La phase clef de préparation des données dans un projet d'Intelligence Artificielle

Bernard Quinio, Antoine Harfouche, Cyril Viallet, Rolande Marciniak

La phase clef de préparation des données dans un projet d'Intelligence Artificielle

Bernard Quinio*
Antoine Harfouche*
Cyril Vialet**
Rolande Marciniak***

***** Universite Paris Nanterre - *CEROS*

****** Université de Perpignan - *HNHP UMR 7194 CNRS-MNHN*

******* Universite Paris Nanterre - *IDHES*

Résumé :

Plusieurs avancées ont eu lieu en Intelligence Artificielle (IA) surtout dans les domaines ou les données sont massives et les questions précises. Peu de recherches traitent des cas plus difficiles tel qu'un faible volume de données et des questions plus ouvertes. L'objectif de cette communication est de montrer comment surmonter les difficultés liées au développement d'une IA pour un de ces cas difficile d'application. Notre recherche intervention dans le cadre d'un projet a pour objectif d'utiliser l'IA pour produire de nouvelles connaissances sur le comportement de l'homme préhistorique. L'approche adoptée dans cette communication repose sur le travail fondateur de Star (1989) concernant l'infrastructure des activités scientifiques, ses caractéristiques mais aussi les étrangetés qu'elle produit. Notre démarche s'appuie aussi sur une approche instrumentale issue des réflexions de Caseau (2018) sur l'IA. Avec ces apports, nous proposons une démarche itérative et en interaction avec l'infrastructure de l'organisation pour appréhender le développement d'IA et nous montrons que la phase de préparation des données, en lien étroit avec la mise en œuvre des algorithmes d'IA, y est déterminante.

Mots clés :

Intelligence Artificielle ; Infrastructure ; Classifications ; Standards ; Objet Frontière, Préparation de données, Projet

1 Introduction

L'intelligence artificielle (IA) comprend un ensemble d'approches, ayant chacune des objectifs plus précis que le raisonnement intelligent ; ces approches concernent la perception, le traitement du langage naturel, la planification, la navigation, la représentation des

connaissances et le raisonnement logique (Caseau 2018). Dans les domaines de la représentation des connaissances et du raisonnement logique, des avancées ont été réalisées surtout dans le cas de données massives et de questions précises portant sur des connaissances structurées (ADT 2018). L'exemple type se trouve dans les applications de reconnaissances des formes s'appuyant sur des réseaux neuronaux. Peu de recherches traitent des cas plus difficiles tel qu'un faible volume de données et des questions plus ouvertes ou des connaissances moins bien structurées. Notre terrain d'étude, qui correspond à un de ces cas, est un projet de recherche qui a pour objectif d'utiliser l'IA pour produire de nouvelles connaissances sur le comportement de l'homme préhistorique.

L'objectif de cette communication est de montrer que la plupart des difficultés liées au développement d'une IA pour ces cas difficiles d'application peuvent être surmontées lors de la phase préparation des données. Il s'agit d'augmenter le volume des données et de les enrichir en cohérence avec l'environnement organisationnel concerné, pour nous un laboratoire de recherche scientifique, et en harmonie avec les technologies d'IA utilisées. Nous avons donc dû rechercher des supports théoriques dans ces deux domaines : construction des savoirs dans les organisations scientifiques et choix des techniques d'IA

L'approche adoptée repose le travail de Star (1989) concernant l'infrastructure des activités scientifiques, ses caractéristiques mais aussi les anomalies qu'elle produit et qui impulsent une dynamique créatrice de nouvelles connaissances. Notre démarche s'appuie aussi sur une approche instrumentale issue des réflexions de Caseau (2018) sur l'IA. Cet auteur propose une matrice permettant de sélectionner les méthodes IA adaptées au problème à résoudre et des cycles itératifs courts pour vérifier l'adéquation de ces méthodes.

Dans la partie suivante (partie 2), nous présentons la théorie de l'infrastructure élaborée par Star ; puis nous l'associons aux méthodes IA les plus adaptées en fonction de leur contexte (partie 3). Ensuite, (partie 4) nous présentons le projet et le positionnons dans les approches IA, décrivons la préparation de données qui a permis la mise à jour de l'infrastructure informationnelle et conceptuelle, présentons l'utilisation de l'IA et les résultats obtenus. Enfin, nous conclurons par une discussion de nos résultats.

