Évaluation de la recherche

RAPPORT D'ÉVALUATION DE L'UNITÉ

ERTIM - Équipe de Recherche Texte, Informatique, Multilinguisme

SOUS TUTELLE DES ÉTABLISSEMENTS ET ORGANISMES :

Institut national des langues et civilisations orientales - Inalco

CAMPAGNE D'ÉVALUATION 2023-2024
VAGUE D

Rapport publié le 19/01/2024

Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur

Au nom du comité d'experts[1] :

Alain Peyraube, Président du Comité

Pour le Hcéres[2] :

Stéphane Le Bouler, président par intérim

En vertu du décret n° 2021-1536 du 29 novembre 2021 :
1 Les rapports d'évaluation « sont signés par le président du comité ». (Article 11, alinéa 2) ;
2 Le président du Hcéres « contresigne les rapports d'évaluation établis par les comités d'experts et signés par leur président. » (Article 8, alinéa 5).

Pour faciliter la lecture du document, les noms employés dans ce rapport pour désigner des fonctions, des métiers ou des responsabilités (expert, chercheur, enseignant-chercheur, professeur, maître de conférences, ingénieur, technicien, directeur, doctorant, etc.) le sont au sens générique et ont une valeur neutre.

Ce rapport est le résultat de l'évaluation du comité d'experts dont la composition est précisée ci-dessous. Les appréciations qu'il contient sont l'expression de la délibération indépendante et collégiale de ce comité. Les données chiffrées de ce rapport sont les données certifiées exactes extraites des fichiers déposés par la tutelle au nom de l'unité.

MEMBRES DU COMITÉ D'EXPERTS

Président :	M. Alain Peyraube, Directeur de recherche émérite, CNRS, Paris
Expert(e)s :	Mme Angélique Amelot, CNRS, Paris (personnel d'appui à la recherche) Mme Claire Doquet, Université de Bordeaux, Bordeaux (représentante du CNU) M. Mathieu Constant, Université de Lorraine, Nancy.

REPRÉSENTANT DU HCÉRES

M. Jean-Luc Minel

REPRÉSENTANTS DES ÉTABLISSEMENTS ET ORGANISMES TUTELLES DE L'UNITÉ DE RECHERCHE

Mme Marianne Fauchereau, Inalco
Mme Rima Sleiman, Inalco
M. Assen Slim, Inalco

CARACTÉRISATION DE L'UNITÉ

- Nom : Équipe de Recherche Texte, Informatique, Multilinguisme
- Acronyme : ERTIM
- Label et numéro : UR 2520
- Composition de l'équipe de direction : M. Damien Nouvel (directeur), M. Mathieu Valette (co-directeur)

PANELS SCIENTIFIQUES DE L'UNITÉ

SHS Sciences humaines et sociales
SHS4 L'esprit humain et sa complexité
SHS5 Cultures et productions culturelles
ST5 Sciences pour l'ingénieur
ST6 STIC (Sciences et technologies de l'information et de la communication)

THÉMATIQUES DE L'UNITÉ

Les thématiques scientifiques de l'unité ont été articulées, au cours de la période écoulée (2017-2022), autour de trois axes, centrés pour l'essentiel sur le Traitement automatique des langues (TAL) :

- Outillage des langues peu dotées
- Humanités numériques
- TAL pour les applications

Considéré comme un domaine d'application de l'Intelligence Artificielle (IA) et lié à des capacités d'outillage et d'analyse pour les études linguistiques, le TAL relève à la fois de l'informatique, des mathématiques et de la linguistique, ce qui explique la diversité des panels scientifiques de l'unité déclinés ci-dessus.

Les tâches spécifiques de ces thématiques sont à la fois l'extraction d'information, l'acquisition et la numérisation de données textuelles et manuscrites électroniques, mais aussi orales avec des enregistrements audios, leur analyse linguistique au niveau de la syntaxe et de la sémantique, la traduction automatique.

Le Document d'autoévaluation (DAE) souligne que le TAL est un secteur fortement concurrentiel au niveau international, investi par des acteurs académiques et industriels dont les géants du numérique, ouvert pour le partage des résultats scientifiques et des innovations technologiques, mais pas toujours pour les données, pourtant indispensables à la recherche. C'est aussi un domaine qui a considérablement évolué ces dernières années, tout comme les recherches sur le numérique et le développement de nouveaux logiciels et algorithmes qui ont eu un impact très important sur les activités scientifiques.

ERTIM travaille en conséquence sur le TAL sans interaction avec des chercheurs en SHS, par exemple dans le développement d'algorithmes ou de modèles pour les langues, la fouille de données, l'annotation automatique, etc., mais aussi sur l'appropriation progressive du numérique par les chercheurs en sciences humaines et sociales (SHS). L'unité leur assure une bonne compréhension des instruments informatiques les plus nouveaux et de leur utilisation : concordanciers et logiciels de statistiques textuelles, outils de transcriptions écrites ou orales, plateformes d'annotation, interfaces de création ou de curation de bases de données, ainsi que des outils d'analyse linguistique (morphologiques et un peu syntaxe) pour la création et la manipulation de modèles de langues. Cette évolution et la résolution de questions techniques a permis à ERTIM de concentrer ses travaux scientifiques sur la mise en œuvre et l'exploitation d'applications de TAL pour les humanités numériques, en particulier pour l'étude des langues et des aires culturelles, qui sont particulièrement nombreuses et typologiquement différentes au sein de la tutelle, l'Institut national des langues et cultures orientales (Inalco).

Le DAE signale aussi que l'unité "ne cherche pas à se positionner sur les recherches théoriques en mathématiques pour le TAL, en particulier pour ce qui concerne l'expérimentation de modèles de langues par IA, comme les réseaux de neurones profonds ou Deep Learning, qui nécessiteraient des compétences pointues et des moyens de calcul conséquents". Elle est à même, cependant, de les exploiter et de les évaluer pour ce qui est de l'annotation de données, de la fouille de textes, de la didactique des langues et de la linguistique de corpus, en un mot des humanités numériques dans leur ensemble.

HISTORIQUE ET LOCALISATION GÉOGRAPHIQUE DE L'UNITÉ

L'UR ERTIM est une ancienne Équipe d'Accueil (EA) créée en 2005 par la fusion de deux centres de recherches propres à l'Inalco : le Centre de Recherche en Ingénierie Multilingue (Crim) et le Centre d'Études et de Recherche en Traitement Automatique des Langues (Certal).

Le Crim proposait deux formations pluridisciplinaires de troisième cycle à vocation professionnelle : un diplôme d'études supérieures spécialisées (DESS) intitulé « Traductique et gestion de l'information » et un DESS « Ingénierie Multilingue ». Le premier d'entre eux insistait davantage sur les aspects linguistiques dans des domaines tels que la traduction et la terminologie tout en proposant également un volet informatique adapté à l'évolution de ces activités et aux nouvelles technologies de l'information. Le second concernait les aspects plus techniques du TAL pour mieux préparer aux professions d'analyste et concepteur d'outils informatiques pour l'acquisition et le transfert des connaissances. Ces DESS ont été remplacés par le diplôme de Master depuis la réforme Licence-Master-Doctorat (LMD) et l'harmonisation européenne des grades universitaires.

Unité propre de l'Inalco, ERTIM dispose de locaux dans la Maison de la recherche de l'Inalco, dans le bâtiment historique situé au 2, rue de Lille, 75007 Paris, construit en 1716 par le marquis de Bacqueville, alors que la rue se nommait encore rue de Bourbon. Racheté et restauré ensuite en 1767 par Jean-Louis Bernage, le lieu prend pour nom l'hôtel de Bernage. L'unité dispose dans cet ensemble de locaux spacieux et confortables, après avoir été contrainte de déménager à deux reprises en 2019 et 2020 dans des sites temporaires près de la place de la République, lorsque la Maison de la recherche a été en travaux de rénovation.

ERTIM a évidemment aussi accès au Pôle des Langues et Civilisations situé au 65, rue des Grands Moulins, 75013 Paris, pour organiser ses diverses activités qui nécessitent des espaces plus importants, et pour avoir accès aux fonds documentaires de l'Inalco, qui sont gérés par la Bibliothèque universitaire des langues et des cultures (BULAC), qui partage ce site des Grands Moulins avec l'Inalco.

