RISIS Research infrastructure for research and innovation policy studies

SEVENTH FRAMEWORK PROGRAMME

CorTexT

IFRIS Institut Francilien Recherche Innovation Société

LabEx Science, Innovation and TEchnology in Society

DOCUMENTION
Training CorTexT-Risis
SHORT COURSE TYPE A (SCA)
10 -12 May 2017

See : http://risis.eu/events/

Organized by CorTexT-Lab Team
OF LISIS Unit in Paris-Est

INRA SCIENCE & IMPACT

UPEM UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

ESIEE PARIS

Marne La Vallée, May 2017

Table of Contents

1. OVERVIEW OF RISIS INFRASTUCTURE 3

2. THE RISIS CORE FACILITY 4

3. DOCUMENTATION OF THE RISIS CORTEXT PLATFORM 5

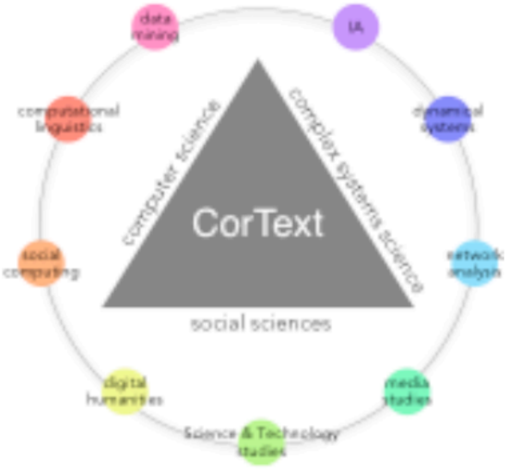
4. MAIN FEATURES OF RISIS CORTEXT 6

CORPUS BUILDING AND EXPLORER 7

TEXTUAL PRE-PROCESSING AND PROCESSING 9

TIME-BASED DATA PROCESSING 14

DATA COMPUTING AND ANALYSIS 16



This documentation has been realized by the CorTexT-Lab Team.

1. OVERVIEW OF RISIS INFRASTUCTURE

Specific Datasets Platforms and Analytical Platforms gathered in the RISIS Infrastructure have the objective to overcome their specific scope, their frame and contents and their openness in order to achieve the goal of opening a common asset and deliver a breakthrough infrastructure in the research domains of humanities and Social Sciences that targets the study and the analysis of Science, Technology and Innovation in Society.

The functional interoperability of those platforms grounds the Infrastructure Project RISIS. The two platforms – Cortext and SMS digital platforms – are complementary in their focus and are committed to realize a joint opening of their facility and to foster the deployment of the infrastructure.

SMS aims at taking advantage of the new (and numerous, even endless) resources provided by the web and to structure a convenient and robust solution to enable linked-data processing under specified format of for data retrieving and enrichments.
CortTexT Manager is an integrated application dedicated to the processing and treatment of various types of corpuses produced by researchers, whether structured or unstructured.

Semantic treatments are important to make the data suited for Cortext. After the opening of each platform as a service, the main challenge of the period has been the opening of a link to circulate between platforms, addressing technical issues that require a selection of methods to use (within existing methods) and specific tailoring on the 2 platforms.

The collaboration between CorTexT and SMS platforms has been crucial for that purpose. This technological collaboration has been opened during 2016 at the achievement of the WP.6 in order to have an inclusive approach of Data Sets Platforms specifications and specificities. Herein below are the main activities that have been defined through Activity Sheets in relation to the objective of developing the RCF:

- Linking efficiently different RISIS datasets for global retrieving and joint analysis from an unique secured workspace
- Providing possibilities for linking and jointly analyzing RISIS datasets with other sources of open linked data
- Providing a rich set of tools to process, analyze and visualize RISIS data in one integrated and scalable platform
- Empowering users with a working space to circulate, store datasets and results and share knowledge that they aim to publish
- Enabling the monitoring of user practices with the aim of reporting and communicating on the RISIS Core Facility.

In this context the present documentation focuses on use of CorTexT Platform to providing a rich set of tools to process, analyze and visualize datasets in one integrated and scalable platform.

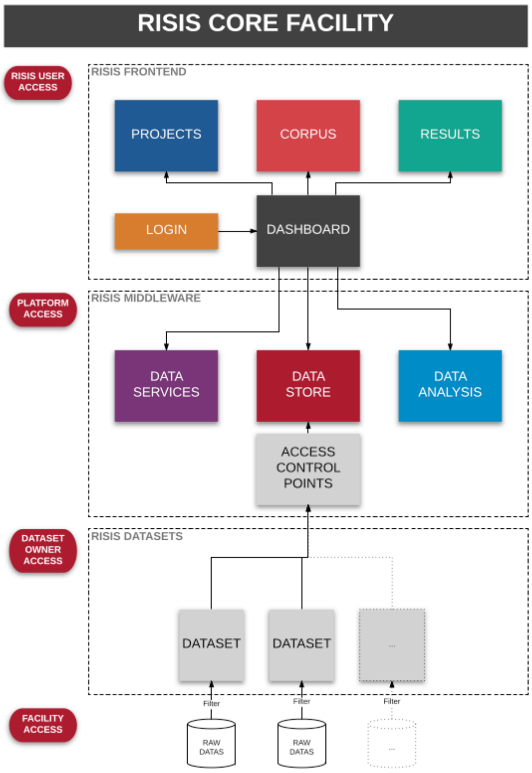
2. THE RISIS CORE FACILITY

The CorTexT Manager Interface proposes a Dashboard to access various projects that can be composed of one or several datasets. It is possible for users to share their project – and therefore their datasets- with colleagues who have been previously recognized as member of the infrastructure, including their rights to access to datasets.

Basics of the Infrastructure are the different RISIS datasets that enter the platform in an unfiltered, or where necessary filtered form (e.g. just extracts of a specific dataset). The platform comprises all routines for data storage, data services and data analysis. Via the RISIS frontend, users with interesting research projects can login to the dashboard and access data, analyze them, and retrieve results and/or extract a newly created data corpus.

The structure inside UPEM is fully secured with its own storage, processing and frontend RISIS servers. All data are double back-upped (inside and outside server location). All transactions inside UPEM are encrypted with HTTPS/SSL connections, preventing user information or data to be visible on network.
Access is restricted at different levels: physical, hardware and software including user access from outside the UPEM infrastructure.

Data are imported into this infrastructure according to an API reference that has been implemented by the RISIS Software Development Team. Data are either hosted at UPEM already or imported via SMS API endpoints (formatted with the API mentioned).
The following scheme illustrates the overall design of the global RISIS infrastructure as a core system.



3. DOCUMENTATION OF THE RISIS CORTEXT PLATFORM

All these features are documented on the CorTexT Manager on line “Help web page” (https://docs.cortext.net/) and a set of video tutorial is offered to users to enable basic actions (https://player.vimeo.com/video/185307813). This web application has been plugged and tuned for the RISIS project and a specific Dashboard has been developed. To summarize, three level of documentation exist:

. A help page describes all the Scripts and the way to process datasets and choose parameters;

. Contextual basic information and warnings are delivered along the scripts proposed to uses within the interface when parametric choice are key for data processing or calculation;

. Some tutorial videos have been realized to help users.

The documentation covers all the scripts described above.

CORPUS BUILDING AND EXPLORER

DATA PARSER
CORPUS EXPLORER
RISIS DATASETS

TEXTUAL PRE-PROCESSING AND PROCESSING

TERM EXTRACTION
CORPUS TERM INDEXER
LIST BUILDER
CORPUS LIST INDEXER

TIME-BASED DATA PROCESSING

DEMOGRAPHY
PERIOD SLICER

DATA COMPUTING AND ANALYSIS

NETWORK ANALYSIS
CORRESPONDENCE ANALYSIS

4. MAIN FEATURES OF RISIS CORTEXT

The main features of users' access in the RISIS CorTexT Manager are:

