

Bienvenue en IC<sup>5</sup>

Printemps 2010

# Technologies pour la documentation et l'indexation dans l'hypermédia

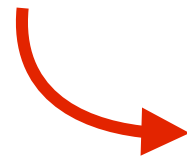
Enseignée par Franck Ghitalla

# Analyse Exploratoire de Données Numériques



Technologies pour la documentation et  
l'indexation dans l'hypermédia

Enseignée par **Franck Ghitalla**



**Fabien Pfaänder**

# Analyse Exploratoire de Données Numériques

Enseignée par **Fabien Pfaänder**

Secondé par Guilhem Fouetillou

& Mathieu Jacomy

Fabien Pfaänder

Poste 4061

Fabien.pfaender@utc.fr

Laboratoire Costech, groupe  
Cognition Research and  
Enactive Design & Webatlas

K100 (PG)

# Sondage

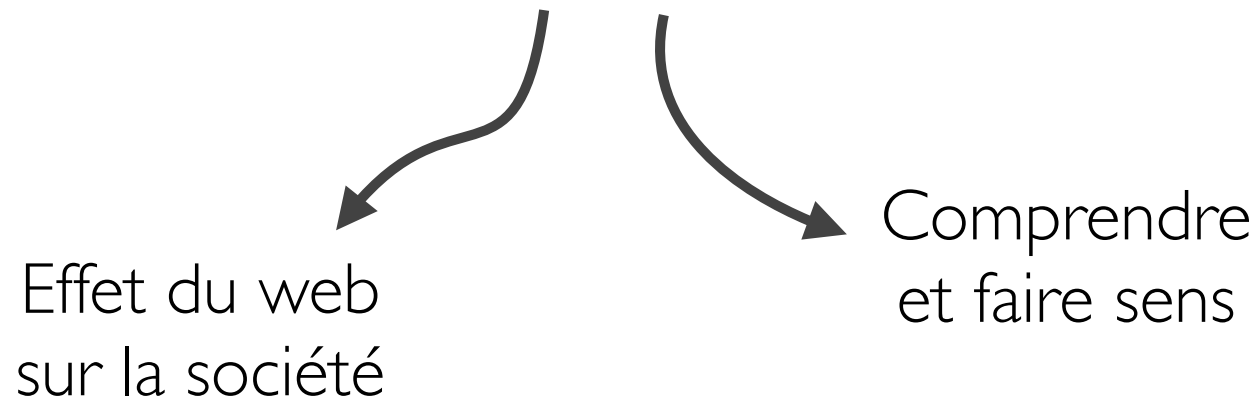
- ▶ Qu'est ce que le Web ?
- ▶ Qu'est ce qu'un crawler ?
- ▶ Qu'est ce qu'une visualisation d'informations ?

