

Buzz Words ...



Détection, Catégorisation, Visualisation; Les Données Massives (Big DATA) dans un contexte eDiscovery

Gwendoline Huret, EMEA expert eDiscovery

Septembre 2015

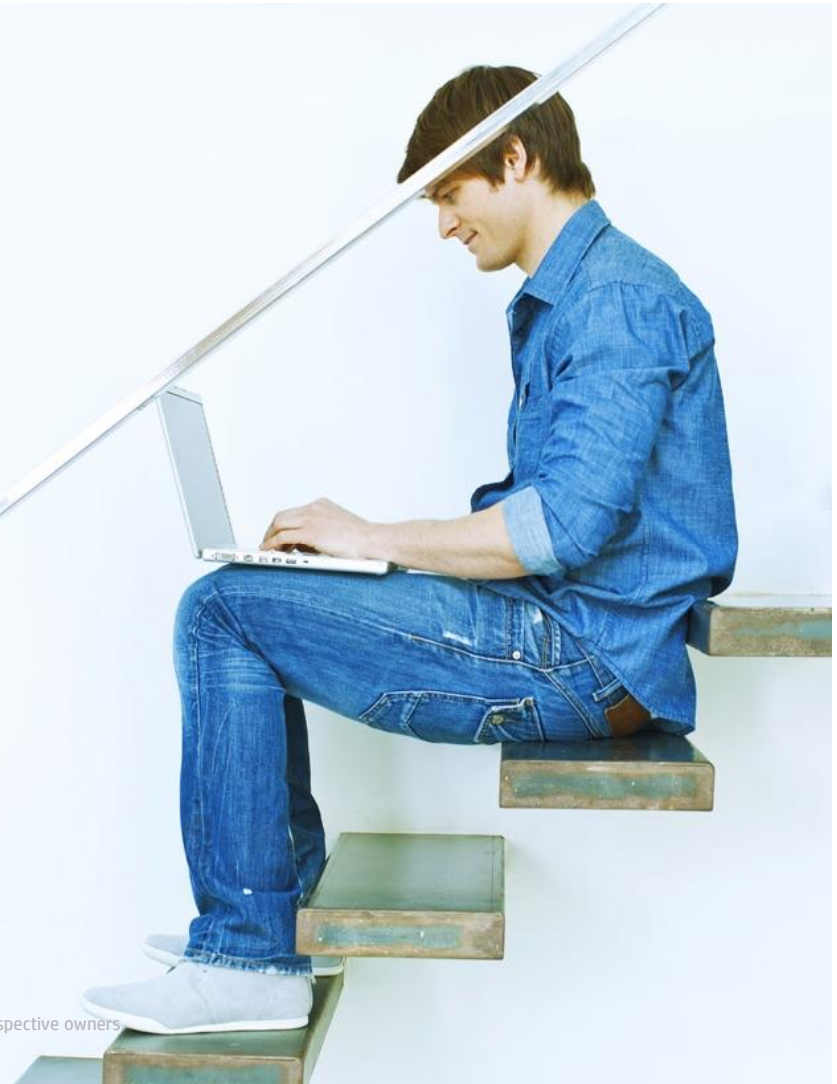
Problématique BIG DATA

- **Traitement Manuel**
- **Traitement Informatique Classique**
 - Par Mots Clés
 - Par Recherche filtrée



