A LIRE AVANT DE RENSEIGNER LE DOCUMENT

Ce document servira à identifier les besoins en collecte et analyse de données pour la création d'indicateurs et à connaître les pratiques en analyse de données dans le cadre de Hubble. Cette synthèse permettra la création du modèle conceptuel pour la plateforme Hubble.

En fonction du travail de collecte et/ou d'analyse que vous avez fait, vous renseignerez uniquement les parties qui concernent votre travail. Nous vous demandons d'être le plus précis possible dans vos descriptions.

==========

Date de rédaction : 27 juin 2016 (actualisé le 06 janvier 2017, révisé le 28 juin 2017)

Nom du rédacteur du document : Jean-François Cerisier, Hassina El Kechaï, Laetitia Pierrot et Lucie Pottier (revu par Laëtitia Pierrot et Sergio Ramirez)
Spécialités : ~~Didactique, Sciences de l'éducation~~, Informatique, Analyse de données, Sciences de l'information et de la communication (*rayer les mentions inutiles*)

Cas d'étude Hubble : TECHNÉ

Scénario hubble :

Personnes impliquées pour la collecte et l'analyse : TECHNÉ (Jean-François Cerisier, Hassina El Kechaï, Laetitia Pierrot, Lucie Pottier et Sergio Ramirez), LIUM (Sébastien Iksal) et Lab-STICC

Période de la collecte : 2015-2016, réitéré en 2016-17

Période de l'analyse : juin-~~novembre~~ 2016 – avril 2017
==========

Dispositif d'apprentissage (Étude de cas de Hubble)

Type de dispositif : Accompagnement scientifique du projet Living Cloud qui fournit des équipements numériques nomades aux élèves du Lycée Pilote Innovant International (LP2I).

Finalité de l'apprentissage : Le projet Living Cloud a comme objectif de transformer les pratiques pédagogiques des enseignants ainsi que les conditions d'apprentissage des élèves. Le rôle du laboratoire Techné est alors d'observer l'impact du projet sur ces deux éléments à travers 5 sous-projets, dont le premier qui est basé principalement sur la collecte et l'analyse de traces. Ce sous-projet 1 a comme vocation de fournir des éléments de compréhension sur les usages numériques des lycéens.

Utilisation du dispositif et fonctionnalités : L'établissement étudié comporte près de 500 élèves en voie générale. Ces élèves sont équipés dès leur rentrée en seconde d'une tablette ou ordinateur portable (ou hybride). En 2015-16, après diffusion d'une autorisation de participation au projet, 161 élèves (76 en seconde, 53 en première et 32 en terminale) ont accepté de fournir leurs données d'utilisation à partir du réseau du lycée (logs du proxy). De plus, 54 élèves ont accepté d'être équipés d'un logiciel de traçage (Kidlogger) permettant de suivre l'activité en ligne et hors ligne, à la fois dans l'établissement et également en dehors.
Sont tracés:

- l'équipement utilisé ;
- les sites, fichiers et dossiers consultés ;
- les applications lancées ;
- la durée et la période d'utilisation.

Ces diverses données ont vocation à alimenter un tableau de bord.
Afin de compléter l'analyse de ces données collectées automatiquement, des entretiens d'explicitation, individuels ou collectifs, sont organisés auprès des élèves volontaires.

Contexte de production de données : Les données du proxy sont disponibles sur la période de l'année scolaire, à savoir de septembre 2015 à juin 2016 pour la première année de collecte. Les logs du logiciel de traçage sont disponibles depuis avril-mai 2016.
En plus de ces données numériques, nous disposons des caractéristiques des élèves (sexe, age, classe et niveau d'étude, PCS…) ainsi que de leur emploi du temps des cours. Ces données sont fournies par le lycée.
Une nouvelle session de collecte des données est mise en place pour l'année scolaire 2016-2017 (mêmes données collectées : proxy / et logs d'un logiciel de traçage)

(Au besoin indiquer les différentes moments de la production (savoir si des données ont été produites sur plusieurs années)

Décrire la problématique posée et les objectifs de l'analyse :

Problématique générale Comprendre les conditions d'appropriation à l'échelle collective du numérique et identifier les types de pratiques numériques des lycéens du LP2I.