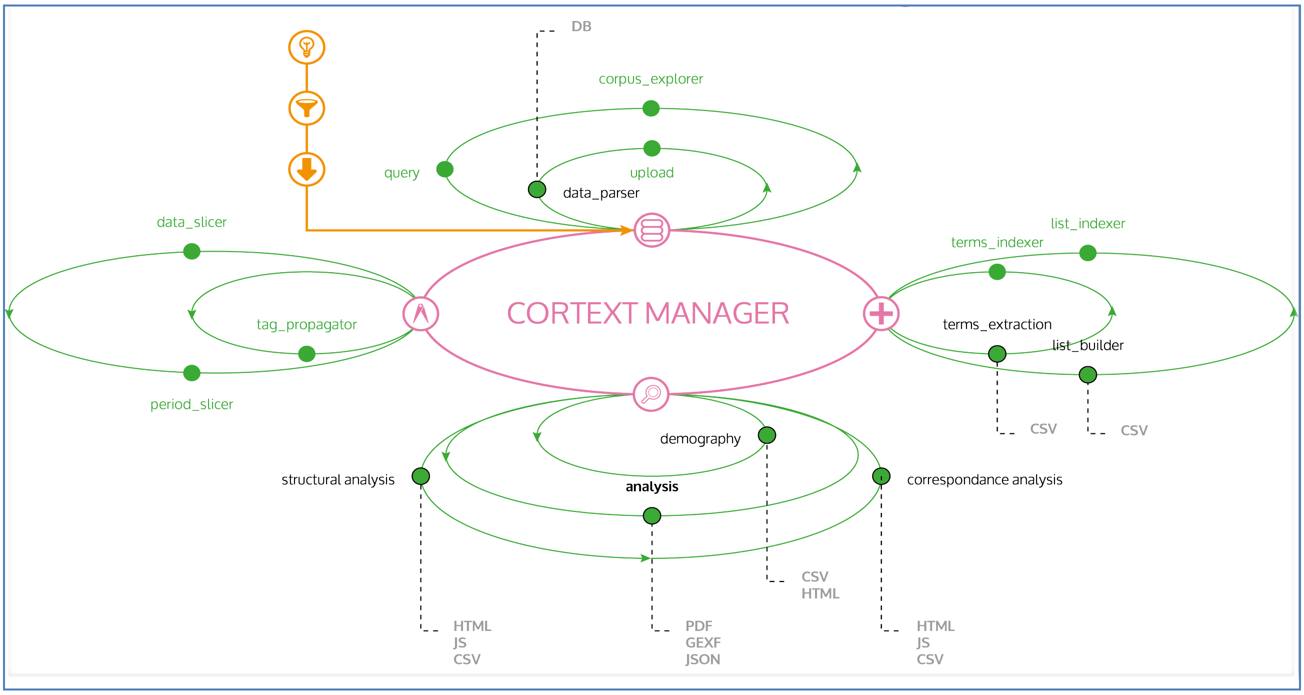
Parsing and exploration of Corpus

Enrichment of corpus through terminological extraction, entity recognition and term list filtering

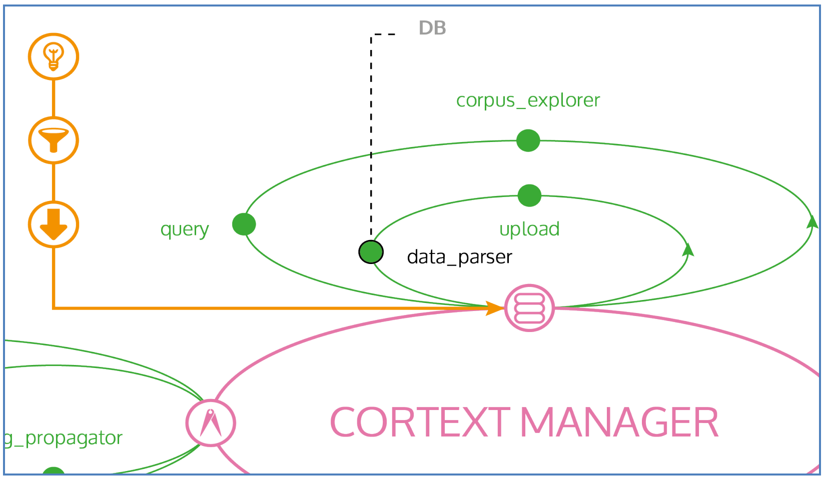
Datasets analysis and visualization of results

Manipulations of time dependent variables

The user' accessibility to the capacity has received a functional analysis of main task in relation to the various scripts that have been assembled in the menu of the interface and organize to empower scenario of analysis. The next sections 5.2 to 5.5 propose the specification of those features.



CORPUS BUILDING AND EXPLORER

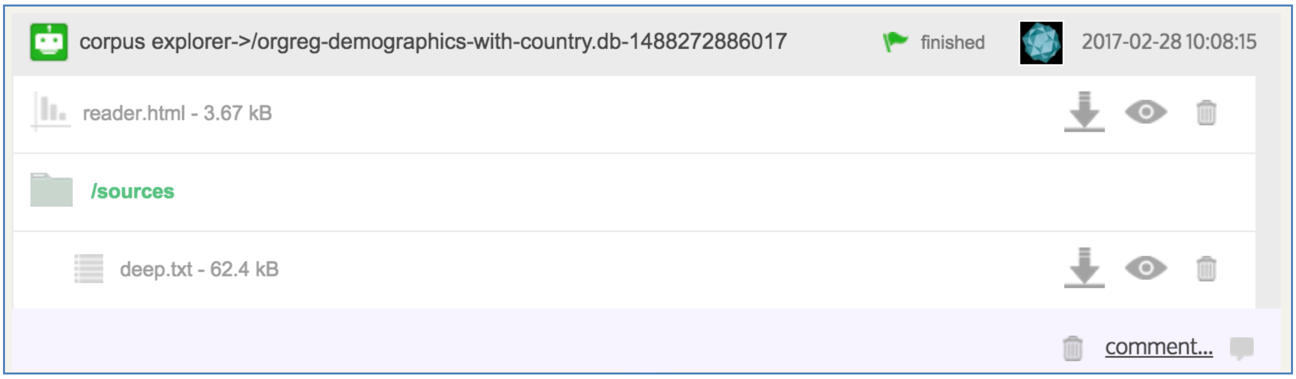


DATA PARSER

The script “Data parser” is a generic parsing script that handles a wide range of data formats: ISI files (as downloaded from the Web Of Science), Factiva datasets, PubMed (in the xml format export), RIS file as provided by scientific platforms such as Google Scholar or Scirus, batches of simple text files or any file formatted in csv format. It is also possible to parse xls files from Excel or Open Office. Europress parser is also available plus other specific database parsers. The capacity of this parser to afford many format of datasets enable a portable solution to link CorTexT Manager application to any type of structured datasets that are available in the Datasets Platform of the RISIS Project. As output, data parser produces a SQLite database (suffixed by a .db).

CORPUS EXPLORER

Moreover a second script called “Corpus Explorer) enables a simple and robust navigation interface within the DB (in a html mode: reader.html): access to sources file is possible to get back to raw data (/sources), a comment can be dropped to users that collaborate on the same project


corpus explorer->/orgreg-demographics-with-country.db-1488272886017 finished 2017-02-28 10:08:15

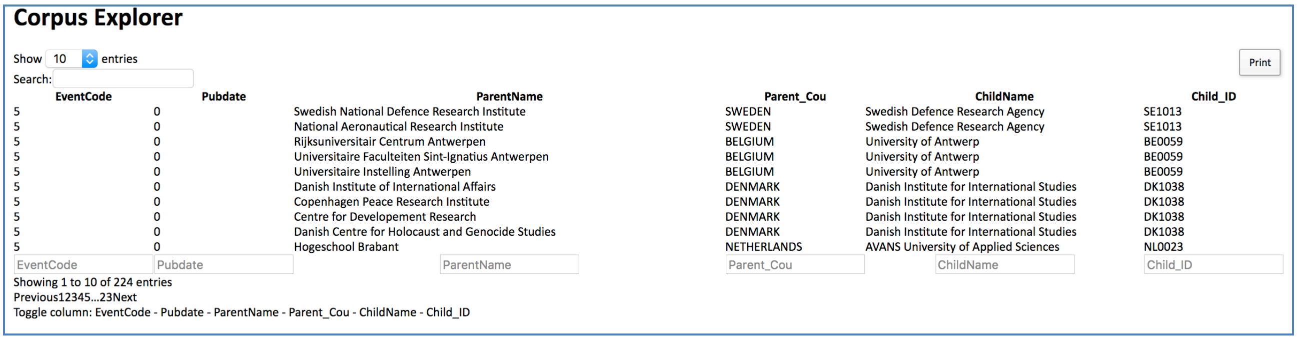
reader.html - 3.67 kB

/sources

deep.txt - 62.4 kB

comment...

When accessing to reader.html page, one is allowed to search, to till, to save and print the selected raw (with a threshold selection)



Corpus Explorer

Show 10 entries

Search:

EventCode	Pubdate	ParentName	Parent_Cou	ChildName	Child_ID
5	0	Swedish National Defence Research Institute	SWEDEN	Swedish Defence Research Agency	SE1013
5	0	National Aeronautical Research Institute	SWEDEN	Swedish Defence Research Agency	SE1013
5	0	Rijksuniversitair Centrum Antwerpen	BELGIUM	University of Antwerp	BE0059
5	0	Universitaire Faculteiten Sint-Ignatius Antwerpen	BELGIUM	University of Antwerp	BE0059
5	0	Universitaire Instelling Antwerpen	BELGIUM	University of Antwerp	BE0059
5	0	Danish Institute of International Affairs	DENMARK	Danish Institute for International Studies	DK1038
5	0	Copenhagen Peace Research Institute	DENMARK	Danish Institute for International Studies	DK1038
5	0	Centre for Developement Research	DENMARK	Danish Institute for International Studies	DK1038
5	0	Danish Centre for Holocaust and Genocide Studies	DENMARK	Danish Institute for International Studies	DK1038
5	0	Hogeschool Brabant	NETHERLANDS	AVANS University of Applied Sciences	NL0023
EventCode	Pubdate	ParentName	Parent_Cou	ChildName	Child_ID

