Utilisation de données à des fins de recherche

Guillaume Piolle
(ex-)CERNA / CentraleSupélec
guillaume.piolle@centralesupelec.fr

École « éthique du numérique », Arcachon
28 septembre 2016

Le GT « accès aux données » de la CERNA

Saisine du Cocor d'Allistene

Accès aux données à des fins de recherche en sciences et technologies du numérique

Les données ont un rôle crucial pour la recherche en sciences du numérique. Leur acquisition, leur disponibilité, leur utilisation peuvent poser des questions éthiques.

Quelles sont les questions éthiques spécifiques :

- Aux données publiquement disponibles (web social notamment) ;
- Aux données confidentielles (vidéo-surveillance, données médicales) ;
- Au stockage, au partage, à la disponibilité de ces données ?

Le GT « accès aux données » de la CERNA

Composition du groupe de travail

- **Christine Balagué** (titulaire de la chaire « réseaux sociaux » à l'IMT, vice-présidence du CNN, CERNA) ;
- **Danièle Bourcier** (DR émérite CNRS, COMETS, CERNA) ;
- **Max Dauchet** (PR émérite Lille 1, président de la CERNA) ;
- **André Loth** (directeur de projet au Ministère des Affaires sociales et de la Santé) ;
- **Guillaume Piolle** (E/C CentraleSupélec, CERNA), animateur du groupe de travail ;
- **Sophie Vulliet-Tavernier** (directeur des relations avec les publics et la recherche de la CNIL, CERNA).

Le GT « accès aux données » de la CERNA

Personnalités auditionnées

- **Serge Abiteboul** (DR Inria, PR affilié à l'ENS Cachan, Académie des Sciences, Conseil National du Numérique) ;
- **Didier Benza** (CIL, RSSI et FSD à Inria) ;
- **François Bourdoncle** (co-fondateur d'Exalead, président de FB& Cie) ;
- **Nozha Boujemaa** (DR Inria, directrice du centre de recherche Inria Saclay Île-de-France, directrice de l'Institut Société Numérique) ;
- **Dominique Cardon** (sociologue à Orange Labs) ;
- **Claude Castelluccia** (DR à Inria) ;
- **Michel Dojat** (DR Inserm, co-responsable du nœud IAM de France Life Imaging) ;
- **Claude Kirchner** (DR Inria, pdt du COERLE, CERNA) ;
- **Nicolas Lechopier** (MC université de Lyon 1).

Les données dans la recherche en numérique

Une manne inépuisable

Les sciences du numérique, informatique en tête, ont cessé d'être des disciplines purement formelles. De nombreux travaux s'appuient sur la disponibilité ou la possibilité d'élaborer des jeux de données d'un **volume**, d'une **diversité** et d'une **richesse** considérables.

L'exploitation de données liées à l'humain, au social, au vivant. . . permet l'émergence d'**algorithmes** et d'**architectures** spécifiques et originaux, ainsi que l'inférence de nouvelles données.

Les données dans la recherche en numérique

Des problématiques nouvelles

L'informatique et ses disciplines associées n'ont pas historiquement vocation à traiter de l'humain.

Absence d'une culture et d'une méthodologie de questionnement éthique (et déontologique).

La manipulation (possiblement en masse) de données ayant trait à l'humain met le chercheur dans une situation de responsabilité (relativement) nouvelle et engendre des risques spécifiques dont il faut avoir conscience.

Les données dans la recherche en numérique

Les questions éthiques naissent à la frontière

- De nombreux projets de recherche sont **inter-** ou **trans-disciplinaires**, impliquant informatique et sociologie, psychologie, philosophie, droit, médecine...
- Les besoins en données proviennent souvent des exigences liées aux autres disciplines ;
- Les co-disciplines peuvent bénéficier d'une expérience antérieure quant à ces problématiques.

Les données dans la recherche en numérique

Les questions éthiques naissent à la frontière

- De nombreux projets de recherche sont **inter-** ou **trans-disciplinaires**, impliquant informatique et sociologie, psychologie, philosophie, droit, médecine...
- Les besoins en données proviennent souvent des exigences liées aux autres disciplines ;
- Les co-disciplines peuvent bénéficier d'une expérience antérieure quant à ces problématiques.