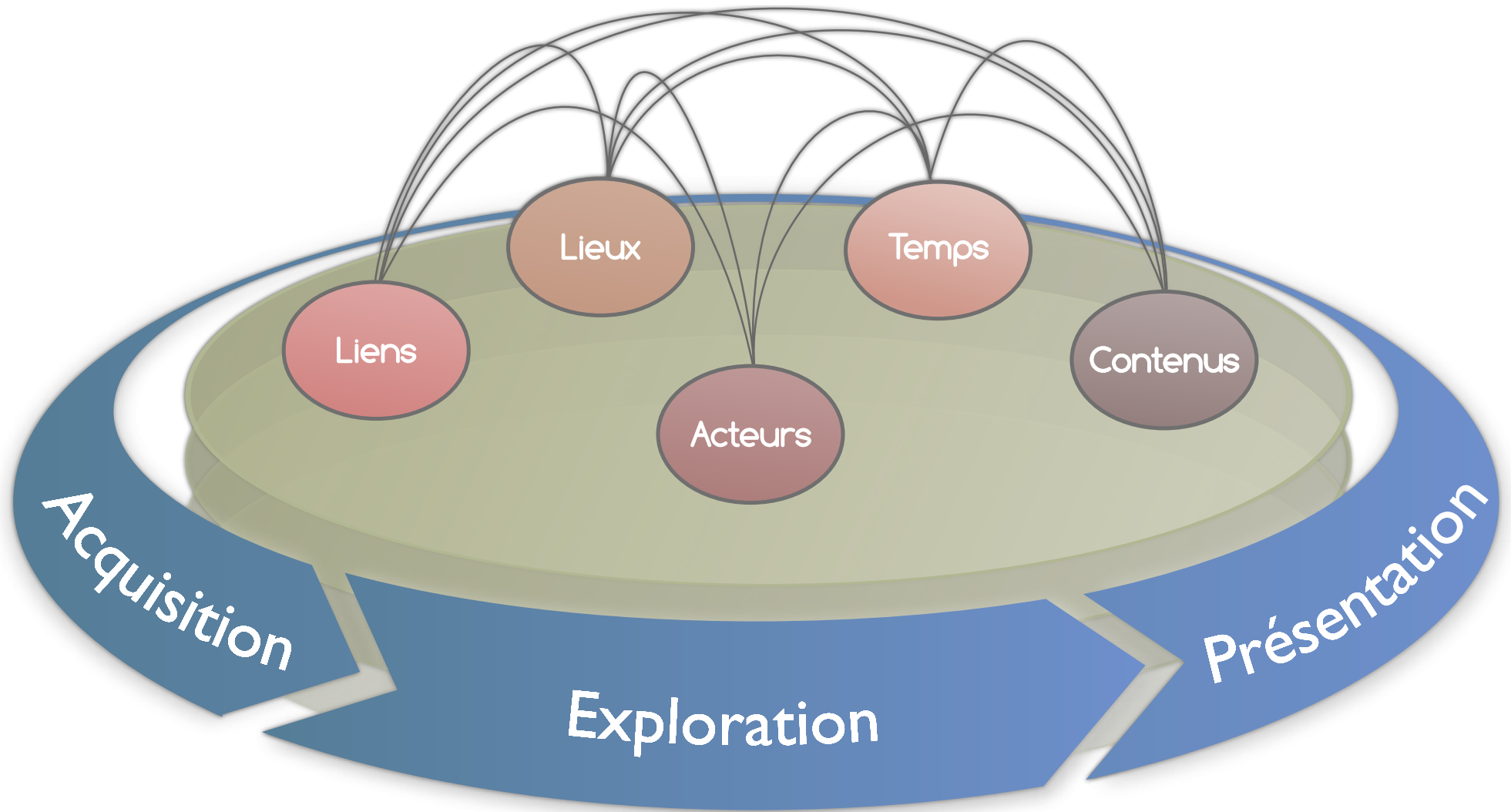
# Exploration et analyse de données hypermédia : Science du web

Nouvelles techniques d'analyse et d'indexation  
du web, découverte de connaissances assistée  
par des visualisations, présentation d'informations.

# Exploration et analyse de données hypermédia : Science du web

Nouvelles techniques d'analyse et d'indexation  
du web, découverte de connaissances assistée  
par des visualisations, présentation d'informations.





# Ce qui sera abordé ce semestre

Présentation de l'UV, mode de validation, organisation et projets

Introduction à l'analyse exploratoire de données (EDA)

Concepts avancés pour l'EDA

Introduction aux web & network science

**Acquisition** : obtenir du contenu web

**Exploration** : petite histoire d'un explorateur du web

**Exploration** : historique des dispositifs de découverte de connaissances

**Exploration** : indicateurs des moteurs de recherche

**Exploration** : la cas du web sémantique

**Présentation** : sémiologie graphique, graphisme d'information

**Présentation** : ergonomie cognitive

En pratique : le web social aujourd'hui

Outils : théorie des graphes, infoviz ou conférence surprise

En pratique : Géomatique et géolocalisation

En pratique : conférence Franck Ghitalla

# Ce qui sera abordé ce semestre

Introduction Navicrawler, discussion projets

navicrawler (suite) et décision projets

Introduction à processing

Application de processing à la dataviz

Conception d'un plugin Gephi

Crawler en perl et visualiser ses résultats

Outils en ligne d'aide à la recherche (insight for search, link:, etc.)