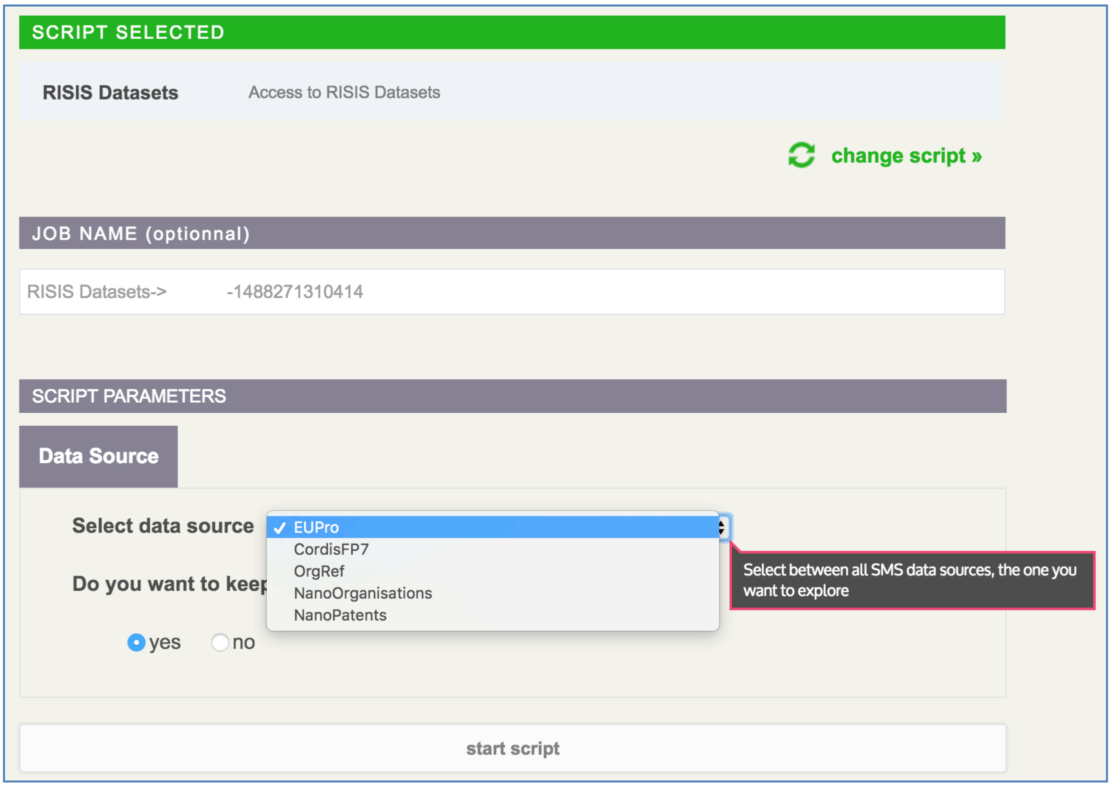
Showing 1 to 10 of 224 entries

Previous12345…23Next

Toggle column: EventCode - Pubdate - ParentName - Parent_Cou - ChildName - Child_ID

RISIS DATASETS

Since the Joint Opening of SMS and CorTexT is supposed to interoperate retrieving and analysis of corpus, we did not develop until now specific parsers to each of the Datasets Platforms. The collaboration between the two facilities supposed that CorTexT and SMS APIs would be connected so that a simple and direct parser of SMS outcomes was to be designed and developed. This has been realized end of 2016. Thus, the Dashboard includes a script called "RISIS Datasets" that enable to connect to all Datasets Platforms as they can be plugged in the DataStore. In the following Snapshot, one can notice that the choice of several Datasets is proposed: EUPro, Cordis FP7 (an extraction of the EUPro database), NanoOrganisations (a rich Database of PRO in the field of NanoScience and technology) and NanoPatents (an extraction of the PATSTAT datasets).



SCRIPT SELECTED

RISIS Datasets — Access to RISIS Datasets

JOB NAME (optionnal)

RISIS Datasets-> -1488271310414

SCRIPT PARAMETERS

Data Source

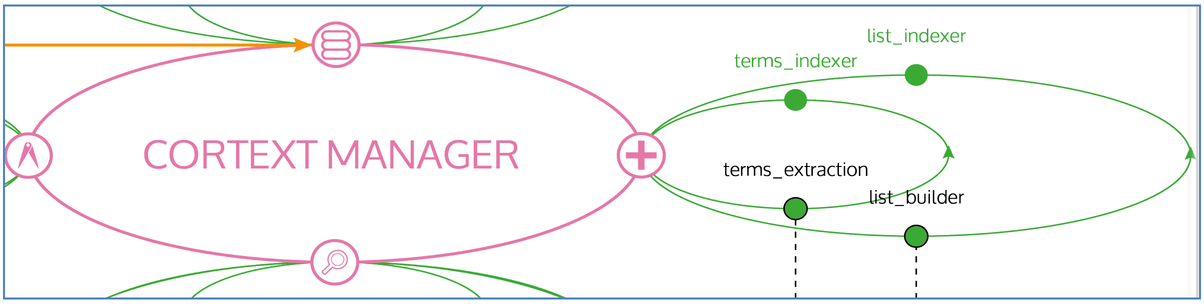
Select data source: ✓ EUPro

Select between all SMS data sources, the one you want to explore

Do you want to keep

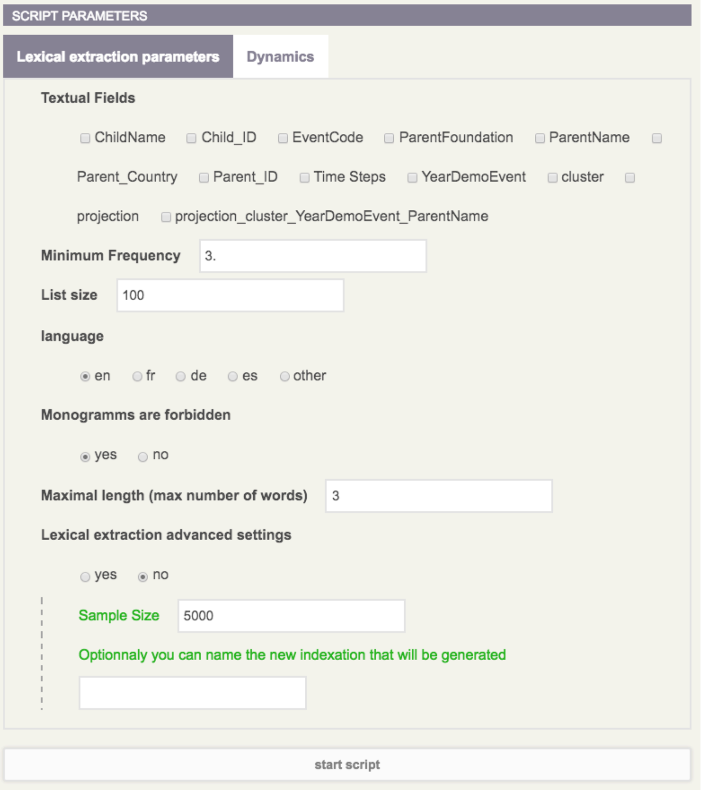
● yes ○ no

TEXTUAL PRE-PROCESSING AND PROCESSING



TERM EXTRACTION

This script is a powerful script, which relays on terminological extraction up-to-date solutions. Automatic multi-terms extraction is a typical task in NLP, yet the existing tools are not always well suited when one wishes to extract only the most salient terms. As specificity computing is time and resource expansive, we have developed an automatic method to extract lists of terms that we suspect to be the best candidates for lexical extension. Thus we will be interested in groups of relevant terms featuring both high unithood and high termhood as defined in (Kageura, K., & Umino, B., 1996).
Terms extraction automatically identifies terms pertaining to a given corpus. In fact, Natural Language Processing (only for English, French, Spanish or German text) tools that we use allow us to identify not only simple terms but also multi-terms (called n-grams). The final result is compiled in a csv file which can be either downloaded and edited offline or with a spreadsheet editor like open office (recommended) or even edited online using the online csv editor provided by CorText (simply click the csveditor.php file).



SCRIPT PARAMETERS

Lexical extraction parameters | Dynamics

Textual Fields

□ ChildName □ Child_ID □ EventCode □ ParentFoundation □ ParentName □ Parent_Country □ Parent_ID □ Time Steps □ YearDemoEvent □ cluster □ projection □ projection_cluster_YearDemoEvent_ParentName

Minimum Frequency 3.

List size 100

language

● en ○ fr ○ de ○ es ○ other

Monogramms are forbidden

● yes ○ no

Maximal length (max number of words) 3

Lexical extraction advanced settings

○ yes ● no

Sample Size 5000

Optionnaly you can name the new indexation that will be generated

start script

Parameters and tuning the terminological extraction

Textual fields definition – Select the textual fields one wish to analyze and index:

Terms list filtering – lexical extraction script aims at identifying the most salient multi- terms according to statistical criteria (see technical description below). You can also exclude any term below a given minimum frequency. Another important parameter is the list size you wish to extract. This parameter may have strong impact on script speed, so it is advised to keep this parameter below 1000

Language – Enable to specify the language of dataset – only French, German, Spanish and English are taken into account. If one selects "Other", no grammatical processing will be applied to the data, meaning only statistical criteria based on collocation of words will be used to derive phrases.

