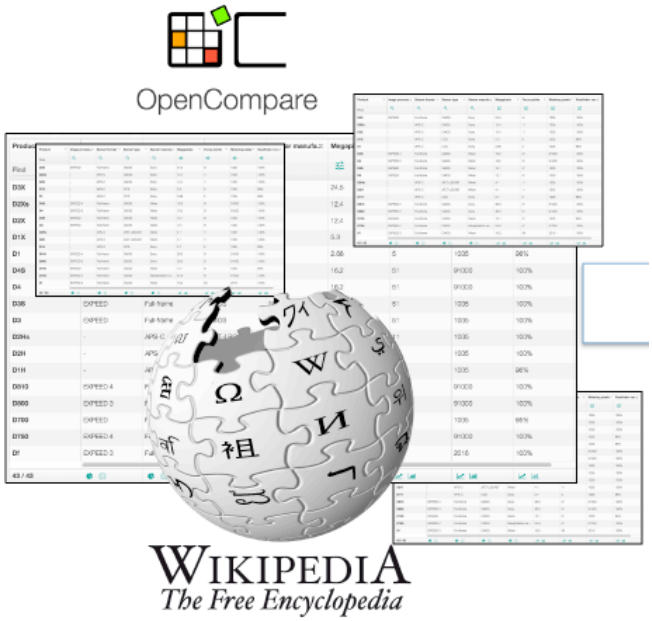
MIAGE 1
Projet Développement Logiciel –
Responsable : Mathieu Acher
Contact : mathieu.acher@irisa.fr
2017/2018

Gestion de projet – Projet Développement Logiciel (PDL)

Projet #4: « WikipediaMatrixAnalysis »

Ressources :
http://mathieuacher.com/teaching/PDL/

Wikipedia est une fantastique source de données. Nous avons extrait des millions de données tabulaires provenant de Wikipedia et nous souhaitons comprendre finement leurs « propriétés ». Cette analyse des données tabulaires nous permettra de concevoir des outils plus adaptés que ceux fournis par défaut dans Wikipedia. C'est en tout cas un des objectifs d'opencompare.

