

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

# Introduction

## Atelier *Humanités numériques*

Jean-Baptiste Camps

École nationale des chartes | Paris, Sciences & Lettres

Casa de Velázquez  
Madrid, 9 octobre 2018

# Plan

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation  
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

1

## Définition ?

- Survol historique
- Des humanités numériques au computational et à la science des données

2

## Collecte et structuration des données

- Interroger et collecter des données depuis des entrepôts (API)
- Construire ses données
- Structuration des données

3

## Interrogation et analyse des données

- Outils d'interrogation
- Méthodes quantitatives : des corrélations à la modélisation
- Quelques outils particuliers

# Faut-il vraiment définir les humanités numériques ?

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation  
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

- interdiscipline ?
- transdiscipline ?
- science auxiliaire ?
- approche  
méthodologique ?
- courant  
historiographique ?
- Mode ?
- source de financement ?
- “cheval de Troie  
néolibéral” ?

Dans les humanités  
numériques francophones,  
beaucoup de temps a été  
pris, dans les années 2010,  
par la question de la  
**définition** des humanités  
numériques...

# Une définition et une distinction

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

## Définition de travail

Mise en œuvre de méthodes numériques en sciences humaines.

## Ne pas confondre

humanités numériques numérique **en** sciences humaines ;

“études numériques” numérique comme **objet d'étude** des  
sciences humaines (y compris de manière très  
traditionnelle) ; *digital studies*.

# Quelques dates : chronologie sélective et personnelle

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

- 1941-9 Robert Busa conçoit le projet d'*Index Thomisticus* appuyé sur des machines. En 1949, il convainc Thomas J. Watson, fondateur d'IBM ;
- 1946 A.D. Booth entreprend la conception d'une "traductrice automatique" ;

# Années 1940-1960

Les jésuites, les bénédictins et les autres

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

Un premier projet d'ampleur : l'indexation et la lemmatisation de l'œuvre de saint Thomas d'Aquin.



*Livia Canestrano au travail sur l'Index Thomisticus (gauche), l'atelier de Gallarate (droite) ; années 1950-1960 ; arch. R. Busa*

CC-BY-NC, <http://melissaterras.blogspot.com/2013/10/for-ada-lovelace-day-father-busas.html>

# Années 1940-1960

## Les jésuites, les bénédictins et les autres

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

Dom J. Froger, La Critique des textes et son automatisation,  
1968.

Depuis vingt ans, exactement depuis qu'en 1946 A.D. Booth entreprit de construire une traductrice automatique, les philologues utilisent de plus en plus fréquemment les ordinateurs : au mois de mai de cette année une revue spécialisée, *Computers and the Humanities*, énumère cent vingt programmes en cours d'exécution, et l'énumération n'est pas exhaustive.

La plupart de ces travaux sont, en dernière analyse, des *index verborum* (...) on peut considérer l'ordinateur comme l'instrument par excellence de tous les travaux qui relèvent plus ou moins directement de la lexicographie. (...)

Il est maintenant courant d'« automatiser » les recherches philologiques ou littéraires, et d'utiliser la machine électronique ou mécanographique pour des études lexicographiques ou stylistiques ; ces procédés ont déjà été employés pour aider à la critique conjecturale, et par exemple combler les lacunes des manuscrits de la mer Morte

En 1960-61, sous la direction de Mme Poyen, deux « programmes » destinés à l'ordinateur Gamma ET Bull ont été établis : l'un (...) pour la collation, l'autre (...) pour la recherche de l'enchaînement des manuscrits d'après leurs variantes.

# Quelques dates : chronologie sélective et personnelle

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation  
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

- 1941-9 Robert Busa conçoit le projet d'*Index Thomisticus* appuyé sur des machines. En 1949, il convainc Thomas J. Watson, fondateur d'IBM ;
- 1946 A.D. Booth entreprend la conception d'une "traductrice automatique" ;
- 1966 création de la revue *Computers and the Humanities* ;
- 1967 Charles Muller soutient à Strasbourg sa thèse *Étude de statistique lexicale : le vocabulaire du théâtre de Pierre Corneille* ;
- 1968 Dom Froger publie *La Critique des textes et son automatisation* ;