Monograms – Give the possibility to exclude monograms (that is terms composed of only one word). It is advised to exclude monograms, as they tend to be less informative terms.

Maximal length (max number of words) – It is possible to limit the maximum number of words encapsulated in a multi-term. Three is reasonable, but feel free to try to identify longer multi-terms.

Advanced settings – These advanced options are described below.

Sampling – For large datasets. Terms extraction will only be based on sample sub-corpuses with a given number of documents (randomly drawn from original successive datasets). Nevertheless, detected terms will be indexed in the whole corpus whatever the sampling strategy.

Frequency (c-value) computation level: The frequency of terms can be computed at the document level (meaning that terms frequency are computed based on the number of distinct documents they appear in) or at the sentence level (default choice).

Specificity score: The selection of most pertinent terms results from a trade-off between their specificity and their frequency (see technical explanations below). By default, specificity is computed as a chi2 score. It is also possible to use a simpler tf.idf score to do so. One can also deactivate the role of specificity in the final ranking of extracted terms such that only top N most frequent terms will be retrieved.

Linguistic pre-processing: One can also totally deactivate the linguistic pre-processing (post-tagging, chunking, stemming). It is useful when treating texts in other languages than English, French, German or Spanish, but also in cases where one does not want to reduce the extraction to a single grammatical class.

Grammatical criterion: By default, noun phrases are identified and extracted but you can also choose to try to identify adjectives or verbs.

Automatically index the corpus: The script first extracts a list of terms and then indexes the corpus. It is possible not to index the corpus by checking this box.

Pivot Words: Only multi-terms containing this string will be extracted.

Starting Character: Only terms starting with this character shall be extracted. It comes in handy if you wish to index hashtags for instance (#). Simply deactivate linguistic pre-processing and make sure to also deactivate linguistic pre-processing in that case.

Time periods – Different lexical extraction process will be applied to the different time periods defined (either from the original time range or from a customized time range if one was computed before). Time slices are either regular (uniform distribution of time steps per period) or homogeneous (uniform distribution of documents per period).

Technical description

The whole processing of textual data can be described as follows: it first relies on classic linguistic processes that ends up defining sets of candidate noun phrases:

POS-tagging: Part-of-Speech Tagging tool first tags every term according to its grammatical type: noun, adjective, verb, adverb, etc.

Chunking: Tags are then used to identify noun phrases in the corpus; a noun phrase can be minimally defined as a list of successive nouns and adjectives. This step builds the set of our possible multi-terms.

Normalizing: We correct small orthographical differences between multi-terms regarding the presence/absence of hyphens. For example: we consider that the multi-terms “single-strand polymer” and “single strand polymer” belong to the same class.

Stemming: Multi-terms can be gathered together if they share the same stem. For example, singular and plurals are automatically grouped into the same class (e.g. “fullerene” and “fullerenes” are two possible forms of the stem: “fullerene”).

Counting: We enumerate every multi-term belonging to a given stemmed class in the whole corpus to obtain their total number of occurrences (frequency). In this step, if two candidate multi-terms are nested, we only increment the frequency of the larger chain. For example if “spherical fullerenes” is found in an abstract, we only increment the multi-stem: “spheric fullerene” but not the smaller stem “fullerene”.

References:

Frantzi, K., & Ananiadou, S. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. International Journal on Digital Libraries

van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. Arxiv preprint arXiv:1109.2058.

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. Terminology, 3(2), 259–289. John Benjamins Publishing Company.

CORPUS TERM INDEXER

This script works hand in hand with the lexical extraction. Actually, by default, it is even automatically launched every time a lexical extraction is executed. Its basic objective is, given a series of textual fields (provided by the user), to index every term found in a given term list csv file (specified by the user).
It then provides more flexibility in the indexation process as users are allowed to edit term list by themselves either by editing their own csv in a spreadsheet editor like open office (recommended) or Google Spreadsheet or by using the online csv editor provided by CorText.
Only the second and third columns are important for launching an indexation. Concretely, user should provide tabulation separated UTF-8 encoded file with no text delimiter (which is already the format generated by CorText lexical extraction). The indexer will proceed as follows. The third column (classically entitled "forms") provides a list (separated by |&|) of strings that will be indexed using the label provided by the second column (entitled "main form"). It means that each time that one of those strings is found the database will store this information. Other available options should be straightforward. They include the possibility to check for case when indexing to only index one occurrence of a term per sentence.

LIST BUILDER

List builder helps user to manage categorical entities and named entities. Not only does it provide lists of most frequent textual entities for a given field but it also creates a list of potentially duplicate entries when raw data are noisy (potentially useful for cited references, names, cited journals, addresses etc.)

Field – User selects the field
List size – Only the most frequent items will be extracted
Proximity Threshold – This setting is necessary for entity normalization. Default value should be fine, but if one feels like some equivalent forms are still missing, one may need to lower this threshold.
Exhaustive search: By default, two strings are assumed strictly different if they share no common words. If one wants to go beyond, one should check the Exhaustive Search option that will extend the search scope.

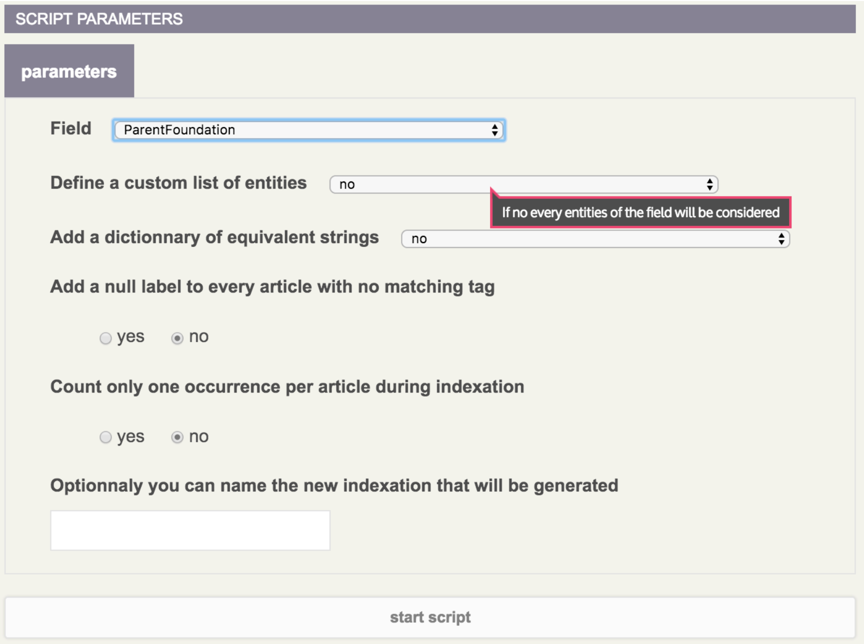
Technical description

First, the set of N most frequent items are selected and exported along with their frequency in a csv file.
Second, every couple whose string-based proximity is above the threshold will be exported in a csv file (equivalence file) compiling a wide range of proximity measure: ngram proximity, Levenshtein ratio, word-level proximity measures (inclusion, jaccard, etc.). Some measures may be more adequate according to the field although Levenshtein distance is usually the most pertinent. Item frequencies are also provided to help focus on the most interesting phrases. Note that when applied to isi-based cited references, only references published the same year may be considered as possible equivalent couples.

Once cleaned, those two files can be later used in corpus list indexer script. A csv editor is also provided to visualize and edit those files directly online.

CORPUS LIST INDEXER

This script is naturally connected to list builder script. It provides users with full control other a set of items that may later get mapped or analyzed. Technically, a new field will be created based on user selection.



