Programme du datathon

JEUDI 5 SEPTEMBRE			
09:00	30 minutes	**Accueil des participants**	• Accueil café • Remplissage des fiches matrice de compétences
09:30	30 minutes	**Introduction du datathon**	• Ice-breaker : débat mouvant (les participants se positionnent de chaque côté d'une ligne suite à une série de questions) • Tour de table rapide (chacun se présente en 10 secondes - nom, prénom, occupation) • Présentation du fonctionnement d'un datathon et rappel du programme des deux jours • Point sur la documentation des projets (présentation des fiches de connaissance et du wiki) • Présentation du data pipeline
10:00	15 minutes	**Présentation des défis**	• Présentation courte des 3 défis par Marin, Annaigh et Lucie (5 minutes par défi)
10:15	95 minutes	**Brainstorming et constitution des défis**	• Rappel du contenu du défi par Marin, Annaigh ou Lucie • Brainstorming : chaque idée est notée au tableau • Vote avec des gommettes pour constituer les idées à partir des idées plébiscitées
12:00	90 minutes	**Déjeuner et café**	
13:30	220 minutes	**Développement en équipe**	Les équipes font avancer leur projet
17:10	20 minutes	**Rapide point d'avancement**	Chaque équipe présente en 2 minutes l'avancement de ses travaux. Possibilité d'une "visite" pour l'équipe OE qui ne participe pas mais voudrait voir le datathon.
17:30	60 minutes	**Développement en équipe**	Les équipes font avancer leur projet
VENDREDI 6 SEPTEMBRE			
09:00	30 minutes	**Accueil des participants**	Accueil café Reprise du travail de groupe
09:30	150 minutes	**Développement en équipe**	Les équipes font avancer leur projet Pause de 5 min pour "Pitch des pitchs"
12:00	90 minutes	**Déjeuner et café**	
13:30	180 minutes	**Développement en équipe**	Les équipes font avancer leur projet
16:30	60 minutes	**Présentation des projets et conclusion**	Présentation des projets en 5 minutes par groupe

OpenEdition **Datathon Read Write Cite**

Lire, Écrire, Citer : deux jours pour comprendre les usages à partir des données

Guide du participant

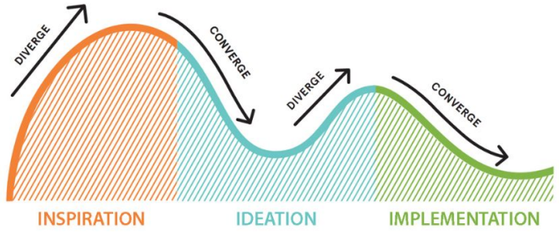
Qu'est-ce qu'un datathon ?

Un hackathon désigne tout événement de durée variable pendant lesquel des personnes très diverses se rassemblent pour résoudre des problèmes, classiquement en développant des outils informatiques mais pas nécessairement. Le datathon a pour particularité de partir de jeux de données mis à disposition des participants pour résoudre ces problèmes.

Le succès d'un datathon réside dans l'addition des compétences et des expertises qui produit un mélange étonnant fondé sur l'intelligence collective et le principe selon lequel « le tout vaut plus que la somme des parties. » Il repose aussi sur l'expérience : un datathon réussi doit d'abord à l'énergie positive qu'il dégage en évitant de dériver vers une culture de la compétition et des attentes trop élevées vis-à-vis des participants. Enfin, l'événement doit rester ouvert à celles et ceux qui n'ont jusqu'alors pas ou peu utilisé de données dans leur vie. C'est l'occasion de mettre les mains dans le camboui des données !

Comment se déroule un datathon ?

Même si chaque événement a ses particularités, on retrouve dans un datathon les trois temps du mouvement créatif définis par Tim Brown d'IDEO.



Inspiration

Cette première phase vise à définir le problème et
consiste à expliciter les défis pour s'assurer qu'ils soient bien compris par les participants. Le public doit pouvoir reformuler les défis proposés et développer une compréhension partagée des enjeux.

