

Sciences des données Sciences sociales

Paul Girard médialab

Lauren Lesnard CDSP DIME-SHS

Martial Foucault CEVIPOF

journée USPC - 30/11/2015

fondé en 2009 par Bruno Latour
méthodes numériques pour les SHS
analyse de controverses,
sociologie du numérique (usage et traces)

Sociologie,
Ingénierie Informatique,
Design de l'information et de l'interaction,
Ingénierie pédagogique

30 personnes, 15 permanents dont 3 enseignants-Chercheurs
Paul Girard médialab

Sociologie computationnelle,
Humanité numérique,
Méthode qualiquantitative,
Analyse visuelle et Exploratoire de Données,
Analyse visuelle de réseau, Web mining,
données complexes (hétérogènes et éparses),
Scientométrie, Usages du numérique,
Médias sociaux.

<http://medialab.sciences-po.fr/fr/publication/>

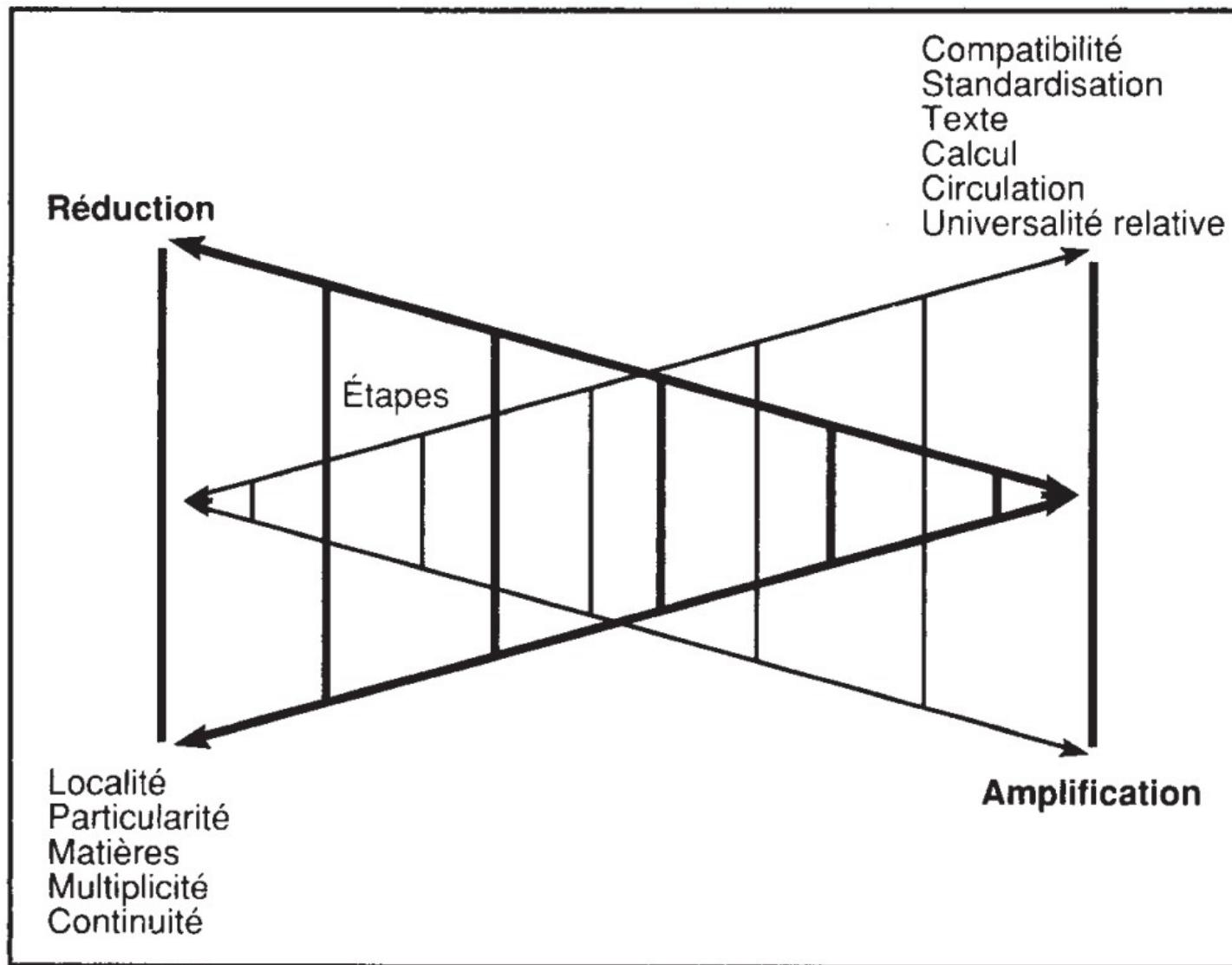


FIGURE 11.22.

(Latour 1993)

Sciences des données

Soin des données

Réversibilité de la chaîne de transformation

Visualisation et Exploration

(McCloud 1993)

richness



legibility

knowledge



information



data

Sociologie, Sciences Politiques, Histoire, Économie, Design de l'information, Ingénierie informatique, Physique, Systèmes complexes, Maths, Statistiques, Linguistique, Histoire de l'art, Cartographie, Géologie, Biologie, Philosophie, Ingénierie pédagogique...

Besoins en infrastructure (stockage, cpu) ?

OUI

FORCCAST : innovation pédagogique en cartographie de controverses

Un ou des masters Data Sciences ?
Data scientists : maths, info, stats
Data humanists : SHS, info, stats, viz

Equipex DIME-SHS
et Centre de données
socio-politiques

1. Besoins scientifiques

- Sciences sociales : rôle des enquêtes
- Enquêtes \Leftrightarrow Données
- Données des sciences sociales
 - Données quantitatives
 - enquêtes par questionnaire (sondage)
 - résultats d'élections
 - bases de données administratives
