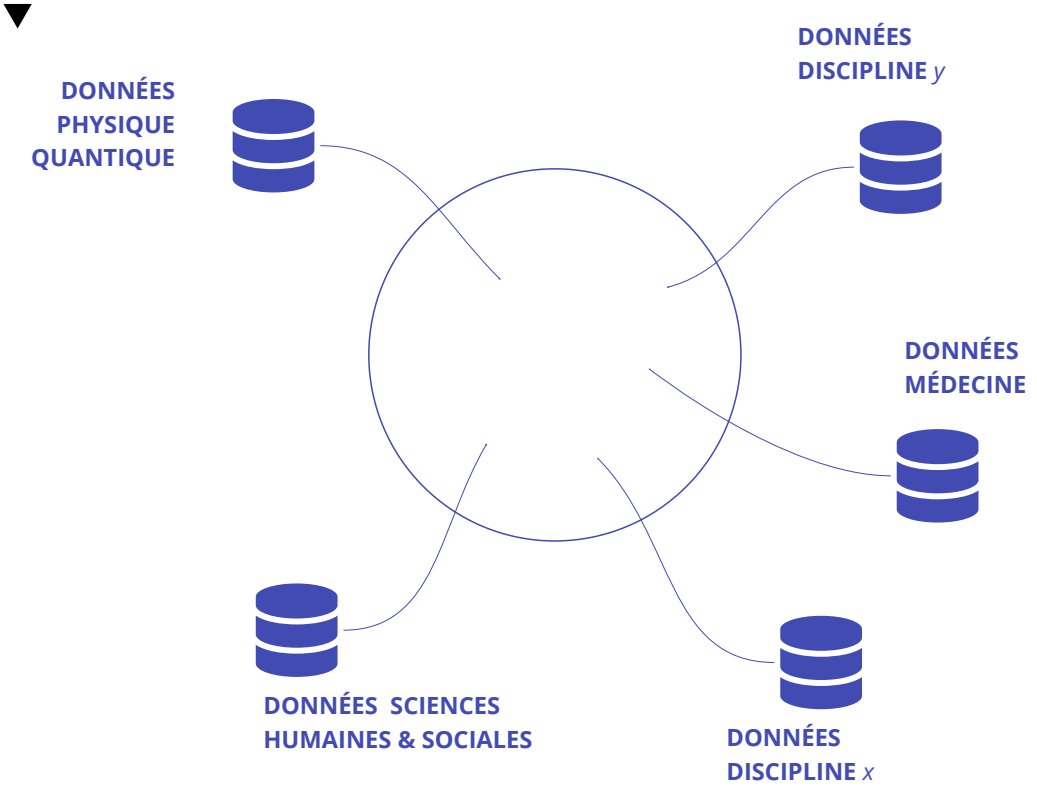
Cette **thésaurisation des données doit pouvoir se faire dans un système d'archivage d'emblée multilingue** (cf. supra).

Cette **thésaurisation de données** ressources peut être **localisée** (une plateforme de collecte et d'agrégation, des appels à projets contributifs) et **ubiquitaire** (un ensemble d'outils dédiés et en partenariat qui permettent de collecter des données).

Nous précisons ainsi dans la suite les modalités d'accès à ces ressources :

- les **appels à contribution** (pour atteindre une quantité / qualité de données suffisante),
- la **plateforme de contribution / de lecture de données**,
- et, de façon plus **ubiquitaire, l'usage des différentes plateformes et plug-ins** mis à disposition des utilisateurs (déjà mentionnés plus haut dans la partie « Moteur spécialisé »).

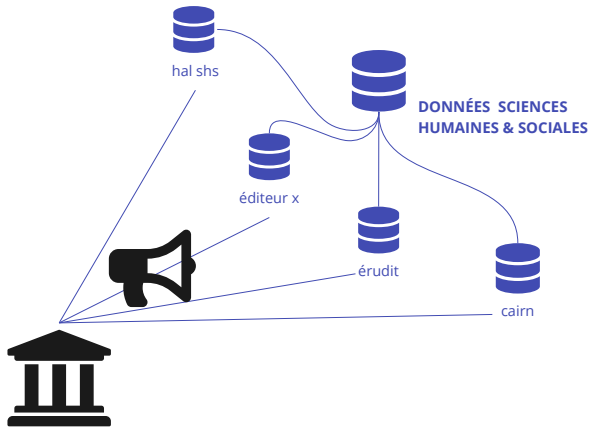
Illustration de principe d'une agrégation de données spécialisées.



b_ Des appels à contribution publics pour participer à la thésaurisation de données scientifiques disciplinaires

Des appels à contribution des éditeurs pour **constituer des bases de connaissances** :

- à partager avec / entre les chercheurs, traducteurs et éditeurs,
- afin d'entrainer le moteur avec des corpus de données disciplinaires de bonne qualité scientifique et d'un volume suffisant.



▲

Illustration d'appels à contribution pour constituer des bases de connaissances minimales sur des thèmes / disciplines.

c_ Une plateforme de dépôt et consultation de ressources linguistiques éditorialisées

Mix de :

- service de création / de partage de base de connaissances,
- base lexicale européenne disciplinaire.