Idéation

La seconde phase consiste à générer une multitude d'idées autour du problème défini dans la phase précédente. Les participants sont incités à explorer de nouvelles solutions aux défis et, après un vote ou une phase de délibération, sélectionner les propositions pour constituer un nombre limité d'équipes.

Implémentation (prototypage)

Enfin, la troisième phase occupe la majeure partie de l'événement puisque les participants prototypent, voire développent, dans un temps limité, les visualisations de données ou services qui tentent de répondre aux défis.

Qui peut participer à un datathon ?

Tout le monde ! Un datathon ne doit pas être réservé aux habitué-e-s aux données. Réutiliser des données de manière pertinente, c'est un sport d'équipe ! Cela demande idéalement des compétences de design, de narration, d'analyse, de traitement et d'exploration des données. Personne n'a vraiment toutes ses compétences à la fois et c'est l'alliance des points de vue qui fait la richesse d'un tel évènement. **Nous comptons sur vous pour mélanger les compétences dans la constitution des équipes.**

*Ce travail a bénéficié d'une aide du gouvernement français au titre du Programme Investissements d'Avenir, Initiative d'Excellence d'Aix-Marseille Université - A*MIDEX. Nous remercions aussi chaleureusement le GIS URFIST qui a financé ce datathon.*

OpenEdition GIS Réseau URFIST A*Midex initiative d'excellence Aix-Marseille

Présentation des défis du datathon

Défi 1 - analyser la fréquentation des plateformes

Dans les analyses de l'impact de la science, l'acte de lecture a souvent disparu. L'impact est généralement réduit à la citation académique, mesurée avec des outils inadaptés que le mouvement des Altmetrics tente de réformer. Pourtant, autour des plateformes numériques d'édition comme OpenEdition, les données pullulent pour analyser l'activité de lecture. Deux sources principales peuvent être exploitées : les logs, des journaux sur lesquels un serveur enregistre l'ensemble de ses activités (des données souvent bruitées par les robots) et les trackers, des scripts qui envoient une information au serveur de collecte des données de fréquentation. S'appuyant sur ces sources massives de données, OpenEdition a développé un détecteur de lecteur inattendu pour identifier les cas où la diffusion en open access a favorisé la diffusion de productions scientifique en SHS au-delà du public habituel de la science, généralement des chercheurs dans la même discipline. Les équipes dans le cadre de ce défi tenteront de documenter les cas de lecteurs inattendus, de reconstituer des profils de visiteurs, de comprendre les usages à partir des données voire de créer des personas.

Défi 2 - Analyses textuelles sur les articles publiés sur Hypothèses

Les carnets de chercheurs d'Hypotheses.org constituent un corpus riche pour comprendre les pratiques communicationnelles et la mise en visibilité des chercheurs en SHS dans le nouvel écosystème scientifique numérique. Le projet HYCAR porté par le GIS Réseau URFIST propose deux axes d'analyse sur ce matériau qui constitueront la base de ce deuxième défi : savoir comment les chercheurs s'approprient ces outils et comprendre comment ces carnets s'articulent aux autres formes de présence numérique (archives ouvertes, réseaux sociaux, blogs de revues). Dans ce défi, nous cherchons à discerner différentes stratégies de publication, de distinguer des comportements types ou encore de comprendre l'articulation entre le carnet et les autres publications des chercheurs. Pour répondre à ces questions, nous pourrons nous appuyer sur des données riches avec le texte de l'ensemble des billets des carnets, l'index disciplinaire des carnets ainsi que la taxonomie des billets.

Défi 3 - Les citations dans et vers les plateformes d'OpenEdition

Les citations constituent aujourd'hui l'étalon de la mesure de l'impact des publications scientifiques. Cette approche est couramment critiquée, car elle réduit l'évaluation de la recherche à une seule et unique dimension. Toutefois, les données des citations peuvent servir à des usages très différents de la seule évaluation. Elles peuvent servir à connaître les points d'intersection entre les disciplines, retracer l'évolution des communautés de recherche, découvrir l'émergence de sujets ou encore comprendre les stratégies de carrière des chercheurs.

