CorTexT Platform is for You

BARBIER Marc
UMR LISIS, INRA
Director of CorTexT Platfomr

INRA SCIENCE & IMPACT

CORTEXT

IFRIS Institut Francilien Recherche Innovation Société

LabEx Sciences, Innovations et TEchniques en Société

RISIS Research infrastructure for research and innovation policy studies

The big picture of challenges

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,…)

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

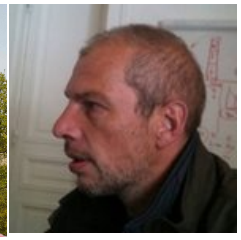
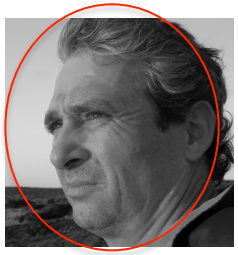
AIMS

- 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

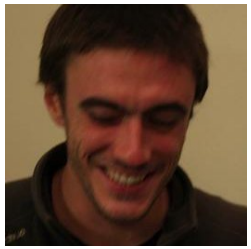
CorTexT Team

CORTEXT

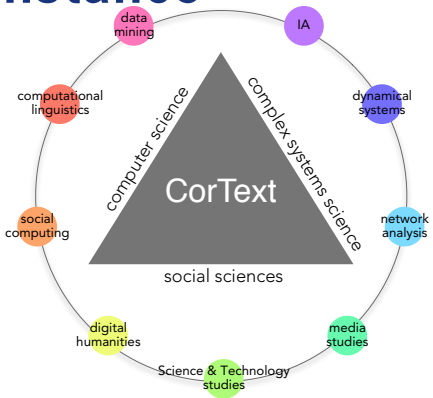
CORTEXT LAB

http://www.cortext.net



Equipe CorTexT Lab

- Barbier Marc (Dir.)
- Anis Arabi
- Breucker Philippe
- Cointet Jean-Philippe
- Duloquin Chloé
- Duong Tam-Kien
- Laurens Patricia
- Martinez Cristian
- Mazières Antoine
- Mogoutov Andreï
- Orsal Guillaum
- Quatrebarbes Constance
- Schoen Antoine
- Villard Lionel



CorTexT Journey

The design and development of a Digital Studies on-line Instrument

"Big-data": also a challenge for Humanities

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 be lasting significance — a clinic 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 in 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.

But science's reliance on digital data extends

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

Ecosystems of Disciplines and Platforms

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

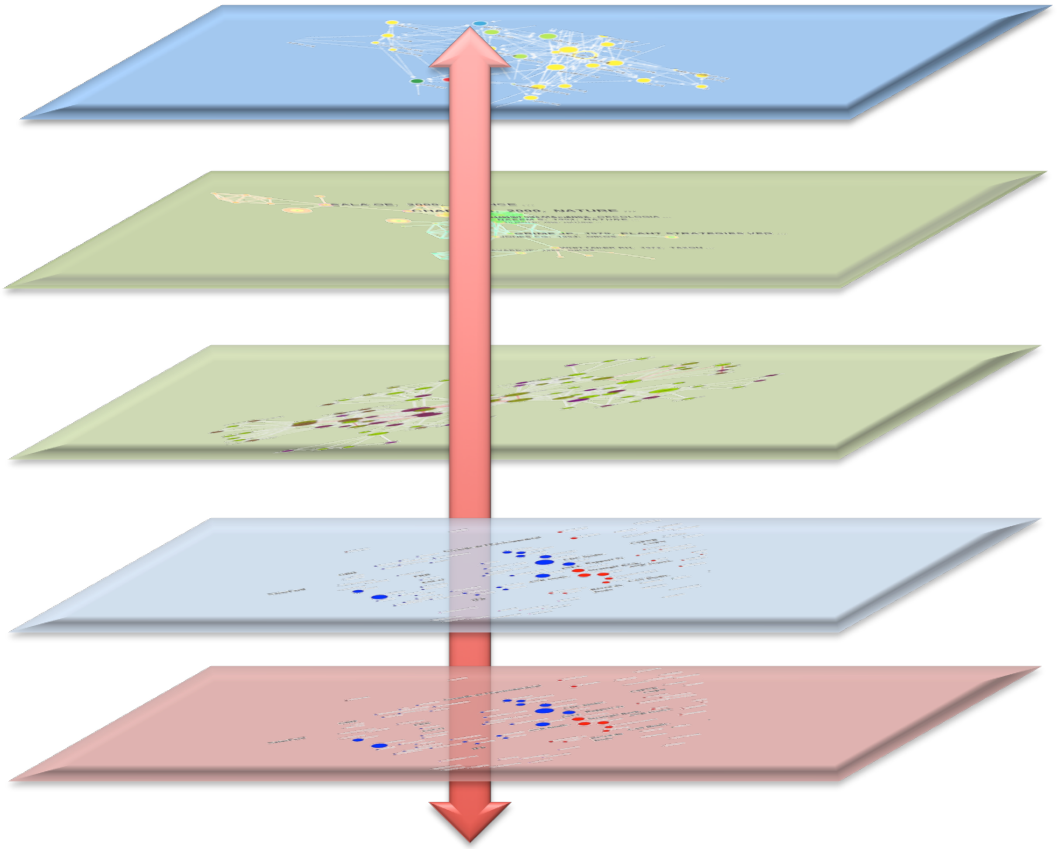
- **Tracking Projects**
- Platforms de Natural Langage Processing
- Plateforms of Science& Technology Mapping
- Digital Humanitis Platforms Plateforme Humanités Digitales

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.

Multi-Levels Heuristic is most needed

Science Dynamics in Layers



Web mapping of instituions => macro view of international institutions

Academic production => identification of main actors, themes and dynamics

Research project => focus on more scientific activity at french scale

Space of experts => where scientists meet policies

Political spaces => Identification of scenes where political discourses are built (Press, Parliament, social-network, blog...)

=> depending on existence and access to relevant datas, existing tools & our tool development capacities

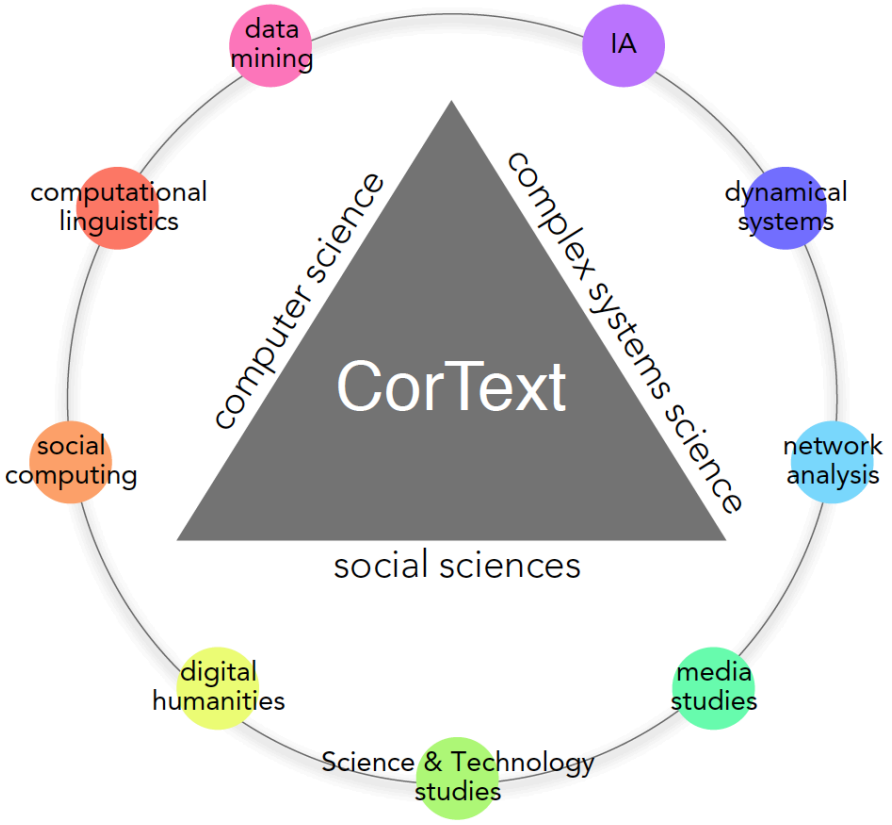
From Akera, SSS, 2009)

CorText Lab

- A digital platform for social scientists conducting *empirical studies* on complex heterogeneous textual corpora in the fields of Science and Technology Studies, Media Studies, political sciences, etc.

- built upon competencies, skills and also research in Computer Science, Complex System Analysis, Natural Language Processing.

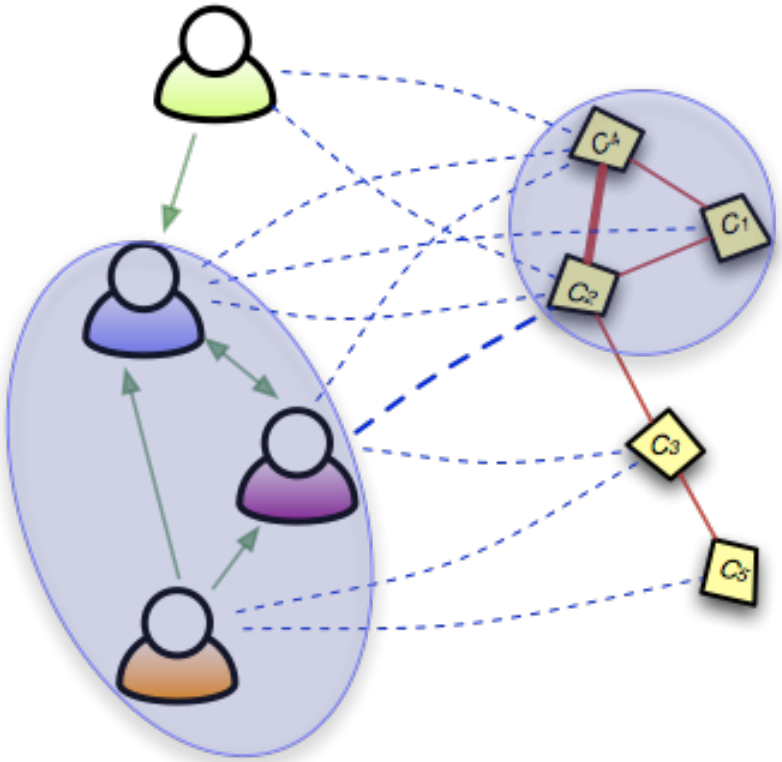
- online and open: http://manager.cortext.net



From Co-Word Analysis to Community Detection and beyond

- Generalized co-occurrences analysis framework mixing people, terms, countries, etc...
- Clustering techniques are being used to circulate from micro to macro