ENVIRONNEMENT DE RECHERCHE DE L'UNITÉ

En tant qu'unité propre de l'Inalco, ERTIM consacre naturellement une grande partie de ses activités à des thématiques liées avec celles de l'établissement. Les échanges avec les autres unités de recherche affiliées à l'Inalco sont nombreux et, comme pour l'établissement dans son ensemble, très variés en termes de thématiques, de disciplines et de langues et cultures.

L'unité occupe cependant une place singulière à l'Inalco. Elle n'est véritablement spécifique à aucune langue et à aucune culture. Elle interagit dans les domaines d'enseignement et de recherche dès qu'il s'agit de méthodes de TAL. Son positionnement en humanités numériques et en technologies d'IA lui font ainsi occuper une place transverse et stratégique concernant plusieurs sujets au cœur de l'établissement.

ERTIM est étroitement liée à la filière d'enseignement « Textes, Informatique, Multilinguisme » (TIM) qui assure une formation permettant l'obtention du diplôme de licence avec une spécialité en langue orientale et en TAL, et du diplôme de Master « Traitement Automatique des Langues ». Ce dernier est organisé selon une première année composée d'enseignements en traitement automatique des langues, suivie de plusieurs parcours en Master 2 (M2), adossés chacun à un ou plusieurs établissements universitaires : Inalco, Université Sorbonne Nouvelle, Université Paris Nanterre. La filière est en effet intégrée aux Filières TAL et Ingénierie Linguistique (PluriTAL) de la Sorbonne, de l'Inalco et de Paris Nanterre. L'objectif de cette formation commune est de donner à des étudiants issus des cursus de langues ou de sciences du langage des bases solides qui leur permettent de s'orienter vers les métiers de l'ingénierie linguistique, et de leur donner les possibilités de choisir entre diverses perspectives : document électronique, ingénierie multilingue, traductique. Cela dit, les parcours du master TAL de l'Inalco sont depuis 2019 "Ingénierie Multilingue (IM), Technologie de la Traduction et Traitement des Données Multilingues (TeTraDoM, anciennement Traductique), Recherche et Développement (R&D)".

ERTIM a aussi bénéficié d'équipements informatiques, notamment la plateforme Cumulus, de la part de l'Initiative d'excellence (Idex) de l'Université Sorbonne Paris Cité (USPC), devenue l'Alliance SPC en 2019 avec l'Inalco comme membre fondateur. L'unité fait également appel à des équipements mutualisés comme l'« infrastructure de recherche » IR* Huma-Num, ou le calculateur Jean Zay. Trois membres de l'unité ont enfin adhéré individuellement au Laboratoire d'excellence (Labex) « Fondements empiriques de la linguistique » (EFL).

EFFECTIFS DE L'UNITÉ : en personnes physiques au 31/12/2022

Catégories de personnel	Effectifs
Professeurs et assimilés	1
Maîtres de conférences et assimilés	3
Directeurs de recherche et assimilés	0
Chargés de recherche et assimilés	0
Personnels d'appui à la recherche	0

Sous-total personnels permanents en activité	**4**
Enseignants-chercheurs et chercheurs non permanents et assimilés	0
Personnels d'appui non permanents	1
Post-doctorants	0
Doctorants	9
Sous-total personnels non permanents en activité	**10**
Total personnels	**14**

RÉPARTITION DES PERMANENTS DE L'UNITÉ PAR EMPLOYEUR : en personnes physiques au 31/12/2022. Les employeurs non tutelles sont regroupés sous l'intitulé « autres ».

Nom de l'employeur	EC	C	PAR
Inalco	4	0	0
Total personnels	**4**	**0**	**0**

AVIS GLOBAL

ERTIM est une unité de taille très modeste. Cette caractéristique essentielle peut être un avantage, et la direction de l'UR a tenté de profiter de cette situation, avec un certain succès. L'unité s'inscrit en effet clairement dans le domaine du TAL, et elle a su trouver sa place dans un domaine à très forte concurrence et en constante et rapide évolution en se positionnant sur des objectifs théoriques spécifiques.

L'unité a identifié des thématiques précises et centré ses activités de recherche dans des domaines de recherche précis, qui rendent dorénavant sa présence incontournable à l'Inalco. Son mode de fonctionnement, simple et efficace, est aussi coordonné avec la filière d'enseignement TIM responsable de la Licence et du Master TAL dans laquelle un grand nombre des membres de l'équipe interviennent.

L'unité s'est ainsi mobilisée pour traiter des questions d'ordre méthodologique appliquées aux données textuelles, écrites ou orales, dans un contexte fortement multilingue. L'unité exerce une veille constante sur les outils de TAL pour la diversité des langues, ce qui lui permet d'anticiper les tendances, d'être proactive et compétitive, en particulier pour les langues qui sont peu dotées, à l'instar du vietnamien, du bambara, du taïwanais, du teochew, et surtout du quechua. Les thématiques choisies sont en phase avec les évolutions du numérique et les besoins actuels en traitement de données linguistiques. Cette stratégie est assurément judicieuse à la fois du fait des évolutions du TAL et du contexte local, le rattachement à l'Inalco favorisant les approches multilingues. Grâce à ce positionnement clair et original, l'unité bénéficie aujourd'hui d'une bonne reconnaissance.

La production scientifique de l'unité est cohérente avec sa taille, ses activités et les thématiques scientifiques sur lesquelles elle intervient (4 ACL, 32 communications dans des congrès, dont plusieurs internationalement reconnus, tels que le 5th IFIP International Workshop on Artificial Intelligence for Knowledge Management (AI4KM) à Melbourne, et l'ACL 2020 à Seattle). Le nombre et le type de publications correspond aux standards dans le domaine du TAL même si le volume de publications des membres permanents de l'unité est assez disparate. La stratégie scientifique d'ERTIM vise à ne pas dépendre des moyens technologiques (algorithmes) pour se concentrer sur une vision à moyen terme de son domaine d'activité principal : les méthodes exploitées, la production de données et d'outils pour des objectifs scientifiques et applicatifs. L'unité a récemment étendu progressivement sa production scientifique à des supports internationaux, tout en restant très présente dans les conférences nationales. Elle cherche ainsi à partager, autant que possible, ses logiciels, données et connaissances en ligne, tout en tenant compte des intérêts des communautés linguistiques concernées.

L'attractivité de l'unité, au niveau national, est attestée par le nombre important de ressources financières complémentaires de la dotation annuelle de base de l'Inalco. ERTIM bénéficie ainsi de plusieurs projets de l'Agence nationale de la recherche (ANR) DALiH et TALD, dont un porté par l'unité, d'un financement issu de l'Idex de USPC, d'une subvention de la Direction générale de l'armement (DGA), projet VITAL, de plusieurs dispositifs Cifre (Convention industrielle de formation par la recherche) pour financer ses doctorants, ainsi que d'autres partenariats. Cela permet à ERTIM de s'appuyer, pour une très large part, sur des financements

externes pour mener à bien ses activités. Les ressources matérielles dont l'unité dispose dans ses locaux de la Maison de la recherche de l'Inalco sont aussi très convenables, qu'il s'agisse du mobilier, des équipements informatiques, du serveur de calcul, des ordinateurs personnels. Au total, l'unité possède incontestablement des ressources adaptées pour l'instant à son profil d'activités.

L'unité, au cours des dernières années, a enfin renforcé, ses relations avec le monde économique et socioculturel. Une dynamique importante s'est développée à l'égard du grand public et du monde associatif à une échelle nationale, mais aussi internationale, à propos de la question de la préservation, de la numérisation et de la revitalisation des langues en danger à travers le monde.

Ces activités se sont appuyées sur des langues spécifiques, en prolongeant les opérations multilingues de l'unité. De nombreuses questions qui sont délaissées par les acteurs du TAL concentrés souvent sur les performances quantitatives des modèles et les algorithmes, ont été abordées en lien avec les données et avec des considérations sociétales et d'aires culturelles.

Le comité d'évaluation (ci-après comité) estime cependant que tout risque dans ce développement harmonieux sur le moyen terme et a fortiori sur le long terme n'est pas écarté. Le risque le plus apparent est lié à la très faible taille de l'unité, même si elle est résolument soutenue par l'établissement qui la prend en compte.

L'unité a vécu un renouvellement générationnel important de ses personnels statutaires depuis 2019, avec le départ de 5 membres, compensé par un recrutement de 3 jeunes chercheurs. Cela a eu pour conséquence de réduire le potentiel d'encadrement. Cette fragilité interne ne peut pas toujours être compensée par des appels à des directeurs de thèse parmi les partenaires extérieurs à l'unité. Le comité recommande en conséquence de tout faire pour essayer de pérenniser le poste d'ingénieur de recherche et de recruter au plus vite un ou une gestionnaire.