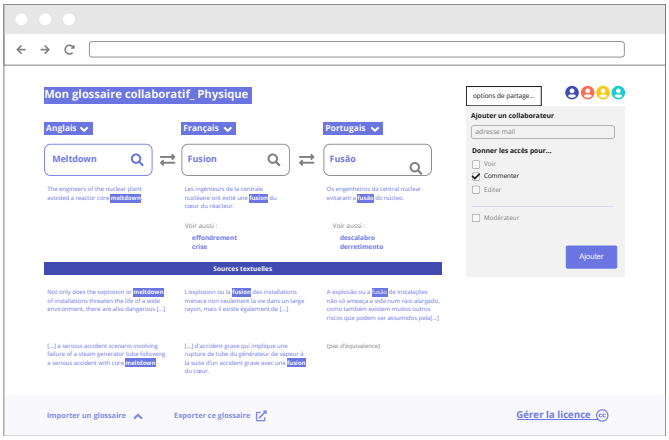
Avec :

- **glossaire open source**,
- **mémoires de traduction**,
- **traduction** (accès sélectifs),
- **corpus pour certains partenaires** (par exemple, ceux qui acceptent la mutualisation, avec un accès qui peut être restreint pour les autres).

Une curation des ressources pourrait être effectuée (par exemple par le MESR) pour garantir la qualité scientifique de la plateforme.

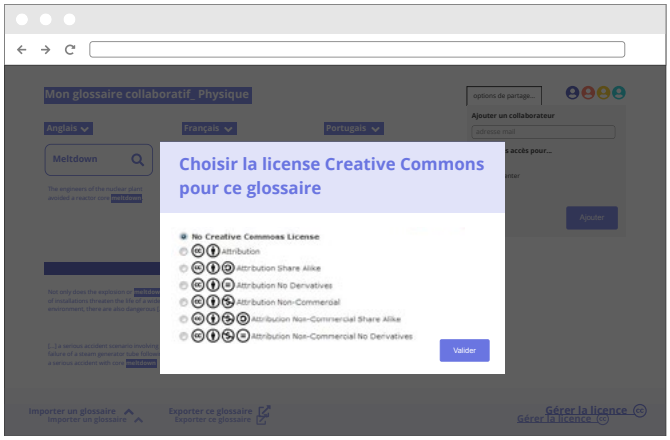
Nous présentons ici cette interface de dépôt et consultation comme une plateforme autonome, afin de l'identifier clairement la fonctionnalité ; en pratique, dans un avenir proche il est probable qu'il soit préférable de **« fusionner » cette interface avec celle de traduction automatique** présentée un peu plus haut pour en faire **une seule et même plateforme « intégrée ».**

Voici ci-dessous quelques **illustrations mettant en avant les fonctionnalités qui nous semblent clefs** pour cette plateforme (les points d'attention sont représentés en bleu).

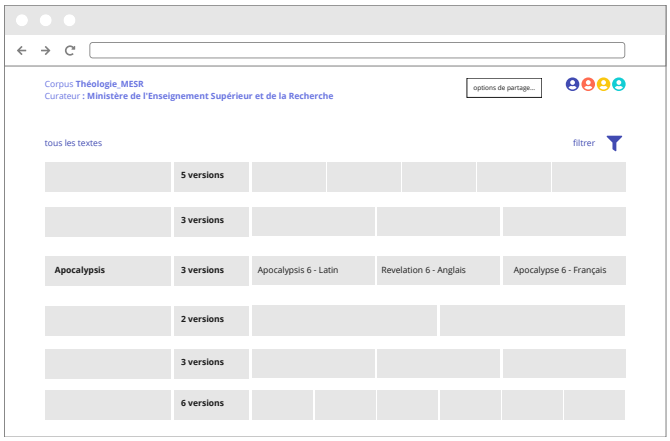


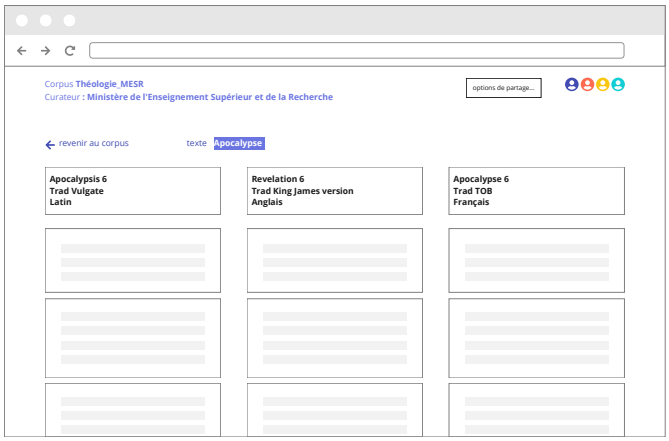
▲

***L'interface de dépôt / consultation** permet de **consulter** des **glossaires**, de **les créer seuls ou à plusieurs**, de **les partager en ligne** avec des collaborateurs sélectionnés, de **les exporter**, ou encore **d'en importer** depuis d'autres sources sous différents formats (par exemple : .csv).*



*Des **licences peuvent être appliquées** sur chacun des glossaires pour en **définir les droits d'utilisation et de partage**.*