- Quels enseignements tirer de ces expériences ?
- Quelles sont les problématiques spécifiques au numérique ?

Les données dans la recherche en numérique

Des questions également appréhendées par le droit

Le droit peut régir la manipulation de données indépendamment d'une activité de recherche (protection des données personnelles, de la vie privée, propriété intellectuelle, régimes de secret. . .)

Dans certains domaines, le droit répond en partie à des préoccupations de nature éthique.

Les données dans la recherche en numérique

Des questions également appréhendées par le droit

Le droit peut régir la manipulation de données indépendamment d'une activité de recherche (protection des données personnelles, de la vie privée, propriété intellectuelle, régimes de secret. . .)

Dans certains domaines, le droit répond en partie à des préoccupations de nature éthique.

- Quelle rôle social pour le chercheur, entre légalisme et transgression ?
- Quelle appréhension par le droit de la recherche **en sciences et technologies du numérique** ?

Protection des données personnelles

La protection des données personnelles régit la manipulation de toute information se rapportant à une personne physique potentiellement identifiable, soit la majeure partie des données traitées par la recherche ou la société civile.

Le cadre juridique s'applique au chercheur, mais prévoit un certain nombre d'exceptions visant à permettre l'accomplissement de ses missions.

Le CIL (Correspondant Informatique et Libertés) de l'établissement est un interlocuteur privilégié pour les questions juridiques relevant de la protection des données personnelles.

Protection des données personnelles

Les données de la recherche et la « coopétition » internationale

L'activité des chercheurs est influencée, d'une manière ou d'une autre, par la protection des données personnelles, la loi « Informatique et Libertés », la CNIL.
Les chercheurs sont parfois amener à s'**auto-censurer** et se trouvent souvent en décalage par rapport aux pratiques observées dans d'autres pays.
$\rightarrow$ perception d'un certain **handicap** (en termes de liberté opérationnelle) des équipes françaises dans certains domaines.

Protection des données personnelles

Les données de la recherche et la « coopétition » internationale

L'activité des chercheurs est influencée, d'une manière ou d'une autre, par la protection des données personnelles, la loi « Informatique et Libertés », la CNIL.
Les chercheurs sont parfois amener à s'**auto-censurer** et se trouvent souvent en décalage par rapport aux pratiques observées dans d'autres pays.
$\rightarrow$ perception d'un certain **handicap** (en termes de liberté opérationnelle) des équipes françaises dans certains domaines.

- Quelle est la part réelle et la part fantasmée dans les contraintes pesant sur les chercheurs ?
- Comment mieux former et informer les chercheurs sur leur cadre réglementaire de travail ?
- Comment établir des protocoles permettant de mener sainement des recherches sur des données potentiellement problématiques ?

Protection des données personnelles

Loyauté de la collecte

Une collecte de données personnelles doit être **loyale**, c'est-à-dire que les personnes concernées doivent en être correctement informées, afin de pouvoir faire valoir leurs droits (notamment d'opposition).

Dans d'autres disciplines, on peut reconnaître le besoin d'une collecte déloyale, pour ne pas introduire de **biais** dans l'étude :

- Information incomplète du participant ;
- Information fausse donnée au participant.

On s'appuie alors sur des protocoles spécifiques offrant un certain nombre de garanties et pouvant inclure un accompagnement personnalisé.

Protection des données personnelles

Loyauté de la collecte

Une collecte de données personnelles doit être **loyale**, c'est-à-dire que les personnes concernées doivent en être correctement informées, afin de pouvoir faire valoir leurs droits (notamment d'opposition).

Dans d'autres disciplines, on peut reconnaître le besoin d'une collecte déloyale, pour ne pas introduire de **biais** dans l'étude :

- Information incomplète du participant ;
- Information fausse donnée au participant.

On s'appuie alors sur des protocoles spécifiques offrant un certain nombre de garanties et pouvant inclure un accompagnement personnalisé.

