Etude

Description de l'e-infrastructure

Visa TM

Vers une infrastructure de services avancés de text mining

2017 / 2019

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

Description de l'e-infrastructure

Livrable Etude – partie 3

I Description de la solution organisationnelle retenue, des missions de la plateforme et des compétences nécessaires pour sa mise en œuvre I

Description du Document

Description de l'e-infrastructure

Lot	Etude
Participants	INIST (CNRS) MaIAGE (INRA)
Date de livraison	31/10/19
Nature : Rapport	Version : 1.0

Contributeurs

	Nom	Organisation
Rédaction	Claire Nédellec Fabienne Kettani	MaIAGE (INRA) INIST (CNRS)
Coordination	Fabienne Kettani	INIST (CNRS)
Relecture	Mouhamadou Ba Sophie Aubin Claude Dahdouh	MaIAGE (INRA) DIST (INRA) INIST (CNRS)

SOMMAIRE

Avertissement

Ce document contient des descriptions des résultats du projet Visa TM. Certaines parties peuvent être soumises à des droits de propriété intellectuelle. Avant réutilisation du contenu, il est nécessaire de contacter le consortium pour approbation.

Acronymes et sigles

TDM	Text and Data Mining
SI	Système d' Information
FAIR	Findable Accessible Interoperable Reusable
IST	Information Scientifique et Technique
MOOC	Massive Open Online Course
ESFRI	European Strategy Forum on Research Infrastructures
TGIR	Très Grande Infrastructure de Recherche
RDA	Research Data Alliance
EOSC	European Open Science Cloud
IA	Intelligence Artificielle

Résumé publiable

Après avoir recensé les besoins de la communauté de recherche par rapport à la fouille de textes dans un premier document, nous avons dressé dans un second une cartographie des différents acteurs impliqués sur ces activités et analysé diverses solutions organisationnelles pour répondre à ces besoins et rassembler ces différents acteurs, en mettant en lumière leurs avantages et freins éventuels.

Après une sélection argumentée d'une solution organisationnelle spécifique dans le contexte actuel de science ouverte, nous allons à ce stade de notre étude expliciter les différentes missions incombant à une plateforme de fouille de textes destinée à répondre de manière optimale aux services attendus aussi bien par les acteurs participants que par les utilisateurs finaux.

Dans ce document, nous avons ainsi répertorié les différentes activités de la plateforme et analysé les interactions entre ses diverses composantes. Ces activités font appel à différents métiers et des compétences nécessaires à leur exercice. La nouveauté de certaines activités a pour conséquence que les contours de certains métiers restent mal définis. Compte-tenu des réponses au questionnaire concernant la formation des répondants à la fouille de textes, il semble nécessaire d'avoir une réflexion sur ce point. Nous avions pu noter, en effet, que beaucoup d'entre eux étaient prêts à se former, que ces formations pouvaient prendre des formes diverses allant de l'autoformation avec l'aide éventuelle de collègues à des formations plus classiques de type universitaire. A l'heure actuelle, un certain nombre de métiers intègrent progressivement des compétences supplémentaires en matière de fouille de textes en fonction des besoins de terrain (comme les professionnels de l'IST par exemple) mais on assiste aussi à l'essor de nouveaux métiers comme celui de *Data scientist.* Un accompagnement des utilisateurs les plus novices par des personnels d'appui peut également être un moyen de faire progresser le développement des compétences en fouille de textes.

Introduction

Partant du choix d'un modèle organisationnel défini à savoir une solution académique semi-centralisée dont nous mettons en avant les avantages ayant déterminé ce choix, nous décrivons ici plus précisément l'e-infrastructure nécessaire pour la mise en place d'une telle solution. Nous détaillons les différentes missions qui relèvent de cette infrastructure afin de répondre aux exigences de production de services de fouille de textes à valeur ajoutée. Nous analysons également les interactions nécessaires entre les différentes composantes de l'infrastructure ainsi qu'avec les utilisateurs. Notre réflexion porte conjointement sur les ressources humaines nécessaires pour le fonctionnement et l'organisation de la plateforme, en termes de métiers mais aussi de compétences. Sur le constat d'une nécessité de montée en compétences sur la fouille de textes ou sur les nouveaux métiers qui s'y rapportent nous avons évoqué les besoins en formation et ébauché un panorama de l'existant.

Description de la solution organisationnelle

La section "Organisation fonctionnelle, quelles alternatives" du document "Acteurs et Scénarios" décrivait trois grand type d'organisations,

- Scénario commercial
- Scénario académique "tout distribué"
- Scénario académique semi-centralisé

La solution académique semi-centralisée favorise la formation, la mutualisation et la réutilisation. Elle permet de répondre rapidement à des besoins divers grâce à la configuration rapide d'applications à partir d'une bibliothèque d'outils partagés et la réutilisation massive de corpus annotés pour l'apprentissage et de ressources sémantiques spécialisées. Elle s'inscrit ainsi dans une démarche de Science Ouverte où les productions des scientifiques bénéficient aux scientifiques avec l'appui d'infrastructures de services.

1.1 Stratégie de science ouverte

Le futur dispositif devra s'inscrire dans le développement d'une stratégie partagée visant à rationaliser et à mutualiser les services de fouille de textes.

1.1.1 Les données

La solution vise en termes de données textuelles, à **rationaliser et mutualiser** l'accès aux **sources documentaires et sémantiques** et à réduire le coût d'ingénierie de la conception de corpus. Cela nécessite que l'accès aux données textuelles soit ouvert, transparent, adapté aux besoins et fiable en termes de qualité, de sécurité juridique et de caractérisation formelle des données.

La solution doit également contribuer à l'ouverture des données, résultats des traitements des services et faciliter leur réutilisation et leur intégration avec d'autres données.
Pour cela des évolutions du contexte d'opération de la plateforme sont nécessaires auxquelles la plateforme doit contribuer à travers l'analyse de ses besoins et ses pratiques dont,

- Changer les modes de publication pour les rendre ouverts
- Rendre les publications accessibles et réutilisables dans des **formats standards** généralisés
- **Agréger**, centraliser, partager et réutiliser des corpus bruts, prétraités, annotés
- Développer et partager des **ressources et modèles sémantiques** spécialisés
- **Combiner et réconcilier** les données par des ontologies de référence (web sémantique)

1.1.2 Les traitements

La solution vise en termes de traitement, à **réutiliser et combiner** les outils de fouille de textes pour les adapter aux besoins. Le document "Focus sur la fouille de textes" montre la fragmentation actuelle en des milliers d'outils et de plateformes techniques.
Elle doit également créer des environnements favorables à **l'expérimentation et la reproduction** qui sont des composantes intrinsèques à l'activité de recherche.

Ces objectifs sont complexes à mettre en œuvre techniquement. Au-delà des **plateformes de traitement et service**, il faut élaborer des **interconnections** durables entre les sources de données, les traitements et les services. Il faut assurer les moyens techniques de **stockage et calcul** de haut débit. L'**interopérabilité** des composants de fouille de textes repose sur la définition d'un langage formel adapté et l'encapsulation des composants.
Elle doit fournir aux utilisateurs sur leurs postes de travail des **modes d'interaction adaptés** à leurs profils spécifiques, par domaines disciplinaires ou par tâche.
Pour plus de finesse et d'**adaptation au besoin**, les solutions de fouille de texte doivent mieux intégrer l'apprentissage automatique et les ressources sémantiques. Cela nécessite que soit arbitré le compromis généricité-adaptation / coût-valeur ajoutée.

En fonction de ces objectifs, les solutions industrielles et académiques présentent des caractéristiques à la fois *complémentaires et supplémentaires*, mais la mutualisation et le partage des résultats de la recherche en fouille de textes, à destination de la recherche académique est une problématique collective. Apporter des solutions de fouille de textes académiques pour des besoins non rentables pour les entreprises du secteur est nécessaire dans tous les domaines de production de la connaissance, y compris non marchands.

La prise en charge du coût de l'ingénierie documentaire, de l'expertise du domaine, de la maintenance du service relèvent des missions des organismes de recherche et parallèlement sont des leviers d'innovation pour l'accès aux publications ou la valorisation des prototypes vers des produits. La solution académique semi-centralisée n'exclut pas des modèles mixtes d'exploitation impliquant par exemple les services développés par des PME innovantes exploitant des ressources libres.

1.2 Interactions entre les acteurs

La future plateforme doit se développer de façon coordonnée avec la stratégie nationale et européenne de développement de services (EOSC), d'infrastructures de recherche (ESFRI, TGIR), de données (RDA, plan S) et de traitements (Plan IA) illustrés par les instruments de la figure 1.

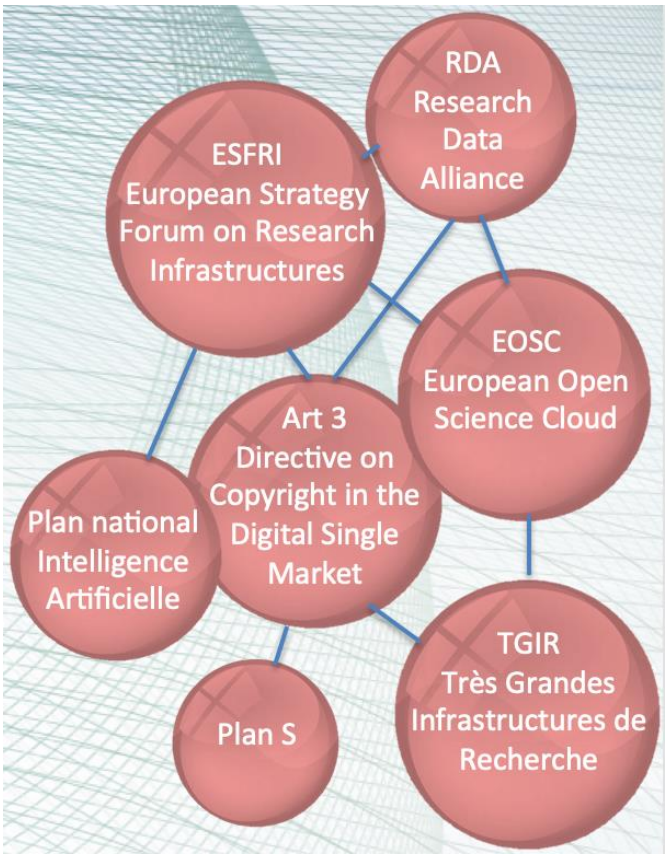


Figure 1. Contexte science ouverte

Plus précisément une solution académique mutualisée représente une solution à la fois réaliste et le meilleur équilibre en terme d'investissement et d'impact à terme parce qu'elle se base sur des solutions techniques existantes qui sont les éléments techniques du projet Visa TM, des modèles de développement déjà éprouvés dans d'autres domaines (communautés Open Source, réseaux IST). Elle s'appuie sur les analyses réalisées par les projets européens Future TDM et OpenMinTeD et les stratégies des partenaires du projet Visa TM, le CNRS, l'INRA et l'Université de Montpellier.

La figure 2 rappelle les acteurs et leurs interactions avec la plateforme identifiés dans le document "Acteurs et Organisations". Rappelons que la mission principale d'un dispositif, d'une e-infrastructure avancée de fouille de textes est de faciliter l'appropriation et l'utilisation des technologies de fouille de textes par les chercheurs et les acteurs de l'appui à la recherche.

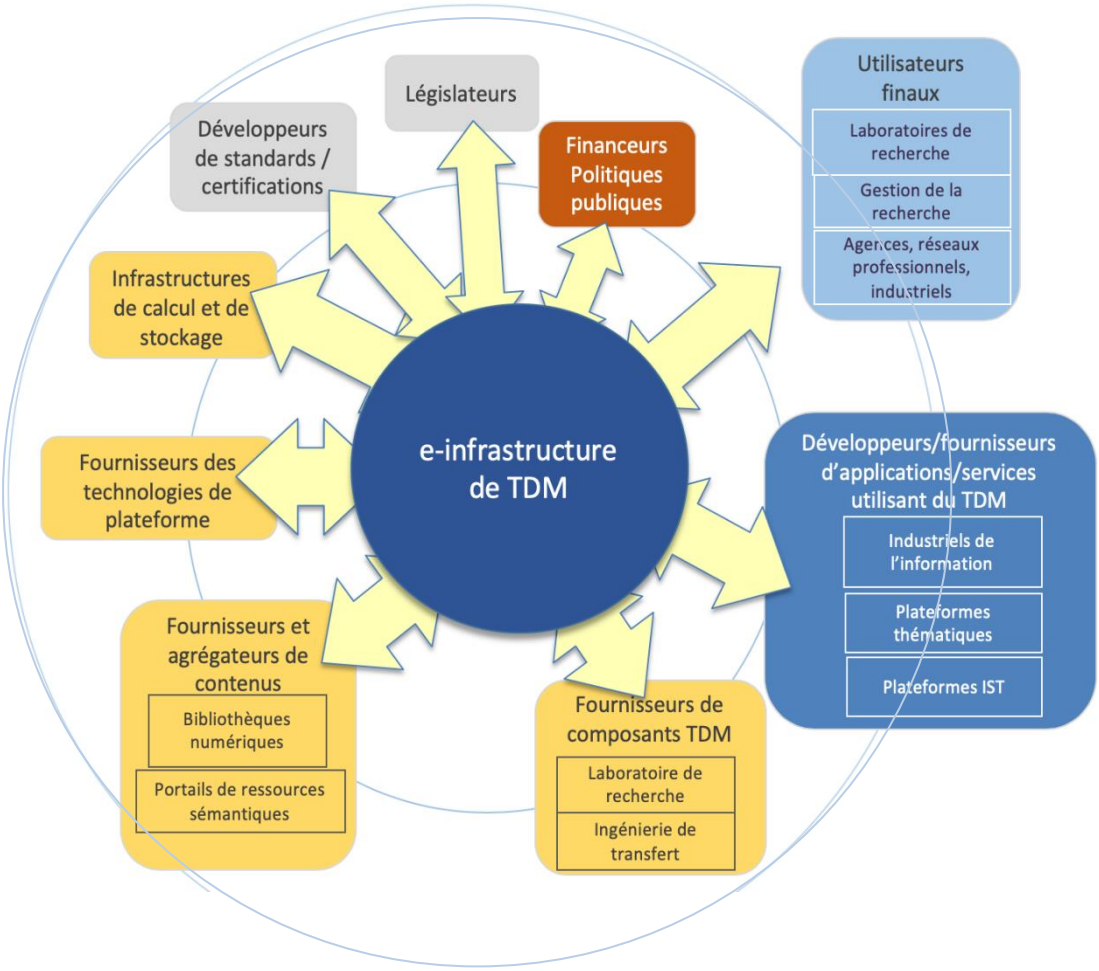


Figure 2. E-infrastructure fouille de textes et acteurs associés

Au-delà de son rôle technique, elle assure expertise, accompagnement et animation en s'appuyant sur son réseau de compétences pour

- assurer le lien entre la plateforme technique et ses utilisateurs
- faciliter son appropriation et s'assurer qu'elle réponde aux besoins des différentes communautés : utilisateurs finaux et intermédiaires, fournisseurs de composants et de contenu.
- contribuer à développer une communauté de pratiques dans le domaine de l'utilisation de la fouille de texte.
- promouvoir l'harmonisation des standards, représentations et pratiques, dans les principes FAIR auprès des différentes communautés concernées (données, publications scientifiques, ontologies, outils logiciels).

