AAPG2021	ACDC		PRCE
Coordonné par	Sylvain Lamprier	48 months	556 355€
CE 23 - Intelligence Artificielle (Axe 5.2)			

ACDC: Apprentissage Contrefactuel pour Data-to-text Contrôlé

Consortium: SU-MLIA, MNHN, ReciTAL, LAMSADE,

Tableau récapitulatif des personnes impliquées dans le projet

Partenaire	Nom	Prénom	Position actuelle	Rôle & implications	Implication (pers.mois)
SU-MLIA	LAMPRIER	Sylvain	MCF HDR	Coordinateur scientifique Responsable WP 0 et 1 Toutes les tâches	20 p.m
SU-MLIA	SOULIER	Laure	MCF	T2.1,T2.2,T2.3, T4.3	11.5 p.m
SU-MLIA	PIWOWARSKI	Benjamin	CR HDR	T1.1,T1.2,T1.3, T2.1,T4.2,T4.3	11.5 p.m
SU-MLIA	GUIGUE	Vincent	MCF HDR	T1.2,T2.1,T2.2,T2.3, T3.1,T3.2,T3.3	7.5 p.m
LAMSADE	ALLAUZEN	Alexandre	PR	Responsable WP 2 T1.2,T2.1,T2.2,T2.3, T3.1,T3.2,T4.2,T4.3	12 p.m
LAMSADE	CHEVALEYRE	Yann	PR	T1.3,T2.1,T2.2,T2.3 T3.1,T3.2,T4.2,T4.3	8 p.m
MNHN	VIGNES LEBBE	Régine	PR	Responsable WP 3 T2.1,T2.2, T3.1, T3.2, T3.3	13.5 p.m
MNHN	BOURGOIN	Thierry	PR	T1.1,T1.2, T3.1, T3.2, T3.3	9 p.m
RECITAL	STAIANO	Jacopo	Chercheur ReciTAL	Responsable WP 4 T1.1,T2.2,T2.3,T4.1, T4.2,T4.3	6 p.m
RECITAL	DURAND	Marie	Chercheuse ReciTAL	T2.3,T4.1,T4.2,4.3	15 p.m

Evolutions de la proposition détaillée par rapport à la pré-proposition

Par rapport à la phase 1, pas de modification importante à signaler. Le budget a été revu légèrement à la hausse (de 520kE à 556kE) pour inclure des coûts de permanents industriels qui avaient été sous-évalués. En terme de consortium, quelques adaptations ont été appliquées en fonction des disponibilités qui ont évolué : Pour RECITAL, G. Moyse et T. Scialom n'y figurent plus (mais restent en soutien important), M. Durand a intégré le consortium (notamment pour ses compétences en collecte de données et évaluation). Pour le LAMSADE, Y. Chevaleyre est venu compléter l'équipe. Du point de vue scientifique, nous avons gardé l'organisation du projet en 4 lots, les WP1 et WP2 ont été sensiblement réorganisés pour améliorer la fluidité des interactions entre ces lots et renforcer l'orientation des solutions vers la tâche finale de data-to-text.

I Contexte, positionnement et objectifs de la proposition

La très grande disponibilité des données est un fait bien établi dans notre société. Que les données proviennent de textes, de traces d'utilisateurs, de capteurs ou encore de bases de connaissances, l'un des défis communs est de comprendre et d'accéder rapidement aux informations contenues dans ces données. Une des réponses à ce défi consiste à générer des synthèses textuelles des données considérées, le langage naturel présentant de nombreux avantages en terme d'interprétabilité, de compositionnalité, d'accessibilité et de transférabilité. Néanmoins, si la génération de résumés pour données textuelles est un problème pour lequel les solutions commencent à être satisfaisantes, la génération de descriptions textuelles dans un cadre plus général (e.g., conditionnelles à des données numériques ou structurées) constitue toujours un problème particulièrement difficile. Ce problème fait référence à un champ émergent dans le domaine du traitement du langage naturel, appelé Data-to-Text, possédant de très nombreuses applications, notamment dans les domaines scientifiques, du journalisme, de la santé, du marketing, de la finance, etc. Une agence

cliente du partenaire du projet RECITAL, qui analyse des rapports d'entreprises pour simuler des stress tests écologiques sur des milliers de produits financiers, nous a rapporté que l'information utile de ces rapports se situait à 60% dans des tableaux, 10% dans des graphiques et seulement 30% dans le texte des rapports. Cet exemple illustre l'importance du problème, que les avancées récentes en apprentissage profond et génération de la langue rendent possible à envisager. Ce projet s'appuie sur ces avancées pour la génération de synthèses textuelles à partir de données tabulaires (bien que les propositions pourraient ultérieurement être étendues à d'autres types de données structurées telles des séries numériques, figures ou graphes), avec un accent particulier porté sur la recherche d'invariance des données d'entrée, l'extraction d'opérateurs de sélection/compression haut-niveau et la personnalisation des sorties produites.

a Objectifs et hypothèses de recherche

Le projet ACDC est centré sur l'extraction, la sélection de contenu et la synthèse textuelle à partir de données tabulaires. Deux grandes problématiques du Data-to-Text concernent l'exploitation de **données tabulaires hétérogènes** (i.e., les contenus d'entrée peuvent être très divers) et la définition d'**opérations de sélection/compression**, en fonction des besoins d'extraction visés. Pour répondre à ces deux grands défis, impliquant la prise en compte de la variabilité à la fois au niveau des entrées (i.e., les tableaux) et des sorties (i.e., les descriptions textuelles) avec supervision limitée, nous proposons de nous appuyer sur des techniques d'**apprentissage profond et par renforcement**, impliquant l'inférence, la manipulation et le décodage de représentations de vues sur les tableaux, que nous appelons opérations d'extraction (de contenu), et sont définies un espace sémantique continu. À l'instar des approches d'augmentation de données largement considérées dans le domaine de la vision par ordinateur pour augmenter les capacités de généralisation des modèles, une proposition de ce projet est de s'appuyer sur la production de **versions contre-factuelles des données** d'apprentissage (e.g., par permutations des lignes et colonnes, transformations selon des espaces sémantiques, etc.), pour définir un espace de représentation d'opérateurs algébriques adaptables en fonction des besoins. Une hypothèse centrale est que ce genre d'espace invariant permettra d'une part de **gagner en robustesse** pour l'identification des contenus dans les données d'entrée, et d'autre part d'améliorer les capacités de composition d'**opérateurs de sélection/compression haut-niveau**, difficiles à extraire dans un cadre d'apprentissage supervisé classique. Dans ce document, nous appellerons ces opérateurs des **opérateurs d'extraction**[1]. L'objectif est de produire des espaces de représentation réguliers, encodant divers types de symétrie sémantique des opérateurs appliqués aux contenus, permettant de **contrôler le mode de compression** des textes générés, en fonction d'un tableau d'entrée. La flexibilité induite permettra également la découverte de nouveaux opérateurs, adaptés aux besoins finaux, pour la génération de textes dans des contextes applicatifs variés. L'inférence d'opérateurs explicites envisagée dans ce projet permettra de mettre en place des **modèles interprétables**, facilitant ainsi l'analyse des synthèses produites, et la **planification de rapports textuels** cohérents, détaillant divers aspects saillants des données d'entrée. Enfin, un objectif du projet sera de permettre la **personnalisation des synthèses générées** en fonction des besoins des utilisateurs finaux.

Les défis que ce projet cible sont : 1) l'**inférence d'opérateurs** d'extraction d'information dans les tableaux, 2) la **gestion de l'hétérogénéité** dans les données d'entrée et 3) la **synthèse contrôlée** de descriptions textuelles. Le projet sera centré sur deux cas d'étude complémentaires, aux propriétés différentes, dans les domaines de l'analyse de données biologiques et l'analyse de documents d'entreprise. Il est articulé autour de 4 lots de travail. Le WP1 s'intéresse à l'apprentissage d'opérateurs et l'extraction de contenu. Le WP2 se focalise sur la planification et la personnalisation des synthèses produites. Le WP3 concerne la production de données supervisées et l'évaluation des synthèses produites par la communauté biologique. Enfin, le WP4 concerne des problématiques de transfert au cas d'étude financier, où les capacités de supervision sont limitées, mais les enjeux économiques considérables.

b Originalité et Positionnement par rapport à l'état de l'art

b1 Positionnement par rapport à l'état de l'art

La tâche émergente de transduction de texte (*"Data-to-Text"*) vise à résumer des données structurées dans une description en langage naturel. Jusqu'à récemment, les efforts visant à extraire la sémantique des

[1] Nous nommons opérateur d'extraction toute opération algébrique et/ou statistique que nous manipulerons sur les tableaux.

données structurées reposaient largement sur des connaissances d'experts et des règles définies manuellement [3, 30]. Avec l'essor des réseaux de neurones, cette tâche s'est naturellement tournée vers les approches du domaine de la traduction automatique neuronale, via l'encodage des tableaux considérés et leur décodage. Les premiers travaux de cette veine ont proposé de s'appuyer sur des représentations clé-valeurs des données [1, 17, 38], permettant alors de considérer les tableaux avec des modèles séquentiels (ex. LSTM). Ce type de représentation présente néanmoins diverses limites pour la prise en compte de structures complexes, regroupant plusieurs entités [38]. Cela a conduit des travaux plus récents à s'appuyer sur des architectures d'encodage contextuel, notamment via l'apprentissage de Transformers hiérarchiques [28]. Pour ce qui est du décodage textuel, des travaux récents sont basés sur de la planification pour garantir des mentions factuelles et cohérentes des enregistrements dans les descriptions générées [19, 25, 39]. Cependant, la totalité des approches du domaine supposent des tableaux très homogènes.

A ce jour, l'approche majeure des modèles de génération Data-to-Text repose sur un apprentissage encodeur-décodeur, fortement supervisé par une maximisation de vraisemblance de descriptions textuelles associées aux tableaux d'entrée. Si cette approche présente l'avantage de permettre la définition implicite du niveau et des modes de sélection/compression de l'information désirés, elle paraît cependant assez limitée selon plusieurs aspects décrits ci-dessous.

Complexité des descriptions attendues. Les jeux de données disponibles ne permettent d'explorer qu'un sous-ensemble limité des enjeux du data-to-text. Lorsque les jeux de données sont centrés sur une seule entité (e.g., E2E [23], WebNLG [7], Wikibio [17], Totto[24]), les descriptions relèvent généralement de la paraphrase des éléments d'un tableau simple, sans raisonnement complexe [38]. A notre connaissance, seul RotoWire [38] approche cet objectif en proposant des tableaux avec plusieurs entités et des descriptions longues sur de multiples entités. Cependant, la mise en relation/comparaison des entités est très limitée : elle relève simplement de la comparaison du score des équipes de basket et de l'identification du meilleur joueur. Le reste de la description correspond à une paraphrase du tableau. Dans ce projet, nous proposons de nous attaquer à des tâches de génération plus complexes pouvant raisonner sur plusieurs entités, avec une **large diversité sémantique** pour leur comparaison.

Extraction de contenu. L'aspect implicite des opérations sur le tableau d'entrée appris par l'ensemble des modèles de data-to-text est limitant dans le cas de raisonnement complexes. Très peu de travaux, pour la tâche de data-to-text, permettent des types d'extraction plus évolués que de simples sélections de cellules (e.g., avec inférence numérique ou sémantique sur les valeurs des tableaux). Ces travaux utilisent principalement une énumération exhaustive d'un ensemble d'opérations possibles, dont les résultats sont intégrés 1) au tableau avant l'opération d'apprentissage [22] ou 2) au cours de l'apprentissage pour déterminer l'opération la plus probable [9]. Cette stratégie induit une forte complexité et requiert des connaissances expertes pour la définition de ces types d'opérateurs haut-niveau dans le cas où l'on souhaite éviter la grande combinatoire de toutes les opérations possibles sur tous les éléments du tableau. Un des points critiques que nous abordons dans le projet réside alors dans la **découverte dynamique des opérateurs à utiliser**, sans exécution exhaustive préalable de toutes opérations possibles. Il est important de noter qu'à ce jour, il existe une large littérature sur l'apprentissage de représentations de mots permettant d'intégrer l'aspect numérique dans la sémantique [8, 20], notamment pour des tâches de questions-réponses [4, 9], ou encore des travaux permettant de réaliser/simplifier des formules mathématiques [14], ainsi que des travaux portant sur l'extraction de programmes (e.g. SQL) à partir d'instructions données en langage naturel [37]. Cependant, aucun travail à ce jour ne s'est intéressé à intégrer cet aspect tout au long du processus d'apprentissage des modèles data-to-text.

