

	<b>Appel à Manifestation d'Intérêt</b> Corpus de recherche « Grand débat national »	<b>2019</b>

Proposition	
<b>Titre en français</b>	<b>GENUINE: Gouvernance guidée par les données de terrain et basée sur l'analyse dynamique</b>

Coordinateur du projet			
<b>Prénom</b>	Genoveva	<b>Nom</b>	Vargas-Solar
<b>Email</b>	genoveva.vargas@imag.fr		
<b>Laboratoire</b>	Laboratoire d'Informatique de Grenoble (LIG UMR 5217)		
<b>Etablissement d'appartenance</b>	CNRS		

Autre(s) partenaire(s)	
<b>Laboratoire, établissement d'appartenance</b>	Mirian HALFELD FERRARI ALVES, Professeur, Université d'Orléans, Laboratoire d'Informatique d'Orléans (LIFO), <a href="mailto:mirian@univ-orleans.fr">mirian@univ-orleans.fr</a> Anais LEFEUVRE-HALFTERMEYER, Maître de Conférences, Université d'Orléans, Laboratoire d'Informatique d'Orléans (LIFO), <a href="mailto:anais.halftermeyer@univ-orleans.fr">anais.halftermeyer@univ-orleans.fr</a>
<b>Laboratoire, établissement d'appartenance</b>	Laurent D'ORAZIO, Professeur, Université de Rennes, IRISA , <a href="mailto:laurent.dorazio@irisa.fr">laurent.dorazio@irisa.fr</a>
<b>Laboratoire, établissement d'appartenance</b>	Jean-Yves ANTOINE, Professeur, Université de Tours, Laboratoire d'Informatique Fondamentale et Appliquée de Tours (LIFAT), <a href="mailto:Jean-Yves.Antoine@univ-tours.fr">Jean-Yves.Antoine@univ-tours.fr</a>

## Présentation de la proposition (1,5 page)

La présentation devra décrire :

- l'objet de recherche ou la question scientifique
- l'originalité de l'approche et pertinence de la méthode utilisée
- les moyens et les outils déployés (ou à déployer) dans le dispositif de recherche
- le cas échéant : les résultats préliminaires

Le succès des politiques publiques et des actions gouvernementales dépend de l'adéquation entre l'agenda gouvernementale et les attentes des citoyens pour qu'ils adhèrent et soutiennent des politiques publiques, des réformes et des nouvelles lois. Il s'agit de chercher des bénéfices collectifs à l'échelle nationale tout en gardant un rôle d'influence dans le contexte international pour aborder des problèmes globaux (par ex. le climat, la sécurité). La communication directe avec les citoyens, la compréhension précise, inclusive et opportune de leurs demandes peut donner lieu à un ajustement intelligent des actions gouvernementales et des politiques publiques. On peut ainsi arriver à une gouvernance harmonieuse et vertueuse. Le "Grand Débat" correspond à la première phase vers ce type de gouvernance, qui consiste à collecter de manière massive et large la vision et l'opinion des différentes couches sociales sur des thèmes clés. Dès l'annonce de ce projet, l'opinion publique s'est interrogée sur la façon dont ces déclarations massives seraient traitées et donneraient lieu à des actions gouvernementales inclusives et concrètes. Le partage agile et souple des données collectées sera crucial à la prise de décision critique, à la communication publique et à la coordination multi-institutionnelle.

Ceci implique la mise à disposition des données adéquates à la vitesse de la pensée (*at speed of thought*), pour que les personnes politiques et les personnes en charge de la prise de décisions puissent planifier, répondre et restaurer la confiance lorsque l'atmosphère sociale devient électrique. Ceci implique aussi, de maintenir les données, de créer des vues historiques de trois aspects associés : les demandes (doléances), les actions et l'évaluation des résultats.

L'ambition de cette proposition se résume par la notion de "analyse-dynamique" (*analytics in motion*) c'est-à-dire, le processus de mobiliser les données pour répondre à des attentes en présence des circonstances qui évoluent rapidement, qui sont conflictuelles et qui donnent lieu à des nouvelles demandes jamais considérées auparavant. L'analyse-dynamique peut encourager une pratique gouvernementale centrée sur les données où les actions répondent à des demandes citoyennes concrètes et donnent lieu à des résultats et aux retours d'expérience *feedback*. Les données peuvent être analysées et utilisées pour simuler des scénarios pour montrer aux citoyens si la réponse à leurs demandes peut donner lieu au bien commun.

### Question scientifique

L'objectif général est de définir un processus d'analyse dynamique pour extraire contenu et connaissance à partir des documents générés par le "Grand Débat". Les objectifs spécifiques sont de développer des outils pour: (i) produire des collections de données nettoyées, organisées, intégrées et contextualisées prêtes pour appliquer des processus d'analyse (environnements de curation de données); (ii) évaluer des requêtes de façon efficace (i.e., à la vitesse de la pensée) et intelligente (moteurs intelligents de requêtes). Les requêtes peuvent servir aux scientifiques de données pour proposer, au-delà des agrégations statistiques ou spatio-temporelles, des analyses fournissant des retours sur les attentes des citoyens en prenant en compte leur contexte (par ex. socio-économique, régional, professionnel).

