SEVENTH FRAMEWORK PROGRAMME

CORTEXT

IFRIS
Institut Francilien
Recherche Innovation Société

LabEx
Science, Innovation
and TEchnology in Society

Cortext.Risis Training

UPEM UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

ESIEE PARIS

INRA SCIENCE & IMPACT

Noisy-Champs 10-12 May 2017

WELCOME

To the training short course

Using the CorTexT-Risis Platform for Research in Science policy and Science-Technology-Studies

A Short Training Course proposed by CorTexT-Lab of the LISIS Unit with the support of IFRIS

May 10th & 12th 2017

Université Paris-Est Marne-la-Vallée

OBJECTIVES

- Being involved in the research communities of RISIS Project
- Getting knowledge and skills to run the Cortext.risis facility
- Being autonomous enough to mobilize Socio-Semantic Analysis on corpus for your own research
- Envisioning future project based on the cortext.risis facility within the RISIS FP7 project or elsewhere

PROGRAMME (1)

- **Day 1 – May 10th, 2017**

13h00: Welcome at *Bois de l'Etang*

13h30h-14h45:

. Introduction : Objectives of the training (Chair MB) and Cortext-Risis in context (Chair MB)

. Presentation of Participants (who & interest)

14h45-16h30:

CorTexT Platform : Context, Organisation and Features of a Digital Platform for SSH (Chair MB)

Principles of Architecture and Login & Access (Chair PB)

Coffeee Break

16h00-17h00 : DATA : Presentation of some RISIS resources (Chair LV)

17h00- 18h00: Setting up Datasets and Parsing in CorTexT Project (Chair LV)

Demo followed by Learning-by-doing on Datasets

PROGRAMME (2)

- **Day 2 – May 11th 2017**

9h-10h30: Metrics of Similarities in networks of itemsets: lecture (Chair JPC)

Work in 3 groups on the same datasets in a shared Project on a Sample Corpus - With assistance

Coffeee Break

11h-13h00: Back to Corpus : how to explore and create lists or time selection (Chair JPC)

Work in 3 groups on the same datasets in a shared Project - With assistance of the team

13h-14h: Lunch

14h-16h: Datasets analysis for structural and dynamics clustering (Chair : JPC)

Work in 3 groups on the same datasets in a shared Project : Graph interpretation and temporal analysis - With assistance of the team

Coffeee Break

16h00-17h00: Setting up individual Project for the next days (Chair : LV) With assistance of the team

17h00-18h00 : Recap, feedback, discussions and follow-up (Chair : MB)

20h: Diner (Paris)

PROGRAMME (3)

Day 3 – May 12th 2017

10 :00-13 :00 : Participants at work with individual training on project

With assistance of the team

Including an Open Coffeee Break

13h-14h: Lunch

14h-15h00: Participants at work with individual training on project

With assistance of the team

15h-16h30: Participants present their achievments and address questions (Chair MB)

16h30-17h00 : Closing of the training (comments, feedback, auto-evaluation) (Chiar MB)

Coffeee Break for living

RESSOURCES

- Wifi available vias Eduroam
- URL : http://cortext.risis.eu
- Repository (google drive): https://goo.gl/1eptRX

SEVENTH FRAMEWORK
PROGRAMME

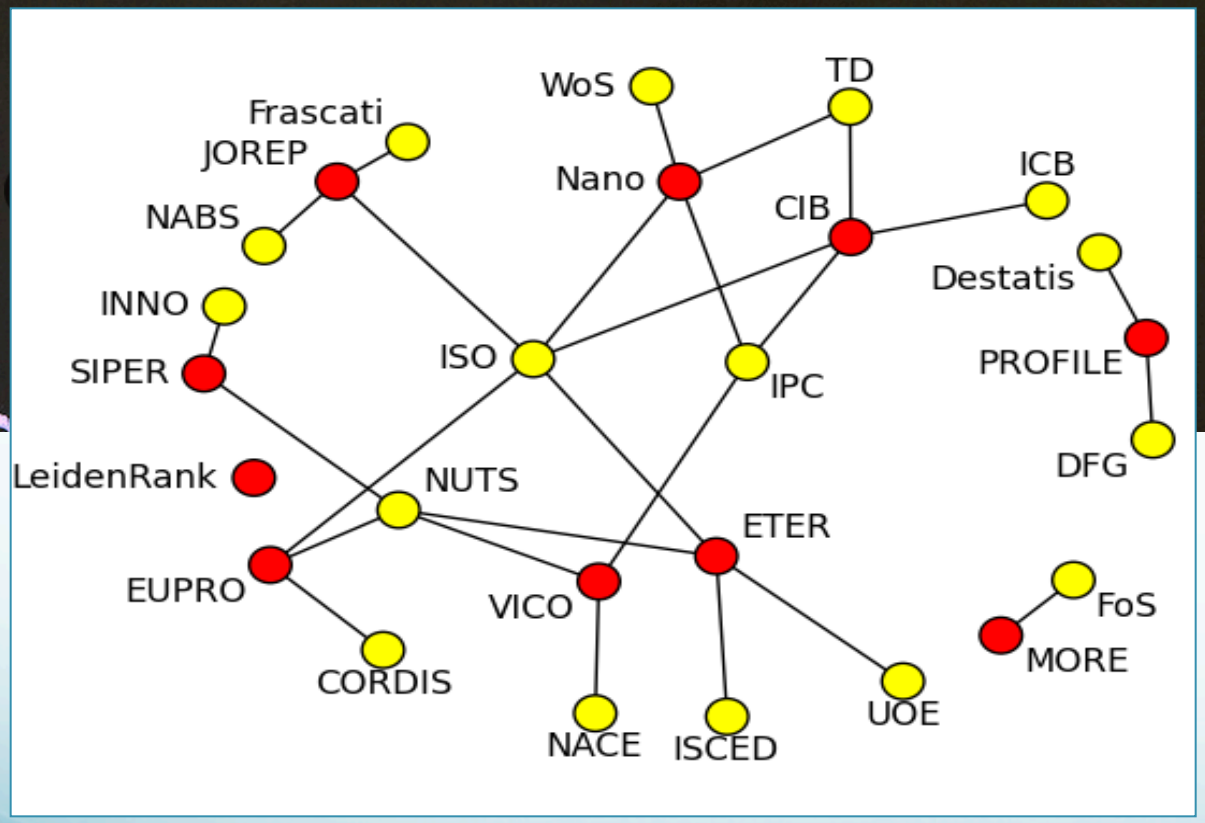
The Cortext.Risis Facility in context of the RISIS FP7 Infrastructure project

RISIS
Research infrastructure for research and innovation policy studies

Contact

About Members Events Datasets Results

RISIS is building a distributed infrastructure for research and innovation dynamics and policies.



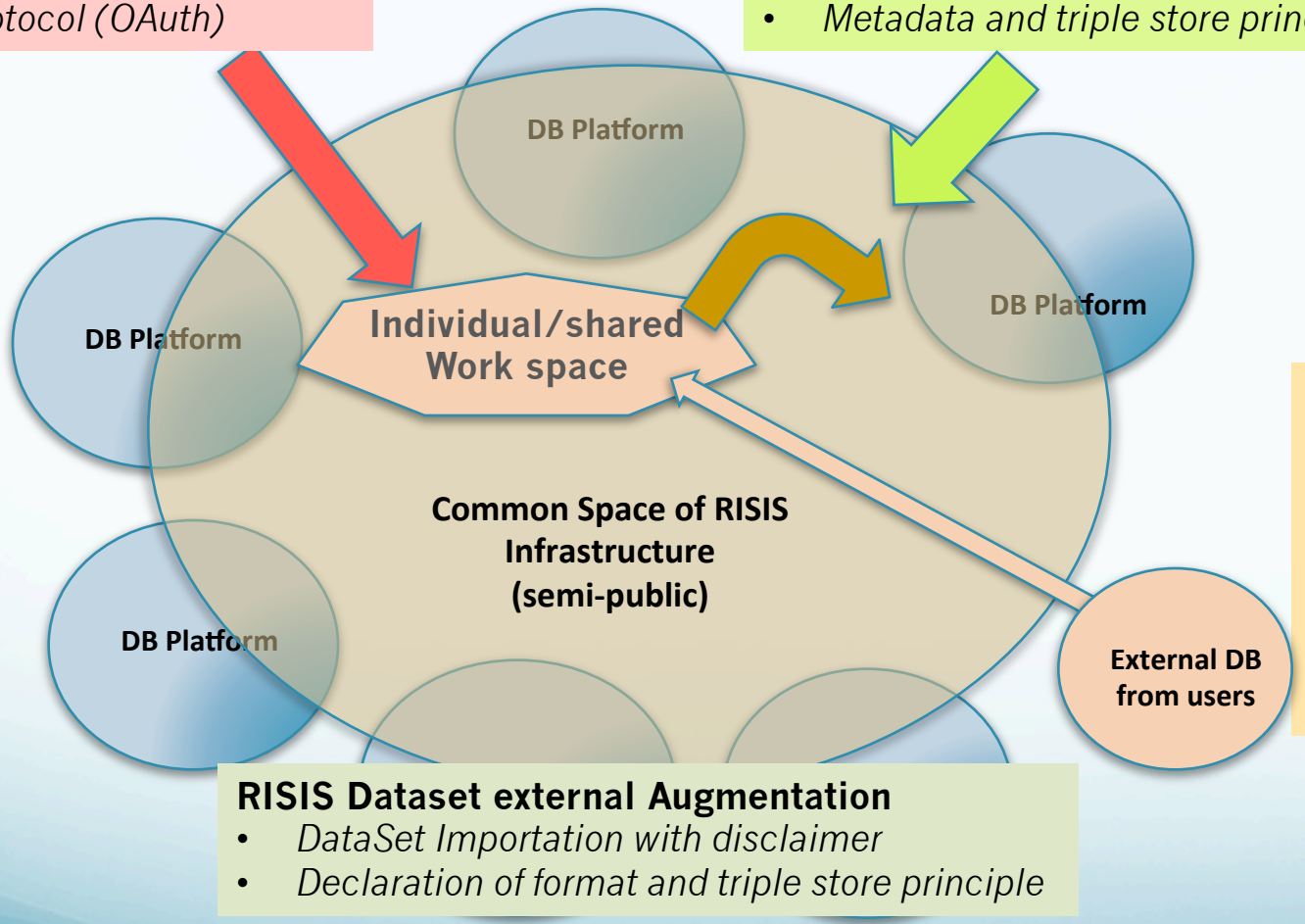
What will be ?