La figure 3 identifie et précise les rôles respectifs des acteurs qui interagissent avec la plateforme. Ils sont déclinés dans la section suivante, "Missions de l'e-infrastructure".

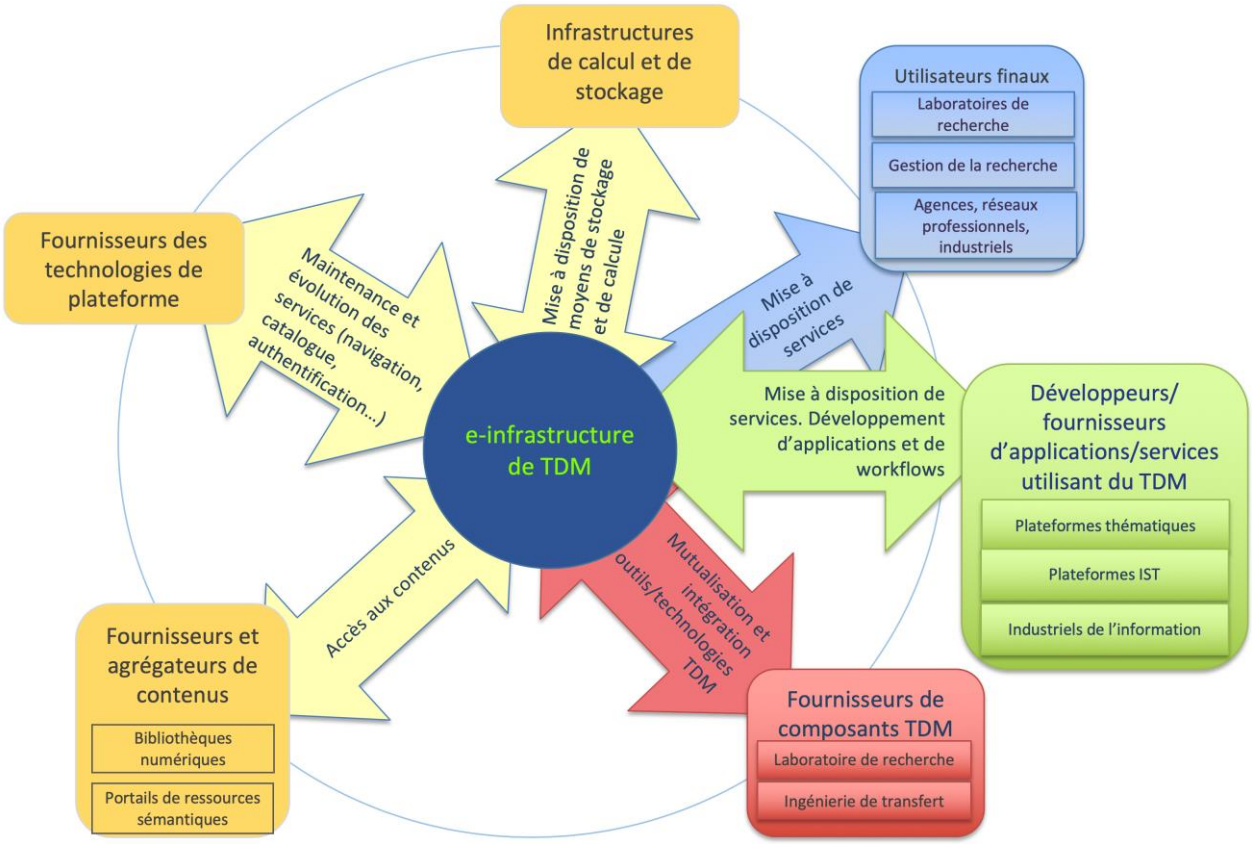


Figure 3. E-infrastructure de fouille de textes et rôle des acteurs associés

Missions d'une e-infrastructure de fouille de textes

Les besoins des utilisateurs finaux en matière de fouille de textes d'une part et les rôles des organisations parties prenantes d'autre part nous donnent un cadre pour définir le champ des activités de la future plateforme. L'efficacité de son fonctionnement se mesurera dans la durée par sa capacité à rationaliser les développements dans un compromis entre le service immédiat et spécifique en réponse à des besoins particuliers et l'investissement à plus long terme dans des solutions techniques et organisationnelle de fond.

Les missions de la plateforme devront donc se décliner en fonction des composantes externes (Figure 3) et de ses besoins de développement interne.

Ce document décrit une vision idéale de l'e-infrastructure cible, en dehors de toute considération de moyens.

Nous nous sommes attachés à décrire ici les fonctionnalités d'une plateforme de fouille de textes ainsi que les compétences et métiers nécessaires pour assurer son bon fonctionnement.

2.1 Objectifs dans l'e-infrastructure

"Une e-infrastructure est une combinaison de technologies numériques (matériel et logiciels), de ressources (données, services, bibliothèques numériques), de communications (protocoles, droits d'accès et réseaux), et de personnes et structures organisationnelles nécessaires à leur exploitation." Source: projet eRosa[1]

L'e-infrastructure que nous visons à mettre en place a pour objectif de faciliter l'appropriation et l'utilisation des technologies de fouille de textes par les chercheurs et les acteurs de l'appui à la recherche (IST, plateformes spécialisées…)

Elle vise également à devenir un point de référence pour l'utilisation de technologies de fouille de textes, à agir en tant que fournisseur de solutions pour les besoins en fouille de textes et à offrir un guichet pour les outils, les services et la formation dans ce domaine.

Sa masse critique lui permet de développer des compétences et savoir-faire et de conserver ces compétences et savoir-faire dans un contexte concurrentiel.

Elle garantit la confidentialité des données textuelles et des traitements.

[1] http://www.erosa.aginfra.eu/

Elle interagit avec les infrastructures et partenaires fournisseurs de contenus et technologies et/ou utilisateurs des services comme précédemment décrit dans la figure 3.

L'e-infrastructure implique principalement les acteurs académiques sans exclure a priori les acteurs commerciaux qui peuvent être utilisateurs et/ou fournisseurs de technologies et de contenus à préciser en fonction du modèle économique de l'e-infrastructure cible.

2.2 Composantes de l'e-infrastructure

Une e-infrastructure de fouille de textes s'appuie sur les composantes suivantes pour répondre aux diverses exigences de service :

Une plateforme technologique
Elle offre un ensemble de ressources bibliographiques et sémantiques, des outils, des workflows et des services de fouille de textes clefs en main et interopérables. Elle offre des espaces virtuels de travail et des mécanismes de création et de partage d'expériences. Elle inclut la documentation sur l'ensemble des ressources, sur les procédures d'exploitation et sur la réglementation. Elle permet ainsi l'exploitation des données et des services pour la fouille de textes à travers un accès transparent à des moyens de calcul et de stockage.

Un réseau ou centre de compétences
Il offre un appui méthodologique et technique, de la formation, de l'expertise et accompagnement sous différentes formes : supports, intervention de consultants, etc.

Des animateurs de communautés et « ambassadeurs »
Recrutés au sein des communautés impliquées et des instances juridiques, de formation, d'agences de moyens, ils constituent un lien entre les différents acteurs intervenant autour de l'e-infrastructure et contribuent à dynamiser les interactions et partages entre ceux-ci.

2.3 Missions de l'e-infrastructure

Ce chapitre va décrire l'ensemble des missions endossées par les composantes de l'infrastructure et les ressources humaines qui s'y rattachent tant du point de vue des activités que du point de vue des métiers et compétences.

2.3.1 Gouvernance et modèle économique [GOUV]

Pour assurer le bon fonctionnement, l'adéquation par rapport aux besoins, la légalité et la pérennité des services, l'e-infrastructure est dotée d'un organe de pilotage et de règles de gouvernance, ainsi que d'un modèle économique adapté à son contexte et à ses ambitions.

Le rôle de l'équipe

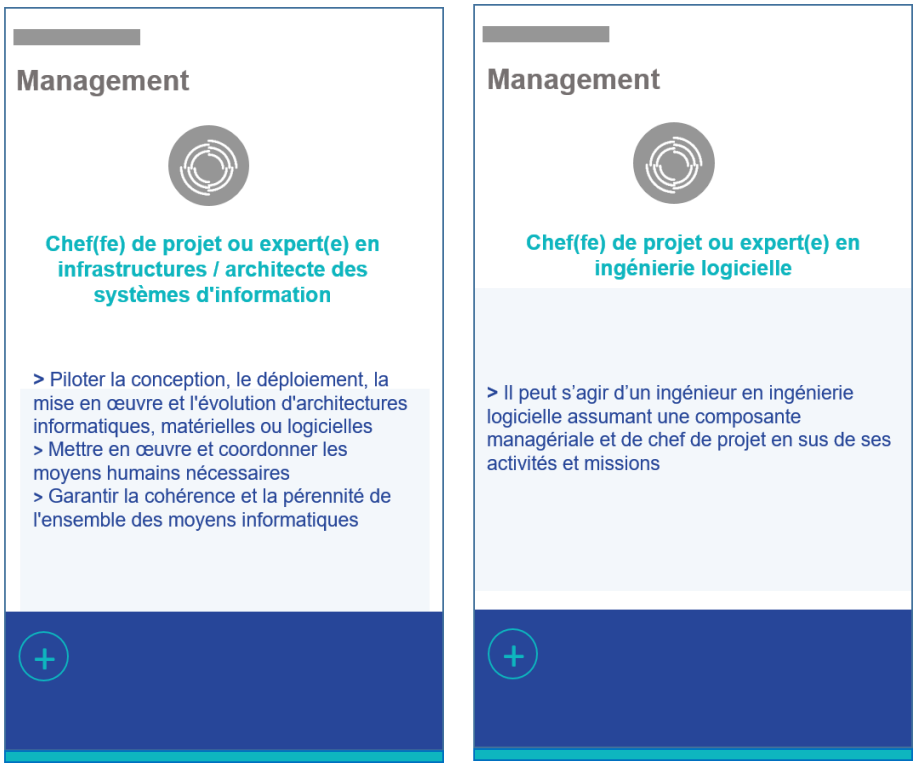
Il s'agit ici d'un véritable organe de pilotage qui :

> définit la politique de l'e-infrastructure (open, FAIR, etc.)
> définit et fait appliquer les règles juridiques et conditions d'exploitation
> en lien avec les équipes assurant les autres missions, met en place les comités de pilotage scientifique, technique, juridique, utilisateurs, etc. nécessaires, les sollicite et prend en compte leurs recommandations
> arbitre les choix technologiques et stratégiques selon les avis des différents comités

Selon le degré de centralisation, l'organe de pilotage pourrait aussi :

> définir et mettre en œuvre le modèle économique
> définir et mettre en œuvre (ou déléguer) les fonctions de gestion financières, humaines et administratives.

Les métiers



2.3.2 Acquisition de contenus [CONT]

L'e-infrastructure met à disposition des utilisateurs les contenus textuels (corpus issus de bibliothèques numériques) et ressources sémantiques utiles. Son objectif est de réduire le nombre et de simplifier les interfaces externes vers les fournisseurs de données textuelles structurées et non structurées pour les utilisateurs. Elle accompagne les contributeurs spontanés qu'elle sollicite en leur donnant les moyens de développer et d'intégrer les connecteurs techniques et de décrire les ressources. Elle définit avec eux les conditions

juridiques de l'utilisation des ressources mutualisées, les formats de ressources et de métadonnées.

