Conception d'un module de spécialisation en gestion des données de la recherche

Thomas DARTIGUEPEYROU
thomas.dartiguepeyrou@etu.hesge.ch
Etudiant
Haute Ecole de gestion de Genève

Lauréline GRANDJEAN
laureline.grandjean@etu.hesge.ch
Etudiante
Haute Ecole de Gestion de Genève

Résumé

C'est afin d'aider les chercheurs et les chercheuses suisses dans les changements liés à la numérisation qu'est né le projet Data Life-Cycle Management. En tant qu'étudiants du Master IS, notre participation à ce projet s'est faite sous la forme de la réalisation d'un pilote de module spécialisé sur le sujet de la gestion des données de la recherche du point de vue des types et formats, module s'insérant au sein d'un Massive Online Open Course (MOOC).

Nous nous sommes concentrés sur les besoins d'une communauté scientifique active en matière de gestion des données de la recherche (GDR). Nous avons déterminé que les besoins des participant-e-s de ce module étaient axés sur une activité quotidienne en gestion des données, tout autant que sur une activité de plus grande ampleur, comme la rédaction d'un plan de gestion des données (DMP) pour les bailleurs de fonds.

Notre module se divise en deux grandes parties : un handout regroupant les information théoriques et pratiques les plus denses, ainsi qu'une série de trois présentations, chacune suivie d'un quiz, assurant un apprentissage efficace. Cette division offre à notre public la possibilité d'acquérir toutes les connaissances dont il a besoin en types et formats des données de la recherche, mais aussi de cibler les informations précises qui lui manquent pour l'une ou l'autre de ses activités de gestion de données.

Le défi pédagogique d'un module s'insérant dans un MOOC est l'implication des participant-e-s, facilitant de fait leur apprentissage. Pour la rédaction de nos présentations, nous avons mobilisé des méthodes pédagogiques modernes telles que le Story Telling, le Nudge ou encore le Positive Reinforcement afin de créer un discours accessible auquel, nous l'espérons, nos utilisateurs-trices pourront s'identifier.

Nous avons souhaité apporter une contribution efficace et cohérente au projet DLCM en proposant un contenu utile, concret, simple de prise en main et facilitant les démarches des chercheurs et chercheuses suisses et étrangers.

Mots-clés

Données de la recherche, DLCM, MOOC, types, formats, enseignement en ligne

1. Types et formats des données de la recherche

Les données de la recherche comportent de nombreux aspects, c'est pourquoi il était tout d'abord important que nous dégagions les deux aspects sur lesquels notre pilote allait porter. Nous n'allions pas aborder la question de la nature (brute, dérivée, formatée, nettoyée, primaire, secondaire, traitée, etc), ni celle des supports (carnets de laboratoire, documents électroniques, papier, logiciels, programmes informatiques, etc) (*Cycle de vie et types de données* website) mais bien celle des **types** (archives, audio, vidéo, bases de données, codes sources, images, langages de programmation, matérielles et physiques, etc) et des **formats numériques** (CSV, TXT, TIF, MP4, etc.) des données de la recherche, bien que tous ces aspects fussent liés les uns aux autres. Afin qu'ils puissent, au besoin, approfondir ces différents aspects en dehors de notre module de formation, nous avons pris le parti de donner à nos participant-e-s des références utiles dans une bibliographie spécifique figurant à la fin de notre *handout*.

1.1. Les types de données de la recherche

Les types de données sont nombreux et hétéroclites. Dans une perspective de formation, il fallait choisir une manière de les présenter qui fasse sens pour les chercheurs, afin qu'ils comprennent et retiennent cet aspect de leurs données. Dans notre revue de littérature, une classification nous est apparue intéressante : il s'agit de la typologie produite par le *Research Information Network* britannique (A. Burnham 2012). On retrouve cette manière de classifier les types dans l'excellent *Fast Guide* de l'EPFL, sur lequel nous nous sommes basés pour créer notre module (Blumer et al. 2019). Cette typologie attribue à chaque type une méthode de collecte ou de production. Comme la méthode de collecte est ce qui préoccupe les chercheurs en premier lieu, réfléchir à celle-ci afin de déterminer le type de données nous a paru être un sens de réflexion somme toute assez naturel pour notre public-cible. Par exemple, les données capturées en temps réel sont des données de type observationnel, alors que les données créées en laboratoire sont des données expérimentales :

Tableau 1: Types des données de la recherche (André 2014)

Type de données	Méthode de collecte	Reproductible ?	Exemples
1. Données d'observation	Capturées en temps réel	NON	Mesure de la salinité d'un océan en un lieu précis, image de la voie lactée, mesure sismique…
2. Données expérimentales	Créées en laboratoire dans des contextes contrôlés	OUI	Séquences génétiques, chromatogrammes, analyses chimiques…
3. Données computationnelles,	Générées par des modèles informatiques		Modélisation du climat, ou de l'économie, ou des écoulements d'air

https://doi.org/10.55790/journals/ressi.2022.e807

de modèles ou de simulations			autour d'une aile d'avion…
4. Données dérivées ou compilées	Résultent d'un traitement, d'une sélection, d'une compilation ou d'une agrégation de données brutes	OUI, mais cela peut être onéreux	Données moyennes de températures, statistiques de populations…
5. Données de référence	Collectées, triées, agrégées puis publiées, elles servent d'éléments canoniques		Données décrivant les objets stellaires compilées dans les publications astronomiques, bases de données de séquences génétiques, archives photographiques, corpus textuels de référence…

1.1. Les formats de données de la recherche

Bien conscients que le format des données peut être imprimé, numérique ou encore physique (A. Burnham 2012), nous avons décidé, en accord avec les objectifs de GDR de la recherche du DLCM, de n'aborder que les formats numériques. Selon la définition canonique (*List of file formats* 2021), les formats numériques des données sont « une façon normalisée d'encoder des données afin de les stocker dans un fichier informatique ». Cette façon normalisée « suit un protocole qui spécifie comment les bits sont utilisés pour encoder l'information sur un support de stockage numérique ».