A NETWORK OF CAPACITIES TO BE TRANSFORMED IN A COMMON

Principles for the Development of a Common Space

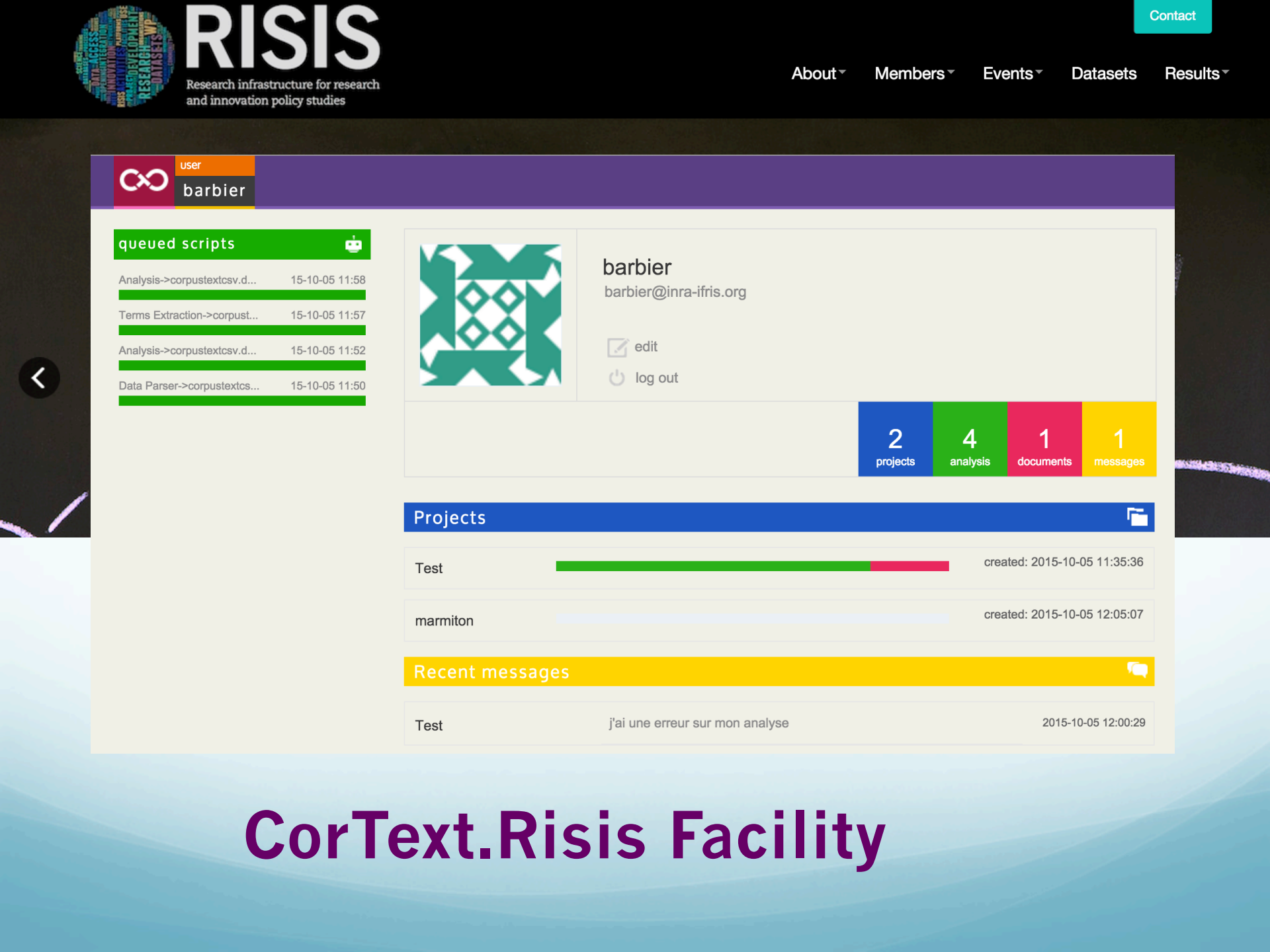
RISIS Unique Entry-Point
- *User Accreditation*
- *Global Authentification Protocol (OAuth)*

RISIS Platform Commons
- *PF Accreditation based on legal rules, robustness and quality principles + common/local division of Data*
- *Global Authentification Protocol (Oauth) acceptance*
- *Metadata and triple store principle*



RISIS Dataset Internal Augmentation
- *Added DataSets to be accredited*
- *Localisation of Dataset in PF*
- *Added value*

RISIS Dataset external Augmentation
- *DataSet Importation with disclaimer*
- *Declaration of format and triple store principle*



CorText.Risis Facility

ROUNDTABLE PRESENTATION

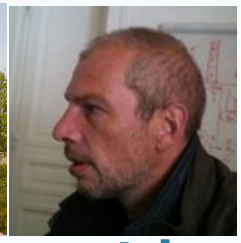
N°	Name	Surname
1	Yousefdehi	Hami
2	Bento	Nuno
3	Yang	LIU
4	Brissaud	Constantin
5	Li	Sisi
6	Ayrapetyan	David
7	Galindo Moreno	Manuel Ricardo
8	Fustec	Klervi
9	Abdelghani	Maddi
10	Rikap	Cecilia
11	Briday	Régis

CorTexT Platform : Context, Organisation and Features of a Digital Platform
An overall view

Using the CorTexT-Risis Platform for Research in Science policy and Science-Technology-Studies