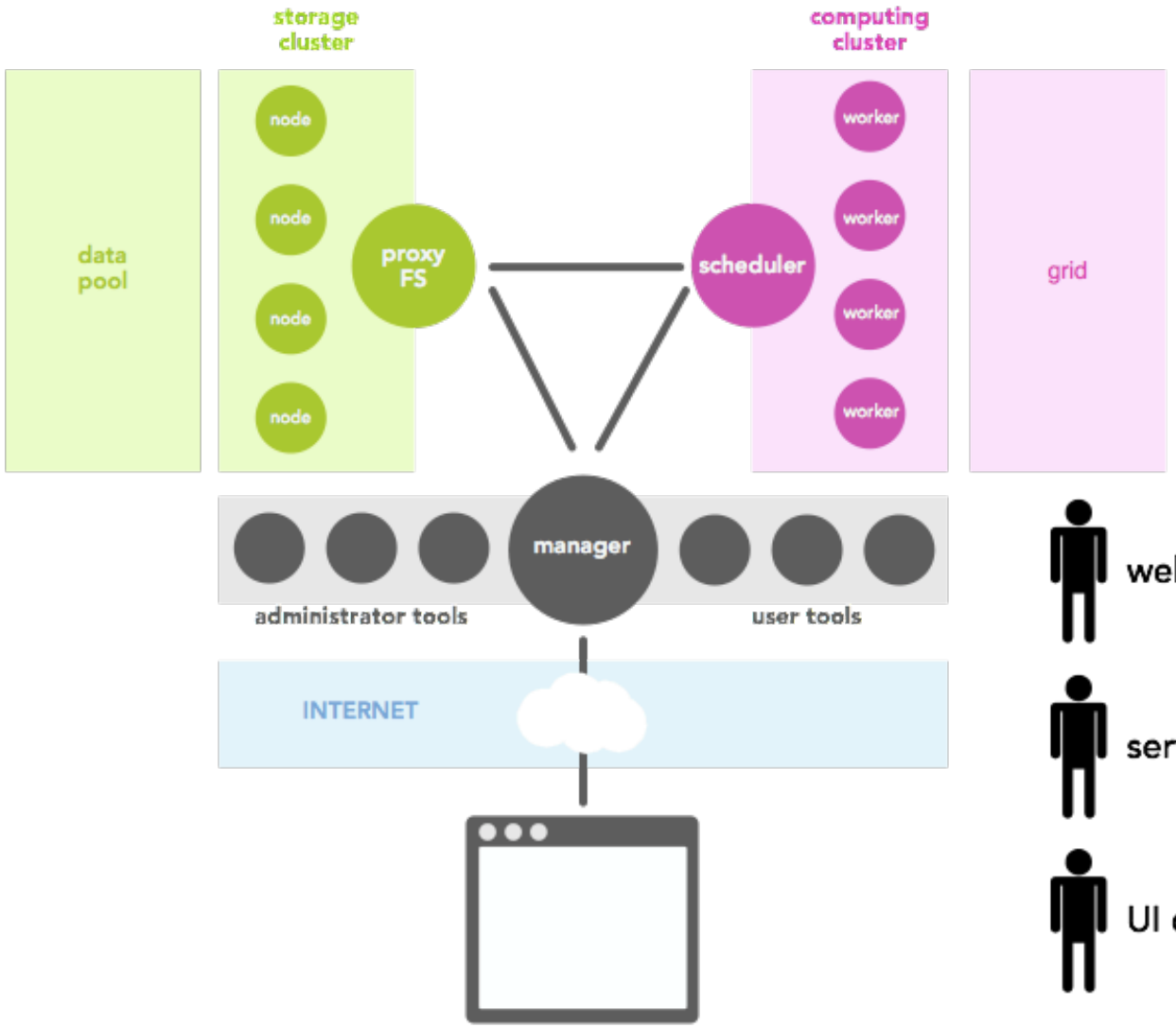
- Observing social dynamics *in-vivo*
- Heterogeneous networks at different levels
- Social dynamics at different time scales



Socio-semantic network and related clusters

Cointet, Jean-Philippe (2009). *Dynamiques sociales et sémantiques dans les communautés de savoirs Morphogenèse et diffusion*, Thèse Ecole Polytechnique Paris.