Si l'attractivité de l'unité est raisonnable au niveau national, il existe une marge de manœuvre importante pour la renforcer à l'international. Le comité encourage ERTIM à s'inscrire davantage dans des réseaux européens et internationaux et même à en susciter la création à travers des projets européens de type « Action Marie Skłodowska-Curie » (AMSC) ou dans d'autres programmes H2020, dont certains appels sont bien adaptés aux compétences de l'équipe.

ÉVALUATION DÉTAILLÉE DE L'UNITÉ

A - PRISE EN COMPTE DES RECOMMANDATIONS DU PRÉCÉDENT RAPPORT

Les recommandations du précédent rapport ont été incontestablement prises en compte, pour la grande majorité d'entre elles.

Il était surtout recommandé à ERTIM d'ouvrir davantage l'unité à l'international, aussi bien pour ce qui est de la production scientifique de ses membres que de la constitution de réseaux. Plusieurs articles ont été publiés dans des conférences internationales renommées en TAL dans le contrat en cours et un événement important « Colloque dédié aux langues minoritaires et aux solutions numériques » (ContribuLing) a été organisé en ligne à l'initiative de l'unité en 2021 avec Wikimédia France et la BULAC. Il a eu une portée internationale indiscutable, autant par la qualité de ses intervenants que par son audience, et il a été reconduit en 2022 dans les locaux de l'Inalco. Ces succès ont conduit à ce qu'une nouvelle édition a eu lieu en mai 2023.

Les résultats sont encourageants. Le fait est que la moitié des 6 000 à 6 500 langues parlées dans le monde sont considérées comme vulnérables ou en danger. ERTIM participe ainsi aux programmes internationaux qui visent à la sauvegarde de ces langues en danger. L'un des leviers majeurs pour la survie de ces langues minoritaires est leur présence effective dans les outils numériques utilisés au quotidien : claviers, reconnaissance vocale, moteurs de recherche, etc. Le développement de ces outils nécessite lui-même la numérisation de nombreuses données linguistiques fournies par des locuteurs : lexiques, dictionnaires, corpus oraux et écrits, ontologies, etc. De nombreux projets ont été initiés ces dernières années pour faciliter cette contribution, dont plusieurs ont été présentés dans les éditions de ContribuLing.

L'ouverture internationale s'est aussi manifestée par le recrutement d'un enseignant-chercheur qui entretient des liens étroits avec des chercheurs de haut niveau en Chine, à Taiwan, et surtout à l'Université polytechnique de Hong Kong. Des publications conjointes sont déjà parues et des projets de coopérations formalisées sont en cours.

Le précédent rapport recommandait aussi à l'unité d'être plus attentive à la durée des thèses des doctorants. La mise en place des comités de suivi au niveau national a permis de répondre en partie à cette demande en sollicitant des avis extérieurs permettant d'assurer un meilleur suivi des doctorants par la formulation d'avis sur l'avancement de leur travail de thèse. Par ailleurs, le nombre de doctorants a été significativement réduit, afin

de mieux les encadrer individuellement. Des critères ont aussi été mis en place informellement en interne, en particulier sur le financement des doctorants et les moyens mis à leur disposition, pour mieux maîtriser les risques de doctorats trop longs ou d'abandons. Sur les 19 doctorants mentionnés dans les données fournies, 6 ont soutenu leur thèse dans la période écoulée, 5 ont abandonné et 8 sont en cours de thèse, démarrée durant la période. Le taux d'encadrement est de 2,7 doctorants par membre statutaire de l'unité. De nombreux doctorats sont dirigés en collaboration avec des personnes externes, qui sont pour une grande partie des chercheurs associés à l'unité.

B - DOMAINES D'ÉVALUATION

DOMAINE 1 : PROFIL, RESSOURCES ET ORGANISATION DE L'UNITÉ

Appréciation sur les objectifs scientifiques de l'unité

L'unité a su trouver sa place dans le domaine du TAL à très forte concurrence et en constante et rapide évolution en se positionnant sur des objectifs théoriques, méthodologiques et applicatifs autour des réseaux sociaux, opinions et arguments, des langues peu dotées, et dans les humanités numériques. Ce positionnement clair et cohérent en TAL, en linguistique et en humanités numériques lui assure une reconnaissance évidente de la part des autres acteurs de ces domaines malgré sa taille modeste, mais celle-ci lui permet sans doute d'avoir une certaine souplesse et de s'adapter plus facilement aux changements stratégiques.

Appréciation sur les ressources de l'unité

L'appel aux ressources de l'environnement dans lequel elle s'inscrit, à savoir l'Inalco, mais surtout la recherche et l'obtention de financements complémentaires, par l'intermédiaire de l'Idex de USPC, des fonds du Labex EFL, permettent à l'unité d'assurer son bon fonctionnement d'un point de vue financier. En termes de ressources humaines toutefois, la très faible taille de l'unité, même si elle est ouvertement soutenue par sa tutelle qui prend en compte cette faiblesse, représente une source de fragilité interne, surtout en ce qui concerne l'encadrement des doctorants, qui n'est pas toujours assuré comme il devrait l'être.

Appréciation sur le fonctionnement de l'unité

Elle est bien identifiée dans son environnement au sein de l'Inalco et des contacts sont établis avec les autres chercheurs de l'institut sur des problématiques TAL appliquées aux langues et aux cultures concernées. La gestion du personnel est conforme à la législation en vigueur ainsi qu'aux directives de l'établissement. La parité hommes-femmes est respectée, et, durant la période de crise sanitaire liée à la pandémie, l'unité a manifesté une attention particulière envers les risques psycho-sociaux, autant pour les doctorants que pour les membres permanents.

1/ L'unité s'est assigné des objectifs scientifiques pertinents.

Points forts et possibilités liées au contexte

L'unité s'inscrit clairement dans le domaine du TAL. Elle a su trouver sa place dans un domaine à très forte concurrence et en constante et rapide évolution en se positionnant sur des objectifs théoriques, méthodologiques et applicatifs autour des réseaux sociaux, opinions et arguments, des langues peu dotées, et dans les humanités numériques.

Le contexte scientifique et technologique a beaucoup évolué ces dix dernières années dans le domaine de prédilection de l'unité ERTIM qui se concentre sur le TAL, avec l'objectif d'identifier des problématiques liées aux données textuelles et à leur exploitation à des fins scientifiques et culturelles. Les difficultés de mise en œuvre des procédures liées au TAL, notamment les approches en apprentissage profond et Deep Learning, qui

nécessitent des compétences pointues mais qui sont souvent indispensables, ont été amoindries grâce aux progrès technologiques qui ont permis à des chercheurs non spécialistes de s'emparer des outils pour leurs travaux. Les analyses linguistiques outillées se sont généralisées.

ERTIM s'est positionnée sur des questions d'ordre méthodologique appliquées aux données textuelles, écrites ou orales, dans un contexte fortement multilingue, comme en témoignent la veille exercée sur les outils de TAL disponibles pour la diversité des langues, en particulier celles qui sont peu dotées, à l'instar du vietnamien, du bambara, quechua, du taïwanais, teochew, etc., ainsi que les éléments figurant dans le portfolio, notamment la thèse et le projet sur la langue quechua.

Cette stratégie était assurément judicieuse à la fois du fait des évolutions du TAL et du contexte local, le rattachement à l'Inalco favorisant les approches multilingues. Grâce à ce positionnement clair et original, l'unité bénéficie aujourd'hui d'une bonne reconnaissance. Les thématiques choisies sont en phase avec les évolutions du numérique et les besoins actuels en traitement de données linguistiques.

Points faibles et risques liés au contexte

L'unité est de petite taille et a subi des changements récents importants dans sa constitution. Plus de la moitié de ses effectifs a quitté l'unité entre 2019 et 2021, dont 1 maître de conférences, 1 professeur de l'enseignement secondaire agrégé (PRAG), 1 IGR (Ingénieur de recherche) et 1 enseignant associé à plein temps (PAST). Ces départs n'ont été que partiellement compensés par l'arrivée récente de trois nouveaux membres, dont 1 maître de conférences, 1 IGR contractuel et ½ personnel d'appui à la recherche (PAR). Cette situation a eu incontestablement un impact négatif sur l'activité scientifique de l'unité pendant une courte période. L'unité a dû être restructurée et ses objectifs redéfinis, avec une nouvelle répartition des responsabilités de chacun.