# Années 1960-1970

Histoire, linguistique, philologie, lexicographie... *quantitatives*

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

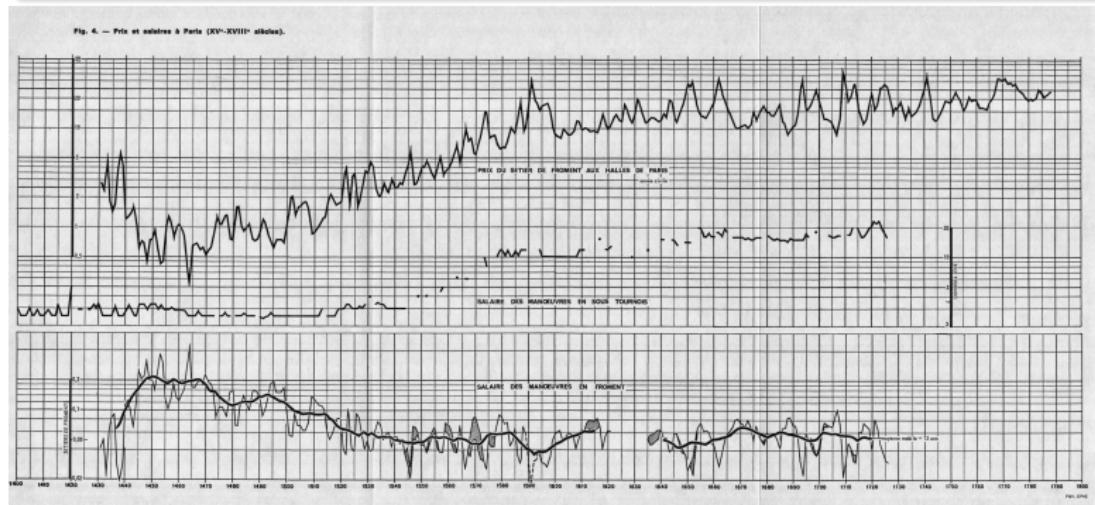
Interrogation  
et analyse des  
données

Outils d'interrogation  
  
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

Emmanuel Le Roy Ladurie (1968)

L'historien de demain sera programmeur ou il ne sera pas.



Micheline Baulant, « Le salaire des ouvriers du bâtiment à Paris, de 1400 à 1726 », *Annales* (1971).

# Quelques dates : chronologie sélective et personnelle

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

1980-... le terme d'*Humanities computing* se répand, avec  
la création de centres dans les univ.  
anglo-saxonnes ;

1982 mise sur le marché du Commodore 64 ;

1987 établissement de la *Text Encoding Initiative* ;

1996 spécification XML ;

# Années 1980-1990

Arrivée du micro-ordinateur | Recul du quantitatif

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers



Il est de bon ton, pour certains, de jouer aux princes de l'intelligence en dédaignant superbement, comme des contingences subalternes, des mesquineries de tâcheron, les exigences de la rigueur et les contraintes de la quantification  
*(Antoine Prost, Douze leçons sur l'histoire, 1996)*

# Années 1980-1990

## Structuration des données et corpus textuels

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

## Les Principes de Poughkeepsie (1987)

Proposer des *Guidelines* pour :

- 1 provide a standard format for data interchange in humanities research.
- 2 suggest principles for the encoding of texts in the same format.
- 3 define a recommended syntax for the format, a metalanguage for the description of text-encoding schemes, describe the new format and representative existing schemes both in that metalanguage and in prose ;
- 4 propose sets of coding conventions suited for various applications.
- 5 include a minimal set of conventions for encoding new texts in the format.

# Quelques dates : chronologie sélective et personnelle

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

1980-... le terme d'*Humanities computing* se répand, avec la création de centres dans les univ. anglo-saxonnes ;

1982 mise sur le marché du Commodore 64 ;

1987 établissement de la *Text Encoding Initiative* ;

1996 spécification XML ;

1989-1993 Tim Berners-Lee propose et mène la création du Web (au CERN) ;

2004 Parution d'*A Companion to Digital Humanities* ;

2007 Jim Gray propose le concept de *data science*.

# Années 2000-2010

De l'*Humanities Computing* aux *Digital Humanities*

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

- Terme de *Digital Humanities* créé par Johanna Drucker, John Unsworth et Jerome McGann, à la fin des années 1990
- remplacer le terme d'« *Humanities Computing* », « too closely associated with computing support services »
- Ce changement devait signifier que ce champ « had emerged from the low- prestige status of a support service into a genuinely intellectual endeavour with its own professional practices, rigorous standards, and exciting theoretical explorations.

# Et maintenant ?

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

## Humanités digitales (?)



Humanités  
computationnelles ou  
tournées vers les  
données

# Des humanités numériques, computationnelles ou intensives en données ?

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computationnel et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

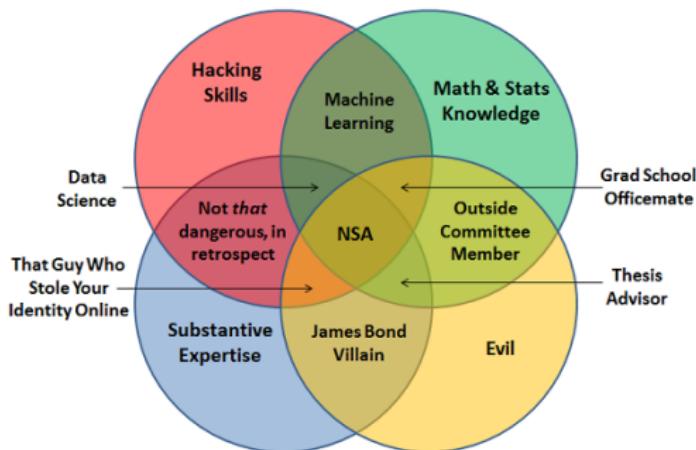
Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers



4 paradigmes de la  
recherche scientifique  
selon Jim Gray

- **expérimentale**
- **théorique**
- **computationnelle**
- **intensive en  
données**

# Plan

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

## 1 Définition ?

- Survol historique
- Des humanités numériques au computationnel et à la science des données

## 2 Collecte et structuration des données

- Interroger et collecter des données depuis des entrepôts (API)
- Construire ses données
- Structuration des données

## 3 Interrogation et analyse des données

- Outils d'interrogation
- Méthodes quantitatives : des corrélations à la modélisation
- Quelques outils particuliers

# Sources de données

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation  
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

## Sources institutionnelles

- bibliothèques numériques : Gallica, E-Codices, ...
- portails d'institutions patrimoniales : data.bnf, ...
- entrepôts de données de la recherche : Zenodo, ...
- portails universitaires ou de laboratoires : University of Oxford Text Archive, ressources de l'IRHT...
- projets divers et variés : sites des projets, s'ils sont encore maintenus ; *Wayback Machine* sinon...

## Autres sources

- données produites par une communauté (e.g. Wikidata...)
- corpus et bases de données payantes (Brepols, Garnier...) ;
- entrepôts de code source : Github, ...

# Questions juridiques

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

## Droits patrimoniaux et domaine public

Droit d'auteur : naît d'un acte original de création.

Cas général : entrée dans le domaine public

70 ans après la mort de l'auteur.

## Copyfraud

(Ré)éditer une œuvre du domaine public *ne crée pas de nouveau droit.*

Se méfier de certaines indications sur les sites commerciaux,  
nulles et non avenues.

# Questions juridiques

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation  
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

## Licence

- libre ?
- licence similaire aux œuvres dérivées ?
- réutilisation commerciale ?
- œuvres dérivées ?

## Exception du droit à la fouille de données

loi du 7 octobre 2016 « Pour une République numérique »

Exception au droit d'auteur :

*Les copies ou reproductions numériques réalisées à partir d'une source licite, en vue de l'exploration de textes et de données incluses ou associées aux écrits scientifiques pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale*

# Comment récupérer les données ?

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

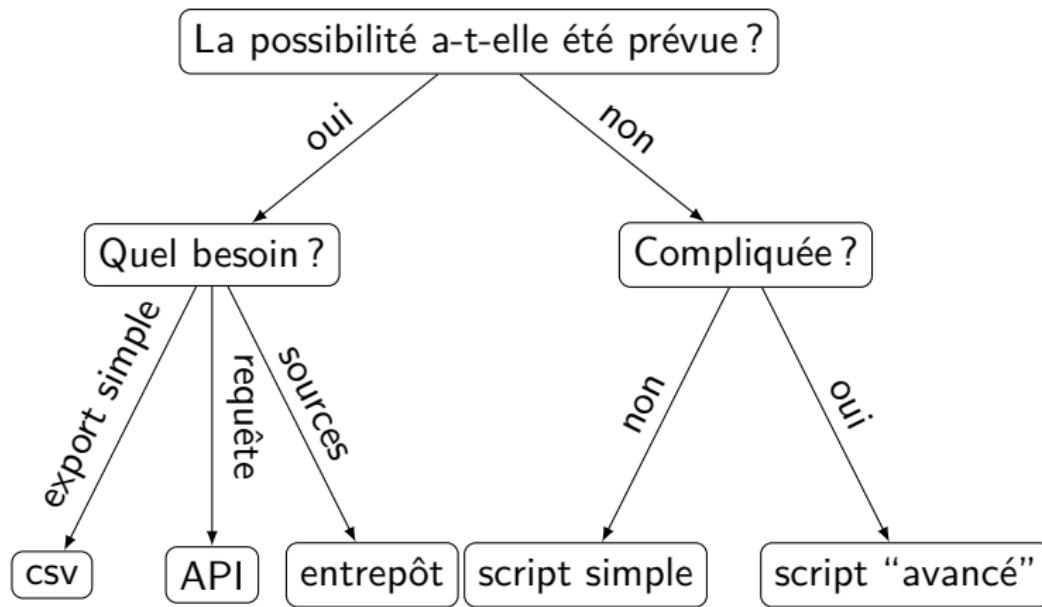
Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers



# Les APIs

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des人文數字  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers



Le portail BnF API et jeux de données décrit et documente l'ensemble des API qui permettent d'interroger et de visualiser les métadonnées des catalogues notamment BnF-Catalogue général, data.bnfr.fr, Gallica, GCFP et les catalogues numériques de la BnF. Les données de la BnF sont également disponibles sous forme de jeux de données (images et textes, métadonnées, statistiques) qui sont directement téléchargeables via le portail.

