55

Le bulletin d'information de l'UPS 2259

focus

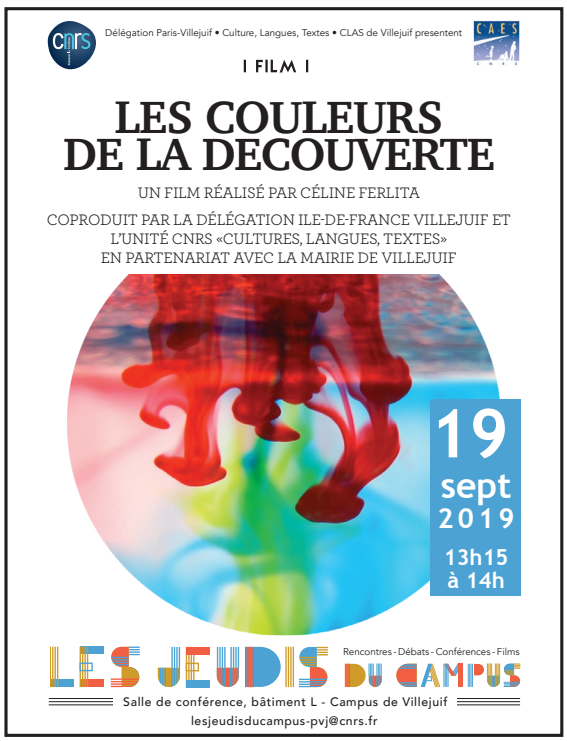
Septembre 2019

Sommaire

Info du mois

Jeudis du campus – 19 Septembre à 13h15

Projection de «Les couleurs de la découverte» (27 min.)
En présence de Céline Ferlita, réalisatrice à « Cultures, Langues Textes » dans la salle de conférence du Bâtiment L.



Coproduit par la délégation CNRS Ile-de-France Villejuif et l'unité CNRS « Cultures, Langues, Textes », en partenariat avec la mairie de Villejuif.

Avec,
Philippe Lognonné (IPGP), Matteo Barsuglia (APC), Mourad Bensidhoum (B3OA), Lluis Mir (Laboratoire de vectorologie et thérapeutiques anticancéreuses), Iordanis Kerenidis (IRIF), Thomas Coudreau et Perola Millman (MPQ), Matthias Beekmann (LISA), Michel Latroche (ICMPE), David Chavalarias (ISCPIF), Martial Foucault (CEVIPOF)

Le film retrace les 5 rencontres inédites entre habitant.es du Val-de-Marne et équipes de recherche de la Délégation, qui ont eu lieu entre mai et juin 2019 à la **Médiathèque Elsa-Triolet.** De quoi explorer par un angle original 10 découvertes scientifiques qui ont marqué la dernière décennie, dans tous les champs disciplinaires investis par le CNRS.

Nous espérons vous voir nombreux dans la salle bâtiment L jeudi 19 Septembre à 13h15.

Vie du CNRS

« Sciences. Bâtir de nouveaux mondes ». Un ouvrage à découvrir pour les 80 ans du CNRS

Depuis 80 ans, nos connaissances bâtissent de nouveaux mondes

Un ouvrage inédit présente, en 80 textes, comment le CNRS et plus généralement la recherche publique ont accompagné les grandes mutations de la société et contribué à bâtir de nouveaux mondes. Publié chez CNRS Editions sous la direction de Denis Guthleben, Sciences. Bâtir de nouveaux mondes sera disponible en librairie à partir du 12 septembre 2019.

https://www.cnrs.fr/fr/sciences-batir-de-nouveaux-mondes-un-ouvrage-decouvrir-pour-les-80-ans-du-cnrs

Service audiovisuel/communication

Mise en ligne sur CANAL U « Des asiatiques en île-de-France : « Nouveaux regards, nouvelles images »»

Ce Colloque a été organisé par Simeng Wang (CERMES 3) dans le cadre du «Tour du CNRS en 80 jours», avec le soutien du CNRS, de la Mairie de Paris et de l'IEA de Paris. Cet événement s'inscrit dans le programme "Emergence(s)" Chinoises en (Île de) France : identifications et identités en mutations. Pour voir toutes nos conférences : https://www.canal-u.tv/producteurs/cnrs_ups2259/conferences

Ouverture du colloque :