2/ L'unité dispose de ressources adaptées à son profil d'activités et à son environnement de recherche et les mobilise.

Points forts et possibilités liées au contexte

Outre la dotation annuelle modeste de 12 000 €/an en moyenne assurée par l'Inalco, calculée en fonction du nombre de ses membres, ERTIM bénéficie de ressources financières complémentaires issues de contrats et de projets collaboratifs de diverses natures : deux programmes de l'ANR, un financement issu de l'Idex de USPC, une subvention de la Direction générale de l'armement (DGA), des dispositifs Cifre pour des contrats doctoraux, ainsi que d'autres partenariats (70 000€/an, en moyenne).

Cela permet à ERTIM de s'appuyer, pour une très large part, sur des financements externes pour mener à bien ses activités. Au vu de la taille de l'unité, ERTIM ne manque pas de moyens financiers.

Une certaine souplesse dans la répartition de cette enveloppe budgétaire a aussi été mise en place, ce qui permet à tous les membres, lorsque les conditions des contrats le permettent, de profiter de ces financements sur contrat, y compris les doctorants. 14 doctorants, sur les 19 doctorants inscrits au cours du contrat écoulé, ont ainsi bénéficié de ces financements.

Les ressources matérielles dont l'unité dispose dans ses locaux de la Maison de la recherche de l'Inalco sont aussi très convenables, qu'il s'agisse du mobilier, des équipements informatiques, du serveur de calcul, des ordinateurs personnels.

Au total, l'unité possède incontestablement des ressources adaptées pour l'instant à son profil d'activités.

Points faibles et risques liés au contexte

Le risque le plus apparent est encore une fois lié à la très faible taille de l'unité, même si elle est résolument soutenue par l'établissement qui prend en compte cette faiblesse.

L'unité a vécu un renouvellement générationnel important de personnels depuis 2019. Cela a eu pour conséquence de réduire le potentiel d'encadrement. Il n'y a plus qu'un seul membre titulaire d'une Habilitation à diriger les recherches (HDR) au lieu de deux auparavant, ce qui était déjà un chiffre faible, compte tenu du nombre de doctorants, de contrats de recherche et de partenariats. Cette fragilité interne ne peut pas toujours être compensée par des appels à des directeurs de thèse parmi les partenaires extérieurs à l'unité.
Le comité a ainsi relevé qu'il y a eu 5 abandons de thèses, dont 4 qui concernent des thèses ayant débuté avant la période d'évaluation, soit un peu plus de 25 % de l'ensemble des thèses au cours de la période d'évaluation. Un accompagnement des nouveaux arrivants pour une montée en responsabilité sera important pour maintenir une bonne activité scientifique et de formation à la recherche.

3/ Les pratiques de l'unité sont conformes aux règles et aux directives définies par ses tutelles en matière de gestion des ressources humaines, de sécurité, d'environnement, de protocoles éthiques et de protection des données ainsi que du patrimoine scientifique.

Points forts et possibilités liées au contexte

La gestion des personnels de l'unité est conforme à la législation et aux directives de l'établissement. La parité hommes-femmes est respectée pour ce qui est des membres permanents : 3 hommes et 3 femmes, en comptant l'ingénieure en contrat à durée déterminée (CDD) de 3 ans et la gestionnaire à mi-temps, Il n'en est pas de même pour les doctorants avec une majorité de 12 femmes pour 8 hommes au cours de la période concernée.

L'ingénieure de l'équipe est disponible pour aider les membres à organiser le stockage des données informatiques sur des supports adéquats, selon le cas individuellement, en interne, au sein de l'établissement, ou en faisant appel à des structures externes telles que Huma-Num. La plupart des membres de l'équipe ont une formation informatique suffisante pour maîtriser les risques.

L'unité organise régulièrement des conseils de laboratoire et des assemblées restreintes. Ces réunions sont renforcées par des rencontres trimestrielles où tous les membres sont conviés, y compris les membres associés, dont le statut reste à déterminer, voir plus loin. Durant la période de crise sanitaire liée à la pandémie Covid-19, l'unité a manifesté une attention particulière envers les risques psycho-sociaux, autant pour les doctorants que pour les membres permanents et le personnel de soutien à la recherche. Des réunions régulières en ligne et des échanges individuels ont été organisés.

Points faibles et risques liés au contexte

Le DAE n'a pas donné d'informations sur les conditions de travail des PAR. Le comité ignorait si l'IGR contractuelle et l'Adjointe technique (AJT) avaient leurs propres bureaux et la possibilité de télétravailler, ou si des membres de l'équipe pouvaient se retrouver en situation de travail isolé, compte tenu de la taille très réduite de l'unité.

La visite de l'unité a permis de clarifier certains points, notamment de constater que les risques de travail isolé n'existent pas.

DOMAINE 2 : ATTRACTIVITÉ

Appréciation sur l'attractivité de l'unité

L'attractivité de l'unité repose essentiellement sur des dynamiques mises en œuvre à partir d'éléments scientifiques et humains. Au regard de la taille d'ERTIM, sa visibilité est satisfaisante dans le domaine du TAL et dans son environnement de l'Inalco. L'unité s'est dotée des moyens nécessaires en répondant à différents appels à projets et en mettant en place des partenariats avec d'autres institutions et entreprises. Elle porte des sujets spécifiques sur les langues peu dotées au sein de l'Inalco qui lui ont permis d'acquérir une visibilité dans son domaine du TAL et en humanités numériques au sein de l'établissement comme à l'échelle nationale, voire, dans une moindre mesure, à l'international, en Asie orientale, par exemple, ouverture qui devra être renforcée.

1/ L'unité est attractive par son rayonnement scientifique et s'insère dans l'espace européen de la recherche.

2/ L'unité est attractive par la qualité de sa politique d'accompagnement des personnels.

3/ L'unité est attractive par la reconnaissance de ses succès à des appels à projets compétitifs.

4/ L'unité est attractive par la qualité de ses équipements et de ses compétences techniques.

Points forts et possibilités liées au contexte pour les quatre références ci-dessus

L'unité possède un bon rayonnement scientifique au regard de sa taille, notamment au niveau national, grâce à ses fortes compétences méthodologiques et applicatives en TAL, grâce aussi à ses travaux sur l'argumentation et sur le développement de ressources pour des langues peu dotées comme le quechua et pour les humanités numériques, de la numérisation à l'analyse des documents.

L'unité a acquis au fil des années cette visibilité nationale et, plus récemment, internationale. Ses chercheurs sont régulièrement invités dans des conférences nationales et internationales et sollicités par des comités éditoriaux, essentiellement en TAL, en linguistique et pour les langues spécifiques dont ils sont experts.
Leur implication dans des colloques internationaux sur le TAL est une constante. ERTIM en a aussi organisé trois dans ses locaux au cours de la période écoulée.
Deux évènements récurrents sont à signaler à ce propos :

- Une participation très active au Hackathon en TAL (HackaTAL), organisé tous les ans depuis 2016, moins centré sur des communications scientifiques, mais plus ouvert à un public varié d'étudiants ou d'industriels. Ces manifestations ont accueilli entre 20 et 40 participants répartis par groupes dans différents ateliers. Après les éditions de 2016 à Paris, de 2017 à Orléans, de 2018 à Rennes, de 2019 à Toulouse et quelques années d'interruption liées à la Covid-19, le HackaTAL a fait son retour en 2023 à Paris.

- Deux éditions de ContribuLing ont eu lieu, en ligne en 2021 et à l'Inalco en 2022. Cet événement, organisé en collaboration avec Wikimédia et la BULAC a permis de réunir des communautés linguistiques autour de la constitution de ressources linguistiques par méthodes contributives. Des sessions méthodologiques ont été proposées, ainsi que des ateliers de formation pour les participants.

À l'Inalco, les membres d'ERTIM sont souvent sollicités pour des questions de recherche et d'enseignement liées au numérique. La question des données pour les langues, qu'il s'agisse de corpus et de dictionnaires, est très régulièrement l'occasion d'interactions, parfois informelles, avec d'autres collègues dans le cadre de l'établissement. Les méthodes pédagogiques pour l'enseignement des langues peuvent également faire appel au numérique et être l'occasion d'échanges avec les membres de l'unité lorsque des processus TAL sont en jeu. Selon les situations, ces sollicitations sont plus ou moins formellement liées au pilotage de la politique scientifique de l'établissement et de ses instances en matière d'enseignement.