Le rôle de l'équipe

- avec **[ANIM]** et grâce à son interaction avec les communautés Ingénierie des Connaissances et Ingénierie Linguistique avec éventuellement des spécialisations thématiques (agriculture, médecine, etc.), elle :
  - collecte les besoins des utilisateurs de la plateforme (types de contenu, qualité, couverture, etc.)
  - identifie les éditeurs scientifiques et intégrateurs de ressources textuelles susceptibles de fournir du contenu sur lequel appliquer la fouille de texte (veille)
  - examine les propositions spontanées des fournisseurs de contenus
- analyse et si besoin négocie les conditions légales (éventuellement financières) concernant les ressources textuelles
- évalue la capacité, les coûts nécessaires pour développer les fonctionnalités de la plateforme et les connexions aux infrastructures/services externes (par rapport à la valeur ajoutée).
- propose et valide avec l'équipe de pilotage les choix ponctuels et plus généralement, les orientations avec **[GOUV]**.
- réalise une veille sur les ressources, les standards, normes et formats
- organise les catalogues de corpus et de ressources (création/révision des catégories, curation des métadonnées par exemple)
- identifie le mode d'obtention des corpus et ressources sémantiques et, avec **[BASE]** contribue à leur mise en œuvre dans la plateforme soit (1) externalisée au fournisseur, soit (2) développée en interne.
- accompagne l'intégration des ressources dans la plateforme par les fournisseurs : par exemple, elle développe les connecteurs techniques (API, systèmes d'authentification, protocole de dialogue ressource (OAI), …) vers et les autres plateformes/services.
- analyse le format de représentation des fournisseurs, assure l'échange avec le format pivot de la plateforme (développe ou intègre les convertisseurs nécessaires) et documente les formats de représentation et le fonctionnement de ces convertisseurs).
- s'assure de la disponibilité de la documentation et des métadonnées des ressources sur la plateforme
- identifie les besoins en termes de visualisation/présentation pour un corpus (annoté)

Les métiers

2.3.3 Mutualisation et intégration d'outils/technologies de fouille de textes [TDM]

L'e-infrastructure définit et implémente les moyens nécessaires aux fournisseurs de composants de fouille de textes pour déclarer ces composants sur la plateforme et les rendre opérationnels et interconnectables avec d'autres composants. Elle maintient, étend et développe les connaissances et l'expérience utile en fouille de textes dans un environnement opérationnel.

Le rôle de l'équipe

- identifie les outils utiles et utilisables (veille sur la fouille de textes et écoute des besoins)
- organise le catalogue de composants (création/révision des catégories, curation des métadonnées par exemple)
- évalue les outils à intégrer en fonction de critères prédéterminés (performance, intégrabilité-interopérabilité, licence)
- évalue la capacité, les coûts nécessaires pour intégrer les outils par rapport à la valeur ajoutée (accompagnement du fournisseur)

- valide avec l'équipe de pilotage les choix ponctuels et plus généralement, les orientations avec la gouvernance **[GOUV]**
- accompagne les fournisseurs de composants / applications (conteneurisation, déploiement, description=métadonnées, maintenance…)
- avec **[ANIM]**, accompagne la production de tutoriels autour de l'utilisation des composants
- met en place les tests et évalue les fonctions offertes par un outil une fois intégré à la plateforme
- s'assure que la documentation est présente, suffisante, adaptée pour chaque fonction d'un outil (documentation technique et documentation utilisateur)
- recueille, gère et corrige les bugs et prend en compte les demandes d'évolution des outils (venant notamment de **[APPS]** et des utilisateurs)
- identifie et publie les lacunes de l'infrastructure vis-à-vis des besoins, en se basant sur une analyse des demandes

Les métiers

Une attention particulière doit être portée à la cohérence technique des composants d'applications, à leur communication entre eux, leur sécurité et à leur bonne intégration à l'infrastructure. Ce point peut être du ressort d'un architecte sur un projet de grande envergure.

2.3.4 Mise à disposition de services basés sur le développement d'applications et de workflows [APPS]

L'e-infrastructure répond aux besoins par la mise à disposition de services et outils de haut niveau. Elle accompagne les utilisateurs en dispensant des conseils sur l'utilisation de ses services, s'assure justement de leur utilisabilité, notamment par leur documentation.

L'e-infrastructure est en mesure de répondre à des demandes *ad hoc* pertinentes couvrant des besoins de développement spécifiques (à la carte) au travers d'applications/services dits *end-to-end*, clefs en main sur la base d'un cahier des charges précis.

Elle peut ainsi fournir :

- des services généraux
- des applications génériques
- des applications spécialisées

Le rôle de l'équipe

- accompagne les utilisateurs de la plateforme qui composent et déploient des applications sous la forme de workflows de fouille de textes par une expertise technique, méthodologique, aide à la documentation, dans les domaines d'application de la fouille de textes.
- organise le catalogue d'applications (création/révision des catégories, curation des métadonnées, par exemple)
- identifie auprès des utilisateurs et transmet à l'équipe de **[BASE]** les besoins d'évolution des fonctions de la plateforme, dont l'outil de composition de workflows (notamment)
- identifie auprès des utilisateurs et transmet aux équipes de **[TDM]** et **[CONT]** les besoins en composants de fouille de textes et en ressources
- identifie les besoins concernant le cycle de vie des applications et transmet à **[BASE]** (besoins de reproductibilité des résultats, accès à l'historique des expérimentations)
- dans le cadre d'un service à la carte :
  - analyse le besoin pour une nouvelle application et évalue l'opportunité et la faisabilité
  - identifie et collecte les ressources numériques, permettant de définir un/des corpus qui seront ensuite décrits et exploités par **[CONT]**
  - conçoit des workflows de fouille de textes (dans la mesure où les interfaces pour le faire sont adaptées à l'utilisateur, sinon le workflow devra être développé par **[TDM]**)
  - s'assure que les composants et les ressources numériques et sémantiques nécessaires sont disponibles et adaptées (si besoin demande leur ajout à **[TDM]** et **[CONT]**)

- garantit la compatibilité des ressources en faisant appel à [TDM] et [CONT]
- compose le(s) workflow(s) de fouille de textes avec les outils mis à disposition
- documente le workflow et met en place les éventuels supports de formation
- si besoin, adapte les ressources et paramètre les composants
- teste le workflow et évalue les sorties avec le commanditaire autant que nécessaire, modifie le workflow et son paramétrage
- teste et évalue les applications en condition réelle avec les données de la plateforme
- documente la nouvelle application
- livre la nouvelle application et si possible la publie sur la plateforme
- assure le suivi de cette application à long terme

Les métiers

« Hybride »

Data scientist

> Connaissances en TAL et en technologies sémantiques et appropriation du domaine de spécialité sur lequel les traitements sont appliqués
> Participation aux différentes briques nécessaires au montage des traitements de fouille de textes
> Maîtrise d'un certain nombre de langages de programmation informatiques (Python ,R, etc.) et de normes et standards
> Connaissances de l'ingénierie des données, veille active et anticipation de toutes les innovations sur l'open source

Très souvent issu d'un cursus en intelligence artificielle, avec des connaissances en machine learning, en statistiques et mathématiques.

Dans le cadre de cette mission il est nécessaire d'allier des compétences métier avec une compétence informatique et de traitement de données. Un recours à un *data scientist* peut donc s'avérer indispensable. Il peut être secondé si besoin par des personnels en mesure de faire l'interface avec des domaines spécifiques.

Data scientist (algorithmes, statistiques, apprentissage automatique…) : nous décrivons ici un modèle de métier vers lequel il conviendrait de tendre sans qu'il soit identifiable à l'heure actuelle concrètement en une personne unique dans les projets de fouille de textes existants. Parmi les activités dans son périmètre :

> extrait des connaissances à partir d'un volume de données
> produit des algorithmes sur les données pour anticiper leur comportement
> recommande des actions
> catégorise les données
> gère les données produites, collectées, et stockées pour les analyser, les exploiter et les transformer afin de créer de la valeur pour les métiers fonctionnels.

La valeur ajoutée s'exerce lorsqu'il crée des données pertinentes nouvelles en recoupant les données existantes.

Remarque : le travail d'un *data scientist* est variable d'un projet à un autre allant de la fouille de textes au développement de modèles prédictifs.

2.3.5 Maintenance et évolution/extension des fonctions de base [BASE]

L'e-infrastructure s'assure du maintien en conditions opérationnelles ainsi que de la mise à jour et de l'évolution du dispositif logiciel nécessaire aux services (outils de découverte de composants et contenus, construction et utilisation de workflows, accès aux résultats). Elle se porte garante :

- de l'environnement de développement
- de l'environnement d'utilisation (interfaces utilisateurs, services d'aide à l'utilisation)
- des applications de gestion et suivi (comptes, autorisations, espaces, accès aux statistiques, rapports, logs etc.)
- des connections aux e-infrastructures clientes (déploiement, mise à jour et mise en production)

Le rôle de l'équipe

- assure la veille sur les technologies utilisées sur des plateformes comparables ou émergentes
- intègre les nouvelles fonctionnalités
- gère l'intégration des contributions externes sur les technologies Open Source
- développe les fonctionnalités nécessaires (interfaces, outils, bases de données etc.)
- évalue la capacité, les coûts nécessaires pour développer les fonctionnalités de la plateforme et les connexions aux infrastructures/services externes (par rapport à la valeur ajoutée).
- propose et valide avec l'équipe de pilotage les choix ponctuels et plus généralement, les orientations avec **[GOUV]**.
- assure la coordination voire la synchronisation avec les plateformes associées
- développe les connecteurs techniques (API, systèmes d'authentification…) vers d'autres e-infrastructures
- organise et fait fonctionner le guichet support en lien avec **[TDM]**, **[CONT]** et **[APPS]**

Les métiers

Management

Chef(fe) de projet ou expert(e) en infrastructures / architecte des systèmes d'information

> Piloter la conception, le déploiement, la mise en œuvre et l'évolution d'architectures informatiques, matérielles ou logicielles
> Mettre en œuvre et coordonner les moyens humains nécessaires
> Garantir la cohérence et la pérennité de l'ensemble des moyens informatiques

Ingénierie logicielle

Ingénieur(e) généraliste (développeur)

> Contribution à une ou plusieurs phases du cycle de vie des logiciels : analyse, développement, qualification, intégration, déploiement dans le respect du cahier des charges, des normes et des règles de sécurité
> Maintenance corrective et évolutive des applications, modules et composants logiciels
> Coopération avec l'Ingénierie de la production

+ Spécialiste support et déploiement (DevOps)

Ingénierie logicielle

Ingénieur(e) généraliste (développeur)

> Contribution à une ou plusieurs phases du cycle de vie des logiciels : analyse, développement, qualification, intégration, déploiement dans le respect du cahier des charges, des normes et des règles de sécurité
> Maintenance corrective et évolutive des applications, modules et composants logiciels
> Coopération avec l'Ingénierie de la production

+ Spécialiste backend et frontend

Administration de données

Administrateur (trice) des systèmes d'information

> Responsable de la maîtrise d'œuvre d'un système d'information (SI)
> Définir et mettre en œuvre des procédures d'exploitation
> Suivi, sécurité et maintien en conditions opérationnelles
> Contribuer au cycle de vie des logiciels et applications

+ Il peut assumer des responsabilités de gestion de projet.

Ingénierie de production

Ingénieur(e) gestionnaire d'applications

> Mise en place d'environnements de production, intégration, développement avec tous les composants nécessaires au bon fonctionnement d'une plateforme
> Préparation et mise en ligne des applications issues, soit de développements internes, soit de la communauté Open source
> Intégration de tous les composants logiciels nécessaires au maintien en conditions opérationnelles des applications (supervision, sauvegarde, arrêt/démarrage/reprise)

Une attention particulière doit être portée à la prise en charge des dysfonctionnements éventuels dans une collaboration interactive avec les utilisateurs. Une prise en compte des contraintes sécuritaires est également indispensable.

2.3.6 Mise à disposition de moyens de stockage et de calcul [INST]

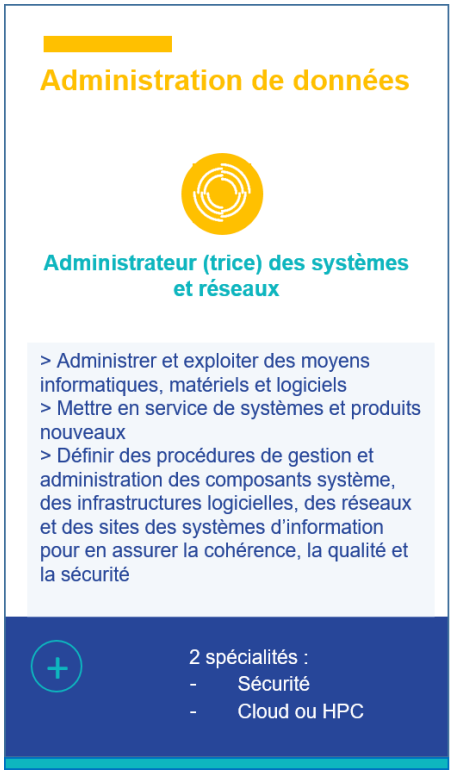
L'e-infrastructure s'assure du maintien en conditions opérationnelles ainsi que de la mise à jour des installations de stockage, de calcul et de réseau et connectivité avec d'autres e-infrastructures.

