{WikiStat}.py

CHOPIN Johann
HENRY Justin
GUIDOUX Alexandre

I.Introduction

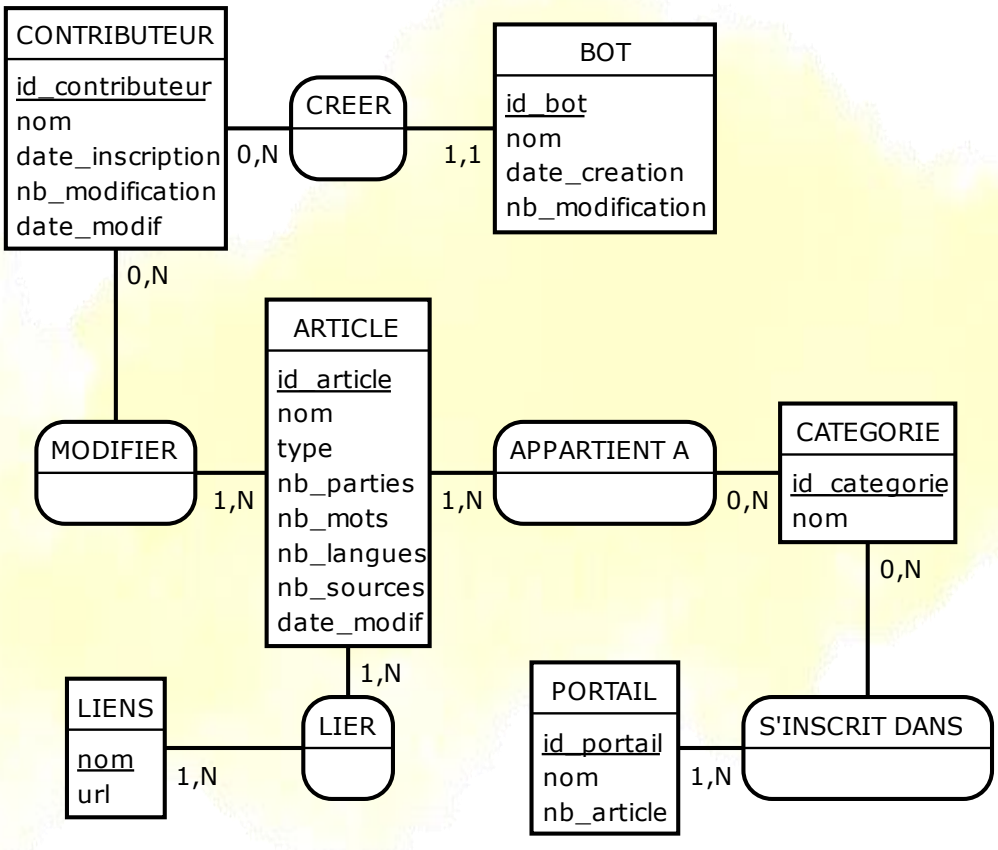
Point de départ

L'information, c'est le pouvoir…

…et c'est pourquoi notre objectif est d'amasser un grand nombre de données, d'avoir quelque chose d'exploitable. Nous aimons programmer en python et la plus grande source de données est le web ; il est tout désigné. Mais quelles informations y extraire ? Nous ciblons un seul site : Wikipédia ; les informations y sont bien organisées et faciles d'accès. Faire un projet pour faire un projet n'a pas vraiment de sens, il faut que ce soit utile et amusant. Notre projet essaye de répondre à ces questions : qu'est-ce qui distingue les articles de qualité et les bons articles, des articles normaux sur Wikipédia ? Comment les caractérise-t-on ?

Conception du MCD

L'*article* est l'élément primordial de Wikipédia. C'est à partir de cette entité que l'on peut extraire le plus de données. On recherche ensuite les entités liées ; on obtient les *catégories*, les *liens*, les *contributeurs ;* puis les *bots* et les *portails*. Cela fait déjà un assez grand volume de données à récupérer et à traiter. Toutes ces données nous renvoient au final ce MCD :



CONTRIBUTEUR (id_contributeur, nom, date_inscription, nb_modification, date_modif)
BOT (id_bot, nom, date_creation, nb_modification, id_contributeur)
MODIFIER (id_article, id_contributeur)
ARTICLE (id_article, nom, type, nb_parties, nb_mots, nb_langues, nb_sources, date_modif)
APPARTIENT A (id_categorie, id_article)
CATEGORIE (id_categorie, nom)
LIENS (nom, url)
LIER (id_article, nom)
PORTAIL (id_portail, nom, nb_article)