L'unité accorde une grande importance à l'accueil des personnels, quel que soit leur statut et la durée prévue de leur activité. Les personnes concernées sont reçues par le directeur de l'unité et l'ingénieure pour leur exposer brièvement l'organisation de ERTIM et l'utilisation des locaux, qui bénéficient de suffisamment d'espace pour qu'un poste de travail soit systématiquement proposé à tous les nouveaux arrivants. Ces derniers sont alors inscrits sur les listes de diffusion générale et restreinte de l'unité.

Pour les enseignants-chercheurs nouvellement recrutés, l'intégration est progressive. L'Inalco a mis en place une décharge d'enseignement la première année et la mise à disposition d'une dotation initiale. Du point de vue scientifique, des propositions sont faites au chercheur accueilli en termes d'implication dans des projets existants et un appui coordonné de la Direction de la recherche, de la valorisation et des études doctorales (DIRVED) est assuré lorsque le chercheur envisage de soumettre un projet.

L'unité organise aussi des séminaires trimestriels et reçoit régulièrement des demandes de séjours de la part de chercheurs étrangers. Interrompu par la Covid-19, cet accueil a été remis progressivement en place depuis l'année 2022.

Pour les personnels qui ne sont pas membres titulaires (invités, stagiaires), lorsqu'une collaboration scientifique durable est envisagée, le statut d'associé peut leur être proposé, afin de maintenir le dialogue et de laisser la porte ouverte aux opportunités de collaborations scientifiques. Le tableur des données montre la variété des affiliations des associés, réseau important pour l'unité, qui lui permet d'étendre son champ d'activités à d'autres domaines ainsi qu'aux partenaires académiques et industriels.

Les questions relevant de l'intégrité scientifique sont discutées en interne. Elles permettent d'évoquer les interrogations liées aux données collectées et aux objectifs des traitements TAL mis en œuvre. La science ouverte est vivement encouragée, en tenant compte des contraintes liées aux données, à propos desquelles des informations ont été apportées en interne, notamment en lien avec le GDR « Linguistique informatique,

formelle et de terrain » (LIFT) consacré aux questions liées aux données et à leur exploitation : propriété intellectuelle, vie privée, diffusion et partage.

L'unité a soumis à plusieurs reprises des dossiers en réponse à des appels à projets européens, dans le cadre du programme Horizons 2020 (H2020) ou des programmes nationaux (ANR, DGA), avec l'appui de la DIRVED. Pour la période concernée, 5 projets ont été retenus sur 8 : 3 programmes ANR, et 2 dans le cadre du dispositif mis en place par la DGA, dont le projet « Régime d'appui à l'innovation duale » (RAPID), porté principalement par l'unité.

Les contrats doctoraux financés sur ressources propres proviennent essentiellement de ces contrats, ainsi que des dispositifs Cifre qui sont régulièrement mis en place avec des entreprises, dont quatre pour la période concernée.

Dans le cadre de l'Idex USPC, l'unité avait bénéficié lors de la période précédente d'un financement de 3 projets, dont la plateforme MultiTAL et l'opération TAL-SHS. La dynamique a été prolongée au-delà de 2017, mais elle a cessé en 2019, l'Inalco n'étant plus membre de cet Idex depuis cette date.

On peut aussi mentionner que le positionnement d'ERTIM sur le traitement automatique des langues aréales dans le périmètre des langues de l'Inalco a accru sa visibilité internationale et a ouvert des portes pour la mise en place de conventions ou projets internationaux, par exemple le projet soutenu par l'ANR « Digitizing Armenian Linguistic Heritage » (DALiH), qui se déroulera de 2022 à 2025.
.
Même si ERTIM n'est pas directement financé par le projet du Ministère de l'enseignement supérieur et de la recherche intitulé « Accompagnement de l'hybridation des formations en langues et civilisations orientales » (AHFLO) porté par l'établissement, les personnels de l'unité y participent activement, en particulier pour ce qui est du chinois ou du quechua.

L'ensemble de ces financements est très largement majoritaire dans les budgets dont dispose l'unité, ce qui lui permet de se doter confortablement de moyens en ingénierie et en équipements. ERTIM est ainsi à même d'assurer les dépenses de fonctionnement pour ses membres qui participent à des colloques et symposiums, ou à des écoles d'été.

L'unité possède des équipements et des compétences technologiques tout à fait suffisants, notamment en informatique au regard des objectifs fixés par l'unité. Elle possède une certaine autonomie vis-à-vis de sa tutelle grâce à l'utilisation de ses ressources propres et d'équipements mutualisés. Ces moyens internes ou mutualisés sont les suivants : machine de calcul NVIDIA GeForce RTX 3090 pour l'apprentissage de modèles de langue par des méthodes de Deep Learning, plateforme Cumulus de l'USPC, outils Huma-Num, supercalculateur du CNRS Jean-Zay.

La présence d'une ingénieure dans l'équipe, dont le poste est financé par l'établissement en remplacement d'un poste permanent, permet de maintenir les infrastructures techniques. Au vu de la taille de l'équipe, l'objectif est de solliciter des moyens externes mutualisés dont il n'est pas nécessaire d'assurer la maintenance. Comme cela est indiqué par ailleurs dans le DAE, ERTIM ne cherche pas à se positionner sur des questions techniques et ne place pas son attractivité dans le domaine du TAL par les infrastructures matérielles ou les innovations logicielles, mais privilégie un positionnement sur les méthodologies. Cette stratégie est un choix lié aux intérêts de l'établissement (langues, humanités numériques) autant qu'à la taille l'équipe. Il est judicieux.

Points faibles et risques liés au contexte pour les quatre références ci-dessus

L'attractivité de l'unité pour les appels à projets est essentiellement nationale. ERTIM ne bénéficie pas encore de projets de recherche d'envergure à dimension européenne ou internationale, malgré des tentatives en réponse à des appels du programme H2020. Les compétences de certains membres d'ERTIM sont suffisantes pour que l'unité envisage de candidater auprès du Conseil européen de la recherche (ERC European Research Council).

L'unité est aussi encouragée, grâce aux liens étroits qu'elle entretient avec des chercheurs en Asie orientale, à élaborer des projets de recherche conjoints, financés par exemple par le Research Grant Committee (RGC) de l'University Grant Committee (UGC) de Hong Kong, ou par le Ministry of Science and Technology (MOST) de Taiwan. Les accords de coopération entre l'ANR et ces deux institutions devraient faciliter la mise au point de tels projets.

L'ingénieure en soutien a une durée de contrat de trois ans. L'attractivité liée aux équipements dépend aussi de la pérennité de ce poste.
L'unité manque de personnels HDR pour l'encadrement doctoral, ce qui pourrait être un frein pour l'accueil de nouveaux doctorants prometteurs. Cependant, il faut aussi reconnaître que l'Inalco mène une politique pro-

active en matière d'incitation des MCF à s'inscrire dans une démarche HDR. La Commission de la recherche donne systématiquement la priorité aux projets HDR dans l'attribution des congés CRCT (5 CRCT en 2021/4 CRCT en 2022/ 9 CERCT en 2023).

Un des derniers points faibles d'ERTIM est assurément le fait que l'Inalco ne fait plus partie de l'idex USPC depuis 2019. La capacité d'obtenir des budgets pour soutenir la recherche académique de l'unité va être réduite.

DOMAINE 3 : PRODUCTION SCIENTIFIQUE

Appréciation sur la production scientifique de l'unité

La production scientifique de l'unité est cohérente avec sa taille, ses activités et les thématiques scientifiques dans lesquelles elle intervient. La stratégie scientifique d'ERTIM vise à ne pas dépendre des moyens technologiques (algorithmes) pour se concentrer sur une vision à moyen terme de son domaine d'activité principal : les méthodes exploitées, la production de données et d'outils pour des objectifs scientifiques et applicatifs. L'unité a étendu progressivement sa production scientifique à des supports internationaux, tout en restant très présente dans les conférences nationales. Elle cherche ainsi à partager, autant que possible, ses logiciels, données et connaissances en ligne, tout en tenant compte des intérêts des communautés linguistiques concernées.

1/ La production scientifique de l'unité satisfait à des critères de qualité.

