APPEL A PROJETS 2018

FÉDÉRATION DE RECHERCHE AGORANTIC

«CULTURE, PATRIMOINES, SOCIÉTÉS NUMÉRIQUES »

Titre : **GO**uvernance des c**O**rpus scientifiques d'étude du **W**eb 2.0

Acronyme : **GoOW**

Porteur : **Guillaume MARREL (LBNC - Science politique)**

Laboratoires : **LBNC :** Martine LE FRIANT (Droit du travail et droit du numérique)
Bérengère GLEIZE (Droit privé et propriété intellectuelle)
Samuel PRISO (Droit international public et droit européen)
Ouassim HAMZAOUI (Science politique)
LIA : Eric SANJUAN (Informatique Décisionnelle)
Pierre JOURLIN (Intelligence Artificielle)
Malel HAJJEM (Recherche d'Information Multilingue)

GoOW est un projet pluridisciplinaire associant des chercheurs en informatique, en droit et en science politique de deux laboratoires de l'UAPV (LIA et LBNC), dont l'objectif est de proposer et de tester des dispositifs de sécurisation juridique et informatique pour l'exploitation en sciences sociales et politique de corpus de données numériques massives notamment issues du microblogging.

Contexte, positionnement et objectif(s)

Toutes les données qui constituent et que génèrent en continu, le Web 2.0 (Tillinac, 2006), sont caractérisées par une grande instabilité, à cause de la labilité de leurs cadres juridiques (Pellegrini & Canevet, 2012), mais aussi du caractère évolutif des techniques informatiques développées pour les saisir[1], ainsi que des opportunités qu'elles constituent en termes d'analyses sociologiques et socio-politiques (Venturini & al., 2014).

C'est dans cet environnement instable que nombre d'institutions universitaires et de recherche scientifique construisent et exploitent de larges corpus de données issus des réseaux sociaux numériques (RSN) et d'autres supports comme les blogs. Autrement dit, un ensemble de matériaux qui contrevient intrinsèquement et de façon croissante, non seulement aux conditions spécifiques d'usage de ces médias aux contenus plus ou moins indexés, mais aussi à la jurisprudence et/ou aux textes juridiques nationaux[2] et

1 *Mission d'expertise sur la fiscalité de l'économie numérique*, Rapport au Ministre de l'économie et des finances, au Ministre du redressement productif, au Ministre délégué chargé du budget et à la Ministre déléguée chargée des petites et moyennes entreprises, de l'innovation et de l'économie numérique établi par PIERRE COLLIN et NICOLAS COLIN, janvier 2013, (http://www.economie.gouv.fr/files/rapport-fiscalite-du-numerique_2013.pdf).
2 Cf. principalement : la loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, *JORF* n°0235, 8 octobre 2016.

européens[3] (Polidori, 2015 ; Foegle, 2017) qui instituent notamment le « droit à l'oubli numérique » (Blanchette & Johnson, 2002 ; Dechenaud & al. 2015).

Aussi, est-il possible que des chercheurs des sciences informatiques et sociales soient demain confrontés à d'intempestives demandes provenant d'utilisateurs de RSN ; et dont l'objet serait l'effacement et le déréférencement de tels ou tels de ces messages « publics ». Plusieurs universités de renom international s'inquiètent déjà de ce que la conservation de ces « archives numériques » par un organisme tiers puisse déboucher sur des plaintes.

À cette évolution de l'environnement des activités numériques, s'ajoute la montée en puissance des logiques de *data lock-in* qui consacrent une tendancielle hégémonie des détenteurs de données et qui démultiplient les enjeux juridiques, scientifiques et politiques de la gouvernance des jeux de données. Prenons-en pour exemple les récentes évolutions de la politique d'usage des données de Twitter, objet d'étude privilégié (Smyrnaios & Ratinaud, 2013) de la « puissance sociale » (Elias, 1990) des RSN dans cet espace public du Web (Cardon, 2010) : sur la dernière décennie, le "petit oiseau" a placé l'accès aux données hors de la portée des capacités d'investissement des acteurs publics de la recherche. Dès 2010 et 2012, et au terme de coûteuses transactions[4], seules les structures à but lucratif comme *Gnip*, *Datasift*, *Topsy*, et *NTT Data* ont obtenu l'accès au *firehose* (le flux exhaustif de ses données brutes[5]) afin de produire des analyses poussées de *datamining* à finalités commerciales. Puis en 2014, Twitter a durci radicalement son modèle économique par le rachat de *Gnip* et la fin de tous les accords conclus précédemment. La firme a pris diverses mesures pour dissuader les pratiques libres d'archivage de *tweets* et assure de façon monopolistique la revente de ces *data* à des prix bien trop élevés pour les utilisateurs à finalités scientifiques et/ou non-directement commerciales. Parmi ces derniers, les institutions de la recherche publique se retrouvent *de facto* empêchées de procéder à une quelconque analyse scientifique des données de ces réseaux, de leur architecture, fonctionnement et usages.

