LE PROJET FULLAB

LISIS (UPEM, INRA, CNRS)

&

DIRECTION DE LA DOCUMENTATION (**ECOLE DES PONTS**)

SÉMINAIRE ISTEX – 25-26 AVRIL 2016

NANCY

OBJECTIFS

AMBITIEUX MAIS ATTEIGNABLES GRÂCE AUX DONNÉES D'ISTEX

OBJET D'ÉTUDE : L'ARTICLE **ET** SON ABSTRACT

- Contexte : évolution des modalités de publication
- Objet d'étude : l'article et son abstract
- Domaine : sciences environnementales

- Objectif global = caractériser l'abstract et étudier :
  - S'il a varié au fil du temps
  - S'il diffère d'un domaine à l'autre

→ **Selon quels critères ?**

ABSTRACT : SIMPLE *TEASER* ?

- Comparer la quantité d'informations de l'abstract avec celle de l'article qu'il résume
  - Structuration
  - Types d'arguments
  - Catégories d'entités nommées
  - Formes linguistiques

- Pour un calcul du **taux de générosité** (à définir…)

→ **Des perspectives d'étude**

EXPLOITER LE TAUX DE GÉNÉROSITÉ

- Selon la discipline : différences de comportement des chercheurs ? D'exigence des revues ?
- Selon le mode de diffusion : en Open Access ou pas ?
- Corrélation Taux de générosité/nombre de citations ?
- Prédominance de certains types d'entités nommées ?
- Existence d'une éventuelle évolution au fil du temps ?

DES APPLICATIONS

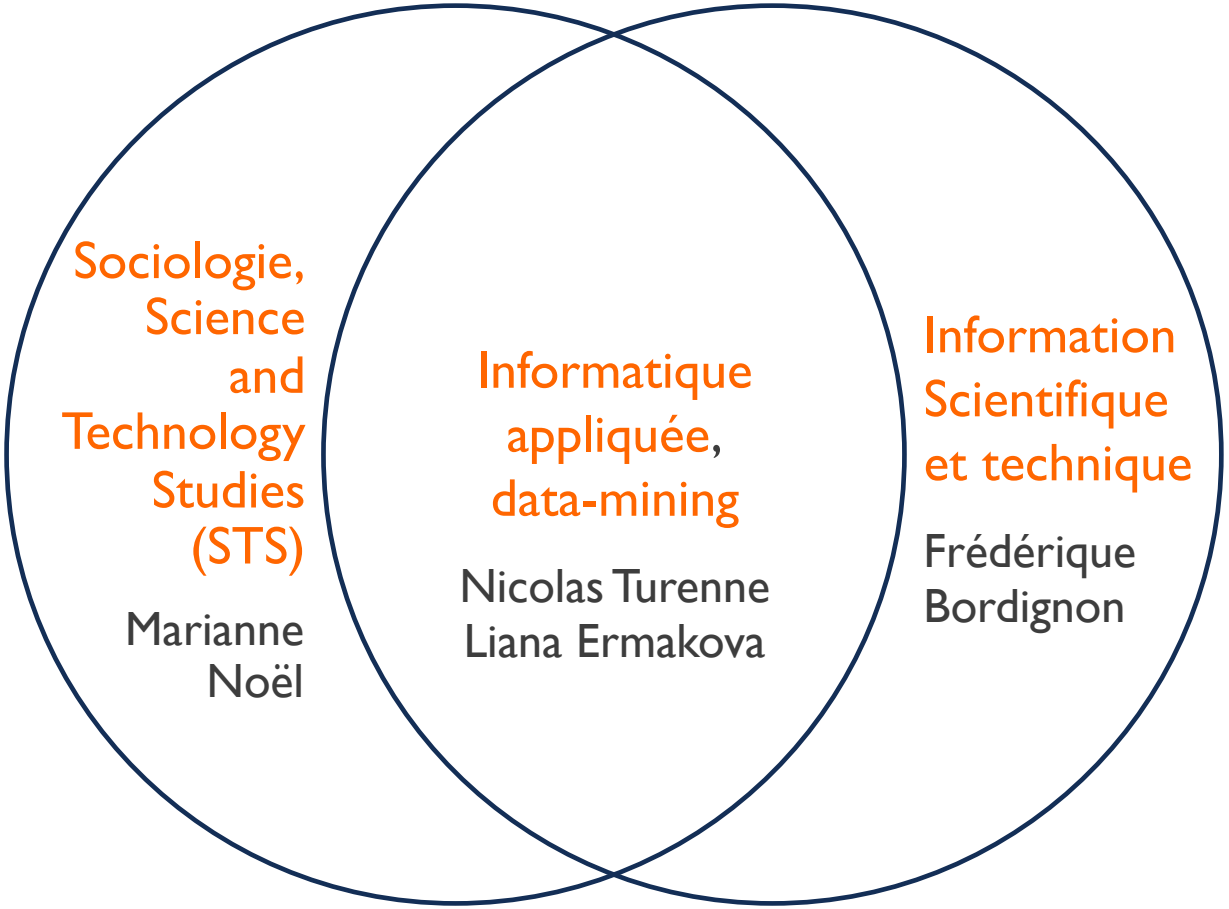
- Pour le professionnel de l'IST :
  - Meilleure connaissance de la littérature scientifique
  - Meilleure connaissance de ses modalités de diffusion
  - Progrès dans la méthodologie de recherche
  - Meilleure stratégie économique (choix des abonnements)

- Pour le chercheur :
  - Clés pour la rédaction efficace d'un abstract
  - Génération automatique

- Pour les pays en voie de développement :

  Des clés pour décider s'il est raisonnable ou non de se passer du full-text

L'OPPORTUNITÉ D'UN TRAVAIL INTERDISCIPLINAIRE



Un noyau informatique,
des productions connexe

ÉTAT DE L'ART : SUR LES MODALITÉS DE PUBLICATION ET LA NATURE DE L'ABSTRACT

- *Abstracting* : né avec l'apparition de l'écriture et des liens sociaux ? Question récurrente et ancienne : celle de l'abondance de la littérature scientifique.

