Atlas IDF SI
PICRI Ritimo / WebAtlas

Compte-rendu d'expérience

Erika Campelo (RITIMO, Rinoceros)
Aude Delavernhe (RITIMO, Rinoceros)
Franck Ghitalla (WebAtlas, UTC)
Suzanne Humberset (RITIMO, Rinoceros)
Mathieu Jacomy (WebAtlas, Telecom ParisTech)

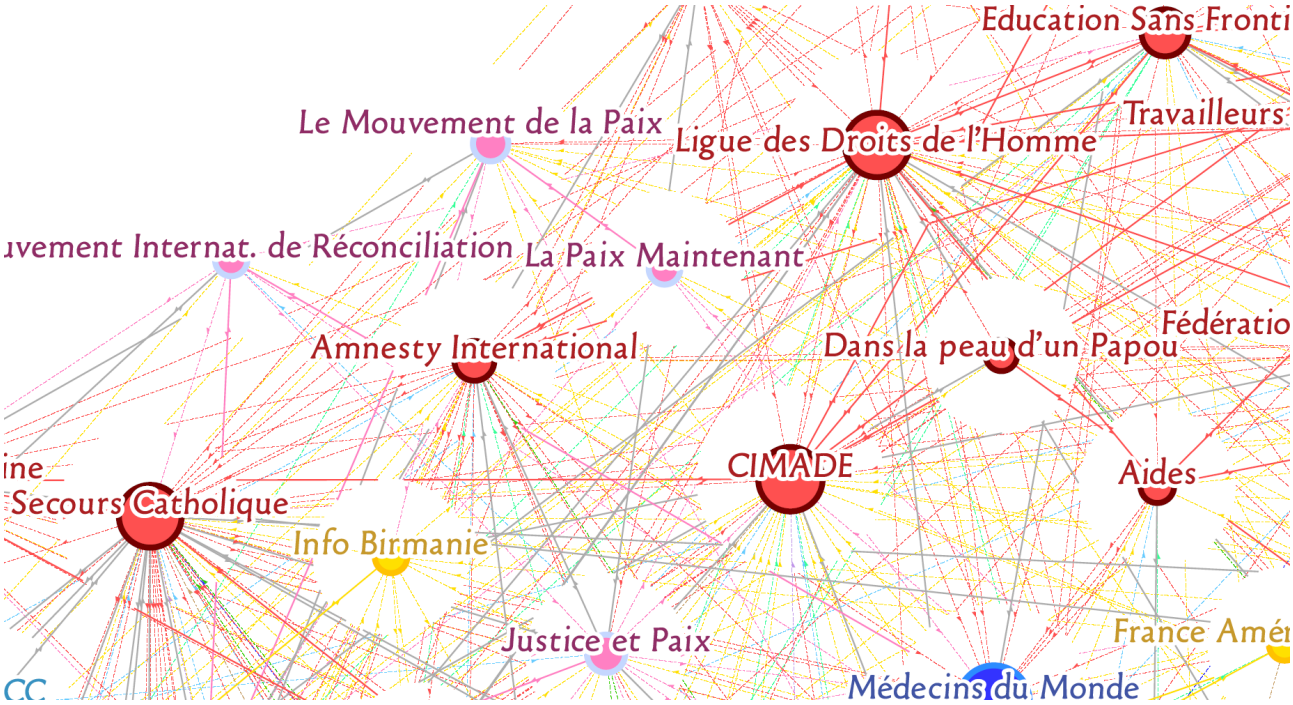


Table des matières

1. Introduction....3
2. Retour sur le projet initial....4
2.1. Rappel et mise en perspective de notre projet initial....4
2.2. L'évolution vers une autre façon de concevoir le projet....5
2.3. Chronologie du projet....7
3. L'Atlas : les documents produits....8
4. Méthode de réalisation de l'Atlas....12
4.1. Brève présentation des outils....12
4.2. Construction du corpus....12
4.2.1. Présentation générale....12
4.2.2. Points de départ....14
4.2.3. Première exploration (pré-corpus)....14
4.2.4. Seconde exploration (circonscription du corpus)....17
4.2.5. Qualification....18
4.2.6. Crawls finaux....19
4.3. Construction des cartes....19
4.3.1. Spatialisation des graphes....19
4.3.2. Manifester les caractéristiques du corpus....20
4.3.3. Extraction des statistiques....20
4.3.4. Finalisation des cartes....21
5. Réalisations technologiques....22
5.1. Le Navicrawler....22
5.2. Gephi....23
6. Retombées et enseignements....26
6.1. Apports et intérêt du projet pour RITIMO....26
6.2. Communauté et associations....26
6.3. La transportabilité des données numériques complexes....28
6.4. En guise de conclusion....30
7. Suites à donner à ce projet....32

1. Introduction

Notre projet, financé par la région Île de France dans le cadre de son programme PICRI, est mené en collaboration par RITIMO, association porteuse du projet (Suzanne Humberset, Erika Campelo et Aude Delavernhe), et WebAtlas, association de recherche initialement installée à l'Université de Technologie de Compiègne sous le nom de RTGI[1] (Franck Ghitalla et Mathieu Jacomy). Vous pourrez trouver toutes informations nécessaires sur nos sites web ritimo.org et webatlas.fr !

L'objectif principal de ce projet est la constitution d'une bibliothèque numérique de notre milieu, du développement durable et solidaire (http://www.rinoceros.org/spip.php?article2506). Comme nous le verrons plus loin, ce n'est pas ce à quoi nous sommes arrivés. Notre principale production, c'est un atlas de la solidarité internationale et du développement durable sur le web francilien, que nous raccourcirons en « Atlas IDF SI ».

Cette étude est le fruit d'une longue, expérimentale, et minutieuse étude du web à plusieurs mains. Et les questions qu'elle pose sont les suivantes : Est-ce que le web correspond à l'image que l'on se fait du milieu de la solidarité internationale ? Quelle est la meilleure stratégie d'occupation du web et comment changer sa place sur la toile quand on est dans ce milieu ?

Notre projet a requis la conception d'outils (et des méthodes scientifiques qui vont avec). Il a également été nécessaire de former les « explorateurs du web » à ces outils, et de mener cette exploration, cette construction d'un corpus.

A l'issue de ce projet nous avons restitué ce travail, et notamment les grandes cartes du web que nous avons produites, aux acteurs de ce milieu et nous avons recueilli leurs réactions quant à la possibilité de s'emparer à plusieurs de la question du web comme territoire informationnel commun.

Mais commençons par le commencement...

1 Une partie de l'équipe RTGI a décidé en 2006 de fonder une entreprise. Il a alors été décidé de céder le nom RTGI à l'entreprise pour que le capital symbolique du groupe puisse l'aider dans sa difficile démarche. C'est ainsi que la « branche recherche », qui a mené ce projet avec RITIMO, a été renommée en « WebAtlas ».

2. Retour sur le projet initial

2.1. Rappel et mise en perspective de notre projet initial

Le concept central de notre proposition initiale, c'est la **bibliothèque numérique**. De quoi s'agit-il ?

Posons le décor. Le web est un vaste océan informationnel dans lequel il est difficile de s'orienter, et la bibliothèque numérique a notamment pour but de pallier à cette désorientation. Si les internautes ont des difficultés à se représenter le web, c'est aussi le cas des personnes qui réalisent des sites web. Peut-on s'installer dans une maison sans s'intéresser à son quartier ? Pourtant on construit bien des sites web sans prendre en compte ceux qui existent déjà, qui vont donner un contexte. Le web est un territoire que l'on aménage, consciemment ou non, collectivement ou individuellement. La bibliothèque numérique est le lieu dédié à l'aménagement des territoires numériques.

La bibliothèque numérique idéale est faite de trois « briques » complémentaires :

1. Un **outil de cartographie du domaine**, capable de forger une représentation du domaine web pour ses acteurs et les personnes qui consultent l'information. Cette représentation, ou cartographie, est la condition première pour qu'il y ait une prise de conscience de l'existence d'un territoire commun.
2. Des **sessions de recherche d'information**. Les informations collectées peuvent être traduites en « répertoires visuels » et interrogées par des requêtes linguistiques classiques. La recherche d'informations répond à la nécessité d'une veille continue, à la façon d'un moteur comme Google que nous savons tous utiliser. La différence est ici que l'on ne cherche pas sur le web « en général », mais à l'intérieur du domaine. Les résultats sont donc plus pertinents ; les sources sont maîtrisées.
3. Un **centre communautaire**, c'est-à-dire un ensemble de fonctionnalités permettant de travailler ensemble. Le but est ici de partager les différentes visions d'un même domaine, d'évaluer la qualité des ressources, de décider d'une orientation. Le travail distribué proposé ici comme modèle s'inspire à l'origine des réseaux sociaux ou « web 2.0 ».

Comme nous le verrons au fil de ce document, ce découpage existe mais nous ne sommes pas parvenus au résultat escompté.

La cartographie du domaine a été notre principal axe de travail – c'est aussi le plus gros investissement en temps. Les outils nécessaires ont été développés et sont disponibles, mais ils constituent une boîte à outil plutôt qu'une boîte noire. Autrement dit, il ne suffit pas d'appuyer sur un bouton pour voir la carte sortir du logiciel. Une certaine maîtrise est nécessaire pour arriver à un résultat satisfaisant, maîtrise technique certes mais aussi maîtrise de l'information. Certaines associations sont spécialisées dans la gestion des ressources documentaires, notamment le réseau RITIMO. Les autres acteurs pourront-ils, pourraient-ils s'investir dans ce travail ?

Les sessions de recherche d'information sont permises par des technologies aujourd'hui bien connues. Pour autant, elles restent lourdes à mettre en place. D'une part cette brique n'est réellement performante que lorsqu'elle est produite et maintenue par un industriel, d'autre part différentes sociétés dont Google proposent gratuitement des services approchant. La question devient alors celle du lieu *physique* d'un centre de calcul (un « gros » PC serveur). Faut-il faire confiance à

Google pour indexer les ressources ? Dans le cas contraire, qui prendra en charge l'hébergement ? Ce questionnement aboutit inévitablement à négocier l'autonomie des associations pour gérer leur patrimoine informationnel ; malheureusement les alternatives et les enjeux sont encore mal connus de la plupart des acteurs (associatifs, mais aussi institutionnels).

Le centre communautaire quant à lui n'a pas vu le jour dans ce projet. Pour diverses raisons liées à la vie du projet, que nous verrons plus loin, il n'a pas dépassé le stade de prototype. Ceci ne nous a pas empêché de réfléchir de près à sa mise en place. Nous butons alors sur le fait que l'ensemble des acteurs ne constituent pas de communauté à proprement parler. Pour que le centre communautaire soit un succès, à quelle sorte de réseau social doit-il s'adresser ?

Au fil du projet nous avons affiné notre vision du web, des acteurs impliqués, et des différentes alternatives techniques. En fin de compte, l'interrogation à laquelle nous nous confrontons est la suivante : **quels outils sont réellement adaptés au mode de fonctionnement de la société civile ?**

2.2. L'évolution vers une autre façon de concevoir le projet

Dans le temps de ce projet, le web a sensiblement évolué. Nous avons assisté à la popularisation du désormais bien connu « web 2.0 ». Parallèlement nous avons appris à mieux nous connaître entre chercheurs et associatifs, à confronter nos différentes façons de travailler. Ces changements nous ont poussé à distinguer le travail d'*expert de l'information* et le travail de *mise à disposition des résultats* auprès de la communauté ; le premier volet étant le plus stratégique et le second le plus problématique.