C'est donc à un double titre que l'UAPV et la FR AGORANTIC sont concernées par cette problématique de *numerical double-bind* (double-contrainte numérique) :

1. En premier lieu, car les chercheurs qui s'y rattachent ont fini par « constituer », au fil des années et des recherches collectives auxquelles ils ont pu participer, toute une série de *corpus* d'archives numériques dont la date d'établissement de certains remonte à plus de 5 ans : projets ANR *CAAS* sur une archive de 20 % du Web en 2009, *ImagiWeb* avec archivage et annotation des microblogs de la campagne présidentielle 2012 et *GAFES* avec archivage des microblogs et des sites liés 2015-2016 sur les festivals, et participation aux campagnes CLEF RebLab dirigées par

3 Cf. notamment : Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données). V. aussi CJUE 13 mai 2014, aff. C-131/12, *Google Spain* c/ *Agencia Española de Protección de Datos*, D. 2014. 1476, note V.-L. Benabou et J. Rochfeld, 1481, note N. Martial-Braz et J. Rochfeld, et 2317, obs. P. Tréfigny ; AJDA 2014. 1147, chron. M. Aubert, E. Broussy et H. Cassagnabère ; AJCT 2014. 502, obs. O. Tambou ; Constitutions 2014. 218, chron. D. de Bellescize ; RTD eur. 2014. 283, édito. J.-P. Jacqué, 879, étude B. Hardy, et 2016. 249, étude O. Tambou ; Rev. UE 2016. 597, étude R. Perray ; A. Debet, *Google Spain* : Droit à l'oubli ou oubli du droit ?, CCE 2014. Étude 13.

4 Pour les seuls accords avec *Gnip* et *Datasift*, il aurait été question, selon la presse spécialisée (*OWNI*, *JDN*), de plus de 250 millions de dollars.

5 C'est-à-dire les textes de tous les *tweets*, ainsi que toutes les autres données liées aux micro-messages.

Julio Gonzalo entre 2013 et 2014 sur la réputation des marques. Aussi sont-ils, en l'espèce, antérieurs aux principaux textes de 2014 et 2016 qui encadrent désormais les obligations juridiques afférentes aux questions numériques.

2. Deuxièmement, en raison de l'intérêt pour les « sociétés numériques » au cœur du projet scientifique de la FR, et qui peut difficilement faire l'économie d'une réflexion approfondie sur la possible marginalisation des acteurs scientifiques dans le régime de régulation des données numériques qui est en train de se mettre progressivement en place.

Organisation et moyens mis en œuvre

A l'aune de cette problématique d'ensemble et de la saillance de ses implications locales, l'ambition de ce projet est de proposer une triple réflexion ambitieuse en termes de sécurisation et de viabilisation de la gouvernance des corpus scientifiques d'étude du Web 2.0, pour l'UAPV et pour les communautés scientifiques nationales et internationales.

1. Dans une optique de **sécurisation juridique**, il s'agit de caractériser l'évolution des cadres légaux qui définissent les conditions particulières d'utilisation des principaux RSN, et de confronter cet environnement juridique évolutif et les stratégies mises en oeuvres par les « géants du web »[6] pour faire face à « l'injonction » qu'il leur est notamment faite de mettre en œuvre les orientations et réglementations européennes les plus récentes en la matière (Barreau, 2016) ; notamment d'utilisation et de transfert des données de leurs utilisateurs européens aux Etats-Unis (Bellanova & De Hert, 2009). Un intérêt tout particulier sera accordé à Twitter, consacré comme « le pouls de l'internet », qui jouit d'une résonance considérable en matière politique et dont provient par ailleurs une partie importante des corpus stockés à l'UAPV. L'interrogation sera double : 1) il s'agit non seulement d'ébaucher une étude du statut juridique des tweets, confrontés aux conditions générales d'utilisation mais aussi, 2) de proposer une analyse comparée des modalités optimales de protection des bases de données créées, au moyen du droit des contrats ou de celui de la propriété intellectuelle (Derclaye, 2006).