Objectifs de l'analyse : Les besoins d'analyse présentés ici sont liés au travail de recherche doctorale portant sur la circulation des pratiques numériques des lycéens. La circulation correspond dans notre cas au passage d'une pratique individuelle (par un lycéen) à une pratique collective (par plusieurs). La circulation renvoie donc à l'évolution des usages dans le temps. Le préalable à la compréhension de cette circulation est l'identification de pratiques. Les besoins d'analyse sont les suivants :

- identifier et caractériser les pratiques numériques des lycéens;
- à partir des actions numériques des lycéens et vues à l'échelle individuelle et collective obtenir la liste des pratiques communes à plusieurs lycéens et propres à certains,
- en fonction du contexte (au lycée (en cours/en dehors des cours) et en dehors du lycée) obtenir la liste des actions

classées pour chaque contexte

- en fonction de l'intention (actions à des fins pédagogiques ou personnelles) obtenir la liste des actions classées par finalité
- en fonction de la récurrence (actions fréquentes ou non) obtenir la liste des actions les plus fréquentes

- identifier des pratiques issues du milieu social
  - en fonction de leur marquage temporel obtenir la liste des pratiques des lycéens marquées dans le temps pour ensuite obtenir la liste des pratiques nouvelles pour des utilisateurs donnés

Production des données :

Décrire le processus de production des données brutes :

- Pour les logs du proxy, collecte systématique des sites consultés par utilisateur selon un login attribué (y compris Pop-up, ad, cookie…).
- Pour les logs de Kidlogger, collecte systématique des actions des utilisateurs par équipement tracé.
- Les données complémentaires (données sociodémographiques des élèves, emplois du temps) sont envoyées en une fois par le lycée au format texte et sont ajoutées à notre base de données.

Extraction des données au format SQL à partir des bases de données hébergées l'Université de Poitiers (Phpmyadmin)

Liste des variables initialement recueillies : Nom et Description

LOGS PROXY :

- datetime indique la date et l'heure de connexion et est exprimé sous la forme JJ/MM/AAAA HH :MM.
- displayed_login est l'identifiant attribué par l'établissement à l'élève, sous la forme NomInitialeprénom.
- ip correspond à l'adresse IP de l'équipement connecté.
- url correspond à l'adresse complète consultée.
- Domain précise le nom de domaine du site consulté.
- web_protocol précise le protocole web utilisé.

LOGS KIDLOGGER :

- ID correspond à l'identifiant unique de l'action enregistrée
- Device_id correspond à l'identifiant associé à l'élève tracé, suivant son équipement
- wcSection_id et wcCategory_id sont des colonnes à ignorer
- log_type indique la nature de l'action. Il peut s'agir d'une application (app), d'une consultation web (url), d'une recherche en ligne (search-query), d'un appel entrant (in_call) ou sortant (out_call), d'une saisie de clavier (keystrokes), d'un dossier consulté (folder), d'une mise en veille (idle) ou d'un message entrant (in_sms) ou sortant (out_sms). Selon l'équipement tracé et l'installation retenue, les appels, messages et recherches ne sont pas tracés et n'apparaissent pas dans les logs.

- Duration précise la durée en secondes d'une action. Ce champ peut avoir pour valeur 0.
- Name qualifie l'action réalisée. En fonction de la classe, ce champ apporte des précisions sur l'action (nom de l'application ouverte ou contenu saisi au clavier ou url consultée).
- Date et time donnent des indications sur la période d'utilisation. Date est exprimé sur le format AAAA-MM-JJ et time HH :MM :SS.