Cette distinction de deux axes de travail tient en bonne partie à l'évolution du contexte, ainsi qu'au fait que le chemin à parcourir nous est apparu plus clairement lors de nos avancées. Pour le dire vite, comme nous construisions le contenu, sa mise en ligne nous est apparue de plus en plus complexe, surtout au vu des mutations du web lui-même.

Avec l'essor du « web 2.0 » une culture du travail distribué a émergé, et de nouveaux outils sont apparus. A l'heure actuelle on trouve notamment des agrégateurs de flux RSS publics, la possibilité d'éditer un moteur « vertical » limité à un corpus (chez Google), et de nombreuses solutions pour implémenter une structure de réseau social, d'hébergement de contenus (Baagz chez Exalead), de wiki... L'avantage de ces outils, c'est qu'ils sont bien intégrés à des systèmes éprouvés, comme Flickr, Myspace ou Google. Leur inconvénient est qu'ils ne sont pas intégrés entre eux, et que la maîtrise de l'information est souvent très limitée (Google n'indexe pas très bien, contrairement à ce qu'on pourrait croire). La question de l'autonomie du système (donc la maîtrise de l'information) et de son *lieu* (virtuel comme physique) se pose toujours, et la bibliothèque numérique est toujours une alternative pertinente. Mais il s'avère nécessaire de rediscuter les principes de conception.

Les associations ont également développé leur capacité à investir les territoires numériques. Certains projets sont à l'initiative d'acteurs émergents sur le web, comme le réseau social dédié aux associations « Ammado »[2] (peu convaincant). Mais la plupart sont à l'initiative d'acteurs déjà bien installés. Citons au plus proche de notre projet le portail « dph »[3] animé par l'équipe Rinoceros à Ritimo, mais aussi un moteur de recherche inter-ressources (Scrutari), des annuaires d'acteurs en ligne... Il s'agit aussi bien de dispositifs originaux conçus pour une fonction précise que d'une aide aux acteurs à s'installer sur le web (formations aux outils du libre, accompagnement pour mettre son site en ligne, à l'aide de « Spip » par exemple). L'intérêt de ces démarches vis-à-vis de notre projet tient au fait qu'elles sont nées des besoins des acteurs, en accord avec leur savoir faire et leurs

2 http://www.ammado.com
3 http://www.d-p-h.info

méthodes de travail, dans la logique de gestion de l'information qui leur est propre (et qui n'est pas la même pour tous). Les projets qui rencontrent le succès ne sont pas les projets « communautaires », ce sont toujours des initiatives locales. Faut-il y voir un signe que l'ensemble des acteurs ne se vit pas, ne veut pas, ou ne peut pas se vivre comme une communauté ? Il en ressort que soit une bibliothèque numérique n'est pas viable, soit elle doit être au contraire le vecteur d'une conscience communautaire, si c'est la direction que souhaite prendre tout ou partie des acteurs.

Par rapport à notre position initiale, nous pensons que l'intégration et la dimension communautaire de la bibliothèque numérique font problème. Pourtant, la question de la maîtrise de l'espace documentaire se pose toujours, si ce n'est plus fortement. L'enjeu le plus important devient alors de **poser la question du territoire** en de bons termes aux acteurs impliqués. C'est un préalable à toute **discussion sur le rôle d'une bibliothèque numérique**.

Parce que ce projet est innovant, parce que ses résultats sont d'un genre nouveau et notamment pour les acteurs de la société civile, il faut avoir *quelque chose à montrer* pour sensibiliser les acteurs à la problématique. Malgré son statut de « tête de réseau » Ritimo ne représente pas la multitude et la multiplicité des acteurs concernés. Son rôle à cet égard est celui d'un pilote, d'une figure exemplaire, d'un relais. Les acteurs de la société civile n'ont pas de représentant désigné, d'institution centrale, de tête de réseau globale. Et il n'est pas question d'y changer quoi que ce soit ! Mais c'est ce qui rend la question du *lieu* de la bibliothèque numérique si difficile, et si cruciale à la fois. La nécessité s'est fait sentir de convaincre, de susciter l'adhésion des acteurs périphériques à notre projet, et de ne pas proposer comme nous l'imaginions une solution « clefs en main ». Et pour engager ce débat, l'enjeu le plus stratégique est la réalisation du contenu susceptible d'alimenter un dispositif tel qu'une bibliothèque numérique. Pour ce travail Ritimo a toute légitimité, puisque son rôle auprès des autres acteurs est notamment de gérer l'information.

Autrement dit notre investissement s'est trouvé limité par le point où l'implication des autres acteurs est indispensable, et ce point est atteint dès la conception d'un dispositif de mise en ligne des résultats. La bibliothèque numérique ne peut exister qu'à la condition d'une bonne implication des associations, c'est-à-dire en réponse à des besoins qui ne sont pas (encore) formulés collectivement. Nous avons choisi d'éviter de proposer un dispositif « tout fait », allant à l'encontre de l'émergence d'une réflexion commune. Mais nous avons réalisé le travail nécessaire pour poser ce débat, et nous l'avons engagé, quitte à proposer des solutions techniques « en kit », des alternatives encore ouvertes et complémentaires même si elles sont parcellaires.

Cette réflexion que nous avons menée chemin faisant a donc abouti à une reformulation des objectifs de notre projet, au phasage en deux axes du travail évoqué plus haut.

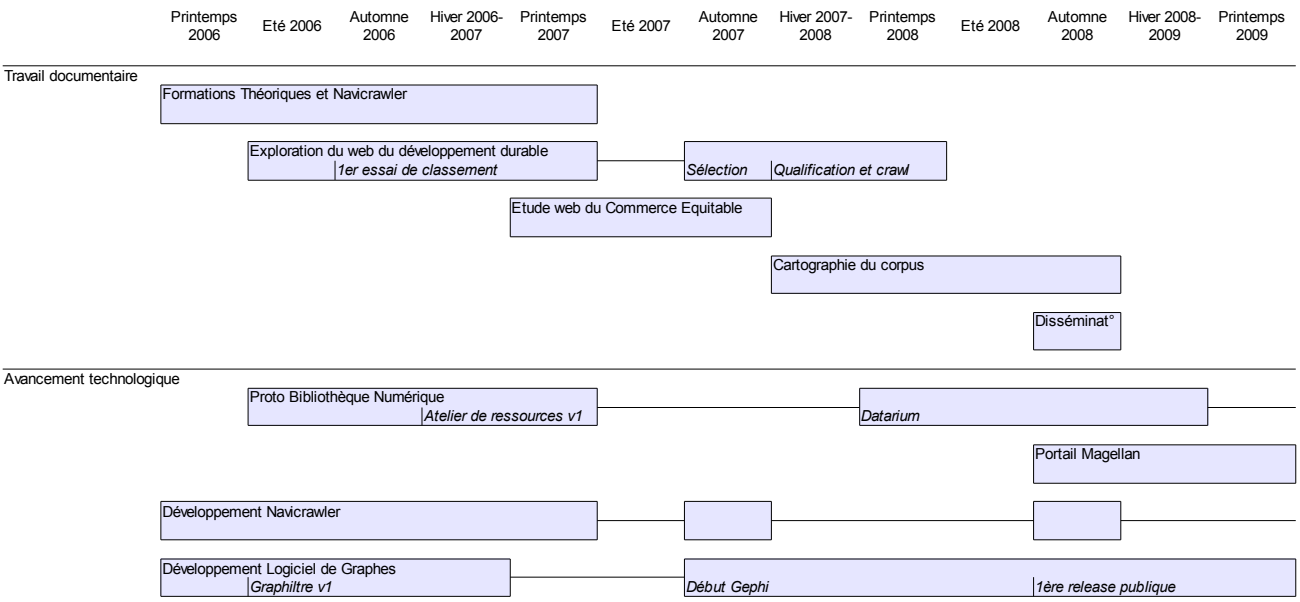
1. **Construire une représentation du territoire numérique**. Collecter les données nécessaires à la construction d'une bibliothèque numérique, les exploiter pour construire des cartographies et des analyses statistiques, et confronter ces résultats aux acteurs impliqués. Ce travail de longue haleine est l'investissement initial indispensable à toute plateforme collective distribuée ; il comprend la réalisation d'outils et de méthodes scientifiques adaptés. Mais il est également un résultat en tant que tel, et les données sont également exploitables par les acteurs selon leurs objectifs propres. La dissémination de ce travail est bénéfique pour le débat sur la maîtrise collective de l'information, car elle pose la question du territoire commun.
2. **Proposer plusieurs solutions techniques pour rendre ces données utiles au collectif**. Nous avons bien sûr la vision d'une bibliothèque numérique, mais d'autres alternatives sont possibles et nous souhaitons voir les acteurs s'impliquer dans le débat. En particulier différents degrés d'engagement collectif sont possibles, et si la bibliothèque numérique est trop communautaire des options plus distribuées existent. C'est au final

une question politique, parce qu'elle touche à une structuration associant la centralité (le lieu physique et symbolique, et l'entité qui le gère) à un régime de collaboration distribué (réseau social « web 2.0 ») ce qui ne fait pas partie des interactions traditionnelles entre acteurs. Notre rôle est alors d'animer le débat, de sensibiliser les acteurs aux enjeux et aux alternatives techniques, et d'amener ainsi un questionnement collectif plus riche sur le devenir commun des acteurs (sur le web, mais aussi au-delà).

2.3. Chronologie du projet

Nous avons séparé la chronologie en deux parties. En haut se trouvent les principales tâches dédiées au travail sur les données. En bas se trouvent les principales tâches dédiées au développement des outils.

Ce graphique n'a pas vocation à donner une vision d'ensemble du projet. Il est simplement proposé pour avoir une idée du rythme et de l'enchaînement de nos avancées. Pour une description plus précise, vous pouvez vous reporter à la partie méthodologique qui détaille notre cheminement.



Chronologie du projet

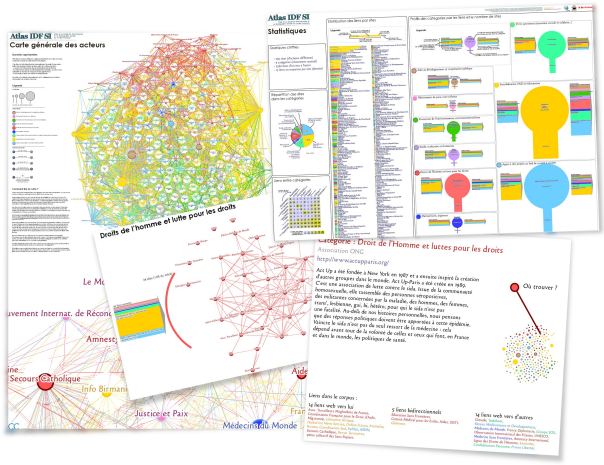
3. L'Atlas : les documents produits

L'étude du domaine de la solidarité internationale et du développement durable sur le web francilien se présente sous la forme d'un **atlas**. C'est-à-dire un recueil de cartographies associées à d'autres documents, comme autant de « tranches » qui composent des vues complémentaires sur un même objet. Ces différents documents sont tour à tour synthétiques ou analytiques, génériques ou détaillés, plutôt graphiques ou plutôt textuels. C'est un atlas construit de la même façon que celui de Mercator (extraits ci-dessous), sauf qu'au lieu de montrer le territoire géographique, il montre le territoire numérique.