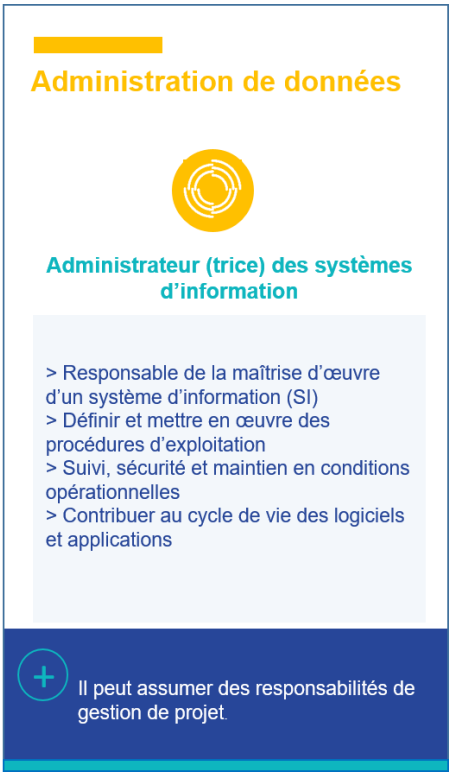
Cette mission peut être déléguée totalement ou en partie à une autre e-infrastructure comme GrNet, EOSC hub, Cines...

Le rôle de l'équipe

- dimensionne, obtient et fait évoluer le matériel informatique nécessaire au bon fonctionnement de la plateforme en fonction de sa charge (nombre d'utilisateurs, pics d'utilisation, coût informatique des traitements) et de la taille des données stockées.
- collecte les besoins, identifie les limites du système, notamment auprès de **[BASE]** et de **[APPS]**
- gère le parc informatique, les systèmes d'exploitation et logiciels des couches basses, les connexions
- déploie, configure et monitore les logiciels nécessaires au déploiement des services
- assure la continuité de service 24/7 et le dépannage des installations
- assure la sécurité des données en accord avec la réglementation et les politiques arrêtées par **[GOUV]**
- assure la pérennité des données selon une politique déterminée avec les autres équipes

Les métiers





2.3.7 Expertise, accompagnement, animation [ANIM]

Afin de fédérer l'ensemble des acteurs impliqués dans les processus de fouille de textes et les inciter à partager leurs expériences, il est nécessaire de mettre en place une activité d'animation autour de la plateforme.

Le rôle de l'équipe

- organise et contribue à l'animation de groupes de compétences autour de la plateforme et au sein des diverses communautés d'acteurs (avec **[TDM]**, **[CONT]**, **[APPS]**)
- promeut les services de l'e-infrastructure auprès des différents acteurs
- avec **[TDM]**, organise des événements pour les communautés TAL, fouille de textes, IA, etc. dans le cadre de conférences scientifiques notamment
- avec **[CONT]**, intervient auprès des éditeurs et des bibliothèques numériques (inscription dans des réseaux, participation à des événements…)
- communique sur les apports de l'e-infrastructure et de la fouille de textes en général, notamment en promouvant les "success stories", applications phare...
- facilite et harmonise l'accompagnement des utilisateurs et la création de documentation par **[BASE]**, **[TDM]**, **[CONT]** et **[APPS]**
- met en place, alimente et anime le site/portail de l'e-infrastructure, qui constitue le point d'accès à la plateforme technique, aux supports de formation, aux diverses informations relatives à l'actualité du domaine, etc.

- crée et alimente les comptes de l'e-infrastructure sur les réseaux sociaux
- contribue à l'organisation et à l'animation des événements internes avec **[GOUV]** pour des plénières et les autres missions sur des sujets plus spécifiques
- contribue à l'organisation et à l'animation des formations autour de la plateforme, notamment avec **[TDM]**, **[CONT]** et **[APPS]**
- accompagne juridiquement les utilisateurs

Les métiers

Communication

Médiateur(trice) scientifique

> Faire comprendre et traduire les concepts de la science pour le grand public
> Dans le cas de la fouille de textes, faire comprendre les apports de ces techniques dans les évolutions scientifiques, l'innovation et donc le transfert vers la société civile

Il ne s'agit pas là d'un véritable métier mais plutôt d'un rôle assumé par des personnes aux parcours assez divers.

Communication

Community manager

> Animer la communauté en ligne : une interface entre l'infrastructure et ses cibles, chargé de fédérer les utilisateurs autour de pôles d'intérêt communs
> Partage et collaboration de ressources numériques et sémantiques et d'outils et de normes et formats relatifs aux opérations de fouille de textes

Il peut se faire aider dans cette tâche par un médiateur scientifique ou assumer seul les deux fonctions éventuellement.

La figure 4 ci-dessous résume globalement les interactions entre les différents acteurs nécessaires à l'exercice des missions de la plateforme :

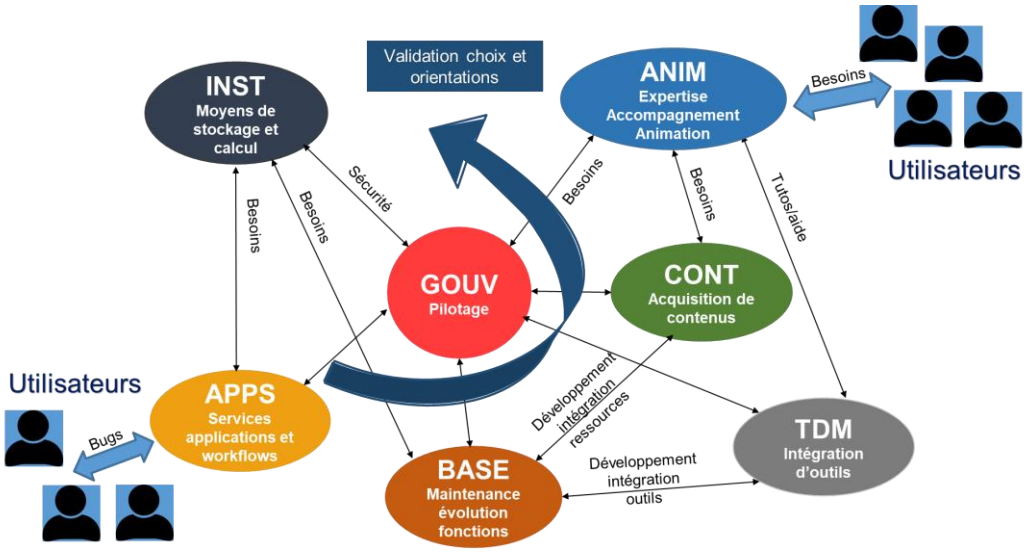
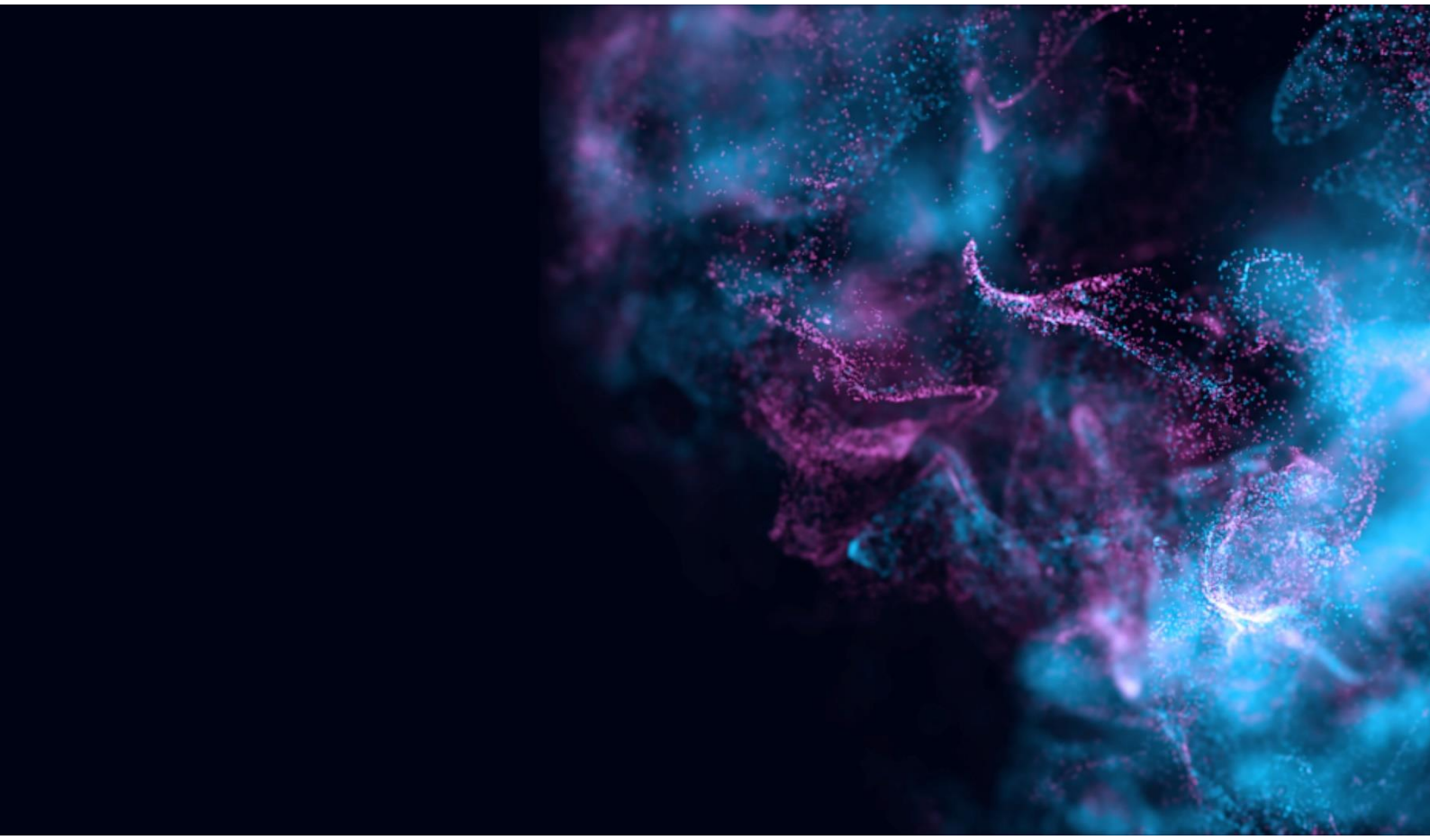


Figure 4. Organisation et missions d'une e-infrastructure de fouille de textes

Nous avons à ce stade de notre étude décrit les différentes missions d'une plateforme de fouille de textes et les différents métiers qui s'y rapportent. Nous avons ainsi déterminé des profils de poste sur chacune de ces différentes missions mais dans les faits ce découpage est probablement moins tranché et on peut imaginer par exemple qu'un ingénieur IST ou un ingénieur fouille de textes puisse endosser le rôle de chef de projet en infrastructures, qu'il puisse également se charger des aspects veille technologique et/ou juridique. Il conviendra donc dans une étape ultérieure de déterminer les nécessités réelles en termes de ressources

humaines, vues sous l'angle du nombre de personnes, pour optimiser le fonctionnement d'une plateforme de fouille de textes. Ce point n'a pas fait l'objet d'une exploration dans notre étude.



Métiers et compétences mis en jeu

Nous avons vu dans le chapitre précédent que la mise en place d'une infrastructure de fouille de textes nécessite d'avoir recours à des métiers très différents et qui peuvent s'incarner dans des profils de poste quelquefois hybrides. Ces métiers peuvent être plutôt habituels et bien décrits tout comme ils peuvent être plutôt nouveaux, en particulier ceux en lien avec les problématiques de l'intelligence artificielle.

Pour faciliter la compréhension et la lecture, nous avons regroupé dans le chapitre précédent l'ensemble des métiers abordés dans des fiches-métier les répertoriant selon différentes familles de métiers.

Nous pouvons ainsi constater qu'une part prépondérante de ces profils appartient aux familles « IST et documentation » ainsi que « Informatique » au travers des fiches Administration de données, Ingénierie logicielle et Ingénierie de production qui sont en charge des briques indispensables au bon fonctionnement d'une e-infrastructure en fouille de textes et dans lesquelles pourront être recrutés les chefs de projet en charge du pilotage.

Un métier un peu à part est celui du *data scientist* qui a émergé ces dernières années dans le sillage du *Big data* et des modalités de son exploitation. D'abord réservé à des personnes ayant endossé ce rôle au gré de leur évolution professionnelle, le *data scientist* est aujourd'hui entré dans les cursus de formation officiels et nous verrons dans le chapitre suivant quelles sont les formations qui y préparent. Doté de connaissances informatiques, mathématiques, en apprentissage automatique, il se doit, dans l'idée que nous nous faisons de l'e-infrastructure décrite ici, d'avoir également une compréhension des enjeux de l'IST et de certaines spécificités de domaine (la fouille de textes en médecine et en agriculture ne répond pas aux mêmes exigences, à commencer par les aspects juridiques).

Les autres familles viennent essentiellement en appui pour l'accompagnement de l'utilisateur vers un usage facilité de la plateforme (formation, tutoriels par exemple) et des actions d'animation visant au partage d'expériences et/ou d'outils.

Un aspect non négligeable est le cadre juridique qui a fait l'objet dans les dernières années d'un certain nombre d'évolutions que nous avons exposées dans le document « Acteurs et besoins ». Le contexte légal en constante évolution rend le soutien juridique d'une plateforme de fouille de textes capital.

Formations

4.1 Les formations actuelles

Ce chapitre dresse un état des lieux des formations proposées à ce jour dans le domaine de la fouille de textes, du text et data mining, des sciences des données de manière plus générale. Il est évidemment bien loin d'être exhaustif tant cette offre est pléthorique et en constante croissance et fait l'objet du document en annexe « Recensement de formations en fouille de textes ».