Plateformes/outils utilisés: Kidlogger / base de données du proxy

Points forts	Points faibles

Stockage des données:

Les logs du proxy sont hébergés en premier lieu sur un serveur de la Région Poitou-Charentes. Ils sont envoyés au format *.CSV dans des fichiers dont les tailles varient entre 10 et 150 Mo. La structure de ces fichiers est la suivante:
datetime;displayed_user;ip;url;domain;protocol
De même, nous avons reçu un fichier *.XSLX, fourni par l'établissement scolaire contenant des informations personnelles de certains étudiants (participants à l'étude).
La nouvelle structure de données proposée divise explicitement les données du proxy de leur forme actuelle en plusieurs tables. Ceci est fait, premièrement, pour assurer le niveau d'anonymisation requis et, deuxièmement, pour faciliter l'analyse statistique.
En effet, il est nécessaire que les données sauvegardées soient dissociées de la personne les détenant. On aboutit à cet objectif en enlevant toute information qui pourrait conduire à une identification immédiate d'un individu.
En plus, si nous créons deux catégories pour les données: "fixes" et "statistiques" (données d'identification et, données à traiter, respectivement), nous pouvons réduire la charge de traitement informatique en limitant les enregistrements (non nécessaires) à retrouver lors d'une requête. Pour ceci, nous concevons une nouvelle structure de données comme un ensemble minimaliste d'informations, suffisamment indépendant afin d'être porteur de sens ou de contenu sans avoir recours à d'autres données".
L'import des données se fait de la manière suivante :

1. Créer un tableau maquette (test_proxy) comme suit: (6 colonnes et 4 indexes):

```
CREATE TABLE `test_proxy` (

		`datetime` TIMESTAMP NULL DEFAULT NULL,

		`username` VARCHAR(16) NULL DEFAULT NULL COLLATE
'utf8_general_ci',

		`ip` VARCHAR(16) NULL DEFAULT NULL COLLATE
'utf8_general_ci',

		`url` TEXT NULL COLLATE 'utf8_general_ci',

		`domain` VARCHAR(100) NULL DEFAULT NULL COLLATE
'utf8_general_ci',

		`protocol` VARCHAR(9) NULL DEFAULT NULL COLLATE
'utf8_general_ci',

		INDEX `datetime` (`datetime`),

		INDEX `user` (`datetime`, `username`, `ip`),

		INDEX `web` (`ip`, `url`(76), `domain`),

		FULLTEXT INDEX `url` (`url`)

)

COLLATE='utf8_unicode_ci'

ENGINE=MyISAM

;
```

2. Importer tous les fichiers CSV disponibles dans ce tableau récemment créé en utilisant ces paramètres (Laisser Heidi choisir l'encodage, normalement et pour ce cas-ci "latin1") :

```
LOAD DATA LOW_PRIORITY LOCAL INFILE 'C:\\Users\\ivan\\Desktop\\Kidlogger live SQL\\import\\Logs proxy 2015-2016\\URL_20160525102813.csv'

INTO TABLE `test`.`test_proxy`

FIELDS TERMINATED BY ';'

OPTIONALLY ENCLOSED BY ""

IGNORE 1 LINES

(`datetime`, `username`, `ip`, `url`, `domain`, `protocol`)

;
```

Si un fichier Excel (*.XLS ou *.XLSX) a été fourni à la place d'un CSV, il est nécessaire de l'exporter en CSV d'abord en utilisant les paramètres suivants :

- Character set : Latin1
- Field delimiter : ;
- Text delimiter : <none>
- Save cell contents
- Don't quote all text cells

Une fois tous les fichiers CSV dans notre tableau maquette, nous procédons au pré-traitement de ceux-ci.

Plateformes/outils utilisés: Base de données MySql moteur MariaDB et HeidiSQL

Points forts de ces plateformes	Points faibles
Prise en main rapide	

Description des pré-traitements:

Au niveau du proxy, en constatant que la longueur moyenne du champ URL est élevée (~129 caractères), nous avons pu conclure qu'une partie très considérable de ces adresses URL avait une origine "non humaine". En réalisant une simple analyse de répartition portant sur les longueurs des adresses URL on a rencontré que la plus commune était de 42 caractères avec plus de 800 000 cas, suivie par celle de 56 avec la moitié (~400 000 cas) et

ensuite 69 caractères avec encore la moitié (~200 000 cas). Ceci dans le contexte de 3738 cas différents (dont la plupart des premiers 60 ont une longueur inférieure à 100 caractères) nous indique que la longueur de l'URL n'est pas un facteur déterminant unique pour filtrer les pratiques "non-humaines".
Or, nous avons essayé de faire l'analyse des adresses URL en limitant leur longueur :