Objectifs

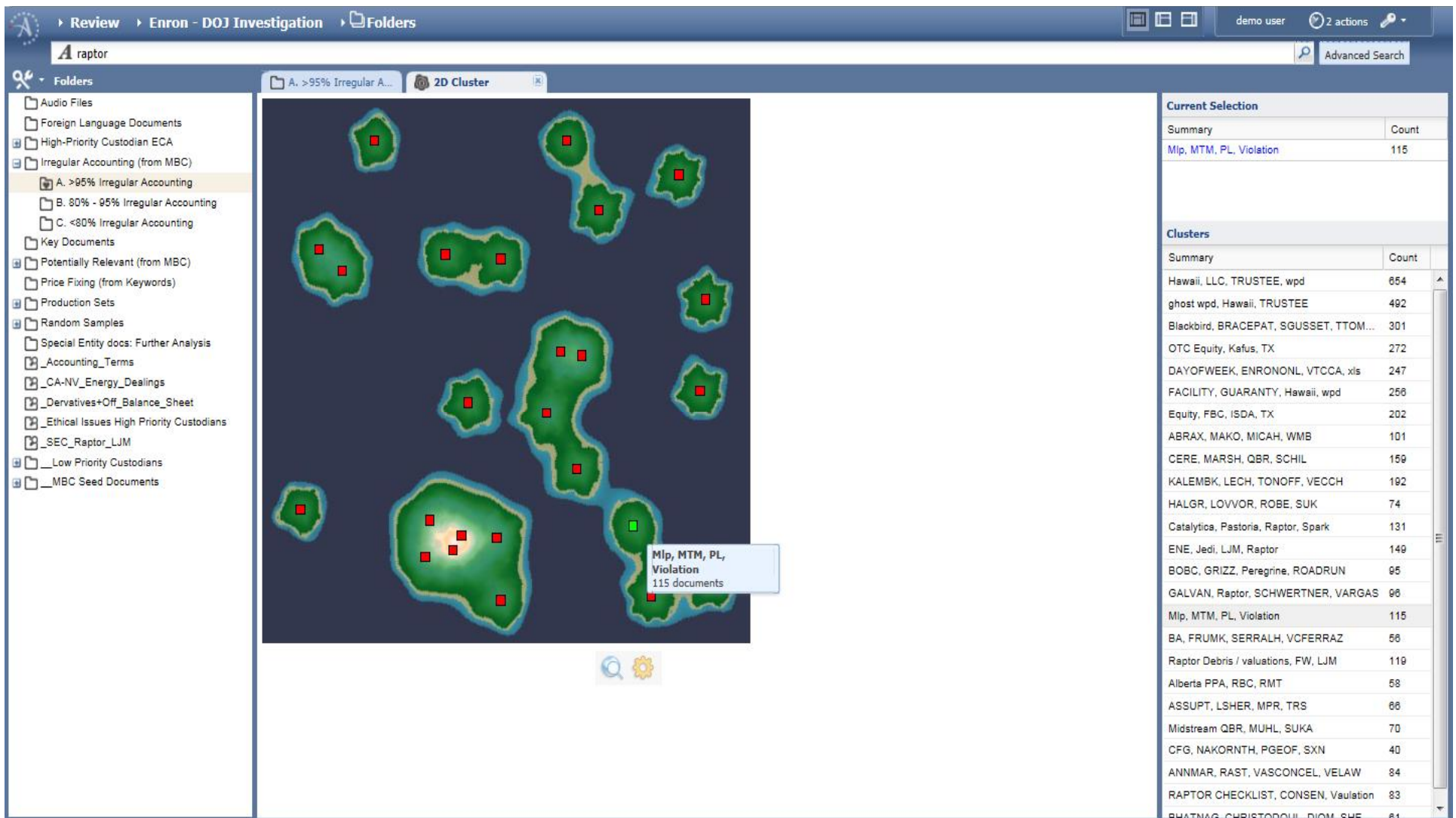
- Réduire la donnée
 - Réduction rapide
 - Réduction automatique
 - Réduction précise



Les méthodes de réduction

- Data Culling (filtrage 1er niveau)
- Déduplication
- Predictive Coding (Codage prédictif)
- Keyword/ Termlist search (mots clés)
- Advanced Search (Recherche approfondie)
- **Visualisation**
(Clustering, Taxonomie, Communication map)





Review

Enron - DOJ Investigation

Enron - DOJ Investigation

Folders

energy

demo user

5 actions

Advanced Search

Folders

Audio Files

Foreign Language Documents

High Priority Custodians (ECA)

1. Mark Taylor

2. Kenneth Lay

3. Sara Shackleton

4. Vince Kaminski

Irregular Accounting (from MBC)

Key Documents

Low Priority Custodians (ECA)

MBC Seed Documents

OCR Documents

Potentially Relevant (from MBC)