Globalement nous y avons trouvé :

- Des formations initiales en présentiel, principalement dispensées par des universités, plutôt longues et diplômantes.
- Des formations courtes (1 à 4 jours) en présentiel (les plus nombreuses), plutôt envisagées comme de la formation continue et dispensées aussi bien par des universités, des écoles d'ingénieur ou de management
- Quelques formations mixtes (*blended learning*)
- Des formations à distance, principalement sous la forme de MOOCs
- Des ressources (supports de formation) mises à disposition par le réseau des Urfist

Les thématiques sont très diverses qu'il s'agisse de cursus longs ou courts et quelquefois très spécialisées (centrées sur des langages informatiques spécifiques par exemple : Python, R,…):

- *Data science*
- *Big data*
- *Open data*
- *Data mining*
- Traitement et analyse d'images
- *Machine learning et deep learning*
- *Text mining*
- Traitement Automatique des Langues
- Analyse et apprentissage statistique
- Intelligence artificielle
- Web sémantique
- Cartographie, data visualisation
- *Crowdsourcing*
- *Data storytelling*

4.2 Quelles formations pour demain ?

Nous avions conclu dans le chapitre précédent qu'une caractéristique de certains métiers de demain dédiés à la fouille de textes était la forme très « hybride » de leurs compétences allant puiser celles-ci dans des disciplines très diverses : sciences de l'information, informatique, mathématiques, etc. Ce constat devrait également se traduire dans les offres de formation de

demain. Par ailleurs, la volonté d'étendre la fouille de textes dans le cadre de la science ouverte nécessitera sans doute d'inclure des bases en termes de connaissances sur la fouille de textes dans tous les cursus diplômants y compris dans l'ensemble des domaines de spécialisation afin de doter les étudiants de capacités à utiliser ces techniques dans leurs travaux de recherche.

Conclusion

Après avoir analysé les dynamiques en jeu au sein d'une plateforme de fouille de textes telles que nous l'avons imaginée, aussi bien entre les contributeurs en interne que ceux interagissant de l'extérieur, nous avons pu décrire les missions primordiales auxquelles elle doit répondre. Ce faisant, nous avons abordé la problématique des métiers nécessaires à son bon fonctionnement et avons été amenés à entrevoir les besoins en formation de demain qui y sont liés. Force est de constater que ces besoins commencent à être pris en compte ces dernières années du côté de l'enseignement supérieur mais que certains profils comme celui de *data scientist* restent encore en tension actuellement en raison d'un manque d'effectifs et du contexte concurrentiel entre le privé et le public. Cette réflexion devra constituer l'un des éléments de l'avancée dans la science ouverte dans les années à venir si l'on souhaite que la recherche française puisse tirer parti le mieux possible des potentialités que lui offre la fouille de textes

Index des figures

Annexes

Annexe 1

VisaTM étude

Missions et compétences nécessaires dans une e-infrastructure de fouille de textes

GOUV
Pilotage

Missions

Définir la politique de l'e-infrastructure
Définir et faire appliquer les règles juridiques et les conditions d'exploitation
Mettre en place les comités de pilotage scientifique, technique, juridique, d'utilisateurs, …nécessaires, les solliciter et prendre en compte leurs recommandations
Arbitrer les choix technologiques et stratégiques selon les avis des différents comités
Définir et mettre en œuvre le modèle économique
Définir et mettre en œuvre (ou déléguer) les fonctions de gestion financières, humaines et administratives

Activités

Pilotage organisationnel
Identifier ou susciter un nouveau besoin
Convaincre de l'intérêt et mobiliser des acteurs
Construire une proposition
Rédiger une proposition
Obtenir les ressources nécessaires (financières, humaines, techniques…)
Construire des partenariats
Evaluer les risques et opportunités
Faire le suivi de projet
Organiser les réunions
Animer les réunions
Etre garant de la qualité
Valoriser le résultat

Pilotage technique
Analyser les besoins et participer à la réalisation du cahier des charges fonctionnel du projet
Assurer une veille technologique
Assurer la conception de la solution au moyen d'expertises approfondies
Définir l'architecture logicielle et/ou matérielle avec le domaine d'application et les experts du domaine

Compétences

Pilotage organisationnel
Capacité à innover
Goût du risque
Connaissance de son environnement professionnel (métier, scientifique et technique)
Capacités à mobiliser les appuis
Animation de collectif
Animation de réunion
Sens de l'organisation
Initiation et conduite de partenariats
Capacité de décision
Compétences relationnelles

Pilotage technique
Concepts et architectures du système d'information et de communication
Méthodes, outils, normes et procédures de la qualité (connaissance approfondie)
Méthode de compétitivité, organisée et créative, visant à la satisfaction du besoin de l'utilisateur, par une démarche spécifique de conception, à la fois fonctionnelle, économique et pluridisciplinaire
Méthode d'analyse des risques (connaissance approfondie)
Sécurité des systèmes d'information et de communication
Anglais technique

Métiers

Management

Chef(fe) de projet ou expert(e) en infrastructures / architecte des systèmes d'information

Management

Chef(fe) de projet ou expert(e) en ingénierie logicielle

CONT
Acquisition de contenus

Missions

Avec **ANIM**:
Collecter les besoins des utilisateurs de la plateforme
Faire une veille sur les éditeurs scientifiques et intégrateurs de ressources textuelles et examiner les propositions spontanées
Analyser les conditions légales et/ou financières concernant ces ressources textuelles
Evaluer la capacité/les coûts par rapport à la valeur ajoutée du développement de fonctionnalités et/ou de connexions à des infrastructures/services externes
Faire de la veille sur les ressources, standards, normes et formats
Organiser les catalogues de ressources et corpus
Accompagner l'intégration de ces ressources dans la plateforme et/ou développer des connecteurs techniques vers d'autres plateformes/services
Assurer les échanges avec le format pivot de la plateforme
Assurer la disponibilité des métadonnées et de la documentation
Identifier les besoins par rapport à la visualisation des corpus
Avec **GOUV**, proposer et valider choix ponctuels et orientations

Activités

Extraire des notices bibliographiques des bases des bibliothèques numériques
Nettoyer les données
Mettre en conformité les formats des documents entre eux et avec le format cible

Compétences

Connaissances en API
Connaissances en formats de documents (métadonnées, xml, pdf,...), normes
Connaissances en conversion et reformatage de documents, langages de traitement et de transformation des données
Connaissances en curation de données
Connaissances en techniques de préprocessing
Connaissance de l'environnement des bibliothèques (usages, logiciels et matériels, format d'échanges des données...)
Utilisation des applications et logiciels documentaires
Connaissances des outils de gestion de bibliothèque numérique

Métiers

IST et documentation

Ingénieur (e) IST spécialiste en GED, ingénierie documentaire

CONT
Acquisition de contenus

Missions

Avec **ANIM**:
Collecter les besoins des utilisateurs de la plateforme
Faire une veille sur les éditeurs scientifiques et intégrateurs de ressources textuelles et examiner les propositions spontanées
Analyser les conditions légales et/ou financières concernant ces ressources textuelles
Evaluer la capacité/les coûts par rapport à la valeur ajoutée du développement de fonctionnalités et/ou de connexions à des infrastructures/services externes
Faire de la veille sur les ressources, standards, normes et formats
Organiser les catalogues de ressources et corpus
Accompagner l'intégration de ces ressources dans la plateforme et/ou développer des connecteurs techniques vers d'autres plateformes/services
Assurer les échanges avec le format pivot de la plateforme
Assurer la disponibilité des métadonnées et de la documentation
Identifier les besoins par rapport à la visualisation des corpus
Avec **GOUV**, proposer et valider choix ponctuels et orientations

Activités

Produire des corpus bibliographiques
Analyse du besoin
Spécification des critères définissant le corpus
Réalisation des requêtes
Extraction du corpus
Post traitement des résultats pour optimiser l'exploitation du corpus en TDM
Exploration du corpus par des métadonnées et documentation de sa constitution et de son contenu pour livraison et diffusion

Compétences

Connaissance approfondie des processus, systèmes et formats liés à la publication scientifique
Connaissance approfondie des formats normalisés de description et de structuration des documents
Connaissance des terminologies pour faciliter les requêtes
Connaissance des entrepôts, bases de données, réservoirs de notices bibliographiques etc. et des plateformes y donnant accès
Connaissance approfondie des langages de requête des bases de données interrogées
Connaissance des outils de statistiques (Excel, R)
Connaissance approfondie des outils d'extraction de corpus
Evaluation des corpus extraits par rapport à des compétences scientifiques de spécialité
Reconnaissance du bruit, silence et capacité à réadapter les requêtes pour améliorer les résultats
Capacité à choisir l'outil d'extraction le plus adapté au besoin
Capacité à être à l'interface de plusieurs domaines d'expertise (domaines scientifiques, applications TDM, systèmes informatiques)

Métiers

IST et documentation

Ingénieur (e) IST spécialiste de la gestion des bibliothèques numériques

CONT
Acquisition de contenus

Missions

Avec **ANIM**:
Collecter les besoins des utilisateurs de la plateforme
Faire une veille sur les éditeurs scientifiques et intégrateurs de ressources textuelles et examiner les propositions spontanées
Analyser les conditions légales et/ou financières concernant ces ressources textuelles
Evaluer la capacité/les coûts par rapport à la valeur ajoutée du développement de fonctionnalités et/ou de connexions à des infrastructures/services externes
Faire de la veille sur les ressources, standards, normes et formats
Organiser les catalogues de ressources et corpus
Accompagner l'intégration de ces ressources dans la plateforme et/ou développer des connecteurs techniques vers d'autres plateformes/services
Assurer les échanges avec le format pivot de la plateforme
Assurer la disponibilité des métadonnées et de la documentation
Identifier les besoins par rapport à la visualisation des corpus
Avec **GOUV**, proposer et valider choix ponctuels et orientations

Activités

Produire des ressources terminologiques, ontologies
Gérer et maintenir ces ressources en collaboration avec des partenaires/experts extérieurs
Garantir la conformité et la qualité des ressources suivant les standards/bonnes pratiques habituels
Valoriser et mutualiser ces ressources dans la communauté scientifique, les réseaux métiers
Contribuer à des travaux en TAL

Compétences

Connaissance des outils informatiques et techniques servant à la constitution et gestion des ressources terminologiques /ontologies
Connaissance des langues de traitement et de transformation des données
Connaissance des normes et standards en terminologie et dans le domaine du web sémantique (891_Référens)
Connaissance des applications des ressources terminologiques/ontologies
Connaissances de base en TAL
Bonne culture générale et scientifique (spécificité de domaine)
Maîtrise de l'anglais (414_Référens)
Aptitude au travail en équipe (186_Référens)
Sens de l'innovation (122_Référens)
Capacité d'adaptation (115_Référens)

Métiers

IST et documentation

Ingénieur (e) IST spécialiste de la gestion des ressources sémantiques

TDM
Intégration d'outils

Missions

Identifier les outils utiles et utilisables (veille sur le TDM et écoute des besoins) en collaboration avec **APPS**
Organiser le catalogue de composants (création/révision des catégories, curation des métadonnées par exemple)
Evaluer les outils à intégrer en fonction de critères prédéterminés (performance, intégrabilité-interopérabilité, licence)
Evaluer la capacité, les coûts nécessaires pour intégrer les outils par rapport à la valeur ajoutée (accompagnement du fournisseur)
Valider avec l'équipe de pilotage les choix ponctuels et plus généralement, les orientations avec **GOUV**
Accompagner les fournisseurs de composants / applications (conteneurisation, déploiement, description=métadonnées, maintenance…)
Accompagner avec **ANIM** la production de tutoriels autour de l'utilisation des composants
Mettre en place les tests et évaluer les fonctions offertes par un outil une fois intégré à la plateforme
S'assurer que la documentation est présente, suffisante, adaptée pour chaque fonction d'un outil (documentation technique et documentation utilisateur)
Recueillir, gérer et corriger les bugs et prendre en compte les demandes d'évolution des outils (venant notamment de APPS et des utilisateurs)
Identifier et publier les lacunes de l'infrastructure vis à vis des besoins, en se basant sur une analyse des demandes

Activités

Veiller au respect des normes, formats adéquats
veiller au respect des procédures, préconisations d'architecture et de qualité
Déterminer les tests à mettre en œuvre et évaluer la faisabilité technique d'une intégration
Assurer un accompagnement technologique des chercheurs et des fournisseurs de technologies désireux d'intégrer leurs outils à la plateforme
Exercer un support à l'intégration de nouveaux composants
Collecter et analyser les demandes d'évolutions

Compétences

Maîtrise des méthodes et outils de développement informatique
Maitrise des techniques d'évaluation des coûts
Audit des applications
Connaissance des méthodes et outils de déploiement
Connaissance des procédures, préconisations d'architecture, d'urbanisme et de qualité
Assistance à formation et dissémination
Connaissance des méthodes et tests d'évaluation de solutions logicielles
Connaissance des architectures techniques

Métiers

Ingénierie logicielle

Ingénieur(e) généraliste (développeur)