2 L'infrastructure conceptuelle, informationnelle et technique

La notion d'Objet Frontière (OF) fondée par Star et Griesemer (1989) a été largement utilisée en gestion mais de manière déconnectée du concept d'infrastructure (Lancini & Sampieri-Teissier 2012) (Trompette et Vinck, 2009). Susan Star écrivait dans une de ses dernières publications que son travail avait été souvent réduit aux OF présentés comme un synonyme de flexibilité interprétative (Star 2010). Or, après l'article fondateur de 1989, le cheminement de Star suggère que l'Objet Frontière (OF) est un moyen permettant de produire des connaissances soit, en les rendant explicites (classification, standardisation) soit, en approfondissant les anomalies détectées. Ces deux activités permettent la mise à jour de l'infrastructure initiale. Il convient ici de distinguer trois types d'infrastructures et des anomalies pouvant y être observées puis de présenter notre acceptation du concept d'OF en lien avec cette infrastructure.

2.1.1 L'infrastructure conceptuelle

L'infrastructure conceptuelle permet aux individus et aux groupes de gérer la connaissance par la construction de classifications et de standards. Le travail sur les classifications et les

standards (les construire, les maintenir et les analyser) fait partie de l'activité scientifique et technique.

La classification est une segmentation spatiale et/ou temporelle du monde (Bowker et Star, 1999). Un système de classification est un ensemble de boîtes dans lesquelles les choses peuvent être rangées afin de réaliser un certain travail : bureaucratique ou de production de connaissances. Cette définition large de la classification délaisse les propriétés d'ordre génétique (remonter aux origines), de mutuelle exclusivité et de complétude et dont on sait que peu de classifications admises les satisfont. Cette définition large relève d'un point de vue pragmatiste plus que formaliste ou puriste (Bowker et Star, 1999, pp. 1-32).

Un standard est un ensemble de règles admises pour la production d'objets. Il s'étend sur plus d'une communauté de pratiques. Les standards sont développés par des organisations professionnelles et/ou des Etats pour faire marcher les choses entre elles. Enfin, les standards ont une inertie significative et peuvent être difficiles et couteux à changer.

2.1.2 L'infrastructure informationnelle

Une grande attention est aussi portée à l'infrastructure informationnelle, à travers les systèmes d'information, les outils de management de la connaissance et depuis peu, l'intelligence artificielle. Des éléments de l'infrastructure conceptuelle sont souvent intégrés dans l'infrastructure informationnelle. Cependant, chaque entité relevant d'une communauté de pratique scientifique mémorise et traite des données en fonction de sa compréhension spécifique du domaine.

2.1.3 L'infrastructure technique

Si l'infrastructure conceptuelle et informationnelle sont largement abordées par Susan Star y compris avec des exemples dans le domaine des Système d'Information, l'auteur a seulement évoqué la couche technique. Nous soulignons ici son importance dans le cas de technologie comme l'IA ou la réalité virtuelle. L'infrastructure technique regroupe les ordinateurs, réseaux, câbles, machines, etc. L'infrastructure technique est conçue pour prendre en charge les infrastructures conceptuelle et informationnelle.

2.1.4 Les anomalies issues des études sur le savoir scientifique

Au cours de ses années de recherche, cinq étrangetés ont attiré l'attention de Susan Star et lui ont servi à construire sa théorie de l'infrastructure. L'anomalie 1 révèle l'ampleur du travail invisible qui sous-tend toute expérience scientifique et qui n'est souvent pas pris en compte. L'anomalie 2 signale qu'une grande richesse d'information peut être écartée et considérée comme sans importance car elle n'entre pas dans les dispositifs prévus de saisie. L'anomalie 3 montre comment s'effectue une transposition d'un champ disciplinaire à un autre qui peut être source de richesses mais aussi créer des incompréhensions. L'anomalie 4 attire l'attention sur les catégories résiduelles qui ne rentrent pas dans les cases préétablies ; s'il y en a trop cela peut demander de revoir les classifications. L'anomalie 5 provient d'une communication difficile entre des groupes différents, notamment concepteurs et utilisateurs.

