Du big data au smart data, l'exploitation des données culturelles dans les bibliothèques et les musées : journée d'étude 2015 de la Fulbi

Catherine Muller, Emmanuel Brandl

Les billets d'EnssibLab
02 mars 2015

Pour l'édition 2015 de sa journée d'étude, « Bib Data, Smart Culture. Exploiter des données dans les bibliothèques, centres de doc, archives et musées », la Fulbi[1] conviait professionnels des bibliothèques, de l'information et de la documentation, ingénieurs, enseignants-chercheurs et consultants à croiser leurs regards sur l'exploitation des données culturelles. D'abord en revenant sur les spécificités et les enjeux des data aujourd'hui en partant de la définition scientifique du concept de "data" et des promesses qu'il contient. Mais aussi à l'aune de la normalisation des données à des fins comparatives et d'enquêtes d'usagers menées à la bibliothèque de Science Po Paris qui sont l'occasion d'aborder concrètement l'intérêt de produire des données sur les usages et les usagers. Ensuite, à partir de la question du « web participatif ». Question approchée à travers des restitutions de projets de crowdsourcing, d'un jeu vidéo grandeur nature organisé à Haguenau par sa médiathèque, ou encore l'analyse d'une enquête universitaire sur les tweets ayant fait le succès de l'évènement « MuseumWeek ».

Enjeux et opportunités des « data »

Dominique Cotte[2], enseignant-chercheur et consultant rappelle avec bon sens que les « data » ne sont pas chose nouvelle, mais ce qui change aujourd'hui, c'est moins l'existence même des « data » que la masse des données aujourd'hui disponibles au traitement et la **manière nouvelle de faire parler ces données**. Il est en effet possible aujourd'hui de faire dialoguer entre eux des silos de données auparavant séparés, participant alors d'une nouvelle manière de produire du sens et de nouvelles connaissances. Le phénomène est appelé à durer et à modifier en profondeur nos activités professionnelles et le rapport que l'on entretient avec elles.

Des "data" pleines de promesses ?

- D'abord d'une promesse « technique » : c'est la promesse dite des « 4 V », volume de données, vitesse de traitement, variété des données recueillies et croisées, et enfin, validité (ou véracité) de ces données.
- Ensuite d'une promesse d'« usages » : les univers de déploiement et d'application sont en effet bien réels, multiples et parfois sensibles. Marketing et vente, comportement clients, santé, épidémiologie, tourisme et territoires, équipement, déplacements, sécurité, détection, etc.

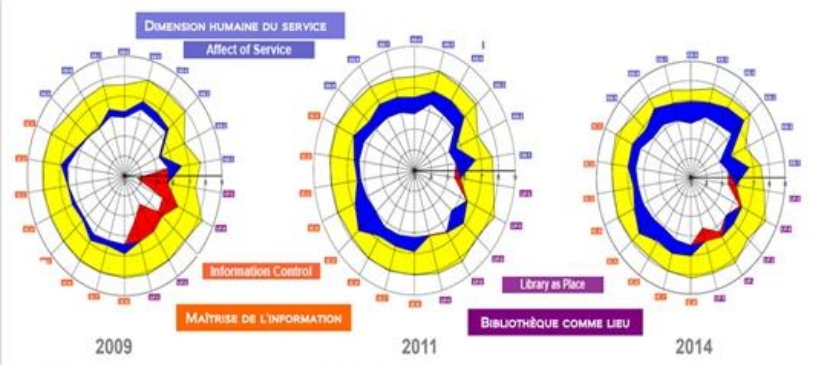
Ces exploitations concernent au premier chef les professionnels de l'information et de la documentation : amélioration du pilotage de projet par une meilleure connaissance de l'environnement, gain également en termes de visibilité et de l'exposition de fonds ou corpus, etc.
La question de la normalisation des données est analysée par le bibliothécaire et membre de la Fulbi, Xavier Guillot[3]. Notion fondamentale qui soulève la question des conditions de recueil et de comparabilité de ces données. En effet, qu'il s'agisse de répondre aux enquêtes propres aux obligations réglementaires des bibliothèques ou à l'enquête statistique générale des bibliothèques universitaire (ESGBU), qu'il s'agisse d'analyser, d'évaluer un établissement pour comparer un objectif avec sa réalisation, ou d'évaluer son impact dans une population cible, les difficultés s'enchainent en constatant rapidement une grande variabilité dans les items, le contenu des items (« Phonogrammes livres » – « Phonogrammes musiques » – « Vidéogrammes » – « livres électroniques seuls » – « Livres électroniques avec support », etc.), ou encore les libellés de paramétrage. Il existe 2 normes pour l'harmonisation des données et une meilleure compréhension des informations fournies par les SI :