BACK OFFICE
Competencies and Organization



system engineer

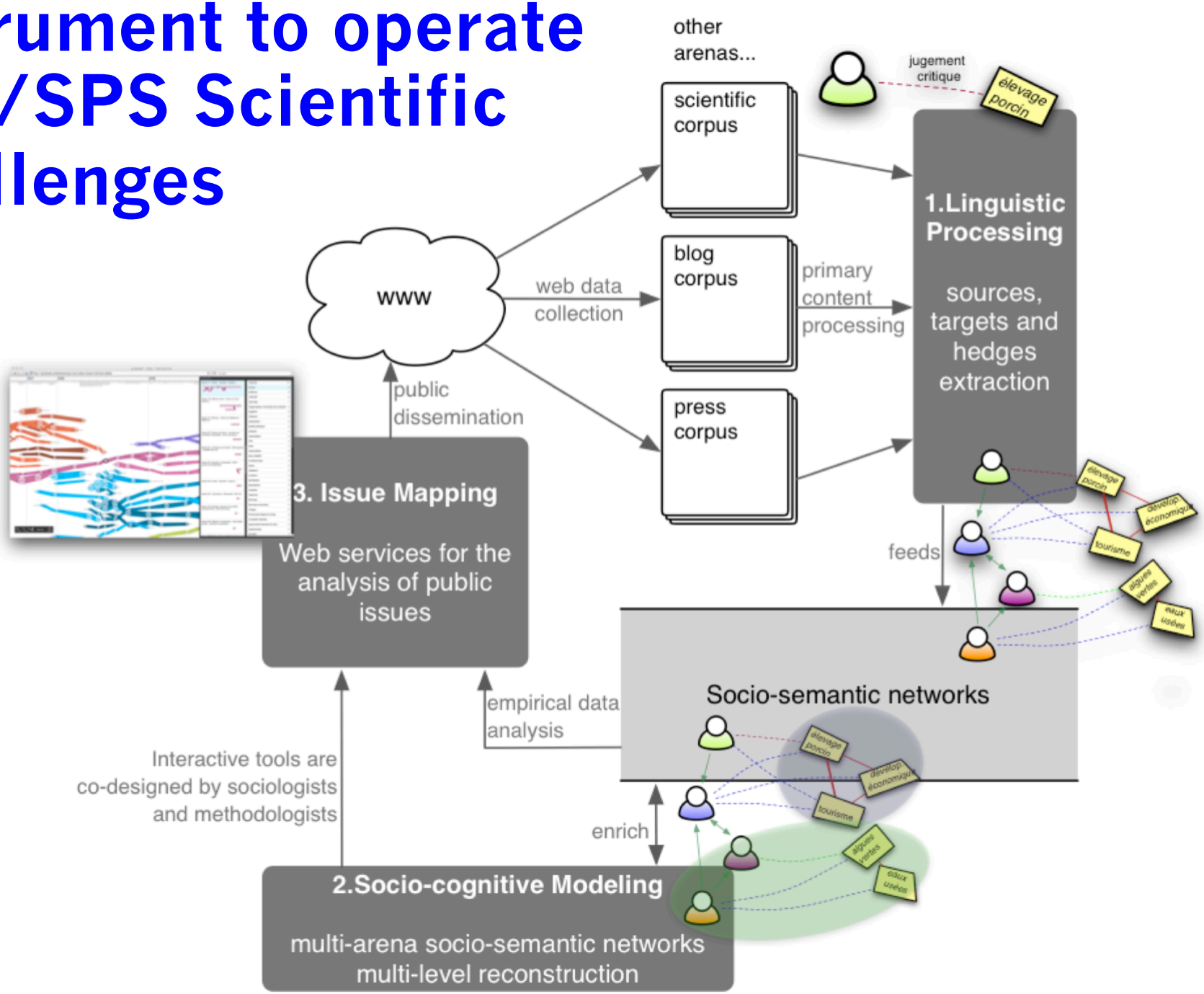
web devloper

methodologist

service design

UI design

Instrument to operate STS/SPS Scientific Challenges

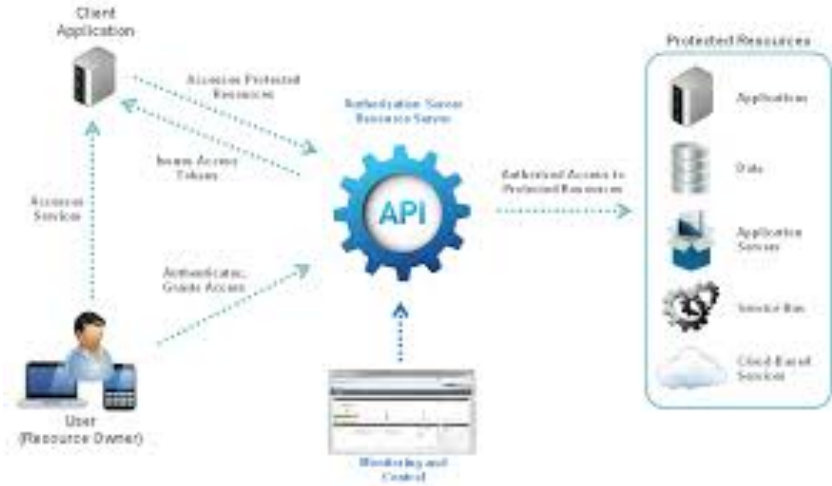


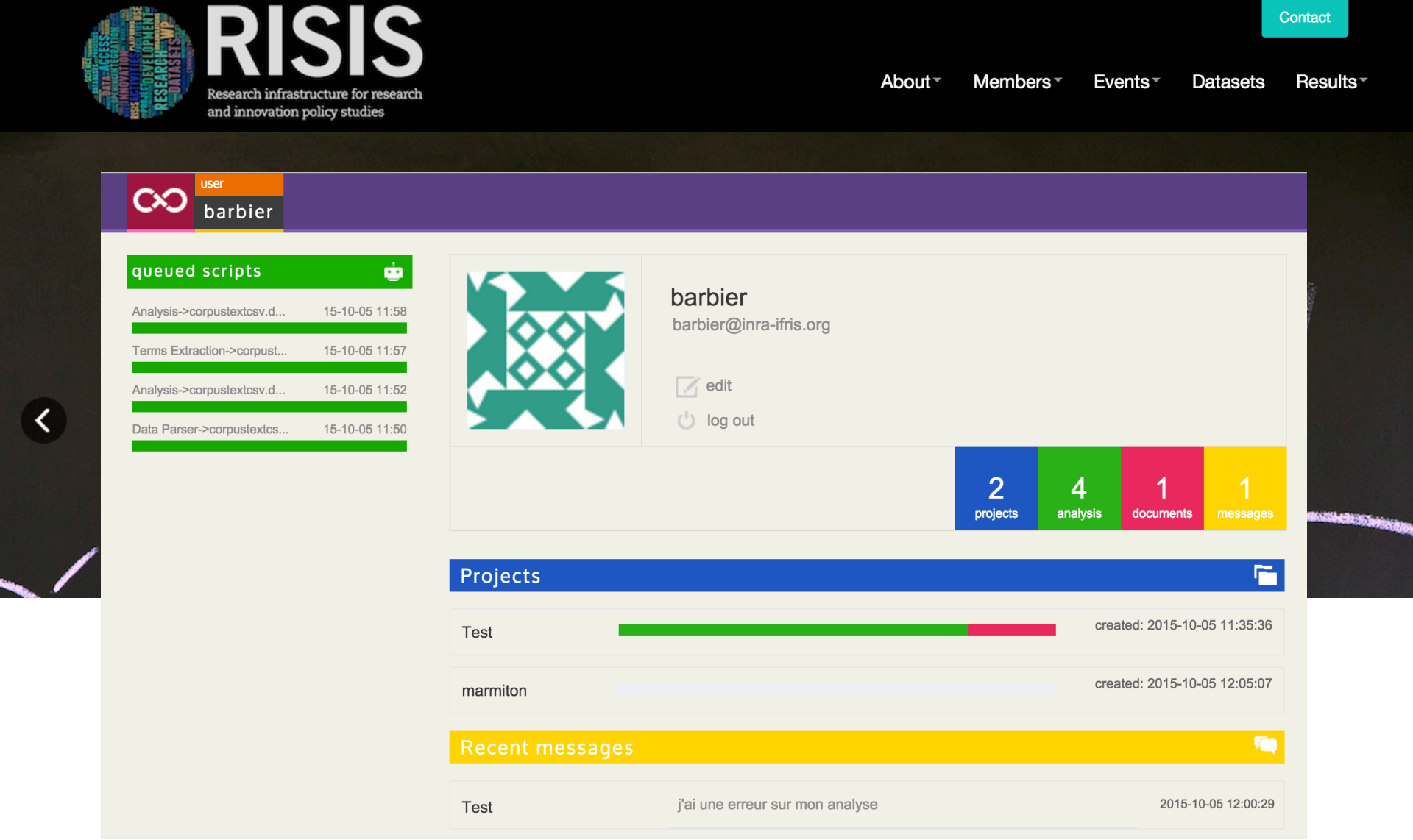
FRONT OFFICE
User – Services and Data

Users

Datasets

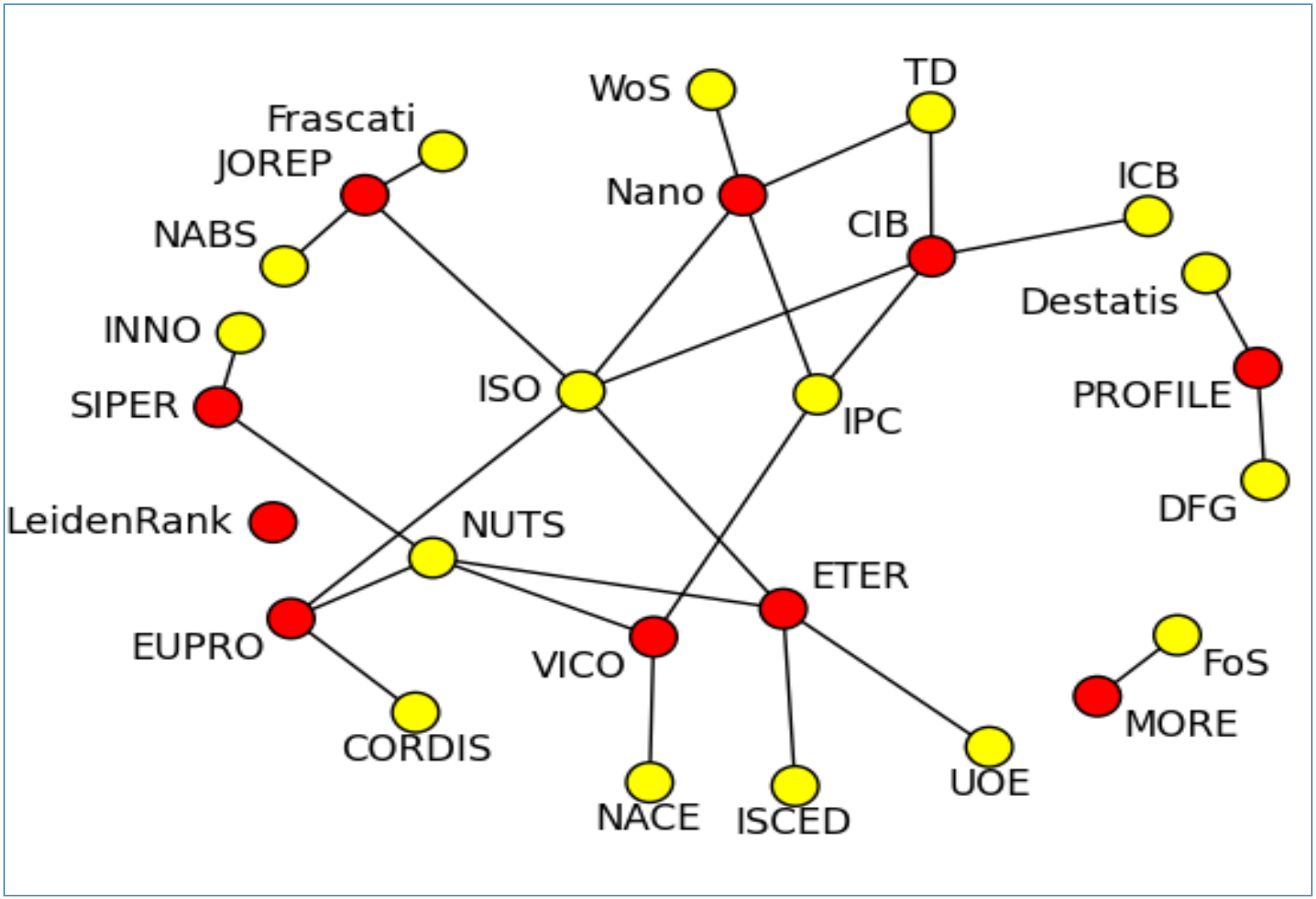
Services





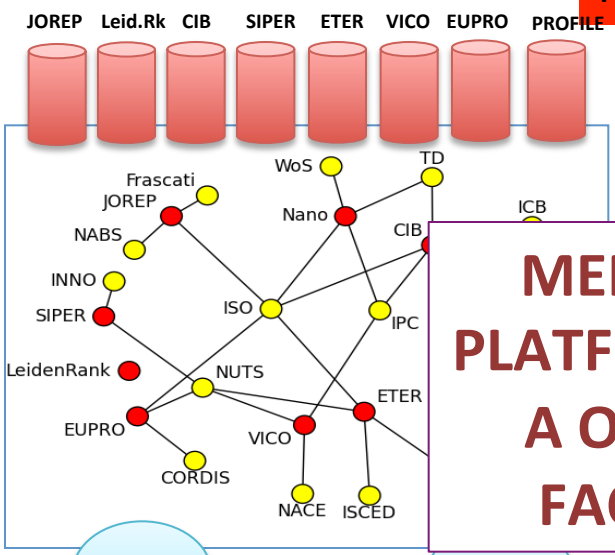
CorText.Risis Facility

A NETWORK OF CAPACITIES TO BE TRANSFORMED IN A EUROPEAN COMMON



Infrastructure Interoperability : an IT Challenge

Global Interoperability Issues



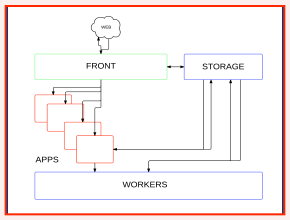
CorTexT

SMS

MERGING PLATFORMS IN A ONLINE FACILITY

Joint Activities for Opening

Datasets Platform



Dedicated analytical resources in DB Platform

Local Scientific CoP

Materiality of Datasets

Cognitive-Silico capacity

Socio-Cognitive capacity

Access & *Protocols*

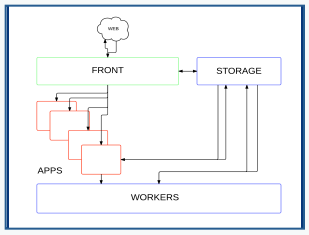
Format & IT *Development*

Meaning & *Training*

Materiality of Analytics

Cognitive-Silico capacity

Socio-Cognitive capacity



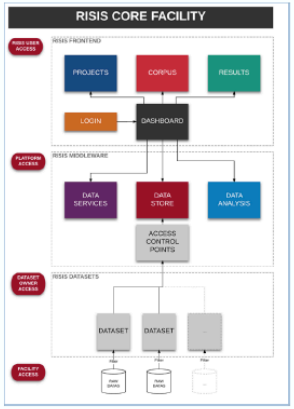
Common analytical resources in Data Sciences PF (Linked Data API, Triple Store, Scripts, Algo., Viz.)