2.1.5 Le rôle des Objets Frontières par rapport à l'infrastructure (OF)

Le concept d'Objet-Frontière (OF) est souvent présenté comme un outil de coordination de groupes humains hétérogènes (Carlile, 2002). Plus précisément, un OF est un objet abstrait ou

concret en lien avec une infrastructure, adapté aux besoins de plusieurs groupes et qui permet à la fois l'autonomie et la communication entre les groupes. (Trompette et Vinck, 2009). Dans notre vision des travaux de Star (Star 2010), un OF répond à des besoins non couverts par la standardisation de l'infrastructure, ce que Star appelle des catégories résiduelles ou alors à des anomalies découvertes dans le fonctionnement de l'organisation. Puis avec l'usage, l'OF peut devenir un nouveau standard ou une nouvelle classification qui va enrichir l'infrastructure.

3 Quelle intelligence artificielle pour quelle question ?

Selon Caseau (2018), il existe deux critères majeurs pour identifier l'approche d'Intelligence Artificielle à appliquer. Ces deux critères sont : 1) la profondeur de la question et 2) le volume de données disponibles. Chacun de ces deux critères correspond à un axe de la matrice que nous proposons de nommer QCDM (Question Connaissance Donnée Méthode) et de relier explicitement au concept d'infrastructure de Star. Le premier axe est lié à l'infrastructure conceptuelle du domaine scientifique. Il y a deux possibilités selon que la question est étroite / précise ou complexe / ouverte. Cet axe est lié au niveau d'abstraction de la question ; plus la question est complexe, plus le niveau d'abstraction est élevé. Le deuxième axe est lié à l'infrastructure informationnelle du domaine scientifique et distingue deux possibilités en fonction du volume important ou non de données existantes. Cela aura un impact important sur le choix des méthodes d'apprentissage : plus le volume de données disponibles est important, plus l'apprentissage automatique pourra être profond.

La matrice QCDM propose donc quatre situations et les relie à quatre familles de méthodes d'IA (ADT 2018). La situation 1 concerne des cas où la question est complexe avec peu de données disponibles. Les méthodes les plus appropriées à cette situation sont les agents intelligents et les multi-agents. La situation 2 concerne des questions simples avec peu de données disponibles. Les méthodes les plus adaptées à cette situation sont les méthodes d'apprentissage automatique classique (machine Learning). La situation 3 concerne des situations où la question de recherche est précise et où les données sont disponibles en grande quantité. Les méthodes d'apprentissage en profondeur et les méthodes d'apprentissage par renforcement sont les mieux adaptées à ce type de situation. La situation 4 concerne des cas avec une question complexe avec un grand volume de données disponibles. Les approches cognitives similaires à IBM Watson sont les plus adaptées à ce type de situation.

	Peu de données (giga-octets)	**Beaucoup de données (téraoctets)**
Question précise	Situation 2 Méthodes traditionnelles simples	Situation 3 Deep Learning
Question ouverte / complexe	Situation 1 Agents intelligents	Situation 4 Approche cognitive - Watson

Figure 1 : Matrice QDCM (Question, Donnée, Connaissance, Méthode)

4 Le PROJET

4.1 Méthodologie de recherche du projet

Cette communication présente une Recherche-Intervention qui a eu lieu au sein du Labo 1. Elle a commencé début 2017 et se finira au premier semestre 2020. L'objectif de cette Recherche-Intervention est de s'inspirer de la théorie générale de Star (1989) et la théorie intermédiaire de Caseau (2018) pour faire progresser le projet et de se nourrir ensuite des résultats du projet pour y faire évoluer ces deux théories. La démarche de Recherche-Intervention a été choisie car elle s'intègre parfaitement avec l'objectif du projet, terrain de recherche, à savoir « créer des solutions technologiques innovantes (Intelligence Artificielle et Réalité Virtuelle) pour étudier les comportements de l'Homme préhistorique. »

L'avantage de cette démarche repose sur le fait que les auteurs de cette communication s'alimentent en permanence des observations participantes qu'ils mènent mais tout en s'inspirant aussi du travail de recherche théorique qui guide leur action et leur interprétation. Du fait de son ambition de produire des connaissances, cette Recherche-Intervention s'inscrit dans une logique prescriptive intégrant une dimension descriptive et une dimension explicative. Au départ, il y a eu une analyse documentaire (Roussel et Wacheux 2005) qui a couvert une large panoplie de documentation existante sur le projet, sur le Labo1 et sur les partenaires. L'investigation prospective effectuée à l'aide de la documentation a été complétée par le recours aux entretiens. Deux types d'entretien ont été adoptés : entretiens non-directifs et entretiens semi-directifs centrés. Les entretiens non-directifs ont été utilisés dans un premier temps pour explorer le contexte et pour stimuler les réactions des acteurs par rapport au projet. Les entretiens semi-directifs centrés ont été réalisés dans une étape ultérieure, et avaient pour but de collecter les informations portant sur la réaction des acteurs par rapport à l'état d'avancement du projet et à l'évolution de leur rôle au sein de ce projet.