- « Statistiques internationales de bibliothèques » - NF ISO 2789 maj Mai 2014[4], laquelle permet d'harmoniser les statistiques au niveau national et de favoriser la comparabilité au niveau international.
- « Indicateurs de performance des bibliothèques » - NF ISO 11620 maj Juin 2014, qui consiste en une série d'indicateurs destinés à caractériser la qualité et l'efficacité des

services fournis par les bibliothèques, évaluer l'efficience des ressources associées, permettant alors d'avoir des règles de calcul communes et de recouper les données de différents équipements.

Toutefois, il reste que la norme ISO 2789 inclut six familles de données, dont les familles « accès », « installations », « personnel », et « dépenses », alors que les SIGB n'incluent pas ces familles, et que les libellés restent variables. C'est pourquoi Xavier Guillot propose plusieurs pistes de réflexion, notamment la création d'un groupe de travail au sein de la Fulbi portant spécifiquement sur ces questions d'harmonisation.

Les enquêtes de satisfaction "*user-oriented*"



© Cécile Touitou - « Libqual+ : Comparaison diachronique »

Cécile Touitou[5] responsable marketing de la bibliothèque de Sciences-Po Paris, nous offre quant à elle un retour approfondi sur les expériences d'enquête « *user-oriented* » menées à la bibliothèque de Science Po Paris. Ici ce sont les usagers qui sont au coeur de la stratégie marketing[6] à la différence des études dites « collections centrées ». Trois principes méthodologiques phares découlent du « volume » et de la « variété » des données recueillies : 1) les données ne parlent pas d'elles-mêmes, 2) il s'agit de rester attentif au fait de ne pas se perdre dans les masses des données collectées et hétérogènes, 3) il faut rechercher l'analyse comparative. Il ressort de cette présentation une certaine inventivité méthodologique à l'oeuvre, sur le modèle de l'enquête américaine de référence « *sweeping the library* », qui consiste à transposer au contexte des bibliothèques une approche ethnographique initialement utilisée par les urbanistes étudiant l'appropriation des espaces commerciaux par les clients. Le dispositif d'enquête de satisfaction en ligne LibQUAL+™[7], bien connu des bibliothèques universitaires, tient compte de ces 3 principes, et même s'il accuse certaines critiques, il permet d'évaluer dans le temps la satisfaction des services et espaces de la bibliothèque , et présente l'avantage d'affiner l'évaluation par "segment de population".

Des big data aux smart data

L'ensemble des interventions a souligné l'importance de l'exploitation pragmatique de tout travail d'enquête : des études certes, mais concrètement pour quoi faire ? C'est toute la question du passage des « *big data* » aux « *smart data* », ou dit autrement, des **masses de données recueillies et traitées à des produits et des services (re)pensés à partir de ces données**. Globalement, ces data visent à mieux connaitre les évolutions des usages, notamment au regard des évolutions technologiques, pour penser la bibliothèque et les services de demain.

Quelques exemples : on constate aujourd'hui que le niveau d'équipement des usagers en « écrans »[8] pose entre autres la question du maintien ou non du service de prêts d'ordinateurs, comme de l'accès aux prises électriques. La méthode du *sweeping* permet par ailleurs de remodeler l'espace des interactions sociales au sein de la bibliothèque selon les comportements d'utilisation des différents types d'usager. L'analyse des types et des taux d'occupation a des répercussions sur les stratégies organisationnelles : quel type de personnel en salle à quels moments ? Les enquêtes auprès des usagers permettent aussi de pondérer certains *a priori* : pour la bibliothèque de Science Po, les demandes d'horaires ne vont pas vers une ouverture 24h/24 mais plutôt vers une ouverture tout au long de l'année, d'où une réflexion pour ouvrir au-delà de 16 semaines à l'année. Par ailleurs, pour les équipements culturels d'un territoire, l'analyse du

déplacement des touristes permet d'appréhender les logiques touristiques et proposer des plaquettes, des parcours, adaptés.