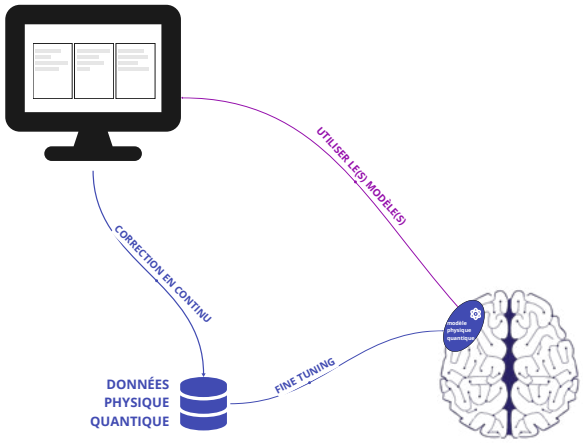




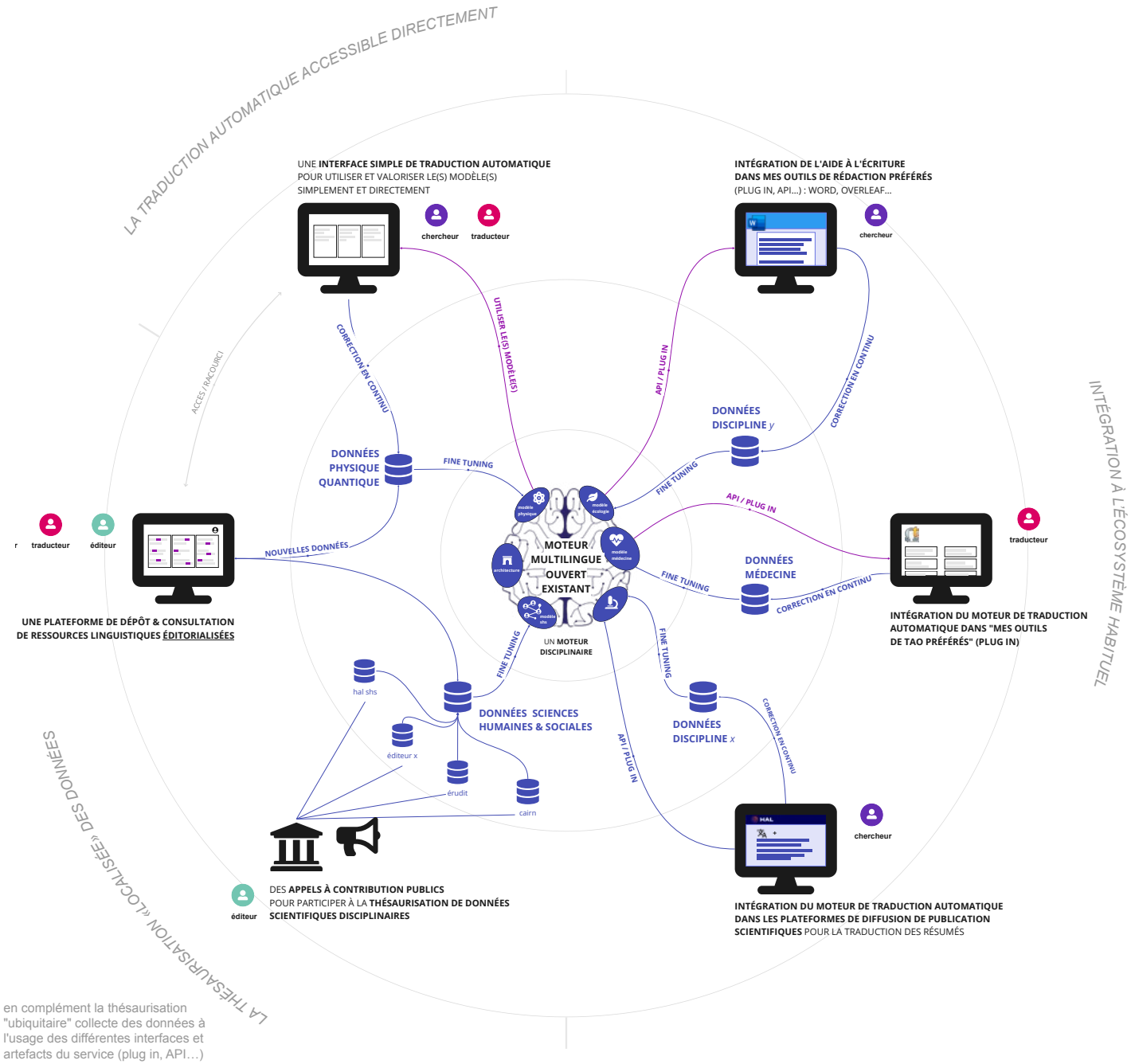
*La plateforme permet de consulter des corpus multilingues. Ces derniers peuvent être **partagés avec des personnes sélectionnées** par le propriétaire.*

d_ Collecte et correction ubiquitaire des données

L'usage des services fournis aux utilisateurs peut permettre de **corriger et de compléter les données au fil de l'eau**. Cette amélioration des données est restituée aux utilisateurs soit par l'intermédiaire de la plateforme d'accès aux données, soit par l'usage du moteur dont la performance est ainsi augmentée (*fine-tuning* régulier).



*Avec l'**accord explicite de l'utilisateur du moteur de TA**, ses données peuvent **contribuer à corriger / améliorer le corpus et ainsi améliorer les performances du moteur**. Ces **données sont également directement accessibles aux utilisateurs** via l'interface d'accès aux données.*



▲

Architecture fonctionnelle de principe des pistes retenues (détaillée).