Plusieurs formats, plusieurs technologies permettent de répondre à la diversité des usages des données de la bibliothèque : alimentation de catalogues, création de nouveaux services innovants, fouille de données, datavisualisation, etc.

Développeurs et métadonneurs, chercheurs et chercheuses, acteurs et actrices du monde de la culture et de la culture du livre, digital humanists, ou encore amateurs et amatrices de culture, les données et métadonnées diffusées par la BnF n'attendent plus que vous !

Accès au BnF API et jeux de données

## Actualités

Mise à disposition de la Très Grande  
Bibliothèque du Chêne Rabelais

Gallica Studio : innovation et créativité  
autour des richesses de Gallica

Mise à disposition des produits  
bibliographiques 2017

Le site [api.bnfr.fr](http://api.bnfr.fr)

## Application Programming Interface

Un ensemble de méthodes et d'outils prédéfinis, qui peuvent servir de base à la construction de requêtes ou de programmes.

Un peu l'équivalent pour programmeur de l'interface graphique pour l'utilisateur.

# Un exemple d'API : IIIF

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

<https://gallica.bnf.fr/iiif/ark:/12148/btv1b9045910f/f0/full/600/scheme://server/prefix/identifier/region/size/rotation/quality.fc>



# Un autre exemple : Canonical Text Services

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

<http://dev.chartes.psl.eu/elec/geste/api/cts?request=GetPa>

```
-<GetPassage>
  -<request>
    <requestName>GetPassage</requestName>
    <requestUrn>urn:cts:froLit:geste.jns11095.transcr_Otin_B:1-100</requestUrn>
  </request>
  -<reply>
    <urn>urn:cts:froLit:geste.jns11095.transcr_Otin_B:1-100</urn>
    -<passage>
      -<TEI xml:id="transcr_Otin_B" type="transcr" corresp="jns11095" xml:lang="fr" version="5" py:pytype="TREE">
        -<text xml:lang="xno" xml:base="urn:cts:froLit:geste.jns11095.transcr_Otin_B">
          -<body>
            -<lg n="1" type="laisse">
              -<l n="1" xml:id="B_1_1">
                -<w lemma="qui" type="POS=PROrel|NOMB.=s|GENRE=m|CAS=n" xml:id="B_w_000001">
                  -<hi rend="initiale filigr 4!">
                    -<choice>
                      <reg>k</reg>
                      <orig>K</orig>
                    </choice>
                  </hi>
                  -<hi rend="initiale">
                    -<choice>
                      <reg>i</reg>
                      <orig>J</orig>
                    </choice>
                  </hi>
                </w>
              -<w lemma="voiloir" tvbe="POS=VERcial|MODE=ind|TEMPS=dst|PERS.=3|NOMB.=s" xml:id="B_w_000002">
```

# Acquisition des données

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation  
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

Si les données n'existent pas sous forme numérique, il faut les construire soi-même.

## Acquisition de données textuelles

- imprimées : OCR (Tesseract, ...);
- manuscrites : transcription, HTR (Transkribus, OCROpy, ...);

## Nettoyage / Post-correction

Des logiciels de post-correction existent,  
fondés sur des dictionnaires, des règles, des corrections en série.

- Antidote (payant) pour les langues contemporaines ;
- PoCoTo (libre, dév. par le CISOCR group, Munich) : permet de créer des modèles pour les anciens états de langue ; <https://github.com/cisocrgroup/PoCoTo>.

# Nettoyer ou fusionner des jeux de données

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

Des outils permettent de faciliter le nettoyage ou la fusion de jeux de données originellement mal structurés ou distincts, ex.  
Dataiku, OpenRefine,...

The screenshot shows the Dataiku Data Science Server (DSS) interface. The top navigation bar includes 'BeDT...', 'Summary', 'Explore', 'Charts', 'Status', 'History', 'Settings', 'PARENTRECIPE', 'Labs', and 'ACTIONS'. Below the navigation is a search bar and a 'DISPLAY' dropdown. The main area is titled 'Viewing dataset sample' with 'Configure sample' and '100000 rows, 7 cols' information. A preview table shows columns: SIGLA, vnum, reportori\_n, rubrica\_3, incipit\_ms, initio, fine. The first few rows of data are visible.