Les formats sont le deuxième aspect des données à considérer lorsqu'un-e chercheur-euse rédige un Data Management Plan (DMP). Le problème est que les formats sont extrêmement nombreux. La check-list pour la gestion des données mis à disposition par DLCM (*Data Management Plan : DLCM* website) liste cinq questions auxquelles les chercheurs doivent être attentifs pour les formats de leur fichiers. Pour notre module, nous en avons ressorti les points suivants :

1. Permet le partage et assure l'accès à long terme aux données
2. Est un format qui répond aux normes du champ disciplinaire
3. Est un format ouvert (non-propriétaire)
4. Lors de conversion d'un format en un autre format, la conversion ne perd ni n'altère aucune donnée.

Trois éléments importants sont à mettre ici en exergue : le respect des phases du cycle de vie des données, le respect des principes FAIR, la préservation des données. Pour le DMP, ces aspects assurent la description, la documentation & qualité, le stockage & sécurité, ainsi que le partage et la conservation à long terme des données (Data Management Plan (DMP) website). Il était donc important pour nous de trouver les grandes lignes dans lesquelles insérer

ces aspects liés aux formats des données de la recherche, afin de les rendre les plus compréhensibles possible aux participant-e-s du module, et surtout afin de faciliter leur choix de formats.

En premier lieu, c'est le type de donnée qui détermine la catégorie de formats dans laquelle puiser. En effet, si la donnée est du texte, il ne sera pas possible de l'encoder en format .mp3, par exemple. Ensuite, l'enjeu de taille est celui des principes FAIR : tous les formats ne répondent pas aux critères de ces principes, d'où l'importance de choisir les bons formats. Rappelons que le FNS invite les chercheurs à s'assurer que leurs travaux répondent aux principes FAIR. Dans l'acronyme, ce sont surtout les deux lettres du milieu qui sont concernées par les formats des données (Blumer et al. 2019) :

- Accessible : le format doit être ouvert (non-propriétaire)
- Interopérable : le format doit être interopérable parmi les différentes plateformes et applications.

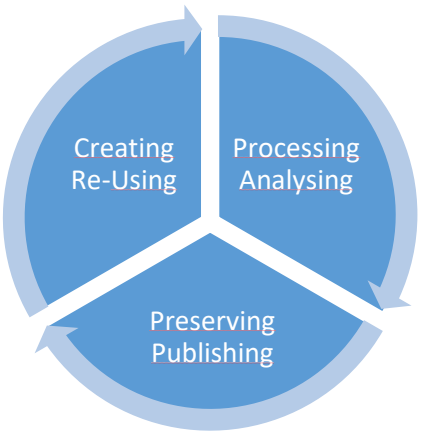
Tableau 2: Principes FAIR et Formats

Findable	**Accessible**	**Interoperable**	**Reusable**
Data and Metadata are easy to find by both Humans and computers.	Humans and computers can readily access or download datasets.	Data from different datasets are prepared to be combined or exchanged.	Published data can be easily combined or replicated in future research.
	Format must be compliant to an open, document standard	Format must be interoperable among diverse platforms and applications, fully published and available royalty-free, fully and independently implementable by multiple software providers on multiple platforms without any intellectual property.	

Enfin, le respect des phases du cycle de vie des données, ainsi que la préservation des données, sont deux enjeux à ne pas négliger. Il est important, lorsqu'on choisit un format adapté au type de ses données et qui répond aux principes FAIR, de s'assurer qu'il soit également adapté à la phase à laquelle on se trouve dans le cycle de vie. Ce sont les phases de collection/création/utilisation des données ainsi que de publication et de préservation qui sont concernées par les choix de formats (et les potentiels changements de formats).

https://doi.org/10.55790/journals/ressi.2022.e807

Figure 1: Cycle de vie des données et choix des formats



L'aspect des principes FAIR et celui des phases du cycle de vie des données fonctionnent bien ensemble puisque les formats qui répondent aux principes FAIR permettent une meilleure réutilisation et une meilleure préservation, en ceci qu'ils permettent de travailler sur différentes plateformes, de collaborer avec plus de personnes, d'éviter les problèmes de licence, ils maximisent la future réutilisabilité des données, et permettent d'être indépendant-e d'un logiciel ou d'une entreprise (Blumer et al. 2020). Ce sont ces avantages en particulier qu'il nous fallait nous assurer de transmettre aux participant-e-s de notre module afin de les éclairer, mais aussi de les inciter à faire le bon choix de formats pour encoder leurs données.

Nous avons enfin choisi un tableau complet des différents formats, classés du meilleur au moins bon, à présenter aux participant-e-s sur le *handout* de notre module. Nous nous sommes basés sur le guide de l'EPFL (Blumer et al. 2019) :

Tableau 3:Les formats des données de la recherche

Type of Data	**Appropriate**	**Acceptable**	**Depreciated**
Tabular (extensive Metadata)	CSV – HDF5	TXT – HTML – TEX – FASTQ – POR	
Tabular (minimal Metadata)	CSV – TAB – ODS – SQL – TSV	XML (if appropriate DTD) – XLSX	XLS – XLSB
Textual / Presentation	TXT – PDF – ODT – ODM – TEX – MD – HTM – XML – EXTXYZ – ODF	PPTX – RTF – DOCX – PDF (with embedded forms) – EPS – IPF	DOC – PPT – DVI – PS
Code / Computation	M – R – PY – IYPNB – RSTUDIO – RMD – NETCDF – AIML	SDD	MAT – RDATA

Image & Spectroscopy	TIF – PNG – SVG – JPEG – FITS	JCAMP – JPG – JP2 – TIF – TIFF – PDF – GIF – BMP – DM3 – OIR – LSM	INDD – AIT – PSD – SPC
Audio	FLAC – WAV – OGG – MXL – MIDI – MEI – HUMDRUM	MP3 – AIF	
Video	MP4 – MJ2 – AVI – MKV	OGM – MP4 – WEBM	WMV – MOV – QT
Geospatial	NETCDF – tabular GIS attribute data – SHP – SHX – DBF – PRJ – SBX – SBN – POSTGIS – TIF – TFW – GEOJSON	MDB – MIF	
3D Structures & Images	X3D – X3DV – X3DB – PDF3D – POV – PDBML	DWG – DXF – PDB	PXP
Generic	XML – JSON – RDF		

Ce tableau est trop complexe pour être affiché dans nos présentations mais il apparaît sur le support de la partie théorique de notre module, le handout, auquel nos participant-e-s peuvent se référer en tout temps.