Vérification factuelle. Du fait de la supervision, la qualité des données (i.e., des descriptions attendues) impacte fortement la qualité des textes générés [29]. Lorsqu'un alignement parfait existe entre le tableau et la description (e.g., E2E, Totto), les modèles n'ont aucune difficulté à générer des descriptions adéquates. Cependant, la simplicité des descriptions attendues (i.e., une phrase sur une seule entité) n'est pas en adéquation avec le réalisme attendu des approches data-to-text dans de nombreux domaines d'application. Lorsqu'un alignement imparfait existe (e.g., WebNLG, RotoWire), la tâche de génération est alors plus compliquée car elle induit l'apprentissage de textes non factuels incluant alors des *hallucinations* vis-à-vis du tableau d'entrée. Divers travaux récents en génération de texte [10, 12] ou data-to-text [6] s'intéressent à cette problématique en s'appuyant sur des techniques intégrant des facteurs de contrôle en complément de

la donnée d'entrée, mais restent limités à des facteurs très généraux [26]. Dans ce projet, notre approche combinée de l'apprentissage d'une algèbre d'extraction (WP1) et de la génération par planification (WP2) a pour objectif de **renforcer les signaux détectés par la supervision** et donc d'améliorer la qualité des textes générées, tout en produisant des **raisonnements complexes explicables** sur de multiples entités.

Personnalisation. Enfin, l'apprentissage essentiellement supervisé des approches de la littérature du data-to-text ne permet pas d'offrir de **contrôle personnalisé du type de compression attendu** dans les synthèses générées. Il est essentiel de souligner que les principales motivations des approches data-to-text résident dans le besoin des utilisateurs d'accéder plus facilement aux données. Or, pour un même tableau, les utilisateurs finaux peuvent avoir des besoins différents. Il nous apparaît alors nécessaire d'aborder la problématique de personnalisation des textes générés. Pour la tâche de data-to-text, elle peut être abordée sous la forme d'une formulation d'un plan de résumé attendu plus ou moins explicite (exprimé en langage naturel ou sous la forme d'entités/d'évènements à énoncer) et plus ou moins complexe (focus sur une entité/un évènement ou séquence d'entités/évènements). En data-to-text, un seul modèle a abordé cet aspect de contrôle avec un facteur contrôlable de niveau d'hallucination acceptable dans la génération [6], mais aucun travail ne s'est concentré sur la personnalisation.

b2 Positionnement par rapport aux projets nationaux et internationaux

La problématique de la génération de langue a été abordée dans des projets au niveau national (ANR LAWBOT 2020, ANR MIDAS 2007) ou international (H2020 IMAGINE). Cependant, la problématique du data-to-text, telle qu'abordée dans ce projet avec la contrainte de tableaux, est très récente (2014/2015). Nous notons divers projets de data-journalisme ou d'archivage de documents, connexes à la problématique du *Document Intelligence* : 1) le projet iCoda à l'INRIA dont les objectifs concernent des tâches de reformulation de requête ou de détection d'entités (*entity linking*), 2) les projets Europeana Newspapers et H2020 NewsEye qui ne traitent en aucun cas les données tabulaires, 3) le projet ANR ARCHIVAL (19-CE38-0011) qui s'intéresse à la problématique de compréhension des documents (non tabulaires) sans objectif de génération.

• **Au niveau national**, nous pouvons recenser le projet ANR intitulé "Générer du texte à partir de données Web sémantiques - WEB-NLG" (2014-2017). Ce projet a posé les bases du data-to-text (méthodes simples sur la façon d'encoder le tableau et de générer du texte, ainsi que la mise à disposition du jeu de données WebNLG [7]). *Notre projet s'inscrit dans la lignée de ce projet mais aborde des défis spécifiques et plus complexes du data-to-text : 1) opérations complexes sur les tableaux, 2) génération adaptée à ces opérations complexes, 3) génération contrôlée et personnalisée à l'utilisateur, et 4) problématique d'adaptation aux domaines.* Récemment, Claire Gardent a obtenu la "Chaire XNLG", pour des objectifs de génération de langue et d'apprentissage profond pour le multilinguisme avec sources multiples. *Ces défis sont complémentaires aux défis abordés dans ce projet, sans recouvrement avec nos contributions.*

• **Au niveau international**, le projet NL4XAI ouvre le champ de l'explicabilité dans les modèles de génération de langage qui est l'une des essences du data-to-text. Nous pouvons également citer deux projets NSF très récents "CAREER : Data-Driven Document Generation #2037519" et "CAREER : Semantic Multi-Task Learning for Generalizable and Interpretable Language Generation #1846185" qui abordent des défis de recherche similaires. *Notre projet diffère dans le sens où nous nous concentrons sur l'apprentissage d'opérateurs complexes pour la compréhension du tableau et la génération (contrôlée) des descriptions.*

En ce qui concerne le domaine de la biologie, notre projet s'inscrit dans la lignée des projets de recherche nationaux (Projets Xper et DbTNT, du LIS-UMR 7205) et internationaux visant à publier des ressources taxonomiques (Catalogue of Life, World Flora online, GBIF, INPN, etc.), mais nous envisageons d'utiliser des méthodes automatiques généralisables, qui produiront des textes rédigés variés, au lieu de fiches standardisées. Ceci répond à l'ambition des projets de production automatisée de faunes et flores numériques en langage naturel, et au-delà, à la production de textes exprimant des particularités remarquables des données.

b3 Caractère Innovant du projet

L'ensemble des approches récentes de data-to-text travaillent donc de manière supervisée, sans représentation explicite des opérateurs d'extraction qu'ils manipulent pour passer du contenu tabulaire global à la synthèse textuelle. Ce projet se démarque car il propose de s'intéresser **à l'expression de ces opérateurs, afin**

de gagner en interprétabilité des modèles, ainsi qu'en capacité de contrôle sur les textes générés. En outre, si dans un cadre figé bien défini, avec de nombreuses ressources pour la supervision, il est possible de s'affranchir de l'expression explicite de ces opérateurs, car le mode de sélection peut être implicitement adapté en fonction des sorties désirées, ce n'est plus envisageable dans un cadre plus large avec une grande hétérogénéité des données d'entrée et des attendus dans un contexte où la supervision est limitée. Notre démarche, en forte rupture avec les approches de la littérature, est donc de chercher à inférer les opérateurs d'extraction de contenu permettant de passer d'un tableau à un texte observé, en ayant pour but d'avoir un apprentissage robuste, qui soit à la fois fortement généralisable et contrôlable par un utilisateur.

Pour répondre à ces besoins, nous proposons **la construction d'un espace latent *sémantique* des opérateurs sur les tableaux**. *Celui-ci correspond à l'aspect novateur majeur sur lequel s'organise le projet*. Une propriété désirée pour cet espace, que l'on cherchera à satisfaire au cours du projet, correspond au fait que de tout opérateur doit être adaptable à tout tableau sur lequel on souhaite l'appliquer. D'un autre côté, on souhaite qu'un maximum de sémantique soit préservée d'un tableau à l'autre pour un même opérateur, afin de disposer d'un espace dans lequel il est pratique d'apprendre des stratégies d'échantillonnage des opérateurs, dont les éléments ont un effet similaire quel que soit le contexte dans lequel ils sont appliqués. On ne souhaite alors pas que les opérateurs, qui correspondent initialement à des requêtes de type SQL par exemple, soient encodés en absolu dans l'espace de représentation Ξ visé, mais plutôt qu'elles soient exprimées selon des proximités dans un référentiel sémantique. Par exemple, on souhaite qu'une opération, correspondant à réaliser la moyenne du poids de tous les orang-outan d'un tableau de poids d'animaux, soit encodée comme devant retourner la moyenne du poids de l'espèce la plus proche de l'orang-outan du tableau sur lequel on l'applique, selon un espace sémantique à définir. De la même manière sur les noms des attributs plutôt que leur valeur, on souhaite qu'un opérateur sur des colonnes de poids puisse s'appliquer sur des colonnes sans nom explicite d'attribut mais dont les valeurs sont exprimées en kg (et inversement). Enfin, on souhaite exprimer une certaine proximité entre les types d'opérations d'extraction utilisés.

Cette proposition, qui correspond à un encodage contextualisable des opérateurs d'extraction, permet à notre espace latent d'être robuste à l'hétérogénéité des tableaux, et d'être efficace lors du transfert du modèle à d'autres données. Lors du décodage d'un opérateur, il est alors toujours possible d'extraire une opération valide, dont nous pourrons apprendre des distributions de probabilités en fonction des tableaux considérés et des textes descriptifs visés. En outre, un espace présentant de telles propriétés permet de considérablement **simplifier les mécanismes d'inférence d'opérateurs** à partir de couples tableau-texte, car une correspondance directe entre les opérateurs et le texte peut être trouvée (i.e., $p(\xi|\tau,\omega) \approx p(\xi|\omega)$, avec τ un tableau et ω la description textuelle correspondante). Afin de tendre vers ce genre d'espace contextualisable, nous définirons divers mécanismes d'encodage-décodage avec coûts adverses, basés sur des architectures neuronales avec réseaux d'attention du type Transformer Network, et des transformations contre-factuelles des tableaux considérés. Des politiques de combinaisons d'opérateurs, séquentielles ou en parallèles, dans cet espace seront également considérées, afin d'élargir le spectre des synthèses de tableau possibles, en évitant l'explosion de la complexité de l'espace de représentation. **De tels travaux, traitant à la fois de l'hétérogénéité (i.e., variabilité des entrées et des attendus) et de la recherche d'opérateurs d'extraction pour la génération textuelle, sont tout à fait novateurs dans le domaine du data-to-text, et même au delà, pour l'extraction de contenu synthétique de manière plus générale.** Nous nous appuyons sur cet espace d'opérateurs pour construire une plateforme complète liée au data-to-text, de la sélection de contenu à sa synthèse textuelle, personnalisable, organisée selon un plan du discours. Les enjeux sont importants pour les deux cas d'étude considérés, dont les communautés ont de forts besoins en analyse de données tabulaires.

c Méthodologie et gestion des risques

Le projet s'articule autour de 4 lots de travail, représentés avec leurs inter-dépendances en figure 1. Le WP1 vise à la mise en place des mécanismes d'extraction de contenu, via l'encodage et la composition d'opérateurs algébriques d'extraction dans les tableaux, pour le décodage de descriptions textuelles simples. Le WP2 s'intéresse à l'enchaînement de ces opérateurs, par l'apprentissage par renforcement de politiques de sélection pour la synthèse, la planification et la personnalisation de textes descriptifs longs. Le WP3 est focalisé sur le cas d'étude biologique, pour l'établissement de jeux de données tableaux-textes alignés, et l'évaluation des approches dans le contexte des centres d'intérêts de la communauté. Le WP4 correspond à