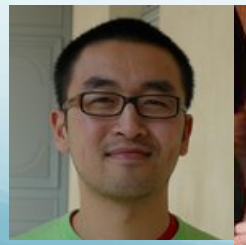
People and Goals

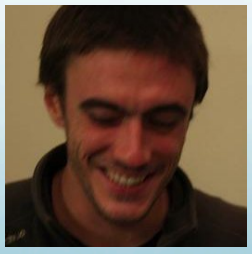






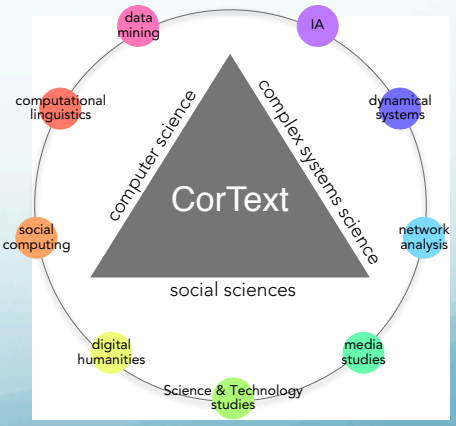






In the name of the CorText Team

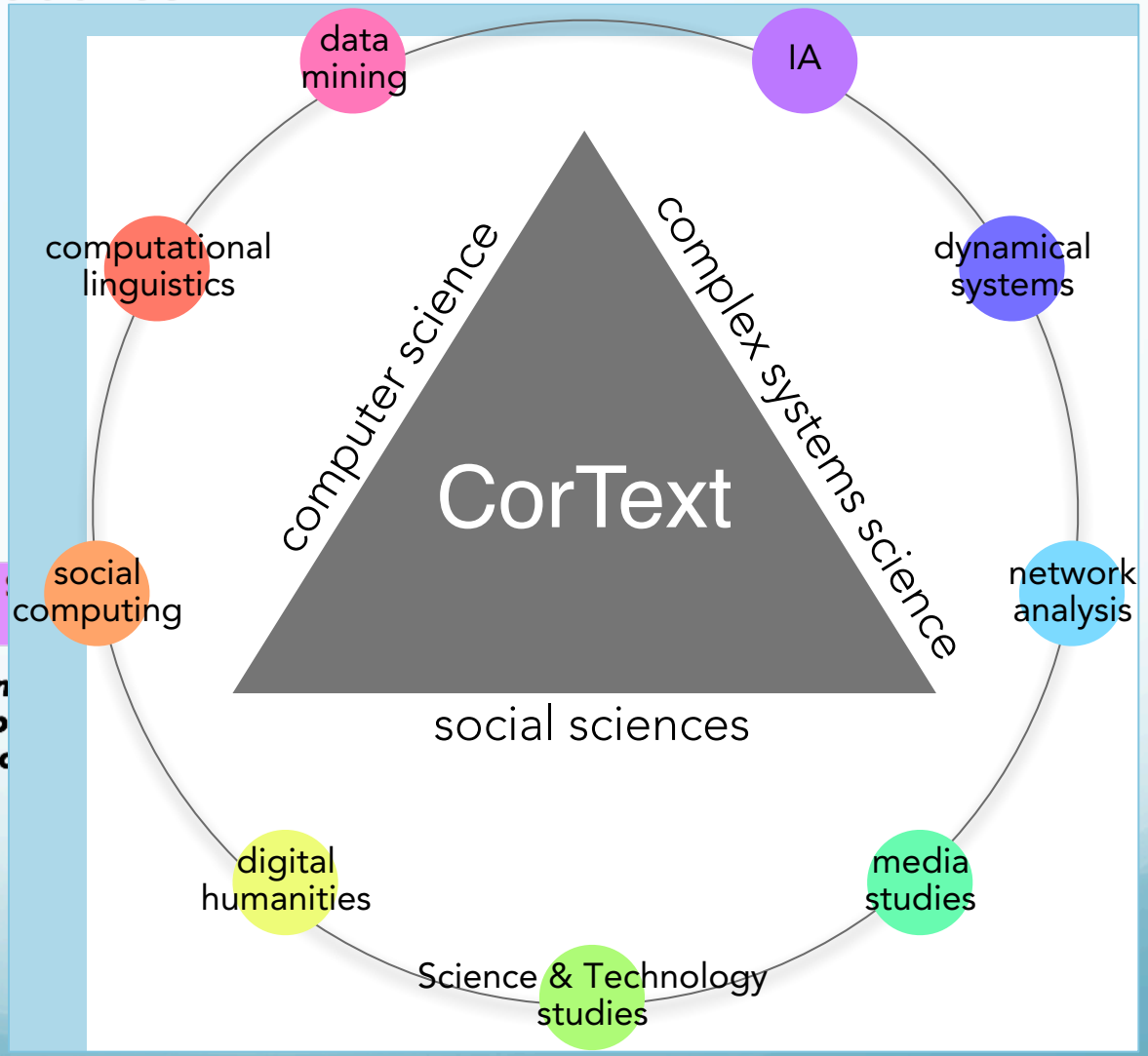
- Barbier Marc (Dir.)
- Breucker Philippe
- Cointet Jean-Philippe
- Duloquin Clhoé
- Duong Tam-Kien
- Laurens Patricia
- Martinez Cristian
- Mazières Antoine
- Mogoutov Andreï
- Schoen Antoine
- Turenne Nicolas
- Villard Lionel



The origin of our world / what could be original?

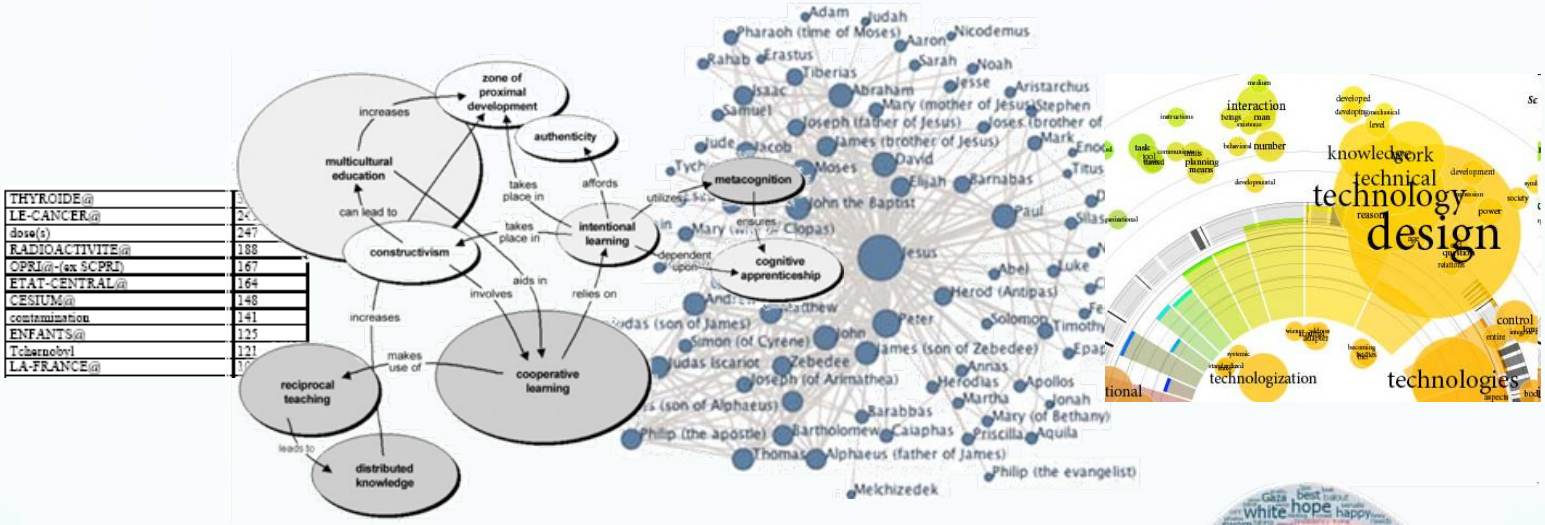
L' alliance fondatrice

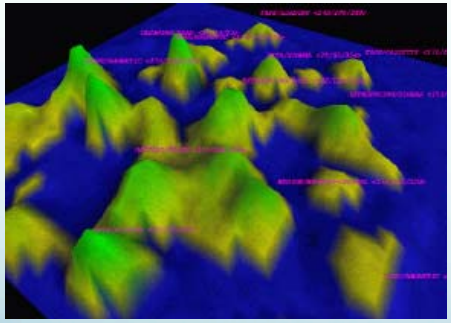
* 1955, Eugene Garfield crée un répertoire interdisciplinaire pour les bibliothécaires regroupant les articles des principaux périodiques et leurs références (Garfield, 1955).
* 1958, fondation de l'Institute for Scientific Information (ISI) et première version papier du Science Citation Index (SCI) en 1963.
* 1963, Solla Price publie "Little science, big science" et impose l'idée d'une mesurabilité de la production scientifique reflet de loi sociales.

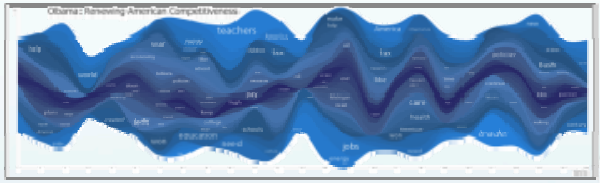


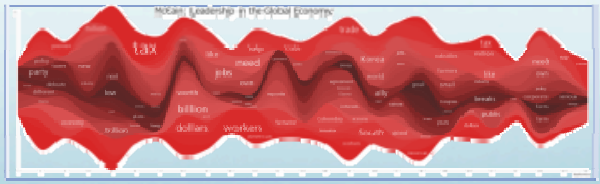
Measurement of text

The Many Ways of vizualising DataSets









www.visualcomplexity.com

Neoformix.com

An Epistemic Challenge for STS Researchers

- Pixelisation of sciences/society debates on the web
- Streams of d@t@ in any production system or business activities
- Time and Space of Research Activities (extraction of massive set of data, artificial experimenting, practices accountability)

Political Changes with Science-in-Society Accountability

Tools & Skills for Science Policy following an Alliance of Artificial Intelligence and Human & Social Sciences: library sciences, scientometrics, research management, collaborative accountability, web design

A technological Challenge for old-IA

Tools & Skills for the design of technological platforms for research: pluridisciplinary work between IT Engineers, Linguistic and Information Science and Human & Social Scientists (historian, sociologist, economist,…)

The challange of "Big Data"

nature Vol 455|4 September 2008

COMMENTARY

How do your data grow?

Scientists need to ensure that their results will be managed for the long haul. Maintaining data takes big organization, says **Clifford Lynch**.