Enjeux et opportunités du « web participatif »

Projets collaboratifs, participatifs et contributifs : de l'intérêt des publics au cœur du crowdsourcing

© Muséum national d'Histoire naturelle « Herbier numérique collaboratif citoyen »

Antoine Courtin[9], ingénieur d'études au Labex parisien *Les passés dans le présent* rappelle le flou qui entoure la terminologie du ***crowdsourcing***, à mi-chemin entre l'animation de communautés et l'indexation collaborative, signe des nombreuses attentes et des enjeux métiers qu'il recèle. Toutefois, s'il existe de nombreux critères de distinction, variables selon le porteur de projet (institution culturelle, archives, collectif), il demeure que les *publics* sont toujours au cœur des projets de *crowdsourcing*. Et c'est bien autour de l'implication des publics que les questions s'organisent : quel cadre donner à cette participation ? Comment donner à la fois envie de participer tout en contrôlant la qualité des contributions (donc en modifiant ou parfois en refusant), et surtout sans perdre les informations sur les données initiales. En définitive, chacun trouve ses solutions, notamment en fonction de plusieurs facteurs : type de contribution, type de document, porteur du projet, objectif du projet.

Pauline Moirez[10], conservatrice du patrimoine à la BnF et experte des techniques documentaires numériques et services en ligne présente **le projet CORRECT**[11] qui est en phase d'expérimentation à la BnF depuis 2012. A l'origine du projet, l'idée d'améliorer les modes textes des documents de Gallica par des contributions d'usagers en s'appuyant sur une plateforme préexistante et des contributeurs wikisourciens. Ce projet s'est appuyé sur une enquête usager qui a permis de pointer un profond changement de mentalités. En effet, alors qu'en 2008, les usagers sont opposés au fait d'intervenir et de modifier les documents eux-mêmes, ils sont très nettement prêts à participer en 2011. Il en va de même du côté des professionnels : s'ils sont opposés en 2009 au fait de pouvoir demander aux usagers de corriger des données issues de la BnF, aujourd'hui les retours sur la plateforme CORRECT sont extrêmement positifs. Il y a eu comme une prise de conscience du fait que **les métadonnées des professionnels et des usagers sont complémentaires et non concurrentes**. Mais ce projet n'a pas suscité l'engouement espéré, seuls 54 documents ont été finalement corrigés (10 000 pages) sur les 1419 ouverts[12]. D'où une reconfiguration du projet visant à favoriser encore plus l'effet « communauté ». Ce projet intègre pour cela trois modules complémentaires : un module de correction, un module technique permettant de fusionner les corrections, mais aussi un réseau social permettant d'échanger sur les corrections et difficultés, de constituer des groupes de travail.

Lisa Chupin[13], doctorante au laboratoire DICEN-IdF CNAM de Paris, restitue un travail de thèse en cours : un **projet d'herbier collaboratif**, les « Herbonautes », site développé par le Muséum d'histoire naturelle de Paris, pour permettre d'aider les chercheurs à transcrire les étiquettes des planches d'herbier qui ont été numérisées. Le travail présenté ici montre finalement combien les projets de crowdsourcing doivent s'adapter aux spécificités des documents qu'ils proposent à la collaboration des internautes, et à la culture de professionnels engagés. En effet, d'un côté si les « herbonautes » sont souvent de véritables érudits, il reste que les contributions doivent toutes être validées par des professionnels. D'où la mise en place d'un dispositif de contrôle des transcriptions en amont des contributions, d'un dispositif pédagogique de tutoriel, et enfin d'un dispositif de contrôle en aval des contributions. Ce qui limite la visibilité des contributions et entraine parfois une perte d'information. C'est pourquoi il y a aujourd'hui une volonté d'intégrer davantage l'espace de la plateforme de crowdsourcing et l'espace des collections pour avoir directement accès aux échanges, et ainsi valoriser la participation en la rendant immédiatement visible. Notons qu'ici aussi le crowdsourcing provoque un changement de mentalités, que Lisa Chupin appelle sans détour une « révolution épistémologique » : de la même façon que pour le projet CORRECT, les professionnels scientifiques doivent apprendre à accepter que des « profanes » interviennent sur leurs documents.