2. Formation en ligne et cadre pédagogique

Les questionnements posés par la GDR se posent tout d'abord sur le fond, comme nous l'avons vu précédemment, par la mise en place de bonnes pratiques et de normes nouvelles afin d'établir un cadre efficace, mais également sur la forme. En effet, si la normalisation et la standardisation de ces pratiques a pour ambition, comme c'est le cas du projet DLCM, de s'étendre à l'ensemble d'un territoire (ici la Suisse), il devient obligatoire que leur enseignement suive la même voie. Il faut dès lors se questionner sur les moyens à mettre en place pour permettre cette uniformisation de l'enseignement en matière de GDR et créer un cadre pédagogique aussi pratique qu'efficace pour mener à bien cet objectif ambitieux.

Là aussi, la révolution numérique et les opportunités qu'elle représente jouent un rôle central. Grâce à elle, il est maintenant possible d'employer des outils capables d'atteindre une population cible aussi vaste et hétéroclite que celle des chercheurs suisses notamment grâce au médium choisi par le projet DLCM : l'enseignement en ligne sous la forme d'un MOOC.

2.1. L'enseignement en ligne

La Suisse n'est certes pas connue pour l'étendue de son territoire, pourtant en atteindre l'ensemble de la population chercheuse représente un défi de taille. En effet, notre pays

https://doi.org/10.55790/journals/ressi.2022.e807

compte de nombreuses universités, hautes-écoles, instituts etc… dispersé sur les 26 cantons d'un pays possédant 4 langues nationales, plus l'anglais, langue internationale et scientifique de référence.

Pour relever ce défi, le projet DLCM compte sur l'enseignement en ligne (Bari et al. 2020). Cette technique ayant déjà fait ses preuves ces vingt dernières années représente de nombreux avantage. Un gain de temps, tout d'abord, car elle permet de toucher de nombreuses personnes en même temps et évite ainsi la fastidieuse répétition de l'enseignement. Un gain d'espace, ensuite, puisqu'elle permet l'enseignement à distance et évite à l'enseignant-e et/ou au(x) participant-e(s) les déplacements inhérents à l'enseignement en présentiel. La crise sanitaire a accéléré l'importance qu'a pris l'enseignement en ligne car, malgré certaines problématiques qui y sont toujours liées et sur lesquelles nous reviendrons plus tard, il a permis à tous les élèves et étudiant-e-s de continuer à suivre leurs cours à un moment où la présence physique traditionnelle dans une salle de classe était devenue risquée, voire interdite.

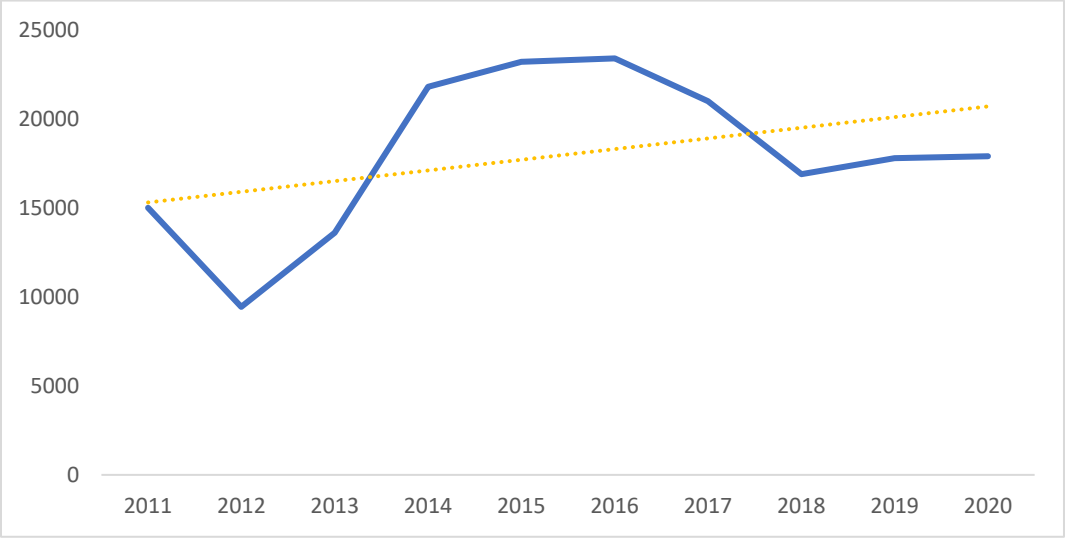
L'enseignement en ligne s'est aujourd'hui imposé comme un médium indispensable. Toute institution, enseignant-e mais aussi étudiant-e et chercheur-euse y est confronté-e. Il est donc parfaitement cohérent, au vu de ses importants avantages, que le projet DLCM l'utilise. Reste cependant à en définir la forme.

2.2. Le MOOC

L'enseignement en ligne offre d'indéniables atouts dans le cadre d'une formation visant à toucher un public largement dispersé et diversifié. Il existe cependant différentes formes d'enseignement en ligne. Il s'est agi dès lors pour nous d'identifier celle qui sera la plus performante en vue de l'objectif du projet DLCM.