Data can be "big" in different ways. National and international projects such as the Large Hadron Collider (LHC) at CERN, Europe's particle-physics laboratory near Geneva in Switzerland, or the Large Synoptic Survey T for northern Chile, are freque way they will challenge the s computation, networking a But research data can also b lasting significance — a clini the observation of a unique e big because of descriptive cha require context such as the exp Because digital data are so e replicated and so recombina tremendous reuse opportuni investigations already under advantage of past investments

To enable reuse, data must b In some cases the effects of d nomic, because experiments h In other cases, data loss repres nity lost forever. Funders now as assets that they are underw the greatest pay-off for their i demand that researchers and host institutions document and implement data-management and data-sharing plans that address the full life cycle of data — including what happens after a grant finishes. Host universities thus find themselves with legal and ethical obligations to provide a legacy of faculty data. Publishers must also identify the most effective ways to connect publications with data and preserve the scientific record.

Developing infrastructure

Managing the life cycle of scientific data presents many challenges. These include deciding responsibilities, funding, resource allocation, what data should be kept and for how long.

In a sense, landmark international projects like the LHC are the least problematic: the costs of data management are explicit in the budget and tend to be dominated by technology expenses that decline over time. These projects also include dedicated personnel; and, although the volume of data is often vast, the streams fit within well defined descriptive schemes.

example, have invested substantially in common infrastructure for a more systematic reliance on data, networks and computation. And there are vast numbers of scientific research projects producing at most a few terabytes per year of big data, or data that can be aggregated into a big-

> Ultimately, the best stewardship of data will come from disciplinary engagement with preservation institutions. General-purpose data management as provided by universities through their research libraries will have its limits. Where there is no natural locus of disciplinary stewardship, universities will need to establish consortia to enable disciplines to create and sustain such engagement[4].

possible here. In 2007, the US National Science Foundation, recognizing the importance of such standards, established the Community Based Data Interoperability Networks (INTEROP) funding programme for the development of tools, standards and data management best practices within specific disciplinary communities. INTEROP should make its first awards this autumn. Although many classes of scientific data aren't ready, or aren't appropriate, for standardization, well chosen investments in standardization show a consistently high pay-off[3].

"The best stewardship of data will come from engagement with preservation institutions."

At the start of the data life cycle, individual scientists will have primary responsibility for stewardship. But longer term, data preservation can only be done by institutions. If data are to be consolidated or shared on a frequent basis, there is a lot to be said for moving to institutional control sooner rather than later. Scientists are not necessarily good data managers and can more fruitfully spend their time doing science

information management tasks to a rotating staff of students and postdocs. Indeed, as specific data sets become distant from current research activities, stewardship can become a tax on scientific productivity.

Scientists need to act responsibly during des working linary stand- nd recording s experimen- allow for data when the data ive metadata experimental racing prov- e from, how ence on other their capture. ocumenting e community y defined and software. It is uch software

hort term, it is use. In a high- or university's network, machines will often be compromised if updates aren't applied; this can mean data destruction or corruption. Disasters such as Hurricane Katrina, which destroyed labs and computing facilities, are important reminders that data need to be backed up frequently and comprehensively in diverse and distant locations. Appropriate use of IT services such as secure storage or hosting from the host institution may be valuable. In the longer term, digital data is at risk from various forms of technological obsolescence (particularly if locally held removable storage media are used). There is a need for new institutional services that can help with all these needs, handling traditional IT issues and information-management issues more familiar to librarians and archivists.

At some point, the primary copy needs to migrate to an institutional service. Today, these services are sparse. In the United Kingdom there are data services associated with several

Changement in our Infrastructures

- For Human and Social Challenges : new "digital librairies", new techniques of text mining, new algorithms of network analysis, and new institutional contexts for Research

A creole Landscape

- **Many sub-disciplines**
- Scientometrics
- Informetrics
- Webometrics
- Webstudies
- Network Studies
- CWS studies
- Information Extraction
- TAL
- Knowledge visualisation

- **Tracking Projects**
- Platforms de Natural Langage Processing
- Plateforms of Science& Technology Mapping
- Digital Humanitis Platforms Plateforme Humanités Digitales

Ecosystems of Platforms

Anderson, S & Blanke, T (2012). Taking the Long View: From e-Science Humanities to Humanities Digital Ecosystems, HISTORICAL SOCIAL RESEARCH-HISTORISCHE SOZIALFORSCHUNG, Volume: 37 Issue: 3 Pages: 147-164.

CORTEXT: Goal and aims

GOAL

To provide a **digital platform available** to « RISIS research groups » and to impact the practices of Research in Science policy and Science-Technology-Studies

AIMS

- to equip scientists with tools that enable them to tackle the complexity of heterogeneous textual corpora dynamics
- to develop innovative analytical methodologies that will bring new insights and renewed capacities to investigate contemporary issues of Sciences Innovation and Technology in Society

The CorTexT.Risis team provides

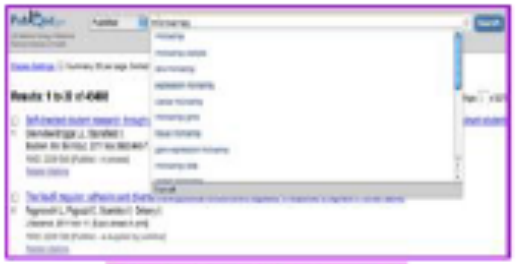
- Tools, process, scripts, procedures and methods encapsulated in an On-Line Open Access Digital Platform: *www.cortext.risis.eu*
- Skills, methods and training competencies to be mobilized in Training Session and projects of RISIS Associated Labs

Main features for for Research in Science policy and Science-Technology-Studies

scientific productions

Web Of Science ISI

Microsoft Academic Search



Medline Pubmed

specific databases

rare disease database

projects database

clinical trials database

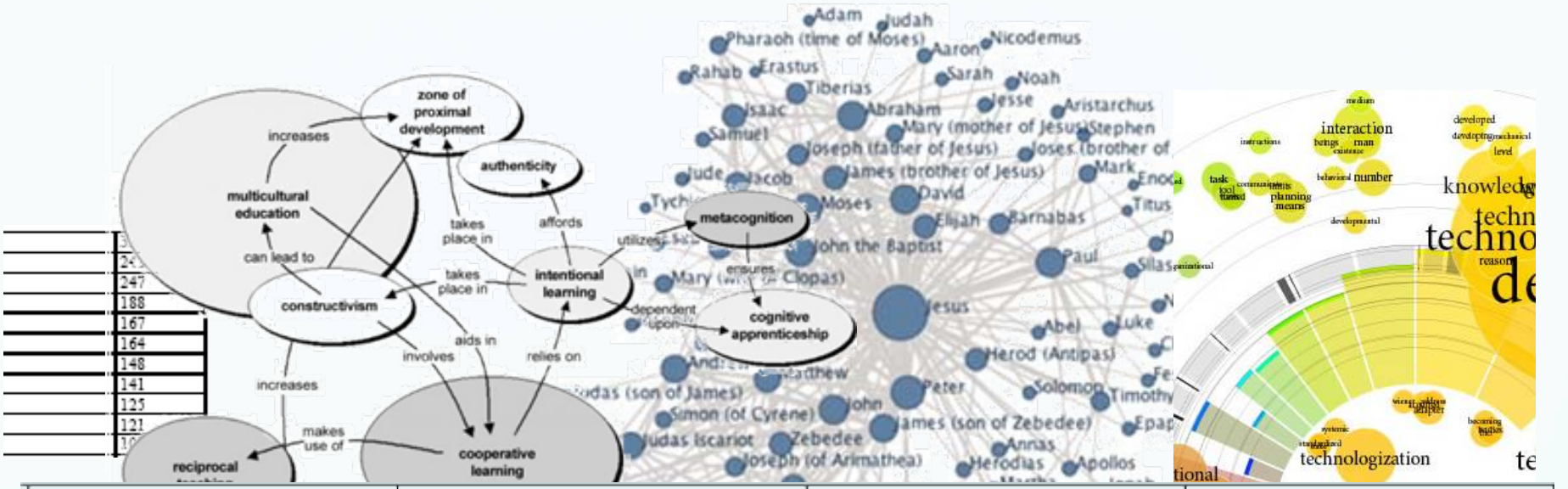
media productions (press+web)