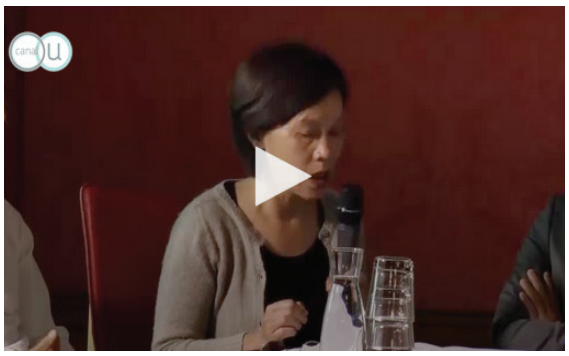


Table-Ronde
"Participations politiques et citoyenneté" :

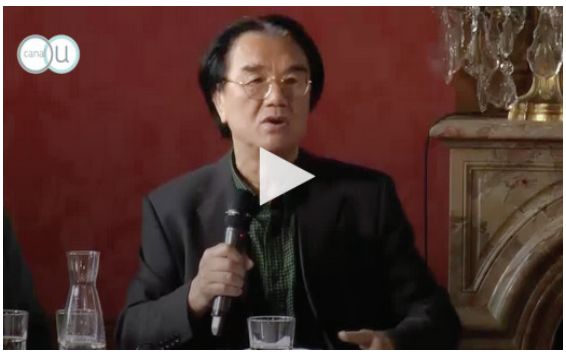


Table-ronde
"Transmissions, générations et mémoire"

Changements au centre de documentation AGH

En cette rentrée, de nombreux changements ont eu lieu sur le campus et en particulier au centre de documentation. D'une part le départ en fanfare du CEH, et celui, à peine plus discret d'Elodie CHACON, responsable du centre de documentation, remplacée dans ces fonctions par **Jean Baptiste MAISTRE.**

C'est donc une équipe restreinte mais toujours aussi dynamique qui accueille les lecteurs et prend soin des collections depuis le premier septembre.

Cette nouvelle configuration demande une réorganisation du service et en particulier de la façon dont nous répondons à vos demandes. Il ne nous est en effet plus possible à deux, d'assurer une ouverture systématique de la bibliothèque chaque jour de la semaine. Nous vous proposons donc de nouveaux horaires d'ouverture :
les mardis et jeudis de 10h00 à 12h00 et de 13h00 à 17h00.

Les autres jours de la semaine, nous répondrons à vos demandes selon nos disponibilités, nous ouvrirons les portes du centre de documentations lorsque nous serons là, **n'hésitez pas alors à nous solliciter**.

Par ailleurs, il vous sera toujours possible de nous contacter par mail pour vous assurer de notre présence au moment où vous comptez venir.
documentation.haudricourt@vjf.cnrs.fr

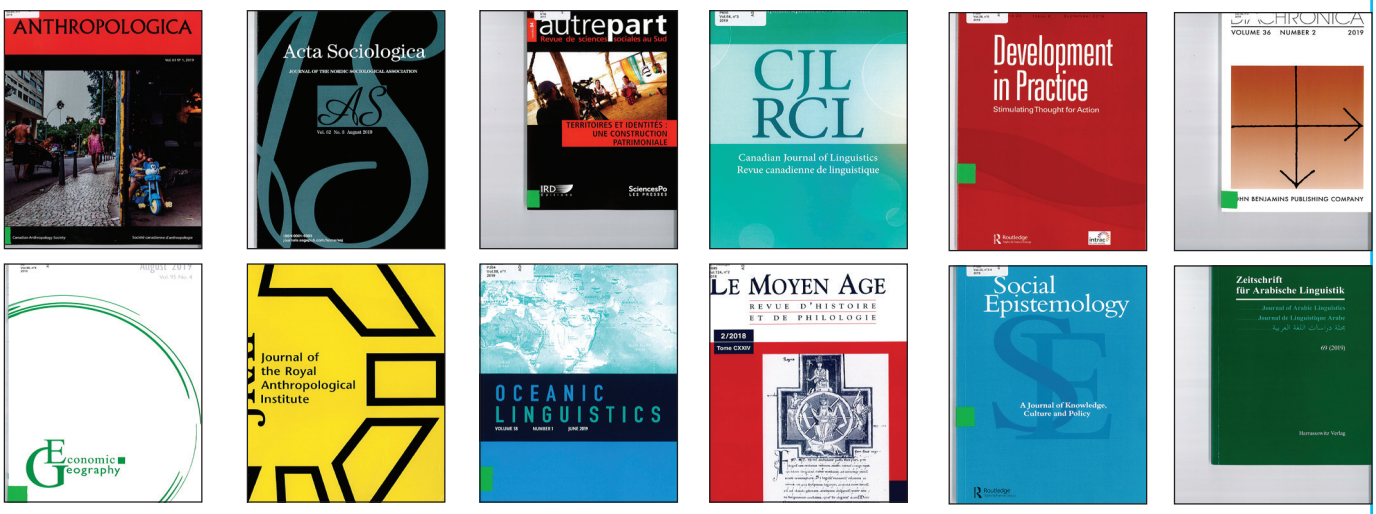
En parallèle, nous travaillons, avec le reste de l'unité, à vous faciliter l'accès aux documents électronique de nos fonds, cela devrait prendre la forme d'une bibliothèque numérique sous **Omeka.**