2. Dans un objectif de **sécurisation informatique**, l'objectif est de proposer une technologie qui, basée sur l'outil FELTS[7] développé par Pierre Jourlin, en remplaçant le contenu des microblogs par des ensembles de références au WikiPedia, satisferait à un double impératif technique, à savoir : 1) élaborer un système de « brouillage » de la traçabilité des messages (et tout particulièrement ceux archivés dans les différents *corpus* scientifiques des équipes de recherche rattachées à l'UAPV), qui garantisse donc de ne pas pouvoir remonter de manière univoque à leurs auteurs, y compris à partir de recherches par mots clefs spécifiques et ce, 2) tout en garantissant les conditions minimales de la poursuite des recherches en sciences sociales.

6 On retrouve derrière cette expression des sociétés comme : *Google*, *Apple*, *Facebook*, *Amazon*, *Microsoft (*GAFAM*)*, *Yahoo*, *Twitter*, *LinkedIn* et d'autres autour du nouveau sigle NATU : Netflix, Airbnb, Tesla, et Uber.
7 Cf. https://github.com/jourlin/FELTS.

3. Depuis une perspective de sciences sociales - de science politique notamment -, il s'agira enfin mettre en chantier trois recherches complémentaires visant tester la **viabilisation de l'exploitation scientifique des données** extraites du Web 2.0, et la **consolidation d'une posture d'épistémologie critique** (Ollion & Boelaert, 2013).

    a. Pour prolonger les questionnements juridiques, nous analyserons les dynamiques socio-politiques de régulation de l'accès et de la gestion des données, et évaluerons de façon prospective la place, le rôle, les marges de manœuvre et les stratégies des acteurs publics de la recherche scientifique dans le cadre du Big Data (Harcourt, 2014).

    b. Pour opérationnaliser les outils informatiques, nous expérimentons les transformations des conditions d'analyse de sciences sociales, dans le nouveau cadre technique d'anonymisation des « traces numériques » (Nguyen, 2014). Pour ce faire, nous prendrons pour objet les pratiques de communication et de manipulation politiques observables à partir des corpus collectés dans Twitter à l'UAPV. Sur Twitter, les campagnes collectives de diffusion manuelle de discours concertés[8] se généralisent dès 2013 autour de l'image d'organismes publics ou privés (université, banques, constructeurs automobiles…) et triomphent avec la campagne de Donald Trump presque intégralement digitalisée. Efficaces, difficiles à détecter et incontrôlables, ces pratiques méritent toute notre attention. Elles feront l'objet d'une sociologie des stratégies collectives de mise en circulation de discours de *disqualification* de personnalités politiques et de *politisation* d'enjeux d'action publique. Nous envisageons deux cas d'étude issus du corpus "Festival" de l'ANR GAFES : 1) le phénomène de débordement de l'occurrence "festival" par l'item "migrant" au moment de la crise migratoire de 2014 et 2015 et qui vient renforcer le processus de mise à l'agenda de la politique im-migratoire ; 2) l'emballement des tweets négatifs au moment de l'édition 2015 du festival de Cannes, qui ont pris pour cible la Garde des Sceaux, Christiane Taubira. Il s'agira de proposer une sociologie-historique des phénomènes de propagation / dissémination / diffusion à partir d'une comparaison avec le même type de phénomène au moment de la campagne 2012 autour des candidats Sarkozy et Hollande, saisi dans le corpus de l'ANR IMAGIWEB[9].

    c. Pour poser les jalons d'une démarche réflexive d'épistémologie critique, nous interrogerons 1) les modes de conceptualisation et de représentation de la nature et du fonctionnement du Web, qui renvoient principalement à trois types de modélisation : épidémiologique, ondulatoire et sismicienne ; et 2) les effets de ces modèles d'interprétation (viralité, vibration,

8 Le choix de ces pratiques spécifiques sur ce réseau en particulier s'explique par la complexité des difficultés légales et techniques qu'elles posent, respectivement en raison de la contradiction entre le respect du droit d'expression aux USA et les injonctions européennes de modération de l'expression publique et de l'inexistence en l'état actuel de technologies (de *datamining* ou d'Intelligence artificielle) permettant de « contrer » un groupe de quelques centaines de personnes diffusant manuellement mais de manière concertée des « éléments de langages » (http://www.casilli.fr/2016/11/20/never-mind-the-algorithms-the-role-of-exploited-digital-labor-and-global-click-farms-in-trumps-election/).

