# Digital Diasporas Atlas
# Exploration and Cartography of Diasporas in Digital Networks

**Dana Diminescu, Mehdi Bourgeois, Matthieu Renault**
Telecom ParisTech, 46 rue Barrault 75013 Paris
FMSH, 190 avenue de France 75013 Paris
dana.diminescu@telecom-paristech.fr, mehdi.bourgeois@gmail.com, matthieu.renault@gmail.com

**Mathieu Jacomy**
Médialab Sciences Po, 13 rue de l'université
75007 Paris
FMSH, 190 avenue de France 75013 Paris
mathieu.jacomy@gmail.com

## Abstract

We will present the digital methodological chain and the tools we developed for building the Digital Diasporas Atlas which aims at mapping and analyzing the occupation of the web by diasporas. Such a chain is composed of four intertwined steps: 1) equipped web exploration and corpus building; 2) data enrichment (location, languages, text-mining); 3) network visualization-manipulation and graph interpretation; 4) collaborative sharing of (raw) data and findings. The Digital Diasporas Atlas takes part in the project of introducing digital methods in humanities: promoting digitally equipped social sciences and developing an engineering informed by humanities.

## Introduction to Digital Diasporas

The Digital Diasporas Atlas project aims at mapping and analyzing the occupation (in a quasi-geopolitical sense) of digital territories by the "connected migrants" (Diminescu, 2008). It is grounded on the assumption that studying the web involves a *commitment to technology*: the digital matters not only as object/field of investigation but also and inseparably as means/tool of research. That is why the design and development of digital methods (not apart but in continuity with the "traditional" methods of social sciences) is a core issue in the building of the atlas. Social and political scientists are permanently involved in the conception, development and improvement of such methods and tools. Indeed, one of the crucial goals of the project is to generate an "interface" between social sciences and computer sciences: in other words, to promote *digitally equipped social sciences* as well as an engineering informed by humanities.

## Methodological Chain

### Step 1: Web Exploration (*Navicrawler*)

In the making of a chapter of the Atlas, the first step is the building (and circumscription) of a *corpus* of websites. The researcher plays a crucial role in this process inasmuch as his knowledge of the fieldwork allows him to discriminate the relevant resources for a given diaspora.

In order to complete this stage of *collection*, the researcher needs to be *equipped*. The identification of relevant websites is achieved semi-automatically thanks to a software called *Navicrawler* that allows to scan *web grounds* through a web-browser (see figure 2). Navicrawler is a Firefox add-on, designed and developed by our research team. The interface is located on the left of the currently browsed page.

The essential functionality of the Navicrawler consists in scraping the out-links of the visited websites (listed and stored as "Next Sites") For each website, the researcher can incorporate it to the corpus (it becomes "In Site") or reject it ("Out Site"). He can also describe the websites by adding tags.

The logic of exploration induced by the Navicrawler is located at the crossroads of browsing and crawling. Unlike automatic crawling, it allows the researcher to perceive the *context of links* (Ghitalla, 2008) and thus to avoid a blackbox effect. At the end of this exploration stage, he is able to export his corpus as a graph in which the nodes represent the websites and the edges stand for the links between them.

### Step 2: Data Enrichment (*Digital Toolbox*)

As already mentioned, the social scientist plays a central role all along the process of corpus building and description/enrichment. However, he can be assisted in the content analysis by automatic tools. Our research team developed a *digital toolbox* that renders possible various "enrichment processes", among which:

- Retrieving from a list of urls the information provided by the registrar (ICANN); about the registrant (owner of the domain name), especially his *geographical location*; about the server hosting the website, etc..
- Text-mining applied on the index of the corpus in order to retrieve *named entities*: persons, organizations, places, etc. (using Open Calais API).

- Recognition of the *languages* used in each website (and the distribution of languages in order to study multilinguism, an important issue in migration studies).

### Step 3: Network Visualization (*Gephi*)

In order to visualize the exploration data, in other words to map the corpus previously built, we use a software of graph visualization named *Gephi*, a project initiated and hosted at its beginnings by our research team. This tool allows to spatialize and manipulate the corpus network. Two types of visualization are available:

- a spatialization based on the physical principle of attraction/repulsion (according to the presence or absence of a link between two nodes)
- a geographical spatialization that uses geocoded data: location of websites' owner, websites' addressees, servers, etc. (especially information retrieved during the data enrichment stage)

By handling the network, by observing its evolution (timeline), by visualizing the place and the connections of a given website, by identifying clusters, by filtering the data by categories, in brief by *interpreting* the graph, the researcher produces various representations (or views) of the corpus that allow him to formulate *hypothesis* of research that will be supported (or not) by other online/offline fieldwork investigations.