Bien entendu, nous espérons que le poste d'Elodie sera remplacé dans des délais raisonnables afin de pouvoir à nouveau élargir nos horaires d'accueil.
Nous vous attendons nombreux au centre de documentation et vous rappelons que le formulaire de réinscription pour l'année 2019-2020 est disponible sur place ou directement sur notre site :
https://www.vjf.cnrs.fr/clt/v3/spip.php?article2

Catalogue AGH

Nouvelles acquisitions

Voici une sélection des documents reçus récemment au Centre et acquis par l'unité ou par les laboratoires partenaires.



N'hésitez pas à faire des propositions tout au long de l'année.

Les prolongements des mots (1) Word Embedding (1)

#TAL #ApprentissageProfond #NLP #DeepLearning

La question « comment représenter les unités du vocabulaire (les mots) dans un espace numérique ? » a toujours occupé les experts en TAL et en intelligence artificielle.

On appelle « dénotation sémantique » le concept qui consiste à représenter une idée ou un objet (signifiant) par un symbole (signifié).

Signifier (symbol)  Signified (idea or thing)

Cette représentation a été interprétée comme une représentation de location « localist representation » ce qui revient à exprimer chaque mot en un vecteur qui contient que des « 0 » et seulement un et un seul « 1 » à l'index qui correspond au mot dans un espace qui contient tout le vocabulaire d'une langue.

voiture = [0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0]

Cette représentation est appelée par « one-hot vector » ou « 1 of n » qui est la première et la plus simple représentation vectorielle des mots « word-vector ».

Cette présentation a évolué mais elle a toujours gardé un problème essentiel appelé par « la malédiction de la dimensionnalité ». En effet, la dimension d'un vecteur de mot est le nombre de toutes les unités du vocabulaire d'une langue. La taille de l'espace dépasse les 500.000 dimensions voir un million de dimensions ce qui rend les traitements inefficaces et très coûtent en termes de calcul et de ressources matérielles.

De plus, chaque mot est représenté comme une entité complétement indépendante car les vecteurs sont orthogonaux et les synonymes par exemples ont des représentations complétement différentes. Par conséquent, ces représentations ne pourraient décrire aucune notion de similarité entre les unités du vocabulaire.

Donc peut-être nous pouvons réduire l'espace de représentation vectorielle vers un espace plus réduit et puis trouver des sous-espaces qui permettront d'encoder les relations entre les mots.

En effet, l'intérêt principal d'avoir des vecteurs de mots word-vector est de coder dans les vecteurs eux-même la possibilité d'avoir des notions de similitude et de différence entre les mots à l'aide des mesures de distance comme Jaccard, Cosinus ou euclidien.

Nous ne pouvons pas coder les vecteurs de mots manuellement, raisons pour lesquels nous utilisons un ensemble de méthode qui permet suite à un entrainement d'apprendre à encoder ces vecteurs.

Tous ces facteurs : la réduction de dimension, la notion de similarité et de différence au sein d'un vecteur et l'ensemble des méthodes qui permettent cette représentation multi-dimensionnelles constituent ce qu'on appelle souvent aujourd'hui « prolongements des mots » ou « word embedding » qui n'est pas une notion récente au contraire de ce qu'on pourrait imaginer.

Alors, la question que nous aborderons lors de nos prochains articles focus : comment pouvons-nous y arriver? Comment pouvons-nous mettre en place une représentation multi-dimensionnelle type word embedding ?

CNRS, UPS 2259, 7 rue Guy Môquet - 94800 Villejuif - Tél : 01 49 58 38 04

Directeur de publication : Bernard Weiss - **Responsable éditoriale :** Céline Ferlita
Création graphique et mise en page : Emmanuelle Seguin et Isabelle Michel

http://www.vjf.cnrs.fr/clt - @Ups2259Cnrs