Les données représentées sont issues de notre travail de *collecte*, de *sélection* et de *description* du corpus (le détail de notre méthode est présenté dans la section suivante). Elles sont organisées de la sorte :

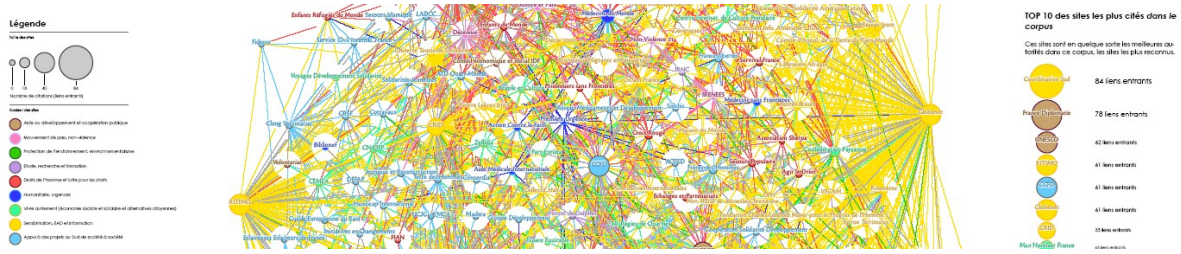
- **259 sites-acteurs** du domaine, et les **3360 liens** entre eux sur le web
- **9 catégories** thématiques dans lesquelles sont classés tous les sites-acteurs
- Une **description de chaque acteur**, sa raison sociale, ses liens et sa catégorie thématique.

Selon qu'on s'intéresse au panorama de l'ensemble, à une catégorie ou à un acteur donné, aux liens hypertextes ou à des distributions statistiques, des représentations graphiques différentes sont pertinentes. Notre atlas permet de répondre à la plupart de ces questions en proposant plusieurs posters (grandes cartographies) et un cahier des acteurs.



Voici quels sont les différents documents qui composent l'Atlas de la Solidarité Internationale et du Développement Durable sur le web francilien :

- **La carte générale des acteurs**



Cette carte représente les 259 sites-acteurs et les 3360 liens entre eux. La carte est en fait un graphe (des ronds reliés par des traits) dont l'aspect graphique a été finement travaillé pour être le plus lisible possible, comme une carte traditionnelle.

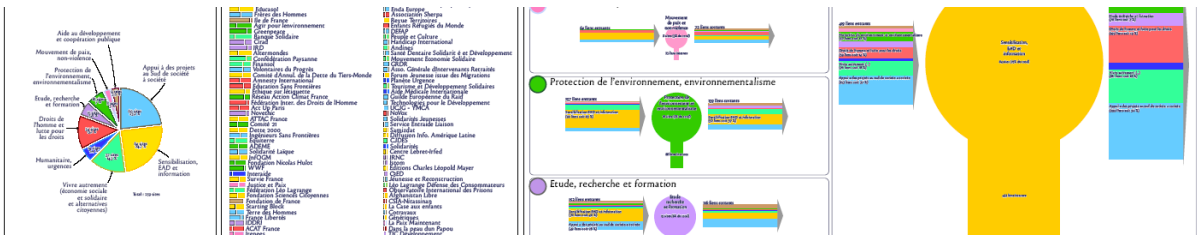
La couleur des sites-acteurs renvoie à leur catégorie thématique, et leur taille indique leur autorité dans le corpus.

La position de chaque site-acteur dépend uniquement des liens qu'il entretient avec les autres, indépendamment de sa catégorie thématique. Pourtant on peut voir des zones de même couleur : c'est que comme l'indique la *théorie des agrégats*, les sites du même thème tendent à se lier entre eux. Or le principe de placement des sites-acteurs fait que ceux qui sont bien liés entre eux se retrouvent dans la même zone de la carte.

Cette carte permet donc aussi bien d'observer la situation d'un site-acteur dans le domaine que d'avoir une idée de la structure des catégories, de leurs proportions et des liens qu'elles entretiennent.

Cette carte est le *panorama* du domaine. Les autres cartes sont consacrées à des aspects précis des données, et permettent d'obtenir des réponses plus précises aux questions que cette carte-ci peut susciter.

- **Le poster des statistiques**



Il ne s'agit pas à proprement parler d'une carte, mais d'un ensemble de représentations graphiques qui montrent différentes distributions statistiques. Néanmoins ces indicateurs ne se présentent pas sous la forme d'une seule collection de chiffres. Ils sont regroupés selon leur nature en représentations graphiques conçues spécifiquement pour être lues facilement, et conjointement avec les autres documents de l'atlas.
Ainsi, chaque élément statistique répond au code-couleur utilisé partout ailleurs pour les catégories (une couleur par catégorie).

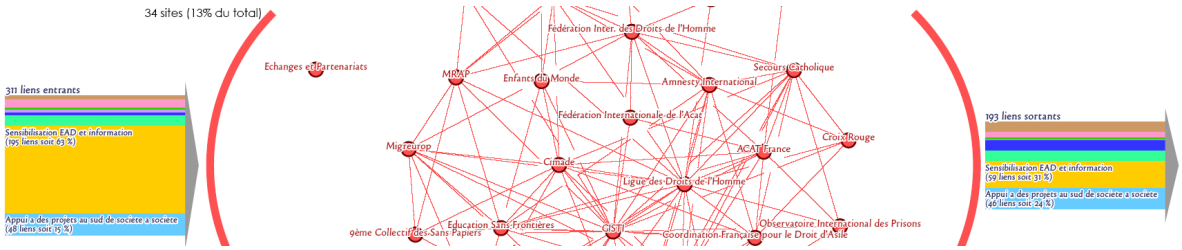
Une première série d'indicateurs décrit la répartition quantitative des acteurs dans les catégories (camembert), ainsi que la distribution des liens entre chaque paire de catégories (tableau coloré).

Une deuxième série d'indicateurs montre la distribution du nombre de liens par site-acteur, typique des réseaux sociaux comme du web. On voit qu'il y a une minorité de sites-acteurs qui regroupent une majorité des liens (dans le corpus) et une grande majorité de sites-acteurs qui ont peu de liens (dans le corpus). C'est une approche hiérarchique du corpus. On peut y distinguer les liens entrants et les liens sortants.

Enfin une troisième série d'indicateurs permet d'avoir une vue efficace de la morphologie de chaque catégorie, et de les comparer. Les indicateurs sont regroupés pour former une sorte « d'ampoule », qui est la signature graphique de la catégorie thématique. Plus le rond est grand et plus il y a de sites dans la catégorie, plus le pied est large et plus il y a de liens internes à la catégorie. De part et d'autre de cette forme se trouvent des flèches qui montrent la quantité de liens entrants, et leur distribution dans les autres catégories.

Ce poster permet de trouver facilement les catégories qui se ressemblent ou se distinguent des autres, et sur quels aspects exactement. On y voit typiquement que les sites « humanitaires » ne se citent pas les uns les autres, alors qu'ils sont très peu nombreux, au contraire des sites « Sensibilisation et EAD », qui bien que très nombreux, sont très liés les uns aux autres. On remarquera aussi, par exemple, que cette dernière catégorie est la seule qui a plus de liens vers les autres que de liens entrants.
Ces indicateurs statistiques traduisent des aspects de la carte générale sous forme de mesures plus pratiques à interpréter.

– **Les cartes de chaque catégorie**

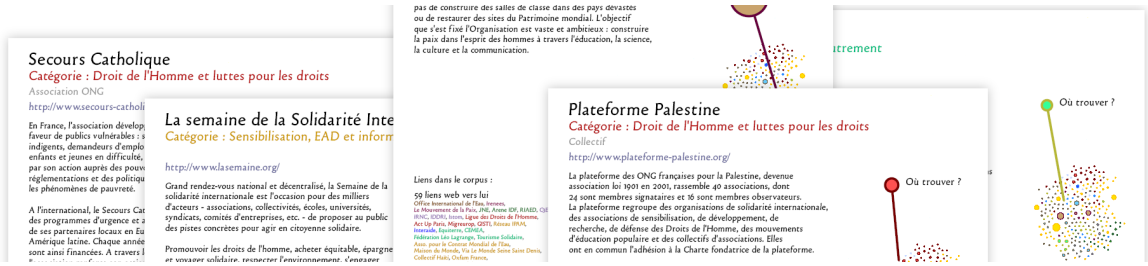


La carte d'une catégorie est construite de la même façon que la carte générale, mais elle ne contient que les sites-acteurs d'une catégorie donnée.
En plus du graphe on peut voir sur la carte de la catégorie les liens qu'elle entretient avec les autres catégories, à la façon du poster de statistique.

Ces cartes sont de bons intermédiaires pour passer de la carte générale au poster de statistiques et vice versa.
Elles permettent d'observer la situation d'un site au sein de sa catégorie, la morphologie de la catégorie (un seul paquet ou des pôles qui se distinguent ?).

Ces cartes répondent aux mêmes règles graphiques et peuvent être comparées les unes aux autres.

– **Le cahier des acteurs**



Chaque page de ce cahier est dédiée à un acteur. Les pages sont par ordre alphabétique du nom de l'acteur.

Ce cahier permet de trouver un acteur que l'on connaît pour le trouver dans la carte générale. Il permet aussi d'en apprendre plus sur un acteur que l'on ne connaît pas mais qu'on a repéré dans la carte générale.

Sur chaque page on trouve les informations suivantes :

- Nom de l'acteur
- Catégories
- Raison sociale
- URL
- Description textuelle en quelques paragraphes (tirée du site)
- Où trouver le site dans la carte générale
- Les liens dans le corpus, en trois colonnes : les sites qui citent cet acteur, les sites qui sont cités par cet acteur, et les sites qui sont les deux à la fois. Ces « voisins » sont de la couleur de leur catégorie.

4. Méthode de réalisation de l'Atlas

4.1. Brève présentation des outils

Un ensemble de logiciels a été développé pour mener à bien ce projet. Cette boîte à outils a été construite en toute rigueur pour répondre à des besoins notamment scientifiques. La vie de ces outils a suivi la même trajectoire que la méthode que nous présentons, à laquelle ils sont associés par nature. Certains se sont avérés essentiels tandis que d'autres ont été abandonnés. Certains projets technologiques ont connu un succès remarquable, et continueront leur vie dans le monde de la recherche en tant que logiciels libres, tandis que d'autres ont connu des difficultés et sont toujours en gestation. Voici un résumé des principaux outils utilisés pour réaliser l'atlas, sachant que nous parlerons plus en profondeur des plus importants dans la section suivante :

- Le Navicrawler, l'outil de base de notre collecte sur le web. Logiciel libre développé par Mathieu Jacomy à WebAtlas, téléchargeable sur webatlas.fr
- Flem, un outil simple mais indispensable pour naviguer dans des listes d'URLs. Logiciel libre développé par Mathieu Bastian à WebAtlas, téléchargeable sur web-mining.fr
- Gephi, un logiciel de manipulation de graphes. Logiciel libre développé par Mathieu Jacomy et Mathieu Bastian à WebAtlas, téléchargeable sur gephi.org
- Atelier de ressources / Datarium. Le cœur d'une bibliothèque numérique collaborative, que nous n'avons pas réussi à rendre fonctionnel. Le prototype développé par Mathieu Jacomy à WebAtlas a été utilisé, mais il est trop limité. L'outil complet développé par Sébastien Heymann et Florian Leclercq à WebAtlas est toujours en gestation. Il n'est pas encore rendu public.
- Baagz, une plateforme collaborative proposée par Exalead. Il est possible de s'y inscrire et de l'utiliser librement sur baagz.com
- Nous avons aussi utilisé les crawlers de l'INA, de RTGI et de Magellan, que nous remercions. Ces crawlers professionnels ne sont pas publics.
- Nous avons en outre développé des mini-outils pour des besoin spécifiques. Ceux-là seront amenés à intégrer le Navicrawler ou Gephi selon leur pertinence.