### Step 4: Collaborative Platform (*Gexf Atlas*)

The Gexf Atlas is a "generic" *collaborative* platform first developed and implemented for hosting the digital diasporas atlas (see Figure 2). It is a tool for publishing and sharing research findings among scientific communities. It is composed of *chapters* (in our case, the various diasporas) and provides for each of them the following data:

- *Maps*: browsable graphs of the corpus, with different views according to the fields of classification.
- *Raw data*: the empirical data (texts, videos, images interviews, etc.) produced/retrieved and used during the research. The Digital Diasporas Atlas is part of the more general "digital humanities" project of providing access and diffusing not only the research results but also the research data.
- *Statistics* : they are automatically generated from both the classification and the graph structure and provide quantitative data about the relations between categories/actors. Statistics help strengthening the hypothesis formulated from the graph visualization.

Let us add that all the steps composing the methodological chain are fundamentally intertwined and applied recursively, in such a way that one could talk about a *circular methodology*. At the "exit" of the circle, the corpus of a digital diaspora is dynamically *archived* in collaboration with one of our research partner.

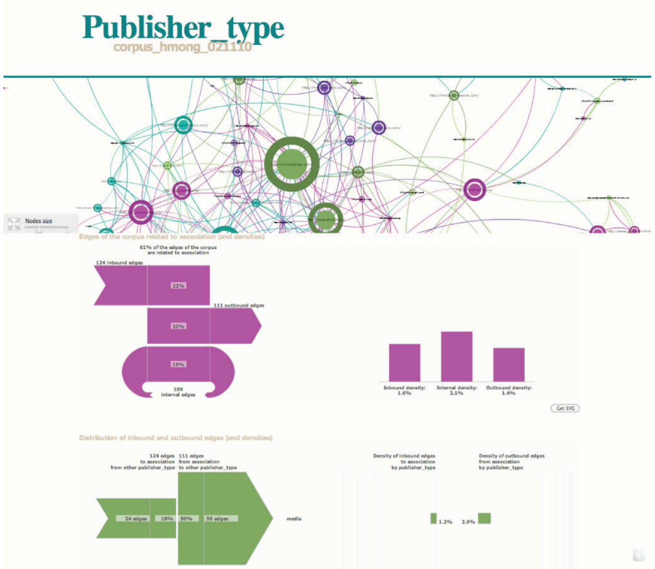
Publisher_type
corpus_hmong_021110
Nodes size
Edges of the corpus related to association (and densities)
61% of the edges of the corpus are related to association
124 inbound edges
22%
111 outbound edges
20%
19%
109 internal edges
Inbound density: 1.6%
Internal density: 2.1%
Outbound density: 1.4%
Get SVG
Distribution of inbound and outbound edges (and densities)
124 edges to association from other publisher_type
111 edges from association to other publisher_type
24 edges
19%
50%
56 edges
media
Density of inbound edges to association by publisher_type
Density of outbound edges from association by publisher_type
1.2%
2.9%

Figure 2: Maps and statistics with the GEXF Atlas

## Work-in-Progress: Dealing with Social Media

The conception of the methodological chain hitherto presented takes its roots in the "web 1.0 age"… before the spreading of web 2.0 applications, especially among migrant communities. Now dealing with the social web, and especially social networking sites demands original methods.

Using *Facebook*'API, we developed a *Social explorer* that allows: 1) to store dynamically (settable frequency) the (egocentric) social graph of a user and its evolution. We especially use it to observe the changes in newly arrived migrants' social networks; 2) to store the social graph corresponding to a Facebook group, in order, notably, to identify "individuals-authorities", subgroups, etc. It helps us to reconsider the structuring of formal and informal migrants organizations. Such graphs are used as supports in traditional interviews.

## References

Diminescu, D., 2008, "The Connected Migrant: an Epistemological Manifesto", *Social Sciences Information*, vol 47, n°4.

Ghitalla, F., 2008, "La carte, Un media entre sémiotique et politique", *Communication et langages*, n°158, pp. 61-75.