Un Massive Online Open Course ou MOOC est un type d'enseignement en ligne consistant en une base pouvant prendre diverses formes et constituant un cours permanent et accessible à tout moment par une vaste population. Très en vogue, le MOOC est aujourd'hui à la pointe de l'enseignement en ligne et est la cible de nombreux articles scientifiques quant à ses qualités mais également ses défauts et comment enrichir ce dernier (Pfeiffer 2015).

Figure 2 : Occurrences du terme "MOOC" dans les articles scientifiques recensés par Google Scholar au cours des dix dernières années



Comme le montre ce graphique, le nombre d'articles scientifiques traitant du MOOC est en constante augmentation au cours de ces dix dernières années. Ce phénomène n'est pas surprenant compte tenu des importants avantages qu'apporte cette méthode d'enseignement en ligne.

Tout d'abord le MOOC est asynchrone, ce qui apporte un bénéfice immense dans le cadre du projet DLCM car cela signifie que tout-e chercheur-euse souhaitant se former sur la question des données de la recherche peut le faire depuis où il-elle se trouve, et à son rythme. Cette qualité permet deux choses : elle enlève de l'équation la question toujours complexe de la synchronisation entre enseignant-e et apprenant-e et, ensuite, elle diminue nettement l'aspect autoritaire et rébarbatif de la formation pour l'apprenant-e. Le MOOC ne s'impose plus comme une formation rigide et obligatoire mais se présente sous la forme d'une ressource, d'un soutien voire d'une solution à une problématique.

Ensuite, le MOOC s'inscrit dans une tendance plus large, riche de promesses et par conséquent chère au cœur de la communauté scientifique, celle de l'Open Acess. En effet, un autre bénéfice important de l'enseignement en ligne est la facilité de communication et de propagation qu'il apporte. L'Open Acess est le mouvement qui encourage cette propagation et dans lequel s'inscrit le projet DLCM puisqu'il s'agit d'ouvrir au maximum l'accès aux données afin d'accélérer les processus scientifiques et ainsi l'avancement des progrès qui peuvent en découler.

Le MOOC, par nature, est une ressource Open. S'il est possible d'en limiter l'accès, ce n'est pas l'idée ici puisque l'intention du projet DLCM est de viser le plus largement possible la communauté scientifique suisse mais aussi, pourquoi pas, étrangère. C'est dans cette perspective que le MOOC s'est imposé car ses qualités en font la meilleure arme possible en matière d'enseignement massif et standardisé. Toute personne intéressée par la GDR pourra accéder à un contenu libre, apportant des connaissances et prônant des normes et des pratiques approuvées par des spécialistes en la matière. Il sera ainsi possible de se former sur cette problématique générale, comme sur un point particulier, dans le temps et le contexte que l'on souhaite. Ce principe d'accessibilité élargie impose cependant une particularité à notre travail, celle de créer un contenu entièrement en anglais. Le but étant ici de viser le public le plus large possible, cette langue était la plus à même d'atteindre un maximum de personnes et surtout de limiter la discrimination liée à l'emploi d'une langue moins internationale.

L'enseignement en ligne et plus particulièrement le MOOC s'imposent comme le meilleur vecteur au vu des objectifs du projet DLCM. La souplesse, l'accessibilité et la disponibilité de ce médium sont sans aucun doutes ses plus grands atouts, il n'est donc pas surprenant qu'il soit particulièrement appuyé dans la littérature scientifique. Ces qualités ne doivent cependant pas effacer les difficultés qu'il fait naître, surtout en comparaison avec une méthode plus « traditionnelle » comme l'enseignement en présentiel.

2.3. Les méthodes pédagogiques

La dématérialisation de l'enseignement soulève principalement deux problématiques.

Tout d'abord, celle de l'interaction. L'apprenant-e n'ayant pas de répondant direct, ce dernier ne peut, par exemple, poser de questions en cas de doute. Ce souci a été au cœur de nos préoccupations. Le but étant de délivrer l'enseignement le plus efficace et précis possible, tout en essayant d'éviter l'emploi d'un format trop rébarbatif, trouver un équilibre entre exhaustivité

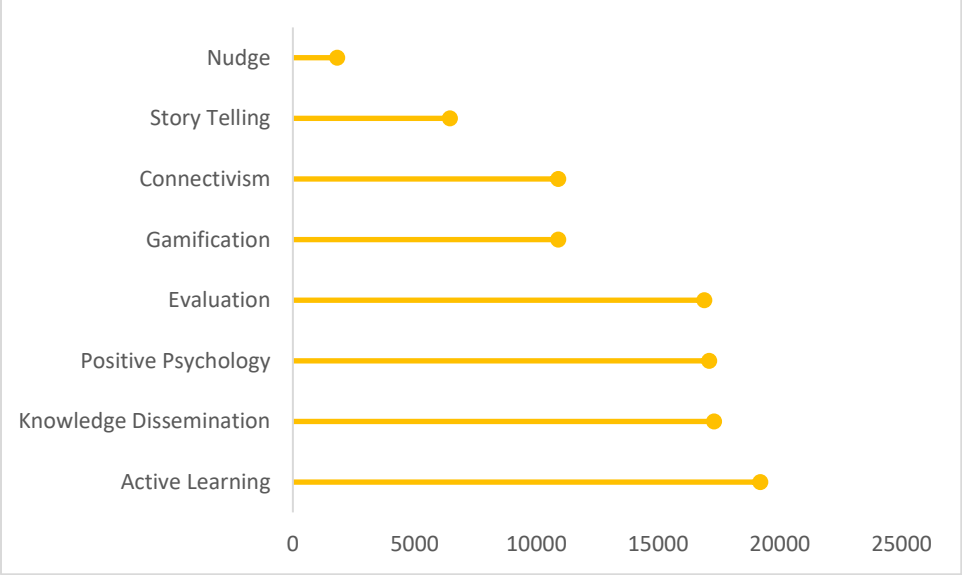
et légèreté a été un point crucial de la mise en place de notre module (Chauhan, Taneja, Goel 2015).