4.2. Construction du corpus

. 4.2.1. Présentation générale

Il s'agit fondamentalement de constituer une **collection** d'éléments que nous allons traiter au fur et à mesure. Chacun de ces éléments est *à la fois un site et un acteur*. L'esprit de ce projet est de profiter des **propriétés du web**, et en ce sens l'objet **"site web"** est l'unité technique que nous allons manipuler, mais au final notre objet est le **réseau d'acteurs**, au delà du web, et en cela l'entité pertinente est **"l'acteur social"**.

Concrètement la plupart des sites auxquels nous sommes confrontés sont associés à un seul et unique acteur. La raison en est que les acteurs cherchent à se présenter sur la toile, et leurs sites font presque systématiquement fonction de vitrine (entre autres). Remarquons toutefois que certains acteurs n'ont pas de site web (et nous ne les avons donc pas collectés), que certains sites sont de pures vitrines tandis que d'autres ont des fonctions documentaires plus poussées (bibliothèques de ressources, lieux de débat...), et que certains sites ont été pris en compte même

s'ils ne correspondent pas à un seul acteur (les sites de campagnes d'information par exemple).

Un site que nous collectons correspond donc en général à un acteur. Techniquement un site est constitué d'un ensemble de **pages** qui contiennent des **liens hypertextes** vers d'autres sites. Nous collectons l'adresse du site et les liens, mais *nous ne collectons pas les contenus documentaires* : ni le texte, ni les images... En revanche nous avons largement consulté ces ressources pour fonder notre description, notre classification des acteurs. Ici nous avons profité du fait qu'un site web est une entité stable, et bien que les pages en elles-mêmes aient évolué dans le temps de l'étude, nous avons pu consulter les sites en permanence pour les analyser, sans avoir besoin de sauvegarder leur contenu puisqu'il était en ligne.

Le moyen de découvrir les sites de notre champ est la méthode d'**extension pas à pas**. Il s'agit d'étendre petit à petit le territoire analysé en suivant les liens hypertextes. A partir d'un site on peut découvrir l'ensemble des autres sites vers lesquels celui-ci a mis des liens hypertextes. Ces sites sont triés, puis on continue l'extension à partir de ceux qui sont pertinents. Cette démarche continue jusqu'à ce qu'un **épuisement des ressources** survienne : on n'arrive plus à découvrir de nouveaux sites pertinents. A ce moment nous avons des moyens d'éprouver l'exhaustivité de notre collecte. Si la collecte est satisfaisante, elle se termine là.

Pour obtenir un corpus de qualité avec cette méthode, il faut tout d'abord bien choisir ses **points de départ**. L'association RITIMO, par son savoir-faire documentaire, a parfaitement pourvu à cette première nécessité. De manière générale il n'y a pas meilleure solution que de faire appel à des experts du domaine.

Ensuite les *explorateurs*, c'est-à-dire les personnes qui explorent le web pour découvrir des ressources et constituer le corpus, sont confrontées à la **profusion des données**. Pour chaque site collecté, on peut s'attendre à trouver entre 10 et 100 sites supplémentaires à analyser, parfois bien plus. Le nombre de sites web à analyser enfle donc à une vitesse impressionnante, qui peut vite décourager. Le meilleur moyen que nous avons de nous prémunir contre l'ensevelissement est de procéder prudemment, patiemment, avec méthode. L'extension pas à pas permet de gérer cette quantité de données ; en effet comme la plupart de ces sites ne sont pas pertinents, il est nécessaire de les trier avant de continuer l'exploration. Comme nous le verrons plus loin, une bonne compréhension de la structure du web permet de s'y retrouver.

Le tri des sites est donc une question importante. Il est nécessaire de déterminer aussi précisément que possible ce qui distingue un site à garder d'un site à éliminer, sachant que des cas-limites se poseront de toute façon. Nous avons donc défini des **critères de sélection** qui reposent autant sur les connaissances de RITIMO que sur notre première confrontation avec le web (pré-étude).

Enfin remarquons qu'il n'est pas aisé de trouver le bon moment d'arrêter l'exploration. En effet comme on étend le territoire d'étude (extension), on élargit en même temps, par un glissement thématique, le champ de ce qui est pertinent. Au démarrage de l'étude les points d'entrée sont totalement pertinents (par définition). Mais la plupart des sites que l'on trouve ne correspond pas totalement aux critères de sélection que l'on se donne a priori. C'est que le web permet justement d'accéder à une multiplicité de points de vue. Il faut trouver un bon équilibre, pas trop sélectif mais suffisamment pour être pertinent. L'exploration nous permet de rechercher cet équilibre et de le trouver. Il s'agit de rechercher l'adéquation entre les critères de sélection et ce que propose le web pour **constituer un corpus cohérent**. Autrement dit nous cherchons à rendre compte de frontières « naturelles » qui existent dans le territoire informationnel. La structure du web nous aide à trouver la bonne limite pour le corpus.

. 4.2.2. Points de départ

Nos points de départ sont issus de trois sources :

- Les acteurs répertoriés par RITIMO. Cette collection d'acteurs a été réduite à ceux qui ont un site web et qui sont en Ile de France. Ceux-là ont encore été triés pour correspondre à nos critères de sélection thématiques.
  182 sites ont été retenus, 431 ont été refusés, et 2165 autres sites à explorer ensuite ont été repérés autour.
- Notre étude du commerce équitable sur le web francophone. Nous avons collaboré avec la plateforme du commerce équitable (collaboration entre PICRI) pour étudier les acteurs du domaine. Lors de cette étude nous avons d'emblée repéré les sites de la solidarité internationale (nous savions qu'une partie des acteurs serait spécifique au champ du commerce équitable). Notons que cette étude s'était déjà « branchée » sur notre exploration préparatoire, très large mais très partielle, du web du développement durable.
  414 sites de départs devaient être analysés. Après une phase d'expansion, 296 sites ont été gardés pour 539 rejetés, et 3192 sites autour à explorer.
- L'étude de la Solidarité Internationale sur le web francophone par Noémi Schneider. Cette étude ayant été réalisée en partie suivant un travail de documentation traditionnel, et en partie à l'aide d'un *focus crawler*, un crawler capable d'analyser le texte pour orienter la collecte.
  Les 106 sites majeurs de l'étude, les plus connectés, ont été analysés.

Il faut remarquer que ces trois sources sont fortement **redondantes**. Les sites majeurs du domaine sont présents dans les trois collections, et c'est bien normal puisque ce sont les sites qui jouissent de **la plus grande autorité** dans le domaine. Néanmoins le fait de recouper les sources permet de s'assurer de la qualité des points de départ.

. 4.2.3. Première exploration (pré-corpus)

A partir de ces sites nous avons commencé à étendre notre collection. *Dans un premier temps nous ne nous sommes pas limités aux sites franciliens.* A ce moment de notre travail, le web francilien était une priorité mais nous tentions si possible d'agréger toutes les ressources pertinentes du web francophone. Mais il y a une autre raison qui motive ce choix, due à la structure du web.

Pour trouver un lien hypertexte, il faut trouver la page dans lequel il est techniquement inscrit. Autrement dit on trouve les liens hypertextes « sortants », mais il est **difficile de trouver les liens « entrants »**. Deux moyens seraient envisageables pour le faire. Premièrement l'administrateur d'un serveur, et souvent le webmestre d'un site, ont les moyens de savoir « d'où » viennent les visiteurs, de quelles pages. On peut donc avoir les liens entrants sur son propre site. Mais ce sont des données d'usages et elles s'accordent mal à notre méthode. Il faut demander à chaque webmestre s'il veut bien partager cette information, cette démarche était trop lourde pour notre projet. L'autre moyen, qui n'est pas une méthode liée aux usages, est d'effectuer un crawl complet du web. En réalité personne n'est capable de le faire à l'heure actuelle, en tout cas pas avec notre niveau d'exigences. Google dispose de la fonctionnalité « link: » qui permet d'obtenir les liens entrants dont dispose le moteur. Mais en pratique ce système fonctionne mal et les données sont largement biaisées, et encore plus largement incomplètes. *Il n'est donc pas possible d'obtenir les liens entrants*, c'est une limite intrinsèque aux méthodes d'analyse du web.

Si l'on dispose d'un corpus cohérent, il est aisé d'obtenir les liens entre les sites de ce corpus. On a donc au moins les liens entrants en provenance du corpus, pour tout site de ce corpus. Mais comment constituer ce corpus ? Nous exploitons dans cette méthode les résultats de la **théorie des**

agrégats. La thèse principale stipule que les ressources qui traitent d'un même thème sont fortement liées entre elles. Autrement dit ces ressources sont plus liées entre elles qu'elles ne le sont avec l'extérieur. Pour cette raison, les liens entrants ont plus de chance de se trouver dans le même domaine que dans un autre. Ce que nous essayons donc de faire, c'est de *capter l'agrégat de notre thématique*, s'il existe. Ainsi on maximise les chances de trouver des ressources malgré l'impossibilité de remonter les liens entrants.

A ce stade de l'exploration, nous ne savons pas si nous avons affaire à un agrégat. Au final nous verrons que notre corpus participera d'un agrégat plus grand, comprenant une bonne partie de son cœur (beaucoup d'acteurs majeurs se situent en Île de France). En revanche nous nous sommes rendus compte très vite que le critère « localisation en Île de France » n'était pas lié à une agrégation des ressources. C'est bien la thématique des sites qui joue sur leur connectivité, pas leur géolocalisation. En conséquence nous avons choisi d'abord de *collecter le maximum de sites de la thématique, pour ne sélectionner qu'ensuite les acteurs franciliens*. L'expérience préalable de WebAtlas confirme ce choix. Hormis lorsqu'on s'intéresse uniquement à un territoire, c'est-à-dire au travers de toute thématique qui se présente (ex. : analyse du web Picard), il est préférable de se concentrer sur une thématique pour détecter un agrégat et profiter de ses propriétés pour constituer son corpus.

Un **travail collaboratif** a permis de faire cette phase du travail. Les membres de WebAtlas comme de RITIMO ont exploré le web chacun de leur côté pour collecter des sites et mettre le tout dans un pot commun. Certains sites ont été ajoutés par un travail documentaire classique : lorsque RITIMO a identifié par ailleurs des acteurs susceptibles de rentrer dans notre étude, les sites correspondants ont été visités. Dans l'optique de collecter un pré-corpus plus large, il était possible aux membres de WebAtlas de collecter malgré leur expertise moindre que celle de RITIMO.

Nous travaillions en échangeant des listes de sites (format tableur). Le *Navicrawler* et *Flem* ont été mobilisés massivement à ce moment. Nous ne disposions pas d'un outil plus adapté pour faire ce travail. Bien que WebAtlas ait conçu le prototype d'un tel outil, ce dernier n'était pas assez puissant pour soutenir la charge de tant de données. Un projet plus ambitieux était en incubation à WebAtlas, appelé *Datarium*, mais la technologie n'était pas suffisamment mûre pour que nous puissions l'utiliser. Cet outil n'est toujours pas opérationnel à la fin de notre projet PICRI, en partie en raison de son ambition même (nous pensons aujourd'hui que cette brique essentielle ne peut être développée sans un appui industriel).