SCRIPT PARAMETERS

parameters

Field: ParentFoundation

Define a custom list of entities: no

If no every entities of the field will be considered

Add a dictionnary of equivalent strings: no

Add a null label to every article with no matching tag

○ yes ● no

Count only one occurrence per article during indexation

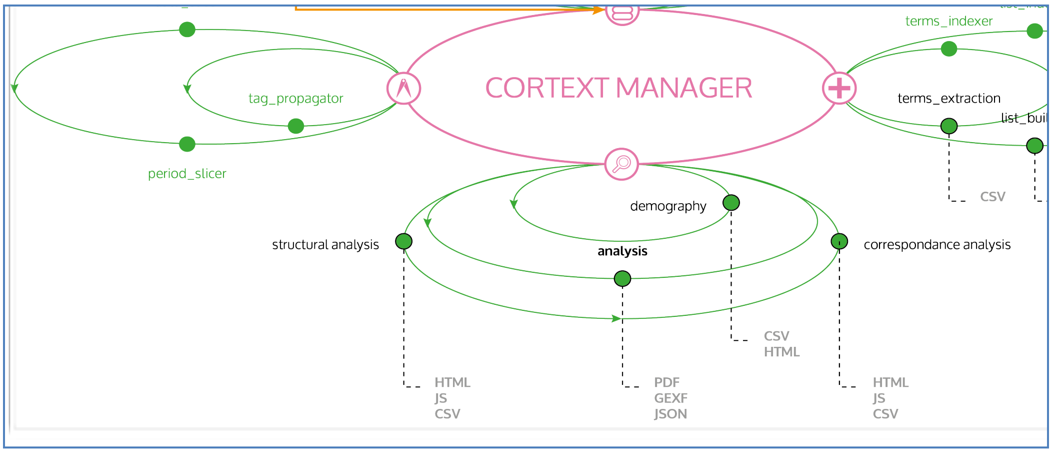
○ yes ● no

Optionnaly you can name the new indexation that will be generated

start script

Field – Select the field you wish to work on
Define a custom list of entities – If yes, one can provide a csv file filled with a list of items that will be specifically indexed in the target field (concretely, only the first column of a tabulated csv file will be considered). If no, every entity present under the chosen field will be indexed.
Add a dictionary of equivalent strings – If yes, one should provide a csv file made of couples of equivalent strings. Entities from the first column of the csv file will be automatically transformed into second column entities.
Add a null label to every article with no matching tag – This will label "null" any field that has none of the tags chosen by the user
Count only one occurrence per article during indexation – This option is useful when one does not wish that several occurrences of the same entry were mentioned for a given document. For instance, if one wants to compute the distribution of articles published by the USA in a scientific database, it may be useful to re-index the Country field first with this option, such that articles written by at least one American author are counted only once. By default, if several scientists with different US affiliations publish a paper, then this article is indexed with several occurrences of USA in the raw database.
Finally, one can give a custom name to the newly generated field.

TIME-BASED DATA PROCESSING

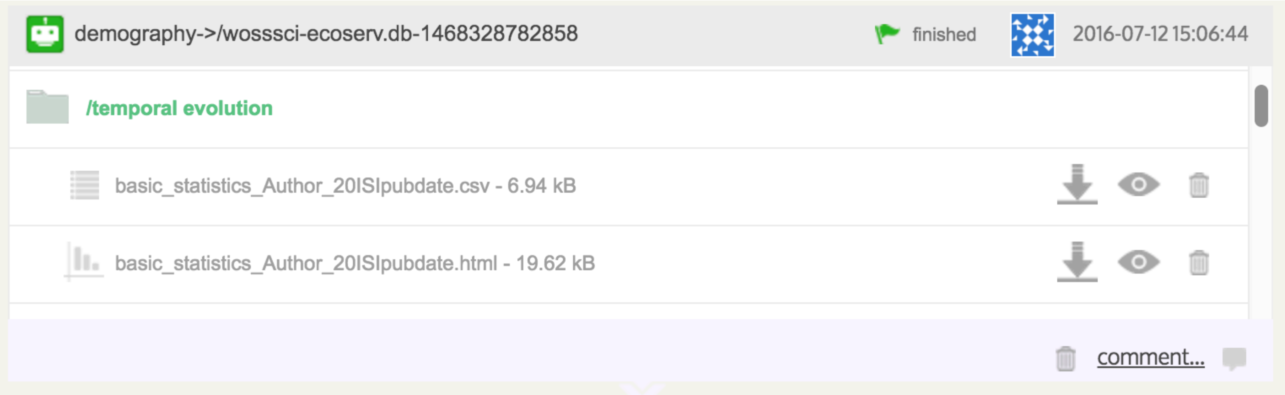


DEMOGRAPHY

Demography processes each field of the corpus and counts the raw evolution of occurrences of the top items. You will simply be asked to specify the number of top items you wish to evaluate, and also whether to use custom or regular time periods. The script creates two directories called “global distributions” and “temporal evolution”.

The first directory “global distribution” simply lists the distribution of items per document and the distribution of documents per item of each field. Those files are useful in order to understand the distributions of the number of authors per article or number of papers written by authors in a scientific database (by selecting the Authors field). Note that distributions are computed over all possible entries in the database, thus ignoring the number of top items to consider.

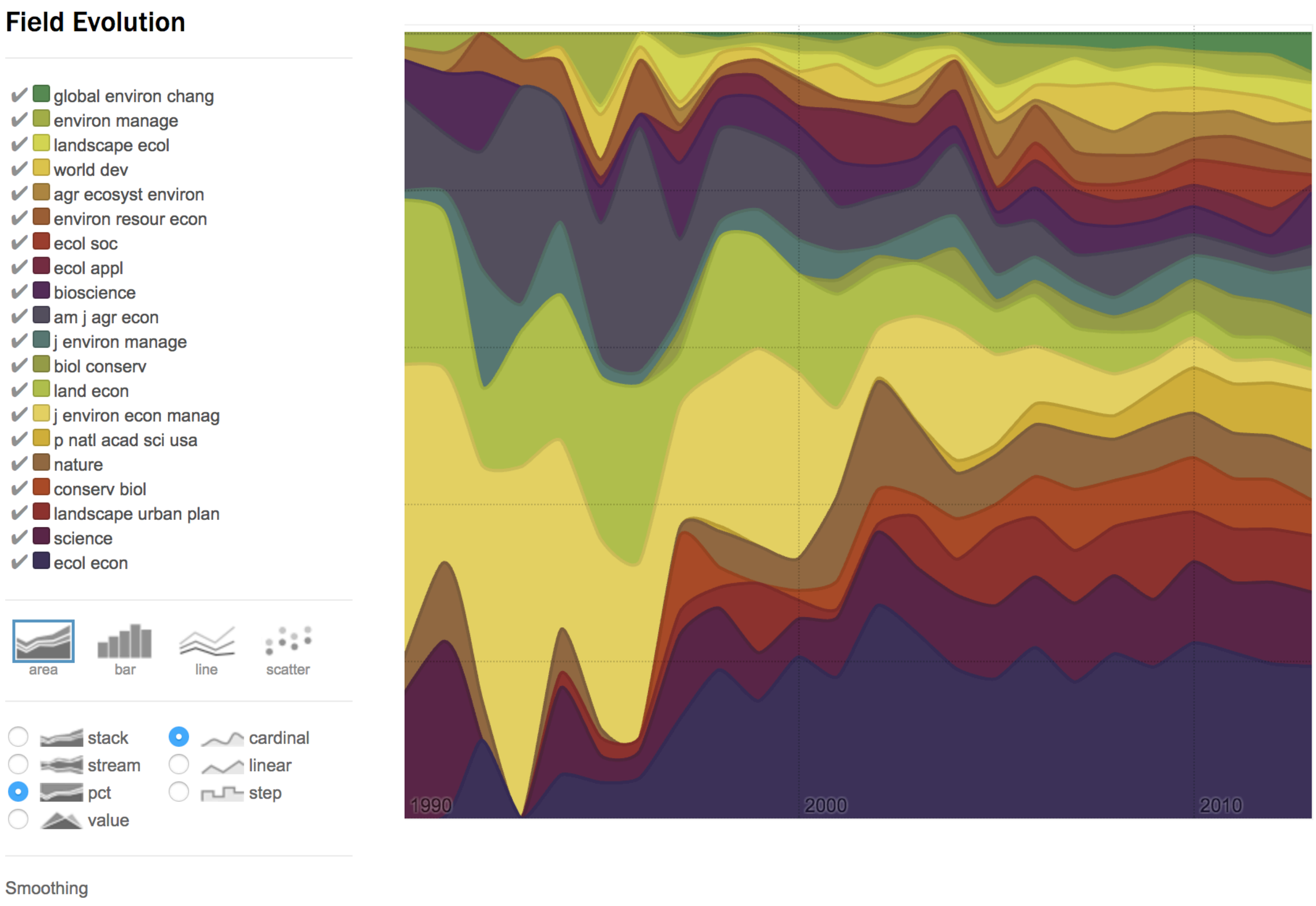
In the “temporal evolution” directory, each field of the corpus will be enumerated over time in a csv file compiling the occurrences at each time step of the top items of the given field (original count of occurrences averaged over 3 or 5 time-steps windows are also available for analysis if raw statistics are too noisy). A dedicated web interface is also provided by clicking the html files to visualize and customize the chart of each chosen field.


demography->/wosssci-ecoserv.db-1468328782858 finished 2016-07-12 15:06:44

/temporal evolution

basic_statistics_Author_20ISIpubdate.csv - 6.94 kB

basic_statistics_Author_20ISIpubdate.html - 19.62 kB



○ stack ● cardinal
○ stream ○ linear
● pct ○ step
○ value

PERIOD SLICER

The script Period Slicer allows users to customize the time periods according to which the corpus is analyzed.

Parameter: When launching Period Slicer, the user needs to input the different time periods, which will be used to split the corpus. Each time period should be separated by semi-colons. A time period is defined by a series of integer values (typically years) separated by commas and between brackets.