Afin de palier ce possible problème au mieux, nous avons adopté plusieurs réponses. Tout d'abord, nous avons divisé notre module en différentes parties, afin de cerner les différentes problématiques liées au types et formats des données de la recherche et ainsi éviter un long enseignement à un-e apprenant-e ne cherchant des informations que sur l'un ou l'autre des aspects du module.

Ensuite, nous avons divisé les sources d'informations en deux vecteurs. Nous avons mis l'accent sur les formes pédagogiques dans les animations commentées et nous avons réservé la masse d'information concrète dans un fichier à part appelé *handout* afin, d'une part, de ne pas submerger nos participant-e-s d'informations complexes pouvant apporter de la confusion dans une partie pédagogique, mais, d'autre part, de ne pas supprimer ces informations très utiles dans chacun des cas particuliers, dans une partie théorique.

En plus de la question de l'interaction, se pose celle du cadre pédagogique. Le MOOC est un sujet qui monte en intérêt au cours de ces dix dernières années. Si beaucoup d'articles scientifiques lui sont consacré, ce n'est pas uniquement pour en vanter ses mérites mais également, et surtout, afin d'identifier ses faiblesses et d'y proposer diverses possibles solutions. Comme nous l'avons relevé, la principale faiblesse du MOOC, et plus largement de beaucoup d'enseignements en ligne, est le manque d'interactivité qu'induit la dématérialisation de l'enseignement. Il nous est dès lors apparu comme crucial d'endiguer au maximum ce procédé afin de maintenir chez notre apprenant-e un véritable sentiment d'enseignement actif et de créer autant que faire se peut une appropriation du contenu dispensé. Cet objectif, déjà compliqué dans un enseignement présentiel, l'est encore bien plus au travers d'un MOOC. Nous nous sommes donc renseignés sur les récentes méthodes conseillées pour ce type d'apprentissage.

Figure 3 : Termes pédagogiques associés au terme "MOOC" dans les articles scientifiques recensés par Google Scholar au cours de ces dix dernières années



De nombreuses méthodes pédagogiques ont été associées au MOOC au cours des dix dernières années. Si toutes ne sont pas directement applicables à notre sujet, il était

intéressant pour nous de les recenser, dans un premier temps pour observer le ou les problèmes qu'elles font apparaître dans ce médium d'apprentissage mais aussi, bien évidemment, afin de s'en inspirer et rendre notre module aussi efficace que possible.

Ainsi nous pouvons constater que le principal problème identifié dans la littérature scientifique est effectivement lié à l'apprentissage actif, limité par le vecteur qu'est le MOOC. Il nous était donc essentiel de mettre l'expérience et la réalité de nos apprenant-e-s au centre de notre travail afin qu'ils-elles puissent réellement comprendre les atouts concrets de notre module de formation pour la recherche.

Pour ce faire, nous sommes servis d'un *story telling* appuyé (Robin 2016). Prendre des exemples concrets, les mettre en forme dans un contexte crédible, soit celui de la rédaction d'un *Data Management Plan* (DMP), élément essentiel à l'obtention de soutiens financiers, nous est apparu comme un puissant moyen d'identification des apprenant-e-s à des personnages fictifs vivant une réalité qui leur est familière.

Cette façon de mettre en avant les bénéfices de bonnes pratiques normalisées en matière de GDR porte un nom : le *Nudge* (Wilde 2016). Il s'agit d'une théorie récente et très en vogue ces dernières années, qui se base sur le renforcement positif d'un comportement que l'on souhaite faire adopter à une population ciblée. Cette méthode nous est apparue ici particulièrement pertinente puisqu'il s'agit de faire adopter un comportement possiblement nouveau et surtout vertueux à un public dont le principal centre d'intérêts demeure bel et bien la recherche, et non la gestion de données (GDR). En s'appuyant sur des élément concrets et réels tels que la rédaction d'un DMP et la visibilité offerte par l'*Open Access* à la base de la GDR prônée par le projet DLCM, nous avons voulu mettre en avant les bienfaits des pratiques que nous mettons en exergue dans notre module et ainsi emporter l'adhésion de notre public qui en comprendra très pragmatiquement l'intérêt.

D'autres méthodes ont également attiré notre attention et nous ont incités à créer un ton léger et un enseignement le plus concis et précis possible. Nous avons voulu insérer un renforcement positif et une imagerie agréable afin de rendre un sujet aisément fastidieux plus agréable et joyeux, dans la ligne de la *positive psychology*. En mettant en avant les bénéfices de la formation et en présentant un sujet complexe comme facilement abordable en gardant des exemples simples et concret ainsi qu'une présentation animée et joyeuse, notre ambition fut de convaincre notre public de leur capacité à atteindre les objectifs fixés par le projet DLCM en matière de GDR.

Ces différentes méthodes pédagogiques n'ont bien entendu pas pu être entièrement intégrées dans notre module. Cependant, leur étude a été d'un apport indéniable et nous ont permis de fixer un cadre et une direction pédagogique précis à notre projet. Ce cadre est constitué de deux axes. Tout d'abord il nous fallait susciter l'intérêt de nos participant-e-s notamment en présentant l'intérêt de notre formation par des exemples concrets, précis et cohérents vis-à-vis de la réalité des chercheurs-euses suisses (Raghuveer et al. 2014). En mettant en avant l'importance d'un DMP bien conçu dans l'obtention de fonds, ainsi que la visibilité offerte par l'*Open Access*, nous nous adressons directement aux préoccupations quotidiennes d'un-e chercheur-euse en leur offrant des outils utiles à leur carrière.

Ensuite, la division des sources d'information nous offre un atout de taille : celui de garder un enseignement léger et concis et, ainsi, de ne pas perdre nos interlocuteurs dans de rébarbatifs détails techniques. La présentation que nous avons créée u un but purement pédagogique et enseigne une méthode, une réflexion qui doit accompagner les chercheurs lors de toute

nouvelle recherche, au moment de la création d'un DMP. Le *handout*, en revanche, fournit des informations techniques plus denses et apporte des éléments théoriques (définitions, listes, schémas…) facilement identifiables dès lors que le message pédagogique de la présentation a été assimilé par l'apprenant-e.

