Fiche cadrage / documentation du challenge 1 d'OpenEdition : Qualifier les cas de fréquentation atypique des contenus OE

Objectif : chaque challenge doit être cadré et documenté en amont de leur présentation le premier jour de la semaine Challenge Data, de manière à être le mieux compris par les étudiants qui travailleront dessus. Ce cadrage doit également vous assurer que leurs contributions seront pertinentes dans la résolution du challenge.
Il y a deux phases principales :

- Le cadrage du challenge
- La documentation du challenge

De manière à homogénéiser les contributions, nous vous demandons de nous fournir les éléments de votre fiche cadrage et documentation du challenge en remplissant les pages suivantes et en nous envoyant la documentation par mail à timothee@datactivi.st
Si vous éprouvez des difficultés ou avez des questions, n'hésitez pas à me contacter.

Date butoir de réception de la fiche Challenge : vendredi 8 novembre

Fiche cadrage Challenge Data 2ème édition

1. Cadrer le challenge : Enjeux (env 10-12 lignes)
A quels grands enjeux répondent ce challenge ? Est-il davantage tourné sur des problématiques d'ouverture de données publiques ou de réutilisation de données déjà ouvertes ? Les données sont-elles au cœur du challenge ou seulement une brique permettant la mise en place d'un service ou d'une politique publique ?

Contexte général :
OpenEdition est une infrastructure complète d'édition électronique au service de la communication scientifique en sciences humaines et sociales. Elle rassemble quatre plateformes complémentaires :

- OpenEdition Journals (anciennement Revues.org) : c'est le plus ancien portail français de revues en ligne. Il est spécialisé en lettres, sciences humaines et sciences sociales. Il accueille aujourd'hui plus de 500 revues en ligne, soit plus de 200 000 articles, dont 95 % sont accessibles en texte intégral
- OpenEdition Books : plateforme de publication de livres dont près des ¾ sont en accès ouvert. En 2019, la plateforme rassemble près de 7 800 livres en sciences humaines et sociales provenant de plus de 90 éditeurs
- Calenda (portail) : il publie des milliers d'annonces d'événements scientifiques : colloques, séminaires, ainsi que des offres d'emploi et des appels à contribution
- Hypothèses.org : plateforme de blogging scientifique. Les chercheurs y créent des « carnets de recherches » dans lesquels ils font état des avancées de leurs recherches. Hypothèses regroupe, en 2019, plus de 3 000 carnets de recherche

Nous contacter : Timothée Gidoin / timothee@datactivi.st / 06 71 53 01 28
Cécile Le Guen / cecile@datactivi.st / 06 84 99 87 00

animés par une communauté de carnetiers de tous pays. L'ensemble des contenus est en libre accès.

Ces plateformes disposent de conseils scientifiques, qui sélectionnent les publications afin d'assurer une qualité scientifique à l'ensemble. En 2018, l'ensemble de ces plateformes, dont les contenus sont majoritairement en libre accès, a reçu 64 millions de visites, provenant du monde entier.

Des services complémentaires sont proposés via les bibliothèques et institutions abonnées (modèle économique de type « freemium », i.e accès libre pour la plupart des contenus OpenEdition et des services payants, comme le téléchargement des articles en pdf).

Contexte du challenge :

Sur la base de logs – *équivalent de journal de bord qui retrace l'historique des événements d'un site* - OpenEdition a développé un outil, dénommé ***Umberto***, qui analyse les fréquentations des contenus du site OpenEdition.

L'outil Umberto détecte les événements atypiques et s'intéresse notamment aux valeurs aberrantes. L'outil devrait ainsi, en théorie, permettre de démontrer l'intérêt de la littérature scientifique pour un public plus large.

En parallèle, développement d'un autre outil, Emile, qui récupère les logs et extrait les IP les plus fréquentés afin de déterminer l'origine de fréquentation (souvent des universités mais parfois aussi des entreprises, des associations…). Cela a permis de construire un jeu de données qui collecte tous les affluents de consultation du site Open Edition (URL, contexte lexical). Environ 7300 affluents ont été collectés, il s'agit désormais de les analyser.

Ainsi, 2151 contenus en ligne (sur plus de 800 000 contenus OpenEdition) ont eu des fréquentations aberrantes de janvier 2017 à juillet 2019. Mais, à ce jour, seulement une petite dizaine de cas ont été documentés, i.e, analysés en profondeur de manière à comprendre les raisons de ces fréquentations élevées.

Ce challenge est le prolongement du défi : « analyser la fréquentation des plateformes » du Datathon Read Write Cite qui a eu lieu en septembre 2019 à Marseille.