04

FEUILLE DE ROUTE MACRO

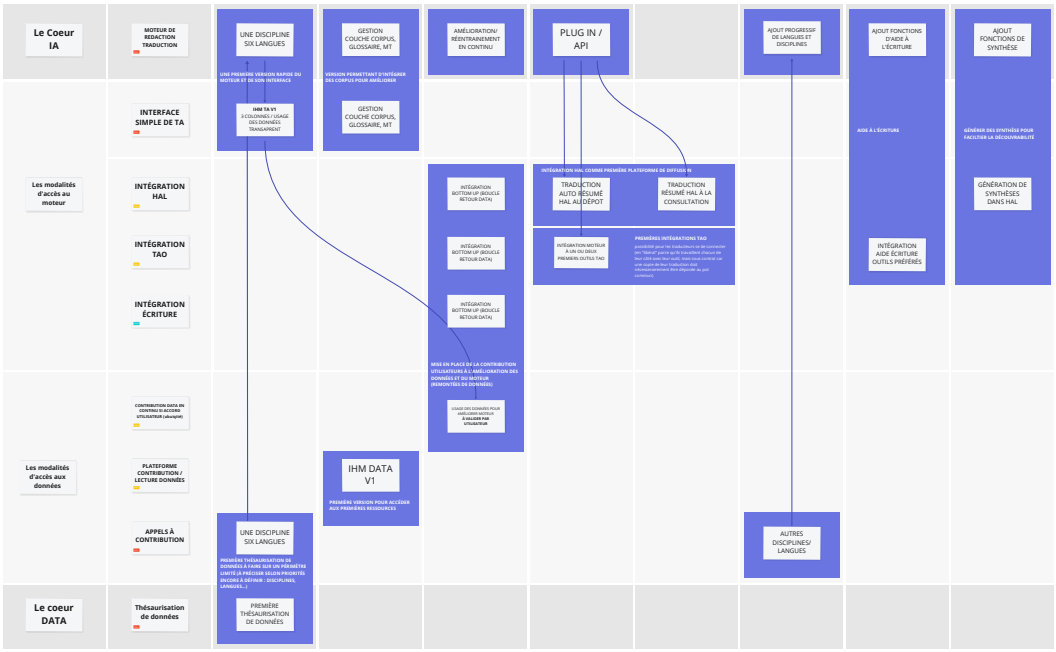
Nous proposons ici une première feuille de route macro afin de prioriser les développements des propositions. De façon très synthétique, il semble **prioritaire de lancer** :

- les développements sur **une première version du moteur de traduction pluridisciplinaire**. Cette première version doit se faire sur la base de premiers corpus avec un périmètre encore à définir : par exemple une discipline (peut-être que les Sciences humaines et sociales ont des besoins plus singuliers et urgents), et quelques langues (par exemple le français, l'anglais, l'espagnol, et quelques autres langues à la demande).
- le développement d'**une interface associée au moteur** pour incarner immédiatement ce dernier et le mobiliser directement. Pour cette première version nous suggérons quelques fonctions clefs, différenciantes de la concurrence (comparaison de modèles de TA, multilingue, importation de glossaires…) ; cette identification de fonctions initiales demande encore d'être précisée en contact avec des utilisateurs (entretiens et sessions collectives de co-conception).
- les **premiers appels à contribution pour la collecte de données disciplinaires** ; ces données sont nécessaires à l'entraînement du moteur pour le spécialiser.

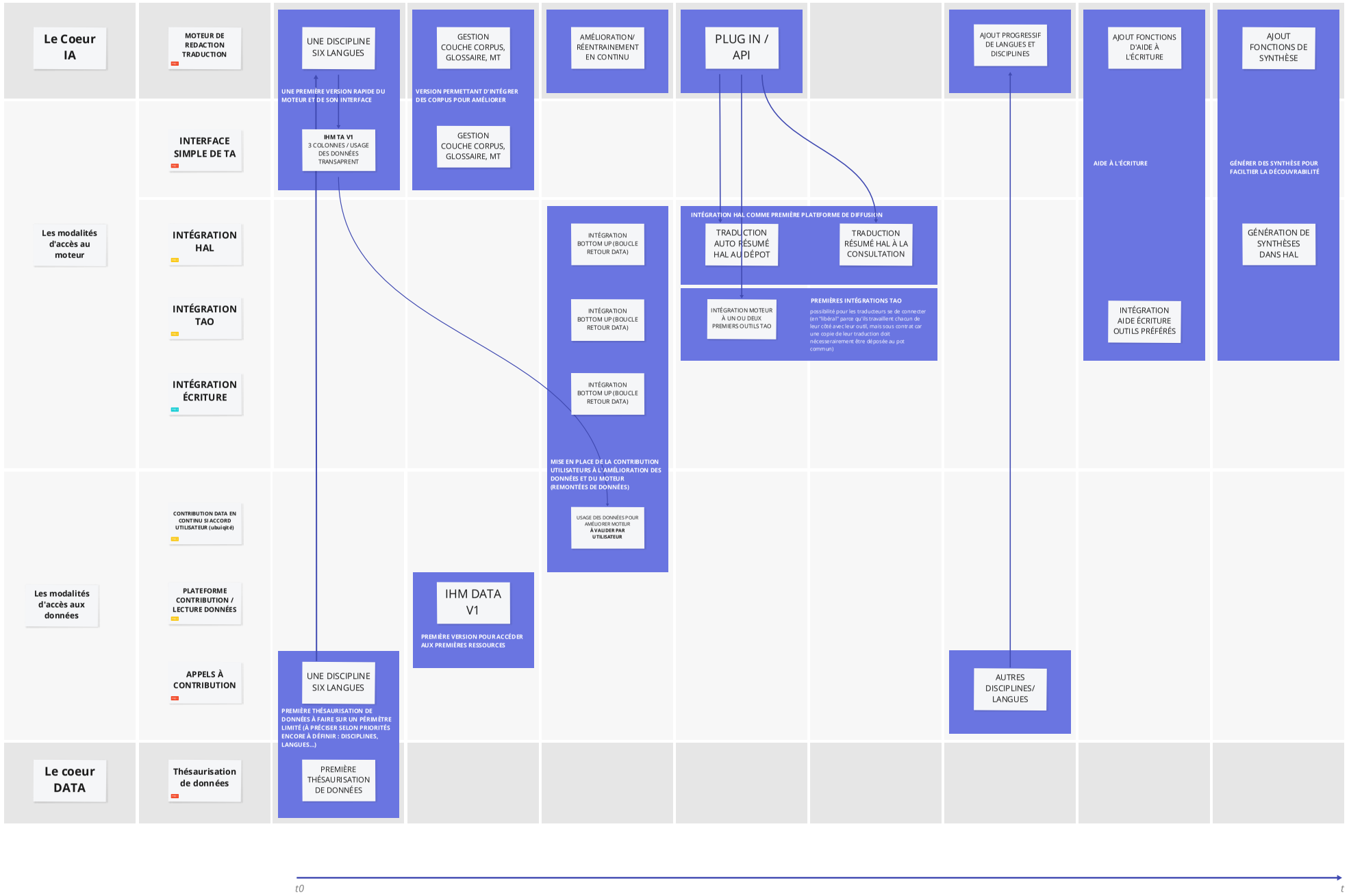
Quelques points ont été abordés tout au long du projet concernant les évolutions possibles tels que, **par exemple, la possibilité de rajouter par la suite un outil collaboratif.** Cependant, il en existe déjà beaucoup, de différentes sortes (traitement de texte, dictionnaire, correcteurs divers, mémoire de traduction, glossaire, interface à plusieurs fenêtres, etc.), et chaque communauté utilise le sien (Scrivener, LaTeX, etc.). Sa construction n'est donc pas prioritaire, sans être non plus exclue. Il faut commencer par un noyau automatique et voir quelles fonctionnalités collaboratives sont susceptibles d'être ajoutées, selon l'utilité qu'elles peuvent avoir, l'adoption (ou le rejet) qu'elles suscitent.

