Data littératie & SHS :
développer des compétences pour l'analyse de données

Béa **ARRUABARRENA**, Gérald **KEMBELLEC**

Ghislaine **CHARTRON**

Laboratoire Dicen-Idf – CNAM

Codata – 14 & 15 mars 2019

DicenIDF

Dispositifs d'Information et de Communication à l'Ère Numérique - Paris Ile de France

le cnam

Contexte

- **Mégadonnées :** essor de l'analyse de données massives
- **Analyse de données :** élargissement à tous les domaines d'activité, y compris à ceux des SHS
- **Emergence de nouveaux profils :** compétences pour produire, collecter, analyser, comprendre, gérer et utiliser les données
- **Vision transverse des données** tournée vers l'usage social: dimensions STI/dimensions SHS

Enjeux SHS

- **Médiation Homme-Données** : les interfaces pour la compréhension des données, les tableaux de bord, les *reportings*, l'éthique, la valeur contextuelle des données…
- **Qualité des données :** filtrage, sélection par rapport aux besoins métiers, référentiels, cohérence, alignements….
- **Pilotage d'un projet Data :** en liens avec les usages de plus en plus prioritaires

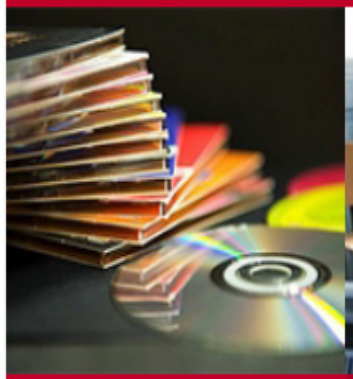
Objectifs & Méthodes

Objectif principal :

Appréhender les compétences requises pour un *Data analyst*, et ce à tous les niveaux de la chaîne de valeur de la donnée

Méthodes :

- Un état de l'art sur le concept de Data littératie : les enjeux liés à l'acculturation aux données
- Analyse à l'appui du référentiel de compétences élaboré pour le Master « Mégadonnées et Analyse sociale » (MéDAS)



Relever les défis de la société de l'information.

MOOC

Formations diplômantes

Titre RNCP I Chef de projet en Ingénierie documentaire

- Master Medas

LP Documentation audiovisuelle/Archives orales et audiovisuelles

LP Documentaliste d'entreprise et métiers de l'Infodoc

Formations certifiantes

Master Sciences humaines et sociales mention humanités numériques Parcours Mégadonnées et analyse sociale (MEDAS)

PRÉSENTATION | PROGRAMME | COMPÉTENCES ET DÉBOUCHÉS | INFORMATIONS PRATIQUES

Publics / conditions d'accès

Formation en alternance :

- Accessible aux étudiants titulaires d'un bac+3 ou d'un titre validant 180 ECTS en sciences

Code diplôme/certificat: MR09500A

120 crédits

Niveau d'entrée

Niveau II (bac+3 bac+4)

Niveau de sortie

Niveau I (bac+5 et plus)

Responsable national

Ghislaine CHARTRON

Responsable opérationnel

Spécificités du Master

- **Master CNAM :** mention Humanités numériques - Habilitation en août 2015/ouverture en 2015
- **Modalité :** en alternance uniquement, CFA du CNAM (en lien avec les entreprises)
- **Accès :** après une licence de SHS ou de STI
- **Spécialisation:** analyse de données
- **Profil :** *Data Analyst, Data scientist*

Data littéracie : état de l'art

- « La capacité à comprendre et à utiliser les données de manière effective **pour la prise de décision** » Mandinach et Gummer (2013) :
- Importance des compétences **mathématiques, informatiques et statistiques** (Wolff et al., 2016).
- Les compétences de base d'une Data littératie « **ne peuvent être considérées de manière isolée** » (Matthew, 2016).

Data littéracie : état de l'art

Compétences transverses basées sur les SHS :

- Epistémologie du *Big data*
- Cadre juridique
- Pilotage spécifique de projets (Shearer, 2000 ; Marr, 2015)
- Gouvernance, la qualité (Arruabarrena, 2018, Koltay, 2016)
- Ethique des données (Matthew, 2016 ; Calzada-Prado et Marzal, 2013)
- Approche critique : développer une culture de la donnée, i.e. la capacité à évaluer de manière éthique et critique les données (Carlson et al., 2011)

Data littéracie : état de l'art

Data littéracie & Information littéracie

- **Maîtrise des données** renvoie à une compétence transverse : la littératie informationnelle (*Information literacy*)
- **Maitrise de l'information :** associée aux usages, au contexte – (i.e. aux connaissances du domaine étudié : les domaines sectoriels et métiers, la stratégie de l'entreprise, etc.)