La méthodologie d'analyse de données s'est focalisée principalement sur la lecture flottante (Robert & Bouillaguet, 1997), sur les observations et sur le croisement des différentes interprétations effectuées par les auteurs de cet article. La lecture flottante a aidé à faire connaitre le contenu des documents en laissant venir à soi certaines orientations et observations qui ont permis de construire l'objet principal de cette recherche. Il s'agit donc de lire les documents et les entretiens et de les relire pour tenter de bien saisir les principale ressemblances et dissemblances. Ensuite, les observations ont permis la description des événements. Les co-observations ont permis le croissement des différentes interprétations présentées dans cette communication.

4.2 Description du projet

Le projet regroupe quatre partenaires qui doivent créer des solutions technologiques innovantes (Intelligence Artificielle et Réalité Virtuelle) pour étudier les comportements préhistoriques. Porteur du projet, le Labo1 est un laboratoire d'archéologie, il apporte au projet son expertise scientifique du domaine. Société1 est une startup spécialisée dans l'intelligence artificielle. Société2 apporte son savoir-faire dans la réalité virtuelle. Enfin, Labo2 est un laboratoire de gestion qui apporte son savoir-faire dans l'intégration et l'analyse des interactions. L'objectif du PROJET est de réaliser un simulateur qui permettra de valider des comportements des hommes préhistoriques dans un environnement immersif reconstitué. Le projet a débuté début 2017, il se finira au premier semestre 2020. A ce jour (décembre 2018), les solutions techniques sont en cours de finalisation et l'IA a été utilisée pour répondre à des questions ciblées des

archéologues. La partie réalité virtuelle du projet, non traitée ici, a permis aux chercheurs d'appréhender par le regard les résultats de l'IA en s'immergeant dans l'information et en leur donnant des outils pour la manipuler.

Le SITE préhistorique concerné est un site majeur, où, depuis la fin des années soixante, chaque artefact ou ossement extrait du site est saisi sur un carnet de fouille puis enregistré dans une base de données. La complexité du projet vient du nombre de disciplines archéologiques qui participent à l'analyse des données issues de la fouille. Neuf chercheurs du Labo1 participent au PROJET et couvrent six disciplines archéologiques différentes ayant chacune ses techniques et ses référentiels de connaissances. Un jeune chercheur archéologue est dédié à quasi temps plein sur le projet.

4.3 L'infrastructure initiale du Labo1 et les anomalies

Pour mettre en place l'IA l'infrastructure initiale doit être prise en compte.

4.3.1 L'infrastructure conceptuelle du Labo1

Les classifications utilisées sont différentes selon les disciplines. Par exemple, pour les outils et les déchets issus de leur fabrication, les préhistoriens utilisent des classifications typologiques. Pour le Labo1, c'est un lexique non publié qui est mobilisé, il est proche de la typologie très référencée de Bordes (1961). Ces classifications permettent de distinguer les différents types d'outils par une analyse des formes de l'outil et des techniques de production ; un racloir sera classifié par le type de retouche appliqué à son bord. Mais ce qui définit un outil est aussi sa fonction (découpe de viande, broyage des os, …) qui est difficilement identifiable au Paléolithique inférieur. De ce fait, des discussions peuvent avoir lieu sur la nature des outils déjà enregistrés dans la base, discussions qui dépendent des classifications utilisées.