Pour Julie Guillaumot[14], responsable du pôle patrimoine à l'Agence régionale du Centre pour le livre, l'image et la culture numérique, le site participatif Ciclic, la création du **service d'archive du film** en 2006 a nécessité de faire appel aux habitants pour atteindre l'objectif de sauvegarder les films qui échappent au dépôt légal[15]. En effet, ces films, qui sont une contribution importante à l'histoire du territoire, ne sont que rarement référencés : par exemple les lieux des vidéos sont souvent inconnus. Ici aussi le dispositif s'adapte au type de document : mise en place pour les internautes d'un service de géolocalisation, de chronolocalisation, de recherche par nom, etc. Mais aussi, puisqu'il s'agit de tranches de vie, de lieux du territoire local, le site de Ciclic insiste sur la mise en place des fonctionnalités *participatives*, comme le fait de pouvoir « résoudre des énigmes » (par des « Sherlock », par exemple identifier les lieux, etc.), de pouvoir « insérer des commentaires » et de cumuler des points (« *gamification* »). Encore une fois, on assiste à une petite révolution : les rôles s'inversent, ce sont les contributeurs qui sont les experts, et non le documentaliste, qui lui n'est pas nécessairement natif du territoire et dont le rôle consiste alors à publier et modérer les commentaires et les énigmes.

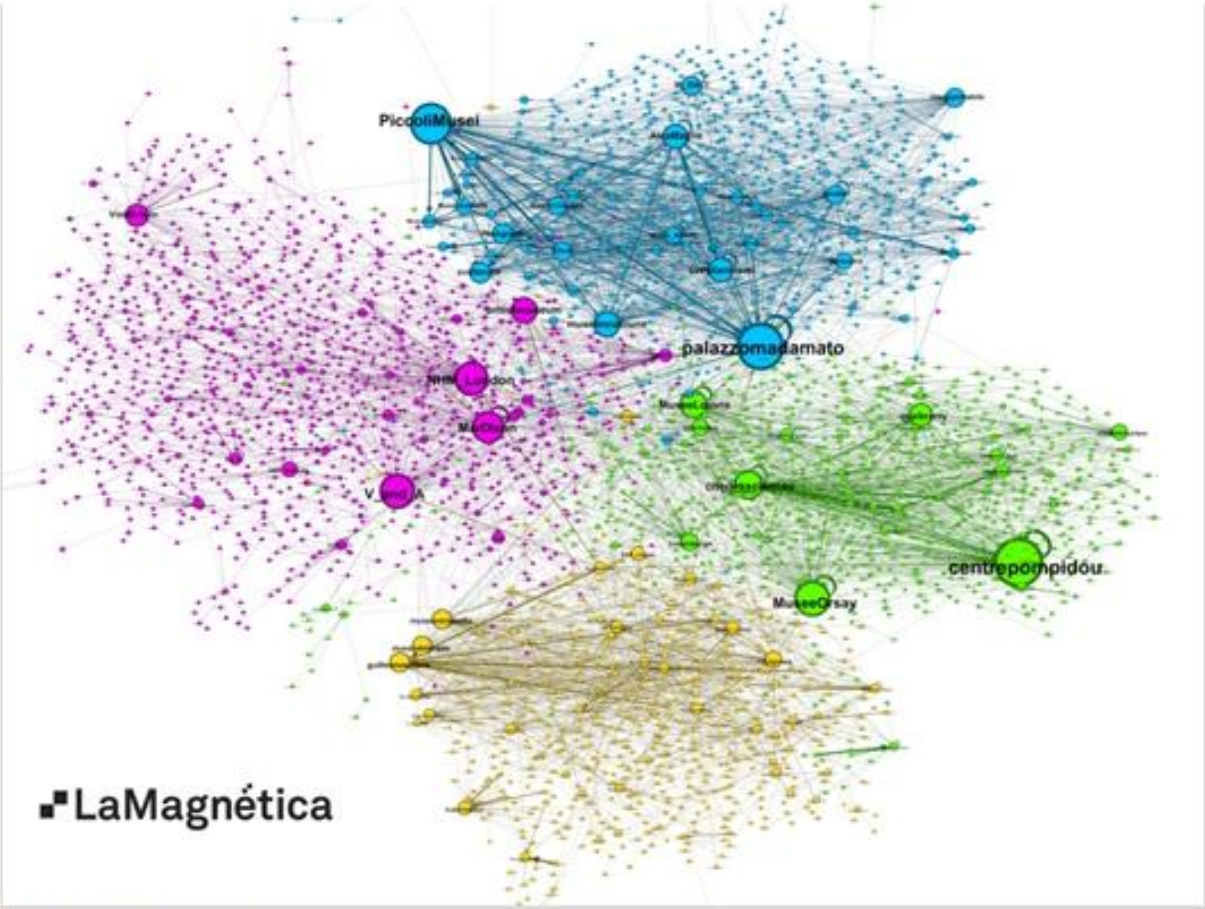
Enfin, le projet de la médiathèque de Haguenau[16] propose une toute autre approche. La dimension participative est envisagée pour la fête des 900 ans de la ville sur la base d'un **jeu vidéo urbain géolocalisé** : le SANDHAAS RUN. Les facteurs pouvant susciter l'intérêt des habitants (lieux et références culturels et historiques), la capacité à toucher un public le plus large possible, mais aussi les aspects ludiques, sont au cœur de la réflexion. La médiathèque s'appuiera pour ce faire sur un partenariat avec la société Xilabs dont le directeur, Ivo Flammer, aura fait une présentation tout à fait convaincante de ses jeux urbains « qui s'incrustent dans votre quotidien », avec géolocalisation et réalité augmentée.