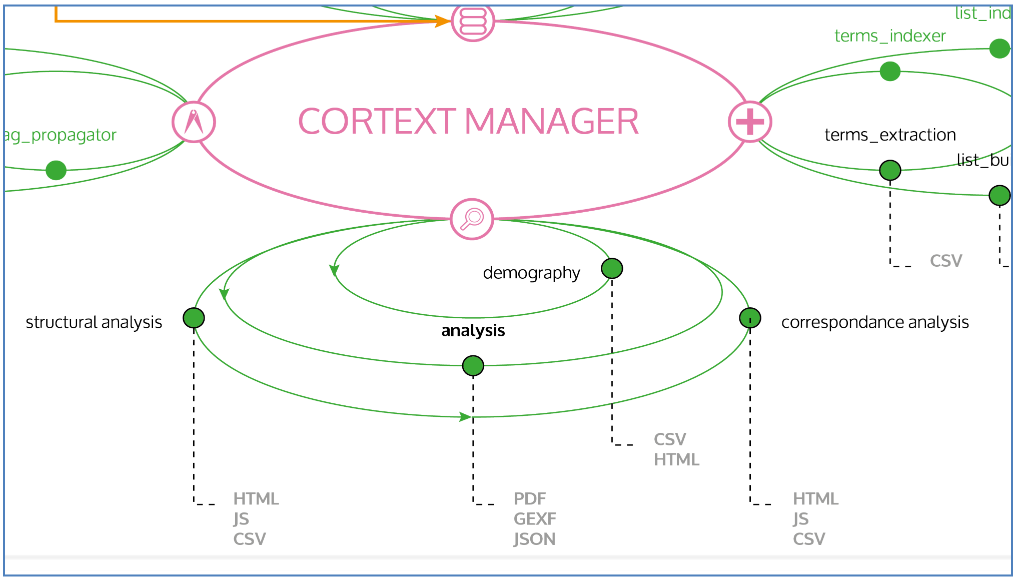
For example [2000,2001];[2002,2003] will defines two time periods: 1: [2000,2001] and 2: [2003:2003]. Screenshot from 2016-08-16 15:16:14

Alternatively, one can use colons to define time ranges: [2000:2009] will define one time period ranging from 2000 to 2009 (included). Commas and colons can be combined in the same time period.

Custom periods defined by Period Slicer can then be used in any script involving time periods (analysis, and terms extractor) simply by choosing “Custom periods” as period’s parameter. Custom periods can also be used as a variable of analysis on its own: choose Periods as field to draw heterogeneous maps involving time periods as nodes.

Period Slicer’s main result is to create a new table in the database to perform further analysis. Moreover a simple csv file (tab separated) containing the number of documents for each time-step is also produced – plotting the evolution of documents is then straightforward and can help to define more accurately a satisfying time range (periods containing a few articles (typically less than 100) could lead to too noisy results).

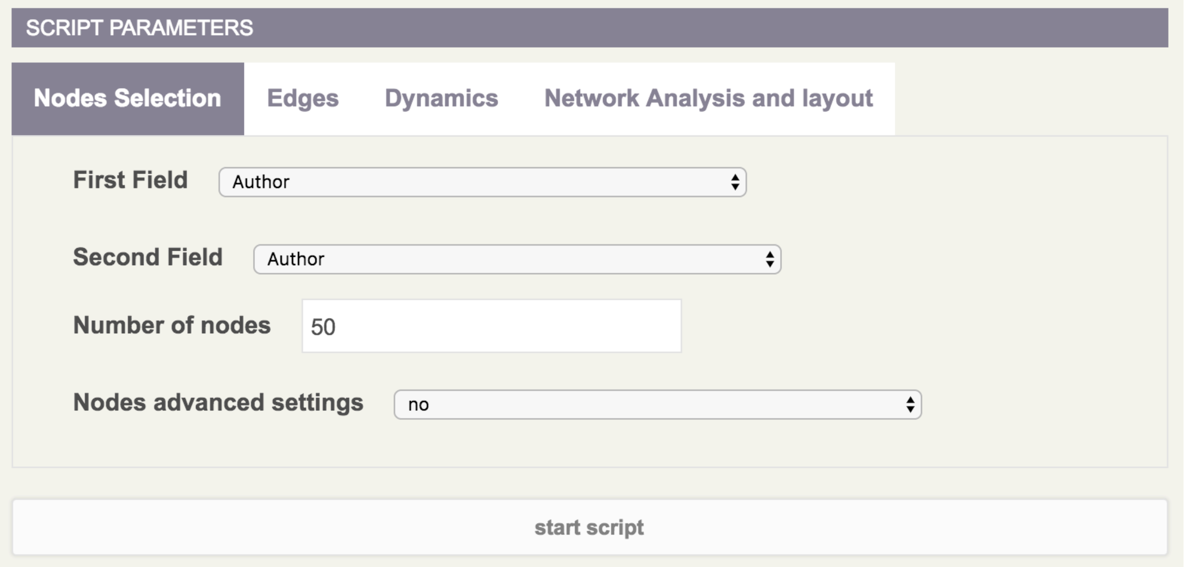
DATA COMPUTING AND ANALYSIS



NETWORK ANALYSIS

This script produces several types of analysis and visualizations. The maps feature homogeneous or heterogeneous nodes (see **nodes selection**) that can be linked according to different types of proximity measures (see **edges**). Different advanced options (see **Network Analysis & Layout**) are also proposed for tagging clusters, producing historical maps, generating "heatmaps" or contingency matrices. Last tubes provide a dynamical perception of maps transformation in time (see **Dynamics**). A web interface also allows browsing maps and editing node and clustering information. It is based on TinaJS explorer developed at ISCPIF. Note that only homogeneous maps are visible with the web interface. Cluster tags and heatmaps are also only visible in the pdf version of the maps.

Node Selection interface



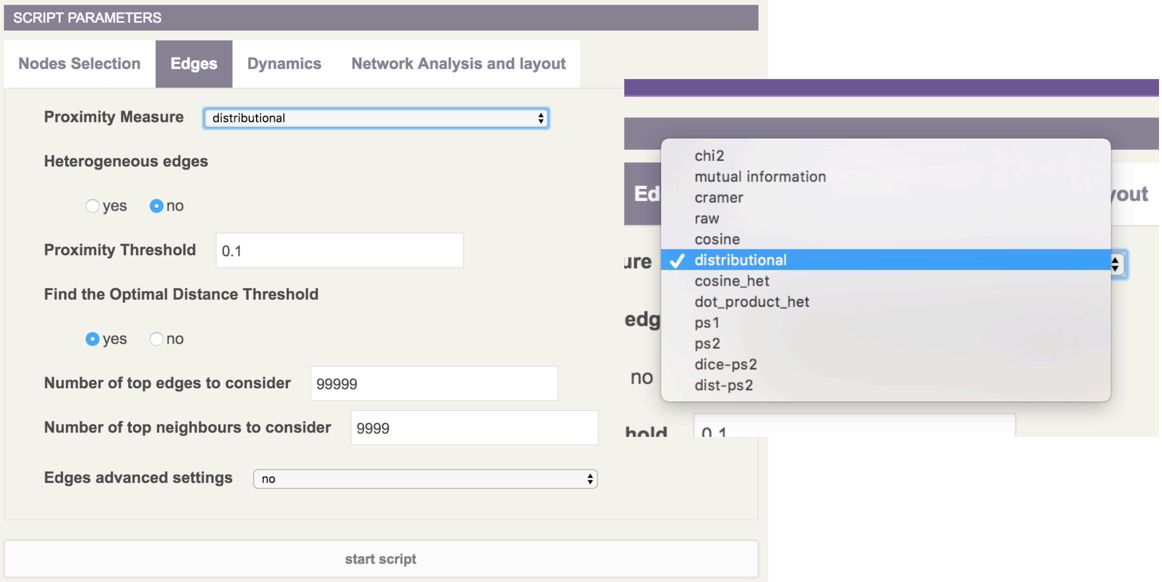
First Field: Author

Second Field: Author

Number of nodes: 50

Nodes advanced settings: no

Edges: Selection of proximity measure



Proximity Measure: distributional

Heterogeneous edges

○ yes ● no

Proximity Threshold: 0.1

Find the Optimal Distance Threshold

● yes ○ no

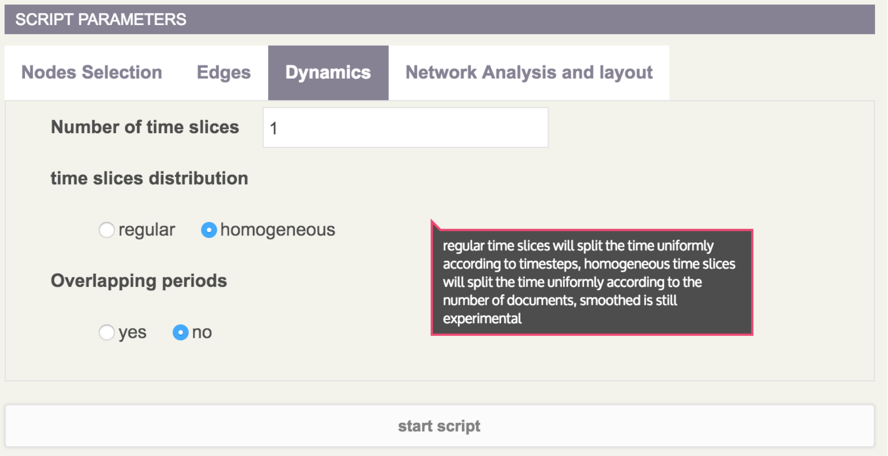
Number of top edges to consider: 99999

Number of top neighbours to consider: 9999

Edges advanced settings: no

chi2
mutual information
cramer
raw
cosine
✓ distributional
cosine_het
dot_product_het
ps1
ps2
dice-ps2
dist-ps2