SIGLA	vnum	reportori_n	rubrica_3	incipit_ms	initio	fine
p_402_0001		BeDT 375.B.A	Pens decapell.	Pens decapell si fo un gentil barri del puel sancta...	009 r	009 r
p_402_0002		BeDT 375.010	Pens decapell.	Harréis è France i fai sepiant uns sos	009 r	009 v
p_402_0003		BeDT 375.001	P. de cap.	Ales mres pres com sular que sircan	009 v	009 v
p_402_0004		BeDT 375.B.B	-	Pens decapell si arret com avies auxdit denan. m...	015 v	006 r
p_402_0005		BeDT 375.020	-	Alosi com sul ca pron de valadores	016 r	006 v
p_402_0006		BeDT 375.018	P. decap.	Qui per resoi cuidar	016 v	017 v
p_402_0007		BeDT 375.011	P. decap.	la non er hem tan pros	017 v	008 v
p_402_0008		BeDT 375.014	Pens de capd.	Léolis amris cui amor ten iolos	018 v	009 v
p_402_0009		BeDT 375.002	P. de cap.	Si ai perdut mon saber	019 r	020 r
p_402_0010		BeDT 375.016	P. decap.	Medis com non pot de vi pensar	020 r	020 v
p_402_0011		BeDT 375.007	P. decap.	De totz chantius sei ieu atsel que plus	020 v	009 r

# Enrichissement des données

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

Quelques exemples d'enrichissements pouvant recevoir le soutien de méthodes computationnelles...

## Lemmatisation

Des lemmatiseurs fondés sur

- des lexiques de forme, ex. TreeTagger ;
- des modèles d'apprentissage machine, ex. Lemming, Pandora,...

Des outils de post-correction, ex. Pandora Post-Correction Editor (<http://dev.chartes.psl.eu/ppa/>).

## Reconnaissance des entités nommées

- SEM, <http://apps.lattice.cnrs.fr/sem/> ;
- Recogito, <https://recogito.pelagios.org> (ex.).

# FAIR Data

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

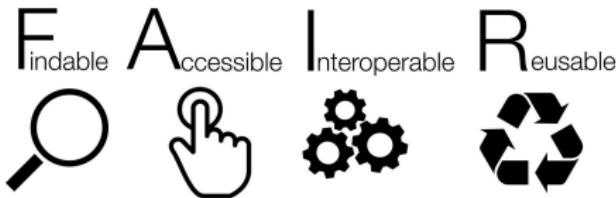
Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation  
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers



(Img. SangyaPundir CC BY-SA 4.0)

- *Findable* : données doivent être faciles à trouver en ligne, y compris sur le long terme (archivage pérenne) ;
- *Accessible* : format ouvert, libre, documenté et compréhensible ;
- *Interoperable* : pas enfermées dans un logiciel propriétaire ; standards, intégrité des données, documentation ;
- *Reusable* : données ouvertes, réutilisables (légalement, techniquement,...).

# Les recommandations de la TEI



```
<TEI>
  <teiHeader>
    <fileDesc/>
  </teiHeader>
  <text>
    <body>
      <p>Salve !</p>
    </body>
  </text>
</TEI>
```

- un cadre conceptuel relatif à la représentation sémantique des textes ;
- conçu par une communauté de chercheurs ;
- ouvert, libre et documenté (Guidelines) ;
- implémenté en XML.

# Lier ses données à des référentiels

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

- référentiels de lieux, personnes, œuvres, titres ...
- jeux d'étiquettes linguistiques de référence ;
- ...

# Plan

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

1

## Définition ?

- Survol historique
- Des humanités numériques au computationnel et à la science des données

2

## Collecte et structuration des données

- Interroger et collecter des données depuis des entrepôts (API)
- Construire ses données
- Structuration des données

3

## Interrogation et analyse des données

- Outils d'interrogation
- Méthodes quantitatives : des corrélations à la modélisation
- Quelques outils particuliers

# Quelques outils d'interrogation de données

## Expressions régulières

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

Comment aller plus loin que le CTRL+F ou le moteur de recherche...



Une expression régulière est un motif qui correspond ou non à une chaîne de caractère donnée.

Ex.

- **dates ?** date avec un ou sans s ;
- **dat.\*** dat suivi de n'importe quoi ou de rien ;
- **\d\d\w\w** 2 chiffres et 2 lettres.

# Quelques outils d'interrogation de données

## Langages de requêtes

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation  
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

Des langages de requête spécifiques permettent d'interroger des données stockées selon un format spécifique.

### SQL

Interrogation de données stockées dans des bases relationnelles.

```
SELECT * FROM table  
WHERE nom = 'JDupont'
```

### XPath et XQuery

Interrogation de données stockées en XML

```
/TEI/text/body//persName[@ref  
→ = 'JDupont01']
```

### SPARQL

Interrogation de données stockées en RDF.