Quelques exemples :

- la **post-édition multilingue collaborative** (faire appel à quelqu'un pour vérifier la traduction de son résumé dans une langue que l'on connaît moins bien, voire pas du tout) ;
- **outil de TAO « léger »**, un **Linguee de la communication scientifique** (dictionnaire contextuel) ;
- possibilité de **corriger la traduction**, car cela améliore le service dans lequel on a confiance ;
- en **bonus** : la **traduction collaborative des contenus**, si cela correspond par exemple au désir des éditeurs, et si certains traducteurs acceptent d'adopter l'interface.



Zoom page suivante


Le Coeur IA
MOTEUR DE REDACTION TRADUCTION
UNE DISCIPLINE SIX LANGUES
GESTION COUCHE CORPUS, GLOSSAIRE, MT
AMÉLIORATION/ RÉENTRAINEMENT EN CONTINU
PLUG IN / API
AJOUT PROGRESSIF DE LANGUES ET DISCIPLINES
AJOUT FONCTIONS D'AIDE À L'ÉCRITURE
AJOUT FONCTIONS DE SYNTHÈSE
UNE PREMIÈRE VERSION RAPIDE DU MOTEUR ET DE SON INTERFACE
VERSION PERMETTANT D'INTÉGRER DES CORPUS POUR AMÉLIORER
Les modalités d'accès au moteur
INTERFACE SIMPLE DE TA
IHM TA V1
3 COLONNES / USAGE DES DONNÉES TRANSAPRENT
GESTION COUCHE CORPUS, GLOSSAIRE, MT
AIDE À L'ÉCRITURE
GÉNÉRER DES SYNTHÈSE POUR FACILTIER LA DÉCOUVRABILITÉ
INTÉGRATION HAL
INTÉGRATION BOTTOM UP (BOUCLE RETOUR DATA)
INTÉGRATION HAL COMME PREMIÈRE PLATEFORME DE DIFFUSION
TRADUCTION AUTO RÉSUMÉ HAL AU DÉPOT
TRADUCTION RÉSUMÉ HAL À LA CONSULTATION
GÉNÉRATION DE SYNTHÈSES DANS HAL
INTÉGRATION TAO
INTÉGRATION BOTTOM UP (BOUCLE RETOUR DATA)
INTÉGRATION MOTEUR À UN OU DEUX PREMIERS OUTILS TAO
PREMIÈRES INTÉGRATIONS TAO
possibilité pour les traducteurs se de connecter (en "libéral" parce qu'ils travaillent chacun de leur côté avec leur outil, mais sous contrat car une copie de leur traduction doit nécessairement être déposée au pot commun)
INTÉGRATION AIDE ÉCRITURE OUTILS PRÉFÉRÉS
INTÉGRATION BOTTOM UP (BOUCLE RETOUR DATA)
MISE EN PLACE DE LA CONTRIBUTION UTILISATEURS À L'AMÉLIORATION DES DONNÉES ET DU MOTEUR (REMONTÉES DE DONNÉES)
USAGE DES DONNÉES POUR AMÉLIORER MOTEUR À VALIDER PAR UTILISATEUR
Les modalités d'accès aux données
CONTRIBUTION DATA EN CONTINU SI ACCORD UTILISATEUR (ubuiqité)
PLATEFORME CONTRIBUTION / LECTURE DONNÉES
IHM DATA V1
PREMIÈRE VERSION POUR ACCÉDER AUX PREMIÈRES RESSOURCES
APPELS À CONTRIBUTION
UNE DISCIPLINE SIX LANGUES
PREMIÈRE THÉSAURISATION DE DONNÉES À FAIRE SUR UN PÉRIMÈTRE LIMITÉ (À PRÉCISER SELON PRIORITÉS ENCORE À DÉFINIR : DISCIPLINES, LANGUES...)
AUTRES DISCIPLINES/ LANGUES
Le coeur DATA
Thésaurisation de données
PREMIÈRE THÉSAURISATION DE DONNÉES
t0
t