La recherche dans le domaine du numérique introduit-elle des besoins similaires, ou bien d'autres motivations pour recourir à une collecte déloyale ?

Protection des données personnelles

Exemple d'expérimentation dans la recherche en vie privée

Le chercheur se place dans un environnement peuplé (centre commercial, gare...) avec une antenne wifi et écoute les « trames d'administration » émises par les smartphones et ordinateurs alentour (pas d'interception ou de capture des communications des utilisateurs).

L'expérience permet de mettre en lumière les informations diffusées automatiquement et les brèches de vie privée qu'elles pourraient permettre.

Protection des données personnelles

Exemple d'expérimentation dans la recherche en vie privée

Le chercheur se place dans un environnement peuplé (centre commercial, gare...) avec une antenne wifi et écoute les « trames d'administration » émises par les smartphones et ordinateurs alentour (pas d'interception ou de capture des communications des utilisateurs).

L'expérience permet de mettre en lumière les informations diffusées automatiquement et les brèches de vie privée qu'elles pourraient permettre.

- Quels moyens a-t-on pour informer ces personnes ?
- Les prive-t-on de l'exercice d'un droit, leur cause-t-on du tort, est-ce une nouvelle source de risque pour elles ?
- L'intérêt de l'étude le justifie-t-il ?
- Pourrait-on reproduire cette étude avec des participants volontaires, « en laboratoire » ?

Protection des données personnelles

Conservation et utilisation pour une nouvelle finalité

Il est possible, à des fins de recherche, de conserver **indéfiniment** des données à caractère personnel et de les réutiliser pour de nouvelles études (même si les données ont été collectées pour une finalité autre que la recherche).

Protection des données personnelles

Conservation et utilisation pour une nouvelle finalité

Il est possible, à des fins de recherche, de conserver **indéfiniment** des données à caractère personnel et de les réutiliser pour de nouvelles études (même si les données ont été collectées pour une finalité autre que la recherche).

- Quelle responsabilité morale pour le chercheur qui décide de « mettre de côté » tel ou tel jeu de données ?
- Comment choisir ce qui doit être conservé ?
- Comment évaluer le risque associé aux données, comment déterminer les mesures de protection à adopter ?
- Quelle responsabilité vis-à-vis des personnes concernées, qui peuvent en théorie ne rien savoir de cette conservation ou réutilisation ? Devrait-on aller au-delà des exigences légales en matière d'information des personnes et de consentement ?

Protection des données personnelles

Consentement et opposition

Un traitement, une collecte ne peuvent normalement avoir lieu qu'avec le **consentement** des personnes concernées ou, plus souvent, en l'absence d'une **opposition** exprimée.

Un consentement peut éventuellement être retiré a posteriori.

Protection des données personnelles

Consentement et opposition

Un traitement, une collecte ne peuvent normalement avoir lieu qu'avec le **consentement** des personnes concernées ou, plus souvent, en l'absence d'une **opposition** exprimée.

Un consentement peut éventuellement être retiré a posteriori.

- Qu'est-ce qu'un consentement éclairé dans tel ou tel contexte, est-ce nécessairement l'outil qui convient ?
- Comment prendre en compte consentement, opposition, rétractation dans la conception des systèmes et dans les expérimentations ?
- Un consentement défectueux, une violation des droits des personnes concernées « souille-t-il » le jeu de données, et avec lui les travaux qui s'appuieront dessus ? Peut-on imaginer une sorte de « labellisation » ou de notation des jeux de données ?

Protection des données personnelles

L'utilisation des données du web social

Informations publiées sur Facebook, Yahoo, Twitter... : mises à disposition par les personnes concernées, sont-elles « rendues publiques », sont-elles librement utilisables ?
Ces données ont un intérêt pour de nombreux champs de recherche.

Diversité des pratiques suivant les disciplines, les cultures déontologiques et les époques :

- Mise à disposition de jeux de données sous forme de « défis » ou de concours ouverts ;
- Extraction sauvage depuis les plates-formes (2003-2008), maintenant rendue techniquement difficile et vue comme déontologiquement problématique ;
- Collaboration avec les plates-formes (Couchsurfing, Skyblog, *Facebook Data Team*, Twitter) : participation à des consortiums de recherche ou ouverture de l'API contre rémunération (Twitter).