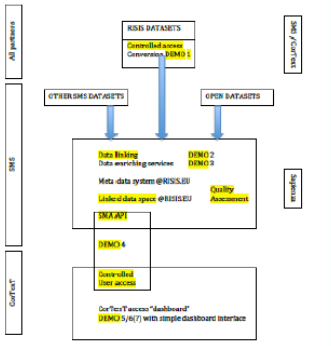
Global Scientific RISIS Domain

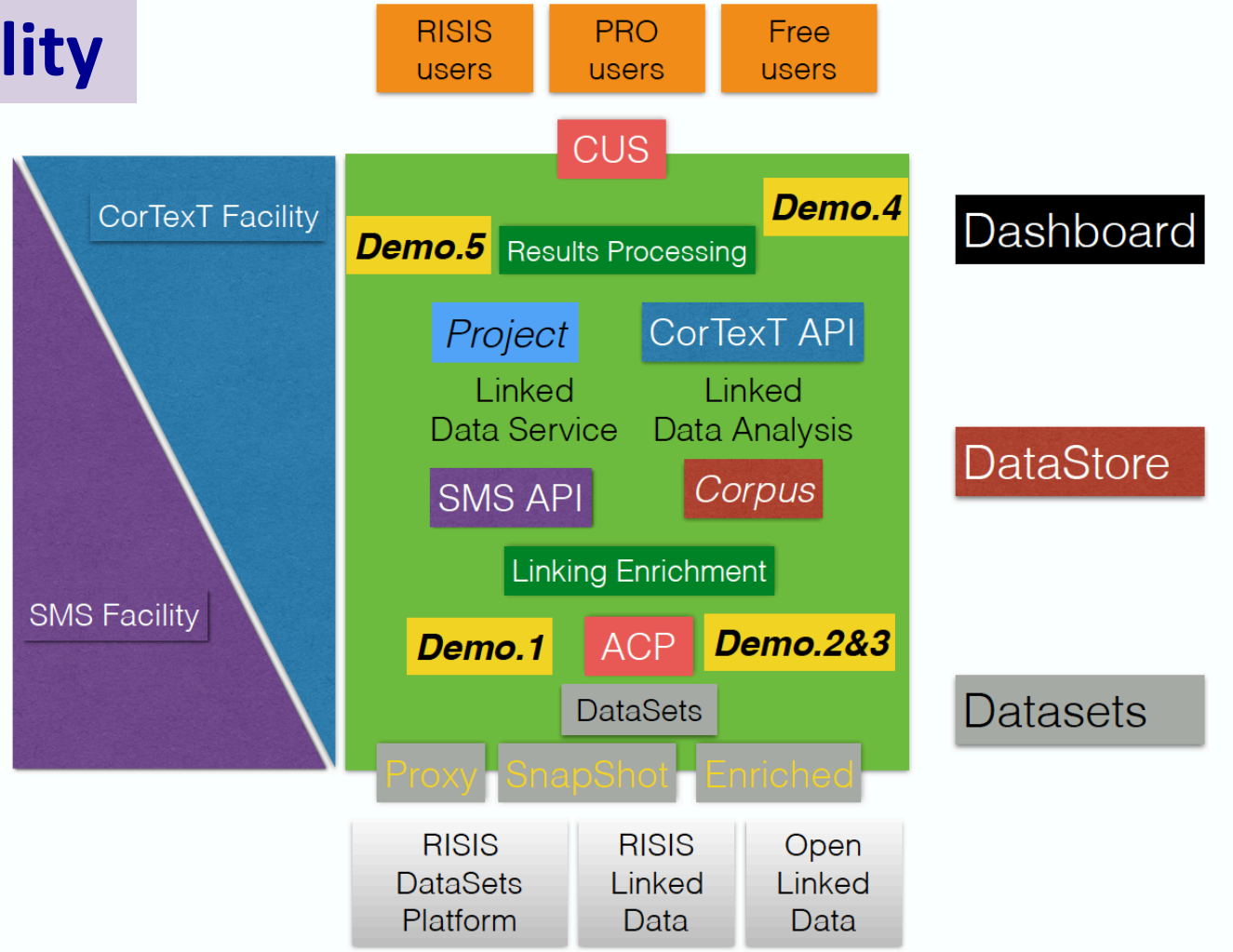
Interoperability and Analytics

RISIS: An opportunity for Interoperability challenges

RISIS Core Facility





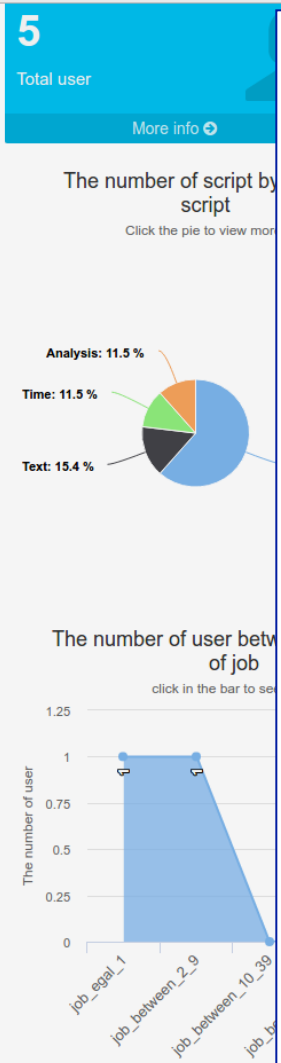


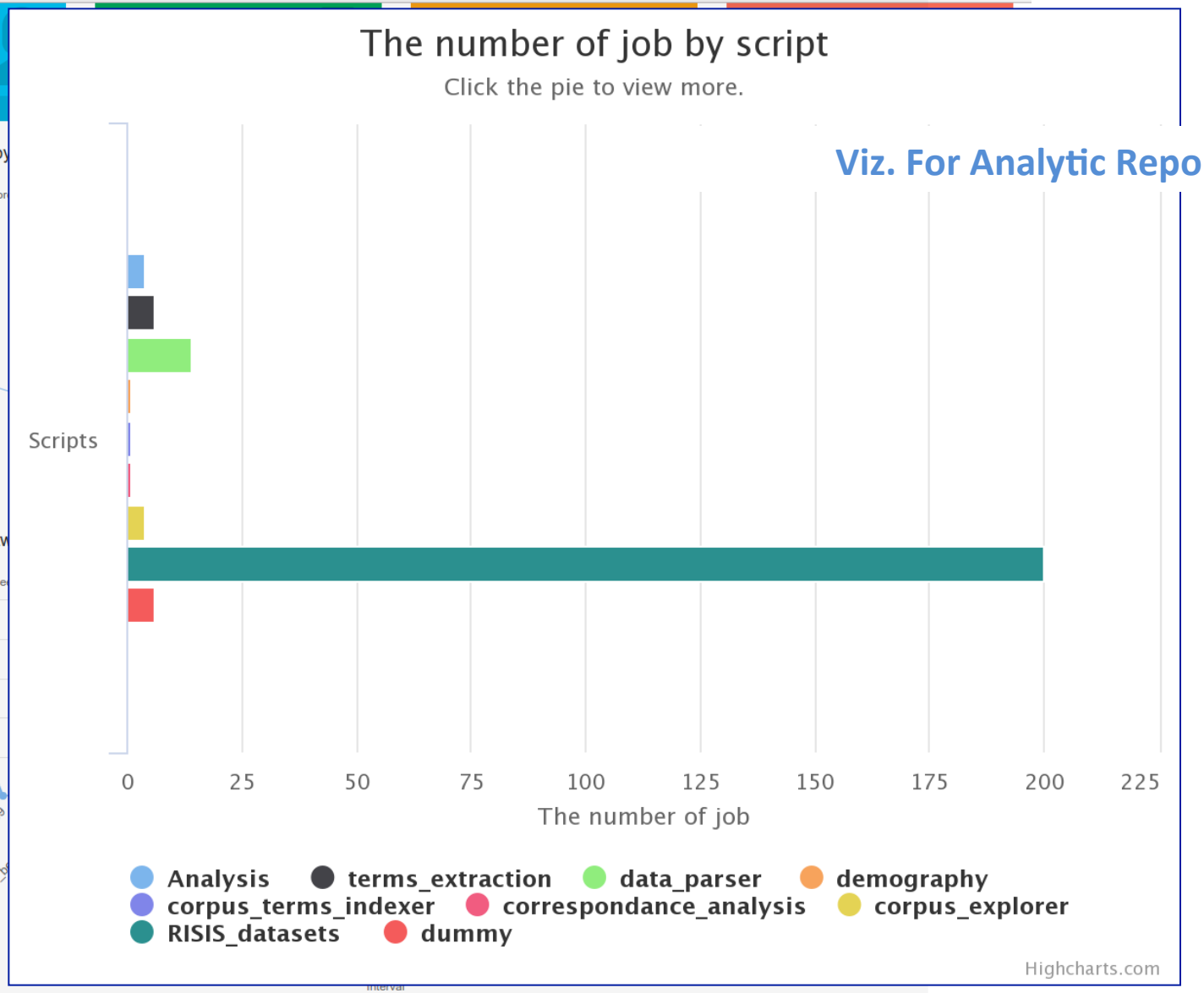
Today

Tomorrow

Towards the *in silico* monitoring of Uses

Monitoring DashBoard



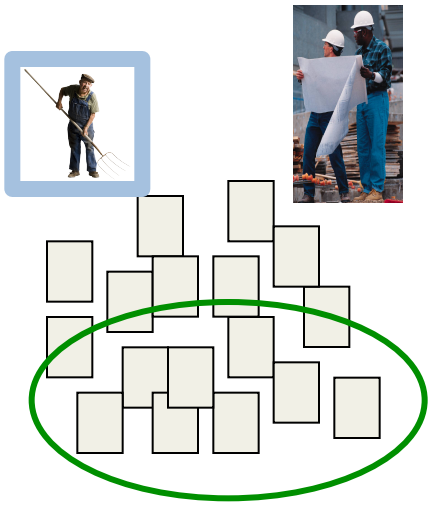


Viz. For Analytic Report

CorTexT Added-Value

Empowering Science Policy Studies and Innovation Studies by the means of data science and Knowledge vizualization

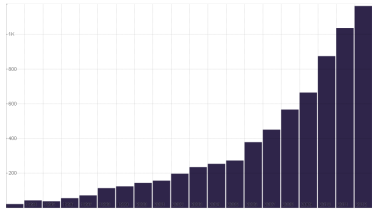
Existing Practices of Datasets and Corpus Processing



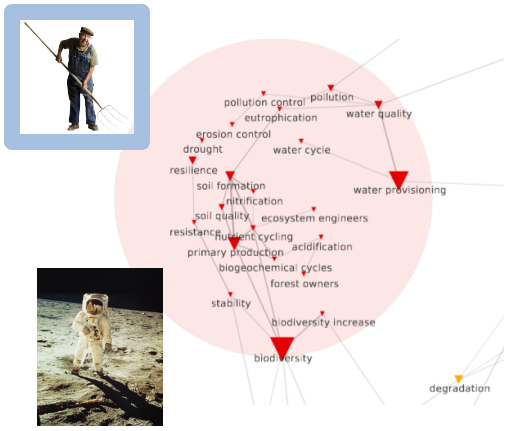
Extraction / Building of a corpus

Selection / Creation of Field descriptors

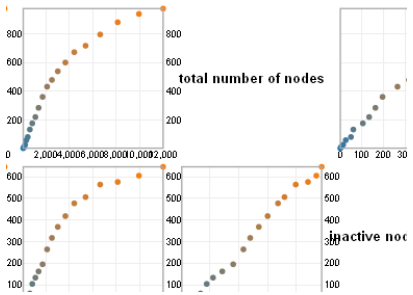
Statistics of Descriptors



Network Analysis and Knowledge Vizualization



Temporal and Structural Analysis and modelisation



Direct

Keywords, Date, Institu., Codes, etc.

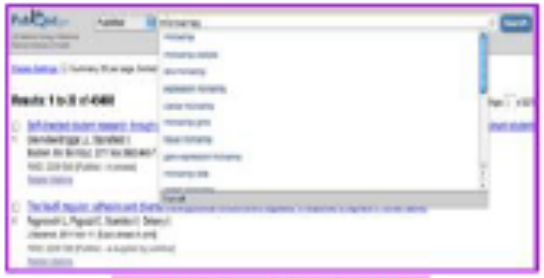
Indirect

Lexical Extraction from Natural Langage Processing

scientific productions

Web Of Science ISI

Microsoft Academic Search



Medline Pubmed

specific databases

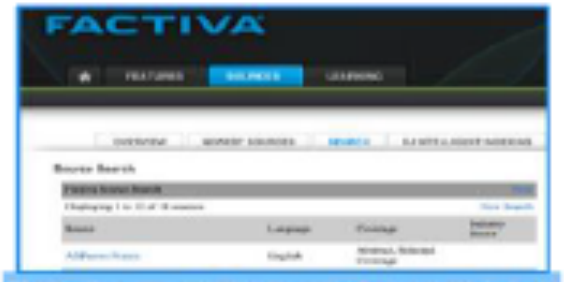
rare disease database

projects database

clinical trials database

media productions (press+web)

web crawler

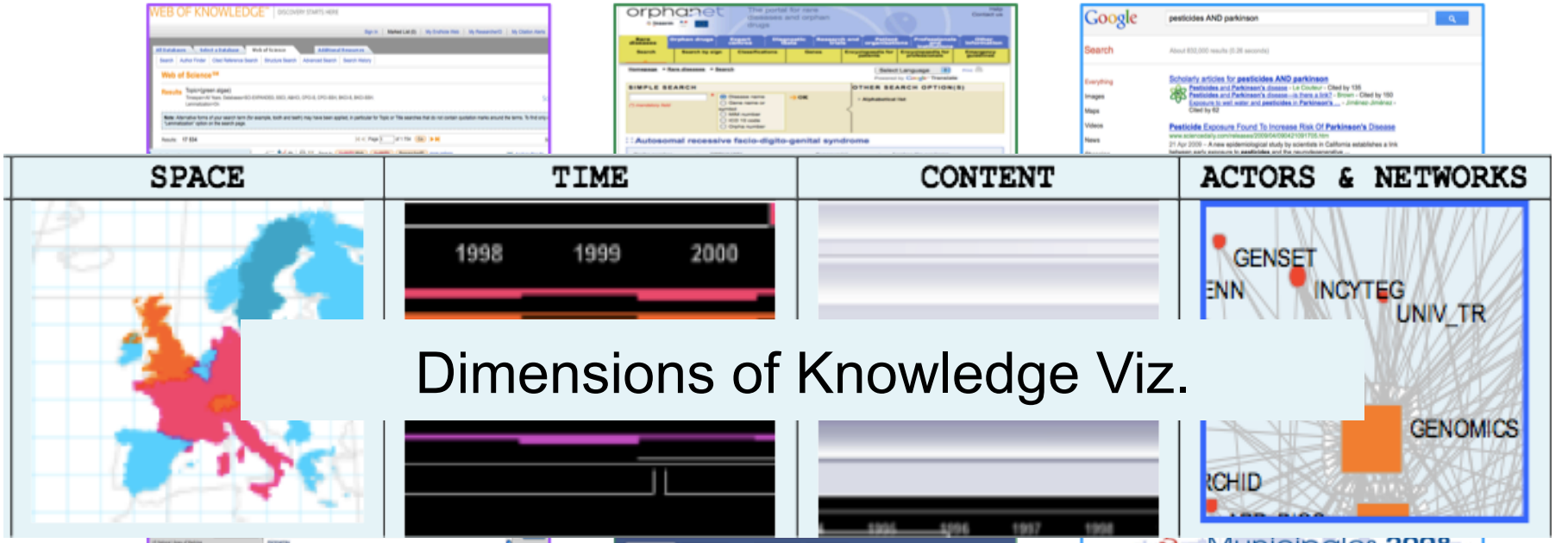


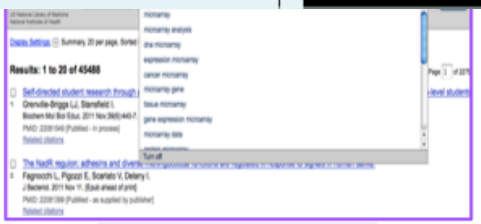
Factiva, press articles archive

online forums

What about vizualization in CorTexT?

scientific productions ***specific databases*** ***Media-web productions***





Medline Pubmed

clinical trials database

online forums

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*)

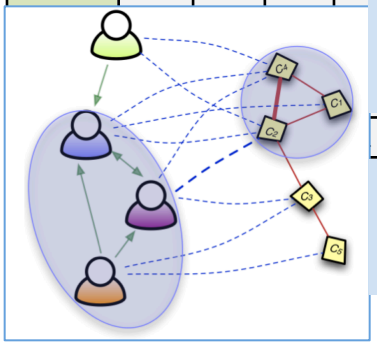
stem	main form	forms	n	C-value	Specificity	Frequency
alga red	red algae	red algae|&|RED ALGAE|&|Red algae|&|red alga	2	703,3	1292,3	464
matter organ	organic matter	organic matter|&|Organic matter				
chlamydomona reinhardtii	Chlamydomonas reinhardtii	Chlamydomonas reinhardtii				
higher plant	higher 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 :

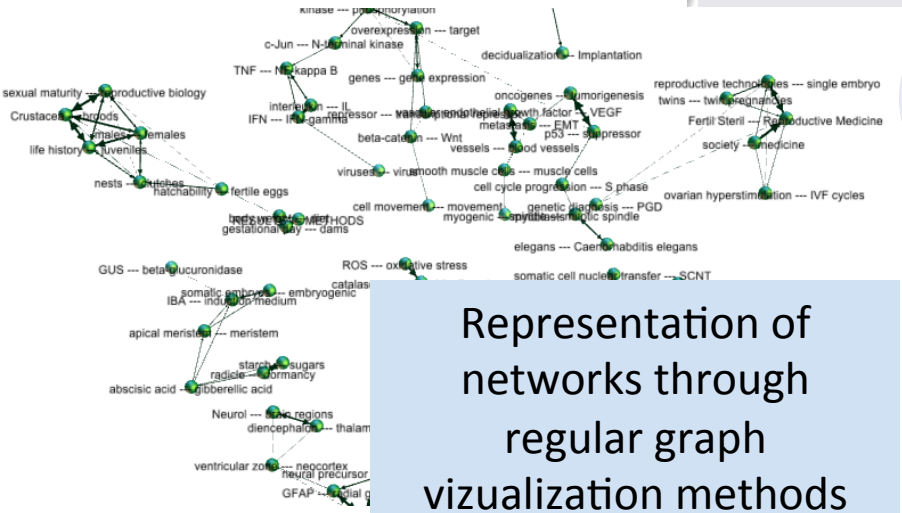
- *Raw cooccurrences : $S_R(i,j) = c_{ij}$*
- *χ^2 score : $S_{\chi^2}(i,j) = \frac{c_{ij} - e_{ij}}{\sqrt{e_{ij}}}$*
- *Mutual Information : $S_{MI}(i,j) = \log(\frac{c_{ij}}{e_{ij}})$*