Ce cadre pédagogique fut au cœur de notre projet car un sujet aussi technique que la GDR peut aisément être indigeste. Aussi nous avons souhaité proposer un module simple et concis, dans l'objectif de toucher le plus largement possible le public cible, tout en étant suffisamment précis et complet pour être utile à tout-e chercheur-euse souhaitant se former à la GDR.

3. Résultats

À la lumière des recherches que nous avons effectuées, tant quant aux connaissances sur les types et les formats que nous devions transmettre à nos participant-e-s, qu'aux concepts pédagogiques et la forme à adopter, nous avons voulu créer un module de spécialisation qui soit correct, clair et utile mais aussi facile à suivre et dont les avantages seraient mis en valeur.

Nous nous sommes tournés vers de la micro-training et avons réalisé trois petites présentations autonomes, néanmoins liées par un ordre progressif. Cela nous a permis à la fois de conserver la précision de nos points (types et formats des données) tout en rendant notre propos plus digeste, adaptable et ludique. À la fin de chaque présentation un petit quiz est proposé aux participant-e-s. Ces quiz offrent à chaque question un choix de réponses dont l'une seule est correcte. Cela permet aux participant-e-s, après avoir été mis en situation relativement passive où ils-elles devaient écouter, comprendre et retenir les informations délivrées, de se tourner vers une activité leur permettant de vérifier si ces informations avaient en effet été comprises et retenues. Enfin, ces trois présentations sont soutenues par un handout sur lequel les définitions, les schémas, les tableaux et les points à retenir sont notés afin que les participant-e-s puissent les lire à leur rythme et les reprendre à tout moment. En outre, les participant-e-s trouvent dans cet exemplier la bibliographie des présentations, mais aussi une bibliographie comportant des sources à explorer pour approfondir un point présenté, ou un point connexe.

Les trois présentations consistent en un support, qui comporte des slides et des animations, ainsi qu'une voix-off. Les aspects visuels des animations se combinent avec l'aspect sonore de la voix, afin de garder l'attention du ou de la participante durant les quelques minutes que dure la présentation.

La première présentation s'intitule « Terminology » et dure 3 minutes 16 secondes. Elle a pour but d'introduire les participant-e-s aux termes et concepts mobilisés. Elle reste théorique, moins tournée vers le story telling que les deux présentations suivantes, et a pour but d'offrir une vue globale du module. Cette présentation se compose de quatre points :

1. Les données de la recherche : il était important de définir ce sur quoi les types et les formats portent. Pour apporter un aspect ludique à la définition, et conserver un angle de présentation concentré sur le concret des participant-e-s, nous avons commencé par énumérer quelques exemples (lab notebook, peer reviews, physical objects…) qui ne sont pas des données de la recherche. La définition que nous en avons ensuite donnée se retrouve sur le handout en support de présentation.

2. Les types des données de la recherche : l'idée ici étant d'introduire nos participant-e-s à la notion, nous n'avons qu'énuméré les cinq types tout en les liant au schéma du cycle de vie des données afin d'opérer la connexion visuelle entre les types et la phase de création/collecte.
3. Les formats des données de la recherche : de la même manière que pour les types, la notion de format est introduite. Cette fois, nous avons voulu opérer plusieurs étapes : à la suite d'une définition canonique (qu'on retrouve sur le handout), apparaît un exemple de la façon dont se présente un format (nom et extension), puis le lien avec les principes FAIR, ainsi que les deux phases du cycle de vie des données (création/collecte, et préservation/publication), faisant le lien avec la phase du cycle de vie présentée dans la slide précédente.
4. La dernière slide annonce les deux présentations suivantes en présentant déjà le premier personnage (Alex), afin de susciter la curiosité de nos participant-e-s. L'objectif ici est d'expliquer le déroulement de notre module, mais aussi de montrer à nos participant-e-s son utilité (comprendre l'importance des types, être capable de les identifier, choisir son format en accord avec le cycle de vie de ses données et les principes FAIR).

La deuxième présentation dure 2 minutes et porte sur les types des données de la recherche. Elle a pour objectif de mieux faire comprendre à nos participant-e-s ce qu'est un type et surtout, comment bien l'identifier. Avec cette présentation, nos deux personnages (Alex et Natacha) entrent en scène, comme opportunité d'offrir des exemples concrets de chercheuses confrontées aux types et aux formats de leurs données, dans le cadre de la rédaction d'un DMP.

1. Dans la première slide, nous présentons nos deux personnages : ce sont deux chercheuses suisses qui doivent rédiger un DMP afin d'obtenir des fonds pour mener à bien leurs recherches. Cette situation fictive très réaliste a pour but d'accrocher notre public. Nous ne nous adressons à lui que par oral. Aucun texte ne se trouve sur cette première slide afin de servir notre méthode pédagogique. En effet, l'idée est ici de créer un puissant Story telling. En supprimant le texte ou tout autre élément externe et superflu à notre diégèse, nous créons un instant de repos pour l'apprenant-e qui dès lors peut se laisser entraîner dans l'histoire qui lui est racontée, et s'y consacrer totalement. Nous espérons, par ce procédé, renforcer l'identification de notre public avec nos personnages, créer un effet cathartique afin de nous assurer du soutien envers ceux-ci. Etant confrontés aux mêmes types de problèmes, notre public souhaite voir nos personnages trouver les solutions qu'eux-mêmes sont en train de chercher. De plus, des ressorts attrayants du Nudge sont également mis en place par l'explication des bienfaits concrets (soutien financier) de la rédaction d'un bon DMP.
2. La deuxième slide concerne Alex : entomologiste, elle travaille sur le son des cigales en Afrique du Sud. Par nature, ses données sont donc observationnelles et non reproductibles. L'exemple d'Alex sert ici de support à notre message. Tout d'abord le texte audio accompagnant cette slide et relatant le choix d'Alex quant à l'identification de type de ses données permet, sans sortir de la diégèse créée par notre exemple, de transmettre à notre public les bonnes questions à se poser afin d'arriver à un résultat correct : de quelles natures sont mes données, à quel type cela correspond et enfin quelle quantité de données prévois-je ? Exposer un exemple concret permet d'éviter

https://doi.org/10.55790/journals/ressi.2022.e807

l'aspect rébarbatif et le trop plein d'informations d'un tableau complet relatant tous les différents types de données (tableau que l'on trouve dans le handout) et maintient notre public dans un Story telling concret et efficace.