Protection des données personnelles

CNIL : délibération « Pages Jaunes »

La formation restreinte considère que la circonstance que des profils personnels sont affichés publiquement sur Internet ne permet pas pour autant à un organisme tiers de procéder à une collecte massive, répétitive et indifférenciée de ces données sans en avertir les personnes concernées.

- Quelles méthodologies pour l'**information**, le recueil du **consentement** (variable en nature et en valeur), pour l'exercice des droits ?
- Comment prendre en compte le problème de la **capacité d'inférence** parfois considérable des algorithmes étudiés ?

Recoupement et réidentification

Si les données manipulées se réfèrent à des personnes physiques, il peut exister un risque d'associer à ces personnes des informations qui devraient rester privées, ou qui pourraient leur causer du tort d'une manière ou d'une autre.

Cela peut être dû au jeu de données initial, aux recoupements entre jeux de données ou au traitement effectué dans le cadre des travaux de recherche.

→ intérêt pour un **assainissement** des données visant à rendre difficile l'identification des personnes

La pseudo-anonymisation d'un jeu de données est rarement efficace à 100 % et les techniques à utiliser dépendent beaucoup de la nature des données

Recoupement et réidentification

Reconnaissance automatique d'auteurs et victimes dans des vidéos pédopornographiques

Utilité sociale indéniable, mais conception problématique : travail de recherche sur des jeux de données « délicats ».

- Comment peut-on exposer ses collègues, ses étudiants à un matériau de cette nature ?
- Quelles **responsabilités** en cas de problèmes de confidentialité ?
- Est-ce pertinent de parler de la notion de **consentement** des personnes concernées pour le traitement des données ?
- Problème juridique de la manipulation de ces données, le chercheur n'agit pas dans le même cadre que la police judiciaire.

Utilité et pertinence de travailler sur des jeux de données de **synthèse** ou de **substitution** ?

Recoupement et réidentification

Reconnaissance et caractérisation de comportements suspects

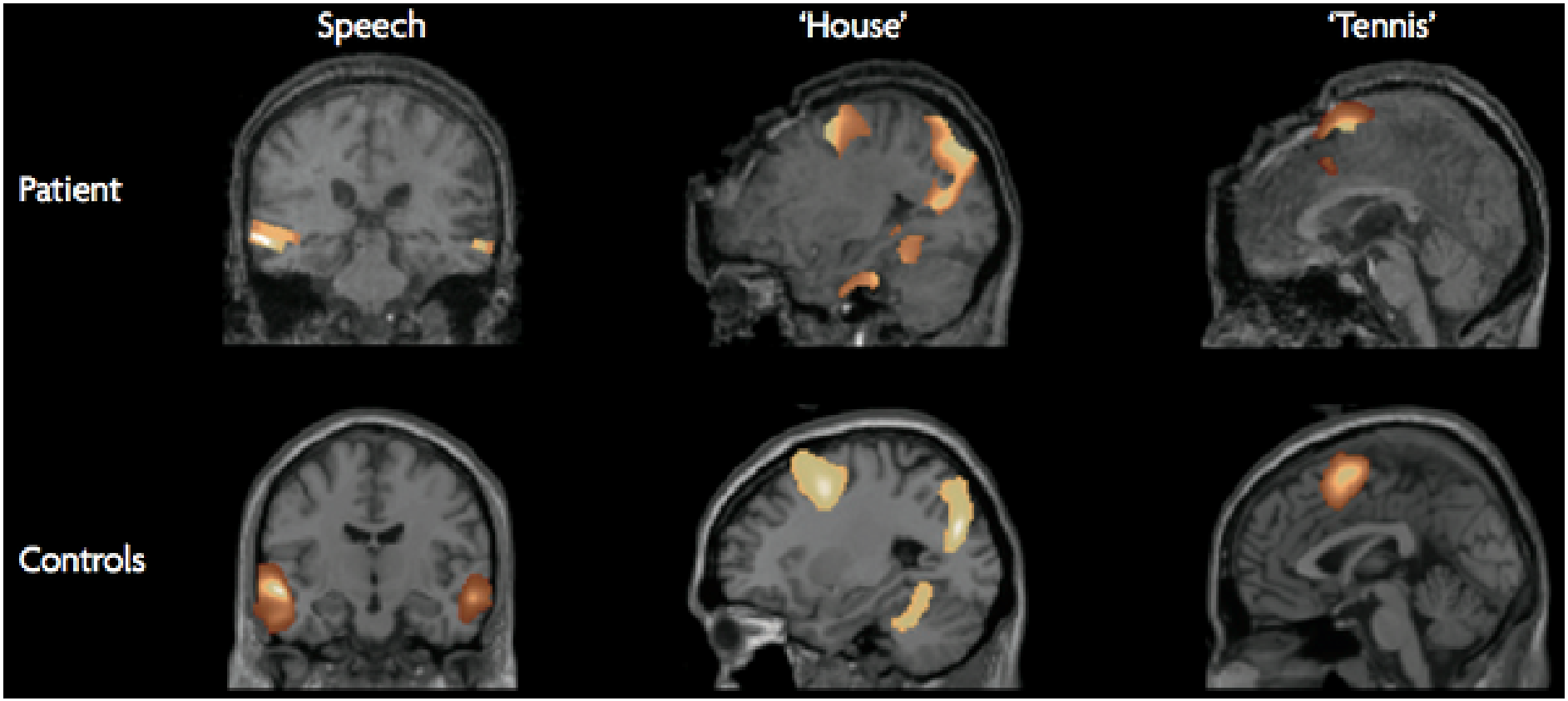
Objectif de prévention d'actions violentes dans un espace public