Indirect Measures of Similarity :

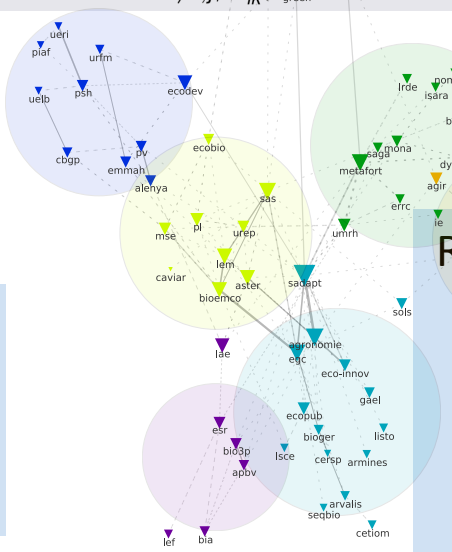
- Mutual Information (distributional) :

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

Choosing the metrics of similarity



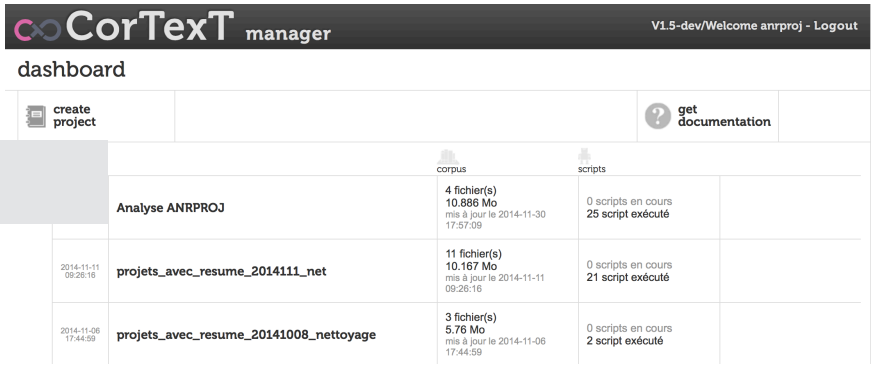
Representation of networks through regular graph vizualization methods



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

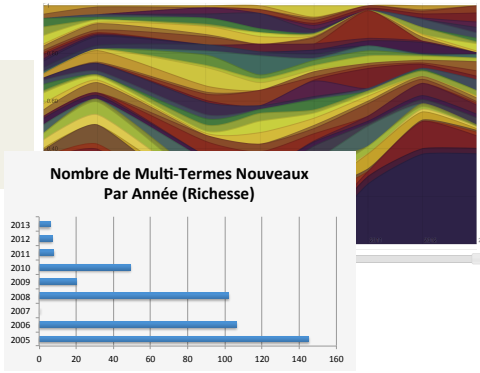
(Louvain Algo. Louvain, method of partition optimization)

From Datasets to Analytics: embedding the platform in contexts of use



Parsing

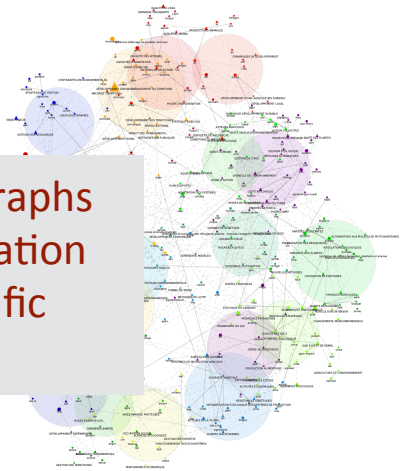
Demography of entities



Terminological Extraction of N-Grams

stem	main form	forms
socio-écologiques système	systèmes socio-écologiques	systèmes socio-écologiques
production végétal	production végétale	production végétale|&|productions végétales
culture diversité système	diversité des systèmes de culture	diversité des systèmes de culture|&|diversité de systèmes de culture
adaptation système	adaptation des systèmes	adaptation des systèmes|&|adaptation aux systèmes
gaz émission	émissions de gaz	émissions de gaz|&|émission de gaz|&|émissions de ces gaz
compromis recherche	recherche de compromis	recherche de compromis|&|recherche d'un compromis
croissance plante	croissance de la plante	croissance de la plante|&|croissance des plantes|&|plantes et la croissance

Selection of relevant N-Grams within Groups (mixed Scientists and Experts)



Analysis of Graphs and Clusterization within Scientific Groups

Presentation and discussion of knowledge extracted from Datasets

RISIS
Research infrastructure for research and innovation policy studies

"DEMOGRAPHY" OF ITEMS AND DISTRIBUTION

Structural analysis of scientific corpus

- Trade-off between open frontiers and the consolidation of a chore community in synthetic biology field
- Cities ISI Field of “Biodiversity” Corpus

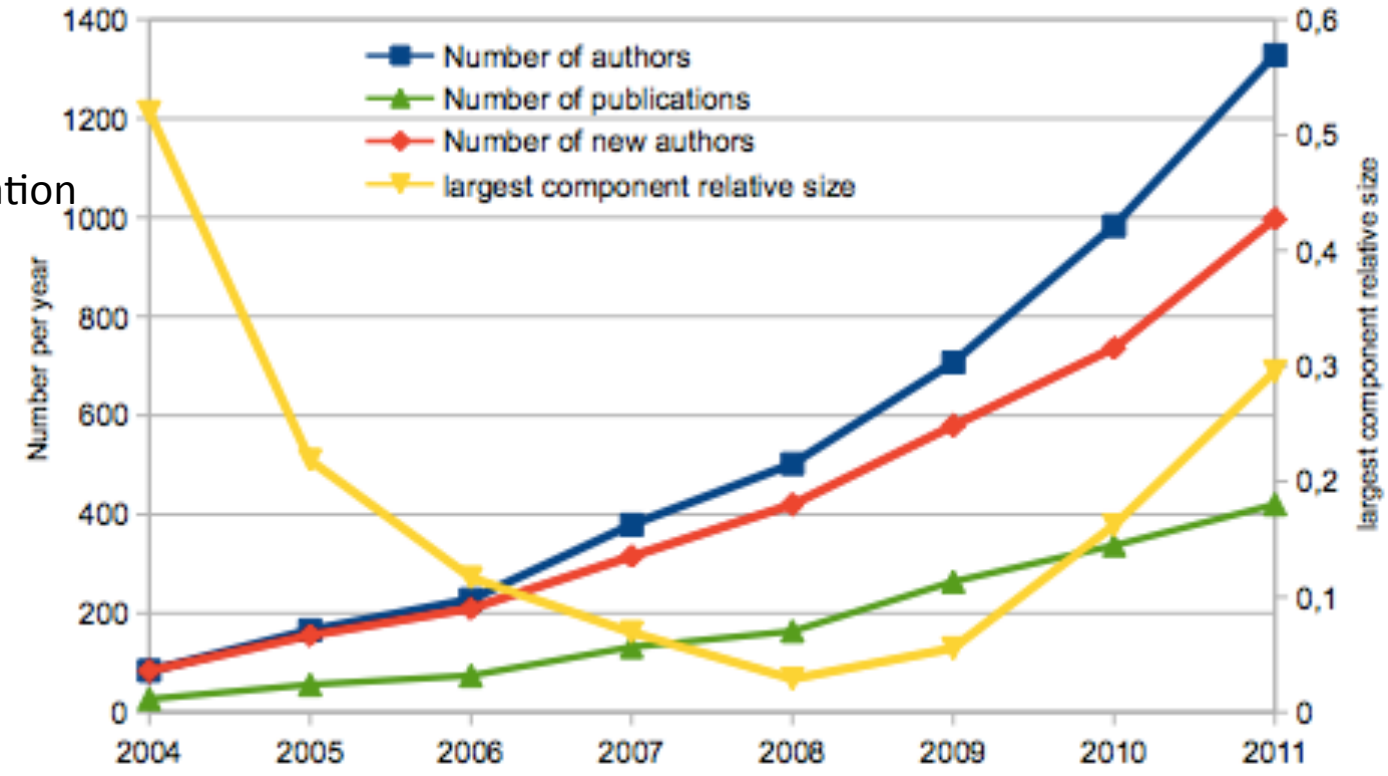
Structural Analysis

Collaboration Network Evolution

Structural analysis of corpus: trends and trade-off

- Trade-off between open frontiers and the consolidation of a chore community in synthetic biology field

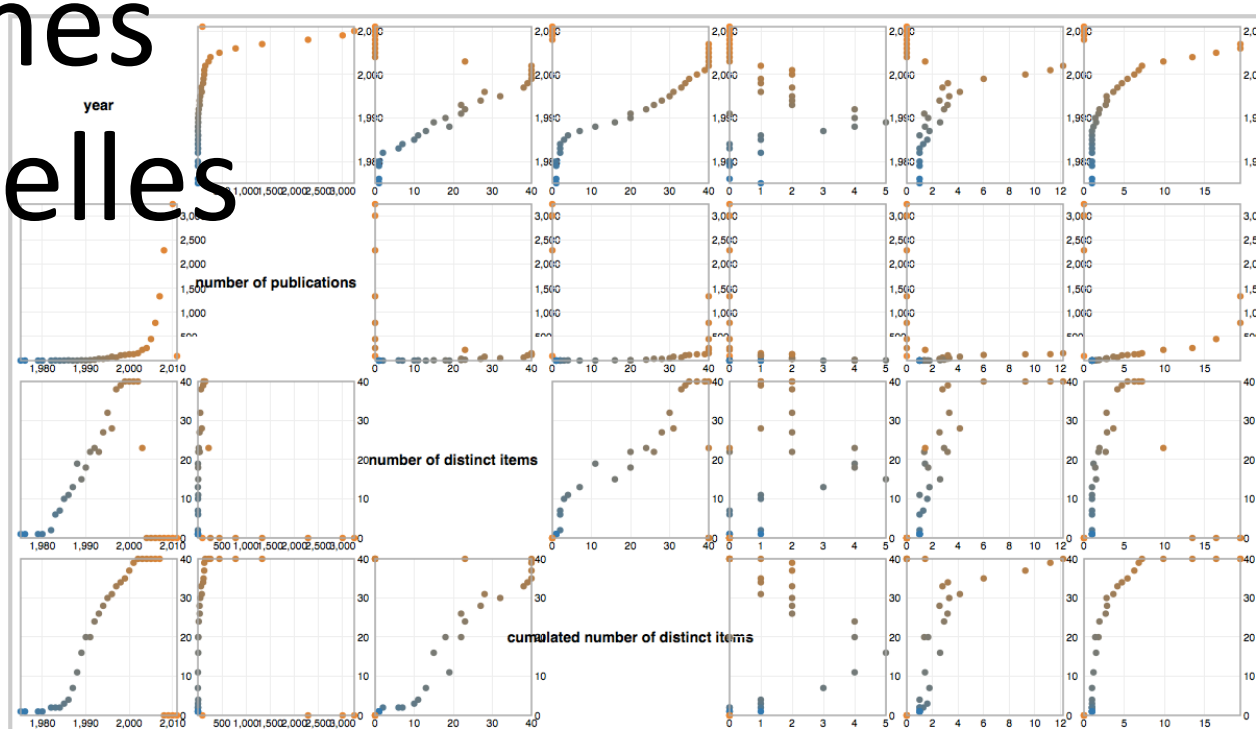
Structural Analysis



Collaboration Network Evolution

Approches fréquentielles

Emphasize on dynamics, circulation and transformation of knowledge production



Topological Phase Transition (collaboration network)