### CQL

Langage d'interrogation de corpus linguistique (*Corpus Query Language*).

```
[word='de.*'][]? [frpos='NP']
```

# Résumer l'information avec des statistiques descriptives

## Introduction

Jean-Baptiste  
Camps

## Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

## Collecte et structuration des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

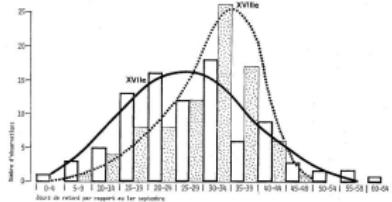
Structuration des  
données

## Interrogation et analyse des données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

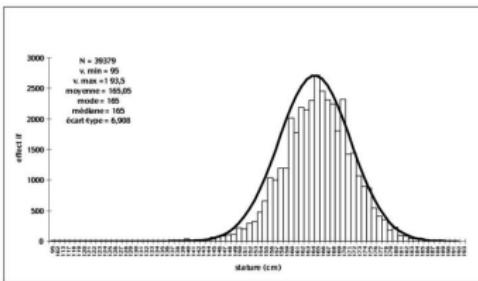
Quelques outils  
particuliers



GRAPH. VI. — Histogramme de répartition des jours de retard des vendanges au XVII<sup>e</sup> siècle et au XVIII<sup>e</sup> siècle.

XVII<sup>e</sup> siècle : nombre d'observations = 90  
moyenne = 28,15 ; écart-type =  $\pm 11,63$ .  
XVIII<sup>e</sup> siècle : nombre d'observations = 83  
moyenne = 29,57 ; écart-type =  $\pm 8,29$ .

La différence entre les moyennes ( $1,42$ ) n'est pas significative à  $p = 0,05$ .



Comment aborder le vaste ensemble de données que l'on a constitué ?  
Une approche **statistique** permet de :

- résumer numériquement ou graphiquement une vaste quantité d'information ;
- avoir des intuitions sur ses données ;
- tester des hypothèses.

# Déetecter des corrélations

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

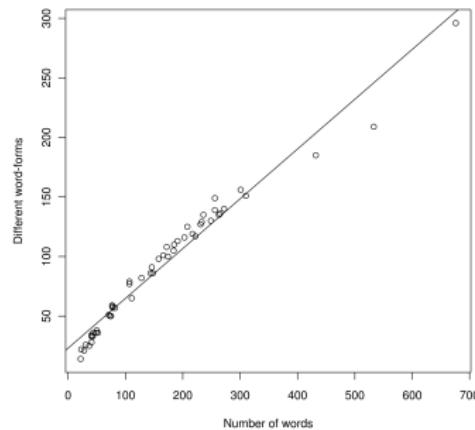
Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers



Identifier quand deux  
variables **fonctionnent  
ensemble**,  
et la **significativité** par  
rapport au hasard.

# Écueils

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

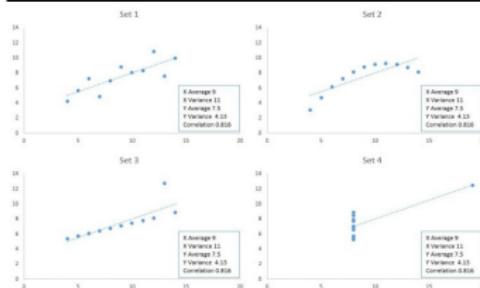
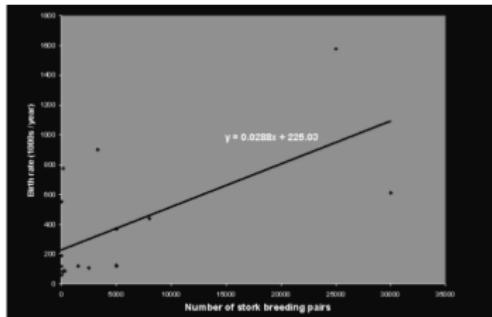
Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers



Cum hoc ergo propter hoc ?

Corrélation n'est pas  
causalité.

Qu'y a-t-il derrière une  
corrélation ?

Des “formes” ou  
comportements très  
différents peuvent être  
caractérisés par le même  
niveau de corrélation.

# Méthodes d'analyse et de visualisation

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

Une bonne partie des méthodes statistiques employées en HN trouvent leur origine dans d'autres champs : économétrie, biostatistiques,...

Elles ont souvent une dimension très géométrique : calculs de distance, nuages de points, inertie...