CONCLUSION DU RAPPORT

Conclusion générale

La réflexion conduite par différents acteurs, à l'issue de plusieurs ateliers, montre que **l'attente** concernant **un service commun de diffusion scientifique multilingue**, passant à la fois par la traduction et la génération de textes, **est réelle**, malgré certaines appréhensions. La **prudence** semble être de mise, en dépit d'un sentiment d'urgence lié à la médiatisation des progrès de l'intelligence artificielle. Comment fédérer les énergies, thésauriser les résultats, mutualiser les risques et les financements, tout en **assurant la confiance** des divers acteurs, et en garantissant la pérennité (donc l'évolutivité) et l'interopérabilité des documents stockés ? Comment initier la dynamique d'un système « donnant-donnant » où chaque acteur accepte de céder certaines de ses contributions en échange de l'accès à celles des autres ? Comment assurer la pérennité et l'interopérabilité des documents ainsi conjoints ? Quelle utilisation de l'intelligence artificielle peut être faite dans cette perspective, et selon quelles conditions et limites ?

Trois principes d'action s'imposent : **agilité**, **compacité**, **pluralité**.

Agir vite, d'abord, dans la mesure où il s'agit de répondre à une demande croissante pour l'aide à la diffusion, à une époque où les développements des applications de l'intelligence artificielle (à condition d'être bien compris et maîtrisés) sont tout à fait prometteurs. Il s'agit d'agir par la traduction ou la génération automatique de métadonnées (notamment les résumés). De façon très utile et féconde, la France pourrait donner à peu de frais un exemple d'ingéniosité, de savoir-faire et d'ouverture scientifique.

Construire un **outil de petite dimension**, modeste par sa taille, mais ambitieux par ses visées, puisqu'il s'agit de mettre à disposition des chercheurs, éditeurs et traducteurs une interface (ou des plug-ins) permettant de donner rapidement, pour une publication, une visibilité dans d'autres espaces linguistiques et culturels (par la traduction, la génération, l'indexation). Cet outil compact, d'emblée placé dans une perspective de développement durable de la diffusion, pourra éventuellement se développer par l'adjonction d'autres fonctionnalités (par exemple collaboratives : glossaire, interface de traduction multilingue, etc.) si le besoin s'en fait sentir.

Inventer un **système pluriel**, parce qu'il s'agit de pouvoir stocker d'emblée des versions multilingues pour chaque document inséré, permettant ainsi une croissance indéfinie du nombre de langues de destination, sans difficulté technique. Au contraire, une conception trop courte, par couples de langues, risque de poser rapidement des problèmes de crise de croissance, dès lors que l'on voudra rajouter à l'anglais initial de nouvelles langues (chinois, arabe, espagnol, portugais, etc.). À noter que la solution technique retenue devra nécessairement être complétée par une réflexion sur les modalités juridiques touchant la production, la transformation, la thésaurisation et la consultation des documents, afin de garantir au mieux les objectifs de confiance et de traçabilité.

LE COMMENTAIRE DE CLAIRE LARSONNEUR

La science, ou plus exactement les sciences, ne sont pas une manufacture de contenu comme les autres, mais au contraire **une activité bien spécifique**. Car et depuis que les fondements de la science moderne ont été posés au XVIIe siècle, la clef d'une contribution scientifique ne réside ni dans son contenu, ni dans la personnalité du ou des auteurs, mais dans le partage des hypothèses, des protocoles expérimentaux et des résultats avec une communauté de chercheurs qui vont s'en emparer, les tester, les discuter. Ce serait une erreur de mettre l'accent sur les contenus et leurs transcriptions dans d'autres langues en oubliant que le cœur de la recherche réside dans les communautés humaines, leurs pratiques et les usages finaux des idées ainsi produites. Or les solutions de traduction automatique grand public et clef en main, portées pour l'essentiel par les opérateurs américains du web, sont inadaptées, voire nocives : la simplification de la communication, l'anonymat, la captation des données, le recours à des corpus généralistes constitués de manière opaque, vont à l'encontre de ce qu'on peut attendre d'une science ouverte. L'initiative portée par le *Ministère de l'Enseignement Supérieur et de la Recherche* ainsi que par l'infrastructure *OPERAS* est ainsi particulièrement **légitime** et bienvenue sur plusieurs plans.

L'originalité de la démarche ici tient à la prise en compte de la **pluralité des acteurs** de la science ouverte : chercheurs, traducteurs professionnels ou bénévoles, relecteurs, éditeurs. La mise en réseau des différents utilisateurs, facilitée par la traçabilité des contributions et la mention des noms des contributeurs est essentielle : outre une meilleure diffusion des propositions scientifiques, **la mise en réseau dynamise les échanges**. Le choix d'un multilinguisme ouvert, non-restreint à des couples de langues, est une autre dimension essentielle à l'heure où nombre de projets scientifiques se montent à l'échelle européenne, selon des périmètres plus divers et inclusifs qu'autrefois. Les technologies linguistiques actuelles de transcription et de traduction des communications quasiment en temps réel, permettent déjà la généralisation de communications dans la langue des auteurs, et pas forcément en anglais. Si l'on veut pleinement correspondre aux pratiques réelles des chercheurs, alors j'aimerais suggérer d'ouvrir l'accès à ces outils (plateforme et plug-ins) aux collègues ressortissants de l'Union européenne, dans la mesure où les enjeux de souveraineté numérique se posent à cette échelle, comme en témoignent la RGPD, la directive sur le copyright ou encore l'AI Act de juin 2023.