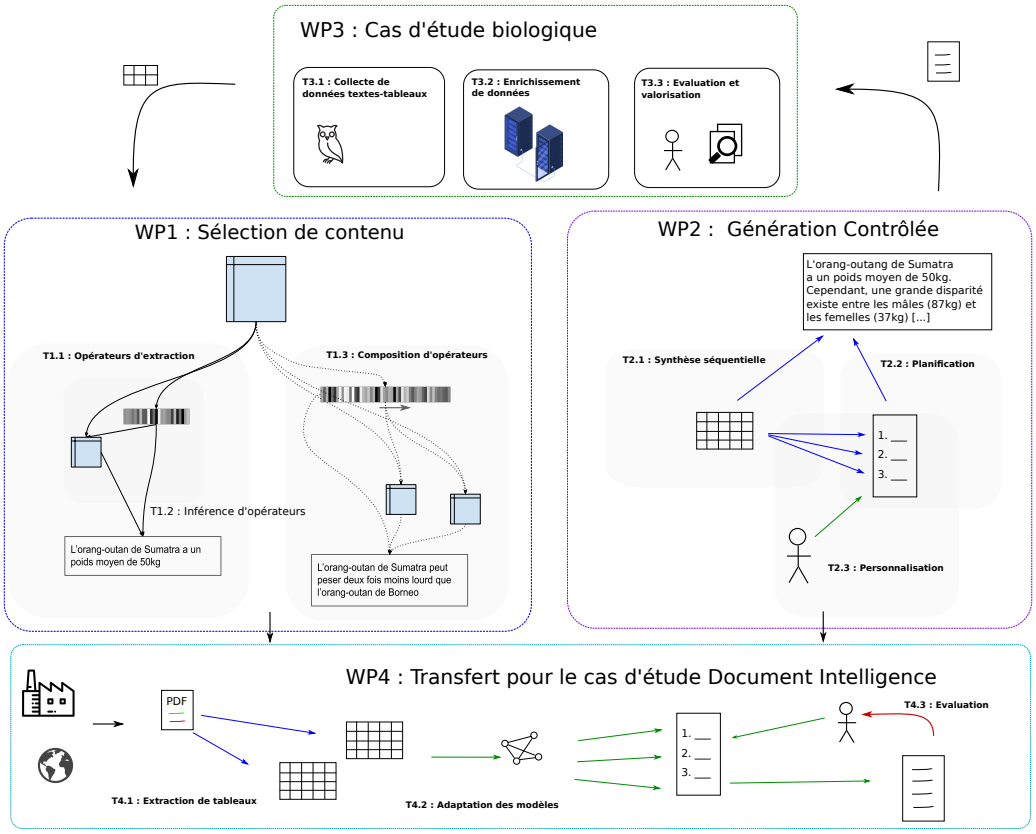

Figure 1. Organisation du Projet

une tâche de transfert des modèles au domaine du Document Intelligence, où la supervision plus difficile. Pour tous ces lots de travail, on considère des données tabulaires, où chaque élément en ligne est associé à des attributs - numériques ou textuels - dont les valeurs sont données en colonne. La figure 2 décrit l'organisation chronologique des différentes tâches sous la forme d'un diagramme de Gantt. Elle reporte également les dates de démarrage des deux thèses prévues pour le projet. La première, dirigée par le partenaire SU-MLIA porte sur l'extraction de contenu (WP1) et l'adaptation inter-domaine. La seconde, dirigée par le LAMSADE porte sur la synthèse contrôlée de textes issus des tableaux (WP2), avec des objectifs de planification et personnalisation. Les deux thèses seront encadrées en co-tutelle entre les partenaires SU-MLIA et LAMSADE pour favoriser les collaborations. Le partenaire RECITAL y prendra part activement pour le déploiement des méthodes développées. Le MNHN y sera également impliqué pour les questions de besoins en données des doctorants, ainsi que l'évaluation qualitative des méthodes.

Work Package 0: Coordination du Projet

Responsable: S. Lamprier (SU-MLIA)

Ce lot vise à assurer le bon déroulement du projet. Le porteur s'engage à assurer un suivi régulier du projet, pour le calendrier des tâches définies pour le projet, ainsi que les instructions de l'ANR. Il suivra les avancées et alertera les différents partenaires si nécessaire. L'objectif est également de permettre une bonne collaboration entre les différents partenaires en garantissant une bonne communication. Les principales actions prévues sont : (1) maintenir un site web dédié au projet, sur lequel l'ensemble des livrables, rapports, jeux de données, logiciels et publications du projet seront rendus disponibles; (2) organiser des réunions régulières avec les participants du projet, si possible en présentiel; (3) animer un groupe de lecture d'articles pour se tenir à jour des avancées des domaines et favoriser les échanges et collaborations entre partenaires.

Livrables: Documents ANR à T0+18, T0+48, Site Web à T0+3, Rapports d'activités à T0+12, T0+24, T0+36, T0+48. **Partenaires:** SU-MLIA (2p.m), MNHM (1p.m), LAMSADE (1p.m), RECITAL (1p.m).

Work Package 1: Extraction de contenu

Responsable: S. Lamprier (SU-MLIA)

Dans ce lot, nous considérons qu'une unité textuelle de description (e.g., une phrase) correspond à une vue sur un tableau global, que l'on peut représenter de manière équivalente par une structure tabulaire associée à une opération algébrique d'extraction. La tâche T1.1 s'intéresse à l'apprentissage de représentations sémantiques ξ dans un espace Ξ des opérations d'extraction dans les tableaux. L'objectif est d'encoder des opérateurs contextualisés, i.e. pour lesquels il est possible d'adapter les effets aux tableaux d'entrée considérés relativement à leur contenu/structure. L'espace de représentation appris servira de base aux tâches T1.2 et T1.3 pour 1) l'inférence de descriptions textuelles via l'apprentissage de la distribution de probabilités d'opérations élémentaires de notre espace appris, et 2) la composition d'opérateurs pour des inférences textuelles plus complexes.

Task 1.1: Apprentissage d'opérateurs d'extraction

Objectifs: À partir d'un générateur d'expressions algébriques dans un format prédéfini (un sous-ensemble de l'algèbre des requêtes SQL) permettant de générer des vues sur les tableaux, nous pouvons apprendre un espace d'encodage sémantique des opérateurs sur les tableaux *de manière auto-supervisée*. Une étude devra être menée sur les types des opérateurs élémentaires, composés d'opérations algébriques unaires (e.g., sélection, projection, etc.) et d'agrégation (e.g., maximum, moyenne, comptage, fréquence, etc.) Nous définirons ainsi une algèbre de transformation dans notre espace, que nous pourrons adapter pour modifier la complexité et l'expressivité de nos modèles.

Description: On considère des couples (τ, s) avec $\tau \in \Gamma$ un tableau d'entités et $s \in S$ une opération élémentaire de l'algèbre telle que s est adaptée à τ. On cherche une fonction d'encodage d'opérateurs (incluant leurs arguments) $e : \Gamma \times S \rightarrow \Xi$ et une fonction de décodage pour un tableau d'entrée $d : \Gamma \times \Xi \rightarrow \tilde{S}$, ainsi qu'une fonction d'interprétation $f : \Gamma \times \tilde{S} \rightarrow \Gamma$, telles que $f(\tau, \hat{s}) \approx s(\tau)$, avec $\hat{s} = d(\tau, e(\tau, s))$ l'opérateur adapté à τ dans un espace de décodage $\tilde{S}$, et $s(\tau)$ l'application de l'opération s à τ selon l'interpréteur à disposition pendant cette phase auto-supervisée. L'idée est de comparer $\hat{s}$ à l'opérateur original s, selon son effet sur τ via la fonction f, qui vise à mimer l'interpréteur selon des éléments de $\tilde{S}$. Cette approche, flexible car ne nécessitant pas de décoder des opérations syntaxiquement valides pour l'interpréteur, permet de définir des coûts dérivables dans l'espace des tableaux d'arrivée. Pour les différentes fonctions, on envisage d'utiliser des réseaux transformeurs neuronaux avec encodage de structure (cellule, colonne et ligne) s'inspirant de ceux employés sur des graphes (e.g., GTNs [40]). La comparaison des sorties tabulaires se fera par des métriques invariantes aux permutations. Dans la suite du projet, on notera $\hat{s}(\tau)$ l'application de l'opérateur décodé au tableau τ selon f.

Un objectif majeur est que l'encodeur utilise le contenu du tableau qui lui est passé en entrée pour encoder l'opérateur dans un espace Ξ, et respecte la sémantique de la requête relativement aux éléments du tableau, plutôt que d'encoder les arguments de manière absolue. Le décodeur vise à adapter tout opérateur issu de l'espace Ξ au tableau sur lequel on souhaite l'appliquer. Pour atteindre ces objectifs, le projet prévoit de considérer des coûts adverses qui tendent à modifier les paramètres de l'encodeur de manière à ce que le décodage selon un tableau contre-factuel de sémantique différente $\hat{\tau}$ produise un opérateur d'effet bien différent sur le tableau original (i.e., $\hat{s}(\tau)$ très différent de $s(\tau)$ pour $\hat{s} = d(\hat{\tau}, e(\tau, s))$). Cela force l'encodeur à se servir de τ. Pour obtenir une sémantique haut-niveau, il faut cependant que l'on évite d'encoder les arguments des opérateurs selon des indicateurs pauvres, du type indices des lignes et colonnes qu'ils manipulent. On propose alors d'encoder les opérateurs en appliquant des perturbations à τ, telles que des permutations de ses lignes et colonnes (voire des modifications de τ selon des symétries sémantiques).

Livrables: Générateur d'opérations élémentaires en SQL (via automates) à T0+3; Métrique de comparaison de tableaux à T0+6; Prototype de transformateur de tableaux selon opérateurs élémentaires à T0+12; **Gestion des risques:** Le principal risque de cette tâche correspond à l'apprentissage de la fonction f. Bien que ce risque reste limité au regard des succès récents d'approches neuronales pour la résolution d'opérations numériques complexes [14], on envisage 2 solutions de repli : 1) $\hat{s}$ est directement comparé à s, 2) $\hat{s}(\tau)$ est obtenu en utilisant l'interpréteur original (moins flexible mais apprentissage simplifié). Une autre difficulté pourrait être liée à la difficulté d'encodage des opérateurs, que l'on pourrait contourner en encodant uniquement leurs arguments. L'expertise de RECITAL pour l'interprétation de tableaux viendra en outre limiter les risques liées à la tâche. **Partenaires:** SU-MLIA (14pm), RECITAL (1pm), MNHN (0.5pm).

Task 1.2: Inférence d'opérateurs d'extraction

Objectifs : L'objectif est de faire le lien entre les tableaux d'entrée et des descriptions textuelles qui leur sont associées en supervision. Selon un jeu de données composé de paires tableau-texte, où les textes sont des descriptions textuelles simples d'un aspect unique des tableaux d'entrée associés (e.g., "le poids moyen d'un orang-outan adulte de Sumatra est de 45kg" à partir d'une table décrivant des individus d'espèces données et indiquant leur poids), il s'agit d'être à même d'inférer les opérations de sélection permettant l'extraction du contenu utile des tableaux, avant de chercher à convertir la vue résultante en texte.

Description : Nous supposons un ensemble d'apprentissage composé de paires (τ, ω), avec ω des descriptions textuelles des tableaux τ associés ; de plus les textes ω considérés dans cette tâche correspondent à des formulations textuelles des extractions de contenu des tableaux d'entrée via des opérations élémentaires telles que définies en tâche T1.1. Autrement dit, il existe un code $\xi \in \Xi$ tel que $\hat{s} = d(\tau, \xi)$ correspond à l'opération à appliquer au tableau τ pour obtenir l'information synthétisée dans ω. L'objectif de la tâche est d'apprendre la distribution de génération textuelle conditionnelle $p(\omega|\tau) = \int_\Xi p(\omega|\tau, \xi)p(\xi|\tau)d\xi$ par maximisation de vraisemblance selon les données d'apprentissage, avec $p(\xi|\tau)$ un prior sur la distribution de codes de transformation selon τ et $p(\omega|\tau, \xi) = p(\omega|\hat{s}(\tau), \hat{s})$ la distribution de probabilité sur l'espace des textes conditionnée au tableau τ et le code de transformation $\xi \in \Xi$. On se propose de traiter le problème de maximisation par inférence variationnelle bayésienne, en échantillonnant ξ selon une distribution apprise $q(\xi|\tau, \omega)$. Le prior $p(\xi|\tau)$ pourra être appris conjointement à $p(\omega|\tau, \xi)$ pour spécifier les opérations à privilégier selon le tableau d'entrée et la tâche considérée, et ainsi permettre la génération de synthèses utiles pour l'utilisateur final.

Livrables: Prototype d'inférence d'opérateurs à T0+12; Générateur de descriptions textuelles à T0+18.
Gestion des risques: Afin de limiter les difficultés issues du problème double d'inférence d'opérateurs et de décodage textuel, on propose d'amorcer le travail avec des données issues de canevas textuels définis par le partenaire MNHN au cours du WP3, visant à générer des textes de manière automatique à partir de leurs bases de données, dont on connaît les opérations de sélection utilisées. Cela permettra de faciliter grandement le travail de décodage textuel avec une base fixe, avant de se confronter à l'inférence d'opérateurs à partir de couples (tableau-texte). On peut aussi pre-apprendre $q(\xi|\tau, \omega) \approx q(\xi|\omega)$ selon des textes non-alignés à des tableaux. **Partenaires:** SU-MLIA (15pm), LAMSADE (2pm), MNHN (0.5pm).