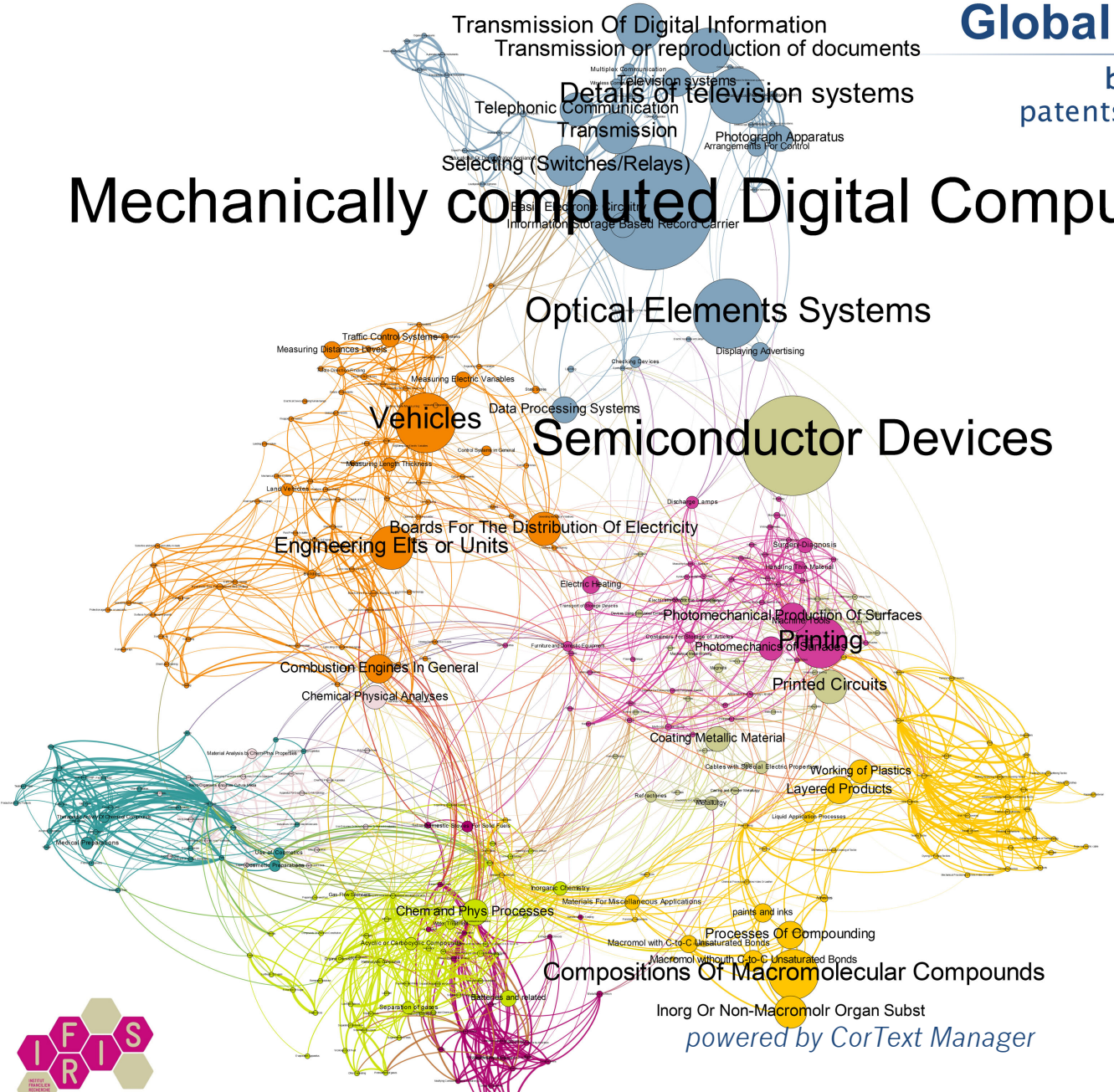
RISIS
Research infrastructure for research and innovation policy studies

LANSCAPE OF N-GRAMS

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

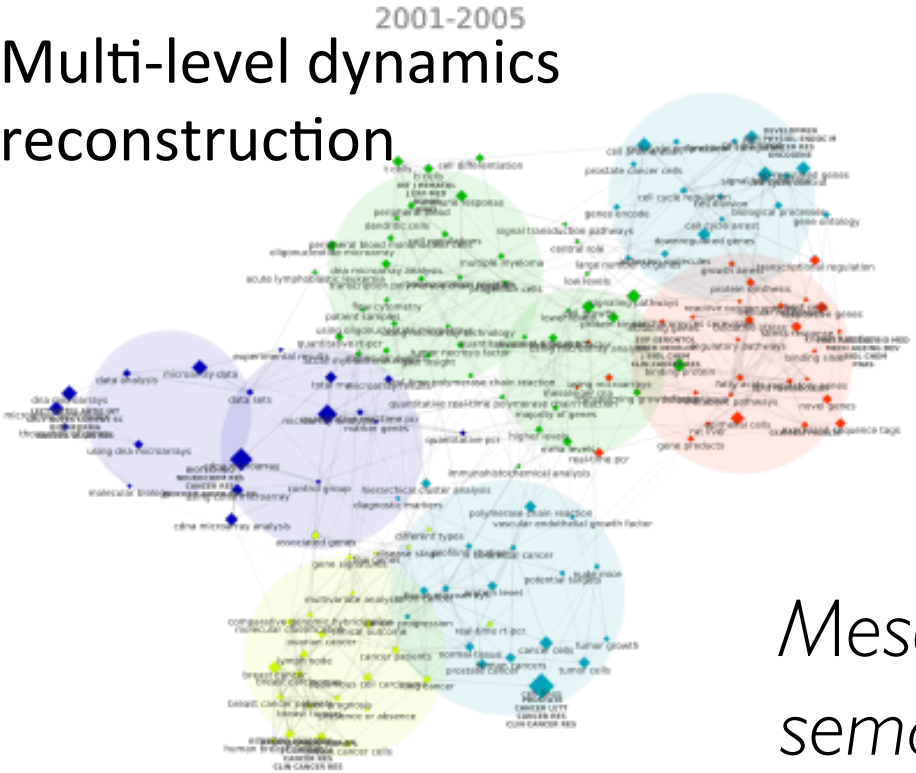
Vehicles
Compositions Of Macromolecular Compounds
Chem and Phys Processes
Printing
Mechanically computed Digital Computers
Semiconductor Devices
Medical Preparations
Domestic Stoves For Solid Fuels
Chemical Physical Analyses

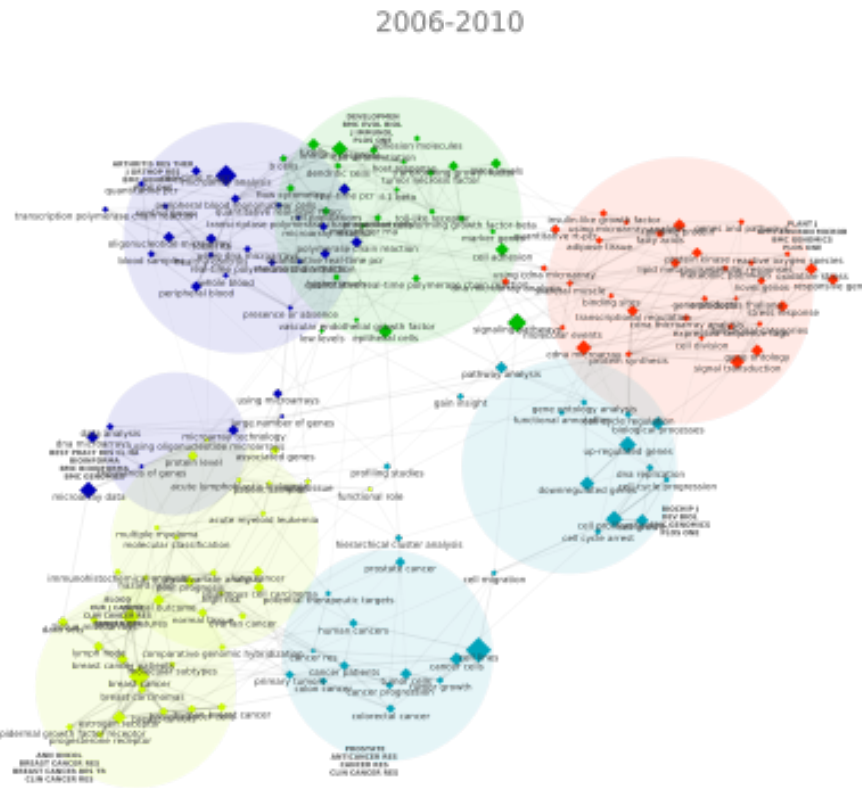
powered by CorText Manager

Evolution of lexical fields

Emphasize on dynamics, circulation and transformation of knowledge production

Multi-level dynamics reconstruction





Meso-dynamics: Evolution of semantic fields between 2 successive periods

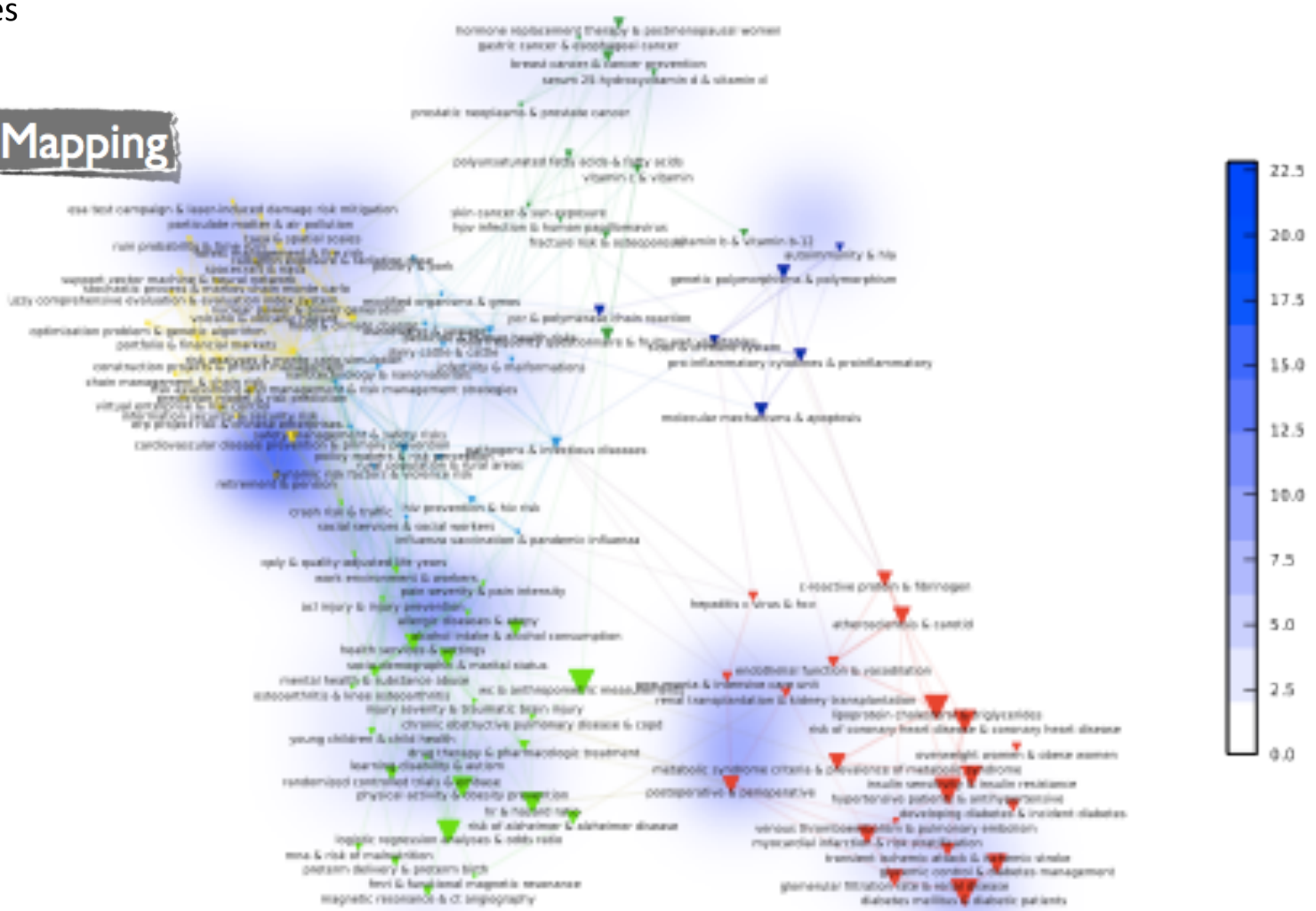
Maps as Spaces

Germany, 2000-2012

- Maps define as a background space which can later serve to project heterogeneous entities