- Par un nombre de caractères fixes, par exemple 42 (la longueur la plus commune) ⇒ On 'perd' des informations avec un 'tronçonnage' de ce type mais on gagne en facilité de calcul.
- Par le nombre de répétitions d'un caractère spécifique de contrôle dans une URL, par exemple l'un de ces quatre: " . = / ⇒ On ne peut pas généraliser que tous les Adresses URL utilisent ces caractères de contrôle; il se peut que certaines adresses URL ne soient même pas limités du tout. Cette méthode est équivalente à une analyse sur le domaine de l'URL.

D'ailleurs, en espérant retrouver un modèle commun sur une colonne autre que 'url', nous avons essayé de travailler sur la colonne 'web_protocol'. Nous avons trouvé que le filtrage par ce champ ne permettait que de séparer notre échantillon en deux grands groupes d'Adresses URL; ceux qui avaient été visités en 'http' et ceux qui avaient été visités en 'https'. Cette information se trouvant déjà au début de l'adresse URL, cette colonne n'apporte aucune information complémentaire et est donc écartée.
D'après les inconvénients imposés par ces méthodes nous avons choisi de caractériser le comportement "non humain" de manière empirique en regardant de tout près les parcours quotidiens de quelques participants. Nous avons pris tous les enregistrements générés par les deux utilisateurs ayant le plus d'interactions et nous avons choisi de suivre ces interactions sur une même journée choisie aléatoirement (2015-11-26).
Après une toute première révision visuelle, nous avons constaté qu'une partie très importante des Adresses URL "non humains" :

- Renvoient vers des sites de publicité ou de traçage.
- Engendrent des cookies.
- Renvoient vers des fichiers qui ont des extensions qui, bien que nécessaires au fonctionnement correct des pages web, ne correspondent pas complètement à des interactions intentionnelles de l'utilisateur humain. Par exemple, *.css, *.js, *.json, *.ini, etc.
- Sont des téléchargements ou des requêtes d'actualisations système / antivirus.

Nous avons pu repérer dans cet ensemble d'Adresses URL certains mots-clés qui caractérisent ces premières conclusions partielles: metrics, pubmatic, meetrics, adserver, affs, adxanox, track, adsystem, googlesyndication, datacollect, vouchercodes, avastupdate, windowsupdate, adledge, googleads, btrll, parmi bien d'autres.
Cependant, le nombre d'interactions "non humaines" restait assez important. En effet, avec ce filtrage on réduisait uniquement de 100 (4043 par rapport à 4158) le nombre d'enregistrements "sales" et visiblement il restait encore des adresses URL "non humaines".
Pour caractériser plus précisément ces adresses URL, nous avons ensuite effectué une visite manuelle de toutes les adresses URL visitées par ces deux participants pendant cette journée et nous en avons créé une catégorisation comportant une description du 'résultat' trouvé au bout de cette adresse et une règle de caractérisation pour en trouver des

similaires.
Avec cette approche minutieuse de nettoyage, nous avons donc pu effectuer une comparaison entre les résultats “sales” (comprenant des Adresses URL “non humains”) et les résultats “propres” (ceux qui se dégageaient en appliquant les règles de filtrage) pour deux (ceux choisis initialement), trois, quatre utilisateurs (ceux qui avaient cardinalement le plus d’enregistrements) et finalement tous les utilisateurs pour la même journée.