Task 1.3: Composition d'opérateurs

Objectifs : L'objectif est d'élargir le spectre de descriptions que l'on est capable de considérer, par l'inférence d'opérateurs composés, impliquant plusieurs opérateurs d'extraction en séquence ou en parallèle. Par exemple, il s'agira d'être capable de générer des descriptions textuelles de plus haut niveau du type "les orangs-outans ont un poids adulte moyen très supérieur à celui des chimpanzés" (impliquant la comparaison de deux vues, l'une sur les orang-outan, l'autre sur les chimpanzés) ou encore "5 espèces de grands singes ont un poids supérieur à celui des orang-outan" (impliquant au moins deux sélections et une jointure).

Description: Alors que dans les tâches précédentes, nous nous étions limités à des opérateurs appliqués au tableau d'entrée uniquement, on considère ici un support Ξ qui inclut également des opérateurs de composition, unaires ou binaires, visant à appliquer des opérations aux vues intermédiaires précédemment définies dans un processus de construction du résultat final. Pour l'encodage de ces opérateurs, on raffinera l'apprentissage de la phase T1.1, en procédant de manière similaire, mais en travaillant avec une liste de vues en entrée de l'encodeur et du décodeur plutôt qu'un tableau unique. L'idée est d'encoder, via $e()$, la sémantique du ou des tableaux auxquels on souhaite appliquer l'opérateur. Cela permet de préparer le travail de la politique de choix des vues à manipuler à chaque instant de manière auto-supervisée, en définissant un espace Ξ adaptable en fonction des données à disposition. Ainsi, à chaque instant t de la composition, selon une séquence de vues $(\tau_0, ..., \tau_{t-1})$ passée en paramètre du décodeur, avec τ_0 le tableau original, le décodeur pourra déterminer quels éléments de la liste utiliser selon une distribution dépendant des informations qu'ils portent. Cette distribution choisit un tableau unique pour les opérateurs unaires, un couple pour les opérateurs binaires (du type union ou produit cartésien). Les encodeurs et décodeurs envisagés devront alors travailler avec des séquences de tableaux, dans une architecture transformeur récurrente par exemple. Bien sûr, il s'agira là aussi d'appliquer des permutations et des coûts adverses lors de l'apprentissage, pour que Ξ soit encodé en sémantique relative et de manière invariante à l'ordre des vues présentées.

Une fois ces opérateurs pré-appris sur une large variété de situations générées, il s'agira ensuite d'apprendre

à les combiner pour synthétiser les descriptions textuelles visées. Pour tout couple d'apprentissage (τ, ω) observé, l'objectif de cette tâche est d'inférer la séquence d'encodage d'opérateurs $o = (\xi_1, \xi_2, ..., \xi_T)$ pour laquelle la composition des fonctions correspondantes produit le résultat permettant de synthétiser ω à partir de la vue résultante, notée τ_T. Pour toute séquence o, on peut extraire une séquence de représentations d'opérations $(\hat{s}_1, \hat{s}_2, ..., \hat{s}_T)$ et une séquence de vues $(\tau_0, \tau_1, \tau_2, ..., \tau_T)$, par application séquentielle des opérateurs de o. L'objectif d'apprentissage sera alors de maximiser $\mathbb{E}_{o\sim q(o|\tau,\omega)}[\log p(\omega|\tau, o) + \log p(o|\tau) - \log q(o|\tau, \omega)]$, avec $q()$ la distribution d'inférence utilisée. On propose de traiter ce problème par apprentissage par renforcement, pour obtenir une politique de combinaison des opérateurs $p(o|\tau)$ adaptée à la tâche pour tout tableau τ de la distribution considérée, permettant finalement la synthèse de descriptions textuelles avancées selon le générateur appris $p(\omega|\tau, o)$.

Livrables : Un générateur d'opérateurs complexes à T0+18, un prototype de stratégie de combinaison simple à T0+24, plus évolué à T0+30. **Gestion des risques:** On se limitera dans un premier temps à des séquences d'opérateurs courtes, limitées à 2 ou 3 opérateurs par exemple. **Partenaires:** SU-MLIA (14pm), LAMSADE (2pm).

Work Package 2: Synthèse Textuelle Contrôlée

Responsable scientifique: A. Allauzen (LAMSADE)

Dans ce lot, nous exploitons les résultats du WP1 pour envisager la génération de textes longs, synthétisant les éléments des tableaux selon différents aspects d'intérêt. Nous envisageons cette génération comme une politique où chaque action correspond à la sélection d'un opérateur d'extraction. Le travail s'articule en trois sous-tâches de difficulté incrémentale. Alors qu'en tâche T2.1, nous nous focalisons sur la génération séquentielle de phrases descriptives, avec prise en compte de dépendances temporelles mais sans organisation préalable du discours dans sa globalité, la tâche T2.2 vise à s'appuyer sur un plan décrivant les différents aspects à aborder en fonction du tableau à décrire. Enfin, la tâche T2.3 vient apporter une plus grande capacité de contrôle aux textes produits, par la mise en place de mécanismes de personnalisation et d'interaction avec l'utilisateur final du système.

Task 2.1: Synthèse séquentielle

Objectifs : L'objectif est d'apprendre à établir des séquences d'opérateurs d'extraction, en les conditionnant à ce qui a déjà été dit du tableau d'entrée. Pour chaque opérateur sélectionné, on peut alors décoder la description textuelle correspondante comme en 1.2 pour former le texte final. Alors que la génération textuelle conditionnée s'est souvent heurté à des problèmes d'effondrement de postérieure dans la littérature, notre hypothèse est que le décodage de descriptions courtes issues d'opérateurs choisis itérativement permettra de contourner le problème en reportant la prise en compte de dépendances temporelles à la recherche d'opérateurs sémantiques prédéfinis, tout en conservant une grande flexibilité.

Description : Nous nous appuierons sur des données sous la forme de couples (τ, ω), où ω est un texte long pouvant se décomposer en unités de description textuelle $\omega_1, ..., \omega_T$ que nous traiterons séquentiellement via nos opérateurs d'extraction. Soit Υ l'ensemble d'opérateurs, simples (définis en tâche T1.1) ou complexes (définis en tâche T1.3). Le but est d'apprendre la distribution $p(\upsilon_n|\upsilon_{1:n-1}, \tau)$, avec $\upsilon_i \in \Upsilon$, en s'appuyant sur les textes et en supposant comme en WP1 que le texte généré ω_t à l'étape t ne dépend que de l'opérateur υ_t correspondant. Ceci nous permet d'estimer la vraisemblance d'un texte via : $p(\omega_{1:t}|\tau) = \int_{\upsilon_{1:t}} \prod_t p(\omega_t|\upsilon_t, \tau)p(\upsilon_t|\upsilon_{1:t-1}, \tau)$. Par rapport aux tâches du WP1, une différence majeure réside dans le conditionnement du choix de l'opérateur utilisé υ_t, selon les opérateurs précédents, qui portent l'information sémantique de ce qui a été dit dans les phrases passées. Afin de permettre ceci, nous emploierons une représentation récurrente hiérarchique basée sur les décodages des opérateurs pour le tableau considéré. De plus, pour éviter de se focaliser sur des vérités terrain textuelles non représentatives de la grande diversité des textes acceptables pour un tableau d'entrée, nous nous appuierons sur des techniques d'apprentissage par renforcement contre-factuelles qui permettent d'apprendre si un texte est vraisemblable ou non [32]. Enfin, des objectifs de gestion des hallucinations des modèles pourront être considérés dans cette tâche, en nous appuyant sur nos travaux pour le contrôle [26] et la vérification factuelle [31], pour la définition de critères à optimiser au cours de l'apprentissage de nos politiques de synthèse textuelle.

Livrables: Prototype de génération sur des textes enchaînant 2 ou 3 phrases à T0+24; Génération de

textes complets à T0+30. **Gestion des risques:** Une manière de gérer les risques liée à cette tâche sera de décoder dans un premier temps les phrases de manière indépendante, puis d'étendre la prise en compte des dépendances temporelles progressivement, dans une approche d'apprentissage par curriculum. Le pre-entraînement sur des données texte seul permettra également de faciliter le décodage textuel. Enfin, si le WP1 prenait du retard, le travail de cette tâche pourrait se baser dans un premier temps sur des opérateurs discrets pré-définis manuellement. **Partenaires:** LAMSADE (13pm), SU-MLIA (6pm), MNHN (0.5pm).

Task 2.2: Planification

Objectifs : Dans la tâche précédente, la prise en compte de l'historique pour la sélection des opérateurs à utiliser à chaque étape permet d'éviter les redites et d'avoir un enchaînement visant à mimer les textes observés avec intégration des dépendances temporelles. Néanmoins, cela ne permet pas une organisation globale du discours en fonction des éléments du tableau : on s'adapte au fur et à mesure du processus mais on n'anticipe pas l'ensemble des éléments utiles du tableau d'entrée pour structurer le texte produit. Cette tâche vise à répondre à cette limite, en cherchant la production d'un plan global selon le tableau d'entrée, permettant de conditionner les séquences d'opérateurs sur les différents aspects à énoncer.

Description : Dans cette tâche, on cherche donc à produire un plan, qui correspond à une séquence de variables latentes $z = (z^1, ..., z^k)$ décrivant les différents aspects à extraire du tableau d'entrée, avec k le nombre d'aspects à considérer. Différentes possibilités seront à envisager : une proposition serait de considérer que chacun des aspects s'applique sur une zone du texte cible prédéfinie. Ainsi, pour un texte cible de T phrases, on définit pour chaque phrase $i \in \{1, \dots, T\}$ un $\gamma_i \in \{1, \dots, k\}$ qui définit l'aspect qui s'applique à la phrase i – avec $\gamma_{i+1} \geq \gamma_i$ pour tout i. Ces zones peuvent correspondre à une répartition uniforme des aspects, ou bien être définies lors de la formation du plan. De manière alternative, à l'instar des approches de renforcement hiérarchique avec options [18], une possibilité à envisager serait que le changement d'aspect soit déterminé dynamiquement, en ajoutant un opérateur de fin d'aspect dans Υ.

Pour la production du plan, diverses pistes sont à explorer. Une possibilité serait de chercher $p(z|\tau)$ qui maximise la reconstruction des textes observés, possiblement via inférence bayésienne selon une distribution variationelle $q(z|\tau, \omega)$ basée sur une segmentation thématique du texte ω. Alors que dans [25], les auteurs supposent qu'un aspect du plan correspond à une entité particulière du tableau extraite selon un réseau d'attention, nous travaillerons plutôt en fonction d'un découpage de l'espace des opérateurs selon le tableau d'entrée, en s'inspirant des approches de recherche de compétences en apprentissage par renforcement [5].

Selon une répartition des aspects donnée $\gamma_1, ..., \gamma_T$, on pourra alors estimer $p(\omega_t|\omega_{1:t-1}, \gamma_t, z, \tau)$, avec γ_t l'indice d'aspect de z en cours pour la phrase t, par marginalisation (et échantillonage de Monte-Carlo) des choix d'opérateurs passés conditionnés à γ_t et z, permettant ainsi de structurer le discours chronologiquement. Lors de l'apprentissage, on pourra considérer un coût d'identification des aspects à partir des phrases décodées ou des opérateurs choisis, afin de forcer l'utilisation des variables z dans les politiques apprises.

Livrables: Prototype avec aspects pré-définis manuellement à T0+30; Prototype avec aspects dynamiques à T0+36. **Gestion des risques:** Une manière de limiter les risques sera de mettre de côté les dépendances temporelles dans un premier temps, et de produire les aspects du plan par des techniques de clustering sur les phrases produites indépendamment, dans une procédure de type EM avec échantillonnage. **Partenaires:** LAMSADE (13pm), SU-MLIA (7pm), RECITAL (1pm), MNHN (0.5pm).