- Question fondamentale de la **prédiction comportementale**, de la nature d'une **recommandation automatique**, du risque d'éliminer l'interprétation humaine, de dériver vers une décision automatique, éventuellement à l'insu des personnes ;
- Comment les jeux de données d'apprentissage/validation sont-ils constitués ? Avec quels biais ? Risque d'**automatisation de préjugés**, **mécanisation de la discrimination** ;
- Droits des personnes ayant participé à la constitution des jeux de données (information sur la finalité, consentement/opposition) ;

Quelle possibilité pour le chercheur de se dégager d'une étude qu'il estime trop problématique du point de vue éthique ? Quel accompagnement des institutions ?

Recoupement et réidentification

Recoupement et réidentification



Partage et réutilisation

Conservation, partage et reproductibilité

Conserver ses données expérimentales est un impératif pour permettre la validation des résultats et la mise en œuvre de recherches complémentaires.

La **reproductibilité des résultats** est, dans les disciplines du numérique, un problème très prégnant, avec des spécificités liées à la propriété intellectuelle.

→ Nécessité du **partage** des données avec d'autres chercheurs ;
→ Intérêt des **plates-formes de partage des données** (ressources mutualisées, meilleure sécurité, large accès, procédures claires. . .).

Partage et réutilisation

Problématiques liées aux plates-formes de partage

- Comment décider d'accorder l'accès à un jeu de données particulier ? En fonction du statut du demandeur ? De sa nationalité ? De la nature de ses travaux ?
- Friction avec le principe de **finalité** de la protection des données personnelles, avec l'obligation d'information des personnes ;
- Comment caractériser la **qualité** des données partagées, quels jeux de données mettre en avant ? Critères scientifiques, critères déontologiques ou « éthiques », liés aux protocoles suivis, à la protection des personnes ?
- Comment évaluer la **sensibilité** des données, comment déterminer la nature et le niveau des mesures de protection associées ?
- Nécessité de permettre l'exécution de **traitements personnalisés** – mais comment ?

Partage et réutilisation

Le projet IAM de *France Life Imaging*

Projet de plate-forme de mise en commun de jeux de données d'**imagerie médicale**, principalement IRM, principalement du cerveau (mais pas uniquement de sujets humains).

Permet de travailler sur des jeux de données plus volumineux, alors que la constitution des bases est longue, difficile et coûteuse.