Administration de données

Administrateur (trice) des systèmes d'information

TDM
Intégration d'outils

Missions

Identifier les outils utiles et utilisables (veille sur le TDM et écoute des besoins) en collaboration avec **APPS**
Organiser le catalogue de composants (création/révision des catégories, curation des métadonnées par exemple)
Evaluer les outils à intégrer en fonction de critères prédéterminés (performance, intégrabilité-interopérabilité, licence)
Evaluer la capacité, les coûts nécessaires pour intégrer les outils par rapport à la valeur ajoutée (accompagnement du fournisseur)
Valider avec l'équipe de pilotage les choix ponctuels et plus généralement, les orientations avec **GOUV**
Accompagner les fournisseurs de composants / applications (conteneurisation, déploiement, description=métadonnées, maintenance…)
Accompagner avec **ANIM** la production de tutoriels autour de l'utilisation des composants
Mettre en place les tests et évaluer les fonctions offertes par un outil une fois intégré à la plateforme
S'assurer que la documentation est présente, suffisante, adaptée pour chaque fonction d'un outil (documentation technique et documentation utilisateur)
Recueillir, gérer et corriger les bugs et prendre en compte les demandes d'évolution des outils (venant notamment de **APPS** et des utilisateurs)
Identifier et publier les lacunes de l'infrastructure vis à vis des besoins, en se basant sur une analyse des demandes

Activités

Animer le développement de services de fouille de textes pour les communautés scientifiques
Définir l'évolution de l'architecture logicielle et/ou matérielle de la plateforme et ses interfaces avec les plateformes de ressources (textuelles et terminologiques en particulier)
Assurer un rôle de conseil et d'expertise ainsi qu'une documentation de la plateforme
Organiser la documentation

Compétences

Connaissances des méthodes, techniques, normes et standards TAL
Connaissance des outils et logiciels utilisés dans la fouille de textes
Connaissance des techniques d'apprentissage automatique supervisées et non supervisées
Veille technologique
Assistance à formation et dissémination

Métiers

Ingénierie logicielle

Ingénieur(e) en ingénierie logicielle spécialiste en fouille de textes

APPS
Services
Applications
Workflows

Missions

Accompagner les utilisateurs de la plateforme qui composent et déploient des applications sous la forme de workflows TDM par une expertise technique, méthodologique, aide à la documentation, dans les domaines d'application de la fouille de textes
Organiser le catalogue d'applications (création/révision des catégories, curation des métadonnées, par exemple)
Identifier auprès des utilisateurs et transmettre à l'équipe de **BASE** les besoins d'évolution des fonctions de la plateforme, dont l'outil de composition de workflows (notamment)
Identifier les besoins concernant le cycle de vie des applications et transmettre à **BASE** (besoins de reproductibilité des résultats, accès à l'historique des expérimentations)
Dans le cadre d'un service à la carte :

- analyser le besoin pour une nouvelle application et évaluer l'opportunité et la faisabilité
- identifier et collecter les ressources numériques, permettant de définir un/des corpus qui seront ensuite décrits et exploités par **CONT**
- concevoir des workflows TDM (dans la mesure où les interfaces pour le faire sont adaptées à l'utilisateur, sinon le workflow devra être développé par l'équipe **TDM**)

Activités

Identifier les besoins et la problématique des utilisateurs de la fouille de textes
Définir une modélisation statistique permettant de répondre à cette problématique
Sourcer et rassembler les données/corpus nécessaires et pertinents pour l'analyse
Extraire des connaissances à partir d'un volume de données
Produire des algorithmes sur les données pour anticiper leur comportement
Catégoriser les données
Gérer les données produites, collectées et stockées pour les analyser, les exploiter et les transformer afin de créer de la valeur ajoutée, en recoupant les données nouvelles avec celles existantes
Animer le développement de services de fouille de textes pour les communautés scientifiques

- s'assurer que les composants et les ressources numériques et sémantiques nécessaires sont disponibles et adaptées (si besoin demander leur ajout à **TDM** et **CONT**)
- garantir la compatibilité des ressources en faisant appel à **TDM** et **CONT**
- composer le(s) workflow(s) TDM avec les outils mis à disposition
- documenter le workflow et mettre en place les éventuels supports de formation
- si besoin, adapter les ressources et paramétrer les composants
- tester et évaluer les applications en condition réelle avec les données de la plateforme
- documenter la nouvelle application
- livrer la nouvelle application et si possible la publier sur la plateforme
- Assurer le suivi de cette application à long terme

Compétences

Connaissances en *text mining*
Connaissances en *machine learning*
Connaissances en TAL
Connaissances en statistiques
Expertise en algorithmie et gestion des bases de données
Connaissances en mathématiques appliquées
Maîtrise de la conduite de projet
Capacité à travailler en équipe et en réseau
Capacité à mener des expérimentations
Capacité d'analyse et de synthèse
Travail en équipe et en réseau
Connaissance de l'environnement de la recherche

Métiers

« Hybride »

Data scientist

BASE
Maintenance
Evolution
fonctions

Missions

Veiller sur les technologies utilisées sur des plateformes comparables ou émergentes
Intégrer de nouvelles fonctionnalités
Gérer l'intégration des contributions externes sur les technologies open source
Développer les fonctionnalités nécessaires (interfaces, outils, bases de données etc.)
Evaluer la capacité, les coûts nécessaires pour développer les fonctionnalités de la plateforme et les connexions aux infrastructures/services externes (par rapport à la valeur ajoutée)
Proposer et valider avec l'équipe de pilotage les choix ponctuels et plus généralement, les orientations avec **GOUV**
Assurer la coordination voire la synchronisation avec les plateformes associées
Développer des connecteurs techniques (API, systèmes d'authentification…) vers d'autres e-infrastructures
Organiser et faire fonctionner le guichet support en lien avec **TDM**, **CONT** et **APPS**

Activités

Participer à l'administration du système d'information en termes de référentiels, règles, démarches, méthodologies et outils
Mutualiser les bonnes pratiques en matière d'utilisation du système d'information du domaine
Contrôler et planifier de manière efficace les modifications d'applicatifs et/ou de logiciels
Assister la maitrise d'ouvrage dans l'élaboration de cahiers des charges

---

Résoudre ou faire remonter les incidents et optimiser les performances
Créer et gérer les comptes utilisateurs
Gérer les dépendances aux composants et autres plateformes
Lancer l'exécution des tâches d'exploitation et assurer le suivi d'exploitation

---

Animer et coordonner
une équipe

Compétences

Administrer les principaux systèmes d'exploitation (Windows/ Linux etc…)
Ecrire des scripts d'automatisation de tâches
Administrer un système de sauvegarde et restaurer les données au besoin
Faire un diagnostic des pannes et rediriger aux besoins les incidents vers les bons interlocuteurs
Rédiger / Ecrire des procédures et former les collègues/utilisateurs
Administrer un système de supervision
Connaître le fonctionnement général d'un serveur de messagerie
Savoir connecter au réseau les principaux composants du SI (Serveurs, baies de disques)
Avoir des notions sur les principaux systèmes de virtualisation du marché
Administrer des serveurs physiques et virtuels

Connaissance des applications et processus métiers
Connaissance approfondie des systèmes de gestion de bases de données, des langages de requête et des outils de programmation
Connaissance des référentiels de bonnes pratiques, les normes, procédures et règles
Connaissance des architectures techniques et logicielles
Savoir anticiper les évolutions fonctionnelles et techniques et accompagner les changements
Savoir « packager » une application
Rédiger et mettre à jour la documentation fonctionnelle et technique
Savoir travailler en équipe et en réseau

Métiers

Administration de données

Administrateur (trice) des systèmes d'information

Ingénierie de production

Ingénieur(e) gestionnaire d'applications

Management

Chef(fe) de projet ou expert(e) en infrastructures / architecte des systèmes d'information

BASE
Maintenance
Evolution
fonctions

Missions

Veiller sur les technologies utilisées sur des plateformes comparables ou émergentes
Intégrer de nouvelles fonctionnalités
Gérer l'intégration des contributions externes sur les technologies open source
Développer les fonctionnalités nécessaires (interfaces, outils, bases de données etc.)
Evaluer la capacité, les coûts nécessaires pour développer les fonctionnalités de la plateforme et les connexions aux infrastructures/services externes (par rapport à la valeur ajoutée)
Proposer et valider avec l'équipe de pilotage les choix ponctuels et plus généralement, les orientations avec GOUV
Assurer la coordination voire la synchronisation avec les plateformes associées
Développer des connecteurs techniques (API, systèmes d'authentification…) vers d'autres e-infrastructures
Organiser et faire fonctionner le guichet support en lien avec TDM, CONT et APPS

Activités

Anticiper les changements et leurs impacts métiers sur le SI et en assurer la promotion par des actions de conseil et de communication

Vérifier la pertinence et la performance fonctionnelle du système d'information
Proposer des évolutions applicatives (fonctionnelles ou techniques)
Participer à la définition et faire appliquer les accords de niveaux de service
Participer à l'élaboration d'outils de consultation, de contrôle et de gestion (scripts, procédures, requêtes, reporting)
Rédiger la documentation fonctionnelle et technique

Compétences

Connaissance d'au moins un langage de programmation
Connaissance des normes et modèles pour le stockage et l'échange de métadonnées
Connaissance d'une ou plusieurs méthodes d'analyse, de conception et de développement
Savoir formaliser des besoins et les transformer en spécifications techniques
Savoir réutiliser et assembler des composants logiciels
Savoir respecter les procédures et les préconisations d'architecture, d'urbanisme et de qualité
Savoir rendre compte régulièrement à sa hiérarchie

Savoir se positionner en développeur backend ou frontend

Métiers

Ingénierie logicielle

Ingénieur(e) généraliste (développeur)

Ingénierie logicielle

Ingénieur(e) généraliste (développeur)

INST

Moyens de stockage et de calcul

Missions

Dimensionner, obtenir et faire évoluer le matériel informatique nécessaire au bon fonctionnement de la plateforme en fonction de sa charge (nombre d'utilisateurs, pics d'utilisation, coût informatique des traitements) et de la taille des données stockées.
Collecter les besoins, identifier les limites du système, notamment auprès de **BASE** et de **APPS**
Gérer le parc informatique, les systèmes d'exploitation et logiciels des couches basses, les connexions
Déployer, configurer et monitorer les logiciels nécessaires au déploiement des services
Assurer la continuité de service 24/7 et le dépannage des installations
Assurer la sécurité des données en accord avec la réglementation et les politiques arrêtées par **GOUV**
Assurer la pérennité des données selon une politique déterminée avec les autres équipes

Activités

Planifier, installer, automatiser, superviser et améliorer les moyens de production
Sécuriser la production
Appliquer des normes et standards de sécurité
Gérer les évolutions et la maintenance des matériels, logiciels et systèmes

Compétences

Connaissance de l'architecture et de l'environnement technique du système d'information
Connaissance des méthodes de mise en production
Connaissances des normes d'exploitation
Connaissances en métrologie et performance
Connaissance approfondie en sécurité des systèmes d'information et de communication
Capacités en diagnostic et résolution de problèmes
Connaissance approfondie en techniques de virtualisation (technologie Cloud)
Connaissance des méthodes, outils, normes et procédures de la qualité
Connaissances en langages de programmation
Maîtrise de l'anglais technique
Veille technologique
Capacité d'évaluation d'une solution informatique
Savoir rédiger de la documentation fonctionnelle et technique
Savoir travailler en équipe
Etre réactif

Métiers

Administration de données

Administrateur (trice) des systèmes et réseaux

ANIM

Expertise
Accompagnement
Animation

Missions

Organiser et contribuer à l'animation de groupes de compétences autour de la plateforme et au sein des diverses communautés d'acteurs (avec TDM, CONT et APPS)
Promouvoir les services de l'e-infrastructure auprès des différents acteurs

- avec TDM organiser des événements pour les communautés TDM, IA, etc. dans le cadre de conférences scientifiques notamment
- avec CONT, auprès des éditeurs et des bibliothèques numériques (inscription dans des réseaux, participation à des événements…)

Communiquer sur les apports de l'e-infrastructure et du TDM en général, notamment en promouvant les "success stories", applications phare...
Faciliter et harmoniser l'accompagnement des utilisateurs et la création de documentation par BASE, TDM, CONT et APPS
Mettre en place, alimenter et animer le site/portail de l'e-infrastructure, qui constitue le point d'accès à la plateforme technique, aux supports de formation, aux diverses informations relatives à l'actualité du domaine, etc.
Créer et alimenter les comptes de l'e-infrastructure sur les réseaux sociaux
Contribuer à l'organisation et à l'animation des événements internes avec GOUV pour des plénières et les autres missions sur des sujets plus spécifiques
Contribuer à l'organisation et à l'animation des formations autour de la plateforme, notamment avec TDM, CONT et APPS
Assurer un accompagnement juridique

Activités

Concevoir la stratégie et le plan de communication autour de la plateforme
Cibler les réseaux sociaux à investir et définir les contenus suivant les acteurs visés et permettant de mettre développer la visibilité de la plateforme
Mettre en place des indicateurs d'efficacité des actions de communication et de satisfaction

Etre en mesure de transmettre le savoir nécessaire à l'utilisation d'une plateforme de façon accessible et compréhensible au public cible
Eveiller la curiosité de ce public pour stimuler son désir de comprendre

Compétences

Communication
Animation de collectif
Initiation et conduite de partenariats (Identifier les partenaires potentiels les plus intéressants pour l'organisation, conclure le partenariat, identifier et mettre en œuvre les moyens de l'entretenir (rencontres, information) ; savoir inscrire son activité dans un réseau dépassant le cadre de l'établissement)
Connaissances sur le TDM et ses usages
Connaissance de l'écosystème de recherche et des ses acteurs
Bonne connaissance des techniques de communication et de leurs supports (rédaction web, réseaux sociaux) ainsi que des outils d'analyse
Bonne pratique rédactionnelle
Avoir le sens de l'animation et une aisance relationnelle
Etre force de proposition
Etre curieux