Un standard important en archéologie est l'ensemble des règles et procédures utilisées pendant la fouille qu'elle soit préventive ou programmée. Il existe deux grands standards : la fouille dite planimétrique ou horizontale et la fouille verticale. Le choix du standard de fouille dépend du temps disponible, du potentiel archéologique, et donc des objectifs de recherches de l'équipe de fouille. Sur le SITE, après avoir procédé à des carottages et au vue des objectifs scientifiques fixés, une fouille planimétrique a été adoptée. Chaque objet est coordonné (replacé dans l'espace) à partir d'un carroyage. Il est ensuite numéroté. La fouille s'effectue en suivant le pendage (la pente) des niveaux archéologiques. Toutes les données issues de la fouille sont inscrites dans un carnet de fouille papier qui sera ensuite ressaisi dans la base de donnée centrale. On ne peut comprendre les données disponibles sans avoir décrit et compris les standards qui ont permis de les créer. Quel que soit le type de fouille, il faut déterminer des origines à partir desquelles seront faites toutes les mesures : le point d'origine, le plan zéro et le Nord de la fouille qui a été placé conventionnellement en direction du fond de la grotte. La maîtrise de la détermination des origines est essentielle pour toutes les simulations 3D ou VR qui seront réalisées ainsi que pour les contrôles par projection 3D des données de la Base centrale. Par ailleurs des standards de mesure spécifiques existent pour certaines disciplines.

4.3.2 L'infrastructure informationnelle

Le cœur de l'infrastructure informationnelle du Labo1 est la base de données centrale qui a été élaborée dès les années soixante-dix. Elle comporte plus de 500 000 objets ou ossements issus des fouilles. La base de données comporte au total, 1 391 635 lignes pour 222 Mo dans une vingtaine de tables. La table principale est la table carnet qui est directement basée sur le carnet

de fouille. Les chercheurs pratiquent tous cette BDD mais, selon leur spécialité, ne l'utilisent pas de la même manière. Par exemple, les sédimentologues utilisent très peu la base qui a été conçue pour les objets et les ossements et pas pour les sédiments. En plus de la base centrale, chaque chercheur développe ses propres données, le plus souvent sur des fichiers Excel qui contiennent ses mesures propres et des calculs faits à partir de références académiques de sa discipline. Le volume de données disponible même en comptant l'ensemble des fichiers personnels est loin de représenter un volume de type BigData qui est de l'ordre de dizaine de Terra Octets.

4.3.3 L'infrastructure technique

La base de données en place fonctionne sur un serveur partagé. L'équipement technique du Labo1 est limité tant en puissance machine qu'en capacité réseau ce qui contraint le travail de certains chercheurs. Les postes de travail en place ne sont pas capables de supporter ni les outils d'IA ni les outils de simulations. Le projet a nécessité de mettre en place un serveur dédié à l'IA dans les locaux de LABO1.

4.3.4 Les anomalies observées pendant le projet

1) Le travail invisible : la fouille s'effectue avec des bénévoles plus ou moins amateurs, aussi peut-il arriver que les responsables de fouilles soient conduits à effectuer des reprises du carnet de fouilles.

2) *L'information écartée* : un des fondements de la pratique archéologique consiste à enregistrer tout au moment de la fouille. Il n'y donc pas d'information écartée.

3) La transposition : plusieurs disciplines utilisent l'actualisme qui consiste à transposer des résultats d'observations scientifiques actuelles (ou subactuelles - récentes) au contexte paléolithique. Par exemple un référentiel WWF actuel est utilisé pour compléter les mesures archéologiques sur la faune et la flore. pour Dans certains cas l'utilisation de l'actualisme pose problèmes.

4) Les catégories résiduelles : dans des fichiers, certains zéro veulent dire « je n'ai pas vu mais il y en a peut-être » et certains autres veulent dire « il n'y en n'a pas ». Or, ces deux types de zéro seront utilisés de la même manière par les logiciels d'analyse de données alors qu'ils sont différenciés par les chercheurs expérimentés.

5) La communication difficile : au cours de la première année de PROJET, la communication entre les ingénieurs IA et les chercheurs du Labo1 s'est avérée difficile. En effet, d'une part les ingénieurs IA étaient peu disponibles et d'autre part, contrairement aux attentes des chercheurs, les ingénieurs IA n'avaient pas réussi à expliquer comment l'outil fonctionnait.

6) Le savoir invisible : nous avons détecté une nouvelle anomalie qui concerne les savoirs tacites (Nonaka 1994) mobilisés par les chercheurs. Un chercheur a ainsi identifié de manière certaine un objet du site dans autre endroit géographique et cela a été confirmé par un autre chercheur très expérimenté. Tous deux étaient persuadés que l'objet venait du site (et ils avaient raison) sans savoir l'expliciter. Ces savoirs tacites ne se manifestent qu'à l'occasion d'un travail invisible ; en effet n'est reconnu comme scientifique (publiable) qu'un travail explicite.