web crawler

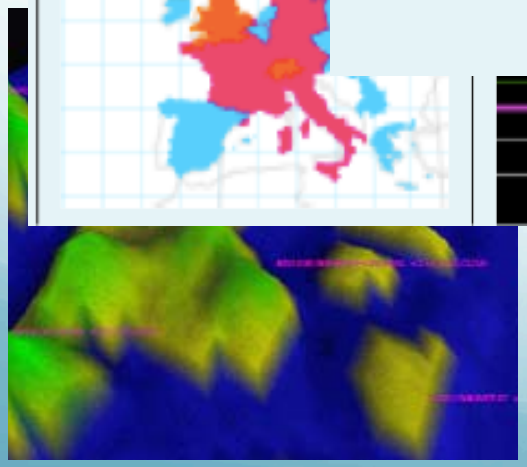
Factiva, press articles archive

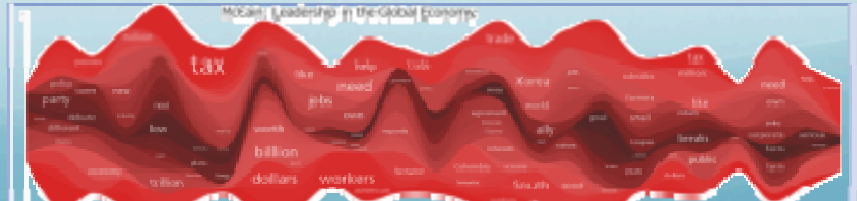
online forums

The Many Ways of vizualising DataSets

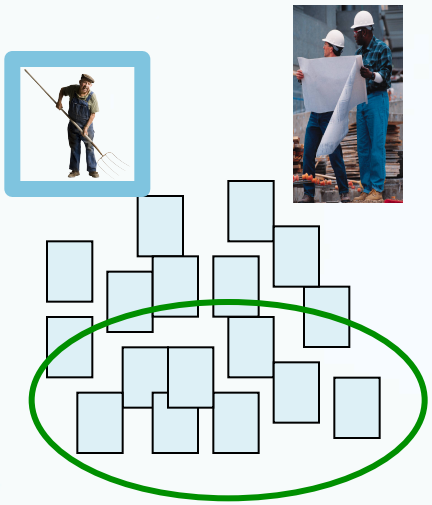


Dimensions of Data Analysis





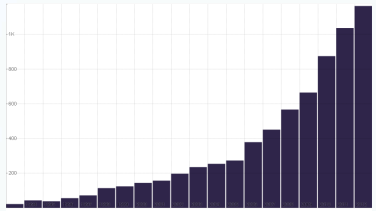
Existing Practices of Datasets and Corpus Processing



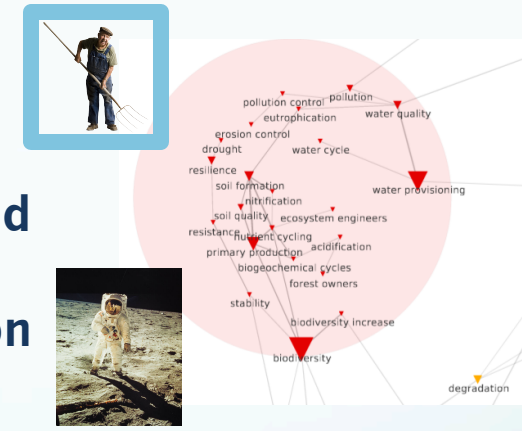
Extraction /Building of a corpus

Selection / Creation of Field descriptors

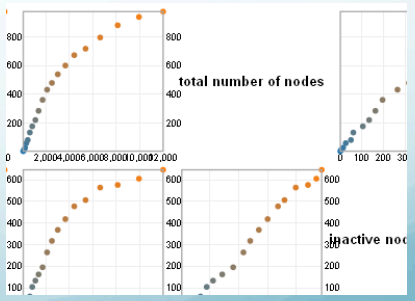
Statistics of Descriptors

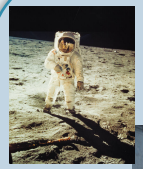


Network Analysis and Knowledge Vizualization



Temporal and Structural Analysis and modelisation





The Scientist

The IT Eng.

The Field Expert

Direct

Keywords, Date, Institu., Codes, etc.

Indirect

Lexical Extraction from Natural Langage Processing

Le modèle de production agricole subit actuellement un changement majeur vers la réduction des intrants polluants, dont les pesticides. Or, réduire l'usage des pesticides tout en maintenant la productivité actuelle ne se fera pas sans innovations techniques et organisationnelles originales.....

Indexation (*Parsing and Tagging*)

The phylogenetic position of the elephant shark (Callorhinchus milii) is particularly

DT JJ NN IN DT NN NN (NNS NN)VBZ RB

relevant to study the evolution of genes and gene regulation in vertebrates.

JJ TO VB DT NN IN NNS CC NN NN IN NNS

Tracking of Nominal Group (*tag chuncking*)

```
gene regulation in vertebrate -> {gene regul vertebr}
phylogenetic position of the elephant shark : {eleph phylogenet posit}
phylogenetic position -> {phylogenet posit}
```

Sampling simple Semantic forms (*stemming & Filtering*)

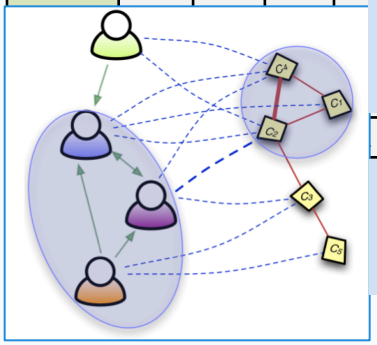
stem	main form	forms	n	C-value	Specificity	Frequency
alga red	red algae	red algae|&|RED ALGAE|&|Red algae|&|red alga	2	703,3	1292,3	464
matter organ	organic matter	organic matter|&|Organic matter				
chlamydomona reinhardtii	Chlamydomonas reinhardtii	Chlamydomonas reinhardtii				
higher plant	higher plants	higher plants|&|HIGHER PLANTS|&|higher plan				
acid amino	amino acids	amino acids|&|amino acid				
lactuca ulva	Ulva lactuca	Ulva lactuca				

Statistics of Occurrences of Simple Forms (*C-Value*))

	Résumé						**Labos**			
	C1	*C2*	*C3*	*C4*	*C5*	*C6*	*Lab1*	*Lab2*	*Lab3*	*Lab4*
Proj1	1	0	2	0	1	0	1	0	1	1
Proj2	1	3	0	1	0	1	1	1	1	1
Proj3	1	1	2	3	1	0	1	1	0	1
Proj4	1	0	0	1	2	3	1	0	1	0

Re-building Datasets in tables (System of interrelated Tables)

Résumé						Labos			
C1	C2	C3	C4	C5	C6	Lab1	Lab2	Lab3	Lab4
	2	3		1	0	1	0		



Tables of related Data

A conceptual framework: co-word analysis

Cooccurrences variables

- *Co-Occurrence matrix C : C_{ij} = number of joint occurrences of i and j in the same document*
- total number of cooccurences of i : $s_i = \sum_{j,j \neq i} c_{ij}$
- global number of co-occurences : $N = \sum_i s_i$
- expected number of coccurrences : $e_{ij} = \frac{s_i s_j}{N}$

	C1	C2	C3	C4	C5
C1		1		1	
C2	1		1	2	
					1
C5			1		

Calculation of co-occurences

Direct Measures of Similarity :

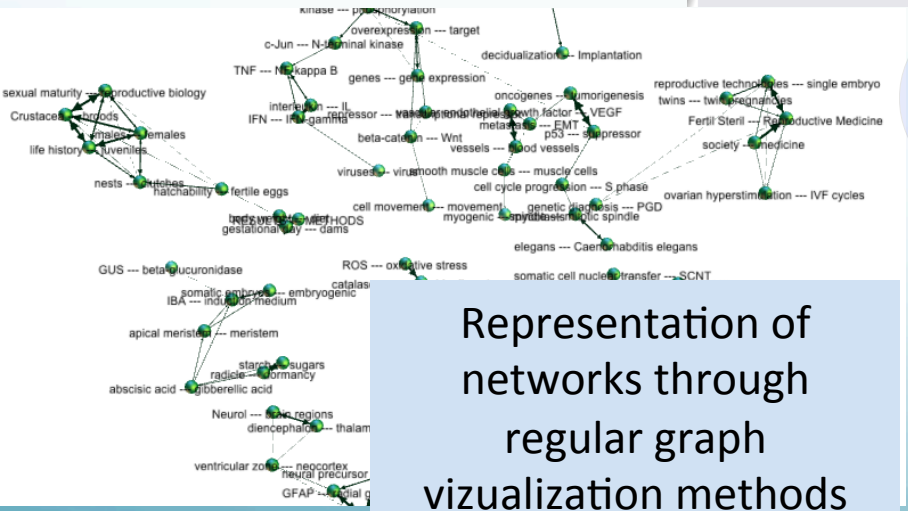
- *Raw cooccurrences : $S_R(i,j) = c_{ij}$*
- *χ^2 score : $S_{\chi^2}(i,j) = \frac{c_{ij} - e_{ij}}{\sqrt{e_{ij}}}$*
- *Mutual Information : $S_{MI}(i,j) = \log(\frac{c_{ij}}{e_{ij}})$*

Indirect Measures of Similarity :

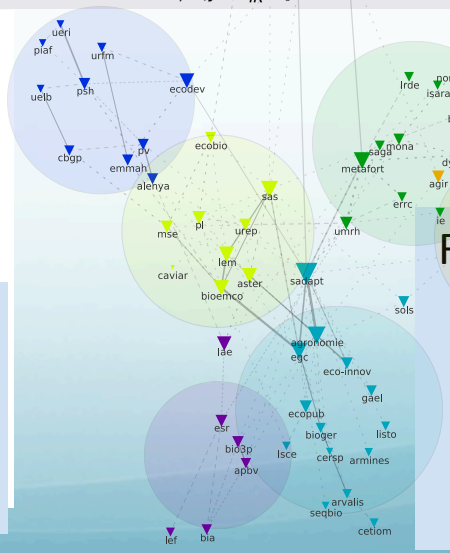
- Mutual Information (distributional) :

$$S_{MId} = \frac{\sum_{k \neq i,j; MI_{ik} > 0} \min(MI_{ik}, MI_{jk})}{\sum_{k \neq i,j; MI_{ik} > 0} MI_{ik}}$$