Connaissances sur la fouille de textes et ses usages
Connaissance de l'écosystème de recherche et des ses acteurs
Maîtrise des techniques de communication orale et écrite
Avoir le sens de l'animation et une aisance relationnelle
Capacité de travail en équipe
Adaptabilité
Autonomie
Disponibilité

Métiers

Communication

Community manager

Communication

Médiateur(trice) scientifique

ANIM

Expertise
Accompagnement
Animation

Missions

Organiser et contribuer à l'animation de groupes de compétences autour de la plateforme et au sein des diverses communautés d'acteurs (avec TDM, CONT et APPS)
Promouvoir les services de l'e-infrastructure auprès des différents acteurs

- avec TDM organiser des événements pour les communautés TDM, IA, etc. dans le cadre de conférences scientifiques notamment
- avec CONT, auprès des éditeurs et des bibliothèques numériques (inscription dans des réseaux, participation à des événements…)

Communiquer sur les apports de l'e-infrastructure et du TDM en général, notamment en promouvant les "success stories", applications phare...
Faciliter et harmoniser l'accompagnement des utilisateurs et la création de documentation par BASE, TDM, CONT et APPS
Mettre en place, alimenter et animer le site/portail de l'e-infrastructure, qui constitue le point d'accès à la plateforme technique, aux supports de formation, aux diverses informations relatives à l'actualité du domaine, etc.
Créer et alimenter les comptes de l'e-infrastructure sur les réseaux sociaux
Contribuer à l'organisation et à l'animation des événements internes avec GOUV pour des plénières et les autres missions sur des sujets plus spécifiques
Contribuer à l'organisation et à l'animation des formations autour de la plateforme, notamment avec TDM, CONT et APPS
Assurer un accompagnement juridique

Activités

Concevoir, réaliser et animer des supports de formation divers sur l'utilisation de la plateforme et adaptés au public cible
Assurer l'assistance aux usagers de ces supports
Evaluer les résultats et les effets des actions de formation

Participer au sein d'une équipe de formation dédiée à l'élaboration d'un ensemble coordonné de ressources d'autoformation multimédia sur la fouille de textes, à destination des utilisateurs de l'e-infrastructure. Ceci dans le but de permettre à l'apprenant de trouver le produit d'apprentissage qui correspond le mieux à ses besoins via une interface de centralisation, lui permettant ainsi de construire son parcours de formation.

Compétences

Connaissances sur le TDM et ses usages
Connaissance de l'écosystème de recherche et des ses acteurs
Maîtrise des techniques de communication orale et écrite
Capacité à analyser un besoin et à y répondre en formalisant un support pédagogique adapté au public ciblé (contenu, support)
Avoir le sens de l'animation et une aisance relationnelle
Disponibilité

Bonne maîtrise des plateformes LMS (Learning Management System)
Bonne maîtrise des outils de conception de tutoriels multimédia
Bonne maîtrise de la scénarisation pédagogique en formation à distance
Compétences en séquençage de vidéos
Connaissance des techniques et outils de médiatisation des contenus pédagogiques
Connaissance des démarches d'apprentissage, d'enseignement et d'évaluation
Bonnes capacités rédactionnelles et aisance à l'oral
Aptitude au travail en équipe et en réseau

Métiers

Formation

Formateur (trice)

Formation

Ingénieur (e) pédagogique multimédia

ANIM
Expertise
Accompagnement
Animation

Missions

Organiser et contribuer à l'animation de groupes de compétences autour de la plateforme et au sein des diverses communautés d'acteurs (avec **TDM**, **CONT** et **APPS**)
Promouvoir les services de l'e-infrastructure auprès des différents acteurs

- avec **TDM** organiser des événements pour les communautés TDM, IA, etc. dans le cadre de conférences scientifiques notamment
- avec **CONT**, auprès des éditeurs et des bibliothèques numériques (inscription dans des réseaux, participation à des événements…)

Communiquer sur les apports de l'e-infrastructure et du TDM en général, notamment en promouvant les "success stories", applications phare...
Faciliter et harmoniser l'accompagnement des utilisateurs et la création de documentation par **BASE**, **TDM**, **CONT** et **APPS**
Mettre en place, alimenter et animer le site/portail de l'e-infrastructure, qui constitue le point d'accès à la plateforme technique, aux supports de formation, aux diverses informations relatives à l'actualité du domaine, etc.
Créer et alimenter les comptes de l'e-infrastructure sur les réseaux sociaux
Contribuer à l'organisation et à l'animation des événements internes avec **GOUV** pour des plénières et les autres missions sur des sujets plus spécifiques
Contribuer à l'organisation et à l'animation des formations autour de la plateforme, notamment avec **TDM**, **CONT** et **APPS**
Assurer un accompagnement juridique

Activités

Elaborer les préconisations juridiques d'un encadrement sécurisé de la fouille de textes et suivre leurs évolutions

Compétences

Connaissance des textes législatifs et règlementaires du droit des systèmes d'information et de la communication, du droit d'auteur, du droit de la propriété littéraire et artistique, du droit de la propriété intellectuelle, du copyright
Connaissance du statut juridique et technique des données
Capacités d'analyse et de rédaction de recommandations juridiques
Veille juridique

Métiers

Juridique

Juriste

Annexe 2

VisaTM étude

Recensement de formations en fouille de textes

Travail collaboratif réalisé à l'INIST-CNRS

PRESENTIEL | FORMATION INITIALE | LONGUE | DIPLOMANTE

Université Paris-Saclay

Cette université présente une offre très large de formations en Data sciences puisqu'on ne dénombre pas moins de 45 formations sur cette thématique. Dans les formations initiales nous pouvons retenir :

Master 2 Traitement de l'Information et Exploitation des Données (TRIED)
https://www.universite-paris-saclay.fr/fr/formation/master/m2-traitement-de-linformation-et-exploitation-des-donnees-tried#presentation-m2

Master 2 Innovation, Marché et Science des Données (IMSD)
https://www.universite-paris-saclay.fr/fr/education/master/m2-innovation-marche-et-science-des-donnees-imsd#presentation-m2

Master 2 Data & Knowledge (D&K)
https://www.universite-paris-saclay.fr/fr/education/master/m2-data-knowledge-d-k#presentation-m2

Master 2 Apprentissage, Information et Contenu (AIC)
https://www.universite-paris-saclay.fr/fr/education/master/m2-apprentissage-information-et-contenu-machine-learning-information-and-content#presentation-m2

Parcours Data Science du diplôme d'Ingénieur ECS
https://www.centralesupelec.fr/fr/cursus-ingenieur-supelec

IODAA (de l'Information à la Décision par l'Analyse et l'Apprentissage)
http://www2.agroparistech.fr/IODAA-De-l-InfOrmation-a-la.html

Big data, biologie et santé
https://www.ensta-paristech.fr/fr/devenir-ingenieur/formation-1ere-annee/enseignements-thematiques

M2 Data Science
https://www.universite-paris-saclay.fr/fr/formation/master/m2-data-science-eit-digital#presentation-m2

M2 Mathématiques / Vision / Apprentissage (MVA)
https://www.universite-paris-saclay.fr/fr/formation/master/m2-mathematiques-vision-apprentissage#presentation-m2

M2 Gestion de données dans un monde numérique (DataScale)
https://www.universite-paris-saclay.fr/fr/education/master/m2-gestion-de-donnees-dans-un-monde-numerique-data-management-in-a-digital-world#presentation-m2

MSc in data sciences & business analytics
https://www.centralesupelec.fr/fr/bienvenue-la-nouvelle-promotion-de-notre-msc-datascience-and-business-analytics

Parcours Data Science du diplôme d'Ingénieur ENSIIE
http://www.math-evry.cnrs.fr/departement/doku.php?id=formation:master:m2ds

Filière Sciences des Données du diplôme d'Ingénieur Télécom ParisTech
https://datascience-x-master-paris-saclay.fr/

Voie Data Science du diplôme d'ingénieur de l'ENSAE ParisTech
https://datascience-x-master-paris-saclay.fr/le-master/presentation/

Université de Lyon 1 - ISFA
https://isfa.univ-lyon1.fr/formation/econometrie-statistiques/econometrie-et-statistiques-784953.kjsp

- Diplôme : Master en Econométrie et statistiques (2 ans)
- Parcours : Sécurité et risque informatique ou Décision Risk-Management

Ces deux parcours incluent des unités sur les statistiques et l'analyse de données, le data mining, le text mining et la visualisation.

Université de Lyon 2
https://dis.univ-lyon2.fr/fr/nos-formations/formations-initiales/m2-sise/specialite-m2-sise-statistique-et-informatique-pour-la-science-des-donnees-parcours-statistique-et-informatique--556814.kjsp

Master 2 SISE (Statistique et Informatique pour la Science des donnéEs) sur 1 an. Parcours Statistique et Informatique.
Formation avancée d'un an sur la data science, avec une forte composante data mining, machine learning et statistique d'une part, informatique et technologies big data d'autre part.

Beaucoup d'autres universités en France introduisent progressivement des cursus autour de ces thématiques, tout comme les écoles d'ingénieurs, les grandes écoles ou les écoles de management.

PRESENTIEL | FORMATION CONTINUE | COURTE | CERTIFIANTE

ENSAE-ENSAI - Formation continue - CEPE

https://www.lecepe.fr/certificats/data-scientist/

Certificat de Data Scientist
Ce certificat de 15 ou 18 jours a pour ambition de permettre, à toute personne souhaitant valoriser la manne de données mise actuellement à sa disposition, d'accroître son champ de

connaissances, d'acquérir un véritable savoir-faire opérationnel et une très bonne maîtrise des techniques d'analyse de données et des outils informatiques nécessaires.

ib (Groupe Cegos)

Formations Les métiers du Big Data

https://www.ib-formation.fr/catalogue/nbs-listing/catref/universib-formations-informatiques-big-data

Formations de 10 à 20 jours.
Cursus Data Scientist
Cursus Data Analyst
Cursus Data Steward

Big Data Foundation Certifiant – Les fondamentaux

https://www.ib-formation.fr/catalogue/nbs-details/catref/universib-formations-informatiques-big-data-les-fondamentaux/ref/mg510/big-data-foundation-certifiant

Formation certifiante de 3 jours permettant de préparer et passer l'examen de certification "Big Data Foundation" de l'EXIN

TELECOM Evolution
http://www.telecom-evolution.fr/fr/formations-certifiantes

Certificat d'Etudes Spécialisées: Data Scientist Data science - Analyse et gestion de grandes masses de données

https://www.telecom-evolution.fr/fr/formations-certifiantes/data-scientist-data-science

Certificat d'Etudes Spécialisées: Intelligence Artificielle

https://www.telecom-evolution.fr/fr/formations-certifiantes/intelligence-artificielle

PRESENTIEL | FORMATION CONTINUE | COURTE

ADBS Paris

https://www.adbs.fr/formations

Formations de 2 à 3 jours.
Analyse de contenu : L'apport de la fouille de textes
Datavisualisation / cartographie pour gérer ses informations
Comprendre les enjeux du web sémantique

CNRS Formation Entreprises
https://cnrsformation.cnrs.fr/

Formations de 1 à 4 jours.

Analyse statistique – langage R :
Apprentissage statistique : théorie et application
Analyse statistique des réseaux
Linkage : analyse conjointe de réseaux et de corpus
Introduction à l'analyse causale
Statistiques pour le Lean 6 Sigma et la production
Langage R : introduction
Analyse de réseaux en sciences humaines et sociales, transport et logistique
Le crowdsourcing : plateformes, applications, algorithmes et perspectives de recherche

Bioinformatique :
Analyses NGS avec R
Analyse avancée de séquences
Bioinformatique pour le traitement de données de séquençage (NGS)
ChIP-seq, RNA-seq et Hi-C : traitement, analyse et visualisation de données
Analyses bioinformatiques avec Python
Scripts en Python pour la bioinformatique et environnement Linux

Traitement et fouille de données, cartographie :
L'analyse sémantique fine des textes et débats : introduction aux traitements automatiques et au modèle cartographique des Atlas sémantiques
Gargantext pour l'analyse exploratoire de grands corpus textuels
Python 3, des fondamentaux aux interfaces graphiques pour l'instrumentation : les communications, la représentation et visualisation de données
Python data analysis for GATE simulations
GATE training on medical imaging (PET, SPECT, CT), dosimetry and radiation therapy - Beginner level
LiDAR : initiation au traitement des données et à l'interprétation archéologique
SIG et archéologie : utilisation du logiciel libre QGIS pour le traitement de données archéologiques spatialisées
Cartographie et SIG en sciences humaines et sociales
Initiation aux SIG en écologie : de la collecte au traitement de données géographiques
Initiation aux SIG et prise en main du logiciel QGIS
Les systèmes d'information géographique (SIG) au service des sciences de l'homme et de l'environnement
Modélisation des réseaux écologiques : initiation au logiciel Graphab
Modélisation des réseaux écologiques : utilisation avancée du logiciel Graphab
Traitement d'images sous ImageJ et les nouveaux logiciels FIJI et ICY : bases conceptuelles et pratiques
ImageJ / FIJI : traitement et analyse d'images de microscopie
Automatisation du traitement d'images : du langage macro (ImageJ, FIJI) à l'intelligence artificielle (Weka, Ilastik, TensorFlow, Keras) et analyse des données (R)
Traitement d'images en Python avec scikit-image

Machine learning et deep learning
Introduction au machine learning et au deep learning, mise en œuvre en Python
Fondements du machine learning et du deep learning
Machine learning sous Python
Deep learning pour le traitement automatique des langues
Apprentissage automatique : introduction aux techniques récentes et mise en pratique sous Python
Apprentissage automatique pour la vision par ordinateur
Apprentissage automatique (machine learning) à base de noyaux

Intelligence artificielle
Intelligence artificielle : état de l'art et applications
Intelligence artificielle et réseaux de neurones artificiels : concepts et applications

Coheris

Formations Business & data intelligence : Big Data, Datamining et Marketing
https://www.coheris.com/formations/customer-data-intelligence/
Le cursus de formation est conçu pour assurer un apprentissage accessible à tout profil de stagiaire. Il ne demande pas de prérequis particulier. L'objectif est d'apporter des connaissances en termes de Big Data & Datamining applicables en marketing, et aider à bâtir des stratégies solides en Business Intelligence. Grâce à des formations axées sur la pédagogie et des exercices pratiques, on apprend comment stocker les données Big Data ou encore quels sont les outils et méthodes de Data mining utilisés en Connaissance client. On apprend également comment concevoir et mettre en œuvre des campagnes marketing performantes en segmentant ses clients par valeur ou par engagement.