4.3. La préparation des données pour le PROJET

Les sources des données sont la base de données centrale augmentée des bases personnelles des chercheurs. Ce sont donc des sources hétérogènes qu'il faut rapprocher et, selon les

disciplines, les données seront sur des supports informatiques différents et dans des formats différents.

Un travail préalable a permis d'identifier les quatre niveaux classiques de données. Premièrement, les données brutes sont les mesures archéologiques enregistrées dans la BDD centrale ou les fichiers de données des chercheurs. Par exemple la largeur d'une phalange de cheval. Deuxièmement, les facettes sont des données brutes agrégées ou mesurées. Par exemple, la largeur déduite du sabot sur la base des mesures des phalanges. Troisièmement, les construits de premier niveau sont élaborés à partir des facettes qui sont interprétées. Par exemple, l'humidité du sol associée à cette largeur de sabot et d'autres facettes. Enfin, les construits de deuxième niveau sont formés des construits précédents. Par exemple, le climat de l'époque considérée.

Puis, l'équipe du Labo1, aidée par Labo2, a créé et mise en œuvre trois types d'Objets Frontières indispensables à l'utilisation ultérieure de l'IA : des cartes cognitives par discipline, des scénarios et des Datasets. Les cartes et les scénarios sont des formats standards d'échange et les Datasets sont des répertoires de traitement (tests) selon la typologie des OF de Star.

Tout d'abord des cartes cognitives par discipline ont été réalisées. Chaque chercheur du Labo1, avec l'aide des membres du Labo2, a réalisé une carte cognitive (mind mapping) de sa discipline en s'appuyant sur ses classifications et standards. Puis ces cartes cognitives ont été confrontées par discipline pour les agréger.

Ensuite, des scénarios de comportement ont été réalisés, toujours sous forme de carte cognitive, en confrontant les cartes par discipline. Ces scénarios relient plusieurs construits de deuxième niveau venant de plusieurs disciplines. Les données brutes sont identifiées dans ces cartes avec le nom des fichiers où elles se trouvent. On voit ici le lien qui a dû être fait entre l'infrastructure conceptuelle et informationnelle. Les questions posées dans les tests d'IA ont été déduits de l'analyse des scénarios.

Enfin, pour chaque test, des Datasets, sous forme de fichier Excel, ont été élaborés, au départ, en suivant un guide conçu par le Labo2 sur la base des interactions avec Société1 et Labo1. Le Dataset est élaboré par les chercheurs des disciplines concernées par le test. Un technicien du Labo1 a effectué un gros travail de reprise des données surtout lorsqu'il fallait utiliser un gros référentiel de données externes venant du WWF. Il faut noter que l'élaboration du Dataset dépend du type d'algo d'IA ; il y a donc une interaction forte entre la préparation des données et l'utilisation de l'IA. Chaque Dataset est finalisé lors de nombreux allers-retours entre l'ingénieur IA et le chercheur archéologue à temps plein sur le projet.

4.4 Mise en œuvre de l'IA

A cet état du projet, l'IA n'a été utilisé que pour faire des prédictions sur des construits de premier niveau uniques élaborés à partir d'une à trois disciplines (Projet1.1, Projet1.2, Projet 1.3). Si le projet global (PROJET) est bien dans la situation 1 de la figure 1, nous n'avons pour l'instant que traité des cas de situation 2. Même si les moyens de mesures indirectes ont permis d'augmenter le volume des données, nous ne sommes pas dans les situations 3 et 4 compte tenu de leur taille.

La technique d'IA utilisée est de type machine Learning avec un classique découpage des données en trois parties : apprentissage, test et prédiction. Cette technique est bien compatible avec la situation 2 de la figure 1. Cinq séries de tests complets ont été effectués à ce jour.