Dynamics: parameterizing temporality



Number of time slices: 1

time slices distribution

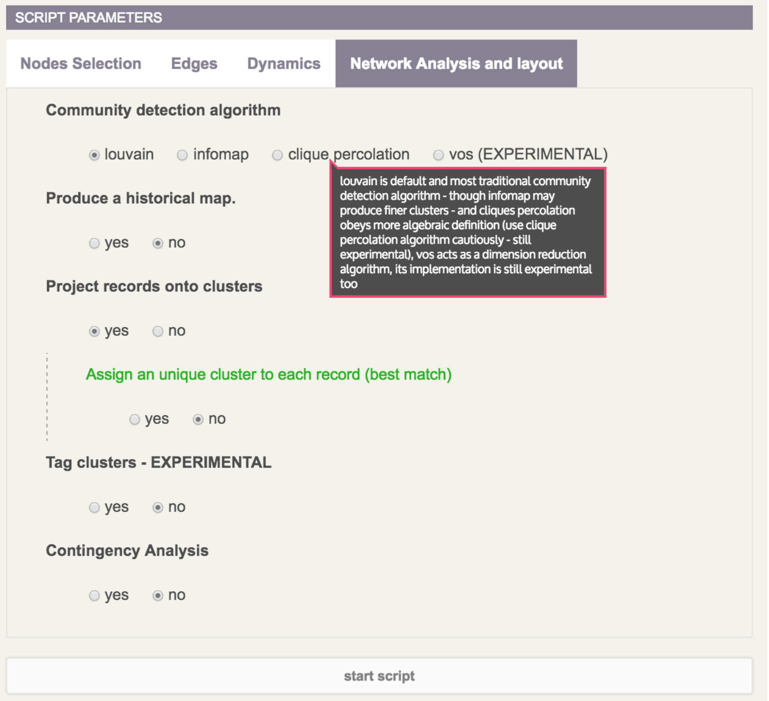
○ regular ● homogeneous

Overlapping periods

○ yes ● no

regular time slices will split the time uniformly according to timesteps, homogeneous time slices will split the time uniformly according to the number of documents, smoothed is still experimental

Network Analysis Layout



Community detection algorithm

● louvain ○ infomap ○ clique percolation ○ vos (EXPERIMENTAL)

louvain is default and most traditional community detection algorithm - though infomap may produce finer clusters - and cliques percolation obeys more algebraic definition (use clique percolation algorithm cautiously - still experimental), vos acts as a dimension reduction algorithm, its implementation is still experimental too

Produce a historical map.

○ yes ● no

Project records onto clusters

● yes ○ no

Assign an unique cluster to each record (best match)

○ yes ● no

Tag clusters - EXPERIMENTAL

○ yes ● no

Contingency Analysis

○ yes ● no

Heterogeneous Network

The underlying rationale behind this analysis is to use a very systematic and symmetric perspective that allows users to produce heterogeneous networks featuring any couple of nodes types. One can mix any two fields: for example from an ISI extraction: authors and their keywords, journals of publication and the journals they cite (journal level inter-citation network), extracted terms and cited journals, countries and cited references, years and terms, etc. Of course it is also possible (and even advised in the first place!) to produce traditional homogeneous networks like: co-authorship, co-words, or co-citation networks. To select the fields you wish to map, simply select the fields (first field and second field) in the node selection tab. The number of nodes to be mapped can also be easily tuned. For computational reasons, the maximal number of nodes is still limited to 500.

The nodes are selected according to their frequency at each time period. For example when mapping a co-authorship network, choosing 50 top items will produce the collaboration network between the 50 most productive authors (in terms of articles production) at each time period. For a research lab vs. keywords map, 50 most productive research labs will be mapped along with the 50 most frequent keywords.

Advanced Fields

Normalization – The top N items are not based on their overall frequency (which is the default behavior). If time dependent is checked, for each time step (typically year) the proportion of articles mentioning every item is computed. Those proportions are added over all the possible time steps resulting in a score that allows to rank all the entities and select the top N nodes, which will be mapped

Map exactly the same set of nodes at each time step – Fixes the set of nodes (only pertinent when several time periods are set). By default, N most frequent entities will be selected at each time period, likely resulting in only partially overlapping nodes being mapped. If you check this option, the N most frequent nodes over the whole time period will be selected.

Freeze nodes positions when computing temporal maps – When computing successive maps, nodes positions are recomputed at each time step by default. Select this option if you prefer to freeze their position starting from the ones computed during the first time period.

Hide isolated nodes – By default, nodes with degree zero are not shown, you can deactivate this option to reveal them (they should then appear on the left side of the map)

Show labels – Names of nodes will be shown or not

Shorten labels – Names of nodes are reduced to set size

Node size scale with their weight – The number of edges of a node determines its size

Node weight – Node weight defined by co-occurrence sum or frequency

Defining time periods

Heterogeneous mapping script also enables users to divide a corpus according to their need. One can work with custom initial time ranges (as defined by period slicer script) or with standard periods (typically a yearly defined corpus). You are then asked in how many slices you wish to divide your corpus (number of time slices) and if you want that the different sub-parts to be equal regarding the number of years (regular) or regarding the number of documents (homogeneous). Each time period will be assigned a sub-corpus gathering all documents produced during this time range from which a map will be built.

By default, periods are sliced to produce a partition of the corpuses overall timespan. However when overlapping periods is activated, time slices intersections are not empty anymore. One time period will intersect with its preceding and succeeding time periods such that half of the time steps or half of the records (if homogeneous slicing has been selected) are shared with either of them.

Sequencing

Whatever the slicing strategy, if several time periods have been selected, the script should output both the set of maps drawn at each time period (click the maps directory in the results pages) and the tube layout representation of clusters transformations in time (click the tube directory in the results pages).

The “tube layout” representation is fully interactive. The width of tubes is proportional to their number of records. Darker tubes are also more robust (more nodes are shared between two consecutive time periods). Simply hover over a cluster to learn about its composition. Note that colors in tube layout are consistent with the ones used for coloring maps.

NETWORK ANALYSIS & LAYOUT

Clusters Detection Method:

The Heterogeneous mapping script automatically identifies locally dense groups of nodes in the network. Different definitions/algorithm of these “communities of nodes” are possible. Users can choose between three popular ways to compute these meso-level structures: Louvain (Blondel et al. 2008), Infomap (Rosvall et al. 2008), clique percolation (Palla et al. 2005). Louvain is the most popular algorithm, while Infomap may succeed in detecting finer-grained communities. Clique percolation’s main advantage is to be interpretable as an algebraic property even though it will tend to exclude poorly connected nodes.

When mapping networks, two options are available to define nodes abscissa (x coordinate). By default, nodes are spatialized in 2d and take positions that optimize the stress produced by network topology (typically two nodes are attracted when connected by a link, the force being proportional to edge weight).

But one can also choose to fix x-position of node according to their “date”. For example, cited references will be positioned according to their publication dates – the network layout is then solely optimized according to the y-axis. This option will produce historical maps such as the one illustrated above and produced by analyzing a corpus of publications about synthetic biology.

In other cases for which a “natural” date is not provided, historical maps are still possible but the “time” at which nodes are positioned then correspond to the date when their number of occurrences reaches 20% of their total frequency over the whole dataset.

Others and Advanced

Project records onto clusters – by default, once the cluster structure of the map has been determined, every article is matched against each cluster composition to assess how close their content are. A document may then be assigned to zero or several clusters at once. Additionally, a new table (whose name starts with “projection_cluster” followed by the field name) will be created in the database. This new table can be used as a new field for further analysis.

Assign a unique cluster to each record (best match) – This option can be activated to limit the maximum number of clusters assigned to each document to 1. The most similar cluster

to a given document is then selected, as long as this similarity is higher than a predetermined threshold (meaning that some documents may still stay unassigned)
Add information from a 3rd variable – This option will produce tags associated to each cluster according to a new dimension in the dataset (to be chosen). A tagging metric should be chosen. Only top N closest tags will appear on the final map (N being an option to be defined in the form). Tagging option is equivalent to computing a new network onto the different clusters that have been identified. Put differently, a heterogeneous network is computed between the cluster field and a second chosen one. For instance, one can compute a journal co-citation network and then tag them with countries (see illustration below). Articles are then projected onto these clusters that become a new kind of variable (field 1). A proximity network between those semantic clusters and institution field can then be computed.   Options are tf (in which institutions are proportionally the most present in the cluster), raw (which is equivalent), chi2 (a chi2 proximity measure is computed between semantic clusters and institutions), Cramer, mutual information (see the description of metrics for more information). Other metrics proposed are rather designed for heatmaps.