Toujours dans l'idée que la clef d'une meilleure efficacité de la recherche tient aux échanges, la fonctionnalité de comparaison des traductions dans les différentes langues est très utile. On pourrait lui ajouter l'affichage de drapeaux d'alerte signalant des problèmes récurrents de traduction (par exemple : faux amis, unités de mesure), drapeaux a priori générés par les équipes, mais aussi occasionnellement par un utilisateur, sur le modèle de Wikipédia. Enfin, il me semble que la clef de la réussite du projet, qui doit effectivement être rapidement mis en place et sur une base simple, tient à sa souplesse. Les pratiques de publication des chercheurs évoluent en effet rapidement : en plus du format classique d'article publié dans une revue, les scientifiques publient désormais aussi des billets sur des blogs comme Hypothèse.org, ou des podcasts sur Youtube. Les conférences sont captées et diffusées en ligne ; les données de la recherche ou les pré-prints sont mis à disposition du public. Il y a donc de plus en plus de supports et de canaux de diffusion et la question de l'articulation entre la voix et le texte se pose déjà. Pour cette raison, et afin de renforcer la pérennité du modèle, il serait judicieux qu'un **comité de pilotage** précisément constitué de représentants des utilisateurs puisse suivre le lancement de l'outil dans sa version test, puis régulièrement évaluer l'outil et faire remonter l'évolution des usages, au moins tous les deux ou trois ans. Il faudrait que ce comité ait lui-même une certaine stabilité, à envisager sur une dizaine d'années avec des renouvellements partiels réguliers.

BIBLIOGRAPHIE & RÉFÉRENCES

Bibliographie de ce rapport

Christophe Dony, Iryna Kuchma, Tomasz Neugebauer, Jean-François Nomine, Milica Ševkušić, and Kathleen Shearer, " Is there a case for accepting machine translated scholarly content in repositories?", COAR (Confederation of Open Access Repositories), 8 mai 2023, http://www.coar-repositories.org/news-updates/is-there-a-case-for-accepting-machine-translated-scholarly-content-in-repositories/

Article Wikipedia : "Base de données orientée documents", consulté le 10 mai 2023, https://fr.wikipedia.org/wiki/Base_de_donn%C3%A9es_orient%C3%A9e_documents

Bibliographie complète

Bouillery, Carine, Marie-Céline Georg, et Elaine Holt. « Édito ». *Traduire. Revue française de la traduction*, no 246 (15 juin 2022) : 3-4.

Cloiseau, Gilles. « La traduction automatique en 2021. Qui, quoi et comment ? Une enquête sociolinguistique ». *Traduire. Revue française de la traduction,* no 246 (15 juin 2022) : 84-97. https://doi.org/10.4000/traduire.2853

Conjard, Lucie, "Ethnographie de l'IA", intervention au séminaire "Code Source", 12 mai 2022, https://listes.services.cnrs.fr/wws/arc/athena/2022-04/msg00013.html

Davat, Ambre," la notion de biais en IA", 08/01 : Séminaire IA Grenoble ; chaire IA, https://www.ethics-ai.fr/ambre-davat-biais-un-concept-technosolutionniste/
Vidéo : https://www.ethics-ai.fr

Fiorini, Susanna. « Traduction automatique et édition scientifique ». *Traduire. Revue française de la traduction*, no 246 (15 juin 2022) : 36-45. https://doi.org/10.4000/traduire.2805

Foti, Markus. « eTranslation. Le système de traduction automatique de la Commission européenne en appui d'une Europe numérique ». Traduit par Noëlle Brunel. *Traduire. Revue française de la traduction,* no 246 (15 juin 2022) : 28-35. https://doi.org/10.4000/traduire.2793

François, Floriane. « Post-édition, mode d'emploi. Retour sur une expérience personnelle ». *Traduire. Revue française de la traduction*, no 246 (15 juin 2022) : 5-9. https://doi.org/10.4000/traduire.2754

Filière, Carole : « Sommaire : n° 9 - La traduction littéraire et SHS à la rencontre des nouvelles technologies de la traduction : enjeux, perspectives et défis (2021) », https://revues.univ-tlse2.fr/lamaindethot/index.php?id=899

Gonse, Angèle. « Traduction automatique et usages sociaux des langues. Quelles conséquences pour la diversité linguistique ? Jean-Claude Beacco et al. (dir.) ». *Traduire. Revue française de la traduction*, no 246 (15 juin 2022) : 24-27.

