

Vers une approche semi-automatique pour la définition de motifs d'argumentation utilisés dans les résumés de projets scientifiques du domaine de la biodiversité

Rosa Cetro*, Marc Barbier **, Philippe Breucker **, Hilde Eggermont****‡, Philippe Gambette*, Tita Kyriacopoulou*, Xavier Le Roux****‡, Claude Martineau*, Nicolas Turenne**@

* Université Paris Est - UMR LIGM (Laboratoire d'Informatique Gaspard Monge)

** INRA Sens, IFRIS, Université Marne la Vallée (France)

*** Belgian Biodiversity Platform, Belgian Science Policy Office

**** Université de Lyon 1 - UMR Ecologie Microbienne

‡ERAnet Biodiversa

@ contact : nturenne.inra@yahoo.fr

Résumé. Nous positionnons notre travail dans le domaine de l'analyse et de la visualisation de données textuelles produites par les scientifiques et réunies en corpus calibré. Ce domaine est reconnu pour sa contribution à la réflexion sur la composition et l'évaluation des politiques scientifiques. Le corpus que nous utilisons est une collection de tous les résumés de projets acceptés dans des guichets d'appels à projet dans le domaine de la biodiversité référencés par le réseau européen BiodivERSA. L'objectif de ce travail ancré dans la sociologie des sciences consiste à mieux comprendre les principales caractéristiques utilisées par les scientifiques pour présenter leur projet et convaincre de ses qualités. Pour cela nous avons utilisé une pluralité d'outils face à la difficulté de dépouiller l'information pour associer le niveau sémantique (structure de l'information) au niveau pragmatique (relations entre les rédacteurs de projet). Notre contribution repose sur un nouveau type d'extraction d'information, hors entités nommées, basé sur l'extraction de motifs d'argumentation. D'une part on remarque que l'usage de ces motifs marque la présence d'arguments dans des résumés de projets, et d'autre part croît avec le temps.

1 Introduction

Les projets de collaboration scientifique ont été très peu étudiés par rapport à la littérature scientifique dans le cadre STS (science and technology studies) /SPS (social and political science). Toutefois, prendre l'entité « projet » comme objet d'étude est particulièrement pertinente à l'heure où l'organisation par projet constitue un modèle de financement et un modèle d'organisation pour la recherche publique. Dans cette contribution nous présentons une proposition d'étude de ce genre textuel, en nous appuyant sur un corpus de projets tirés de la base de données Biodiversa (www.biodiversa.org/8).

Une approche pour analyser un contenu textuel consiste à étudier la distribution des termes clés et d'en faire soit un inventaire soit un regroupement pour établir des tendances thématiques décrit par Lent et Agrawal (1997), Turenne et Barbier (2004), Mei et Zai (2005), Song et Kim (2012). Cette problématique de structuration documentaire est intéressante pour