Conclusions de stage Digital Africa (13/04/2020-31/07/2020)

Paolo Monteiro de Macedo – Le 24/07/2020

Créer une infrastructure de données ouvertes pour le développement d’entreprises numériques à fort impact social en Afrique

**Motivation au projet d’infrastructure de donnée ouvert**

Après plusieurs discussions de définition de projets, j’ai été recruté en vue de la création d’une infrastructure de données ouvertes destinées à des entreprises africaines capables d’utiliser ces données comme levier de croissance. Si le périmètre exact de l’infrastructure ainsi que les données cibles restaient indéfinis, ce projet de Digital Africa repose sur la conviction que les données ouvertes sont une facilité essentielle au développement d’entreprises numériques, porteuses d’une transformation technologique, économique et sociale, capables de répondre à certaines problématiques propres aux pays africains. Ainsi, les axiomes initiaux de ce projet peuvent être définis comme tel :

1. La structuration et le soutien à l’entrepreneuriat est susceptible d’accélérer la croissance économique et le développement d’une protection sociale adaptée aux peuples africains ;
2. Parmi ces entreprises, les projets fondés sur des technologies numériques sont susceptibles d’avoir un impact positif plus large et à moindre coût ;
3. Le développement de ces projets entrepreneuriaux est freiné par un accès insuffisant à des données essentielles, existantes mais non structurées, non centralisées ;
4. Le coût d’accès à ces données essentielles est excessif en rapport à la propension à payer pour ce service pour que des initiatives privées puissent se développer par elles-mêmes pour répondre au besoin existant. En outre, les effets de réseaux et la nature de bien commun des données rend optimal l’allocation de cette infrastructure par une entité semi-publique ;
5. Les entités publiques africaines n’ont pas engagé une démarche dans le sens du développement d’une telle solution à l’échelle continentale et ne sont pas susceptibles de la développer.

Le point 1 relève de la politique économique de développement dont les ressorts ne relèvent pas du spectre d’action direct de Digital Africa. Le point 2 semble manifeste par les propriétés technologiques du numérique et l’expérience quotidien bien qu’il exclut les éléments sous-jacents à la réussite d’un écosystème numérique performant (telle qu’une industrie capable de produire les biens de consommation sur lesquels se fondent les services des technologies numériques). Les points 3 et 4 sont des hypothèses de travail que Digital Africa se propose de démontrer par une évaluation rigoureuse de son projet d’infrastructure de donnée. Le point 5 enfin relève de l’économie politique.

**Mise en œuvre du projet**

Afin de mener à bien ce projet d’infrastructure de données, il convenait dans un premier temps d’évaluer si le besoin existait effectivement. A cet effet, Digital Africa dispose d’une base de données de réponse à un questionnaire, le « challenge des 1000 », qui avait été envoyé à destination de jeunes entreprises africaines en vue de la participation au sommet Afrique-France de juin 2020. Cette base de données contient notamment une description détaillée des produits proposés par chacune de ces 3400 jeunes entreprises ainsi qu’une catégorisation par secteur et par pays.

Après un nettoyage approprié des informations contenues dans la base de données, nous avons pu engager un travail d’analyse des informations contenues dans le texte de description de ces entreprises. En effet, les labels de catégories nous semblaient trop grossiers pour pouvoir obtenir des informations pertinentes sur chacune des observations (« agriculture », « énergie »…). Nous souhaitions connaître, de manière automatisée, le type d’activité précis qu’exerce chacune de ces entreprises afin de pouvoir identifier au mieux les besoins sur un secteur pertinent. Une catégorie trop large représente des modèles d’affaire très distincts les uns des autres qui ne correspondent pas du tout aux mêmes réalités. Cependant, la difficulté est que nous ne connaissons pas, a priori, les différents sous-secteurs d’activité au sein de chacune de ces catégories d’activité. Afin de pallier cette difficulté nous utilisons un algorithme de regroupement des observations, au sein d’une catégorie, par la proximité du langage utilisé dans la description de l’activité de chacune des observations. Ainsi, deux entreprises qui travaillent dans la numérisation de l’activité agricole sont susceptibles d’utiliser un langage plus « proche » dans la description de leur produit qu’une exploitation agricole avec une entreprise d’objets connectés pour de l’agriculture de précision. Nous avons opéré cette analyse pour le secteur de l’agriculture et de la mobilité.