Choosing the metrics of similarity



Representation of networks through regular graph vizualization methods



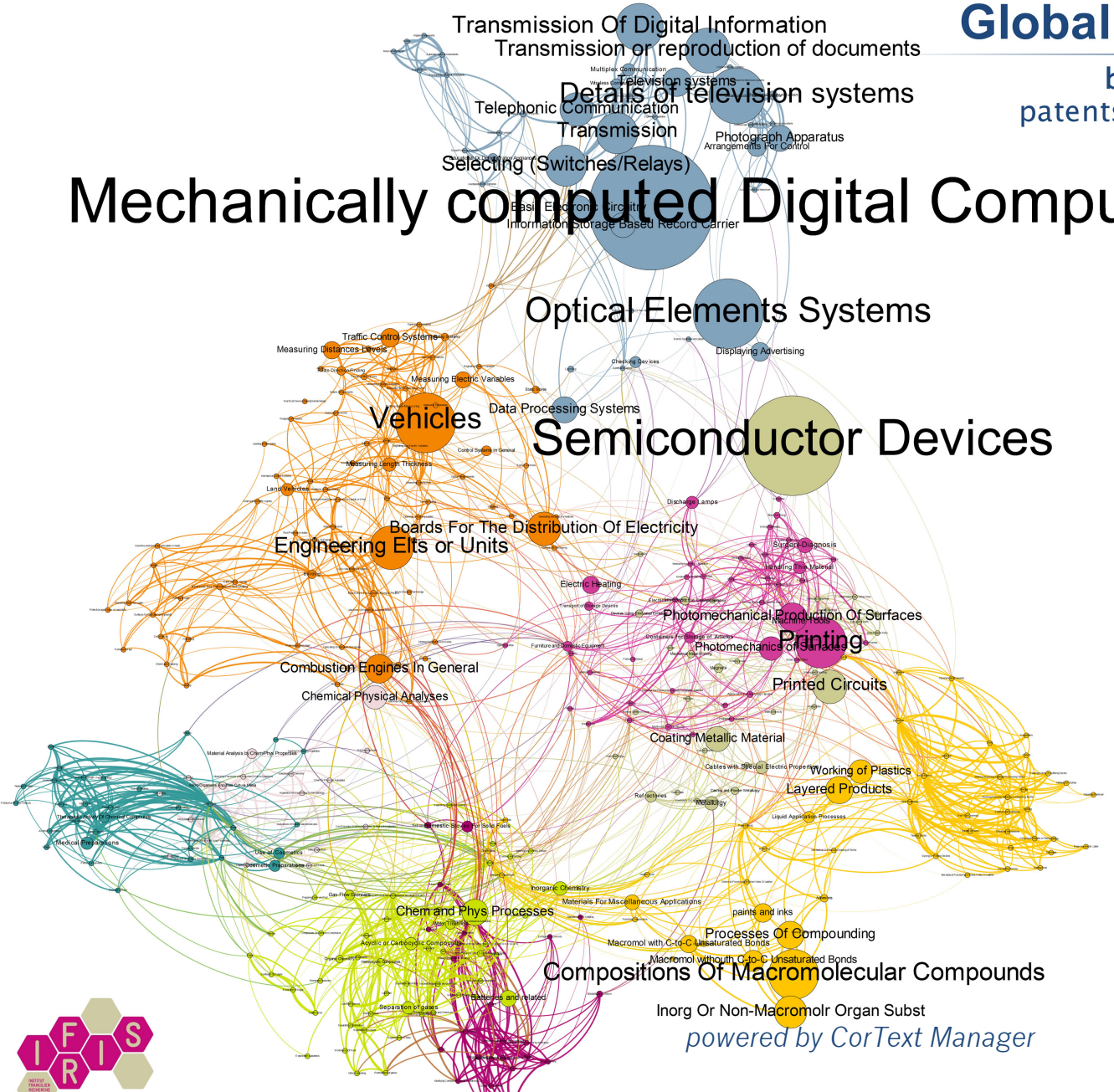
Representation of Clusters of Nodes-Nodes relations thanks to Community Detection Methods

(Louvain Algo. Louvain, method of partition optimization)

Maps of Technology through Patents DB

Global Map of Technology

based on IPC co-occurrence
patents applied worldwide,1996-2005



3 099 093 priority patents

Clustered with CorText Manager

Visualisation created with Gephi

9 technological clusters

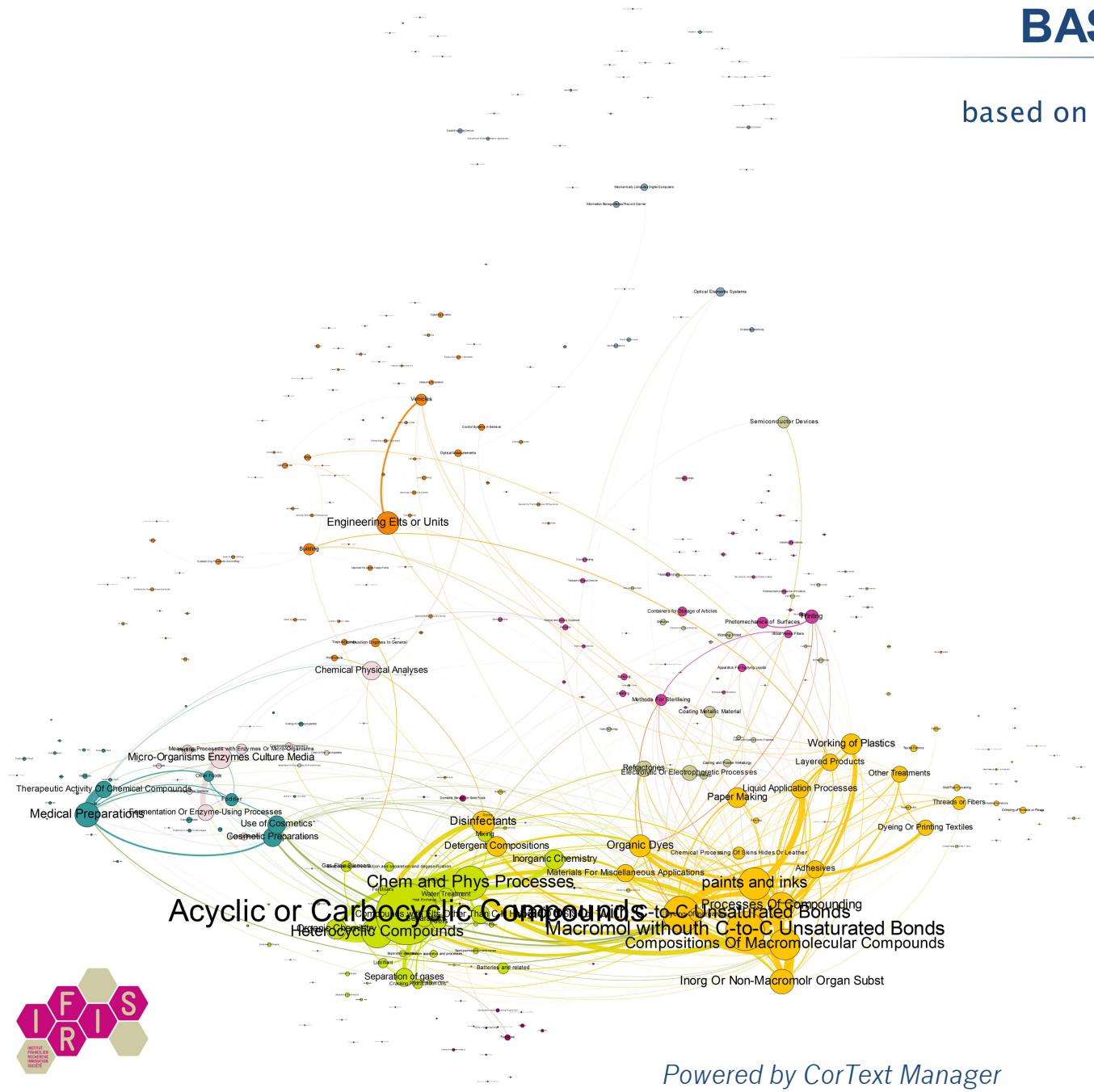
Vehicles
Compositions Of Macromolecular Compounds
Chem and Phys Processes
Printing
Mechanically computed Digital Computers
Semiconductor Devices
Medical Preparations
Domestic Stoves For Solid Fuels
Chemical Physical Analyses

powered by CorText Manager

BASF patents portolio

Global Map Of Technology
based on IPC co-occurrence,1996-2005

9 792 priority patents



9 792 priority patents

Clustered with CorText Manager

Visualisation created with Gephi

9 technological clusters

- Vehicles
- Compositions Of Macromolecular Compounds
- Chem and Phys Processes
- Printing
- Mechanically computed Digital Computers
- Semiconductor Devices
- Medical Preparations
- Domestic Stoves For Solid Fuels
- Chemical Physical Analyses