Production Sets

Random Samples

Redactions

Smart Folder: Accounting Terms

Smart Folder: CA-NV Energy Dealings

Smart Folder: Derivatives + Off Balance Sheet

Smart Folder: Ethical Issues High Priority Custodians

Smart Folder: SEC Raptor LM

Special Entity Docs: Further Analysis

TEST Saved

Tiffed Documents

3. Sara Shackleton

Link Map

Actions

"Sara Shackleton" <sara.shackleton@enron.com>

"Carol St. Clair" <carol.st.clair@enron.com>

"Sheila Glover" <sheila.glover@enron.com>

"Brent Hendry" <brent.hendry@enron.com>

Current Selection Panel

User	Domain	
"AKASHA.R.BIBB" <AKASHA.R.BIBB@ENRON.COM>	ENRON	
"ANNE.C.KOEHLER" <ANNE.C.KOEHLER@ENRON.COM>	ENRON	
"ARONOWITZ ALAN" <ALAN.ARONOWITZ@ENRON.COM>	ENRON	
"BAILEY SUSAN" <SUSAN.BAILEY@ENRON.COM>	ENRON	
"BRACKETT DEBBIE R." <DEBBIE.R.BRACKETT@ENRON.COM>	ENRON	
"BRADFORD WILLIAM S." <WILLIAM.S.BRADFORD@ENRON.COM>	ENRON	
"BRENDA.L.FUNK" <BRENDA.L.FUNK@ENRON.COM>	ENRON	
"BRENT HENDRY" <BRENT.HENDRY@ENRON.COM>	ENRON	
"CAROL.ST.CLAIR@ENRON.COM"	ENRON	

All Users

Add Filter Add Column Save Report Export

Email	Document Count
"ANDREA.V.REED" <ANDREA.V.REED@ENRON.COM>	36
"HOPE VARGAS" <HOPE.VARGAS@ENRON.COM>	29
"MICHAEL.S.GALVA" <MICHAEL.S.GALVA@ENRON.COM>	25
"LISA MELLENCAMP" <LISA.MELLENCAMP@ENRON.COM>	176
"SARA SHACKLETON" <SARA.SHACKLETON@ENRON.COM>	3,852
"BARTON CLARK" <BARTON.CLARK@ENRON.COM>	90
"BRENDA WHITEHEAD" <BRENDA.WHITEHEAD@ENRON.COM>	88
"BRIAN.SCHWERTNER" <BRIAN.SCHWERTNER@ENRON.COM>	21
"MARY.COOK@ENRON.COM"	127
"BRENT HENDRY" <BRENT.HENDRY@ENRON.COM>	661

Strong information and weak information

Large Words and small accounts information is strong information without "context"

“A heavy element and the only metal that is liquid at standard conditions for temperature and pressure with the symbol Hg and atomic number 80, commonly known as quicksilver”

[illegible]

*Lanthanide series

lanthanum 57	cerium 58	praseodymium 59	neodymium 60	promethium 61	samarium 62	euporium 63	gadolinium 64	terbium 65	dysprosium 66	holmium 67	erbium 68	thulium 69	ytterbium 70
La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb
57	58	59	60	61	62	63	64	65	66	67	68	69	70
138.905	140.12	140.907	144.24	144.912	150.36	151.964	157.25	158.925	162.50	164.930	167.259	168.930	173.054
actinium 89	thorium 90	protactinium 91	uranium 92	neptunium 93	plutonium 94	americium 95	curium 96	berkelium 97	californium 98	einsteinium 99	fermium 100	mendelevium 101	nobelium 102
Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No
89	90	91	92	93	94	95	96	97	98	99	100	101	102
227.03	232.0377	231.036	238.02891	237.04817	244.0642	243.0613	247.0763	247.07125	251.08295	252.0838	257.1037	258.10528	259.10698

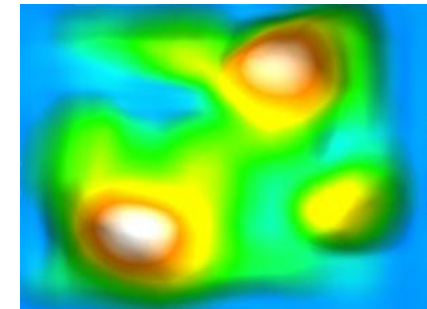
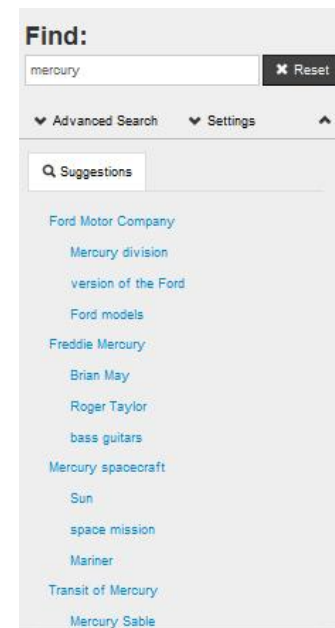
* * Actinide series

With high certainty, experience and element!