Après l’identification de ces sous-catégories au sein d’une large catégorie, l’algorithme qui procède à ce regroupement d’observation (nommé « clusterisation ») n’est pas exempt d’erreur. En outre, si nous sommes capables de regrouper grossièrement des entreprises par la proximité du langage utilisé dans leur description, nous ne connaissons pas bien le type d’activité auquel correspond chaque « sac » (cluster) d’observations. Ainsi, le travail suivant a été l’identification précise de certains types de modèles d’affaire au sein de ces ensembles d’observations regroupé par notre algorithme de clusterisation, par la lecture directe de la description de ces entreprises. Nous avons pu de la sorte identifier 3 modèles d’affaires pour les entreprises pour lesquelles le langage était le plus proche de sujets liés au numérique pour la catégorie agriculture : les plateformes de mise en relation entre producteurs agricoles et consommateurs ou entreprises ; les entreprises d’objets connectés qui font de l’agriculture de précision à l’aide des données collectées ; les entreprises de conseil plus général à destination des agriculteurs, en gestion financière ou agronomique. Nous avons décidé de nous intéresser aux plateformes qui représentaient l’échantillon le plus large et celles pour lesquelles nous pensions qu’un besoin en données ouvertes était le plus susceptible d’exister.

Une fois focalisé sur ce sous-sous-ensemble d’entreprises (des entreprises qui travaillent dans le secteur de l’agriculture, en lien avec le numérique, qui sont des plateformes), nous pouvions désormais identifier un besoin en données ouvertes. Pour cela, nous avons définis certaines hypothèses sur les données qui seraient susceptibles d’intéresser ces acteurs et nous avons interrogé 3 startups qui semblaient les plus représentatives des utilisateurs potentiels de l’infrastructure de donnée. Ces interviews nous ont permis de valider l’existence d’un besoin pour ces entreprises et un jeu de données qui semblait particulièrement pertinent, à savoir un répertoire d’agriculteurs sur un territoire avec une localisation pour chacun d’entre eux.

Le procédé suivant aurait été la recherche de ces données afin de les mettre à disposition sur l’infrastructure de données ouverte, l’évaluation de l’exploitation par les utilisateurs et l’impact des données sur leur croissance. Pour trouver ces données, une piste intéressante consiste en l’utilisation d’imagerie satellite afin de réaliser de la reconnaissance de parcelles d’exploitation qui permet d’inférer sur la présence d’agriculteurs. Toutefois ce processus semblait excessivement complexe sans un jeu de données satellites pré-travaillées qui devrait être produit par un partenaire dans les prochains mois.

En l’absence des données sources utiles pour créer le répertoire et le mettre à disposition sur l’infrastructure de données, nous avons travaillé sur l’amélioration de notre labellisation des entreprises à partir du texte de description. Le processus initial consistait dans le regroupement d’observation par la proximité du texte puis la définition de sous-catégories (voire sous-sous-catégories) d’activité à partir de ces clusters. La réflexion suivante consistait à prendre ce processus dans le sens contraire. Lorsque nous connaissons une sous-catégorie d’activité, comment pouvons-nous identifier les entreprises qui correspondent à cette sous-catégorie à partir de leur texte de description ? En bref, l’objectif était de créer un moteur de recherche d’entreprises à partir de sous-catégories, lorsque les entreprises n’ont pas été, a priori, définies comme appartenant à cette sous-catégorie. Cela permettra notamment d’augmenter l’échantillon, une fois une sous-catégorie identifiée, des entreprises correspondant à cette sous-catégorie. Ce procédé d’identification des entreprises permettrait notamment de réaliser des questionnaires plus massifs, d’interroger et d’agréger les réponses de toutes les entreprises correspondant à une sous-catégorie pour avoir un point de vue plus représentatif du besoin en données ouvertes.

L’ensemble des étapes sont encore à leur balbutiement et doivent être encore largement améliorées notamment dans leur automatisation. Néanmoins, à mesure que la base de données d’entreprise s’accroitra, que les sous-catégories identifiées seront nombreuses et les algorithmes performants et précis, plus la procédure sera efficiente dans la définition de besoin en données ouvertes. On peut résumer ce procédé comme suit :

Identification des autres observations correspondant au modèle d’affaire d’intérêt pour agrandir l’échantillon.