Une mise en perspective historique

Bertrand Müller[17], directeur de recherche au CNRS plaide pour un travail d'historicisation de la data. Les data sont le produit de chaines de constructions successives ou simultanées, mais aussi spécifiques, entre des domaines de traitements de données liés mais distincts : sociologues, historiens, archivistes, documentalistes, chercheurs, informaticiens, tous ces acteurs interviennent dans la définition et la structuration des données, auxquelles sont associées des significations différentes. **Les données ne se livrent donc jamais d'elles-mêmes, isolées d'un contexte**, détachées de significations et de constructions préalables, contrairement à ce que

pourrait laisser entendre l'approche informatique, qui considère la « data » précisément comme une « donnée » qui se livre d'elle-même au chercheur, seule la « méta donnée » devenant alors signifiante. Pour mener l'enquête et interpréter des résultats, il nous faut donc garder en tête qu'**une « data » est toujours l'objet d'un travail de construction préalable**, à travers des catégories, des classements, des choix méthodologiques et épistémologiques.

Une analyse de projet



© Museumweek, LaMagnética

Après cette approche historique, Brigitte Juanals [18] et Jean-Luc Minel [19], tous deux enseignants-chercheurs à Paris 10, et Antoine Courtin, se sont penchés de concert sur une analyse du succès du MuseumWeek[20], également nommée la « semaine des musées sur Twitter ». La particularité de cet évènement tient au fait que l'initiative revient à une douzaine de Community Managers de musées français qui ont réussi à entraîner dans leur sillage, *via Twitter*, 630 musées à travers le monde : ainsi que le titrait le Huffington Post, "***Pour sa 1ère édition, la MuseumWeek a rassemblé du 24 au 30 mars plus de 100 musées français, 630 en Europe, à travers 260.000 tweets publiés autour du hashtag #MuseumWeek. Un événement clé dans l'histoire du web social".*** L'analyse s'appuie sur une méthodologie spécifique et inventive, notamment dans la combinaison de méthodes de recueil et d'analyse des données, et procède à une évaluation à la fois quantitative et qualitative de tweets : analyse comparative et linguistique de 550 tweets originaux, identification des principales thématiques, jeu d'annotations.
Cette étude livre un grand nombre de résultats originaux en identifiant à la fois la distribution des tweets (institutionnels *vs* individuels), les catégories de participants illustrées par une typologie, mais aussi en livrant une analyse quantitative sur la participation respective de chacun des 12 musées initiaux et en analysant la fonction même jouée par les hashtags (les hashtags qui

jouent sur le plan émotionnel sont les plus utilisés), montrant ainsi que la nature du tweet (RT, commentaires, tweets initiaux, etc.) reflète la conception du rôle de *community manager* de chaque institution.

Pour conclure

Pour conclure, il ressort de ces expériences d'exploitation de données culturelles un sentiment de *changement de paradigme*. Ces témoignages attestent du passage d'un modèle de spécificités professionnelles auparavant relativement isolées, à un monde dans lequel les frontières entre savoirs (profane *vs* professionnel), connaissances, fonctions, s'atténuent (sans disparaitre complètement), pour une meilleure complémentarité des savoirs maîtrisés ainsi qu'une meilleure prise en compte de la contribution des uns et des autres dans la production de la connaissance.