Nous avons alors rencontré une **difficulté**. Notre première exploration n'arrivait pas au stade d'épuisement des ressources. Non seulement nous trouvions toujours de nouvelles ressources pertinentes, mais d'une part nous voyions de moins en moins clairement les limites de notre domaine, et d'autre part la collecte devenait de plus en plus difficile. Ce phénomène est la bête noire de la personne qui tente de circonscrire un corpus. Il serait bien commode que l'épuisement du domaine survienne d'un coup, mais la plupart du temps il arrive petit à petit. Au fur et à mesure qu'on avance dans l'extension du corpus, on trouve de moins en moins de ressources et ça devient de plus en plus difficile (nous donnerons plus de précisions plus bas). La position difficile dans laquelle nous nous trouvions correspond à un moment où l'on trouve encore trop de ressources pour décréter l'épuisement, mais où la collecte devient trop pénible. Typiquement il s'agit du cas où pour chaque journée passée à explorer des sites, on ramène une douzaine de sites pertinents (ce qui est peu) mais qui s'avèrent si importants qu'on est obligé de continuer pour ne pas en rater d'autres importants. La raison de cet état de fait est souvent que l'on s'est laissé dériver vers des thématiques trop larges (« topic drifting » dans la littérature scientifique anglo-saxonne : ce phénomène est bien connu car il advient également aux crawlers).

Nous avons donc décidé de nous donner des **critères plus fermes** de sélection du corpus. C'est à ce stade que nous avons décidé de nous concentrer sur les ressources franciliennes. Nous

avons donc visité tous les sites collectés pour déterminer s'ils étaient ou non basés en Île de France. Nous avons par ailleurs resserré la thématique autour de la Solidarité Internationale, parce que le Développement Durable est un critère bien trop large ; en effet on trouve de nombreux acteurs qui s'en revendiquent, mais pour beaucoup il n'est pas évident de comprendre précisément pourquoi ! Au final le critère retenu sera « l'action à l'international », qui s'avère être le plus discriminant pour la coloration que nous souhaitions donner (en effet nous ne voulions pas adopter un regard naïf quand au développement durable). Mais cette formulation ne viendra que dans un second temps.

Il faut bien saisir que la difficulté de cette phase tient au fait que l'on cherche simultanément à comprendre les thèmes présents sur le web et à déterminer des critères discriminants. L'intérêt est de *coller* au terrain-web pour rendre compte au mieux de l'information. Mais le prix à payer est un long tâtonnement jusqu'à ce que se décantent les axes d'étude pertinents. A cet égard il nous a été utile de séparer clairement le processus de **sélection** et le processus de **qualification** de l'information. En effet, à ce stade les critères de sélection sont encore flous. Même après avoir resserré nos critères, nous n'étions pas encore capables de les formuler clairement. Cette situation est tout-à-fait normale, elle est justement due au fait que c'est le regard expert qui permet de circonscrire le corpus, et non des critères posés *a priori* dont il serait impossible d'évaluer la pertinence. La qualité de cette méthode tient justement à la confrontation entre des experts et leur terrain, comme il est de mise classiquement en sciences humaines (c'est un travail similaire à de la sociologie qualitative, avec une dimension plus quantitative ; Bruno Latour appelle ce type de méthode « quali-quantitative »).

Pour nous y retrouver dans cette première exploration, nous avons naturellement décrit nos ressources. Ne disposant pas de grille d'analyse, nous avons rajouté autant de catégories qu'il le fallait pour décrire les sites-acteurs. Nos critères ayant été plutôt trop larges, de nombreuses catégories ont été utilisées. Les voici :

- **Types d'acteurs**
  - Institut
  - Campagne
  - ONG
  - Collectif
  - Site internet
  - Association
  - Institutionnel
  - Université/centre de recherche
  - Entreprise
  - Presse/média
  - Perso
  - Tête de Réseau
  - Mixte
- **Échelles d'action**
  - France et International
  - International
  - en France
- **Domaines d'action**
  - Droits des peuples
  - Féministe
  - Droits de l'homme
  - Biens publics mondiaux
  - Eau
  - Éducation populaire
  - Volontariat
  - Altermondialiste
  - Responsabilité sociale des entreprises
  - Justice sociale
  - Voyage solidaire
  - Paix
  - Commerce équitable
  - Immigration
  - Humanitaire
  - Décroissance
  - Développement durable
  - Environnement
  - Mondialisation culturelle
  - Mondialisation
  - Agriculture
  - Racisme
  - Education au développement
  - Économie solidaire
  - Solidarité internationale

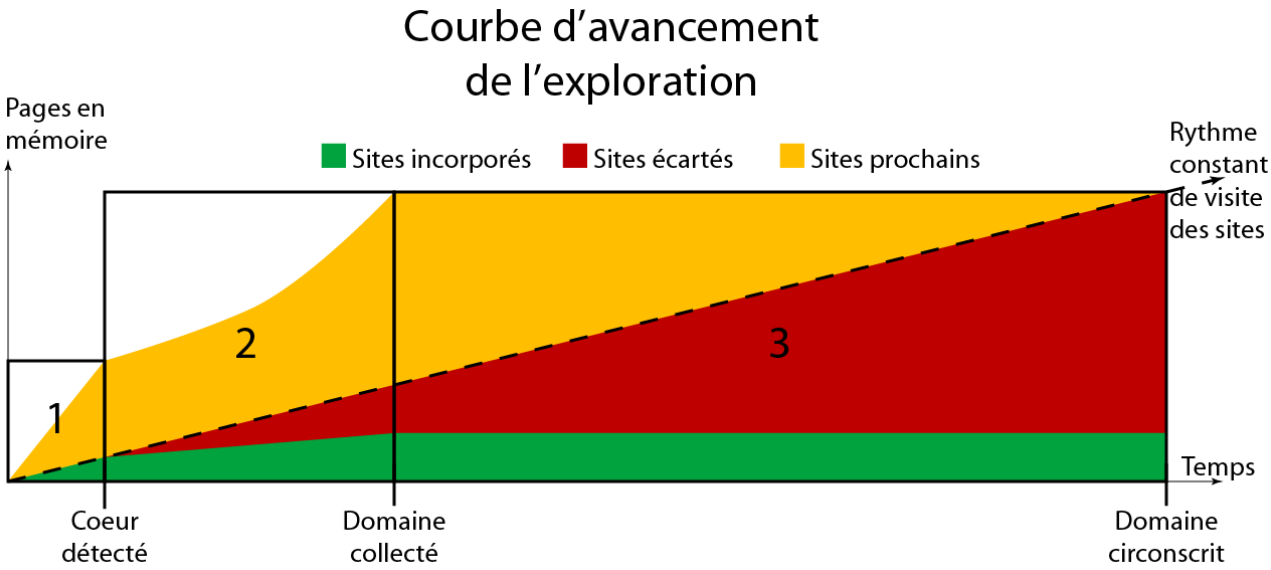
Le rôle de ces catégories a été de nous permettre de faire le point sur l'information présente sur le web, et de choisir en conséquence comment resserrer nos critères de sélection. Ces catégories permettent de **sélectionner** l'information. Lorsqu'il sera question plus loin de **qualifier**

l'information, d'autres catégories seront en jeu. En effet la qualification ayant lieu sur un corpus circonscrit, selon des critères déterminés, il y a un contexte qui donne sens à ces catégories. Tandis que les catégories ci-dessus sont strictement descriptives, elles n'ont pas de contexte. Aussi étonnant que cela puisse paraître, il est plus utile de laisser tomber ces catégories et de recommencer le classement de zéro, pour ne pas mélanger la sélection et la qualification.

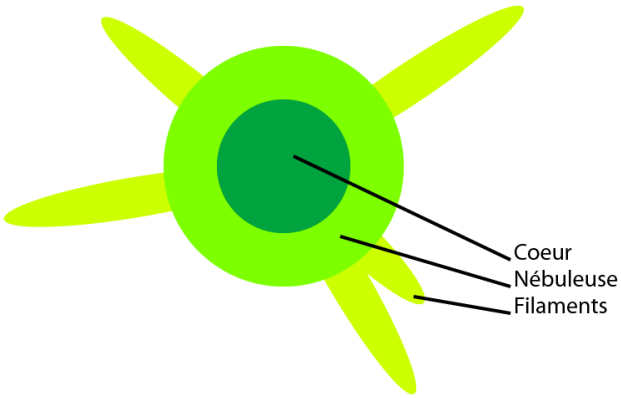
. 4.2.4. Seconde exploration (circonscription du corpus)

Nous avons donc resserré nos critères et restreint notre corpus en cours de constitution. Nous avons repris l'exploration et nous sommes arrivés au stade d'épuisement du domaine.

Nous avons abouti à une sélection de 400 sites, pour plus d'un millier rejetés. Autour de ces 400 sites, 10 000 sites restaient encore à explorer. Bien que l'épuisement soit survenu, nous avons cherché à être sûrs que nous ne rations pas de sites importants. Nous allons donc nous pencher un peu sur ce qui se passe au moment de l'épuisement, à l'aide de la courbe suivante.



Au début de l'extension du corpus, ce sont les sites proches des points de départ qui sont analysés. Comme nos points de départs étaient de bonne qualité, la plupart des sites trouvés rentraient bien dans nos critères. Cette phase est donc facile, parce qu'on découvre de nombreuses ressources pertinentes. Mais rapidement, la plupart des sites « majeurs » est collectée. Le « cœur » a été pris et les sites que l'on trouve appartiennent à la « nébuleuse », qui entoure le cœur.



Le propre de la **nébuleuse** est d'être moins dense que le cœur. Autrement dit, à partir de ce moment (entre 1 et 2 dans la courbe), on trouve de plus en plus de sites non-pertinents. Dans la

courbe, le nombre de sites en rouge augmente sérieusement. L'épuisement survient lorsque le domaine a été globalement collecté, c'est-à-dire que le cœur et la nébuleuse sont collectés (et un gros effort serait nécessaire pour collecter les filaments, ce qui est souvent superflu). Dans la courbe ceci correspond à la limite entre la phase 2 et la phase 3. Pour comprendre ce qui suit, il faut comprendre en quoi consisterait cette phase 3, qu'on ne fait presque jamais tant elle est laborieuse. On a vu plus haut que le Navicrawler détecte des sites en périphérie de ceux qu'on a collectés. Ce sont normalement ceux qu'il faut explorer pour découvrir de nouvelles ressources. Lorsque le principe d'épuisement survient, on se rend compte que ces sites à explorer ne contiennent plus de ressources pertinentes. Le filon est épuisé. Si l'on souhaitait étudier la **frontière** du corpus en toute rigueur, il faudrait trier tous ces sites. Mais ils sont la plupart du temps au moins dix fois plus nombreux que les sites collectés !

Dans notre cas nous avons utilisé un **raccourci**, pour évaluer les 10 000 sites autour de nos 400 sélectionnés. A partir du moment de l'épuisement, moment où il est probable que le domaine a été globalement collecté, nous avons directement cherché à évaluer la présence de sites que nous aurions manqués. Le raccourci repose sur l'hypothèse suivante. A partir du moment où l'on est sûr qu'une grande partie du domaine a été collectée (puisqu'on est arrivé à l'épuisement), il est peu probable que si un site est pertinent, il ne soit pas cité par plusieurs sites que l'on a. Autrement dit on regarde dans les sites qui restent à explorer, ceux qui sont cités plusieurs fois. Concrètement, on élimine tous les sites à explorer qui sont cités une seule fois. Ceci divise par 5 à 10 le nombre de sites à explorer. Nous avons effectué cette approximation et nous avons trié les sites restants. Parmi ceux-ci nous en avons trouvé peu de pertinents. Nous avons donc considéré que l'essentiel du domaine était collecté (exhaustivité suffisante).