Task 2.3: Personnalisation

Objectifs : L'idée est de conditionner le plan à une requête de l'utilisateur, exprimée en langage naturel (e.g., "je veux un résumé axé sur les orang-outans et leurs particularités vis à vis des autres grands singes du tableau d'entrée"). Une interaction peut ensuite être mise en place pour raffiner le texte généré en fonction d'aspects du tableau à renforcer ou atténuer. Cette tâche, plus exploratoire, est à la croisée des travaux en résumé interactif [34, 35] et de génération contrôlée pour le data-to-text [26]. L'objectif est de fournir une capacité de contrôle renforcée à l'utilisateur.

Description : Le travail de cette tâche se décompose en deux étapes. Dans un premier temps, nous chercherons à interpréter des requêtes utilisateur en langage naturel, pour conditionner la formation du plan défini en tâche T2.2. Étant donnés un tableau τ et une question q, le but est d'apprendre la distribution $p(z|q, \tau)$, permettant de contrôler la sélection des opérateurs v à utiliser. On pourra s'inspirer pour cette

étape des travaux récents pour la traduction de requêtes en langage naturel en requêtes SQL [13], en les adaptant à notre contexte de sélection d'opérateurs d'extraction encodés dans un espace sémantique.

Dans un second temps, on envisage de prendre en compte des retours de l'utilisateur par rapport au texte généré. Deux niveaux d'interaction sont envisagés, à savoir un simple pointage d'un aspect i du plan que l'utilisateur souhaite remplacer, menant à un nouvel échantillonage de l'aspect du plan correspondant selon $p(z_i|\tau, q)$, et un texte spécifiant l'aspect à privilégier pour remplacement, via une description q_i en langage naturel, intervenant alors sur une distribution $p(z_i|\tau, q, q_i)$ à définir. On pourra également envisager de raffiner les distributions de sélection des opérateurs en fonction de l'utilisateur, dans un contexte d'apprentissage continu.

Livrables: Générateur conditionné à T0+36; Générateur interactif simple à T0+40; Générateur interactif renforcé (avec re-spécification des intérêts) à T0+46. **Gestion des risques:** Le risque pour s'être tâche est d'être trop dépendant de la précédente. Néanmoins, une solution de rempli serait dans un premier temps de personnaliser sur des phrases uniques, en spécifiant les besoins en langage naturel pour conditionner les distributions de T1.2. Un autre risque est lié à l'interaction qui nécessite des utilisateurs finaux. Le travail pourra néanmoins être amorcé par la définition de bots heuristiques simulant des comportements humains plausibles. **Partenaires:** LAMSADE (13pm), SU-MLIA (7pm), RECITAL (2.5pm).

Work Package 3: Validation Expérimentale : Cas d'étude Biologique

Responsable: R. VIGNES LEBBE (MNHN)

Le premier cas d'étude considéré, en partenariat avec le MNHN, concerne le domaine de la biologie, où l'objectif est de générer des descriptions textuelles de **ressources taxonomiques** décrivant des groupes de plantes et/ou d'animaux par leurs caractéristiques. Les descriptions textuelles visées impliquent bien souvent des opérateurs d'extraction bien plus évolués et variables que les jeux de données publics de la communauté du Data-to-Text. Elles permettront de mettre en place la plateforme de génération visée, dans un cadre bien supervisé, et à fort potentiel pour la communauté biologique.

Task 3.1: Collecte et génération de données

Objectifs L'objectif est l'extraction de tableaux à partir de plateformes du domaine, et leur association à des descriptions textuelles simples et factuelles, en volumes conséquents.

Description Nous choisissons en priorité de cibler des données sur des insectes, mais des jeux de données sur d'autres groupes d'organismes vivants sont aussi disponibles. Ces jeux de données proviendront d'une part de la base de données FLOW[2] et d'autre part de plusieurs contenus matriciels gérés sur la plateforme Xper3[3]. Un travail de sélection des contenus Xper3 sera effectué, pour ne garder que les contenus suffisamment précis, exhaustifs et cohérents. Les tableaux exportés se présenteront sous la forme de vues à extraire de bases de données, avec en lignes des espèces et en colonnes des variables numériques ou qualitatives, reportant des informations de classification, de distributions géographiques, de bibliographie, etc.

En plus des tableaux de données, nous fournirons des textes correspondant à des résultats de requêtes sur ces données. Pour FLOW nous disposons d'ores et déjà de 360 textes d'envergure décrivant des groupes supérieurs au niveau genre dans la classification biologique. Ces textes ont été générés automatiquement à partir de canevas de textes à trous sur la base des données de FLOW. Pour le projet ACDC, nous compléterons les canevas précédents pour générer des textes plus simples mais plus nombreux aux niveaux inférieurs des genres et des espèces (2500 textes de genres et 14000 textes d'espèces). Pour les jeux de données Xper3 nous fournirons des descriptions en langage naturel provenant soit de la littérature à l'origine des jeux de données, soit générés automatiquement. Un travail en collaboration avec les partenaires sera nécessaire pour adapter les données fournies aux besoins des lots WP1 et WP2 et en fonction de la progression du projet. Par exemple, on pourra transposer des canevas existants à des tableaux d'entités dont les noms des attributs sont différents mais conservent la même sémantique. De la même manière, on pourra par exemple changer des unités des valeurs numériques pour challenger les systèmes.

Livrables: Jeux de données tabulaires à T0+3; Corpus (tableau-texte) à T0+9; Rapports expérimentaux à T0+12. **Gestion des risques:** Risques négligeables car le MNHN dispose des données et de l'expérience sur la génération de textes à partir de canevas à trous. Bien que disposant de données sur d'autres organismes

[2] Fulgoromorpha List On the Web http://hemiptera-databases.org/flow/ [3] www.xper3.fr

biologiques, nous ciblons des insectes ce qui est en cohérence avec l'expertise des participants MNHN. **Partenaires:** MNHN (12pm), SU-MLIA (1.5pm), LAMSADE (0.5pm).

Task 3.2: Enrichissement des données

Objectifs L'objectif est d'étendre le spectre de données de la tâche T3.1 avec la mise en oeuvre de plusieurs opérateurs en interaction, en série ou en parallèle. Cette étape nécessitera l'enrichissement des données tabulaires initiales (sur FLOW) et la production de textes plus complets (associés aux données Xper3).

Description Alors que les textes de la tâche T3.1 correspondent à des retranscriptions partielles simples et factuelles des jeux de données tabulaires considérés, nous visons dans cette tâche à produire des textes manipulant des opérateurs plus complexes, élaborés par des opérateurs d'extraction ou de comparaison agissant de concert. A ce stade, nous fournirons des exemples de phrases utiles aux biologistes, qui serviront de base pour la production de canevas permettant leur production en masse pour divers contextes.

Dans un premier temps, ces phrases exprimeront des contenus élaborés à partir de données factuelles agrégées à partir d'une colonne des tableaux, du type "Les espèces du genre 'G' ont toutes des ailes réduites", ou en comparant éventuellement les données de plusieurs colonnes (e.g., "Les spécimens de l'espèce X mesurent en moyenne 5 mm et ont les ailes antérieures et l'abdomen de la même couleur"). Dans un second temps, il s'agira de définir des types de phrases demandant un certain raisonnement sur les tableaux étudiés. Les exemples précédents pourront ainsi évoluer par exemple en "Toutes les espèces du genre G ont des ailes réduites dans une famille F où elles sont normalement développées", qui exprime une observation de singularité d'un groupe d'individus dans les données, ou "le spécimen N de l'espèce X mesure 7mm alors que la moyenne dans l'espèce est 5mm", qui exprime un avis sur un problème possible de catégorisation. La définition de ces types d'observation en concertation avec les partenaires, permettra de mettre en place des procédures pour l'extraction automatique de ce genre de phrases, afin de produire des jeux de données conséquents pour alimenter les modèles d'apprentissage. Des plans d'organisation du discours seront également imaginés pour structurer des enchaînements de canevas de génération, déclenchés selon des règles mises en place par les experts du domaine.

Livrables: Jeux de données enrichis associant données tabulaires, canevas et textes correspondants aux 2500 genres et 14000 espèces. Corpus de descriptions agrégatives simples à T0+12 et impliquant des opérateurs complexe à T0+21. Corpus de textes longs à T0+15. **Gestion des risques:** Risques négligeables sur le plan méthodologique car il suffira d'enrichir la tâche T3.1 selon des principes qui auront été éprouvés et reposant sur une expertise reconnue des participants MNHN. Les risques sont limités par l'accroissement progressif de la complexité des systèmes de règles à définir. **Partenaires:** MNHN (12pm), SU-MLIA (1.5pm), LAMSADE (1.5pm).

Task 3.3: Évaluation et valorisation des synthèses produites

Objectifs : Interprétation et valorisation des résultats produits par les systèmes. Intégration de ces résultats dans les portails en ligne FLOW et Xper3. Évaluation de l'impact pour la communauté scientifique biologique.

Description : Sur la base des résultats obtenus en tache T3.2, une analyse fine des résultats sera nécessaire afin d'identifier précisément la pertinence des singularités qui auront été relevées. On évaluera dans un premier temps la sémantique et la véracité des phrases produites par une analyse manuelle. Un travail d'annotation des résultats obtenus selon divers critères sera réalisée, en terme de véracité factuelle, fluidité du langage et également intérêt pour la communauté. On évaluera ensuite l'impact pour la communauté scientifique biologique, par l'intégration de ces résultats dans les portails web FLOW et Xper3. Dans la base FLOW, ils intégreront et enrichiront les textes à trous déjà disponibles, signalant ainsi clairement si nécessaire des besoins supplémentaires de données ou d'analyse aux utilisateurs de la base. Les communautés de systématiciens des groupes d'insectes traités seront invitées à une évaluation qualitative (notation et annotation libre) des textes proposés, selon un système de feedback déjà en place dans FLOW et Xper.

En permettant la personnalisation des synthèses générées, via les systèmes produits en T2.3, l'objectif sera d'amener les biologistes consultant FLOW ou Xper à exercer une analyse critique des données mises à leur disposition selon leurs besoins courants. Les phrases générées feront ressortir les singularités du dataset et alerteront par "avis" soit sur un jeu problématique des données concernées (incohérence, données incomplètes ou absentes), soit sur des particularités biologiques vraies qui pourront alors les orienter vers de nouvelles problématiques et directions de recherches intéressantes à aborder et à analyser, et où il semblerait

plus pertinent de porter les efforts des taxonomistes. Une analyse des comportements face à l'outil sera menée selon des traces d'usage enregistrées.

Livrables: Sites web FLOW et Xper3 enrichis par une API d'interrogation de tableaux à T0+33. Ajout de fonctionnalités de personnalisation à T0+39. Rapport d'analyse des usages à T0+36 et T0+45. **Gestion des risques:** Le seul risque associé concerne un possible usage trop faible des APIs par la communauté. On veillera à communiquer de manière conséquente dans les congrès et sur les plateformes, afin de montrer l'intérêt que les chercheurs ont à utiliser l'outil et aider son développement. On envisage un déploiement par version (version stable/version beta). **Partenaires:** MNHN (13.5pm), SU-MLIA (1.5pm).

Work Package 4: Transfert : Cas d'étude Document Intelligence

Responsable: J.Staiano (RECITAL)

Ce lot concerne notre second cas d'étude, pour la synthèse textuelle de tableaux issus de **documents d'entreprises** (rapports annuels, RSE,...) à laquelle le partenaire du projet RECITAL est confronté. Les données de supervision étant plus difficiles à obtenir que pour le WP3, un défi de ce lot concernera l'adaptation de modèles pré-entraînés à ce nouveau domaine. Une attention particulière sera portée sur la mise en oeuvre d'évaluations avec des utilisateurs finaux.

Task 4.1: Extraction de données

Objectifs Sélection des documents sources, pré-traitement et extraction semi-automatique des tableaux et des textes correspondants.

Description Dans le cas d'usage financier, nous envisageons l'utilisation de ressources incluses dans des rapports d'entreprise, au format pdf, comportant des tableaux financiers accompagnés de textes discutant des contenus de ces tableaux. Ces rapports, pour l'essentiel du domaine public, sont obtenus par le partenaire RECITAL auprès de ses clients, dont l'activité est de les analyser pour formuler des décisions motivées en lien avec des demandes de subventions (e.g., écologiques). La détection et la segmentation d'éléments dans ces documents seront effectués via des logiciels propriétaires déjà développés par RECITAL.