En dépit de l'existence des résultats académiques et industriels capables de traiter et d'analyser des masses de documents, la plupart ne sont pas capables de donner des réponses "en temps-réel" et qui mettent en évidence le contexte et les conditions dans lesquelles ces résultats ont été obtenus. Deux concepts clés émergent : **conscience du contexte** (*context awareness*) et **réponses opportunes**. Notre proposition répond à ces défis qui sont au cœur d'une analyse dynamique utile pour des différentes organisations, agences, institutions gouvernementales qui désirent exploiter les données du Grand Débat.

## **Approche : description, originalité et pertinence**

Notre approche est basée sur l'expertise de trois domaines : le traitement automatique du langage (LIFO and LIFAT), l'expression et la ré-écriture des requêtes basées sur le contexte (LIFO), et le traitement optimisé des données et l'exécution efficace de requêtes sur les nuages (IRISA et LIG). L'approche proposée intègre d'une façon **novatrice**, le traitement automatique du langage (TAL) dans un contexte de curation de données, pour rendre utilisables (dans des formats pouvant être analysées) et maintenir des données collectées pendant les débats. Nous proposons une nouvelle notion nommée **requêtes basées sur le langage naturel et consciente du contexte** qui seront exécutées de manière efficace par des moteurs. Les réponses seront destinées pour des acteurs différents (citoyens, preneurs de décisions, agences sociales). Il s'agira des résultats quantitatifs obtenus par des statistiques et des agrégations. Les requêtes pourront demander des prédictions, des modèles ou des classifications, par exemple extraire des communautés d'intérêt, générer des fiches spatio-temporelles des préoccupations-actions organisées par types de communautés. Ces requêtes pourront également être calibrées par des contraintes du contexte comme la tendance politique, le contexte socio-économique, le rôle dans la société ou le gouvernement. Les résultats seront produits au moment opportun en appliquant un processus d'analyse dynamique et en dépit de leur coût de calcul. Nous proposons des environnements basés sur des infrastructures de calcul adaptées pouvant réaliser de manière efficace des différentes tâches du processus d'analyse dynamique. En plus des capacités de calcul et de stockage nécessaires, il sera question de proposer des stratégies d'optimisation des requêtes. Notre approche **revisitera les techniques d'optimisation de requêtes et des processus** en utilisant l'apprentissage automatique.

## **Moyens et outils déployés**

Les actions seront organisées autour de deux lots de travail où nous nous appuierons sur un capital humain et des résultats de recherche antérieurs :

**Lot 1 - Curation de collections de données :** l'objectif est de créer un environnement logiciel pour rendre les données utilisables pour des tâches d'analyse dynamique. L'expertise du LIFO et LIFAT pour traiter le contenu textuel exprimé en langage naturel sera clé. Les experts du LIFO et du LIFAT dans le domaine TAL travaillent sur l'application du traitement automatique des langues naturelles à l'analyse approfondie du discours politique, médiatique et citoyen dans de grandes masses de données textuelles. Dans le cadre du projet ANR TALAD<sup>[1]</sup> ils développent des outils pour faciliter et enrichir la démarche de fouille quantitative et qualitative menée par l'analyste du discours. L'équipe du LIG mettra à disposition CURARE, un système de curation de données qui fournit des outils intégrés à un environnement cloud pour assister les analystes des données à trouver quelle est la meilleure collection pour accomplir un objectif analytique donné.

**Lot 2 – Ré-écriture des requêtes basées sur le langage naturel guidées par le contexte :** l'objectif sera de réécrire des requêtes exprimées en langage naturelle vers des représentations qui puissent être optimisées et exécutées de façon efficace sur des architectures de calcul cibles. Le sous-groupe BD mène déjà des recherches dans le cadre de l'analyse et manipulation des données RDF ou des bases de données graphe. Les propositions du sous-groupe BD englobent les requêtes et mises à jour liées à un contexte prédéfini (éventuellement interactif), les requêtes étendues (agrégation, récursivité, calcul itérative, etc.) facilitant l'analyse des graphes, l'interrogation de données à grande échelle. L'IRISA a notamment proposé des structures de données à bases de Filtre de Bloom pour le traitement de jointures et de requêtes récursives dans des environnements massivement parallèles (basé sur le paradigme MapReduce).

Les partenaires de cette proposition participent au réseau DOING (Données Intelligentes) qui rassemble des informaticiens et linguistes pour traiter l'extraction d'information dans des textes pour l'acquisition, l'exploitation et maintenance des bases de connaissances dans un domaine donné. Ce projet pourra donc être potentiellement enrichi par d'autres membres de DOING.

## **Conclusion**

Appliquer des stratégies originales basées sur notre expertise permettra d'assurer la proposition de solutions utiles pour résoudre de problèmes complexes livrés pour une audience

ample avec une connaissance technique limitée mais qui souhaite comprendre les conditions dans lesquelles les résultats ont été obtenus.

---

[1]<https://web.u-cergy.fr/anr-talad/>