HeatMap

Heterogeneous Mapping

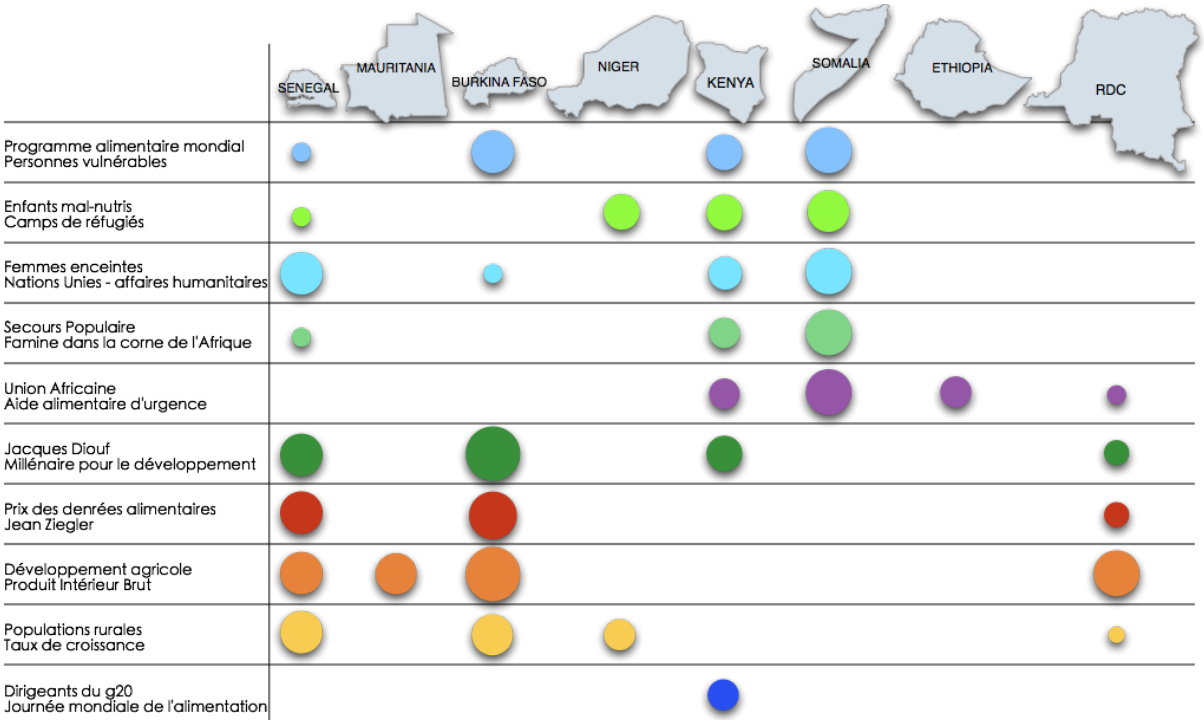


Countries & High-Level Cognitive Map

heterogeneous multi-level networks

Generalized co-occurrences analysis framework mixing people, terms, countries, etc...

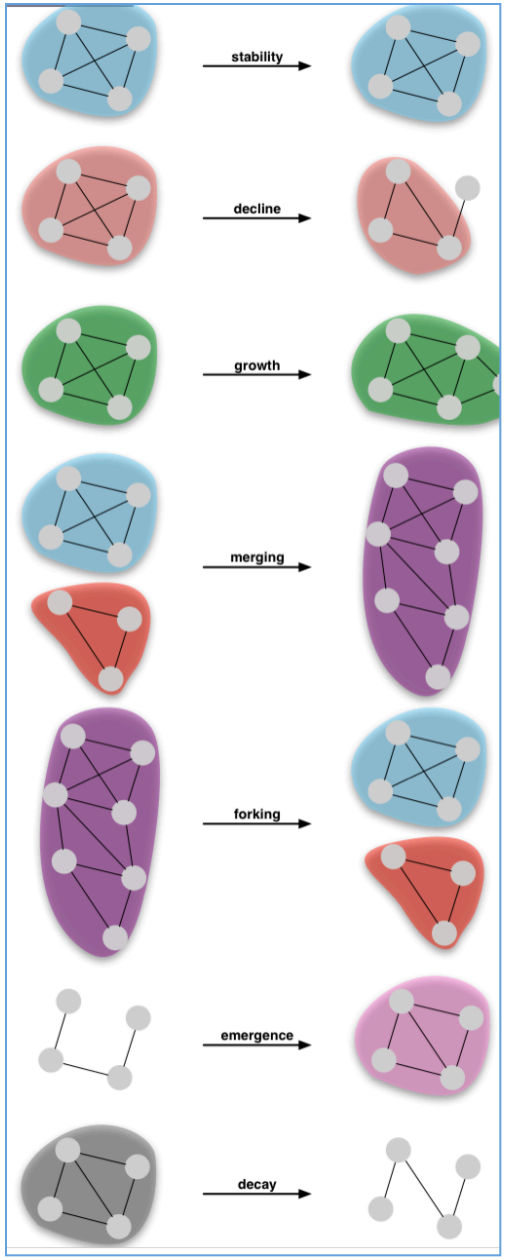
clustering techniques are being used to circulate from single piece of texts to clusters

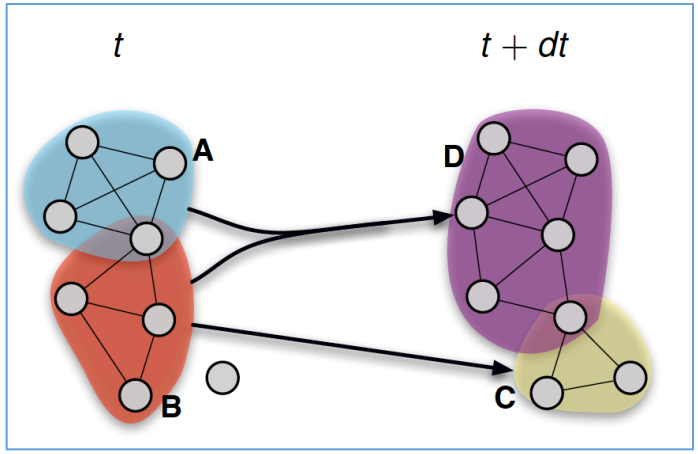


Intensity of issues in different countries (Food Security project)

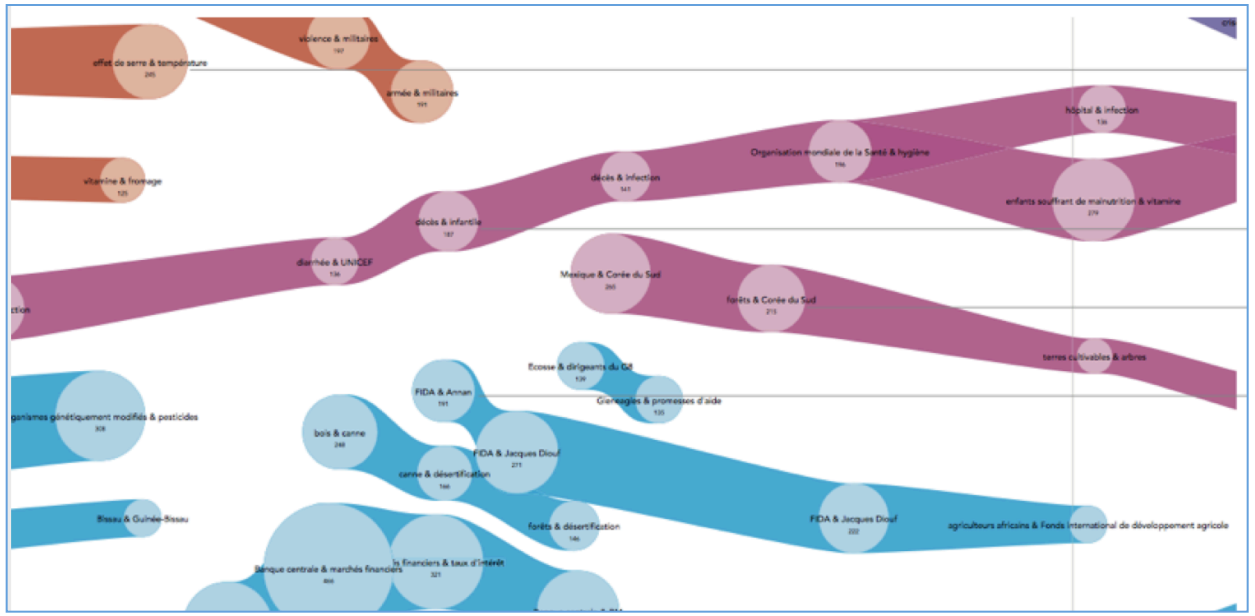
RISIS
Research infrastructure for research and innovation policy studies

DYNAMIC CLUSTERING





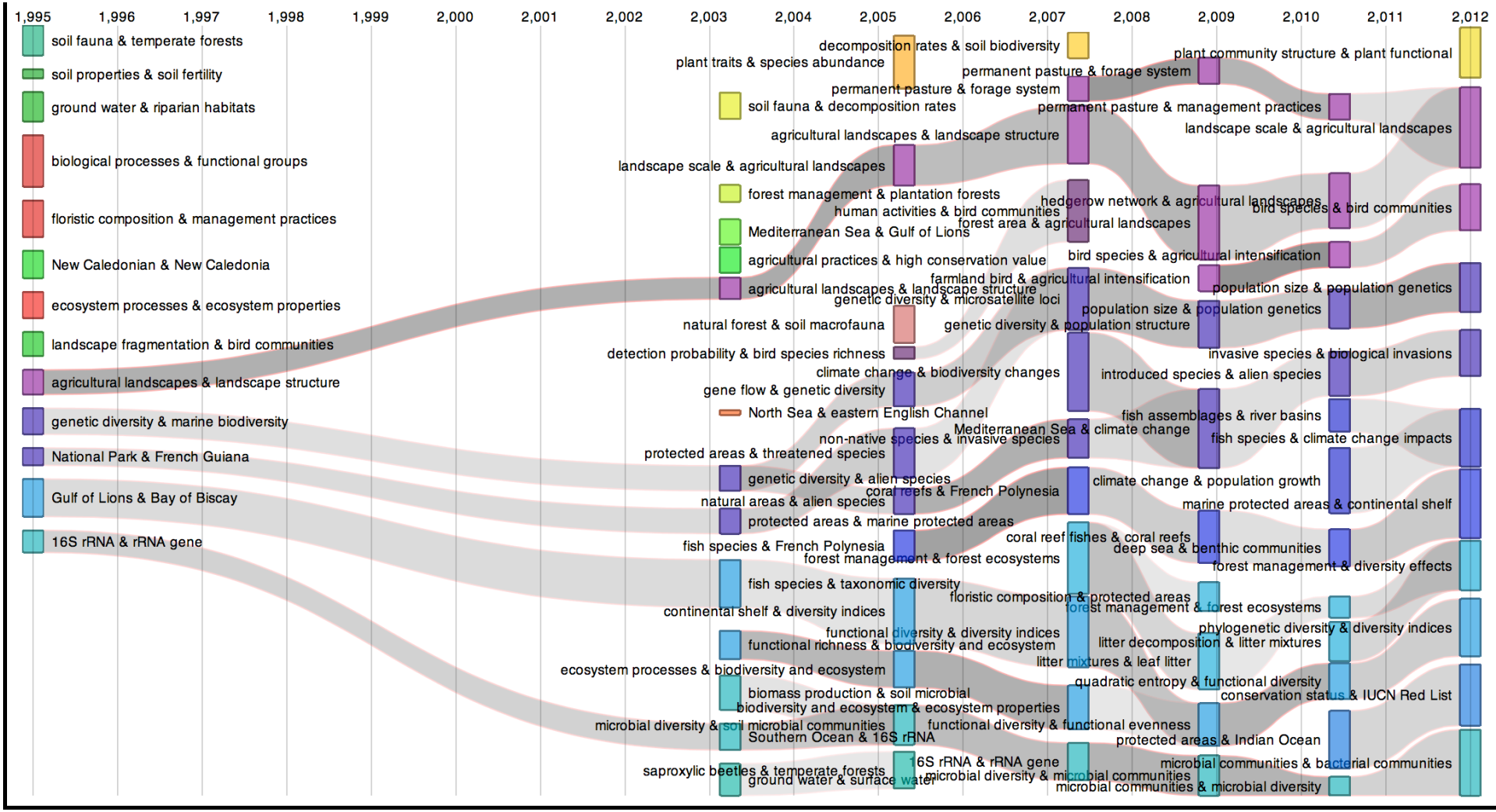
The epistemology of Dynamic Clustering



ti

F(t)