A cette étape, le corpus est **circonscrit**.

. 4.2.5. Qualification

Une fois le corpus circonscrit, il est nécessaire de qualifier les sites-acteurs pour exploiter les données (cartes et statistiques). Cette étape est simple dans son principe. Pourtant c'est l'occasion pour l'équipe de revenir sur les données, de mieux les comprendre, en tout cas de les comprendre autrement. Comme il n'y a plus l'enjeu de sélectionner ou non les sites, c'est l'occasion de s'intéresser à la spécificité de chaque acteur. Et ce faisant, de **remettre en cause** le travail, de réévaluer le corpus.

Comme nous l'avons évoqué plus haut, la description initiale doit être reprise au moment de la qualification. Après réflexion sur le rôle de ces catégories, nous avons décidé de nous limiter à une dizaine d'entre elles. En effet il n'en faut pas plus si l'on veut produire des visualisations claires, et il n'en faut pas moins si l'on veut rendre compte des **composantes principales** du champ. La dizaine nous a semblé à tous un compromis satisfaisant. Les 9 catégories finalement retenues sont les suivantes :

- Aide au développement, coopération publique
- Sensibilisation, EAD et information
- Vivre autrement (économie sociale et solidaire, alternatives citoyennes)
- Protection de l'environnement, environnementalisme
- Appui à des projets au sud, de société à société
- Humanitaire, urgence
- Mouvement de paix, non-violence
- Droits de l'homme et lutte pour les droits
- Étude, recherche et formation

On peut s'en rendre compte en comparant avec la liste inscrite plus haut, ces catégories sont

à la fois plus précises et plus consistantes que les catégories descriptives initiales. Elles sont le fruit d'une réflexion sur le **découpage pertinent** des ressources, pertinent par rapport au web et pertinent par rapport aux relations qu'entretiennent les thèmes entre eux. Certains regroupements ont été effectués parce que les ressources de deux sous-thèmes étaient très liées de fait dans le web, d'autres regroupements sont imputables à la similarité entre sous-thèmes.

A ce stade, nous avons mis à plat nos critères de sélection et nous avons encore affiné notre corpus. C'est seulement à ce moment que nous avons explicité le critère « action à l'international ». Nous avons éprouvé la nécessité de retravailler le corpus et des 400 sites pré-sélectionnés nous en avons gardé seulement 259. Il n'y a pas eu d'étape de sélection à proprement parler, il s'est simplement avéré chemin faisant que certains sites-acteurs n'étaient pas en cohérence avec les autres, et nous les avons éliminés.

D'autres données ont aussi **enrichi** le corpus. La raison sociale tout d'abord, nous a semblé un élément important. En outre nous avons décrit chaque site-acteur par un court texte de présentation que nous sommes allés chercher dans chaque site web. Enfin n'oublions pas que chaque site-acteur est décrit par une adresse web et un nom. Ainsi à l'issue de ce travail nous avons l'équivalent d'une base de donnée décrivant notre corpus.

Ce travail collaboratif a été effectué sur la plateforme *Baagz* proposée par Exalead. Nous avons détourné cet outil de sa vocation première afin de disposer d'un lieu collaboratif pleinement fonctionnel.

Remarquons enfin que nous avons réalisé une première cartographie en cours de route pour évaluer les données et discuter ensemble de la meilleure façon de les présenter sous forme de carte.

. 4.2.6. Crawls finaux

Notre corpus a été collecté avec le Navicrawler et à ce titre nous disposons d'une partie des liens hypertextes. Mais le Navicrawler n'est pas le meilleur outil pour récolter ces liens, et nous souhaitions maximiser la qualité du corpus. Nous avons donc procédé à des crawls avec différents outils qui nous ont aimablement été prêtés. Bien que nous ne soyons pas des adeptes des crawlers pour la constitution de corpus, ce sont des outils essentiels lorsque l'on dispose d'un corpus déjà constitué.

Nous avons tout d'abord effectué un crawl avec le crawler de l'entreprise RTGI. Certains sites n'ayant pas de liens, nous avons revérifié ceci avec le Navicrawler. Puis nous avons demandé à l'équipe Dépôt Légal du Web à l'INA de nous faire un crawl avec leur propre technologie, ainsi qu'à l'équipe Magellan associée à WebAtlas. Nous avons fusionné ces quatre sources.

Le corpus final est donc constitué des données qualifiées associées aux données des crawls. Le format de ces données est un **graphe enrichi**. Nous avons arrêté là le travail de construction des données pour passer à l'étape de visualisation et d'analyse, dont le rôle est de transmettre ces données aux personnes de façon lisible.

. *4.3. Construction des cartes*

. 4.3.1. Spatialisation des graphes

Pour ce travail nous avons utilisé le logiciel Gephi développé au sein de WebAtlas. Ce logiciel est typiquement fait pour traiter le type de graphe enrichi dont nous disposons ici.

Nous avons utilisé un **algorithme** développé à WebAtlas pour spatialiser le graphe. Son principe est simple : les nœuds du graphe (les sites-acteurs) se repoussent tandis que les arcs (les

liens hypertextes) retiennent les nœuds à leurs extrémités. Ce système fonctionne comme si les nœuds étaient des aimants qui se repoussent et que les arcs étaient des élastiques. Tout le graphe se met à bouger à partir d'une configuration initiale aléatoire, jusqu'à ce que les nœuds soient tous dans une position d'équilibre par rapport aux autres. Lorsque la position d'**équilibre** est atteinte, on arrête l'algorithme.

Notre algorithme dispose de **fonctionnalités supplémentaires** que nous avons utilisées ici. Premièrement, plus un nœud a de liens, plus il repousse ses voisins. Deuxièmement, plus un nœud a de liens sortants, moins il y a de force de traction dans ces nœuds. Troisièmement, nous avons ajouté une force de gravité pour homogénéiser le graphe. En outre, les différents **paramètres** de l'algorithme ont été ajustés afin d'obtenir le résultat le plus lisible.

En l'état, les nœuds se recouvraient encore trop pour que la carte soit lisible. Nous avons donc utilisé à la suite un autre algorithme qui amène les nœuds à se repousser lorsqu'ils sont trop proches les uns des autres.

A l'issue de ces traitements, nous avons spatialisé le graphe qui donnera ensuite la carte.

. 4.3.2. Manifester les caractéristiques du corpus

La spatialisation ne suffit pas. Il faut aussi travailler l'apparence des nœuds et des arcs pour rendre compte des données associées au graphe.

La principale information à manifester est la catégorie de chaque nœud. Nous avons donc choisi ensemble une **coloration** pour chaque catégorie, et nous avons colorisé chaque nœud de la sorte (chaque catégorie est doté d'une couleur de fond et d'une couleur de rebord, pour qu'on puisse bien les distinguer). En outre, nous avons donné à chaque nœud une **taille**, en fonction du nombre de liens entrants. En effet comme il est plus facile de mettre des liens vers d'autres sites que d'en obtenir de leur part, le nombre de liens entrants (ou nombre de citations) est un indicateur remarquable.

Enfin nous avons exporté ce graphe en tant qu'**image vectorielle** selon une procédure que nous avons élaborée spécifiquement pour ce travail, afin que la carte soit aussi lisible et complète que possible.

. 4.3.3. Extraction des statistiques

A l'aide d'outils que nous avons construits pour l'occasion, nous avons extrait un ensemble d'indicateurs permettant différentes façons de lire les données. Les statistiques se séparent en deux lots : les statistiques générales, et les statistiques propres à une catégorie donnée.

Les **statistiques générales** permettent de comprendre la structure de l'ensemble du corpus. Il s'agit par exemple des sites-acteurs les plus référencés dans le corpus, ou de ceux qui ont le plus de liens sortants dans le corpus. La liste des sites-acteurs classés du plus connecté au moins connecté est également importante. Citons enfin la répartition des sites-acteurs dans les catégories, et le tableau des liens entre ces catégories.

Les **statistiques propres à une catégorie** permettent de comprendre sa composition et ses relations avec les autres catégories. Il s'agit aussi bien du nombre de sites et du nombre de liens internes à la catégorie, que du nombre et de la répartition de ses liens entrants, de ses liens sortants.

Nous avons travaillé ces statistiques pour qu'elles ne soient pas de simples listes d'indicateurs, mais des éléments visuels cohérents que l'on puisse lire facilement, notamment pour comparer le profil des catégories entre elles.

. 4.3.4. Finalisation des cartes

Enfin, n'oublions pas l'étape importante de mise en page des cartes. Nous avons choisi ensemble quelles cartes produire, et nous avons associé nos différents objets graphiques pour produire une série de cartes ainsi qu'un cahier.

Les cartes disposent toutes d'un titre, d'un cartouche et des logos des partenaires. La plupart des cartes sont construites autour d'un graphe travaillé comme expliqué plus haut. A ce graphe est toujours associé une légende et différents éléments statistiques qui permettent de le comprendre.

Ces cartes que nous avons présentées plus haut existent en numérique au format PDF, mais elles ne sont lisibles qu'imprimées en A0 ou plus grand.

5. Réalisations technologiques

L'activité scientifique d'étude du web, encore très jeune au moment de ce bilan, nécessite des outils. De *nouveaux* outils, c'est-à-dire des outils à inventer. Ce projet est bien sûr un projet d'innovation, même s'il n'est pas que cela.

Si ce projet s'arrête ici, il en restera différentes choses. Des résultats naturellement, et les documents qui les présentent : l'atlas. Un savoir-faire et des liens tissés entre la société civile et la recherche, et pas seulement entre les deux partenaires. Et des **outils logiciels**, réalisés pour faire ce travail, qui sont pour la plupart des projets pérennes, et à la disposition des deux communautés, la société civile et la recherche, pour de futurs travaux.

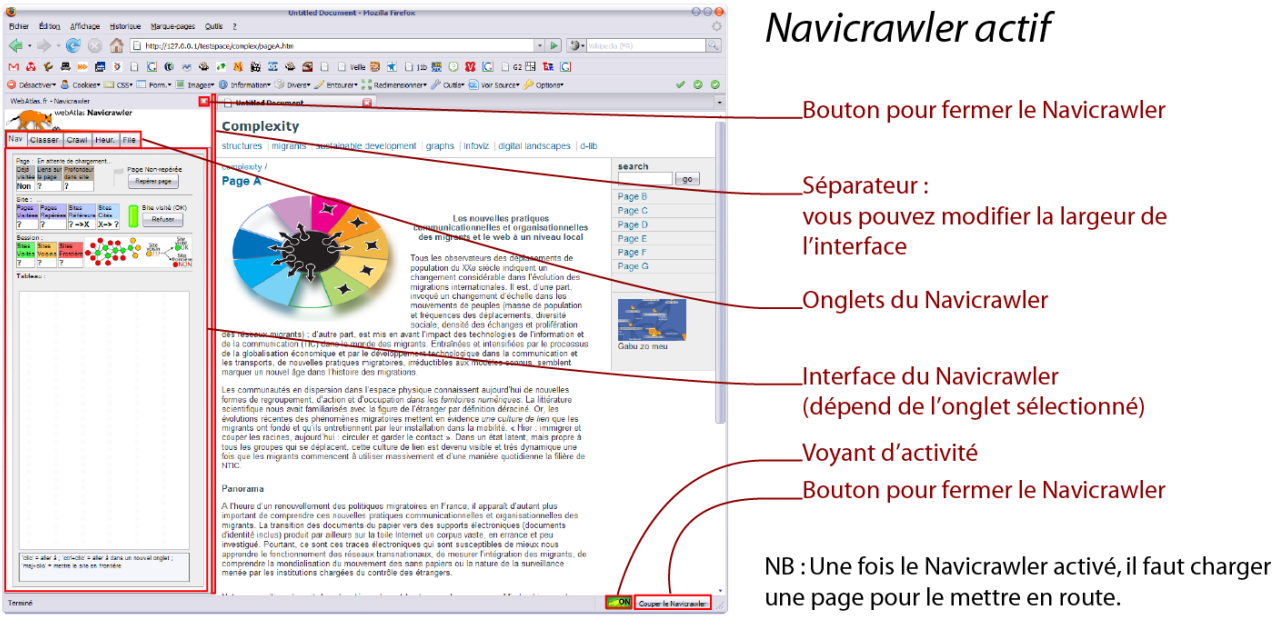
Fruits de l'innovation, les logiciels portent en eux la marque de notre approche, de l'histoire de notre projet, des personnes qui ont travaillé pour les rendre utiles. Les personnes de RITIMO ont en quelque sorte « joué les cobayes » de ces outils. Il est vrai que nous avons tous fait les frais d'outils incomplets, bancals, difficiles à utiliser parfois. Mais ce faisant il a été possible de les enrichir, les ajuster, les améliorer. C'est en ce sens que les personnes de RITIMO ont contribué à construire ces outils. C'est une démarche de conception participative ; notre expérience commune s'y est incrustée au fil du temps.