Heatmap – Heatmaps allow to overlay on a given network(let's say a co-citation network on Facebook: domains are linked when they are oftentimes being shared by the same users) the distribution of presence of an entity taken from a different field (for example the gender of the Facebook user). Both the new field and the variable have to be indicated (in our case male). The algorithm computes for every node on the map its specificity with this modality: are men more likely to share links about 9gag.tv ? Possible metrics are the same than the one available for tagging clusters, plus chi2_dir, cramer_dir and cool deviation that are more useful in this setting. Chi2_dir and Cramer_dir correspond to the classic chi2 and cramer measures except they will also allow the user to observe negative correlations, generating a blue area in the final visualization. Cooc_deviation also measures how distant the number of citations of 9gag.tv by men is from what it should have been if this domain was uniformly distributed among men and women (relatively to their respective numbers). If positive let's say 2: this means that the number of citations by men is twice what should be expected. If negative let's say – 2: it means twice the number of citations of 9gag.tv by men would have been needed to reach its expected theoretical number. The final visualization averages the different specificity scores measured at each point of the network to produce a heatmap. Note finally that you can compute heatmaps over time. The background map will not change and shall still depend on the dynamical settings set in panel 3. But the distribution of the variable plotted on the heatmap will depend on the time range you chose (one needs to define a time period over which successive heatmaps will be computed)
Replace circles with alpha-shapes – This option changes the final layout of the communities around nodes. Instead of circles whose sizes are proportional to the number of articles assigned to a given cluster (if this option was activated), alpha-shapes are drawn for a more "organic" outcome.
Automatic Intertemporal Threshold – This refers to the threshold value used to create inter-temporal links when constructing river networks (tubes). This threshold is computed such that the total number of bifurcation in the final river network scales with the square root of the number of clusters overall (in all time ranges). Nevertheless one can manually tweak the parameter to only consider stronger or weaker links connecting temporally successive clusters.

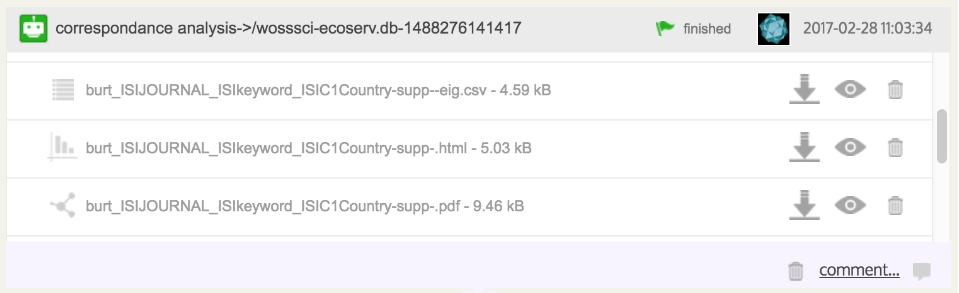
Small cluster Embedding in the river network – A special procedure absorbs smaller clusters in the river network that may tend to stay isolated otherwise. This is not part of the original algorithm described in (Rule et al, 2015) but it still gives good empirical results.
Hide orphan clusters in the phylogeny – By default every cluster is shown, even isolated in the river network. If one changes this option to yes, only dynamically connected clusters will appear in the tube layout and disconnected clusters will be colored grey in every map

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech, 10008.
Martin Rosvall, & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences of the United States of America, 105(4), 1118–1123.
Palla, G., Derenyi, I., Farkas, I. J., & Vicsek, T. A. (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435, 814.
Rule, A., Cointet, J. P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. Proceedings of the National Academy of Sciences, 112(35), 10837-10844.

CORRESPONDANCE ANALYSIS

Based on a resource of the R-project Library Factominer, Multi Correspondence Analysis script has been implemented in the RISIS infrastructure. Users can choose a set of categorical fields that will be used as active fields to perform the correspondence analysis, on top of which supplementary fields can be added.



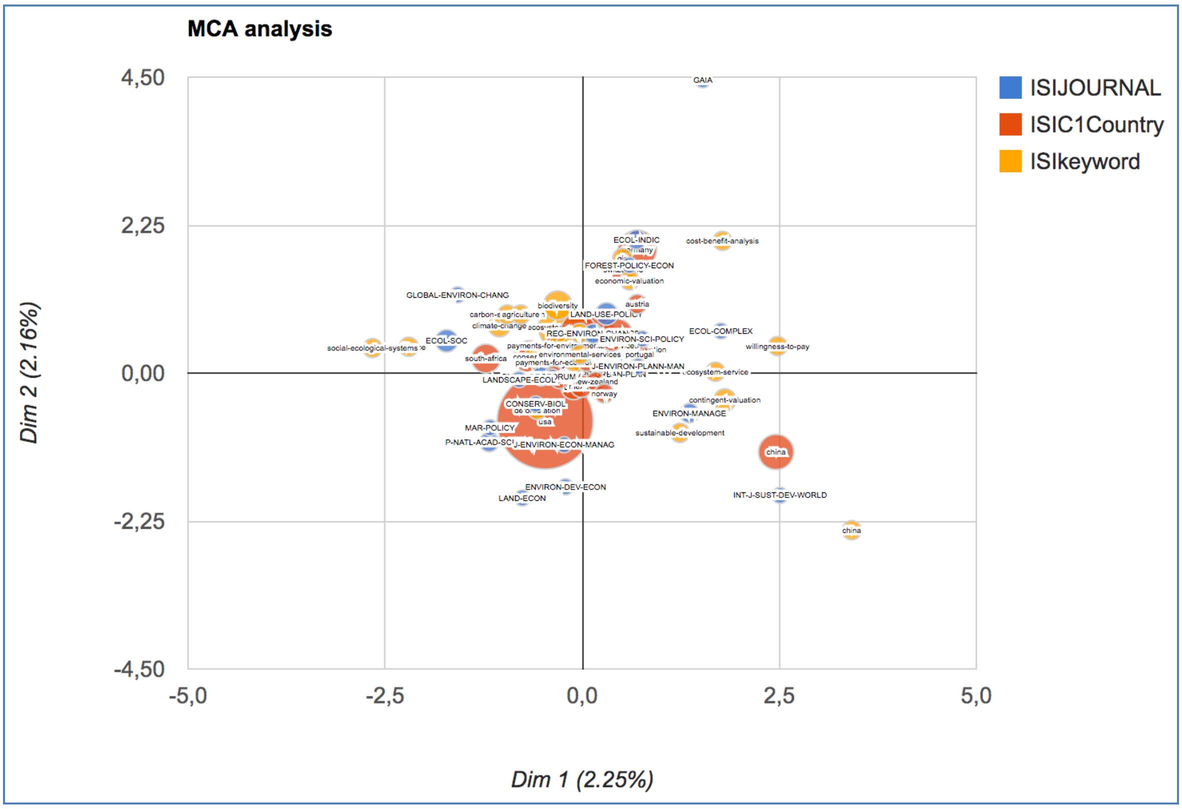
correspondance analysis->/wosssci-ecoserv.db-1488276141417 finished 2017-02-28 11:03:34

burt_ISIJOURNAL_ISIkeyword_ISIC1Country-supp--eig.csv - 4.59 kB

burt_ISIJOURNAL_ISIkeyword_ISIC1Country-supp-.html - 5.03 kB

burt_ISIJOURNAL_ISIkeyword_ISIC1Country-supp-.pdf - 9.46 kB

Interactive bubble charts (powered by Google charts) allow to visually analyzing the positioning of modalities in the three main dimensions of factorial space. The representation is also available as a simpler pdf file. Additionally Burt tables which describe the original data before Correspondence Analysis is performed are delivered and can be used for further use in a third party dedicated software.
The correspondence analysis script is quite straightforward as such but it is also designed to give users a way to compare recent network-based maps with the older tradition of geometrical statistics that eventually share the same rationale about the critical role of context.


MCA analysis
ISIJOURNAL
ISIC1Country
ISIkeyword
4,50
2,25
0,00
-2,25
-4,50
-5,0
-2,5
0,0
2,5
5,0
Dim 2 (2.16%)
Dim 1 (2.25%)
GAIA
ECOL-INDIC
cost-benefit-analysis
FOREST-POLICY-ECON
economic-valuation
GLOBAL-ENVIRON-CHANG
biodiversity
austria
LAND-USE-POLICY
carbon-s agriculture
climate-change
REG-ENVIRON-CHANGE
ECOL-COMPLEX
ENVIRON-SCI-POLICY
ECOL-SOC
social-ecological-systems
willingness-to-pay
payments-for-environmental-services
environmental-services
portugal
south-africa
payments-for-ecosystem-services
J-ENVIRON-PLANN-MAN
ecosystem-service
LANDSCAPE-ECOL
new-zealand
norway
contingent-valuation
CONSERV-BIOL
deforestation
usa
ENVIRON-MANAGE
MAR-POLICY
sustainable-development
P-NATL-ACAD-SCI
J-ENVIRON-ECON-MANAG
china
ENVIRON-DEV-ECON
LAND-ECON
INT-J-SUST-DEV-WORLD
china