SimDiversity: Similarity-reduced measures of diversity in the social and natural sciences
Projet Spark FNS 2020 (F.Bavaud, SLI-Lettres et IGD-FGSE, UNIL)

RESUME GRAND PUBLIC
Les mesures classiques de la variété d'une configuration (entropie de Shannon et variantes) sont basées sur le recensement de *catégories* ou *types* distincts, ainsi que de leur fréquence d'apparition. Or, quel que soit le niveau de granularité retenu, ces types ne sont pas totalement distincts, mais présentent des *similarités* plus ou moins prononcées, propres à *réduire* la valeur de la variété classique:

- les mots présentent des similarités sémantiques (variété lexicale dans un texte)
- les lieux géographiques sont co-fréquentés (variété spatiale)
- les espèces sont morphologiquement ou génétiquement apparentées (biodiversité)
- les votes des parlementaires affiliés à des formations politiques distinctes coïncident souvent (variété politique).

CONTENU ET OBJECTIFS DU TRAVAIL DE RECHERCHE
Le projet SimDiversity est consacré à l'étude des mesures de variété tenant compte de la *similarité entre types*, en particulier des deux mesures que sont l'*entropie réduite*, récemment introduite en biodiversité, et une nouvelle quantité, l'*entropie effective*, dont le calcul, itératif, révèle de forts liens avec la mécanique statistique et les transitions de phase, le clustering et la quantification vectorielle, le transport optimal, et les modèles de perception et de choix en psychologie.
L'entropie effective est toujours concave (i.e. possède une décomposition intra/inter-groupes), au contraire de l'entropie réduite dont la concavité dépend de conditions supplémentaires sur la matrice des similarités, lesquelles restent à caractériser complètement.

Dans son volet empirique, le projet SimDiversity collectera, en dialogue avec des spécialistes de chaque discipline (linguistique, géographie, biologie, politologie), des jeux de données de grande taille, sur lesquels sont calculées et analysées ces mesures de diversité réduites, du point de vue de leur comportement numérique ainsi que de leur interprétation disciplinaire.

CONTEXTE SCIENTIFIQUE ET SOCIAL DU TRAVAIL DE RECHERCHE
L'objectif à terme du travail est de contribuer à promouvoir et unifier l'usage de ces *mesures de variété réduite* au-delà des particularités disciplinaires, et d'encourager leur diffusion dans la pratique générale en Analyse des Données.

MOTS-CLES : effective diversity, similarities between types, information theory, phase transitions, textual richness, political diversity, similarity-reduced biodiversity