Ces outils ont été portés par WebAtlas, et ils n'ont bien sûr pas été développés que pour ce projet-ci – c'est la condition de leur qualité, de leur fiabilité, de leur durabilité. Ils n'en sont pas moins d'importants objets « matériels », nonobstant leur nature logicielle, que ce projet lèguera à ses successeurs. En ce sens ils font partie du patrimoine de ce PICRI, et c'est pourquoi nous leur consacrons cette section.

. *5.1. Le Navicrawler*

La fonction principale de cet outil est de construire un *corpus* de sites web. Il est né du besoin de fabriquer des collections cohérentes de sites web, de façon à ce que la personne qui fait le corpus comprenne ce qu'il y a dedans. C'est cette approche qui le différencie d'un « crawler » classique. Il ne s'agit donc pas de collecter une grande quantité de sites et de la trier ensuite, mais au contraire d'aller soi-même explorer le web et de sélectionner ce qui constitue le corpus en même temps qu'on consulte les pages. C'est une approche constructiviste qui est plus en accord avec les méthodologies classiques en sciences humaines, la plupart des outils étant descriptivistes.

Le Navicrawler est une extension pour le navigateur libre et open source Firefox. C'est-à-dire qu'on peut rajouter le Navicrawler à son Firefox pour explorer le web. Il s'ouvre alors dans une fenêtre latérale, à côté de la page de navigation, et on peut y obtenir différentes informations sur la page où on navigue à ce moment là (liens hypertextes, vers quels sites etc.) et choisir ou non de garder le site dans le corpus. On peut également décrire les données en ajoutant des libellés aux sites. Il est fait pour exporter des graphes qui donneront des cartographies avec un autre logiciel, mais on peut aussi l'utiliser pour faire des tableaux Excel.



Le Navicrawler nous a bien sûr permis de constituer notre corpus, et de faire nos premières cartographies. Disposer d'un tel outil est nécessaire pour évaluer la qualité des données. C'est un peu le « couteau suisse » de la personne qui souhaite faire de la cartographie du web. Son principe est simple, et il est moins performant qu'un crawler sophistiqué. En revanche sa simplicité est un avantage lorsqu'il s'agit de vérifier certains détails des données. Il donne des résultats à la mesure de la personne qui l'utilise, et il est donc indiqué pour un sociologue comme pour un spécialiste d'un domaine documentaire, dans notre cas les personnes de Ritimo.

Le travail avec le Navicrawler est souvent fastidieux, car cet outil n'est en fin de compte qu'une aide pour effectuer le travail à la main. Mais il permet une grande qualité, sans laquelle les données se biaisent facilement. A notre connaissance, aucun corpus de sites web n'a été aussi finement décrit et analysé que le nôtre, parce que ce sont les personnes les plus qualifiées qui ont exploré et analysé le web pour construire un corpus de qualité. L'innovation du Navicrawler est justement de permettre aux spécialistes de l'information, et non pas de l'informatique, de se confronter au web.

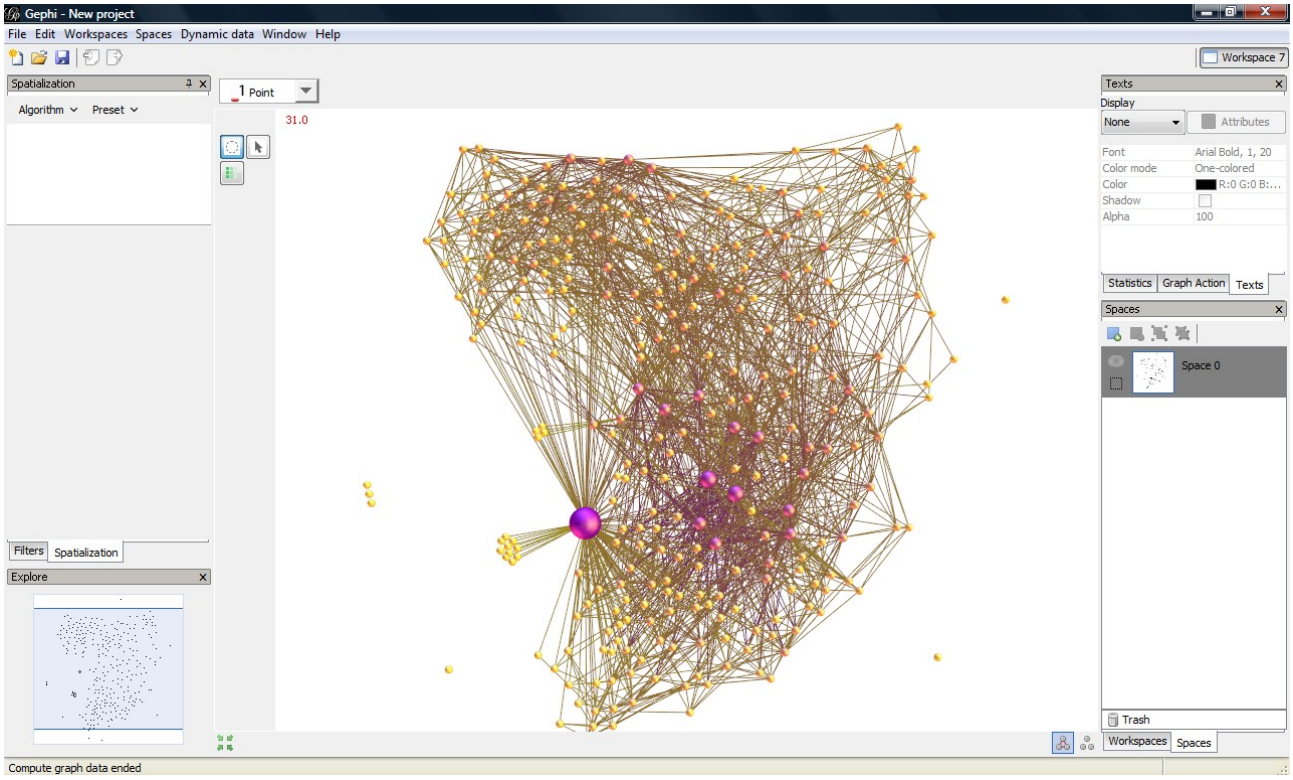
Le Navicrawler est un logiciel libre et open source sous licence GPL 3, téléchargeable sur le site de WebAtlas (avec sa documentation). Il est développé principalement par Mathieu Jacomy, depuis 2005, et continue d'évoluer chaque mois pour répondre aux besoins de ses utilisateurs. Il a été soutenu par Orange Labs, et bien sûr ce projet PICRI a permis de le développer dans d'excellentes conditions. Il est utilisé par le programme TIC-Migrations de la Fondation Maison des Sciences de l'Homme, par des chercheurs en sciences humaines d'Orange Labs, par les étudiants de Bruno Latour à Sciences Po pour effectuer des cartographies de controverses, et par d'autres personnes qui l'ont librement téléchargé. WebAtlas effectue des formations, pour des chercheurs mais aussi pour les personnes de la société civile qui le souhaitent, dans la continuité de ce projet.

. 5.2. Gephi

Gephi est un outil de spatialisation de graphes, taillé sur mesure pour les besoins de WebAtlas. Ce logiciel libre et open source, codé en Java et accéléré par OpenGL, nous permet de passer des données collectées aux cartographies. Il a été initialement développé par Mathieu Jacomy au sein de WebAtlas sous le nom de Graphiltre, et il est désormais sous la responsabilité de Mathieu Bastian et Sébastien Heymann au sein d'une association. Le site gephi.org permet de s'informer sur

ce logiciel et de le télécharger.

La fonction centrale de Gephi est de spatialiser un graphe. Rappelons que les données collectées sur le web peuvent être projetées sous forme de graphe : des nœuds reliés par des arcs. Comme la littérature scientifique le prévoit, ces graphes sont de type « réseau à invariance d'échelle ». Cette catégorie de graphes est la signature de ce qu'on appelle parfois la « complexité » ; ce sont des objets difficiles à traiter car ils ne sont pas hiérarchiques comme des arbres, mais ils ne sont pas aléatoires non-plus, et il y a de fortes disparités dans leur densité. A ce jour il n'existe aucune méthode déterministe pour représenter ces graphes. C'est la raison pour laquelle nous construisons des cartographies : nous avons besoin de construire des algorithmes au cas par cas et de manipuler longuement ces graphes pour en obtenir une visualisation satisfaisante, au bénéfice d'un résultat qu'aucune autre méthode ne permet d'obtenir, et qui nous aide à comprendre la complexité des données. C'est pour faire ce travail artisanal que nous avons conçu Gephi.

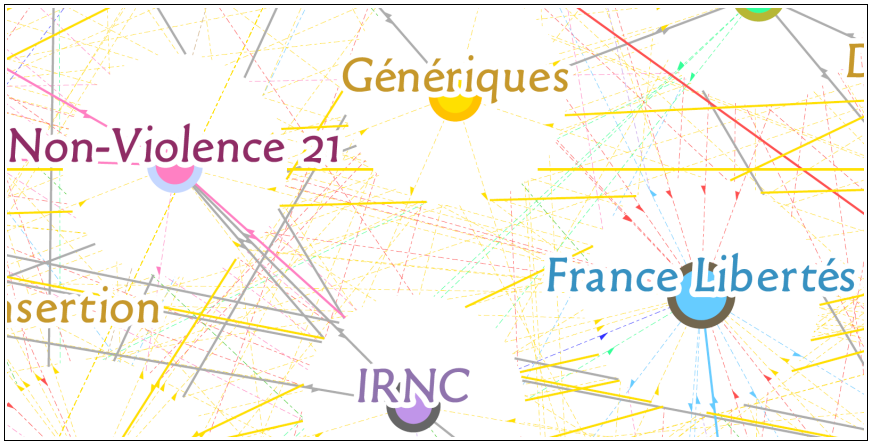


Gephi a évolué en même temps que nos besoins pour réaliser des cartographies. Nous avons notamment développé l'algorithme *ForceAtlas* spécifiquement pour réaliser des cartes, et nous l'avons amélioré continûment. Pour les besoins de ce projet nous lui avons ajouté des paramètres qui n'ont pas de sens dans une optique purement mathématique, mais qui sont nécessaires pour réaliser une cartographie : empêcher les nœuds de se recouvrir, forcer les *hubs* (sites portails) à se dégager sur les côtés du graphe (car les autorités sont plus légitimement centrales)... L'intérêt de Gephi par rapport à d'autres logiciels de graphes est qu'il est possible de manipuler à fond les paramètres des algorithmes pour arriver à un résultat lisible.