2/ La production scientifique de l'unité est proportionnée à son potentiel de recherche et correctement répartie entre ses personnels.

3/ La production scientifique de l'unité respecte les principes de l'intégrité scientifique, de l'éthique et de la science ouverte. Elle est conforme aux directives applicables dans ce domaine.

Points forts et possibilités liées au contexte pour les trois références ci-dessus

La production scientifique de l'unité repose, pour une grande partie sur un savoir-faire méthodologique, en particulier dans la collecte, la fouille et l'annotation de données et de corpus, en sémantique textuelle, en analyse du discours. Ces opérations sont réalisées par les chercheurs à l'aide d'outils logiciels et d'algorithmes de TAL. L'objectif est de poser un regard scientifique critique et expert sur les enjeux théoriques et applicatifs des processus TAL en jeu et de leurs utilisations des données.

L'unité a une politique de publication tout à fait standard pour la communauté TAL. Elle publie essentiellement dans des supports d'audience nationale ou internationale sous la forme de communications dans des colloques et congrès qui sont publiées ensuite, souvent sous forme d'articles, dans les Actes de colloques et congrès. On peut citer ici des articles parus dans des publications de haut niveau international comm ACL (*Association for computational linguistics*), IJNPL (*International joint conference on natural language processing*), LREC (*Language resources and evaluation conference*), EMNLP (*Empirical methods in natural language processing*).

La production scientifique de l'unité, quantitativement, se résume à 2 recueils de communications liées à des colloques ou ateliers, 5 posters, 30 communications qui associent très fréquemment des chercheurs statutaires et des doctorants de l'unité, 4 articles dans des revues à comité de lecture, dont 1 par un doctorant de l'unité, 13 vidéos. 6 thèses ont aussi été soutenues durant la période.

En résumé, l'unité est de plus en plus présente dans les compétitions internationales du TAL, ce qui lui permet de mieux asseoir ses compétences en matière d'algorithmes, d'exploitation de jeux de données et d'évaluation des résultats, tout en privilégiant les opérations qui portent sur des données textuelles des diverses langues de l'Inalco, notamment les langues « peu dotées ».

La diffusion des connaissances dans ERTIM repose pour une grande partie sur les événements organisés : présentations scientifiques lors des réunions de l'unité, des séminaires, des soutenances de doctorat ou de master en TAL. Plusieurs projets impliquent des sous-groupes de l'unité, qui portent les thématiques scientifiques

correspondantes. La forte cohésion de l'équipe dans ses activités d'enseignement, dans l'occupation de locaux confortables, favorise beaucoup cette voie de discussion et de partage de connaissances.

Des propositions sont faites régulièrement à tous les personnels, pour favoriser les implications sur les sujets dont ils sont experts, sous forme de participation à des projets ou de responsabilités ciblées. Les participations aux congrès et colloques sont encouragées et ne sont pas soumises à l'obligation formelle d'une publication. L'ingénieure de l'équipe est à la disposition des doctorants et autres personnels afin d'apporter une alternative aux directeurs et de les former au processus de préparation d'une publication scientifique, autant du point de vue théorique que pour l'utilisation des outils adéquats pour l'analyse des résultats et la rédaction des articles.

Pour garantir l'intégrité scientifique, le partage des productions scientifiques est encouragé et peu d'articles sont rédigés par un seul auteur. Des recommandations et des conseils sont formulés aux jeunes chercheurs concernant les supports de diffusion et le choix des co-auteurs des articles. Des sessions de présentation des travaux sont organisées, par sous-groupes ou en réunion d'unité, afin de favoriser les échanges et les interactions entre membres de l'équipe et de mieux vérifier la solidité des résultats obtenus.

Les questions relevant de l'intégrité scientifique, discutées en interne, interrogent les données collectées et les objectifs des traitements TAL mis en œuvre. La question des langues peu dotées est particulièrement sensible et pose de nombreuses questions éthiques concernant l'équilibre entre le travail scientifique et l'apport des travaux pour les communautés linguistiques concernées. Le questionnement sur l'appropriation des données et les liens avec les industriels du domaine, les actions visant à la préservation, la numérisation et la revitalisation des langues concernées sont des préoccupations constantes d'ERTIM.

La science ouverte est conseillée, en raison des contraintes liées aux données et à leur exploitation. La très large majorité des publications de l'unité sont réalisées sur des supports gratuits et ouverts, ce qui favorise le partage des connaissances. Lorsque les logiciels issus des travaux de recherche sont fiables et robustes, ils sont déposés sur des plateformes adéquates (gitLab, gitHub). Une volonté de mettre en ligne les données de recherche est manifeste, même si elle rencontre souvent des difficultés liées au respect de la vie privée des informateurs et à la propriété intellectuelle. Les initiatives contributives sont promues au sein de l'unité, comme l'illustre l'organisation des conférences ContribuLing en partenariat avec WikiMedia France.

Points faibles et risques liés au contexte pour les trois références ci-dessus

Le nombre et le type de publications correspondent aux standards dans le domaine du TAL, eu égard à l'effectif réduit de l'unité. Le volume de publications des membres permanents de l'unité est néanmoins disparate. Cela peut sans doute s'expliquer par les changements récents dans la composition de l'équipe avec plusieurs départs en cours de période et l'arrivée de nouveaux recrutés auxquels il faut laisser le temps pour s'intégrer. La petite taille de l'unité permet difficilement d'absorber ces changements sans impacter la production scientifique de l'unité.

Il semble aussi que toutes les publications de l'unité n'ont pas été téléchargées sur la plateforme « Hyper Article en Ligne » (HAL), notamment pour certains membres, alors même que l'Inalco s'est engagé fermement dans le domaine de la science ouverte en créant une vice-présidence adjointe à la recherche déléguée à la science ouverte et aux humanités numériques.

DOMAINE 4 : INSCRIPTION DES ACTIVITÉS DE RECHERCHE DANS LA SOCIÉTÉ

Appréciation sur l'inscription des activités de recherche de l'unité dans la société

ERTIM a de nombreux liens avec le monde économique et socioculturel. Plusieurs travaux sont réalisés en collaboration avec des entreprises, ce qui apporte des cas d'application pratiques aux activités de l'équipe et des financements associés. Cet aspect s'est élargi ces dernières années aux activités liées à des communautés linguistiques, permettant également à l'unité de se positionner sur la question de la préservation, de la numérisation et de la revitalisation des langues en danger à travers le monde.

1/ L'unité se distingue par la qualité et la quantité de ses interactions avec le monde non-académique.

2/ L'unité développe des produits à destination du monde culturel, économique et social.

3/ L'unité partage ses connaissances avec le grand public et intervient dans des débats de société.

Points forts et possibilités liées au contexte pour les trois références ci-dessus

D'un point de vue socio-économique, l'unité interagit abondamment avec des entreprises du domaine sur des problématiques spécifiques. Elle joue ainsi simultanément un rôle d'expert et de garant scientifique indépendant de projets menés en collaboration ou en partenariat. Cela est illustré par les quatre dispositifs Cifre et les partenariats en Recherche et développement (RD) avec des entreprises, notamment avec le Centre national d'études spatiales (Cnes) ou avec Électricité de France (EDF). Ces projets interactifs avec des acteurs extérieurs reposent majoritairement sur des applications spécifiques du TAL : annotation ou détection d'éléments spécifiques aux domaines considérés, systèmes de dialogue multimodaux, caractérisation et détection d'influence dans les réseaux, etc. Ils apportent des financements complémentaires, mais aussi de nouvelles thématiques qui s'inscrivent dans les travaux scientifiques de l'unité.

Le nombre de thèses de doctorat financées par des ressources propres, notamment par le dispositif Cifre (voir ci-dessus), témoigne d'une interaction forte avec le monde non académique

Plusieurs autres projets impliquent des interactions plus ou moins formelles avec le monde associatif, notamment Wikimédia, qui soutiennent les liens avec les communautés linguistiques concernées et rendent l'unité active sur les défis culturels et sociétaux de la préservation des langues.

L'unité est aussi très active dans le domaine de la science participative. ERTIM intervient dans des communautés linguistiques et porte les initiatives de mise en place de projets collaboratifs, comme l'illustre l'organisation des Journées ContribuLing.

Dans son partenariat avec EDF, l'unité a été impliquée dans l'exploitation d'un vaste corpus de contributions citoyennes lors des grands débats nationaux en 2019, avec une emphase consacrée au nucléaire.