# Régression

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des人文數學  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

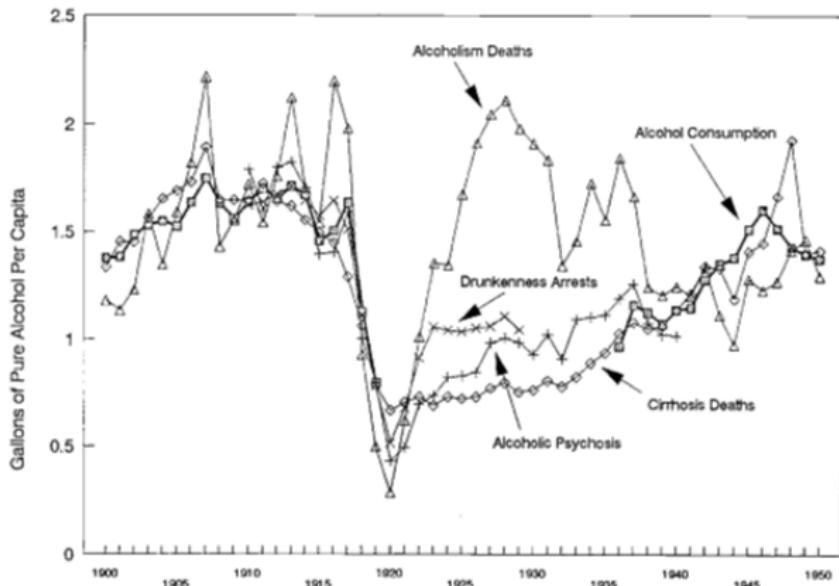
Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

Figure 1: Estimated Alcohol Consumption



J.A. Miron et J. Zwiebel, *Alcohol consumption during prohibition*,  
1991.

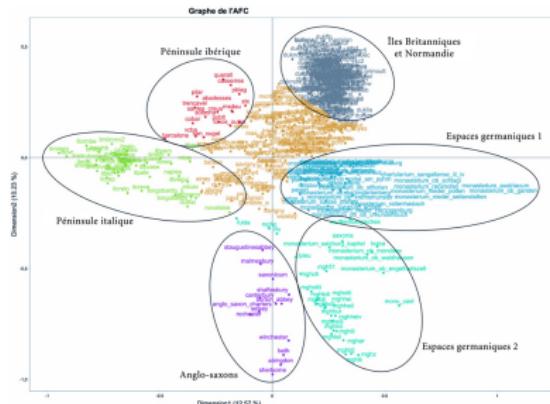
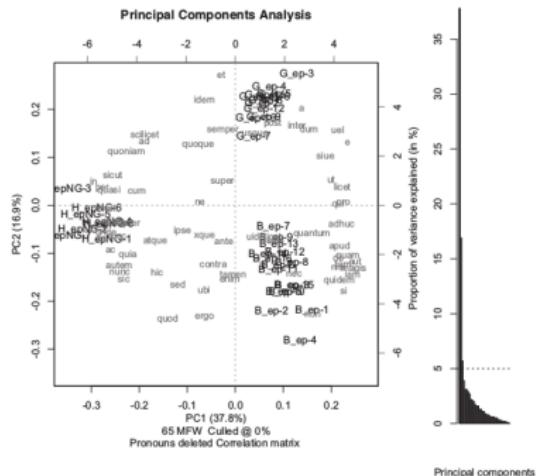
## Analyse par réduction de la dimensionnalité

## Introduction

### Définition ?

## Outils d'interrogation

### Méthodes quantitatives : des corrélations à la modélisation



N. Perreux, *L'Écriture du Monde...*, 2016.

M. Kestemont *et al.*, Collaborative authorship..., 2015.

# Partitionnement de données

## Introduction

Jean-Baptiste  
Camps

## Définition ?

Survol historique

Des人文學  
numériques au  
computational et à  
la science des  
données

## Collecte et structuration des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

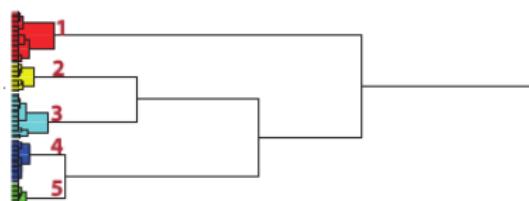
## Interrogation et analyse des données

Outils d'interrogation

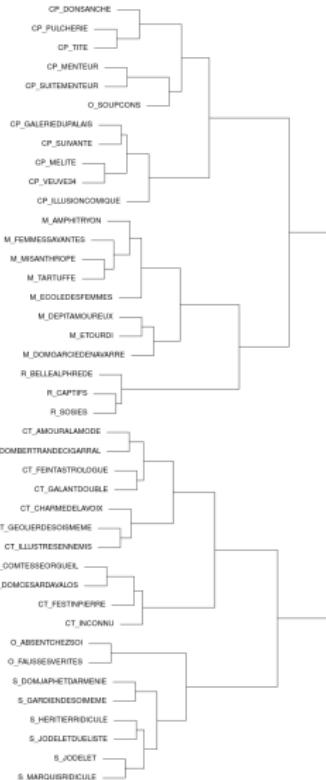
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

### DEES 1983



H. Goebel, *L'aménagement scripturaire...*, 2011.



# Quelques limites

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

## Spécificités des données des SHS

- données fragmentaires, manquantes ;
- complexité ;
- dimensionnalité ;
- non linéarité (comme en physique).