Base de données : collection d’entreprises

MA 1

MA 2

MA 3

Modèles d’affaires fortement congruents, besoin cohérent à cette échelle, peu d’observations

Processus à automatiser

Sous-catégorie D

Sous-catégorie C

Sous-catégorie B

Sous-catégorie A

Catégorie A

Observations nombreuses, sous-catégories diverses et non-congruentes

Algorithme de clusterisation

Sous-catégories plus congruentes, observations moins nombreuses, thématique ambiguë

Algorithme d’identification (search engine)

Modèle d’affaire considéré d’intérêt

Questionnaire de besoin : identification du besoin

**Conclusion**

La durée de mon stage ne m’a malheureusement de parvenir au bout du processus qui consiste en la fourniture, sur une infrastructure, des données identifiées comme répondant au besoin. Pour autant, la définition d’un procédé cohérent et global qui permet de systématiser la définition d’un besoin pour un sous-secteur et un modèle d’affaire en particulier semble être déjà en soi une avancée considérable. Des institutions publiques, dans l’idéal supranationales, peuvent se saisir de cette expression de besoin construite pour proposer le produit nécessaire à la réponse à ce besoin.

En outre, si Digital Africa considère qu’il est de son ressort de répondre à ces besoins multiples qu’elle fait émerger, un travail considérable mais passionnant s’ouvre pour l’association. Répondre à ces besoins demandera sans aucun doute des moyens importants. Pour autant, des économies d’échelles sont prévisibles et des effets de réseaux positifs pourront probablement amortir l’investissement. Il est peut-être du seul ressort d’une entité publique de prendre de tels risques.

Cependant un tel projet restera difficilement évaluable par la nature nécessairement commune des données. Les externalités positives sont peu quantifiables et suivre le trajet d’un jeu de donnée ainsi que son importance dans la réussite globale d’une entreprise relève du défi. Néanmoins, disposer d’éléments objectifs pour évaluer la réussite d’un tel projet semble absolument nécessaire afin de prouver la nécessité de telles infrastructures dans le développement d’un écosystème numérique, et par conséquent économique et social, sain. La nature difficilement évaluable et potentiellement coûteuse de l’investissement demande donc une conviction complète sur le bien-fondé de ce dernier. Seule une équipe totalement motivée, convaincue de la réussite finale du projet, soudée et efficiente pourra mener à bien ce projet et démontrer les fruits, sans aucun doute nombreux, d’une infrastructure de données ouvertes. Cette posture radicale doit probablement être maintenue en silo, imperméable aux circonstances externes sans rapport avec la réussite finale du projet, sur un temps prédéfini. En ce sens, le modèle des startups d’Etat est probablement une inspiration qui pourrait convenir à cette expérimentation avec un porteur de projet entièrement investi dans la réussite finale du produit et une équipe technique idoine.

Une solution alternative serait de reposer sur l’existant et travailler à l’optimisation du procédé permettant l’identification de besoin en données ouvertes en laissant la production de données à d’autres partenaires, en renseignant sur les besoins remontant. Pour cela, plusieurs points d’amélioration sont déjà clairement définis. Un point d’amélioration utile serait également la hiérarchisation des besoins à partir d’indicateurs objectifs (taille d’un marché, capacité à répondre à un défi du développement durable etc.). Cette solution est peut-être moins ambitieuse mais repose sur des fondations solides et sa réussite semble plus assurée. Il conviendrait néanmoins d’évaluer si un tel produit intéresserait effectivement d’autres partenaires, susceptibles par la suite de répondre au besoin exprimé. Digital Africa fonctionnerait alors comme un agrégateur de besoins sectoriels, utile pour la définition de la politique économique de pays africains.

Ce choix stratégique est complexe et doit prendre en considération aussi bien les moyens alloués au projet, les équipes, la définition claire de la mission de Digital Africa et le soutien obtenu par la direction dans l’une ou l’autre direction. Cependant, il me semble que quel que soit la direction choisie, pour réussir, elle doit être totale et ne pas badiner avec une multiplicité d’objectifs.