2. Cadrer le challenge : Problématique (une phrase) *

La problématique doit pouvoir tenir en une phrase. Elle peut être formulée sous la forme d'une question. ***Ce sera l'intitulé du challenge.*** *La précédente définition des enjeux devrait vous permettre de dessiner plus facilement la problématique.*

Qui sont les principaux lecteurs atypiques des contenus OpenEdition et quels usages en font-ils ?

3. Cadrer le challenge : les attendus (env 8-10 lignes) *

Nous contacter : Timothée Gidoin / timothee@datactivi.st / 06 71 53 01 28
Cécile Le Guen / cecile@datactivi.st / 06 84 99 87 00

Qu'attendez-vous des étudiants ? Comment peuvent-ils vous aider à répondre à la problématique, notamment en profitant de leur œil "neuf", de leurs connaissances de l'action publique et de l'accompagnement en design et open data qu'ils recevront ? Explicitez vos attentes tant sur les réflexions qu'ils auront que sur le ou les livrables qu'ils vous remettront à l'issue de la semaine. L'objectif de cette partie est de s'assurer que leurs travaux vous seront utiles et conformes à vos attentes. Quelle forme souhaitez-vous que prenne le livrable des étudiants ? Cartographie des données, prototype de service, rédaction d'un cahier des charges fonctionnel d'un service basé sur des données, réaliser une "mini" AMO à une politique d'open data, faire un benchmark des usages ou des communautés d'utilisateurs, l'analyse / croisement de données ouvertes… La dimension purement technique ne doit pas être dominante mais elle peut, bien sûr, être présente.

Il vous est demandé d'explorer et de qualifier le jeu de données des affluents qui amènent des cas inattendus de consultation et de produire une synthèse sur des cas particulièrement pertinents permettant de faire un plaidoyer de l'Open Access.
Vous serez accompagnés pour cela de Pierre-Carl Langlais, docteur en sciences de l'information et de la communication, qui a été à l'origine du jeu de données des affluents.

Au-delà de la qualification du jeu de données des affluents, il faudra également produire une synthèse générale récapitulant la diversité des cas étudiés, dans une optique de plaidoyer. Poster, vidéo, cartographie, infographie, dataviz… A vous d'être créatifs et de concevoir la ou les meilleures manières de restituer ces informations.

4. Documenter le challenge : bases de données ou jeux de données *
Listez ci-dessous les jeux de données ou bases de données publiques que vous souhaitez mettre à disposition des étudiants, que vous en soyez le producteur ou non.
Si tout est accessible publiquement, via des portails open data ou des plateformes, il suffit de lister certains liens URL et de mettre une petite phrase explicative.

- *Umberto, le détecteur de lecteurs d'Open Edition (avec la base de données des 2151 cas à analyser) :* https://analytics.huma-num.fr/Huma-num/umberto_oe/

5. Documenter le challenge : documents
Cela peut prendre la forme de documents publics, rapports, études, présentations internes. Ceux-ci pourront soit être transmis par lien URL (s'ils sont accessibles publiquement), soit directement dans un drive que nous mettrons à disposition du groupe.
Si des documents internes sont communiqués, nous rappellerons aux étudiants que ceux-ci ne doivent en aucun cas être partagés ou diffusés en dehors de ce Challenge.

Les Posters :

- Poster « Usages d'OpenEdition dans des organisations non-académiques » (et en version pdf)
- Catégorisation des usages de références savantes en contexte non-académique. Le cas de forum ou l'écriture inattendue

Nous contacter : Timothée Gidoin / timothee@datactivi.st / 06 71 53 01 28
Cécile Le Guen / cecile@datactivi.st / 06 84 99 87 00

- Template de poster

Exemples d'histoires racontées par les affluents :

- Histoire 1 : Forum de modélisme en allemand. Recherche de source pour une reproduction de la Grande Réale (galère Louis XIV) >>> iconographie du CRCV (en FR)

- Histoire 2 : Discussion autour du Mercato 2018/2019 et retombées économiques pour l'OM >>> longue citation de FCS (+ autres liens académiques)

- Histoire 3. Latinistes qui s'interrogent sur l'usage possible du neutre en latin pour écriture inclusive (intersexe et transgenre) >>> Carnet Hypo intersexe (hermaphrodisme)

- Bonus. Geneanet utilise la plateforme Journals comme ressource pour son moteur de recherche

Pour aller plus loin (notamment sur l'autre challenge OpenEdition) : Le dossier avec l'ensemble des fiches de connaissance

Documenter le challenge : liens externes

La documentation peut aussi être nourrie de liens vers des initiatives externes ou des articles qui peuvent nourrir leur réflexion ou leur permettre de démarrer un benchmark.

- Datathon Read Write Cite (Septembre 2019)
- Le site OpenEdition

Nous contacter : Timothée Gidoin / timothee@datactivi.st / 06 71 53 01 28
Cécile Le Guen / cecile@datactivi.st / 06 84 99 87 00