D'autres projets ont conduit à la conception de logiciels, qui restent le plus souvent à l'état de prototypes, mais qui peuvent être partagés sur des réseaux adéquats (sites personnels, gitHub, gitLab). Le projet MultiTAL, en sus des informations mises à disposition des communautés cherchant des logiciels TAL sur des langues très peu dotées, a de plus permis à l'unité d'accompagner la création d'une entreprise dans le domaine des nouvelles technologies. Située à Marseille, PostLab développe une plateforme digitale pour ouvrir l'accès à l'écosystème de l'IA académique. Après cinq ans d'existence, l'entreprise est aujourd'hui toujours active.

Les liens tissés avec les communautés linguistiques permettent de partager les résultats obtenus. Ceux-ci peuvent être d'ordre scientifique, mais apportent aussi des jeux de données, corpus oraux ou écrits, à destination du monde socio-économique. L'unité attache une grande importance à ce que ces résultats soient, autant que possible, largement diffusés et partagés dans les communautés concernées.

Les deux événements importants organisés par l'unité, ContribuLing et le HackaTAL, ont été tous les deux ouverts au grand public, au-delà des milieux du TAL. ContribuLing a organisé des ateliers et des conférences centrées sur les méthodes de collecte de ressources linguistiques, multilingues ou monolingues. Ils ont attiré des communautés linguistiques ciblées et ont permis de tisser des liens dans plusieurs réseaux ouverts au grand public, notamment la fondation Wikimédia. Le HackaTAL, organisé tous les ans entre 2016 et 2019, et à nouveau en 2023, est un atelier générique sur les défis à relever dans le domaine du TAL, également destiné à tout public.

Points faibles et risques liés au contexte pour les trois références ci-dessus

Au regard de la taille modeste de l'unité, les points faibles sur ses activités concernant le monde socio-économique non académique sont quasiment inexistants. On peut toutefois mentionner qu'il semble que deux thèses financées par des dispositifs Cifre n'ont pas été menées à terme. Les raisons invoquées par la direction lors de la visite de l'unité, à savoir les difficultés dues à la pandémie, ne suffisent pas à expliquer ce point faible.

ANALYSE DE LA TRAJECTOIRE DE L'UNITÉ

Rappelons d'abord les axes de recherche retenus pour la période 2017-2022 : 1/ Sémantique de corpus pour les applications TAL ; 2/ Acquisition de connaissances ; 3/ Technologies éducatives et apprentissage des langues ; 4/ Corpus et multilinguisme (axe transversal). Au cours de la période écoulée, chacun de ces axes a bénéficié d'une activité scientifique conséquente et a été l'occasion de travaux scientifiques que l'on retrouve dans les divers éléments mentionnés dans le DAE (projets, publications), qui signale aussi que le troisième axe portant sur les technologies éducatives et l'apprentissage des langues a été l'objet de moins d'activités en fin de période.

Les thématiques scientifiques d'ERTIM ont peu varié sur la période et ont été l'objet de travaux théoriques et appliqués dont l'ambition peut être mise en correspondance avec la taille de l'équipe, en particulier pour ce qui concerne le TAL, pour lequel l'unité s'est efforcée d'avoir un positionnement raisonnable et stratégique pour sa taille et ses moyens. Ce choix est directement lié à l'évolution rapide des algorithmes et modèles d'IA et des moyens qui y sont consacrés par les acteurs industriels du domaine.

Comme mentionné *supra*, la période a aussi été marquée par un nombre important de départs et d'arrivées ainsi qu'à un changement de direction, ce qui a nécessité une réorganisation et une remise en place d'une dynamique dans un contexte marqué par la pandémie et le travail à distance. L'unité a bien surmonté cette phase transitoire et a retrouvé une bonne stabilité en personnels et en moyens au cours de l'année 2021.

L'unité a aussi renforcé ses relations avec le monde économique et une dynamique importante s'est développée à l'égard du grand public et du monde associatif à une échelle nationale et internationale. Ces activités se sont appuyées sur des langues spécifiques, en prolongeant les opérations multilingues de l'unité. De nombreuses questions qui sont souvent délaissées par les acteurs du TAL (qui se concentrent souvent sur les performances quantitatives des modèles et algorithmes) ont été abordées en lien avec les données et des considérations sociétales et d'aires culturelles.

Quatre nouveaux axes sont proposés pour la période 2022-2027 :

1. Humanités numériques

Les recherches en sciences humaines et sociales et dans les disciplines littéraires sont désormais fortement encouragées à utiliser des outils numériques. À l'Inalco, l'unité est particulièrement sollicitée dans plusieurs domaines, sciences sociales, sciences politiques, histoire, arts, pour appliquer des méthodes de TAL et de linguistique de corpus. L'objectif est la mise au point de méthodes ou d'algorithmes pour des finalités dans tous les domaines des humanités numériques, à des fins de recherche ou d'exploitation.

2. Diversité des langues

En cohérence avec la mission conservatoire des langues du monde, portée par l'Inalco, ERTIM s'intéresse à la question de la diversité des langues du point de vue de leur traitement automatique. L'accent est mis sur les langues dites « peu dotées » et les langues d'héritage. Deux angles d'approche sont envisagés : d'une part, étudier ce que ces langues posent comme questions spécifiques pour le TAL et pour la nécessaire adaptation des méthodes (corpus et autres ressources limités, effets de l'apprentissage par transfert, absence de standardisation...) ; d'autre part, montrer ce que le TAL peut apporter dans la description et la préservation de ces langues. Une langue sans outils de traitement automatique ou mal prise en charge par les systèmes informatiques est aujourd'hui d'autant plus menacée.

3. Méthodologie du TAL

Les travaux scientifiques menés à ERTIM font majoritairement appel à des méthodes de TAL qui nécessitent une expertise pour mieux concevoir, évaluer et exploiter les tâches réalisées. Les progrès technologiques en traitement automatique de données sont indéniables et ces dernières années, la mise au point et l'exploitation de méthodes relevant de l'IA s'appuyant sur des réseaux de neurones pour construire des modèles de langues, sont devenus des questions incontournables en TAL, qui font l'objet de nombreux travaux de recherche par la communauté scientifique. L'unité travaillera autant sur la mise au point que sur l'exploitation de ces technologies TAL, pour réaliser des études, mettre au point des algorithmes, ou pour constituer des ressources numériques, souvent déclinées par langues et par tâche.

4. Acquisition de l'information linguistique

Les travaux de cet axe vont s'inscrire dans le domaine de la linguistique informatique. L'unité veillera à identifier, modéliser, extraire des informations pertinentes à la compétence linguistique à partir de données de performance, qui peuvent provenir de corpus ou, plus récemment, des représentations créées par des modèles de langues à partir de corpus. Dans ce contexte, la problématique de la distinction entre généralisation (statistique, linguistique) et la mémorisation lexicale présente un intérêt particulier. ERTIM cherchera à mettre en œuvre des méthodologies qui permettent de distiller des schémas linguistiques à partir de données, et de vérifier leur validité, en portant un regard critique sur les approches purement expérimentales (les « probes » par exemple) et une attention toute particulière à formuler des hypothèses linguistiques qui facilitent l'interprétation des expériences.

RECOMMANDATIONS À L'UNITÉ

Recommandations concernant le domaine 1 : Profil, ressources et organisation de l'unité

L'unité ERTIM manque de personnels stables dans le domaine du soutien à la recherche. Actuellement, l'équipe n'a pas de gestionnaire, et l'ingénieure de recherche est contractuelle. Le comité recommande à la tutelle de poursuivre son action de pérennisation du poste d'ingénieur de recherche et de recruter au plus vite un ou une gestionnaire.

L'unité ne comportant qu'un professeur, il est également urgent qu'au moins un de ses membres obtienne l'habilitation à diriger des recherches pour éviter que des demandes d'encadrement de thèse qui seraient fructueuses pour l'équipe restent sans suite. Les huit doctorats actuellement encadrés, qui correspondent aux forces d'encadrement de l'équipe, sont un effectif assez faible. Même si les doctorants d'ERTIM renvoient des avis très positifs sur le suivi dont ils bénéficient et sur leurs conditions de travail, cet effectif ne devrait pas se réduire encore sous peine de menacer le dynamisme du groupe de doctorants.

Pour ce qui est des budgets de recherche, le comité encourage ERTIM à poursuivre sa politique d'augmentation de ses ressources par la quête de financements, qu'il s'agisse de partenariats avec le privé ou de réponse à des appels à projets.