 - toute autre base de données constituées à des fins de recherche en sciences sociales
 - Données qualitatives
 - entretiens individuels ou collectifs
 - photos
 - vidéos
 - carnets ethnographique

1. Besoins scientifiques

- Accéder à des corpus de données
 - Qui doivent être constitués
 - Enquêtes par questionnaire en ligne
 - Corpus web (scraping, crawling, etc.)
 - Qui existent déjà
 - Enquêtes par questionnaire
 - Enquêtes qualitatives
- Données souvent personnelles
- Données qui peuvent être sensibles (au sens de la Cnil)

1. Besoins scientifiques

- Accéder à des corpus de données **Solutions**
 - Qui doivent être constitués
 - Enquêtes par questionnaire en ligne
 - Corpus web (scraping, crawling, etc.) **Dime-Web (médialab)**
 - Qui existent déjà
 - Enquêtes par questionnaire **CDSP (lien TGIR Progedo)**
 - Enquêtes qualitatives **Dime-quali (bequali)**
- Données souvent personnelles
- Données qui peuvent être sensibles (au sens de la Cnil)

2. Interdisciplinarité

- Méthodes des sciences sociales communes à plusieurs disciplines
- Exemples d'enquêtes Elipss réalisées
 - Pratiques culturelles (sociologie)
 - Fécondité, contraception (démographie/santé publique)
 - Valeurs, environnement, énergie (sociologie/producteur énergie)
 - Dynamiques de mobilisation politique (science politique)
 - Santé et travail (histoire)
 - Architecture et obésité (architecture)
- Enquêtes proposées
 - Débats sur le changement climatique (science politique et linguistique)
 - Culture financière et comportements patrimoniaux (économie)
 - Hiérarchisation, délimitation et identification des échelles territoriales (géographie)
 - Rapport des citoyens à la justice (science politique, droit, histoire)
 - Modélisation de la violence verbale (sciences du langage)
 - Sociologie des séries télévisées (sociologie)