Powered by CorText Manager

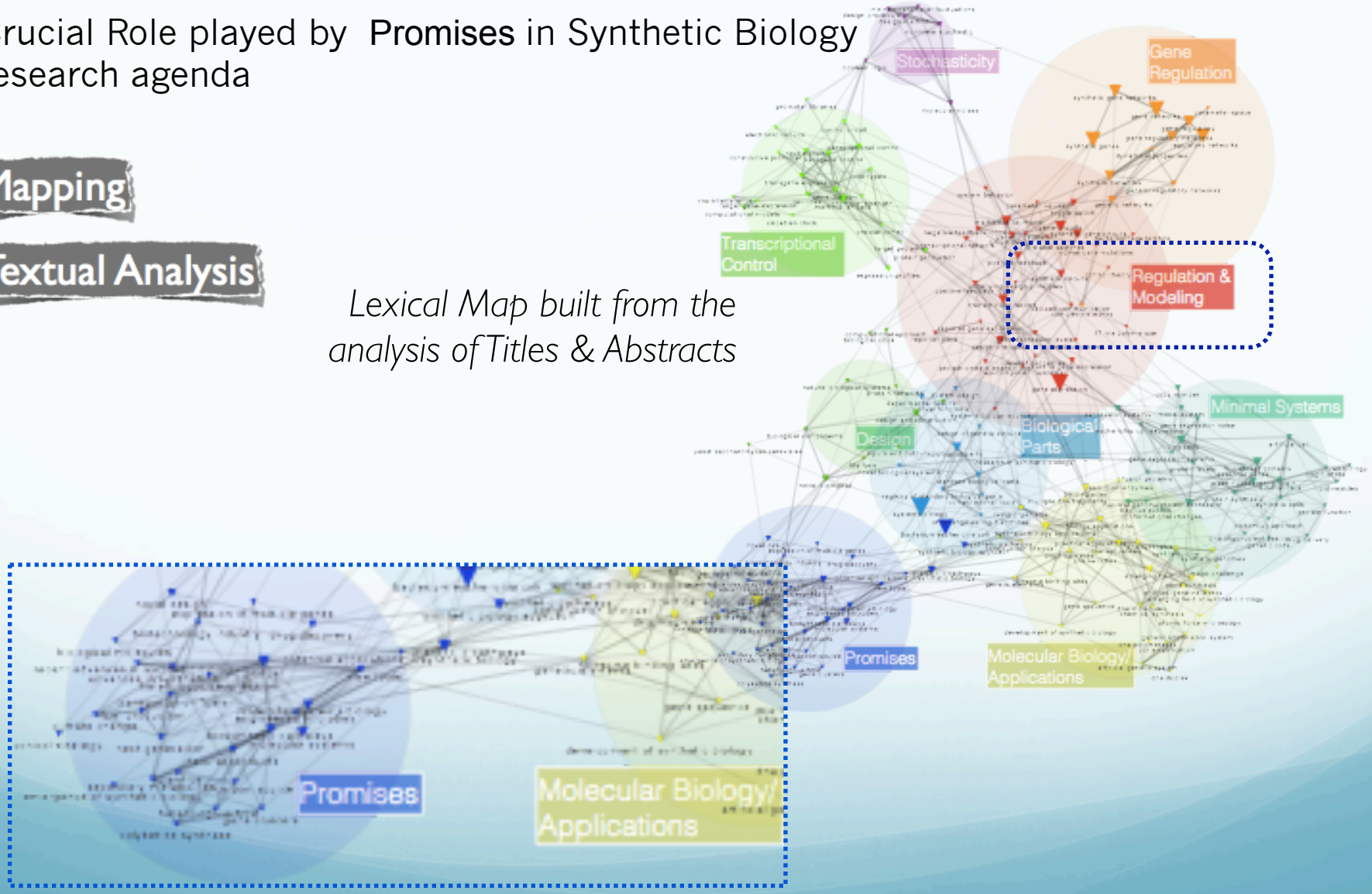
Discourse Analysis

- Crucial Role played by **Promises** in Synthetic Biology research agenda

Mapping

Textual Analysis

Lexical Map built from the analysis of Titles & Abstracts

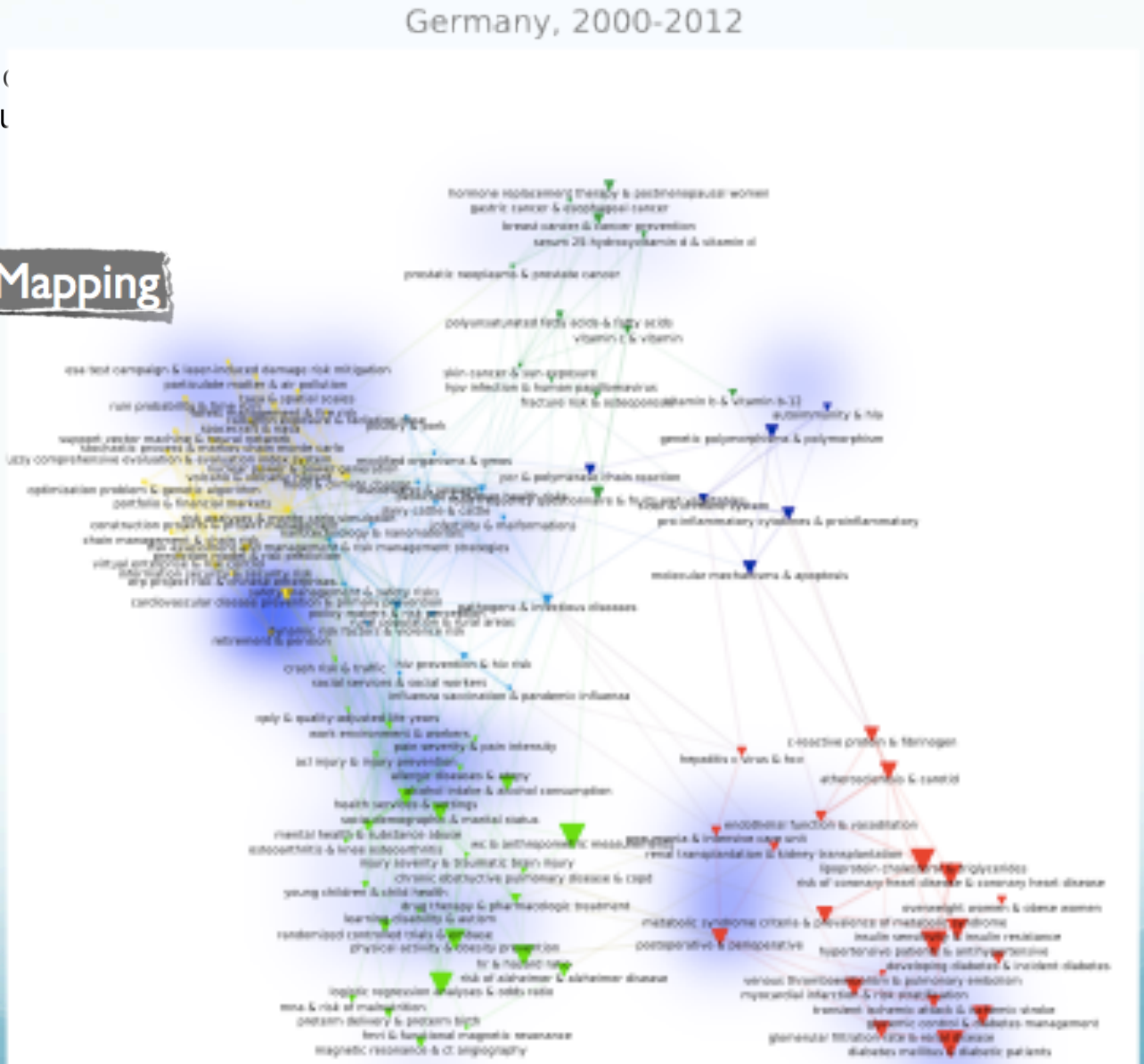


Maps as Spaces

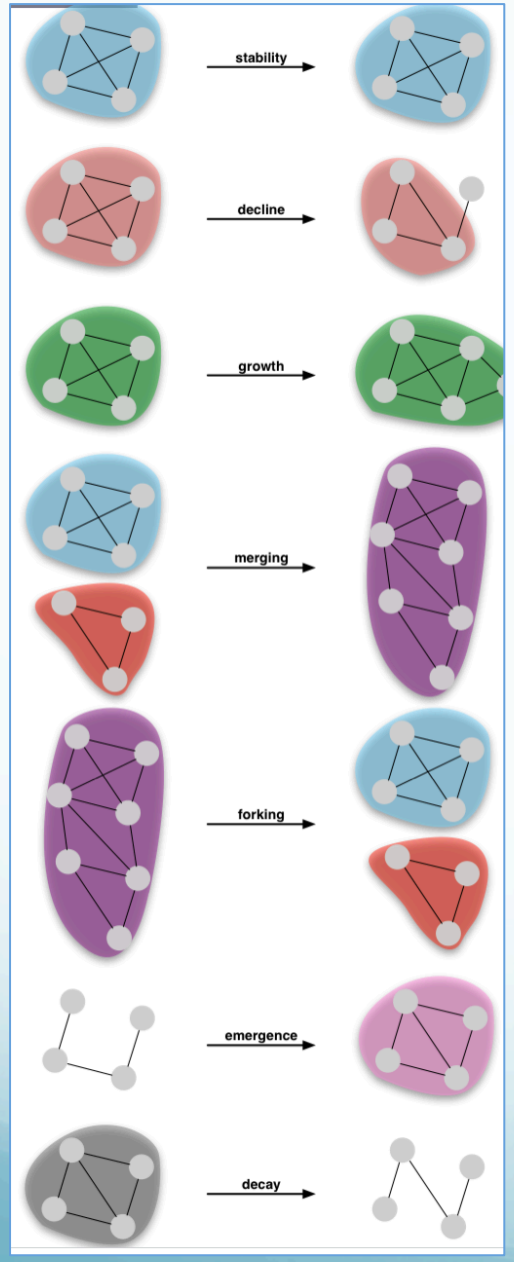
- Maps define as a ba
project heterogeneou