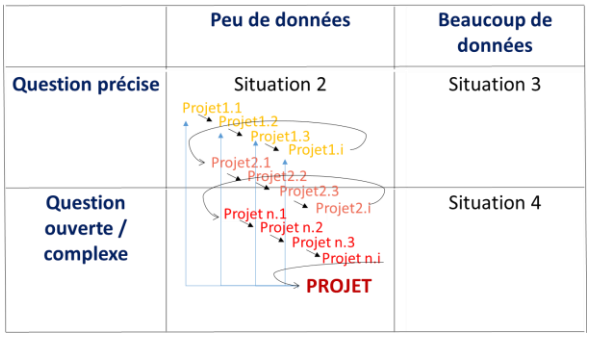

Figure 2 : Evolution du projet

A chaque test, une fois l'algorithme d'IA mis au point ce qui peut demander de nombreux allers-retours, un fichier est créé avec les commandes interprétées et les résultats obtenus. Ces résultats sont discutés d'abord par l'ingénieur de Société1 et le chercheur du Labo1 à temps plein sur le projet. Puis les résultats sont présentés lors de réunions mensuelles avec un représentant du Labo2 en observateur. Enfin des réunions plénières du Labo1 permettent d'échanger sur ces résultats avec tous les chercheurs associés au PROJET. L'équipe de Société1 a développé récemment un fichier de résultats dont les paramètres sont modifiables à distance, les chercheurs du Labo1 auront ainsi la possibilité d'interagir directement avec l'outil dans les prochaines semaines (Projet2.1, …).

4.5 Les difficultés et anomalies rencontrées et les solutions trouvées

Deux difficultés importantes ont été repérées dès le début du projet. Premièrement, le volume de données disponible est faible avec des Dataset ayant peu de lignes (individus) pour beaucoup de colonnes (attributs) avec en plus de nombreux trous entre lignes et colonnes. La Société1 a été obligé de rechercher les algos d'IA de type machine Learning les plus adaptés à cette difficulté.

Deuxièmement, dans les tests, même simples, les archéologues disposent de variables mesurables (par exemple, faunes et flores présentes) mais ils ont aussi besoin de variables non mesurables (par exemple, biomes ou climat à l'époque concernée). Pour réaliser l'apprentissage de l'algorithme d'IA, il est indispensable de trouver un moyen de mesure indirecte pour ces dernières variables. Après discussions et débats, les archéologues ont accepté d'utiliser deux moyens de mesures indirectes : premièrement l'utilisation de données actuelles externes (référentiels externes) quand l'actualisme est jugé acceptable, deuxièmement du dire d'expert donné par plusieurs chercheurs de Labo1 sous forme d'une échelle de valeur numérique. Il faut noter que cela permet aussi d'augmenter un peu le volume des données traitées par l'algo.

La première anomalie (travail invisible) a été prise en compte pour la partie Réalité virtuelle non présentée ici. Disons juste que les chercheurs ont insisté pour voir dans la réalité virtuelle les scans des pages de carnets papiers où sont notées toutes les corrections effectuées.

Les deux mesures indirectes des variables (référentiel externe, dire d'expert) ont été très difficiles à accepter par le LABO2 car elles sont liées à deux anomalies : anomalie 3 de transposition pour les référentiels externes utilisés dans l'actualisme et anomalie 6 de savoir invisible pour l'utilisation du dire d'Expert. L'acceptation des mesures indirectes a demandé de longuement revenir sur ces deux anomalies lors des réunions d'équipe du Labo1.

L'anomalie 4 des catégories résiduelles a dû être affrontée pour mettre en œuvre la technique de gestion des trous dans les données par l'IA. Plusieurs techniques existent : supprimer l'individu ou l'attribut, prendre la valeur moyenne ou utiliser le principe d'imputation des plus proches (K NL). Le choix de la technique dépendra de la résolution conceptuelle de l'anomalie 4.

L'anomalie 5 de communication difficile a aussi eu un impact direct sur le projet. Dans un premier temps, sans des explications claires sur le fonctionnement des outils de l'IA, les chercheurs du LABO1 ont été dubitatifs devant les premiers résultats des tests. Puis le nouvel ingénieur IA a pris le temps d'expliquer en détail le raisonnement de l'outil et les variables essentielles que l'algo a utilisé pour obtenir son résultat. Les résultats devenant explicables, ils ont beaucoup plus intéressé les chercheurs. Cela a même entraîné l'utilisation de nouveaux outils (LIME, SHAP), proposés par l'ingénieur de la Société1, explicitement dédiés à l'explicabilité des résultats de l'IA. Les résultats étant plus explicables, ils donnent lieu actuellement à des discussions plus riches dans le Labo1 et à un début de diffusion externe (publication et travail avec un autre labo). De plus la mise à disposition des chercheurs d'un fichier de résultats modifiables permettant de rejouer les tests va améliorer l'appropriation par les chercheurs de l'outil IA.