## Des corrélations à la modélisation

Rendre compte des agents et de leurs comportements, de leurs interactions.

- systèmes dynamiques, multi-agents ;
- transitions de phase ?

Des interactions simples peuvent mener à beaucoup de complexité.

# Ex. modèle de Schelling et ségrégation urbaine

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

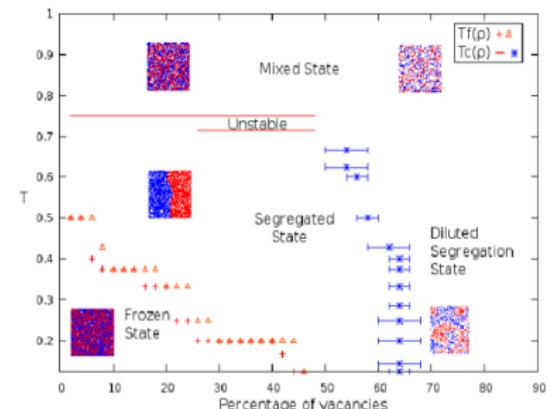
Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

#	O	O	##	#	#	#	O	O	O	#
#	#	O	O	#	#	O	O	O	O	#
#	#	O	#	#	O	O	#	#	#	#
#	#	O	#	O	O	O	#			
#	#	O	#	#	#	O	O	#	#	#
#	O	#	#	#	#	#	#	#	#	#
#	#	O	#	#	O	O	O	#	#	#
#	#	O	#	#	#	O	#	#	#	#
#	#	O	#	#	#	O	#	#	#	#
#	#	O	#	#	#	O	#	#	#	#
O	#	#					O	#	O	O

T.C. Schelling, *Jour. Math. Soc.* 1  
(1971)



Gauvin et al., *European Physical Journal B*, 70, 293–304 (2009)

# Ex. dynamique de population des manuscrits

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique

Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation  
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers



- “birth”/copy rate increases from bottom to top
- “death”/decimation rate increases from left to right
- increments = 0.1, from 0 to 1
- average fraction of bifurcations (over  $10^2$  realisations, 500 time steps each)

# Ex. analyse de réseaux

## Introduction

Jean-Baptiste  
Camps

## Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

## Collecte et structuration des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

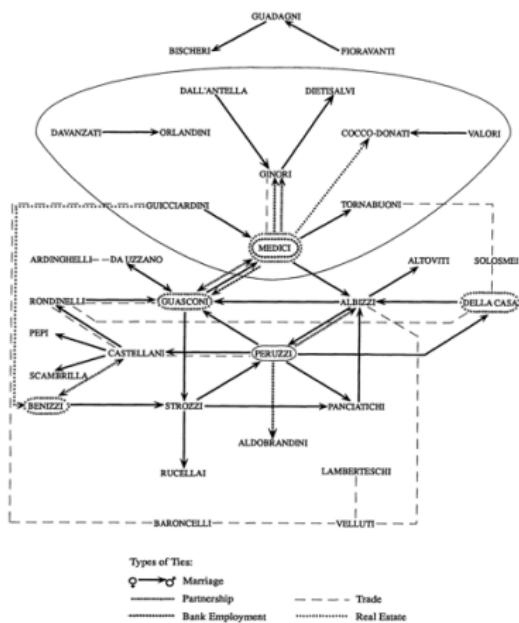
Construire ses  
données

Structuration des  
données

## Interrogation et analyse des données

Outils d'interrogation  
Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers



## Fondements théoriques

- mathématiques : théorie des graphes,
- sociaux : structure sociale vue comme un ensemble de relations entre individus.

## Variété des applications

- réseaux sociaux, clientèle, ...
- réseaux de citations, emprunts, liens ;
- réseaux de correspondance ;

# En guise de conclusion...

Introduction

Jean-Baptiste  
Camps

Définition ?

Survol historique  
Des humanités  
numériques au  
computational et à  
la science des  
données

Collecte et  
structuration  
des données

Interroger et collecter  
des données depuis  
des entrepôts (API)

Construire ses  
données

Structuration des  
données

Interrogation  
et analyse des  
données

Outils d'interrogation

Méthodes  
quantitatives : des  
corrélations à la  
modélisation

Quelques outils  
particuliers

## Pourquoi se lancer dans les humanités numériques ?

- ① Parce que vous êtes le plus à même de le faire : vous connaissez votre sujet, vos données,...
- ② Parce qu'elles peuvent vous apporter un regard différent et de nouveaux résultats sur vos questions de recherche.

## Un conseil

Apprendre à manier des logiciels spécialisés est nécessaire, mais, le meilleur investissement sur le long terme,

- ① se familiariser avec un langage de programmation ;
- ② acquérir des notions de stats et de mathématiques.