Grass, Thierry. « L'erreur n'est pas humaine ». *Traduire. Revue française de la traduction*, no 246 (15 juin 2022) : 10-23. https://doi.org/10.4000/traduire.2763

Hurot, Laura. « Vers une slow translation ? Ralentir pour mieux traduire». *Traduire. Revue française de la traduction*, no 246 (15 juin 2022) : 109-17. https://doi.org/10.4000/traduire.2869

Kosmopoulos, Christine, Natacha Aveline, Colette Cauvin-Reymond, Bernard Elissalde, Maria Gravari-Barbas, Margaux Hardy, Nathalie Lemarchand, et al. « CybergeoNet – Traductions scientifiques ». *Cybergeo : European Journal of Geography*, 23 février 2022. http://journals.openedition.org/cybergeo/38309#tocto2n4

Lacour Philippe ; Bénel Aurélien- TraduXio Project : Latest Upgrades and Feedback, jdmdh :6733 - *Journal of Data Mining & Digital Humanities*, 8 janvier 2021, Atelier Digit_Hum - https://doi.org/10.46298/jdmdh.6733

Mion, Enrico Antonio. « Un dialogue de sourds ». *Traduire. Revue française de la traduction*, no 246 (15 juin 2022) : 46-54. https://doi.org/10.4000/traduire.2814

Poibeau, Thierry (2019). Babel 2.0. *Où va la traduction automatique*. Odile Jacob, Paris. https://www.odilejacob.fr/catalogue/sciences/informatique/babel-20_9782738148490.php

Russo, Nicola Pascal. « Faut-il craindre l'interprétation automatique ? De la fiction à la réalité, le point sur les « traducteurs électroniques ». *Traduire. Revue française de la traduction*, no 246 (15 juin 2022) : 98-108. https://doi.org/10.4000/traduire.2863

Tagand, Hélène. « Traducteurs, quel est votre métier ? » *Traduire. Revue française de la traduction*, no 246 (15 juin 2022) : 55-64. https://doi.org/10.4000/traduire.2823

Vidrequin, Magali. « Acceptabilité de la traduction automatique. Le cas de la post-édition chez les traducteurs médicaux ». *Traduire. Revue française de la traduction*, no 246 (15 juin 2022) : 77-83. https://doi.org/10.4000/traduire.2844

Villani, Cédric, "Donner un sens à l'Intelligence artificielle", rapport du 28 mars 2018, https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf

Autres ressources : blogs, colloques, séminaires, émissions (ordre chronologique)

- Blog d'Adrian Acolier (The morning paper).
  Article sur les plongements de mots (word embeddings) :
  https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

- Blog de Mark Liberman. Article sur les limites de la traduction neuronale en ce qui concerne les expressions idiomatiques :
  http://languagelog.ldc.upenn.edu/nll/?p=40602

- "Les humanités numériques en langue", Atelier DigitHum, ENS Paris, 17 **octobre 2019**, https://digithum.huma-num.fr/atelier/2019/

- Xavier De La Porte, avec T. Poibeau (invité). Comment la traduction automatique s'est-elle mise à (mieux) marcher ?
  Podcast "Le Monde a changé", France Inter, 11 **septembre 2020**.
  https://www.radiofrance.fr/franceinter/podcasts/le-code-a-change/la-traduction-automatique-avec-thierry-poibeau-9802626

- Translation Spaces - Volume 9, Issue 1, **2020**
  https://www.jbe-platform.com/content/journals/2211372x/9/1

- "IA et cancer : le diagnostic infaillible ?", émission radio du 21 **avril 2021**, https://www.franceculture.fr/emissions/la-methode-scientifique/la-methode-scientifique-emission-du-mercredi-21-avril-2021

- Séminaire "Objectivité et big data en médecine", **avril-juin 2021**,
  http://poincare.univ-lorraine.fr/fr/seminaire-objectivite-et-big-data-en-medecine-objectivite-graphique-en-sciences

- Colloque "Qu'est ce qui échappe à l'IA ?",'Ecole Polytechnique, 20-21 **septembre 2021**, https://www.polytechnique.edu/actualites/retour-sur-le-colloque-quest-ce-qui-echappe-lintelligence-artificielle

- “Le numérique dans les sciences humaines : édition et visualisation”, congrès de l’ACFAS (association canadienne pour la recherche francophone), 9-10 **mai 2022**, https://www.acfas.ca/evenements/congres/programme/89/300/310/c?ancre=23314

- “La formation en traduction à l’ère du numérique”, congrès de l’ACFAS (association canadienne pour la recherche francophone), 12-13 **mai 2022**, https://www.acfas.ca/evenements/congres/programme/89/300/306/c

- Colloque : “L’IA au prisme des sciences humaines et sociales”, Paris, EHESS, 13-14 **oct 2022**, https://www.ehess.fr/fr/journ%C3%A9es-d%C3%A9tude/lia-prisme-sciences-humaines-et-sociales

- Colloque : “Objectivité et big data en médecine”, Univ. de Strasbourg, 14 et 15 **déc. 2022**, https://www.misha.fr/agenda/evenement/colloque-objectivite-big-data-et-medecine-objectivite-graphique-en-sciences

- “Après ChatGPT : où en est-on avec les modèles de langage ?”, demie journée d’étude (org. Thierry Poibeau), ENS Paris, 11 **janvier 2023**, 14h, https://www.risc.cnrs.fr/echos/420146

- Séminaire et carnet de recherche “CulturIA”, **2022-3** https://cultureia.hypotheses.org/703

- Rapport de l’enquête 2022 sur les pratiques professionnelles en traduction, Société Française des Traducteurs, **juillet 2022** : https://www.sft.fr/sites/default/files/2022-11/2022_SFT_resultats-enquete-statistiques-metiers-de-la-traduction.pdf

- Barbin, Franck ; Hernández Morin, Katell ; Phuez-Favris, Gaëlle (2022) «Rédaction et traduction des métadonnées. Guide à destination des auteurs et comités éditoriaux de revues» [guide issu du projet Optimice] : https://doi.org/10.34847/nkl.1b145pz7

FIN DU DOCUMENT