Compréhension du ‘sens’ de l’information

Utilisation d’algorithmes de comparaison et de modélisation statistique de manière à ‘**comprendre**’ le sens de la donnée

- **Modèle formé à partir des données reçues et non contraint par des dictionnaires pré-établis**
- **Indépendance de la langue**
(Traitement des mots comme symboles)
- **Analyse de sentiments.**



$$P(\theta | x) = \frac{P(x | \theta) \cdot P(\theta)}{\sum_{\theta' \in \Theta} P(x | \theta') \cdot P(\theta')}$$

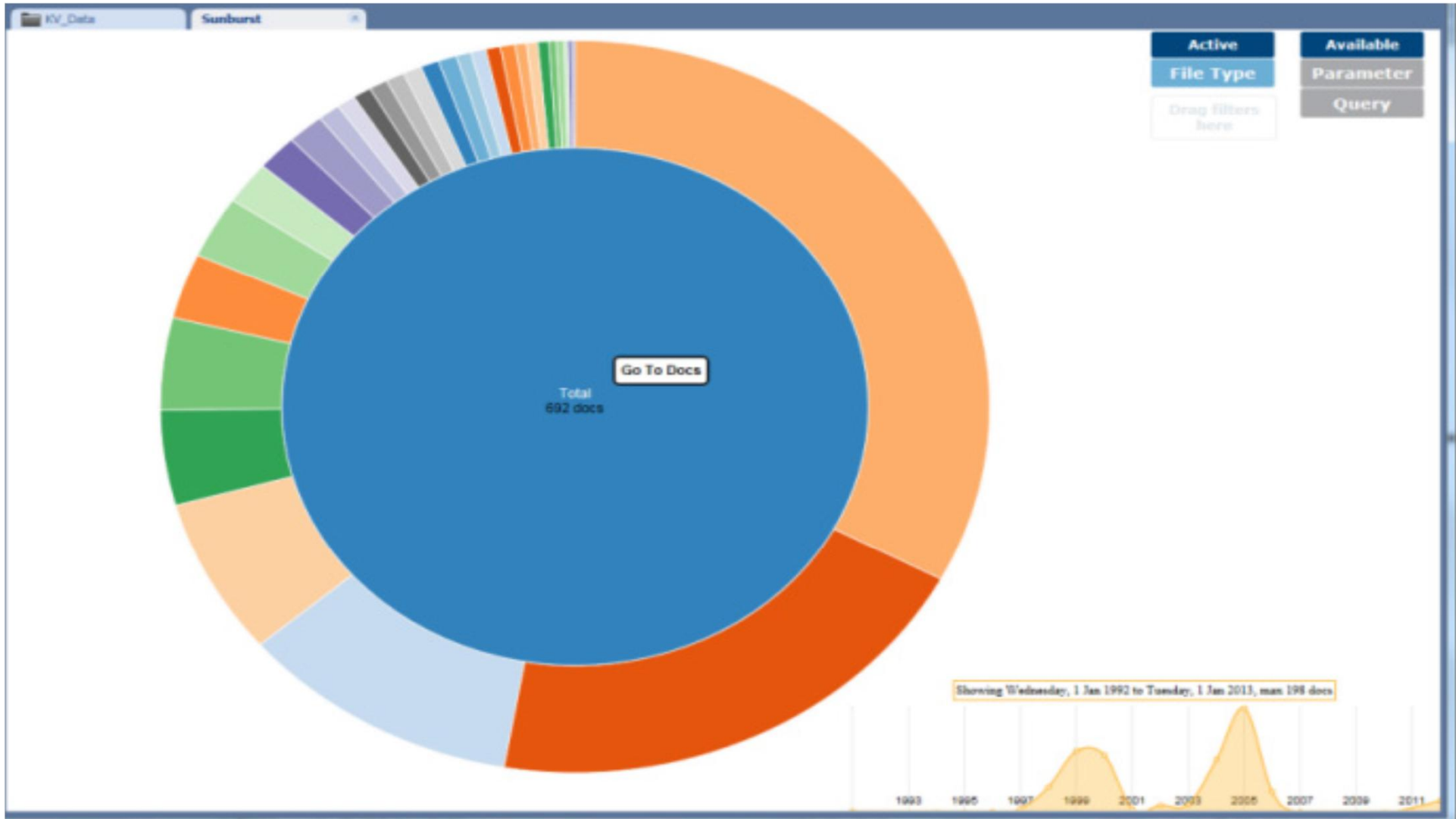
**Information Theory
and Bayesian Inference**