Nombre de participants pris en compte	Nombre d’enregistrements “sales”	Nombre d’enregistrements “propres”	% “humain”
2	4158	526	2%
3	4287	604	4%
4	7373	2761	38%
161	102750	63448	38%

À partir des résultats représentants le pourcentage restant, nous avons pu constater qu’il était possible de reconstruire un parcours humainement réel des visites effectuées pour 2, 3 et 4 participants. De même, l’utilisation de la même liste de règles de nettoyage obtenue à partir de l’analyse de traces de deux participants a prouvé être très fiable pour le reste des participants.
Nous avons pu en dégager que, soit les pratiques des participants de tous les niveaux ont fortement été liées ce jour-là, soit la liste de règles trouvée élimine effectivement plus de 60% des comportements “non humains”. Des analyses successives sur les autres lots de données reçus ont confirmé l’effectivité de nos listes de nettoyage, atteignant des taux d’élimination d’enregistrements non humains de 65%. Cette élimination de données qui n’intéressent pas notre étude facilite énormément la définition des possibles profils d’utilisateurs mais surtout diminue le volume de données que nous devrons traiter. Ceci entraîne des temps de calcul plus courts et donc une réactivité importante lors de nos essais techniques et une disponibilité de résultats statistiques améliorée pour les analyses quantitatives successives.

Plateformes/outils utilisés: Base de données MySql moteur MariaDB et HeidiSQL

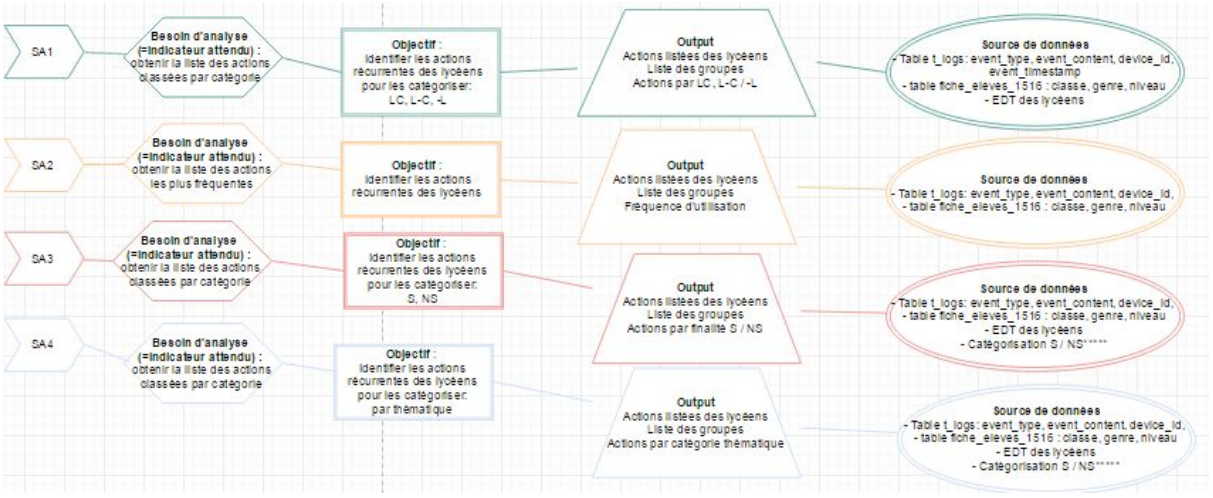
Points forts	Points faibles

Description des analyses : (Faire une description de chacune des analyses conduites)
Liste des variables : Nom et Description

Liste des méthodes mise en œuvre :
Mode opératoire technique, logiciels utilisés

Quatre scénarios d'analyse ont été testés, pour lister les actions des lycéens en fonction du :

- contexte [SA1]
- de la temporalité [SA2]
- de la nature d'intention [SA3]
- de la catégorie thématique [SA4]



Le [SA1] dépend du cadre spatio-temporel dans lequel est l'action enregistrée. Il est défini en croisant le marquage spatio-temporel (timestamp ou time dans les traces) avec l'emploi du temps des élèves. Trois variables sont retenues :