Mise à disposition de chaînes de traitement standard, documentation sur la production de chaque jeu de données.

A priori à disposition de n'importe quel projet de recherche (sur autorisation de la CNIL).

Partage et réutilisation

Questions liées à la plate-forme IAM

- Nécessité de trouver un modèle économique correspondant aux intérêts, droits et investissements de chacun ;
- Risque de réidentification difficile à évaluer, pas toujours bien étudié ;
- Anonymat ou rétractation du consentement ?
- Communiquer sur le respect des droits des personnes pendant la collecte et le traitement ?

Partage et réutilisation

Recherche en sécurité informatique : exploitation et partage de traces d'attaques

Source d'informations essentielle pour la connaissance de la menace informatique, la conception des moyens de détection/protection et leur évaluation.

- Question de la loyauté de la collecte et des droits des personnes d'une manière générale (intérêt d'une pseudo-anonymisation, d'un assainissement ?) ;
- Cadre juridique des données liées aux infractions ;
- Difficulté de l'obtention de jeux de données pertinents et exploitables ;
- Valeur stratégique particulière de ces jeux de données.

Beaucoup de raisons qui rendent le partage difficile, malgré les besoins de la communauté.

Sujets connexes

Anonymat dans la conception de technologies numériques

Question récurrente de l'**anonymisation** des jeux de données expérimentales, du travail sur des bases de données anonymes ou anonymisées

Fait l'objet d'une saisine et d'un groupe de travail de la CERNA

- Quelle évolution historique et culturelle pour les notions d'identité, d'anonymat, de pseudonymat ?
- État de l'art techniques et possibles recommandations pour :
  - Protection de l'anonymat, gestion des identités multiples, des pseudonymes ;
  - Anonymisation, pseudonymisation, assainissement de données publiques ;
  - Mesure de la qualité d'une anonymisation et des risques de réidentification.

Prédiction comportementale

Importance du concept de prédiction

Une théorie scientifique est se caractérise par sa capacité à énoncer des prédictions. L'objectif de la plupart des disciplines est de prédire un résultat vérifiable à partir de conditions initiales connues.

La prédiction permet de **planifier** et de **concevoir** pour infléchir le résultat attendu dans le sens d'un plus grand bien (on espère !)

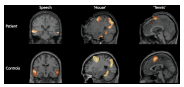
- Lorsqu'elle s'applique à l'humain, la notion de prédiction peut entrer en conflit avec le concept de libre arbitre ;
- Le fait qu'une partie seulement d'une population ait accès à des résultats prédictifs peut engendrer des inéquités de diverses natures ;
- Il est tentant de construire des prédictions purement statistiques pouvant fournir des résultats erronés.

Manipulation de code malveillant

Il est essentiel pour la recherche en sécurité informatique de pouvoir disposer d'un corpus de souches virales riche et tenu à jour.
Idéalement, ce corpus devrait être partagé entre les chercheurs du domaine pour permettre une collaboration la plus large et la plus efficace possible.

- Les logiciels malveillants sont des données nuisibles en elles-mêmes, qui ne nécessitent pas nécessairement une action humaine pour causer du tort ;
- Rendre public du code malveillant n'est normalement pas permis, et pourrait avoir un effet négatif (encouragement à utiliser ces logiciels ou à en produire de nouveaux) ;
- Les nécessaires restrictions sur l'accès à ces bases de données peut introduire des inéquités au sein de la communauté scientifique.

Ressources iconographiques

-  Owen, A. M. & Coleman, M. R., *Functional neuroimaging of the vegetative state*. Nature Reviews Neuroscience, 9(3), 2008, pp. 235-243 (`http://www.wbic.cam.ac.uk/Members/icrg/research/documents/owen-coleman-nnr-2008.pdf`) ;
-  El Guendouz, F., Hamoune, N. et Hommadi, A., *Syndrome alterne du tronc cérébral révélant un prolactinome géant*, Research fr 2014 ;1 :1092, 2014 (`http://www.labome.fr/research/Giant-prolactinoma-revealed-by-Alternating-hemiplegia.html`).