5 Discussion et conclusion

L'objectif de cette communication est de montrer qu'il est possible de surmonter des difficultés liées au développement d'une IA dans le contexte d'un faible volume de données et de question ouverte et ceci lors de la phase de préparation des données.

Le cadre théorique de Star est adapté au domaine de la recherche scientifique et l'analyse de l'infrastructure et de l'utilisation des Objets Frontières (OF) nous semblent tout à fait pertinente dans le cas de notre projet IA. Nous avons détecté une sixième anomalie concernant le savoir invisible et de ce fait enrichi le travail de Star et mieux structuré l'infrastructure.

Les résultats de notre recherche montrent que quand l'infrastructure n'est pas prête pour soutenir un projet IA, l'adoption d'Objets Frontières permet une mise à jour de cette infrastructure. Cette mise à jour transforme et enrichit l'infrastructure conceptuelle, informationnelle et technique rendant ce genre de projet réalisable d'une manière incrémentale. Cette dynamique transformationnelle des infrastructures est aussi alimentée par les anomalies qui sont inhérentes à tout travail scientifique.

Dans le cas du PROJET, l'infrastructure conceptuelle a été enrichie par le référentiel externe créé par le WWF et utilisé par les paléontologues et les palynologues. La mise en œuvre de ce référentiel externe a soulevé de nombreuses questions pour trouver les correspondances les plus adéquates entre des espèces du quaternaire et des espèces actuelles. La conception et l'élaboration des Datasets nécessaires à l'IA peuvent être considérées comme un nouveau standard. Les Datasets sont venus enrichir l'infrastructure informationnelle. Le système de fichier IA interactif et un nouveau serveur sont venus enrichir l'infrastructure technique.

Notre recherche est une première réflexion sur la transformation de l'infrastructure générée par l'adoption de l'IA. D'autres recherches suivront pour analyser la création et la diffusion des connaissances pour les chercheurs de Labo1 ainsi que l'utilisation des Objets Frontières étudiés en relation avec la création des connaissances.

6 Références

ADT (2018) Renouveau de l'Intelligence Artificielle et de l'apprentissage automatique ; Rapport de l'Académie des Technologies, Mars 2018, Paris, 102 pages

Bordes. F. (1961) Typologie du Paléolithique ancien et moyen, Ed. Delmas, Bordeaux 85 p.

Bowker G. C. et Star S. L., (1998), « Building Information Infrastructures for Social Worlds_ The Role of Classifications and Standards», *Community Computing and Support Systems*, LNCS 1519, pp. 231-248

Bowker G. C. Star S. L., (1999), *Sorting Things Out Classification and its consequence*, The MIT Press, 377 p.

Carlile, P. R. (2002), "A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development", Organization Science, Vol. 13, n°4, p. 442-455

Caseau (2018) « Accompagner la dissémination de l'Intelligence Artificielle pour en tirer parti », Enjeux Numérique, N°1 Mars 2018, pp. 9- 14

Lancini A. et Sampieri-Teissier N. (2012) « Contribution des Objets-Frontière (OF) à la Gestion des Connaissances (GC) : analyse des dépendances dans un bloc opératoire », Systèmes d'informations et management, n°4 Vol. 17, 8-37.

Nonaka I. (1994), "A Dynamic Theory of Organizational Knowledge Creation", Organization Science, Vol.5, n°1, pp.14-37

Robert, A.D., & Bouillaguet, A. (1997). L'analyse de contenu. Que sais-je ? France : PUF.

Roussel P. et Wacheux F. 2005, Management des ressources humaines : Méthodes de recherche en sciences humaines et sociales, Editions de boeck, 440 pages.

Star S. L. et Griesemer J.R., (1989), « Institutional Ecology, Translations and Boundary Objects : Amateurs and Professionals in Berkekey's Museum of Vertebrate Zoology, 1907-39», *Social Studies of Science*, Vol. 19, pp. 387-420

Star S. L., (2010), « Ceci n'est pas un objet frontière ! », *Revue d'anthropologie des connaissances*, Vol 4, N°1, pp. 18-35.

Trompette P. & Vinck D. (2009) « Retour sur la notion d'objet frontière » ; Revue d'anthropologie des connaissances ; 2009 Vol. 3, n° 1 pp 5 à 27