Representing the Dynamic of Calculated Clusters (Domain: Biodversity)



Section 3

- CorTexT live: some demo of analysis from datasets to publications (T=3x10mn)
- Demo: on fait des trucs sur de jeux de données qui ont abouti à des publis dans un contexte plutot SPS et Scientometrics (N=3)

Examples of Contexts of use: the intermediary-work of the STI-scientist

- Science Mapping for Science Policy Studies
- Ecology of Knowledge & Infrastructure
- Scientific Community Landscape Modeling
- Characterizing the emergence of scientific communities
- Mapping Issue Framing
- Digital Public Spaces & Politics
- Spatial dynamics & Knowledge production

Tailoring CorTexT use to Context (1)

CorTexT as a platform for

Scientific Facts Production for STS analysis of Biodiversity infratructure

ENVIRONMENTAL SCIENCE & POLICY 38 (2014) 254–262

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/envsci

Making taxonomy environmentally relevant. Insights from an All Taxa Biodiversity Inventory

Celine Granjou [a,*], *Isabelle Mauz* [a], *Marc Barbier* [b,c], *Philippe Breucker* [b,c]

[a] IRSTEA, Grenoble, 2 Rue de la Papeterie, BP 76, 38 402 St Martin d'Heres Cedex, France
[b] INRA Sens, IFRIS, Marne la Vallée, France
[c] UR Sciences en Société, 5 bd Descartes, F-77454 Marne-la-Vallée Cedex 02, France

ARTICLE INFO

Article history:
Received 25 June 2013
Received in revised form
25 November 2013
Accepted 11 January 2014
Available online 5 February 2014

Keywords:
Biodiversity
Inventory
Taxonomy
Nature conservation
Collaborative arrangements
Data

ABSTRACT

For several decades taxonomy has been marginalized in academic labs and universities. Today, rising concerns over biodiversity and ecosystem services are creating an unprecedented opportunity for it to be viewed as a crucially relevant field. This article aims to scrutinize how the biodiversity concerns entail new collaboration designs between taxonomists and nature managers and between taxonomists and ecologists. Our key point is that taxonomy's environmental relevance is not given: instead, taxonomic data have to be made relevant by taxonomists and their partners in specific collaborative and organizational arrangements. The article draws on an empirical study of an All Taxa Biodiversity Inventory (ATBI) in a national park in the French Alps, including an ethnographic survey combined with scientometric analysis. It was found that the collaboration initiated in the ATBI between taxonomists, ecologists and the park managers was paved with disappointments and reorientations because it partly failed to address the tension between a taxonomic and an ecological approach to the relevance of taxonomic data. The rise of biodiversity and ecosystem services concerns constitutes a "double-edged sword" for taxonomists: while there is greater opportunity for taxonomists to render their work visible through new research collaboration arrangements with ecologists, it also entails a risk that they remain mere data providers for nature managers and ecologists interested in ecosystem functioning.

© 2014 Elsevier Ltd. All rights reserved.

Tailoring CorTexT use to Context (2)

CorTexT as a platform for

Empowering scientific Communitie' s

Strategic thinking

Chapter 4

Textual analysis and scientometric mapping of the dynamic knowledge in and around the IFSA community

Marc Barbier, Marianne Bompart, Véronique Garandel-Batifol, and Andréi Mogoutov

Abstract *Using the proceedings of six European IFSA Symposia, we analysed the themes that were central in these Symposia as well as trends from a number of papers and authors. We then assessed the wider domain of agricultural research based on a corpus extracted from the CAB and SCI databases of the Web of Knowledge. The co-word analysis allows the generation of maps which graphically represent how keywords are linked, and allows the identification of thematic clusters. The dynamic of keywords in the period 1991–2007 was also analysed, thus allowing the identification of keywords which were of central importance during different periods. This showed how themes such as sustainability emerged, disappeared and re-emerged under different guises. The various analyses are provided to further the reflexivity of the IFSA community, especially regarding its publication practices and thus its efforts to make results from Farming Systems Research more widely available.*

M. Barbier (✉)
INRA Sens, IFRIS, Marne la Vallée, France
e-mail: marc.barbier@grignon.inra.fr

M. Bompart • V. Garandel-Batifol

Tailoring CorTexT use to Context (3)

CorTexT as a platform for

Research Domain positioning and analysis for institutional strategy

Ecosystem Services ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Ecosystem Services

journal homepage: www.elsevier.com/locate/ecoser

The place of agricultural sciences in the literature on ecosystem services

Elise Tancoigne [a,*], Marc Barbier [b], Jean-Philippe Cointet [b,c], Guy Richard [a]

[a] INRA, UAR1155 Département Environnement et Agronomie, Centre de Recherche Val de Loire, 2163 avenue de la Pomme de Pin, CS 40001, Ardon, F-45075 Orléans cedex 2, France

[b] INRA, UR1326 SenS, IFRIS, Université Marne-la-Vallée, Cité Descartes, Champs-sur-Marne, 5 boulevard Descartes, F-77454 Marne-la-Vallée Cedex 02, France

[c] Institut des Systèmes Complexes – Paris Ile-de-France (ISC-PIF), 113 rue Nationale, 75013 Paris, France

ARTICLE INFO

Article history:
Received 6 February 2014
Received in revised form
1 July 2014
Accepted 23 July 2014

Keywords:
Bibliometric analysis
Semantic networks
Ecosystem services
Agricultural sciences
Socio-ecosystem

ABSTRACT

We performed a quantitative and qualitative analysis of the scientific literature on ecosystem services in order to help tracing a research agenda for agricultural sciences. The ecosystem services concept now lies at the heart of current developments to address global environmental change. Do agricultural sciences generate knowledge that covers this emerging theme? An analysis of scientific production allowed us to return to the ecological origins of this concept and see how little it has been appropriated by agricultural sciences until now, despite major focus on the issue of agro-ecosystems in the literature. Agricultural sciences tend to be more active in the field of environmental services, defined as services rendered by humans to ecosystems. The main studied services are those which have already been clearly identified and which act in synergy. Less attention is paid to the antagonisms between different services. These findings call for the implementation of agricultural research programmes that will consider the socio-agro-ecosystem as a whole and broaden the traditional issues addressed by agricultural sciences. We insist on three main management and operational issues that needs to be overcome if this is to be done: working at the landscape scale, increasing inter-disciplinary collaborations and take uncertainties into account.

© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-SA license (http://creativecommons.org/licenses/by-nc-sa/3.0/).

Tancoigne, E., et al., The place of agricultural sciences in the literature on ecosystem services. Ecosystem Services (2014), http://dx.doi.org/10.1016/j.ecoser.2014.07.004i

Tailoring CorTexT use to Context (4)

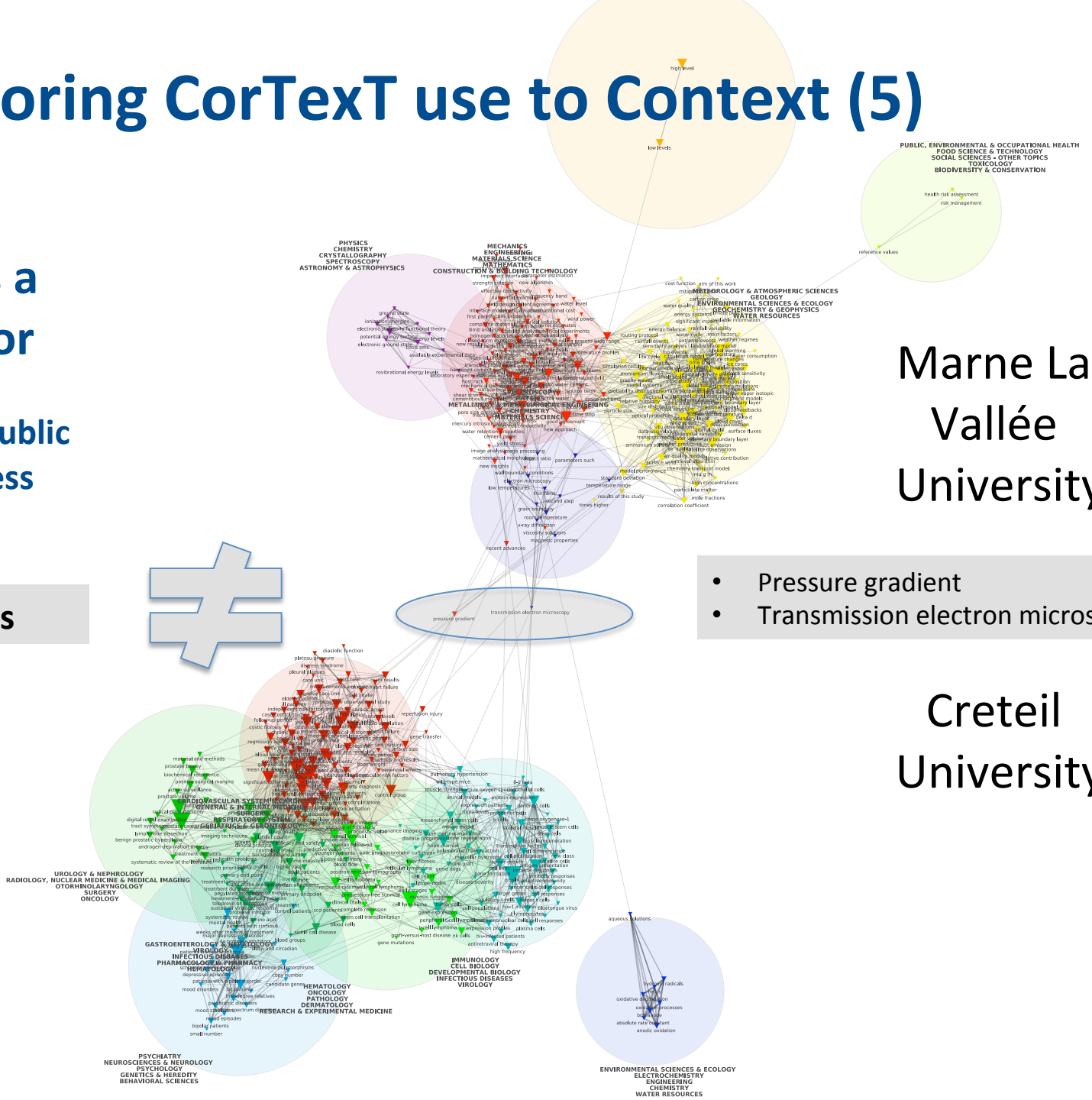
CorTexT as a platform for

Impact analysis of Public Research Programs

http://www.agence-nationale-recherche.fr/fileadmin/documents/2016/Rapport-Impact-Agro_ANR-Inra.pdf

Tailoring CorTexT use to Context (5)

CorTexT as a platform for Institutional Investments in Public Science Business



Marne La Vallée University

- Pressure gradient
- Transmission electron microscopy

Fronts of sciences

Bibliome UPE Map of Extracted Terms in abstratct and projection of Subject Category

Creteil University

Pole UPE 2012-2015
Top 500 Termes extraits
Métrique: Distributional
7 nearest nodes
Tagging: 5 SUJ, Chi2