Notes

[1] La Fulbi est la Fédération des utilisateurs de logiciels pour les bibliothèques, documentation et information.
[2] Enseignant-chercheur à l'Université de Lille 3 et consultant au Cabinet Ourouk, Dominique Cotte est également animateur du GT exploratoire du GFII « Big data, smart data ». Titre de l'intervention : « Enjeux des data pour les industries de l'information et de la connaissance ».
[3] Xavier Guillot en poste à la Médiathèque départementale du Puy-de-Dôme est également membre du CUTO, Le Club des UTilisateurs d'Orphée. Titre de l'intervention : « Normaliser les données utilisateurs des SIG ».
[4] Voir P. Roswitha, « Les indicateurs de qualité pour les bibliothèques nationales », *Bulletin des bibliothèques de France*, n° 6, 2013.
[5] Titre de l'intervention : « Les enquêtes auprès des usagers : mieux connaitre les usages et les pratiques, mesurer la satisfaction et les attentes ».
[6] Comme en écho à l'intervention de X. Guillot, Cécile Toitou signalera à ce titre la norme ISO 16439 « Méthodes et procédures pour évaluer l'impact des bibliothèques » qui définit l'impact social et économique des bibliothèques sur leur territoire d'action.
[7] Pour une présentation complète, voir H. Coste, « LibQUAL+ », *Bulletin des bibliothèques de France* , n° 1, 2013.
[8] Ordinateur, smartphone, etc., tant et si bien que sur les tables de travail, on trouve parfois plus d'écrans que de livres.
[9] Titre de l'intervention : « Les projets de crowdsourcing dans les institutions culturelles : Retours d'expériences »
[10] Titre de l'intervention : « Crowdsourcing à la BnF : une approche expérimentale ».
[11] Projet CORRECT : "CORRection et Enrichissement Collaboratifs de Textes". Les objectifs de la plateforme sont : mettre à disposition des utilisateurs des outils de correction et d'enrichissement des documents numérisés et s'appuyer sur un réseau social pour soutenir et organiser la collaboration. Le projet réunit 9 partenaires : Orange, BnF, Jamespot, Urbilog, I2S, ISEP, INSA Lyon, Université Lyon 1, Université Paris 8.
[12] Ce qui selon Pauline Moirez tient à un triple problème : problème de relais institutionnel, ingérence de la BnF dans la « communauté » des wikisourciens, qui préfèrent intervenir sur des documents numérisés par eux-mêmes, et enfin, problème de format : en sortie, le plein texte n'est pas compatible avec le format « alto » pour la recherche plein texte de Gallica. Le format « alto » permettant d'assigner une adresse à chaque terme reconnu par l'OCR (« optical

character recognition » ou reconnaissance optique de caractères) au sein d'un document. Ce qui est primordial en termes de format de sortie pour les projets de crowdsourcing.
[13] Titre de l'intervention : « Les dispositifs de transcription collaboratifs d'herbiers. La question de la valorisation et de l'intégration des données produites par les internautes aux collections »
[14] Titre de l'intervention : « Les usages participatifs du site memoire.ciclic.fr »
[15] Seuls donc sont concernés les films qui n'ont pas reçu de visa d'exploitation du CNC : films amateurs, associatifs, militants.
[16] Présenté par Maïlis Frebillot et Thomas Schlotter, bibliothécaires. Titre de l'intervention : « Sandhaas Run. Le projet de jeu géo-localisé de la Ville de Haguenau ».
[17] Titre de l'intervention : « Valoriser les données de la science ».
[18] Brigitte Juanals est maître de conférences habilitée à diriger des recherches en Sciences de l'information et de la communication à l'Université Paris Ouest Nanterre La Défense, chercheur à l'UMR 7114 MoDyCo et chercheur associé au LabSIC.
[19] Jean-Luc Minel est professeur des universités à l'Université Paris Ouest Nanterre La Défense et chercheur à l'UMR 7114 MoDyCo.
[20] Titre de l'intervention : « Musées et data : de l'exposition des données patrimoniales à l'analyse des données sur les réseaux sociaux numériques ».