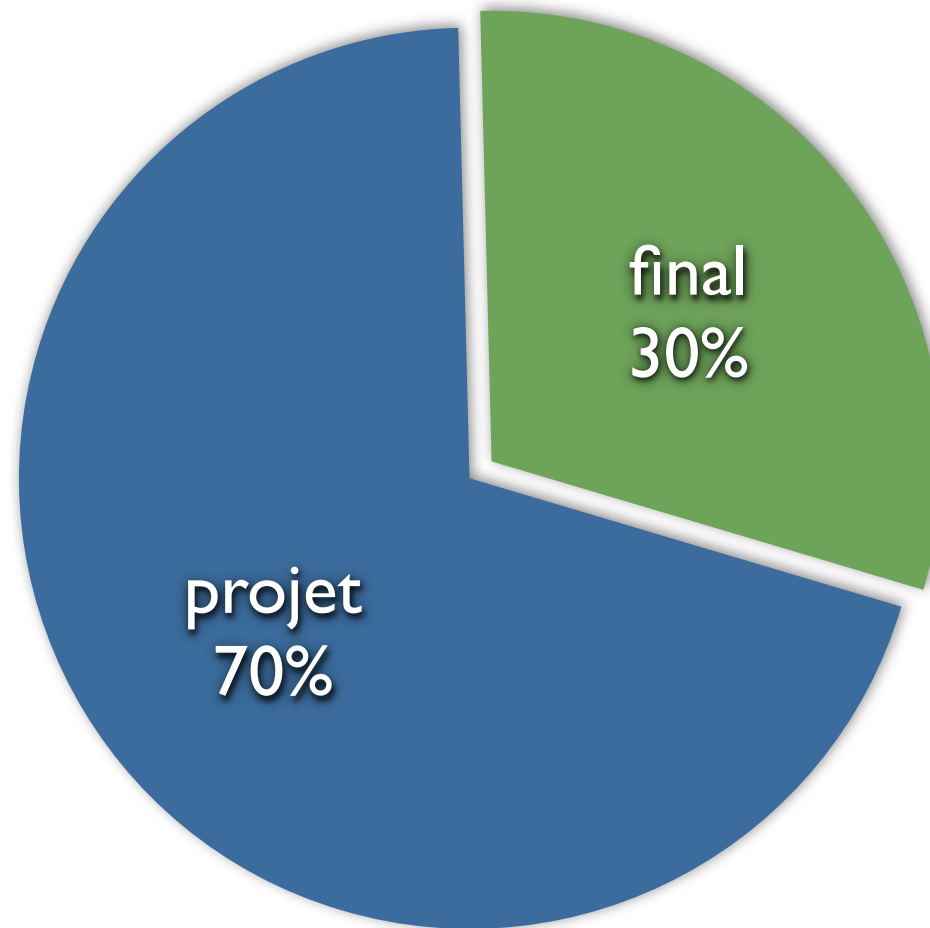
Séance l' madness et bonnes pratiques de la présentation orale

Construction des graphismes d'informations

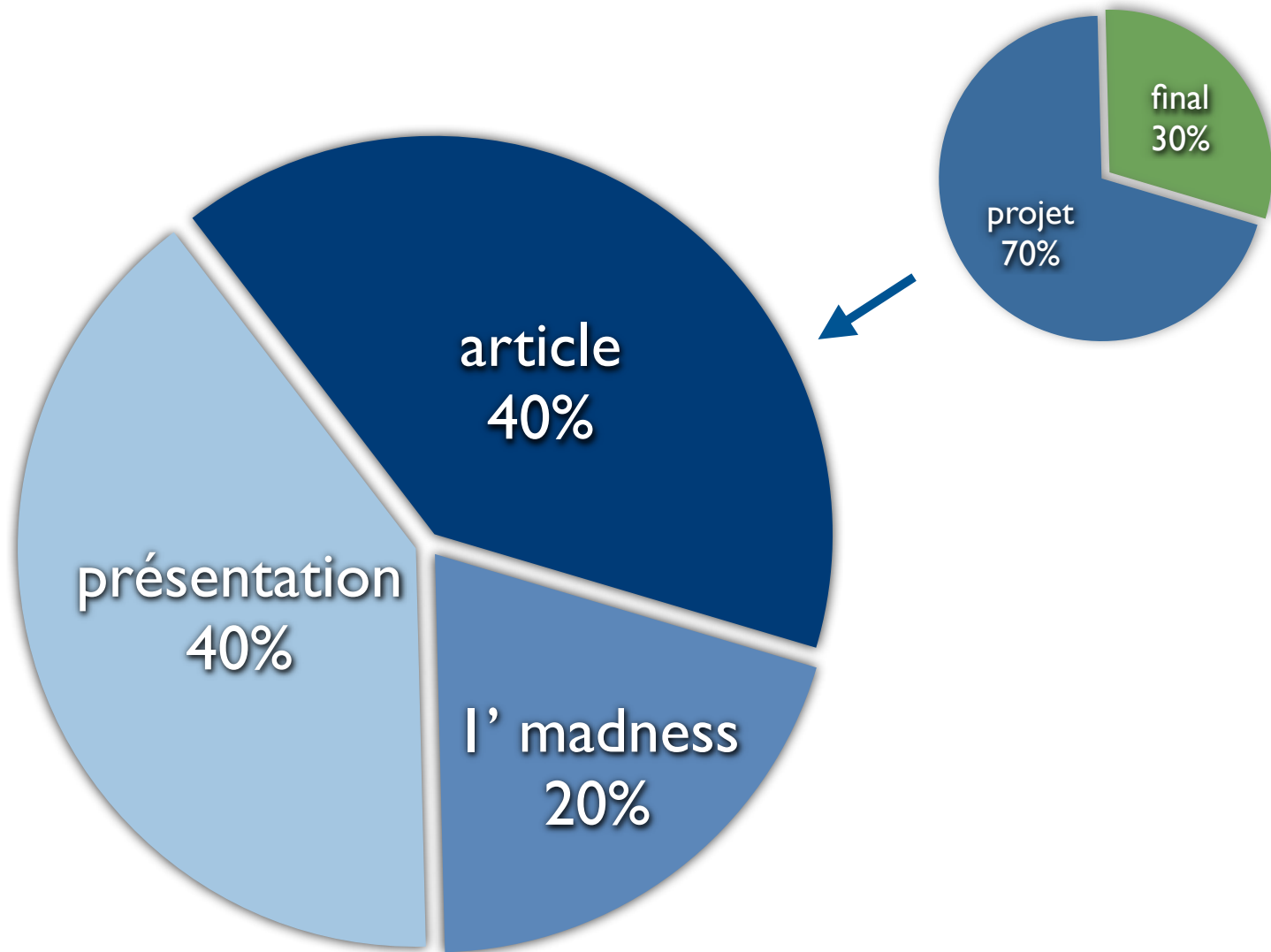
Outils en ligne pour le graphisme d'information (many eyes, viz, etc.)