- Polysémie du mot : aperçu, *brief*, compendium, *digest*, résumé, *summary*, *survey*, *synopsis*,…

- Pratiques disciplinaro-dépendantes.

- Les contraintes extérieures, en particulier les consignes données par l'éditeur, en formatent-elles l'écriture ?

How to construct a *Nature* summary paragraph

Annotated example taken from *Nature* 435, 114–118 (5 May 2005).

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline, may be included in the first paragraph if the editor considers that the accessibility of the paper is significantly enhanced by their inclusion. Under these circumstances, the length of the paragraph can be up to 300 words. (This example is 190 words without the final section, and 250 words with it).

During cell division, mitotic spindles are assembled by microtubule-based motor proteins[1,2]. The bipolar organization of spindles is essential for proper segregation of chromosomes, and requires plus-end-directed homotetrameric motor proteins of the widely conserved kinesin-5 (BimC) family[3]. Hypotheses for bipolar spindle formation include the 'push–pull mitotic muscle' model, in which kinesin-5 and opposing motor proteins act between overlapping microtubules[2,4,5]. However, the precise roles of kinesin-5 during this process are unknown. Here we show that the vertebrate kinesin-5 Eg5 drives the sliding of microtubules depending on their relative orientation. We found in controlled *in vitro* assays that Eg5 has the remarkable capability of simultaneously moving at ~20 nm s^{-1} towards the plus-ends of each of the two microtubules it crosslinks. For anti-parallel microtubules, this results in relative sliding at ~40 nm s^{-1}, comparable to spindle pole separation rates *in vivo*[6]. Furthermore, we found that Eg5 can tether microtubule plus-ends, suggesting an additional microtubule-binding mode for Eg5. Our results demonstrate how members of the kinesin-5 family are likely to function in mitosis, pushing apart interpolar microtubules as well as recruiting microtubules into bundles that are subsequently polarized by relative sliding. We anticipate our assay to be a starting point for more sophisticated *in vitro* models of mitotic spindles. For example, the individual and combined action of multiple mitotic motors could be tested, including minus-end-directed motors opposing Eg5 motility. Furthermore, Eg5 inhibition is a major target of anti-cancer drug development, and a well-defined and quantitative assay for motor function will be relevant for such developments.

ÉTAT DE L'ART : SUR L'EXTRACTION DE CONNAISSANCES

- Approche d'extractions d'entités nommées
(semi-supervisé : règles, thésaurus, supervisé : CRF)

- Approche d'extraction de multi-termes
(non-supervisé : n-grammes, semi-supervisé : règles)

- Approche de résumé automatique
(sélection de phrases par pondération lexicale)

- Approche d'annotation du discours
(modèle et corpus RST, *rhetorical structure theory*)

LISTE DES TÂCHES

Revue de la littérature et analyse des développements existants d'ISTEX

Travail empirique et automatisé de constitution des corpus (informatique, environnement, chimi

Etude des indices de diversité lexicale et définition d'un **indice de générosité**

Extraction d'Entités Nommées (T1) : utilisation du logiciel x.ent, création des dictionnaires, créa
de grammaires locales, applications des grammaires de schémas argumentatifs

Extractions terminologiques ISTEX (T2) : utilisations des outils d'extractions ISTEX. Analyse
différentielle Résumés / Corps des publications (T3) : définitions de métriques de comparaison,
visualisations chronologiques comparatives (diagrammes en barres, boites à moustaches)

Etudes de cas : applications de T1, T2 et T3 sur différents corpus et validation avec des experts

Création d'un site web : mise en place d'un site web qui intègre l'accès aux sous-corpus, ainsi qu
briques logicielles de traitement de données (R notamment)

Présentation sur le site d'ISTEX des expérimentations

BESOINS

RESSOURCES HUMAINES, DONNÉES ET OUTILS

- Post-Doc
- Moyens de calcul distribué

RÉSULTATS ESPÉRÉS

AUX CONFINS DE PLUSIEURS DISCIPLINES

PRODUIRE UN OUTIL DE TEXT-MINING

- Constitution d'un ou plusieurs corpus à partir des données ISTEX
- Elaboration d'un cadre analytique
- Elaboration d'un outil de traitement, prototypage
- Création d'une interface *user-friendly*

PRODUIRE DE LA CONNAISSANCE

- Publications dans des revues du champ des STS, de l'histoire et de la sociologie des sciences, des sciences de l'information-communication
- En rapport avec les professionnels de l'IST

MERCI DE VOTRE ATTENTION

Frédérique Bordignon

Liana Ermakova

Marianne Noël

Nicolas Turenne