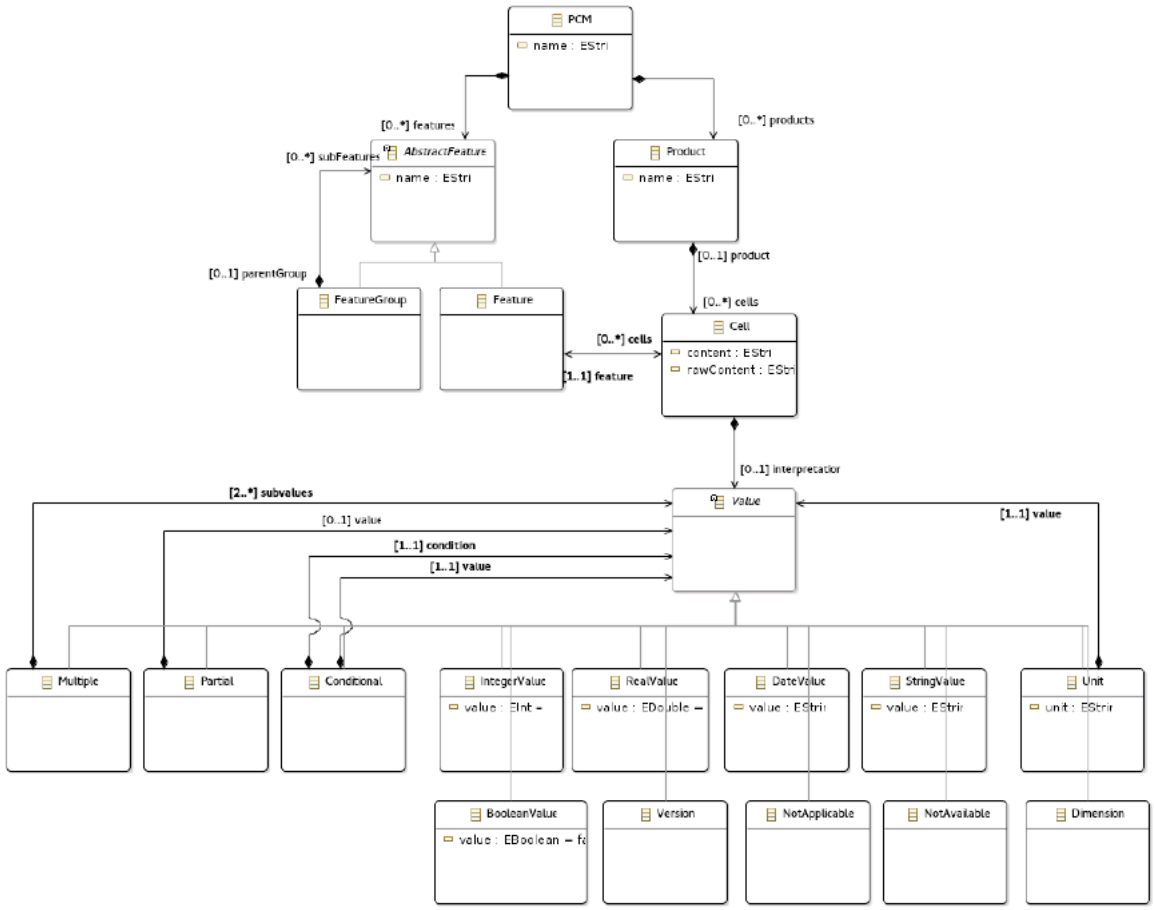


Au cœur de la problématique est la définition du « format » censé représenter les matrices. Le format permet de structurer l'information d'une matrice, et n'importe quel outil (configurateur, tri, édition, etc.) s'appuie sur ce format. Autrement dit, la conception du format est cruciale pour la qualité des futurs outils.

Plusieurs tentatives de définition du format PCM (pour Product Comparison Matrix) ont été effectuées dans le passé (cf par exemple Chapitre 4 et 5 de la thèse de Guillaume Bécan disponible ici : https://tel.archives-ouvertes.fr/tel-01416129/)

L'objectif de ce projet de PDL est d'analyser des milliers de PCM provenant de Wikipedia et de produire des recommendations argumentées sur le format PCM: Est-ce que certains concepts sont manquants? Est-ce que les données de Wikipedia sont correctement encodés dans ce format?

Nous utiliserons l'API Java d'opencompare, un projet de mise en route est disponible ici: https://github.com/OpenCompare/getting-started Le diagramme de classes ci-dessous donne la structure du format PCM et de l'API Java. A noter que nous utilisons un format PCM qui est different de celui utilisé actuellement sur le site web (le projet #1 sera l'occasion de comparer les deux formats)



Concrètement, des milliers de matrices provenant de Wikipedia, au format PCM, vous sont fournis. Le but est d'adresser certaines questions comme:

- Quels sont les noms de "features"/"produits" qui reviennent souvent?
- Quelles sont les tailles des matrices?

- Quelles sont les valeurs de cellules les plus fréquentes?
- Quelles sont les valeurs de cellule qui sont le plus souvent ensemble dans une colonne ou une ligne?
- Y a-t-il une correlation entre le "type" prédominant d'un "feature" et les valeurs des cellules?
- Est-ce que les valeurs d'une colonne sont "homogènes" (i.e., du meme "type")?
- Est-ce qu'on peut idenifier des matrices "similaires"?
- Dans quels cas la procedure d'extraction est défectueuse?
- …

L'analyse s'effectuera de deux manières:

- Automatiquement, avec l'API Java d'OpenCompare : https://github.com/OpenCompare/getting-started pour pouvoir lire des matrices et donc les traiter/analyser à large échelle. Il est recommandé d'utiliser des outils statistiques dédiés (en complément du Java) pour répondre à certaines questions (on pourra utiliser R ou Python par exemple, éventuellement Excel)
- Manuellement: il faudra très certainement examiner certaines matrices individuellement. De plus, il faudra interpreter les données (statistiques) que vous aurez collectées grâce à votre analyse automatique.

Comment commencer ?

Etudier l'API d'opencompare : https://github.com/OpenCompare/getting-started

Utiliser l'API sur certains PCMs et commencer le travail d'analyse.

Les archives PCM fournies:

- http://mathieuacher.com/teaching/PDL/201617/modelPCM-Wikipedia-10102016.zip
- archives tar.gz disponibles ici : http://mathieuacher.com/teaching/PDL/
- d'autres seront fournies en cours de projet

Attention : ne pas nécessairement analyser les milliers de PCM à la première itération : commencer par analyser un « sample » (ou corpus) réduit. Vous ferez passer à l'échelle votre solution plus tard dans le projet