Ateliers projet

# Comment obtenir IC05 ?



# Comment obtenir IC05 ?



# Organisation du projet

Avec un ou des partenaires  
avec des rôles clairement définis !

1. **Présentation** des projets le 22 février
2. **Choix définitif** le 9 mars
3. **1 minute madness** le 27 avril
4. **Exposé ~25min** sous forme d'un cours le 15 Juin
5. **Article de blog de 1500 signes** au plus tard le  
vendredi 18 juin 23h59

Où trouver plus de renseignements ?

**Sur le web !**

[www.utc.fr/~wic05/](http://www.utc.fr/~wic05/)

# Présentation des projets proposés

4 personnes :

- responsable communication
- responsable visualisations
- responsable exploration
- responsable enquête / méthodologie

## **Sujet 1 : hot map**

Observer l'organisation des communautés liées au sexe sur internet et notamment le positionnement de l'érotisme/porno, gratuit/payant, texte/video, etc. Comment le sexe s'organise-t-il sur le web ? Qui tire les ficelles ?

## **Sujet 2 : Dalai la'map**

Que pense le web du damaï lama chinois et du damaï lama tibétin ? comment sont-ils perçus dans le monde ?

## **Sujet 3 : Islamap**

Projet regroupant geovisualisation et structure du web avec pour objectif de dresser un portrait de l'islam sur le web et où se trouvent les acteurs. Sunnites, Chiïtes, Esbolah, Fatah, etc. Comment sont-ils organisés sur le web ?

## **Sujet 4 : pub map**

Le web est-il le nouvel eldorado de la publicité ? Ces sites que l'on supprime d'ordinaire parce qu'ils polluent les cartes, sont-ils eux-mêmes organisés ? Où mènent la pub et à qui profite-t-elle ?

## **Sujet 5 : Is there a dr in the house**

Comment un diagnostic différentiel peut-il être visualisé ? La visualisation peut-elle aider à la prise de décision ? Quelles méthodes sont les meilleures, quels diagnostics sont les plus chers, les plus dangereux, etc.

# Présentation des projets libres

2 à 4 personnes avec des rôles bien définis

## **Sujet développement**

Développer un plugin pour un logiciel existant (Firefox, Gephi) ou un crawler spécifique (API, Forum, etc.) utilisable par les autres. Un exemple d'exploration basique utilisant le plugin ou le crawler est **obligatoire** !

## **Sujet infoviz**

Concevoir une visualisation de données pour manifester un phénomène ou aider à l'exploration de données. Ici encore, un exemple basique d'exploration rendu possible par la visualisation est **obligatoire** !

## **Sujet constitution de corpus**

Choisir un sujet donné ou une source de données définie et l'explorer pour en fournir la structure. Exemples : télécharger les comptes rendus de séance de l'assemblée nationale et déterminer la proportion de parole gauche/droite, homme/femme, etc. ou bien établir la cartographie web du génocide des inuits au Canada, ou du point de croix.

## **Sujet Search Engine**

Construire un moteur de recherche spécifique, vertical/horizontal, au contenu spécialisé ou non, avec un corpus finalisé ou non. On pourra s'appuyer notamment sur des services en lignes (google CSE, Yahoo Boss, Yahoo Pipes) pour proposer le moteur de recherche.

Excellent semestre à tous !