Data littéracie : état de l'art

Data littéracie & Analyse de données

- Appréhension des possibilité et des limites de l'analyse de données (datamining)
  - Les biais liés à l'analyse de données massives
  - Les « méga-erreurs » (Liu et al., 2016)
- Savoir différencier et associer les approches statistiques/analytiques et exploratoires
- Importance de la qualité des données

Data littéracie : état de l'art

Data littéracie et datavisualisation

- Maitrise du traitement visuel des données (Littéracie visuelle)
- Capacité à communiquer ses analyses
- Impact stratégique : Interprétation et diffusion des analyses
- Sémiologie graphique (Bertin, 1967)

Compétences ciblées

- Savoir développer des stratégies data à partir des données dans des contextes variés.
- Gérer et manager des projets de données articulés avec des besoins métier et des besoins sociétaux.
- Savoir traiter des données quantitatives et/ou qualitatives
- Mettre en place des solutions de fouille de données.
- Participer à des projets de Data sciences (description, prédiction, prescription)

Compétence n°1 :

Collecter et préparer des données

- **Requêter des bases de données**, en exporter les données pour constituer son corpus ;
- ***Scraping* :** développer des scripts spécifiques (python par exemple) ;
- **Modéliser :** S'approprier les méthodes de modélisations de données structurées (Merise, UML) et non structurées (NoSQL), ainsi que leurs encodages (utf-8, iso-8859-1…) et formats (SQL, JSON, CSV, XML...) afin de pouvoir les extraire et les réexploiter ;
- **Nettoyer/Préparer**
  -Normaliser les données, pouvoir traiter des ensembles de données cohérents
  -Encoder, transcoder des fichiers dans divers formats (Excel, CSV, JSON)
  -Réconcilier les données en regroupant des jeux issus de différentes sources et en les enrichissant à partir de référentiels d'autorités.

Compétence n°2 :

Analyse des données pour produire des informations, des indicateurs utiles aux organisations, à des individus.

- Analyser des données via des **méthodes statistiques** dans une démarche descriptive ou analytique afin d'affirmer ou d'infirmer des hypothèses et/ou stratégies en lien avec les besoins des métiers d'une organisation

- Analyser des données via des **méthodes de *Datamining***, de *Machine Learning* et d'Intelligence artificielle dans une démarche exploratoire (logiciel R, Python) : ces méthodes fonctionnent sur l'exploration et la découverte de relations entre données ou de modèles, de manière heuristique sans hypothèses forcément définies au préalable.

Compétence n°3 :

Pilotage de la qualité des données au regard des valeurs sociétales

- Développer une analyse **critique et éthique** sur les données afin de constituer des corpus qui seront validés par les organisations ;
- Sélectionner les données en respectant la **législation** en vigueur (RGDP), pouvant être diffusées en interne ou en externe
- Gérer les données (*Master Data Management & DPM*)
  -Concevoir ou sélectionner des **référentiels**
  -Décrire les données et intégrer des **métadonnées** ;
- **Mémoire**: établir un plan de gestion des données et/ou des livrables documentant les données et les processus associés (*Data-Book).*

Compétence n°4 :

Gestion d'un projet « data » dans une organisation.

- **Diagnostiquer**, auditer, repenser la structuration de la data de l'entreprise ;
- **Dialoguer** avec les clients/métiers et analyser leurs besoins et les usages associés ;
- Assurer la **communication** en présentant les données sous différentes formes visuelles (tableaux, graphiques, ...) ;
- **Interpréter** les résultats et rédiger un livrable de synthèse préconisant des solutions d'aide à la décision, à l'optimisation de la stratégie de l'organisation ;
- Identifier et développer de **nouveaux leviers associés à l'usage de données.**

Projets data : le *data book*

Un Data Book :

- Un livrable, recueil d'information d'un projet Data
- Un outil de capitalisation du projet
- Une documentation de référence pour la reprise du projet data

Module du Databook	Livrables associés au module, selon le modèle CRISP_DM ajusté
0 Guide et terminologie	Terminologie
1 Feuille de route projet data	Contexte, priorisation des usages, objectifs métier, critères de succès métier, inventaire des ressources, exigences, hypothèses et contraintes, risques et contingences, coûts et bénéfices, maquette initiale de datavisualisation, plan projet, évaluation initiale des outils et techniques
2 Méthode de rationalisation de l'inclusion/exclusion des données	Méthode de rationalisation de l'inclusion / excusion des données
3 Rapport d'exploration	Rapport d'exploration
4 Périmètre d'investigation	Cible d'analyse Data Mining, Critères de succès Data Mining
5 Qualification des données source	Rapport de collecte des données et description des données collectées, dictionnaire des données, rapport de qualité des données
6 Structure de la matrice d'apprentissage	Rapport de nettoyage des données, attributs dérivés, enregistrements générés, matrice d'analyse agrégée, matrice d'analyse reformatée pour les analyses, description des traitements des données
7 Benchmark résultats analytiques	Techniques de modélisation, critères d'évaluation de la modélisation, Test Design, paramétrages, modèle(s), description du (des) modèle(s), évaluation du modèle, paramétrages révisés
8 Appropriation des résultats métier	Support de présentation des résultats, évaluation des résultats par rapport aux critères de succès métier, résultat opérationnel, revue critique du process, liste des actions possibles, décision
9 Feuille de route usage (pour chaque usage direct)	Plan de déploiement, plan de pilotage et de maintenance
10 Capitalisation de connaissances (ensemble des usages indirects)	Retour d'expérience et documentation