L'extraction de tableaux dans des pdf est néanmoins une tâche difficile, pour lequel le partenaire RECITAL mettra en oeuvre toute son expertise pour définir un pipeline de traitement efficace, via notamment le développement de méthodes d'apprentissage basées sur des annotations humaines et des prédicteurs appris à identifier les éventuelles incohérences de structure et de sémantique dans les tableaux produits. Un travail conséquent concernera également la détection et le filtrage des textes associés, en apprenant à identifier les textes discutant d'éléments du tableau considéré. Cela pourra se faire par exemple par la mise en oeuvre de techniques de génération de questions et de réponse automatique à des questions [27].

Livrables Jeux de données associant données tabulaires et exemples textuels à T0+12; **Gestion des risques:** Le risque le plus important concerne la présence de données trop bruitées dans les jeux fournis. Ce risque est limité par l'expertise du partenaire et le recrutement d'un ingénieur dont une tâche importante concernera le filtrage et la correction manuelle des données extraites. **Partenaires:** RECITAL (12pm), SU-MLIA (0.5pm).

Task 4.2: Adaptation au domaine

Objectifs Sachant que les tableaux du domaine financier sont possiblement relativement bruités, nous souhaitons nous appuyer sur des connaissances acquises à partir des entraînements réalisés sur des domaines mieux pourvus, tels que ceux considérés en WP3. Notre hypothèse est qu'il est possible de définir un transfert important des stratégies de sélection d'opérateurs au domaine financier, car les besoins d'extraction d'informations saillantes dans les deux types de corpus peuvent pour une large part être similaires (e.g., on cherche des valeurs maximum ou moyennes, ou à distinguer des groupes d'éléments qui se distinguent des autres, quel que soit le domaine d'application).

Description Pour permettre cela, nous proposons de chercher à affiner les réseaux pré-appris sur le corpus biologique, de manière à ce qu'ils soient capables de conserver de bonnes performances sur le corpus biologique, tout en atteignant de nouvelles capacités sur le domaine financier. Une possibilité pour éviter le déséquilibre brutal des modèles (e.g., catastrophic forgetting) serait d'instiller des exemples du domaine financier de manière progressive, selon des stratégies de planification adaptables aux performances sur les deux corpus [21]. Une autre approche, que l'on souhaite privilégier, serait de se baser sur des méthodes de distillation de modèles, qui visent à distiller de nouvelles compétences à une politique commune, tout en conservant les spécificités des deux tâches dans des réseaux distincts, de manière à éviter l'interférence

des gradients, bien connue pour les problèmes de renforcement multi-tâches par exemple [36]. Il s'agira d'apprendre de nouveaux réseaux spécifiques au corpus financier, régularisés via divergence des distributions (e.g. KL divergence) par rapport aux réseaux biologiques. On pourrait également imaginer un apprentissage conjoint avec une architecture centrale et deux architectures spécifiques, avec distillations des connaissances communes dans la partie centrale. Des contraintes inspirées de CycleGAN [41] peuvent également être considérées, avec recherche de fonctions de transports entre les distributions $p(\xi|\tau)$ conditionnées par les tableaux issus des deux corpus. Enfin, il sera possible de favoriser encore l'adaptation en pré-entraînant les modèles de langue au domaine financier, à la manière de [11] pour le transfert data-to-text multi-lingue.
Livrables: Modèles du WP1 adaptés au domaine à T0+24; Modèles du WP2 adaptés au domaine à T0+42. **Gestion des risques:** Étant donné le large spectre de possibilités envisagées, les risques sont assez faibles. La mise à disposition des outils à des utilisateurs finaux, clients de RECITAL, permettra par ailleurs d'exploiter des retours en contexte et d'ainsi affiner les performances des modèles. **Partenaires:** SU-MLIA (10.5pm), RECITAL (4pm), LAMSADE (2pm).

Task 4.3: Évaluation et impact pour le domaine

Objectifs: Cette tâche se focalise sur l'évaluation des générations sur des documents du domaine financier ainsi que de leur impact potentiel sur des tâches applicatives pertinentes pour RECITAL (i.e., recherche d'information, question-réponse).
Description; Afin d'évaluer la capacité de transfert sur le domaine financier, nous proposons d'exploiter une méthodologie d'évaluation originale, que nous avons proposé récemment [31]. Un travail préliminaire est en cours pour son extension au data-to-text [27] mais de nombreux enjeux sont encore restants. Il s'agira de s'appuyer sur les jeux de données produits en T4.1 pour adapter et étendre une approche d'évaluation de synthèses textuelles par des systèmes de génération de question et réponses à des questions : les questions générées à partir des descriptions produites doivent pouvoir être répondues à partir des tableaux et inversement. L'expertise de RECITAL en évaluation des modèles permettra de compléter cette étude par la mise en place d'indicateurs statistiques de qualité textuelle.

L'évaluation d'impact pour le domaine se fera au travers de la mise à disposition des outils via les plate-formes RECITAL. Un des enjeux de RECITAL vise à favoriser l'accès à l'information de documents structurés (incluant des tableaux) via des tâches de recherche d'informations ou de question-réponse. Une des hypothèses est que le data-to-text peut faciliter l'accès à l'information de tels documents en se servant de la description générée comme intermédiaire pour identifier l'information pertinente. Les descriptions générées peuvent alors être indexées pour un système de recherche d'information ou alors être lues par un système de compréhension/raisonnement pour la génération de réponse à des questions.

Dans un second temps, l'évaluation pour le domaine concernera la mise à disposition des outils de synthèse textuelle sur la plateforme, permettant l'interaction des clients de RECITAL avec ces modèles, et la production de synthèses à la demande, en fonction des contenus des rapports analysés. Une analyse des comportements des utilisateurs face aux systèmes sera menée. Cette évaluation d'impact concernera bien entendu également un volet personnalisation, permettant aux utilisateurs de définir les aspects qui les intéressent en fonction de leurs besoins.

Livrables: Rapport d'évaluation selon des modèles question-réponse à T0+36; Intégration aux outils de RI de RECITAL à T0+36; Mise à disposition des systèmes via une API d'interrogation des documents à T0+42; Rapports d'usages à T0+48. **Gestion des risques :** Comme pour la tâche T3.3, le principal risque concerne le manque d'usage de la part des utilisateurs finaux. Une attention particulière sera portée à l'ergonomie des outils mis à disposition, permettant une bonne prise en main de la part des utilisateurs. Les personnes impliquées prendront en compte les retours des utilisateurs dans une perspective d'amélioration constante de la plateforme, dans l'intérêt des relations de RECITAL avec ses clients. **Partenaires:** RECITAL (11.5pm), LAMSADE (7pm), SU-MLIA (4pm).

II Organisation du projet et moyens demandés

a Coordinateur scientifique et consortium

Le consortium réunit de fortes compétences en apprentissage profond et par renforcement pour la modélisation de données non structurées (SU-MLIA : S. Lamprier; LAMSADE: Yann Chevaleyre), le data-to-text et

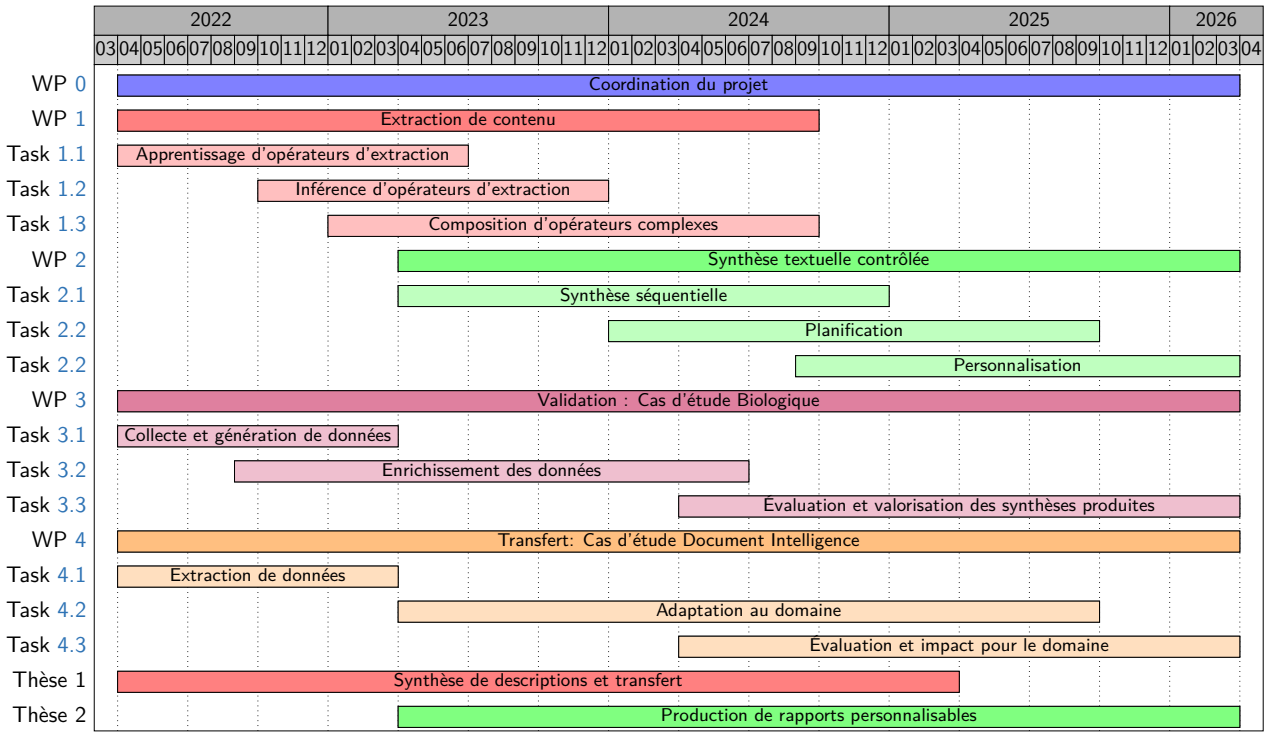

Figure 2. Diagrame de Gantt du projet.

Chercheur	Person.month	Appel, agence bourse allouée,	Titre du projet	Coordinateur scientifique	Début–Fin
Sylvain Lamprier	12,5	ANR	COST	Lynda Tamine	2018-2022
Alexandre Allauzen	12	ANR	SPEED	Lionel Mathelin	2020-2024
Régine Vignes Lebbe	8	ANR	e-Col+	Pierre-Yves Gagnier	à venir (2021-2029)

Table 1. Implication des coordinateurs scientifiques dans des projets en cours

la recherche d'information (SU-MLIA: B. Piwowarski, L. Soulier, V. Guigue; RECITAL: J.Staiano) et la génération du langage naturel (SU-MLIA : S. Lamprier, B. Piwowarski; LAMSADE : A. Allauzen et RECITAL: J. Staiano). Chez RECITAL, M.Durand est spécialisée en collecte et traitements de données en ligne, et l'évaluation des modèles en contexte industriel. Le MNHN (R.Vignes Lebbe et T.Bourgoin) vient compléter le consortium, en y apportant sa grande expertise scientifique dans le domaine de la biologie, pour la spécification des attendus, la constitution des ressources et la validation des sorties générées. Alors que SU-MLIA, le LAMSADE et le MNHN sont des partenaires académiques, RECITAL est une PME dont l'activité R&D est centrée sur le traitement automatique du langage. Son implication permet de confronter les avancées du projet à des cas d'usages industriels avec enjeux très importants dans le domaine du Document Intelligence. Son expertise à la fois pour les aspects techniques et pratiques sera un atout majeur pour la bonne connexion du projet avec les besoins sociaux-économique du monde de l'entreprise actuel. SU-MLIA a déjà collaboré avec RECITAL et le MNHN dans le cadre de divers encadrements de thèses, dont les productions ont été particulièrement fructueuses. Les publications récentes des membres du projet attestent de leurs compétences dans le domaine [16, 15, 2, 28, 29, 33, 32].