9 Nombre de twittos, stratégies concertée, ciblage des locus, délimitation des éléments de langage, temporalité de l'action…

réplique…) sur la place des acteurs et de leurs stratégies, dans l'analyse sociologique (Boullier, 2015a & b).

Impact et retombées

1. Etablissement de procédures et d'outils d'anonymisation des corpus issus du Web 2.0 permettant d'assurer techniquement la sécurité juridique de nos établissements et de garantir des conditions d'exploitation compatibles avec la recherche en sciences sociales, présentés et éprouvés lors d'un colloque international fin 2018.

2. Publications des études de cas (issus des corpus collectés à l'UAPV) de pratiques concertées de communication politique sur les RSN et validant l'efficacité des procédures d'anonymisation.

3. Mise en place d'un Observatoire des évolutions et des enjeux affectant la gouvernance des corpus scientifiques d'étude du Web 2.0 (ObGoOW), qui se distinguerait de la plupart des structures existantes traitant des données numériques dans le domaine de l'enseignement supérieur et la recherche (Observatoire numérique de l'Enseignement supérieur, Agence nationale des usages des TICE...), en se proposant d'être une ressource de veille juridique, technique et socio-politique s'adressant spécifiquement aux chercheurs de sciences sociales (selon la localisation de leur résidence administrative de rattachement et de leur statut) souhaitant utiliser les données provenant des principaux RSN. L'observatoire aurait pour finalité en interne, de faciliter l'exploitation, au moyen d'outils de visualisation, de traitement par R, des corpus déjà existants à l'UAPV. Partenariat envisageable : Observatoire des mondes numériques en sciences humaines (OMNSH).

4. Consolidation de la réflexion à l'origine des deux options du master *d'Informatique* d'Avignon (*e-commerce* pour la spécialité *Ingénierie systèmes* et option *Web 2.0* pour la spécialité *Réseaux*), ainsi que du master *Gouvernance numérique* proposé par l'UFRip DEG, dont les équipes entendent développer de manière concertée une expertise sur l'identification et le traitement des enjeux stratégiques émergents relatifs au pilotage des « data » et à la gouvernance des outils et des données numériques, en partant de la problématique croissante, que constitue pour les organisations publiques et privées, l'optimisation de la maîtrise de leur « patrimoine informationnel ».

5. À partir notamment de la réflexion d'ordre épistémologique engagée sur les modes de conceptualisation du Web 2.0, reformulation de la problématique de *numerical double-bind* de la gouvernance des corpus scientifiques d'étude du Web 2.0 en vue de solliciter d'autres financements (PEPS/Jeune chercheur/FEDER/ANR…).

Dimension interdisciplinaire et cohérence par rapport à l'axe identitaire

Le projet relève d'une problématique interdisciplinaire structurante pour l'établissement, la fédération de recherche et les laboratoires SHS de l'UAPV : celle des

conditions techniques, juridiques et éthiques d'exploitation scientifique des data du Web 2.0. Cette problématique transversale, centrée sur la dimension "société numérique" de l'axe identitaire, est susceptible d'intéresser l'ensemble des chercheurs mobilisant les Big et/ou Open Data sur les thématiques de la culture et de ses publics, du patrimoine muséal, littéraire et historique, de l'espace ou encore du politique. Cette problématique n'a encore jamais été financée par un appel à projet Agorantic. Dans cette phase initiale et préparatoire, la démarche associe au sein du LBNC des chercheurs de sciences juridiques et de sciences sociales du politique, ainsi que des chercheurs en informatique du LIA. Le projet est positionné sur trois des cinq axes de la FR : les deux axes transversaux non-thématiques 1 et 5 et l'axe 3 porteur des enjeux juridiques, éthiques et politiques du numérique :

1- Méthodologies et Interdisciplinarité

3- Politique(s), transparence et éthique

5- Structuration et exploitation de corpus

Dimension internationale du projet et partenariats extérieurs envisagés

Construit à partir des corpus de l'UAPV dans l'environnement français, le projet de recherche a vocation à intégrer les règles juridiques internationales et européennes ainsi que des corpus de données francophones d'origine tunisienne, et non francophones notamment états-uniennes. Il vise par ailleurs à développer des approches comparatives, en particulier avec l'équipe de l'UNERD de Madrid à l'origine de CLEF RepLab 2012-2014 menée par Julio Gonzalo[10].

Autre partenariat : Observatoire des mondes numériques en sciences humaines (OMNSH).