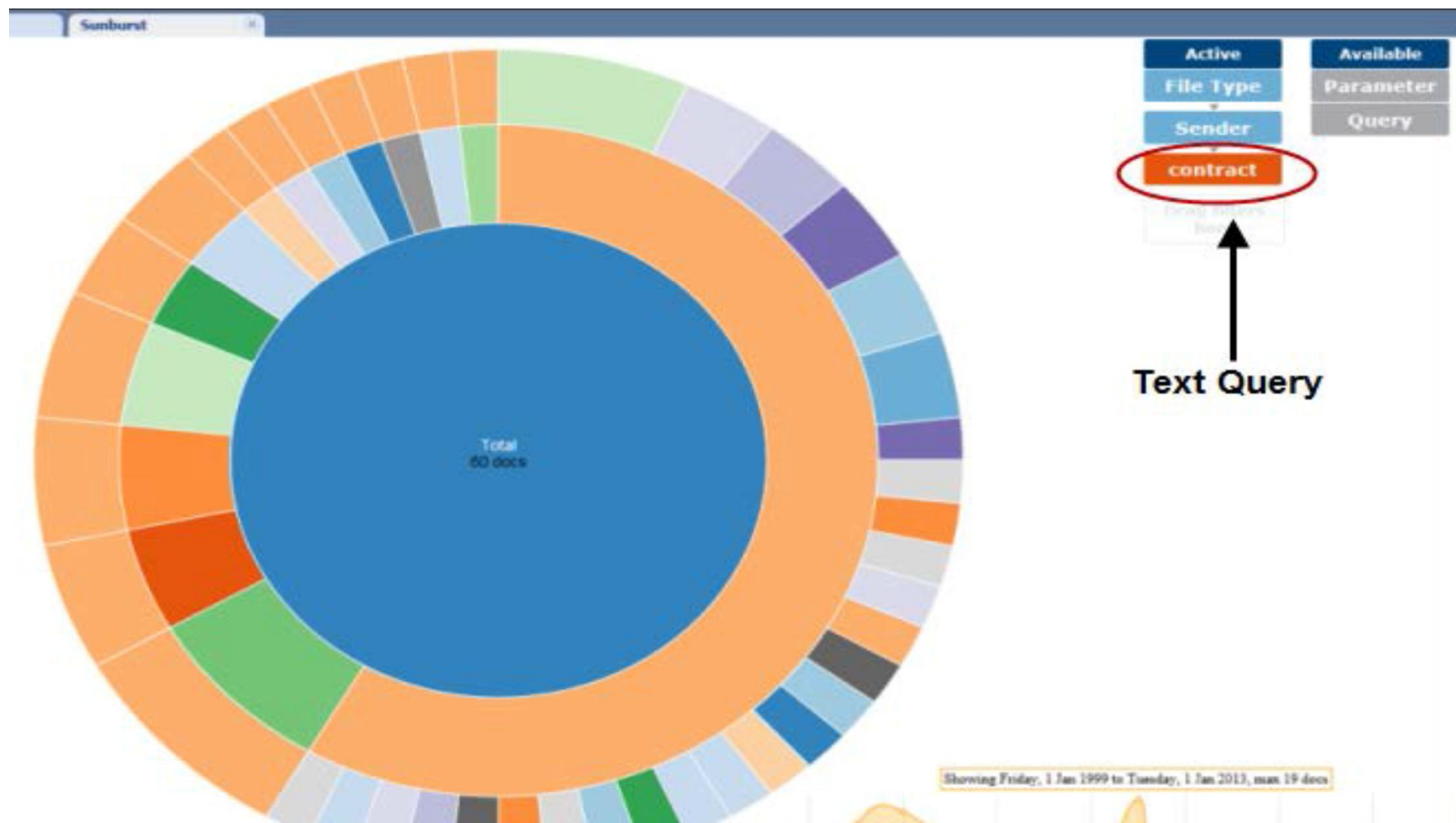
$$H = -\sum p_i \cdot \log_2(p_i)$$

Implications technologiques

- **Indexation rapide** des données massives
- **Indexation complète** (non limitée par les formats et dépôts)
- **Indexation intelligente** (métadonnées et contenu)
- **Interface flexible** de rendu visuel







Merci