Sylvain Lamprier est Maître de Conférences à Sorbonne Université (équipe MLIA) depuis 2009. Il a soutenu son HDR en septembre 2020, sur le thème de l'apprentissage et l'inférence bayésienne pour l'extraction d'information dans les données sociales, notamment textuelles. Il a publié divers articles dans des conférences et revues du domaine (ICML, NEURIPS, IJCAI, JMLR, etc.), et participé activement au montage et réalisation de nombreux projets, notammment ANR (FRAGRANCES, LOCUST, COST, etc.). Depuis un certain nombre d'années, il s'est investi dans le domaine de l'apprentissage profond par renforcement, avec l'encadrement de plusieurs thèses (dont une avec le partenaire RECITAL).

SU-MLIA L'équipe MLIA de Sorbonne Université est spécialisée en apprentissage statistique et apprentissage profond. C'est l'une des entités leader en apprentissage profond en France. Sa recherche va de la conception théorique aux développements algorithmiques, pour de nombreux domaines d'application tels que la vision par ordinateur, le traitement du langage naturel et l'analyse de données complexes. L'apprentissage de représentation, l'inférence bayésienne et l'apprentissage par renforcement pour la génération de données structurées sont au coeur de ses recherches depuis de nombreuses années. S. Lamprier, porteur du projet, sera impliqué dans toutes les étapes du projet. L. Soulier, par son expertise en data-to-text et gestion des hallucinations sera impliquée principalement dans les tâches du WP2. B. Piwowarski se focalisera plus intensément sur le WP1, autour des questions d'apprentissage d'opérateurs. V. Guigue, déjà en collaboration avec le MNHN pour d'autres projets, permettra de fluidifier les échanges entre les WPs.

LAMSADE Le LAMSADE, créé en 1974, au sein de l'Université Paris Dauphine est un laboratoire d'Informatique initialement dédié à l'aide à la décision et la recherche opérationelle. Le projet MILES (Machine Intelligence and LEarning System) rassemble les chercheurs s'intéressant à l'apprentissage machine et venant de différents horizons comme par exemple la théorie des jeux, les mathématiques appliqués et le traitement automatique des langues et des images. Alexandre Allauzen est professeur à l'École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI), effectuant sa recherche au LAMSADE et à l'Université Paris-Dauphine. Son principal sujet de recherche est l'application du deep-learning au traitement automatique des langues et en particulier sur les modèles génératifs. Yann Chevaleyre est professeur à l'Université Paris-Dauphine et il s'intéresse à différents champs de l'apprentissage machine: de l'apprentissage par renforcement à la robustesse des modèles. Ils seront tous deux impliqués sur l'ensemble des WPs du projet, avec un investissement plus important sur le WP2.

MNHN Le Museum national d'Histoire naturelle (MNHN) est un établissement de recherche scientifique et de diffusion de la culture naturaliste. L'UMR ISYEB (Institut de Systématique, Evolution, Biodiversité) est une unité majeure des recherches en Systématique en France. Cette discipline construit les systèmes de description et de classification indispensables à tous les travaux sur le vivant. Au sein de cette UMR l'équipe LIS (Informatique et Systématique) détient une expertise en systèmes d'information et développement de méthodes d'aide à l'identification. Régine Vignes Lebbe, bioinformaticienne, est directrice de l'équipe LIS, professeur, et conceptrice des méthodes et algorithmes mis en oeuvre dans la plateforme métier Xper3. Elle est aussi impliquée dans l'équipe française du GBIF (Système mondial d'information sur la biodiversité) et dans les programmes du MNHN d'informatique pour la Biodiversité. Thierry Bourgoin est professeur MNHN, systématicien et entomologiste, membre du Comité exécutif du Consortium européen de Taxonomie (CETAF) et directeur adjoint du Consortium international Species 2000/Catalogue of Life. Il coordonne le développement de bases de données DbTNT du MNHN-Paris, dont le contenu servira de pilote pour les applications du projet ACDC. Tous deux se concentreront sur le WP3, en collaboration avec WP1 et WP2.

RECITAL Fondée en 2017, RECITAL applique les très récents progrès du traitement automatique du langage (TAL / NLP) aux documents non structurés (documents, emails, contrats..). Les solutions de Document Intelligence de RECITAL permettent aux entreprises de transformer leurs données non structurées en avantages compétitifs en fluidifiant plusieurs cas d'usage (recherche d'informations, analyse de contrats, traitement de flux de mails, évaluation des risques, process de Due Diligence). RECITAL, dont le chiffre d'affaire en 2020 avoisinait le million d'euros, compte 30 collaborateurs dont 6 docteurs, 3 doctorants CIFRE, et plus de 40 000 utilisateurs actifs. Son équipe de R&D publie régulièrement dans les meilleures conférences internationales (NeurIPS, ICML, ACL, EMNLP..) et contribue au développement de l'état de l'art dans le domaine. Jacopo Staiano est un chercheur confirmé, responsable scientifique chez RECITAL, dans le domaine du traitement de la langue et l'apprentissage profond (h-number: 22). Il sera impliqué sur les WP 1, 2 et 4 pour assurer des développements en forte collaboration avec les partenaires académiques. Marie Durand est une chercheuse (docteure depuis 2013), développant chez RECITAL des méthodes de collecte de données, de traitement de ces données, et d'évaluation des modèles sur ces données. Elle sera très impliquée, avec l'ingénieur recruté, sur les différents aspects du WP4.

b Moyens mis en oeuvre et demandés

L'aide demandée, s'élevant à 556 355 euros, concerne en grande partie le financement de ressources humaines. Pour les partenaires académiques, en coût marginal, il s'agit de personnel non permanent (deux

thèses de doctorat co-supervisées SU-MLIA / LAMSADE et 18 mois d'IE pour le MNHN). Pour le partenaire privé RECITAL, en coût complet, l'aide couvre (à 45%) les coûts des deux permanents impliqués, ainsi qu'une année d'ingénieur. Le reste de l'aide concerne principalement des frais d'environnement, un peu de missions pour les académiques pour les déplacements en conférence (28kE) et des ressources matérielles, qui consisteront principalement en l'achat d'ordinateurs personnels et GPUs pour les doctorants (34kE). Les ressources de calcul seront complétés par des demandes d'accès au centre Sorbonne pour l'IA et au portail Jean Zay, ainsi que les ressources GPU déjà disponibles dans les équipes.

- **Sorbonne Université - MLIA (SU-MLIA)**: 1 thèse de doctorat (110 kE) sur WP1 et WP4 (36 pm) et un stage (3,5 kE) sur WP1 (6pm) . Resources machines: serveur GPU (3 cartes + serveur) et 2 PCs (15kE), Missions: conférences (10 kE). Coûts administratifs: 12 %.
- **LAMSADE**: Personnel: Idem SU-MLIA mais thèse et stage sur WP2 et WP4. Resources machines, missions et coûts administratifs: Idem SU-MLIA.
- **Muséum National d'Histoire Naturelle - ISYEB (MNHN)** Recrutement d'un IE bioinformaticien supervisé par les 2 personnels MNHN (66kE, 18pm) : T3.1: 6 pm pour l'extraction et le filtrage des données; T3.2: 5 pm pour l'enrichissement des données; T3.3: 7 pm pour la mise en place de methodologies d'évaluation et des APIs sur les plateformes visées. Ressources : 2 PCs (4kE). Missions : conférences (8kE). Coûts administratifs: 12 %.
- **RECITAL**: 2 chercheurs permanents (21pm) pour le suivi du projet, la conception et le développement des méthodes, 1 année d'ingénieur (70kE, 12pm) : T4.1: 7 pm pour la collecte et le filtrage des données dans les documents d'entreprise et T4.3: 5 pm pour le développement et le déploiement de l'API d'interogation sur la plateforme RECITAL. Frais d'environnnement: 68%.

	SU-MLIA	LAMSADE	MNHN	RECITAL
Personnel permanent	324 866€	199 600 €	244 887€	140 000€
Personnel non permanent	113 498€	113 498€	66 000€	70 000€
Coûts des instruments et du matériel	15 000€	15 000€	4 000€	0€
Missions	10 000€	10 000€	8 000€	0€
Frais d'environnement	16 619€	16 619€	9 360€	142 800€
Sous-total	**155 117€**	**155 117€**	**87 360€**	**352 800€**
Aide demandée		**556 355€**		

Table 2. Moyens demandés par partenaire (l'aide totale inclut 100% des éléments éligibles dans ce tableau, sauf pour RECITAL dont l'aide ne correspond qu'à 45% du sous-total).

III Impact et bénéfices du projet

a Adéquation à l'axe scientifique

Le projet PRCE ACDC vise l'axe 5.2 "Intelligence artificielle" du domaine "Sciences du numérique" de l'appel à projet générique selon plusieurs aspects :

- le développement de modèles basées sur l'apprentissage de représentation et le transfert, l'apprentissage profond et par renforcement.
- l'exploitation de données variées (texte et tableau), structurées et d'une grande variabilité
- la tâche adressée est complexe, elle relève à la fois du traitement automatique de la langue et de l'apprentissage statistique, avec pour objectif final d'interagir avec des utilisateurs.

De plus, l'implication d'un partenaire industriel tout au long du projet avec une application finale dédiée au *Document Intelligence* rentre tout à fait dans le cadre d'un projet PRCE.

b Impact scientifique

b1 Pour les domaines du Machine Learning / Natural Language Processing

Si l'on n'ambitionne pas dans ce projet d'atteindre un niveau humain pour interpréter des tableaux de données, nous sommes convaincus que les méthodes que l'on envisage auront un fort impact pour la communauté scientifique, car ils définissent des mécanismes d'adaptation haut-niveau pour la **compréhension des données**, dans les cadres applicatifs visés. Les avancées récentes en apprentissage profond (e.g. transformeurs structurels), nous permettent d'envisager sereinement ce genre d'objectifs, qui constitueront un pas important pour la communauté vers des **systèmes généralisables et personnalisables**, dont l'apprentissage ne se contente pas d'imiter les sorties observées mais recherche à combiner des stratégies d'extraction complexes

pour répondre à des besoins peu définis. Pour la communauté TAL, ce genre d'avancée est cruciale pour définir divers types de **systèmes guidés par les données** à disposition, plutôt que d'apprendre à simplement imiter des humains. Le séquençage et la **planification d'opérateurs** comme proposés en WP2 du projet est une proposition à fort potentiel pour **dépasser les problématiques d'effondrement de postérieure ou de biais d'exposition** auxquels sont très souvent confronté les systèmes de génération de la langue naturelle. Enfin, le domaine de l'apprentissage statistique est confronté à un besoin grandissant de méthodes capables d'**expliquer les décisions** qu'elles prennent (xAI), portées par diverses politiques pour la protection des individus face aux machines. Un reproche très souvent fait aux architectures neuronales concerne leur opacité, ce projet apporte un élément de réponse important à cette critique, par la définition de modèles d'extraction et de verbalisation basés sur des opérateurs explicites, dont on peut interpréter la sémantique, tout en conservant de grandes capacités d'expressivité.

b2 Pour le domaine biologique

Les travaux de NLP sur les données de biodiversité se sont jusqu'à présent souciés d'extraire des entités à partir de texte. Mais on dispose aujourd'hui de plus en plus de données structurées sous forme de tableaux. Les projets du GBIF, du CoL ou du World Flora-on-line sont des exemples d'initiatives internationales où la construction automatique de phrases et de textes les combinant serait un plus important et attendu à court terme, puis leur interrogation en language naturel à moyen terme. Au delà des applications directes du LIS (Xper3 et FLOW), ACDC montre ainsi un fort potentiel d'impact pour les domaines scientifiques sur la biodiversité et, face **aux besoins grandissants de devoir nommer et tracer l'origine naturelle des composants dans tous les domaines** de notre société. En outre, les capacités de contrôle de l'architecture envisagée favoriseront **la découverte de connaissances à partir des bases de données considérées, par sérendipité.** Pour les biologistes, au-delà d'**une rapidité, d'une évolutivité et d'une généricité dans la production de synthèses textuelles en langage naturel à partir de tableaux de données**, l'IA pourra ainsi également apporter une **aide à la décision** en exprimant des informations remarquables pouvant suggérer des orientations pour des recherches à mener.

c Impact socio-économique

Le domaine émergent du Data-to-Text répond à des enjeux sociétaux très forts. **Les besoins d'analyse et d'extraction de contenus issus de données tabulaires sont omniprésents dans de très nombreuses applications**, notamment dans les domaines du journalisme, de la santé, du marketing, de la finance, etc. Un des premiers exemples d'application fut la publication d'un article du *"Los Angeles Time"*, généré automatiquement à partir de données numériques sismiques. D'autres exemples ont concerné le suivi des flux numériques (bourse, billetterie, suivi de la population...), l'assistance aux diagnostic médicaux ou encore le soutien d'enfants en difficulté d'élocution (par exemple, pour les aider à mieux retranscrire leurs journées).