3. Il en va de même pour notre second exemple, celui de Natacha. Ce second exemple permet de renforcer nos concepts pédagogiques en créant un effet de répétition qui marque la logique de notre méthode d'identification des types de données de la recherche (par la méthode de collecte/création), la rendant par là-même plus claire qu'avec un seul exemple. De plus, Natacha s'inscrit dans un domaine scientifique totalement différent de celui d'Alex. En effet, Natacha travaille sur la littérature romantique suisse. Ses ressources et données sont littéraires, différentes de celles d'Alex qui est issue d'une branche de sciences empiriques. L'exemple de Natacha permet donc un plus large spectre d'identification au sein de notre public. De plus, il montre aussi que la rédaction d'un DMP (et donc la formation dispensée ici) est utile à tout type de chercheur-euse, provenant de tout type de discipline.
4. La fin de cette présentation adopte un aspect bien plus théorique. C'est cependant en connaissance de cause que nous sortons, à ce moment, de notre Story telling. Alex et Natacha ayant rempli leurs rôles (identification/exemplification, incitation douce), nous concluons cette présentation en présentant trois points importants des types de données : identifier le type de données est une condition nécessaire pour pouvoir choisir un format adéquat, identifier le type de données ne peut se faire sans savoir où on se trouve dans le cycle de vie des données, et enfin il est important de re évaluer le type et surtout les formats de ses données selon où on en est dans le cycle de vie.

La troisième et dernière présentation de notre module dure 4 minutes 12 secondes porte sur les formats des données, avec Alex et Natacha qui assurent la continuité de la narration. Dans le chapitre des formats se trouvent deux bonnes pratiques vers lesquelles nous avions pour but de diriger nos participant-e-s, celle de choisir un format adapté au cycle de vie des données concernées ainsi que celle de choisir un format qui répond aux critères des principes FAIR.

1. Sur le même modèle que la première slide de la précédente présentation, cette slide ne comporte que très peu de texte : elle a simplement pour fonction d'accrocher l'attention du ou de la participante, et d'introduire Alex et Natacha en rappelant le type de données qui leur incombe (observationnelles et de référence). Ensuite, la suite immédiate de la présentation est annoncée oralement : Alex et Natacha vont devoir choisir le bon format pour encoder leurs données, dont la première étape est de choisir un format en adéquation avec les principes FAIR.
2. La connaissance de ces principes est nécessaire à la compréhension de ce qu'est un bon ou un mauvais choix de format. Nous avons ainsi explicité l'acronyme, en gardant un visuel simple afin d'inciter le ou la participante à se concentrer sur les explications données à l'oral (il ou elle peut se référer aux tableaux et définition sur son handout). En lien avec la théorie du Nudge, nous énonçons également les bienfaits que le respect de ces principes peut lui apporter (comme le fait de faciliter la visibilité de ses recherches). En soulignant les avantages des principes F.A.I.R et plus largement de l'Open Acess, voulons renforcer l'adhésion de nos participant-e-s en tant qu'individus mais également en tant que membres de la communauté scientifique. Cette idée est soutenue par la présence graphique de nos deux personnages, ensemble sur la même slide.

https://doi.org/10.55790/journals/ressi.2022.e807

3. Les types de données d'Alex et de Natacha sont mis en corrélation avec les différents formats qui pourraient leur convenir, afin d'exemplifier concrètement ce qu'un choix de format représente à ce stade de leur travail. Les participant-e-s sont à ce moment invités à se référer au tableau complet des différents formats sur leur handout pour en prendre connaissance à leur rythme. L'importance des principes FAIR est rappelée, et le respect des phases du cycle de vie des données est introduit comme deuxième point à considérer lors du choix d'un format, en expliquant que le format doit être re évalué à chaque phase de vie de la donnée. Il s'agit ici de s'assurer que le format est viable pour la création/collecte de données, ou capable d'assurer la préservation/publication de données. Ces points sont mentionnés à l'oral, afin de ne pas surcharger visuellement la slide, sachant qu'ils sont listés dans le handout pour être utilisés comme rappels lors d'une situation concrète de choix de format.
4. En cette fin de module, nous nous sommes permis une slide bien plus chargée que les précédentes. L'idée est ici d'offrir un support plus marqué à notre public en revenant sur les apprentissages dispensés sous forme d'une liste de questions. Si toutes ces questions trouvent une réponse, alors le format choisi est correct. Cette slide permet donc au ou à la participante de vérifier que les données qu'il ou elle possède déjà sont bien encodées et lui permet aussi de faire un bilan sur ce qu'il ou elle est supposé-e maîtriser grâce à cette formation. Ces questions doivent à présent faire partie de ses réflexes lors de ses futures collectes de données et l'accompagner à tout moment dans son travail, tels les nuages encadrant visuellement nos personnages. Enfin, les participant-e-s sont encouragé-es à ajouter des questions utiles dans leurs domaines respectifs, afin de créer un véritable pense-bête à l'aide de leur handout.
5. Par renforcement de nos méthodes pédagogiques (Positive psychology, Nudge, Story telling) nous terminons nos présentations avec nos deux personnages, heureuses d'avoir maîtrisé l'encodage de leurs données. Elles ont bien identifié les types de leurs données et choisi d'excellents formats, ainsi elles ont obtenu le soutien financier nécessaire à leurs activités de chercheuses, et s'assurent de mener leur travail de manière qu'il soit visible et profite au progrès de leur discipline. Nos personnages sont bien évidemment heureux, leur mission est accomplie tout comme, nous l'espérons, sera celle de nos apprenant-e-s.