3. Besoins en infrastructures

- Collecte données
 - Crawl, scraping (+ CPU, mémoire vive)
 - Questionnaires par Internet
- Stockage données
 - Redondance
 - Liens avec le Cines indispensables (mutualisations !)
- Mise à disposition données (sécurité)

4. Liens avec la formation

- Forts pour Dime-Web (cf. Forccast)
- Forts mais locaux pour Dime-quali
- À développer pour Dime-quant (une expérience jusqu'à présent)

5. Laboratoires autres structures concernés

- Consortium Dime-SHS : Sciences Po (porteur), U Paris Descartes, Ined, TGIR Progedo, GENES, EHESS, Telecom ParisTech
- Potentiellement tout labo de science sociales
- Et au delà

6. Participation programmes nationaux

- TGIR Progedo
- TGIR Huma-Num

7. Liens industriels, européens

- Contrat/partenariat avec opérateur téléphonie (Dime-quanti/Elipss)
- Montage projet européen (Dime-quanti/Elipss)

CEVIPOF :

Des projets et des défis

Panel Electoral 2017,
Politique de la Terre
et autres

Big Data : révolution méthodologique

- **Science politique : 2 approches conventionnelles**

- Approche qualitative (entretiens)
- Approche quantitative (enquêtes, base de données longitudinales ou cross-section)

- **Science politique : 2 nouvelles approches**

- Approche expérimentales (terrain ou laboratoire)
- Approche Big Data (exploration de nouveaux phénomènes, problème de l'identification de la causalité)

Big Data : CEVIPOF

- **Panel Électoral 2017** : enquêtes en ligne durant 20 mois (nov. 2015 ⇒ juin 2017) auprès de 10 000 mêmes répondants (à partir de 27 000 personnes interrogées) :
 - 8 millions de réponses individuelles
 - 40 millions de données électorales et sociodémographiques collectées

Big Data : CEVIPOF

- **Panel Électoral 2017 :**

- Collaboration avec *Stanford University* pour collecter durant la campagne 2017 l'ensemble des communications (sites web, blogs,...) des candidats et partis politiques.
- Département de linguistique de *Stanford University* : programme de recherche en intelligence artificielle et bibliothèque pour archivage et stockage
- Préparation d'une collaboration CEVIPOF-MédiaLab sur traitement de telles données

Big Data : CEVIPOF

• Projet Politique de la Terre (USPC)

- Axe 3 : Dynamique des zones critiques et conflits d'urbanisation .

Quatre thème de recherches :

- Pollinisation et urbanisation (Isabelle Dajoz)
- Mobilisation sociales environnementales: le cas de l'agriculture à Plaine Commune (Nathalie Blanc)
- Les réseaux sociaux : exemple d'Ermont (Nicolas Douay)
- Les inégalités territoriales de santé et environnement, Participation électorale et expression du vote (Jean Chiche et Daniel Boy)

Big Data : CEVIPOF

•Projet sondothèque

- Convaincre les instituts de sondage de mettre à disposition les données individuelles anonymes d'enquête d'opinion publique dans un système d'interrogation unique en ligne
- Développer des outils de visualisation de millions d'observations

Besoins

- Recrutement de data scientists, développeurs
- Faciliter des lieux d'échanges et travail sur les protocoles de collecte et d'exploitation de données
- Utiliser les BIG DATA pour améliorer nos protocoles d'enquête et développer des expérimentations en ligne (cf. projet ERC Yann Algan à Sciences Po)

Formation

- Disponibilité des chercheurs et ingénieurs de recherche CEVIPOF pour des masters communs “Data Science”
- Quelques noms et compétences
 - Chercheurs : Bruno Cautrès, Jean Chiche, Martial Foucault, Brigitte Le Roux
 - Ingénieurs : Diégo Antolinos-Basso, Flora Chanvril,
- Une approche méthodologique : les Ateliers d’analyse géométrique des données
- Un enjeu : la formation des formateurs ?