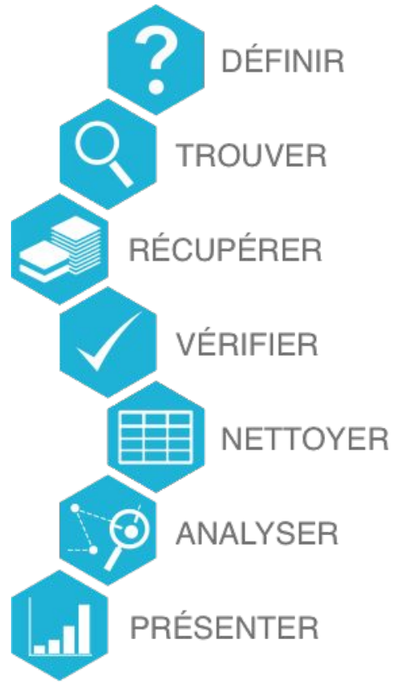
Pour répondre à ces questions, le projet OpenCitations propose près des 14 millions liens de citations entre des publications identifiées par un Crossref DOI. Si les citations académiques "d'articles à articles" constituent une valeur évaluative de la recherche et les indicateurs précisés ci-dessus, elles se restreignent au domaine universitaire. Les indicateurs regroupés sous le terme d'Almétrics intègrent à leurs mesures la présence de citations sur d'autres sphères sociales (réseaux sociaux numérique, site web...). Cependant, ces interfaces ne proposent qu'une lecture quantitative de ces pratiques. Nous proposons pour ce défi, d'explorer les contenus citant OpenEdition dans cette "littérature grise". Pour cela nous mettrons à disposition les données constituées pour le projet "Appropriation du savoir ouvert" (Lucie Loubère & Fidélia Ibekwe). Des sous ensembles de textes pourront vous êtres proposés pour permettre une exploration plus ciblée sur des pistes non explorées.

Comment réutiliser les données ?

Quelles sont les grandes étapes de la réutilisation d'un jeu de données ?

Le Data Pipeline développé par School of Data (schoolofdata.org/methodology) propose une méthologie en sept étapes qui décrit la réutilisation d'un jeu de données :

- **Définir** : trouver l'angle et définir l'objectif du travail réalisé avec les données
- **Trouver :** identifier les données à mobiliser
- **Récupérer :** obtenir les données dans un format exploitable par les machines
- **Vérifier :** s'assurer de l'intégrité et de la fiabilité des données
- **Nettoyer :** corriger les erreurs et adapter les données à la finalité
- **Analyser :** interpréter les données de manière itérative
- **Présenter** : raconter une histoire avec les données analysées et visualisées.



Où trouver les meilleurs visualisations de données pour inspiration ?

Data Journalism Awards (datajournalismawards.org/) : la sélection annuelle des meilleurs projets internationaux de visualisation de données

Flowing Data (https://flowingdata.com/) : un blog qui référence de manière très régulière les meilleures visualisations de donnée

Information is beautiful (https://informationisbeautiful.net/) : le blog de David McCandless, designer spécialisé dans la visualisation de données

The Functional Art (http://www.thefunctionalart.com/) : le blog du chercheur Alberto Cairo spécialisé dans la visualisation de données

Comment accéder aux données pendant le datathon ?

Vous trouverez sur vos tables pendant le datathon des fiches données qui décrivent :

- Production du corpus
- Contenu du corpus
- Processus de collecte
- Traitement des données
- Modalités de diffusion des données
- Considérations légales et éthiques
- Source des données.

Les fiches du catalogue de données sont aussi disponibles sur le wiki du datathon : https://datathon-openedition.frama.wiki/.