Formations de 1 à 2 jours.
Introduction au Big Data & à l'intelligence artificielle
Fondamentaux du Data Mining
La Dataviz pour faire parler vos données

Coheris Analytics SPAD : La Formation Data-Mining
https://www.coheris.com/formations/analytics/data-mining/
La formation data-mining vise à aider les entreprises et sociétés d'études qui souhaitent acquérir de nouvelles connaissances stratégiques et opérationnelles à partir de leurs bases de données.
L'outil étudié lors de la formation, Coheris Analytics SPAD, est une référence en matière d'analyse de données et data-mining, bénéficiant de plus de 35 ans de Recherche et Développement. C'est un logiciel largement utilisé tant dans le monde professionnel que dans le domaine de l'enseignement et de la recherche. Il constitue une référence sur le marché français en analyse de données quantitatives, qualitatives, textuelles ou non structurées.
Le cycle d'apprentissage regroupe 9 sessions qui couvrent tous les champs du Data Mining.
Toutes ces sessions associent présentation détaillée des méthodes et mise en pratique afin de tirer pleinement partie de Coheris SPAD. Des rappels théoriques sont fournis afin de maximiser la maîtrise des méthodes présentées.

Formations de 1 à 2 jours.
Statistiques descriptives et data management
Estimation et tests statistiques
Prédiction d'une quantité
Séries chronologiques
Les méthodes statistiques du data mining
Les méthodes machine learning du data mining niveau 2
Analyses exploratoires et typologies
Text mining

ENSAE-ENSAI - formation continue - CEPE
https://www.lecepe.fr/formations/data-science/

Formations Data science

Gamme de formations (de 1 à 3 jours) pour comprendre les enjeux du Big Data et s'initier aux techniques de ce domaine en plein développement.

Enjeux :
Panorama du Big Data
Enjeux juridiques du Big Data
Panorama des méthodes de Data mining
Analyse des réseaux
Les données structurées sur le web
Méthodes avancées de Data Mining

Méthodes et outils :
Mettre en œuvre et utiliser les outils informatiques des Big Data
R pour la data science
Spark pour la data science
Modélisation et initiation au machine learning
Machine learning
Les fondamentaux du Deep learning
Visualisation des données
Visualisation et cartographie pour le web
Statistique textuelle pour le Text Mining
Réduction de dimension et classification non supervisée (Clustering)
Techniques de scoring

GFII Paris
https://www.gfii.fr/fr/formation

Formations de 2 à 4 jours.

Gestion de données et traitement intelligents
Introduction au web sémantique et au web de données

Web sémantique et web de données : mise en œuvre d'un projet
Text-mining et classification automatisée pour l'indexation, la création de base de données et l'analyse de sentiments

TELECOM Evolution
https://www.telecom-evolution.fr/fr/formations-courtes
Intelligence Artificielle : Attentes économiques et défis scientifiques
Big data : enjeux stratégiques et défis technologiques
Data Science avec Python
Data science : introduction au machine learning
Machine Learning avancé
Big Data : panorama des infrastructures et architectures distribuées
Data Science dans le Cloud : Big Data, statistiques et Machine Learning
Visualisation d'information (InfoVis)
Extraction d'informations du Web
Text Mining
Opinion mining : e-reputation et recommandation

ib (Groupe Cegos)
https://www.ib-formation.fr/catalogue/nbs-listing/catref/universib-formations-informatiques-big-data

Formations de 2 à 4 jours.

Les fondamentaux
Big Data – L'essentiel (séminaire)
Big Data – Enjeux et perspectives
Big Data – Les fondamentaux de l'analyse de données
Big Data – Conception et pilotage de projets

Analyse, restitution et visualisation de données
Les fondamentaux des statistiques appliquées
Les fondamentaux de l'analyse statistique avec R
Analyse statistique avancée avec R
Big Data - Analyse, Data Visualisation et introduction au Data StoryTelling pour la restitution de données
Big Data – Mise en œuvre pratique d'une solution complète d'analyse des données
Analyse et visualisation de données avec Power BI

Intelligence artificielle, Data Science
Séminaire : Machine Learning – La synthèse
Séminaire : Data Science – les fondamentaux
Séminaire : Intelligence Artificielle (IA – La synthèse
Data Science – Mise en œuvre du Machine Learning
Data Science – Mise en œuvre du Deep Learning

Serda Formation
https://www.formation-serda.com/gestion-des-data

Formations de 1 à 2 jours

Gestion des data
Web sémantique et Linked open data
Text mining et analyse de contenus
Intelligence augmentée et automatisation des processus documentaires
Intelligence artificielle, machine learning : découverte et applications
Big data : l'état de l'art
Bonnes pratiques de la visualisation graphique de données (dataviz)
Mise en récit des données (data storytelling)

STAT4DECISION Conseil et formation Data science
https://www.stat4decision.com/fr/formations/

Formations de 1 à 3 jours

Python
Python pour la data science
Data visualisation avec Python
Analyse textuelle avec Python
Spark avec Python
Deep Learning avec Python

R
Logiciel R pour la data science
Data mining et machine learning avec R
Développement d'applications web avec R shiny
Visualisation avec R
Séries temporelles avec R

Big Data et Open Data
Les fondamentaux du big data
Les fondamentaux du machine learning et du deep learning
Data science – bonnes pratiques et outils
Open Data
Culture générale de la donnée - De l'open data au big data

Logiciels de Data Science
Logiciel KNIME
Logiciel RapidMiner
Dataiku DSS
Langage Julia

Statistique et analyse de données

Statistique et analyse de données
Statistique et analyse de données avec XLSTAT
XLSTAT – XLSTAT-R et programmation avec XLSTAT
Approche PLS (PLS Path Modeling)
Régression PLS

Télécom Evolution
http://www.telecom-evolution.fr/fr/domaines/big-data

Big data
Big data : enjeux stratégiques et défis technologiques
Data Science avec Python
Data science : introduction au machine Learning
Big Data : panorama des infrastructures et architectures distribuées
Data Science dans le Cloud : Big Data, statistiques et Machine Learning
Visualisation d'information (InfoVis)
Extraction d'informations du Web
Text Mining
Opinion mining : e-reputation et recommendation
Concevoir et piloter un projet big data
Intelligence Artificielle : Attentes économiques et défis scientifiques

Réseau des Urfist
https://sygefor.reseau-urfist.fr

Formations de 1 à 2 jours et ressources disponibles (supports de formation)

Urfist Bordeaux

Python :
Premiers pas en programmation avec Python : initiation
Premiers pas en machine learning avec Python : méthodes d'intelligence artificielle pour l'analyse de données
R :
Analyse statistique et création de graphiques avec Stat R : initiation
Quantifier l'incertain et gérer des problèmes d'inférence : initiation aux statistiques bayésiennes
Concevoir des rapports automatisés sous R : de RMarkdown à Bookdown
Faire de la classification (clustering) des variables sous R. Prise en main du package ClustOfVar
Perfectionner sa pratique de Stat R : régression linéaire simple et multiple, analyse de variance
Créer des modèles prédictifs avec R en présence de données manquantes : arbres de décision
Dataviz avancée sous R : bien choisir son package, mettre en œuvre une stratégie de visualisation pour ses données
Collecter des données du web pour créer ses jeux de données : web scraping et APIs avec R
Analyse de corpus, cartographie :
Construire et valoriser un corpus spécialisé à partir des collections pluridisciplinaires Istex

Explorer un corpus documentaire issu d'Istex à l'aide de l'outil de cartographie Cillex
Formation de formateurs à la cartographie de corpus avec Gargantext

Urfist Lyon

Gargantext / EasISTEX : Fouille textuelle avec Gargantext sur les corpus ISTEX
Construire et valoriser un corpus spécialisé à partir des collections pluridisciplinaires Istex
Nettoyer et transformer des données avec Openrefine : des premiers pas aux usages avancés
Introduction à RStudio

Urfist Méditerranée

Ateliers pratiques sur l'usage des ressources Istex

Urfist Paris

Formations
Atelier pratique sur l'usage des ressources Istex
Cartographie et visualisation de données
La visualisation des données avec Gephi

Ressources
Nettoyer et transformer des données avec OpenRefine : des premiers pas aux usages avancés
http://urfist.chartes.psl.eu/ressources/nettoyer-et-transformer-des-donnees-avec-openrefine-des-premiers-pas-aux-usages-avances

Urfist Rennes

Ressources
Journée d'étude Big et Open Data : supports des conférences et vidéos en ligne
https://www.sites.univ-rennes2.fr/urfist/blog/2015/06/journee-detude-big-et-open-data-les-supports-des-conferences-en-ligne

Introduction aux problématiques et aux solutions du Traitement Automatique des Langues (TAL)
https://www.sites.univ-rennes2.fr/urfist/ressources/introduction-aux-problematiques-et-aux-solutions-du-traitement-automatique-des-langues-ta

Analyser les textes avec Iramuteq, outil de statistique lexicale
https://www.sites.univ-rennes2.fr/urfist/ressources/analyser-les-textes-avec-iramuteq-outil-de-statistique-lexicale?destination=ressources%3Ffield_tag_tid%3DAll%26keys%3D%26field_update_date_value%255Bvalue%255D%3D%26page%3D3

Cartographier le web avec Hyphe, visualiser les réseaux avec Gephi
https://www.sites.univ-rennes2.fr/urfist/ressources/cartographier-le-web-avec-hyphe-visualiser-les-reseaux-avec-gephi?destination=ressources%3Ffield_tag_tid%3DAll%26keys%3D%26field_update_date_value%255Bvalue%255D%3D%26page%3D4

Urfist Strasbourg

Ressources

Introduction to R
http://urfist.unistra.fr/uploads/media/slidesRIntro.pdf
L'analyse du discours assistée par ordinateur: la méthode ALCESTE et le logiciel IRAMUTEQ
http://urfist.unistra.fr/uploads/media/diapo_formation_URFIST.V2018.pdf
Introduction à la statistique textuelle
http://urfist.unistra.fr/uploads/media/1StatTextV2.pdf
Traitements de données d'enquêtes et initiation au logiciel SPAD
http://urfist.unistra.fr/uploads/media/cahier_traitements_de_donnees_d_enquetes_-J-Paul_Villette_URFIST-_2_et_7_novembre_2016.pdf

MIXTE (BLENDED LEARNING) | FORMATION CONTINUE | COURTE

ib (Groupe Cegos)

Formations mixtes de 3 à 4 jours.

Big Data - Les fondamentaux de l'analyse des données
https://www.ib-formation.fr/catalogue/nbs-details/catref/universib-formations-informatiques-big-data-les-fondamentaux/ref/bd540/big-data-les-fondamentaux-de-lanalyse-des-donnees
Propose après le présentiel un vidéocast et deux vidéo-tutos.

Big Data - Mise en œuvre pratique d'une solution complète d'analyse des données
https://www.ib-formation.fr/catalogue/nbs-details/catref/universib-formations-informatiques-big-data-les-fondamentaux/ref/bd550/big-data-mise-en-oeuvre-pratique-dune-solution-complete-danalyse-des-donnees
Avant le présentiel : un quiz de consolidation des prérequis
Après le présentiel:

- Un quiz pédagogique pour évaluer vos acquis et approfondir les sujets de votre choix
- Des vidéocasts pour revenir sur les points clés de la formation
- Des vidéos-tutos pour vous accompagner dans l'utilisation des outils du Big Data

A DISTANCE | FORMATION CONTINUE | COURTE

FUN MOOC

MOOC : Recherche reproductible : principes méthodologiques pour une science transparente (48 h sur 6 semaines)
https://www.fun-mooc.fr/courses/course-v1:inria+41016+session02/about

MOOC : Fondamentaux pour le Big data
https://www.fun-mooc.fr/courses/course-v1:MinesTelecom+04006+session10/about

Deep Learning
https://www.fun-mooc.fr/courses/course-v1:CNAM+01031+session02/about

Analyse des données multidimensionnelles
https://www.fun-mooc.fr/courses/course-v1:agrocampusouest+40001S05+session05/about

Données et algorithmes
https://www.fun-mooc.fr/courses/course-v1:u-cergy+156003+session01/about

TELECOM Evolution

MOOC : Fondamentaux pour le Big data (24 h sur 7 semaines)
https://www.telecom-evolution.fr/fr/e-learning/fondamentaux-pour-le-big-data

MOOC ouvert en continu.

Nous constatons également l'arrivée de formations privées sur ces créneaux spécifiques comme **Data University- Institut de sciences des données** https://datauniversity.fr/