Le domaine en plein essor du *Document Intelligence* que nous considérons en WP4 de ce projet présente des enjeux considérables pour **le traitement d'informations critiques dans les secteurs de la finance, de la gestion des risques et du suivi des réglementations**. L'analyse et l'interprétation des données produites par les entreprises est un enjeu crucial pour la réglementation, le suivi, l'analyse et l'amélioration du fonctionnement de structures géantes et mondialisées dont l'influence est aujourd'hui considérable. L'importance de ces activités est telle que d'autres structures importantes privées (Deloitte, KPMG, Gide Loyrette Nouel, etc.) ou publiques (juridictions ou commissions spécifiques) sont elles-mêmes spécialisées dans cette analyse. Or une majorité de l'information utile dans ces documents est présentée dans des tableaux. Leur analyse est donc cruciale pour permettre une meilleure transparence dans l'activité de structures supra nationales qui publient de l'information dans une quantité telle que les acteurs responsables de leur analyse ne peuvent la réaliser efficacement. Les impacts à court terme du projet correspondent à **l'intégration de méthodes efficaces pour extraire les éléments importants des tableaux de manière facilement interprétable par des auditeurs financiers**, comme ceux des clients de RECITAL, qui sont confrontés à l'analyse de longs rapport pour prendre des décisions de subvention importantes en fonction des politiques courantes. À plus long terme on peut envisager que ce genre de projet aboutira à des **systèmes capables d'interpréter seuls des rapports entiers et émettre un avis circonstancié, avec prise en compte de la multi-modalité des documents analysés**. Ce projet est un pas important dans ce sens et ouvre de nombreuses possibilités pour le futur.

d Actions de transfert socio-économique

Le projet inclut un partenaire, RECITAL, en lien avec de nombreux clients dont les besoins sont d'analyser des rapports dont 60% des informations utiles sont incluses dans des tableaux (selon leurs retours). Actuellement ces clients utilisent la plateforme RECITAL pour effectuer des recherches dans des textes, par des systèmes de recherche d'information classiques, et à travers des systèmes de question-réponse travaillant sur des données textuelles. Le développement des méthodes du projet ont un fort potentiel pour le partenaire, et bien au delà, tant le marché pour ce genre d'innovations est gigantesque. De **possibles brevets** pourraient émerger du projet selon les résultats obtenus, pour protéger l'effort de recherche français et défendre les PMEs qui font face à d'énormes multi-nationales dans le domaine. En outre, le transfert des innovations du projet se fera par la **mise à disposition des outils sur diverses plateformes en ligne**, permettant le chargement de rapports et retournant une API d'interrogation des données tabulaires de ces documents.

e Valorisation scientifique et diffusion

La diffusion des résultats sera assurée par la **publication scientifique** des travaux du projet dans des conférences nationales et internationales de premier plan, en Machine Learning (ICML, Neurips, ICLR, etc.) et en traitement de la langue (EMNLP, ACL, NAACL,etc.). La diffusion des résultats également au travers d'évènements nationaux comme la conférence en recherche d'information CORIA, la conférence nationale en traitement de la langue TALN ou encore la conférence nationale en apprentissage automatique CAP. Dans la communauté biologique, on visera des publications dans les conférences internationales du TDWG et pour la problématique *Document Intelligence*, ICDAR.

Pour renforcer la diffusion scientifique, les partenaires s'impliqueront dans l'**organisation de challenges internationaux** comme SemEval ou NTCIR, qui assureront une grande visibilité et la réutilisation des jeux de données et des méthodes employées. Nous envisageons également l'organisation d'un **workshop data-to-text** dans le cadre de conférences internationales de premier plan, e.g. ICML ou EMNLP.

Les partenaires exploiteront également leurs **réseaux** pour promouvoir la diffusion des méthodes et logiciels mis au point : plateforme RECITAL, sites Webs MHNH, MLIA, LAMSADE. Afin d'assurer une dissémination accrue des développements du projet, on veillera à la **mise à disposition des logiciels et prototypes sur des plateformes de partage** de code comme Github ou GitLab. Des annonces **Twitter** seront publiées sur les comptes des différents partenaires.

L'expertise acquise dans le cadre du projet ACDC, ainsi que les jeux de données qui en seront issus, viendront **enrichir les enseignements dans divers programmes de Master en Intelligence artificielle et apprentissage automatique**, dans lesquels sont impliqués plusieurs partenaires du projet et au delà. **L'encadrement de stages et de thèses** dans le cadre de ce projet participeront à la formation des étudiants dans le domaine de l'intelligence artificielle.

Enfin, les partenaires s'engagent à participer à des **rencontres entre industriels et universitaires** au cours d'évènements comme les Meetups (Paris Machine learning ou WiMLDSParis) ou la fête de la science.

Références des Partenaires

[2] E. Delasalles, S. Lamprier, and L. Denoyer. "Learning Dynamic Author Representations with Temporal Language Models". In: *ICDM*. 2019.

[15] S. Lauly, Y. Zheng, A. Allauzen, and H. Larochelle. "Document Neural Autoregressive Distribution Estimation". In: *J. Mach. Learn. Res.* 18 (2017).

[16] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab. "FlauBERT: Unsupervised Language Model Pre-training for French". In: *LREC*. 2020.

[26] C. Rebuffel, M. Roberti, L. Soulier, G. Scoutheeten, R. Cancelliere, and P. Gallinari. *Controlling Hallucinations at Word Level in Data-to-Text Generation*. 2021.

[27] C. Rebuffel, T. Scialom, L. Soulier, B. Piwowarski, S. Lamprier, J. Staiano, G. Scoutheeten, and P. Gallinari. *Data-QuestEval: A Referenceless Metric for Data to Text Semantic Evaluation*. 2021.

[28] C. Rebuffel, L. Soulier, G. Scoutheeten, and P. Gallinari. "A Hierarchical Model for Data-to-Text Generation". In: *ECIR 2020*. 2020.

[29] C. Rebuffel, L. Soulier, G. Scoutheeten, and P. Gallinari. "PARENTing via Model-Agnostic Reinforcement Learning to Correct Pathological Behaviors in Data-to-Text Generation". In: *INLG*. 2020.

[31] T. Scialom, P.-A. Dray, P. Gallinari, S. Lamprier, B. Piwowarski, J. Staiano, and A. Wang. "SAFEval: Summarization Asks for Fact-based Evaluation". In: *arXiv preprint arXiv:2103.12693* (2021).

[32] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano. "ColdGANs: Taming Language GANs with Cautious Sampling Strategies". In: *NEURIPS* (2020).

[33] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano. "Discriminative Adversarial Search for Abstractive Summarization". In: *ICML* (2020).

Références autres

[1] S. Agarwal and M. Dymetman. "A surprisingly effective out-of-the-box char2char model on the E2E NLG Challenge dataset". In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 2017.

[3] D. Deng, Y. Jiang, G. Li, J. Li, and C. Yu. "Scalable column concept determination for web tables using large knowledge bases". en. In: *Proceedings of the VLDB Endowment* 6.13 (2013).

[4] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. "DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs". In: *NAACL-HLT*. 2019.

[5] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. "Diversity is all you need: Learning skills without a reward function". In: *arXiv preprint arXiv:1802.06070* (2018).

[6] K. Filippova. "Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data". In: *Findings of EMNLP*. 2020.

[7] C. Gardent and L. Perez-Beltrachini. "A Statistical, Grammar-Based Approach to Microplanning". In: *Computational Linguistics* 43.1 (2017).

[8] M. Geva, A. Gupta, and J. Berant. "Injecting Numerical Reasoning Skills into Language Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. 2020.

[10] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. "Toward Controlled Generation of Text". In: *ICML*. 2017.

[11] M. Kale and A. Rastogi. "Text-to-Text Pre-Training for Data-to-Text Tasks". In: *Proceedings of the 13th International Conference on Natural Language Generation*. 2020.

[12] Y. Kikuchi, G. Neubig, R. Sasano, H. Takamura, and M. Okumura. "Controlling output length in neural encoder-decoders". In: *EMNLP*. 2016.

[13] H. Kim, B.-H. So, W.-S. Han, and H. Lee. "Natural language to SQL: Where are we today?" In: *Proceedings of the VLDB Endowment* 13.10 (2020).

[14] G. Lample and F. Charton. "Deep Learning for Symbolic Mathematics". In: *CoRR* abs/1912.01412 (2019).

[17] R. Lebret, D. Grangier, and M. Auli. "Neural Text Generation from Structured Data with Application to the Biography Domain". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016.

[18] A. Léon and L. Denoyer. "Options Discovery with Budgeted Reinforcement Learning". In: *CoRR* abs/1611.06824 (2016).

[19] L. Li and X. Wan. "Point Precisely: Towards Ensuring the Precision of Data in Generated Texts Using Delayed Copy Mechanism". In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2018.

[20] B. Y. Lin, S. Lee, R. Khanna, and X. Ren. "Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-trained Language Models". In: *CoRR* abs/2005.00683 (2020).

[21] C. de Masson d'Autume, S. Ruder, L. Kong, and D. Yogatama. "Episodic Memory in Lifelong Language Learning". In: *NeurIPS*. 2019.

[22] F. Nie, J. Wang, J.-G. Yao, R. Pan, and C.-Y. Lin. "Operation-guided Neural Networks for High Fidelity Data-To-Text Generation". In: *EMNLP*. 2018.

[23] J. Novikova, O. Dušek, and V. Rieser. "The E2E Dataset: New Challenges For End-to-End Generation". en. In: *arXiv:1706.09254 [cs]* (2017). arXiv: 1706.09254.

[24] A. P. Parikh, X. Wang, S. Gehrmann, M. Faruqui, B. Dhingra, D. Yang, and D. Das. "ToTTo: A Controlled Table-To-Text Generation Dataset". In: *EMNLP*. 2020.

[25] R. Puduppully, L. Dong, and M. Lapata. "Data-to-Text Generation with Content Selection and Planning". In: *AAAI*. 2018.

[30] E. Reiter, S. Sripada, J. Hunter, J. Yu, and I. Davy. "Choosing Words in Computer-generated Weather Forecasts". In: *Artif. Intell.* 167.1-2 (2005).

[34] O. Shapira, R. Pasunuru, H. Ronen, M. Bansal, Y. Amsterdamer, and I. Dagan. "Evaluating Interactive Summarization: an Expansion-Based Framework". In: *arXiv preprint arXiv:2009.08380* (2020).

[35] O. Shapira, H. Ronen, M. Adler, Y. Amsterdamer, J. Bar-Ilan, and I. Dagan. "Interactive abstractive summarization for event news tweets". In: *EMNLP: System Demonstrations*. 2017.

[36] Y. W. Teh, V. Bapst, W. M. Czarnecki, J. Quan, J. Kirkpatrick, R. Hadsell, N. Heess, and R. Pascanu. "Distral: Robust multitask reinforcement learning". In: *arXiv preprint arXiv:1707.04175* (2017).

[37] B. Wang, M. Lapata, and I. Titov. "Learning from Executions for Semantic Parsing". In: *arXiv:2104.05819* (2021).

[38] S. Wiseman, S. Shieber, and A. Rush. "Challenges in Data-to-Document Generation". In: *Empirical Methods in Natural Language Processing*. 2017.

[39] S. Wiseman, S. Shieber, and A. Rush. "Learning Neural Templates for Text Generation". In: *Empirical Methods in Natural Language Processing*. 2018.

[40] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim. "Graph Transformer Networks". In: *CoRR* abs/1911.06455 (2019).

[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017.