Recommandations concernant le domaine 2 : Attractivité

L'attractivité est d'autant plus nécessaire à l'équipe ERTIM que son effectif en chercheurs titulaires est restreint. Élargir l'équipe est d'ailleurs une priorité de la direction actuelle. L'attractivité d'ERTIM est bien réelle puisque l'unité reçoit régulièrement des demandes d'accueil de la part de chercheurs étrangers. Le comité recommande à ERTIM d'accepter de manière plus large ces demandes, de manière à faire montre de sa capacité à fédérer et à renforcer son attractivité à une échelle plus locale.

Le comité encourage également ERTIM à s'inscrire davantage dans des réseaux européens et même à en susciter la création à travers des projets européens de type « Action Marie Skłodowska-Curie » ou dans d'autres programmes H2020, dont certains appels sont bien adaptés aux compétences de l'équipe.

Recommandations concernant le domaine 3 : Production scientifique

Le comité encourage ERTIM à maintenir son niveau de production scientifique, tout à fait dans les normes des domaines dans lesquels travaille l'unité. Les nombreuses collaborations actuellement engagées avec des chercheurs et des enseignants-chercheurs d'autres institutions et la reconnaissance de la qualité des travaux des membres d'ERTIM permettent une bonne diffusion de leurs travaux. Ils sont rarement recensés, toutefois, de manière significative dans les revues les plus reconnues en sciences du langage dans la communauté scientifique internationale.

Les colloques récurrents organisés par ERTIM, notamment la conférence Contribuling, sont également une source de collaborations fructueuses pour l'équipe, et le comité l'encourage à poursuivre ce type de collaboration pour s'installer de manière encore plus évidente comme équipe de référence en TAL et en sciences du texte.

Recommandations concernant le domaine 4 : Inscription des activités de recherche dans la société

Le comité encourage l'équipe ERTIM à maintenir son ancrage dans le monde non académique, repérable notamment par les contrats public-privé existants et un taux important de contrats doctoraux dans le cadre du dispositif Cifre : 1/3 des doctorants sont financés de la sorte.

DÉROULEMENT DES ENTRETIENS

DATE

Début : 11 septembre 2023 à 09h00

Fin : 11 septembre 2023 à 18h00

Entretiens réalisés : en présentiel

PROGRAMME DES ENTRETIENS

9H30-10h :	Accueil et présentation du comité de visite
Plénière 10h-10h30 :	Réunion à huis-clos avec les tutelles – Mme Rima Sleiman, Vice-présidente de la Recherche de l'Inalco
10h30-11h00 :	Session plénière – Présentation du laboratoire : MM. Mathieu Valette et Damien Nouvel
11h0-11h30 :	Portofolio et trajectoire – M. Damien Nouvel
11h30-11h45 :	pause
11h45-12h05 :	Présentation de l'axe « Humanités Numériques » – M. Mathieu Valette
12h05-12h25 :	Présentation de l'axe « Diversité des Langues » – M. Pierre Magistry et Mme Ilaine Wang
12h25-12h45 :	Présentation de l'axe « Méthodologie du TAL » – M. Damien Nouvel
12h45-13h05 :	Présentation de l'axe « Acquisition de l'information linguistique » – Mme Kata Gabor
13h05-14h00 :	Pause Déjeuner
14h-14h30 :	Réunion à huis clos avec les doctorants
14h30-15h :	Réunion à huis clos avec la direction actuelle et la future direction
15h00-15h30 :	Réunion à huis clos du comité en présence du conseiller scientifique

Fin de la visite

POINTS PARTICULIERS À MENTIONNER

Le DAE souligne à de nombreuses reprises qu'ERTIM est une unité de taille modeste, mais rappelle aussi que cette taille limitée peut être un avantage. Par son positionnement original et central au sein de son établissement (l'Inalco), l'unité a des thématiques bien identifiées, une activité centrée sur ses domaines de recherche précis, enfin un mode de fonctionnement simple et efficace, coordonné avec la filière d'enseignement TIM responsable de la licence et du master TAL dans laquelle un grand nombre des membres de l'équipe interviennent.

Le DAE revient aussi à la fin de son rapport sur le fait qu'ERTIM a connu un renouvellement important de ses personnels statutaires au cours de la période : 5 départs qui étaient prévisibles et 3 nouveaux recrutements qui ont dû être intégrés progressivement dans l'unité. Cela a nécessité un travail important de réorganisation qui s'est déroulé sans trop de difficultés. Depuis la fin du premier semestre de 2021, les effectifs sont stables et, même s'ils ne sont pas au niveau de ceux dont disposait l'unité en 2017, le fonctionnement d'ERIM est aujourd'hui satisfaisant, mais seulement à court terme.

OBSERVATIONS GÉNÉRALES DES TUTELLES

Maison de la Recherche- Inalco

2, rue de Lille

75007 Paris

Paris, le 14 /01/ 2023

A l'attention du Haut Conseil à l'Evaluation de la Recherche et de l'Enseignement Supérieur

PUR250024525 - ERTIM - Équipe de recherche : textes, informatique, multilinguisme.

Objet : Observations de portée générale

L'Inalco adresse ses vifs remerciements au comité HCERES pour son engagement significatif et pour la qualité et la précision de son travail d'évaluation des activités de l'unité de recherche ERTIM. Les observations et les recommandations du comité permettront de contextualiser les activités de l'équipe et fourniront des éléments essentiels pour renforcer sa vision future. Après avoir pris connaissance du rapport d'évaluation ainsi que des remarques constructives qu'il propose, l'établissement propose d'apporter les quelques précisions suivantes :

- p. 4 l'affirmation "ERTIM travaille en conséquence sur l'appropriation progressive du numérique par les chercheurs en sciences humaines et sociales (SHS)" n'est pas totalement complète. **Précision :** L'équipe aussi des travaux qui portent sur le TAL uniquement sans interaction avec des chercheurs en SHS, par exemple le développement d'algorithmes ou de modèles pour les langues, la fouille de données, l'annotation automatique, etc.

- p. 4 la liste "concordanciers et logiciels de statistiques textuelles, outils de transcriptions écrites ou orales, plateformes d'annotation, interfaces de création ou de curation de bases de données" n'est pas tout à fait complète.
Précision : il manque notamment des outils d'analyse linguistique (morphologiques et un peu syntaxe) ainsi que la création et la manipulation de modèles de langues

- p. 13 "L'unité manque de personnels HDR pour l'encadrement doctoral, ce qui pourrait être un frein pour l'accueil de nouveaux doctorants prometteurs."

Précision : L'Inalco mène une politique pro-active en matière d'incitation des MCF à s'inscrire dans une démarche HDR. La Commission de la recherche donne systématiquement la priorité aux projet HDR dans l'attribution des congés CRCT (5 CRCT en 2021/4 CRCT en 2022/ 9 CERCT en 2023).

- p. 14: "Il semble aussi que toutes les publications de l'unité n'ont pas été téléchargées sur la plateforme « Hyper Article en Ligne » (HAL), notamment pour certains membres."
Précision : Il est à noter que l'Inalco s'est engagé fermement dans le domaine de la Science ouverte en créant une vice-présidence adjointe à la recherche déléguée à la science ouverte et aux humanités numériques. Un appui aux enseignants-chercheurs est apporté notamment sous la forme de « Halathons » (séance d'aide et de formation à la mise en ligne des publications sur HAL).

- p. 15 : "Les raisons invoquées par la direction lors de la visite de l'unité, à savoir les difficultés dues à la pandémie, ne suffisent pas à expliquer ce point faible."
Précision : Cette affirmation pourrait être modérée en mentionnant un sous-encadrement qui a été évoqué avec l'équipe pendant la visite du comité, même si cela n'a peut-être effectivement pas été indiqué lors de la discussion spécifique aux CIFRE.

Rima Sleiman
Vice-présidence de la Recherche

Assen Slim
Vice-président adjoint de la Recherche

Les rapports d'évaluation du Hcéres
sont consultables en ligne : www.hceres.fr

Évaluation des universités et des écoles
Évaluation des unités de recherche
Évaluation des formations
Évaluation des organismes nationaux de recherche
Évaluation et accréditation internationales

Évaluation • Recherche • Enseignement supérieur
Hcéres

2 rue Albert Einstein
75013 Paris, France
T.33 (0)1 55 55 60 10