- "LC" pour les périodes au lycée de cours (du lundi au vendredi, sauf durant les vacances, de 09h à 17h00, en dehors des récréations et pauses)
- "L-C" pour les périodes au lycée en dehors des cours (du lundi au au vendredi, sauf durant les vacances, de 09h00 à 17h, en dehors des cours)
- "-L" pour les périodes en dehors du lycée (du lundi au vendredi avant 09h et après 17h, les samedis et dimanches, et les périodes de vacances)

Le [SA2] dépend de la temporalité dans lequel est l'action enregistrée. Il est défini en tenant compte de la durée de l'action enregistrée (exprimée en seconde) et de la fréquence de l'action par rapport aux autres actions de l'élève. Quatre variables sont retenues :

- "RC" pour les actions courtes (inférieures à la moyenne de la durée totale des actions d'un élève) et peu fréquentes (valeurs inférieures à la médiane de l'effectif total des actions d'un élève)
- "RL" pour les actions longues (supérieures à la moyenne de la durée totale des actions d'un élève) et peu fréquentes (valeurs inférieures à la médiane de l'effectif total des actions d'un élève)
- "SC" pour les actions courtes (inférieures à la moyenne de la durée totale des actions d'un élève) et fréquentes (valeurs supérieures à la médiane de l'effectif total des actions d'un élève)
- "SL" pour les actions longues (supérieures à la moyenne de la durée totale des actions d'un élève) et peu fréquentes (valeurs supérieures à la médiane de l'effectif total des actions d'un élève)

Le [SA3] dépend de la nature d'intention des actions des élèves. Elle est déterminée en tenant compte de la temporalité de l'action, de la catégorie de l'action et du contexte. Les deux natures ont été définies avec l'aide des élèves reçus en entretiens individuels ou collectifs. Deux variables sont retenues :

- actions réalisées pour des apprentissages scolaires
- actions réalisées pour des raisons personnelles.

Le [SA4] dépend de la nature thématique des actions des élèves. Cette nature thématique est obtenue par une méthode mixte, composée de quatre méthodes complémentaires :

1. les actions les plus fréquentes sont catégorisées manuellement en fonction d'une liste finie de catégories-type (communication/création/divertissement/recherche d'informations/transaction/surf).
2. dans le même temps, les url ou applications (anti-virus, publicités lancées en popup...) jugées non pertinentes pour l'étude sont identifiées et mises de côté
3. les actions les plus fréquentes sont catégorisées en interrogeant automatiquement par un script un service en ligne (Fortiguard web filtering service), en fonction d'une liste finie de catégories-types (voir la liste ici : http://help.fortinet.com/fos50hlp/54/Content/FortiOS/fortigate-security-profiles-54/Web_Filter/FortiGuard%20Web%20Filtering%20Service.htm#)
4. un regroupement des catégories proposées manuellement et automatiquement est réalisé pour aboutir à une nouvelle liste de catégories.

La liste des catégories retenues est la suivante, où seules les catégories sous-lignées sont conservées pour l'analyse :

tag	catégorie
1	Stockage
2	Outils de communication
3	Réseaux sociaux
4	Jeux
5	Consultation Audio ou Video
6	Megaportails
7	Transactions
8	Malware
9	Metric
10	Téléchargement
11	Organisation
12	ID_Navigateur
13	ID_Références
14	ID_Lifestyle
15	ID_Lecteur de documents
16	ID_Disciplinaire
17	ID_Autres info-doc

18	BOC_Bureautique
19	BOC_Création multimédia
20	BOC_Ressource disciplinaire
21	BOC_Autre outil de création
22	A_Flashplayer
23	A_Java
24	-
25	NOT URL, NOT APP

Note : ce scénario d'analyse est le prolongement du nettoyage réalisé sur les logs du proxy.
Les quatre scénarios décrits sont obtenus par des requêtes SQL (SELECT).
Résultats obtenus:

Points forts des analyses	Points faibles des analyses

Description des données produites au cours du traitement

Objectif de la création de ces nouvelles données :

Mode de calcul de ces variables :

Description des nouvelles variables : Nom et description

Nom	description

Description des Itérations

Pourquoi le processus d'analyse a été itéré ?

Points forts des itérations	Points faibles des itérations