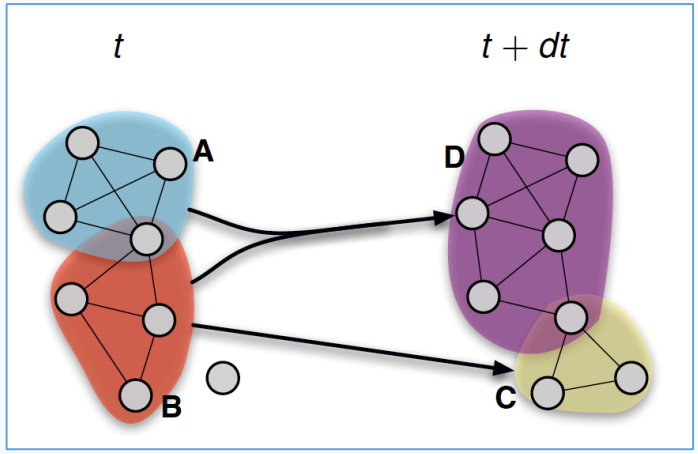
HeatMap

Heterogeneous Mapping

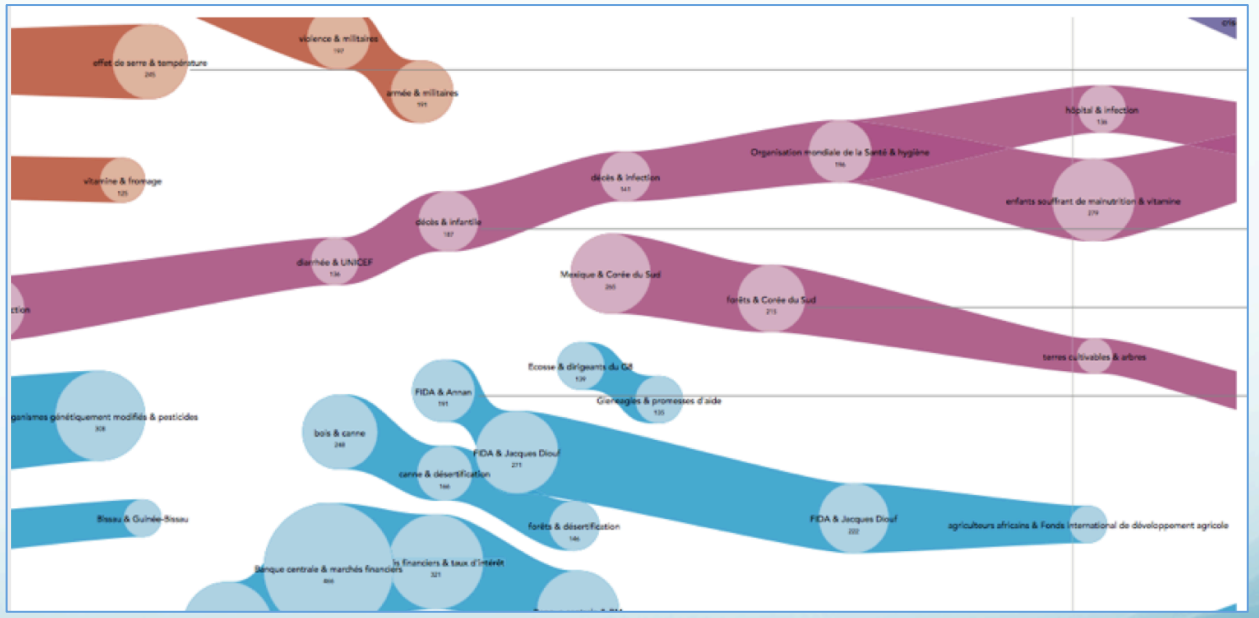


Countries & High-Level





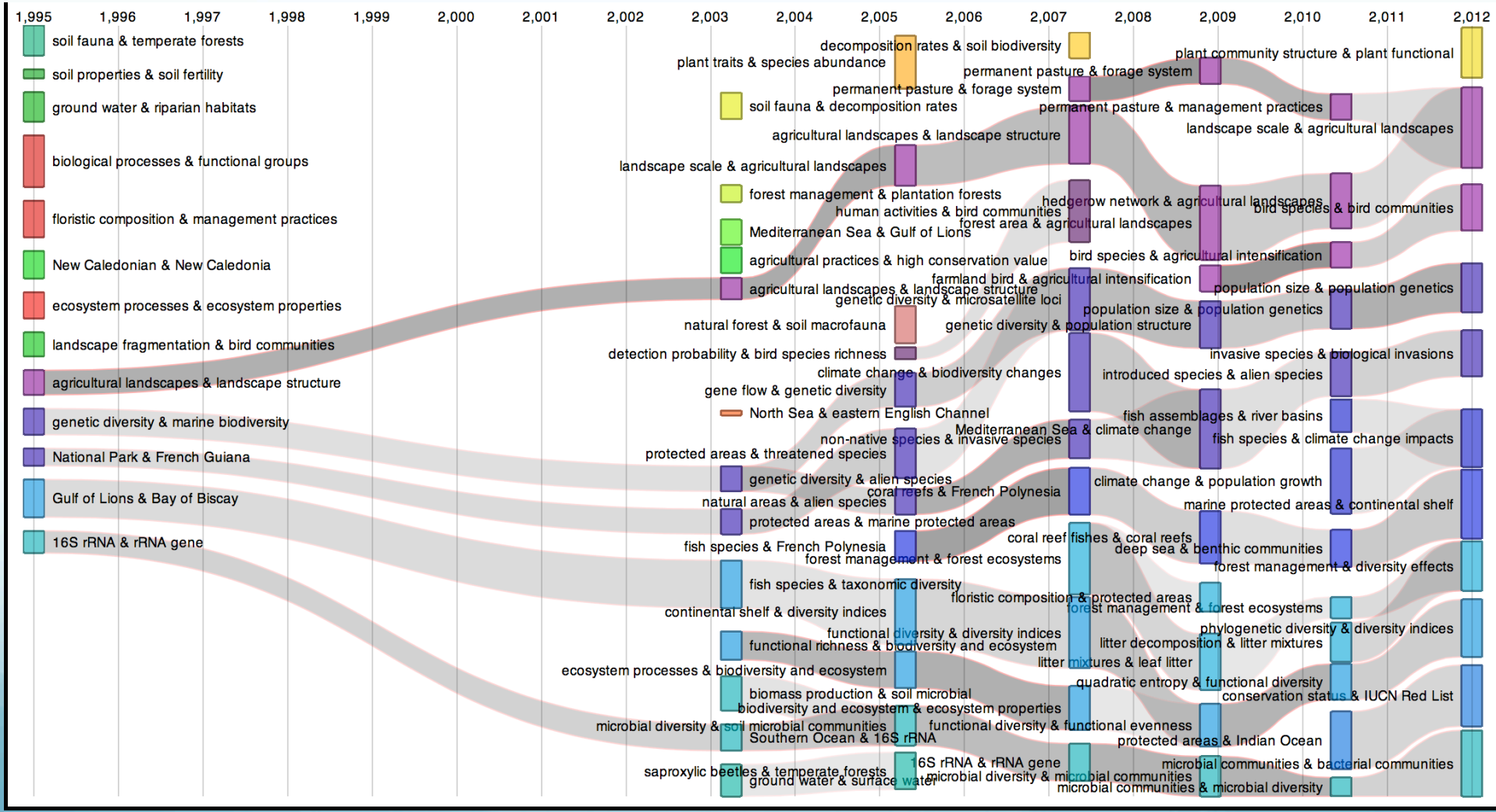
The epistemology of Dynamic Clustering



ti

F(t)

Representing the Dynamic of Calculated Clusters



See the: TRAINING BOOK

Delivered in the Training Repository