Modules de data book et livrables (A. Nesvijevskaia, 2019)

MDM & Plan de gestion de données

Le Master data mangement est une méthodologie de management des données qui a pour objectif la création de référentiel de données s'appuyant sur **les données de références de l'organisation** indispensable à ses activités

Le data management plan (DMP) est de plus en plus demandé dans la recherche scientifique, dans les appels à projets financés sur fonds européens.

L'objectif est ainsi de documenter la manière dont seront produites ou collectées les données au cours et à l'issue d'un processus de recherche, en s'attachant notamment à définir comment elles **seront décrites, partagées, protégées puis conservées**

Compétence n°5 :

Conception des **services** de données pour transformer des activités sociétales.

- Identifier la valeur des données dans différents contextes d'activités, et savoir proposer des **transformations des activités**
- Identifier le **périmètre du service** (cerner ses limites) et évaluer sa capacité à résoudre des problématiques posées par des individus ou des collectifs
- Développer des **interfaces Homme-Données** : Élaborer des tableaux de bord, diffuser les données en restituant les résultats par des techniques de visualisation

Conclusion

Data littéracie @ SHS : développer des compétences pour l'analyse de données

- **Compétences à la fois techniques, informatiques et statistiques** (compétences 1 et 2)
- **Compétences en SHS** indispensables à la compréhension des données et l'appréhension de l**eurs usages en contexte pour la société** : compétences en lien avec l'éthique, et le juridique, mais aussi avec le pilotage de la qualité des données (compétence n°3), de méthodes de gestion de projets « data » (compétence n°4) et de conception de services de données (compétence n°5).
- **Interactions étroites entre littératies informationnelle, data et numérique.** Ces interactions sont aux fondements d'une réflexivité critique que sous-tend cette nouvelle forme de littératie (Buschman, 2009)

Cette recherche permet des réajustements progressifs du référentiel du Master, nécessaires dans un contexte en évolution rapide.

Références

- Arruabarrena (2018), Big data & éthique : la qualité des données en débat. In *L'Ethique en contexte info-communicationnel numérique* (Dir L. Balicco, E. Broudoux, G. Chartron, V. Clavier, I. Paillart), De Boeck, pp. 25-18.
- Bertin, J., *Sémiologie graphique*, Paris, Mouton/Gauthier-Villars, 1967.
- Calzada Prado, J. and Marzal, M.A. (2013), Incorporating Data literacy into information literacy programs: core competencies and contents, *Libri* , Vol. 63 n°2, pp. 123-134.
- Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining Data Information Literacy Needs: A Study of Students and Research Faculty. Portal: *Libraries and the Academy*, Vol. 11, n°2, 629–657.
- Koltay, T. (2016). Data governance, Data literacy and the management of data quality. *IFLA Journal*, Vol. 42, n°4, pp. 303-312.
- Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, n° 115, pp. 134-142.
- Mandinach, E. B., & Gummer, E. S. (2013). A Systemic View of Implementing Data literacy in Educator Preparation. *Educational Researcher*, Vol. 42, n°1, pp. 30–37. http://doi.org/10.3102/0013189X12459803
- Marr, Bernard (2015). *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. John Wiley & Sons.
- Matthews, P. (2016) Data literacy conceptions, community capabilities. *The Journal of Community Informatics*, Vol. 12, n°3.
- Nesvijevskaia, A. (2019) *Phénomène Big Data en entreprise : processus projet, médiations et impact sur les indicateurs de génération de valeur* (thèse CNAM à paraître, s.d. G. Chartron)
- Shearer, Colin. 2000. The CRISP-DM model : the new blueprint for data mining. *Journal of Data Warehousing*, Fall 2000, A 101 communications Publication edition, sect. Vol. 5, n°4.
- Wolff, Annika & Moore, John & Zdráhal, Zdenek & Hlosta, Martin & Kuzilek, Jakub. (2016). *Data literacy for learning analytics*, pp. 500-501. 10.1145/2883851.2883864.