4. Conclusion

L'identification des connaissances utiles aux chercheurs-euses en matière de types et de formats de données de la recherche, ainsi que le choix de méthodes pédagogiques pertinentes pour notre module nous ont permis d'aboutir à la création d'un pilote de formation que nous espérons adéquat, tant pour le projet DLCM que pour notre public-cible.

L'importance de l'identification du type des données par la méthode de collecte/création était à souligner afin que le choix du bon format soit le plus aisé possible. En outre il était primordial de faire comprendre aux participant-e-s les avantages à respecter les principes FAIR ainsi qu'à observer les phases du cycle de vie des données, tant à un niveau individuel que pour la communauté scientifique dans son entier. De cette manière, nous espérons les inciter à observer de bonnes pratiques.

Ces connaissances et ces bonnes pratiques n'auraient pas pu être transmises sans une bonne compréhension de notre part du MOOC et des défis liés à l'enseignement en ligne. La *Positive psychology*, le *Story telling* et le *Nudge* ont été à la base des décisions que nous avons prises lors de la construction de notre module. Nous avons ainsi pu concevoir trois petites présentations sous forme de slides animées et commentées par une voix-off, présentations que nous avons voulues simples, capable d'éveiller l'intérêt et surtout claires. La narration de l'histoire de deux chercheuses permettant l'identification de nos participant-e-s et offrant l'opportunité de présenter des exemples concrets s'avère un bon moyen, selon nous, de délivrer les connaissances de façon claire, ludique et incitative. Enfin, diviser les sources d'information sur deux supports (présentations et *handout*) nous a permis de nous concentrer sur l'aspect pédagogique, tout en ne négligeant jamais l'apport théorique indispensable à la complétude des connaissances que nous devions apporter à nos apprenant-e-s.

Bibliographie

A. BURNHAM, 2012. Research Data - Definitions. 2012. P. 5.

BARI, Manon, BEZZI, Manuela, GUIRLET, Marielle et MAKHLOUF-SHABOU, Basma (dir.), 2020. *Formation et éducation en gestion des données de recherche du point de vue du projet DLCM : dispositifs d'e-learning* [en ligne]. , TRMASID 25. Haute école de gestion de Genève. [Consulté le 25 avril 2021]. Consulté à l'adresse : https://doc.rero.ch/record/328462

BLUMER, Eliane Ninfa, CHAPTINEL, Jérôme Julien, MASSON, Antoine, REICHLER, Fantin, SAMATH, Sitthida, VARRATO, Francesco et MILFORT, Frank, 2019. EPFL Library Research Data Management Fastguides. [en ligne]. 11 février 2019. [Consulté le 26 avril 2021]. Consulté à l'adresse : https://zenodo.org/record/3327830#.YIaDvaE682w

BLUMER, Eliane, SAMATH, Sitthida, VARRATO, Francesco et BOREL, Alain, 2020. Optimizing your research data management. [en ligne]. 28 avril 2020. [Consulté le 26 avril 2021]. DOI 10.5281/zenodo.3773657. Consulté à l'adresse: https://zenodo.org/record/3773657

CHAUHAN, Jyoti, TANEJA, Shilpi et GOEL, Anita, 2015. Enhancing MOOC with Augmented Reality, Adaptive Learning and Gamification. In : *2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)*. Octobre 2015. p. 348-353.

Cycle de vie et types de données, sans date. [en ligne]. [Consulté le 25 avril 2021]. Consulté à l'adresse : https://www.unil.ch/openscience/fr/home/menuinst/open-research-data/les-donnees-de-recherche/cycle-de-vie-et-types-de-donnees.html

Data Management Plan : DLCM, sans date. [en ligne]. [Consulté le 13 janvier 2022]. Consulté à l'adresse : https://www.dlcm.ch/resources/dlcm-dmp

Data Management Plan (DMP), sans date. [en ligne]. [Consulté le 14 janvier 2022]. Consulté à l'adresse : https://www.unil.ch/openscience/fr/home/menuinst/open-research-data/gerer-les-donnees-de-recherche-research-data-management/data-management-plan-dmp.html

List of file formats, 2021. *Wikipedia* [en ligne]. [Consulté le 9 décembre 2021]. Consulté à l'adresse : https://en.wikipedia.org/w/index.php?title=List_of_file_formats&oldid=1058510989

PFEIFFER, Laetitia, 2015. *MOOC, COOC : La formation professionnelle à l'ère du digital*. Dunod. ISBN 978-2-10-072972-2.

RAGHUVEER, V. R., TRIPATHY, B. K., SINGH, Taranveer et KHANNA, Saarthak, 2014. Reinforcement learning approach towards effective content recommendation in MOOC environments. In: *2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE)* [en ligne]. Patiala, India: IEEE. Décembre 2014. p. 285-289. [Consulté le 14 janvier 2022]. ISBN 978-1-4799-6876-3. Consulté à l'adresse: http://ieeexplore.ieee.org/document/7020289/

ROBIN, Bernard R., 2016. The Power of Digital Storytelling to Support Teaching and Learning. *Digital Education Review*. 15 décembre 2016. P. 17-29. DOI 10.1344/der.2016.30.17-29.

WILDE, Adriana, 2016. Understanding persuasive technologies to improve completion rates in MOOCs. In: [en ligne]. 7 juin 2016. [Consulté le 14 janvier 2022]. Consulté à l'adresse: https://research-repository.st-andrews.ac.uk/handle/10023/12206

5. Tables des figures et tableaux