Budget prévisionnel

Agor@ntic :	8 000€
Cofinancement :	9 000€
Total :	17 000€

Cofinancements sollicités

- AAP *"Colloques et manifestations scientifiques"* UAPV 2018
- AAP interne LBNC 2018

10 https://scholar.google.com/citations?user=opFCmpYAAAAJ

Annexes

Budget prévisionnel (€)

Dépenses		Recettes	
Nature	**Montant**	**Origine**	**Montant**
Fonctionnement			
Mission équipe	1 000		
Frais colloque	8 000		
Personnel			
1 stage master 6 mois	3 000	**Agor@ntic**	**3 000**
1 ingénieur d'étude 50% - 3 mois	5 000	**Agor@ntic**	**5 000**
Total projet	**17 000**		
Montant pour Agor@ntic			**8 000**

Références

- Catherine Barreau, « Le marché unique numérique et la régulation des données personnelles », *Annales des Mines - Réalités industrielles*, 3, août 2016, pp. 37-41.

- Rocco Bellanova & Paul De Hert, « Protection des données personnelles et mesures de sécurité : vers une perspective transatlantique », *Cultures & Conflits*, 74, 2009, pp. 63-80.

- Jean-François Blanchette, Deborah Johnson, « Data retention and the panoptic society : The social benefits of forgetfulness », *The Information Society*, 18.1, 2002, pp. 33-45.

- Dominique Boullier, « Les sciences sociales face aux traces du *big data*. Société, opinion ou vibrations ? », *Revue française de science politique*, 65/5, 2015, pp. 805-828.

- Dominique Boullier, « Vie et mort des sciences sociales avec le big data », *Socio*, 4, 2015, pp. 19-37.

- Dominique Cardon, *La démocratie internet. Promesses et limites*, Paris, Seuil, 2010.

- Conseil d'État, *Le numérique et les droits fondamentaux. Étude annuelle*, Paris, La documentation Française, 2014.

- Anne Debet, *Google Spain* : Droit à l'oubli ou oubli du droit ?, CCE 2014. Étude 13.

- David Dechenaud & al., *Le droit à l'oubli numérique – Données nominatives – Approche comparée, Larcier 2015.*

- Estelle Derclaye, « Une analyse économique de la protection contractuelle des bases de données », *Reflets et perspectives de la vie économique*, XIV, 4,, 2006, pp. 49-73.

- Norbert Elias, *La dynamique de l'occident*, Paris, Calmann-Levy, 1990.

- Jean-Philippe Foegle, « La CJUE, magicienne européenne du 'droit à l'oubli' numérique », *La Revue des droits de l'homme*, mis en ligne le 16 juin 2014; http://revdh.revues.org/840.

- Bernard E. Harcourt, « Governing, Exchanging, Securing: Big Data and the Production of Digital Knowledge », *Columbia Public Law Research Paper*, 14-390, 2014; http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2443515.

- Benjamin Nguyen, « Techniques d'anonymisation », *Statistiques et sociétés*, 2/4, 2014, pp. 43-50.

- Etienne Ollion et Julien Boelaert, « Au-delà des *big data* », *Sociologie*, 6/3, 2015, mis en ligne le 20 janvier 2016; http://sociologie.revues.org/2613.

- François Pellegrini, Sébastien Canevet, « Le droit du numérique : une histoire à préserver », *Vers un Musée de l'Informatique et de la Société Numérique en France ?*, Actes du colloque « Vers un Musée de l'Informatique et de la Société Numérique en France ? », Paris, Conservatoire National des Arts et Métiers, pp.61-76; http://minf.cnam.fr/Papiers-Verifies/Colloque-MINF-2012.pdf.

- Marion Polidori, « L'arrêt *Google Spain* de la CJUE du 13 mai 2014 et le droit à l'oubli », *Civitas Europa*, 34/1, 2015, pp. 243-266.

- Nikos Smyrnaios & Paul Ratinaud, « Comment articuler analyse des réseaux et des discours sur *Twitter* », *tic&société*, 7, 2, 2013 ; http://ticetsociete.revues.org/1578.

- Jean Tillinac, « Le web 2.0 ou l'avènement du client ouvrier », *Quaderni*, 60, 2006, pp. 19-24.

- Tommaso Venturini, Dominique Cardon & Jean-Philippe Cointet, « Présentation (Méthodes digitales. Approches quali/quanti des données numériques) », *Réseaux*, 188/6, 2014, pp. 9-21.