Nous avons implémenté dans Gephi un système de filtres qui permet d'intervenir sur l'aspect graphique du graphe pour en faire une carte. Typiquement, il s'agit de mettre des couleurs aux nœuds selon leur catégorie thématique, et de leur donner une taille proportionnelle au nombre de liens qui pointent vers eux (dans le corpus). Mais les filtres sont bien plus génériques que ça et permettent la plupart des opérations nécessaires pour fabriquer une représentation visuelle

complète, au sens de Jacques Bertin.

Au regard des autres projets de WebAtlas, notre PICRI est particulièrement exigeant quant à la lisibilité des cartes pour le grand public. Nous avons donc développé de nouvelles façons de représenter les nœuds et les arcs pour répondre à cette exigence. Comment représenter tous les arcs sans pour autant parasiter la lecture des nœuds, de façon à ce qu'on puisse toujours savoir qui est connecté à qui ? Notre innovation repose sur l'ajout d'une zone vierge autour de chaque nœud, qui n'empêche pas de suivre les liens mais qui ajoute une meilleure lisibilité (voir image ci-dessous). Ce nouveau type d'export de graphes est maintenant disponible pour tous dans Gephi.



6. Retombées et enseignements

. 6.1. Apports et intérêt du projet pour RITIMO

Ce projet a apporté à Ritimo une connaissance des outils en train de se mettre en place dans le domaine du Web et pas simplement ceux utilisés dans le cadre du projet. La rencontre avec des chercheurs n'a pas été toujours simple mais a toujours été enrichissante. Pas simple parce que Ritimo a du s'approprier des outils logiciels innovants souvent complexes et en cours de stabilisation. Au début, les problèmes de compréhension commune de l'étude du web ont été peturbées par des problèmes de vocabulaire, de méthodes de travail.

Mais au final, la collaboration chercheurs et animateurs associatifs travaillant dans le domaine de l'information et du web a permis à chaque partie de prendre du recul par rapport à des pratiques de travail, des façons de mener des projets. Pour Ritimo cela a participé grandement à prendre la mesure de l'enjeu du Web comme outil et support central, à comprendre la nécessité de l'investir, de profiter des outils innovants. Cela a donné à l'association des pistes sur la façon d'y prendre une place de référence comme diffuseur d'informations.

Connaître sa place, les complémentarités avec d'autres, être acteur pour être bien placé donc référencé est une question centrale pour Ritimo depuis qu'une partie des publics des bibliothèques et des centres de documentation se tournent plutôt vers la toile. Maîtriser cette géographie complexe pour mieux diffuser nos outils, élargir nos publics, devenir plus collaboratif est un pari pour l'avenir. Ritimo comme acteur de la solidarité internationale a pu nouer de véritables partenariats dans le monde par la valorisation et la dissémination des points de vue du Sud et ce grâce à la maîtrise des technologies de l'information et de la communication et de l'appropriation des innovations en cours.

L'intérêt a bien entendu été aussi de pouvoir explorer le web de notre milieu par l'utilisation du navicrawler. Cela nous a fait connaître beaucoup d'acteurs, nous a fourni de multiples liens intéressants. Ce projet a ainsi participé à l'amélioration de notre bibliothèque internationale rinoceros et à son enrichissement.

Parallélement, Ritimo a participé à la création d'une Confédération de sites ressources pour une Démocratie Mondiale (Coredem) qui a permis d'apprendre à utiliser des outils libres et coopératifs comme le wiki et de créer un moteur de recherche (Scrutari) propre à cette communauté mais disponible pour d'autres.

Cette prise de conscience et l'étude elle-même avec les cartes comme résultat visible ont donné à Ritimo une légitimité dans le milieu francilien du développement durable et de la solidarité internationale pour parler, conseiller, accompagner les animateurs de sites soucieux de leur efficacité et de leur référencement.

L'intérêt de ce projet et des dispositifs PICRI plus largement a été très manifeste lors de la journée de restitution du 25 novembre 2008. Cet intérêt se poursuit avec les acteurs à qui nous présentons les cartes et le travail de recherche.

. 6.2. Communauté et associations

Nous avons confronté notre travail aux acteurs concernés par la carte au cours d'une journée de restitution de nos travaux. Plusieurs associations ont répondu présent à notre invitation et nous avons donc eu l'occasion de discuter de l'intérêt et de la pertinence de nos travaux.

Première constatation, c'est un travail difficile à recevoir quand on est un acteur représenté

dans la cartographie. L'objet « carte » intrigue et suscite l'intérêt, mais il est difficile de « lire » des informations intéressantes au premier coup d'œil. Les premières réactions consistent à se chercher « soi-même » dans la carte, à regarder comment on est classé et à qui on est relié. A cet égard les acteurs ont été satisfaits, ce qui témoigne de la qualité de notre travail de sélection et de qualification. La réaction qui vient ensuite est la recherche des personnes avec lesquelles on est lié. Là encore les réactions sont bonnes mais nous avons observé plusieurs cas particuliers. D'abord les acteurs posent la question de l'adéquation entre les liens dans la carte et les liens réels ; c'est l'occasion de faire le point de la représentativité d'un site web. Ensuite certains acteurs attendus n'étaient pas dans la carte. Aucun acteur majeur n'est manquant mais les choix que nous avions effectués au moment de la sélection sont bien sûr discutables. Nous arrivons là au point où **un acteur seul ne peut définir précisément un corpus**. Il a alors été question de la capacité de la communauté des acteurs à s'emparer d'un tel travail pour faire vivre une Bibliothèque Numérique.

Ce qui nous mène à notre seconde constatation : les acteurs associatifs ne se vivent pas comme une communauté. (Note de WebAtlas : il faut dire que lorsque nous analysons des corpus web, nous avons l'habitude de désigner sous le terme de « communauté » tout ensemble d'acteurs très densément lié.) C'est assurément abusif, et les acteurs de notre domaine n'entendent pas le terme « communauté » de cette oreille. Pourtant la carte, en tant que graphe de relations, est indéniablement dense, et en ce sens nous pensons que c'est le signe d'une activité « communautaire » significative. Cependant il s'avère que le bon mot n'est pas « communautaire » mais « **collectif** ». Quelle est la différence, qu'est-ce que ça change ?

Nous pensions que la grande quantité de liens, et donc d'interactions entre acteurs, justifiait qu'on parle de « communautés ». Mais en fait ce n'est pas la quantité d'interactions qui compte ici, c'est leur nature. Désigner les acteurs par le mot « communauté » sous-entend qu'il existe une certaine unité, une direction commune. Au contraire nos associations se vivent comme une multitude, comme un milieu varié dans lequel deux acteurs ont chacun une histoire différente, une vocation distincte, une action particulière. Un principe d'indépendance très puissant est à l'œuvre, mais qui n'empêche pas les associations d'interagir massivement. Les divergences n'empêchent pas les collaborations, et dans ce réseau les acteurs n'ont pas pour habitude de s'ignorer. Or étonnamment, il est courant que des acteurs « s'ignorent volontairement » au point de refuser de mettre des liens vers les autres.

Nous considérons qu'une institution, par définition, est une entité qui ne pointe pas vers d'autres acteurs (ou plus précisément vers des acteurs plus « petits » qu'elle, surtout s'ils sont non-institutionnels). C'est ce qu'on observe dans notre carte où France Diplomatie jouit de liens entrants depuis 78 acteurs de notre corpus, tandis que seulement 38 sont cités. Mais ce serait encore plus flagrant si nous étudions les liens avec de plus grosses institutions qui ne sont pas dans notre corpus. Un autre cas est bien connu. Lorsque deux acteurs sont en compétition, ils ont tendance à s'ignorer l'un l'autre. Les entreprises ne mettent jamais de liens vers leurs concurrents, mais on retrouve ce comportement partout où des entités se disputent une ressource limitée. C'est par exemple le cas des sites d'information, et notamment des sites de journaux, qui se lient peu les uns les autres. On observe ce cas très localement dans notre carte, pour les sites humanitaires. Deux traits les rapprochent des sites d'entreprise : ce sont des sites vitrines et non d'information, et ils s'ignorent entre eux. En dehors des acteurs institutionnels et humanitaires, les acteurs sont remarquablement liés.

Donc si l'on résume, nous n'avons pas affaire à une communauté, mais pour autant peu d'acteurs s'ignorent. Autrement dit nos acteurs ne forment pas une communauté mais leur stratégie observée (explicite ou non) est massivement coopérative. **Les acteurs n'embrassent pas de destinée commune mais ils ne sont pas non plus en compétition pour une ressource limitée, fut-elle l'accès** (avoir plus de liens entrants, donc plus de visiteurs et de « ranking » dans les moteurs de recherche). Les acteurs de notre domaine partagent en bonne entente un territoire

informationnel commun.

Nous sommes donc « au bon endroit » pour proposer un travail collectif sur le territoire partagé par tous. A cet égard notre projet a reçu bon accueil lors de la journée de restitution. En revanche notre proposition de Bibliothèque Numérique butte sur le fait qu'un objet centralisé s'accorde mal aux logiques distribuées des acteurs. La B.N. s'adresse à une communauté, pas à une collectivité, ou alors pas sous la même forme. Nous avions misé sur le fait qu'un service en ligne était suffisamment ouvert pour que des pratiques se diffusent librement. Mais tout service en ligne repose tout de même sur des données centralisées, « quelque part ». Il y a donc forcément un acteur pour gérer, pour arbitrer... Cet acteur dispose qu'on le veuille ou non d'une forme de propriété ou au moins de responsabilité sur le système. Or c'est bien là que la dimension non-communautaire de nos acteurs se ressent. Car dans cette logique le partage se traduit non pas par l'utilisation d'un service commun, mais par la duplication et la diffusion des données. Une circulation sur un mode informationnel, à la façon dont on partage des mails ou des documents numériques, et non à la façon dont on se coopte sur un réseau social comme Facebook ou Myspace.

. *6.3. La transportabilité des données numériques complexes*

Dans notre projet comme dans d'autres, le fruit de bien des efforts se retrouve sous la forme d'une base de données.

Pour l'ingénieur, la mobilisation d'un tel outil n'est pas chose triviale. C'est un travail de spécialiste que de concevoir une structure de données, et le dispositif capable de l'accueillir. Dans une démarche scientifique telle que la nôtre, la question est la suivante : **comment construire des données qui réfèrent au terrain en toute rigueur ?** Les bases de données sont des outils avec lesquels il est possible de faire énormément de choses à faible coût, en argent mais surtout en temps. C'est pourquoi on les utilise. En revanche ces qualités ont un prix. Les bases de données sont facilement manipulables, calculables, mais ce sont des espaces étriqués, fortement contraints. C'est parce qu'ils sont si contraints qu'on peut faire tant de choses avec. Mais le terrain qu'on étudie, lui, est souvent tout l'inverse. Il est vaste et aux contours imprécis, il est multivoque, ambigu, intransportable, difficile à mesurer. C'est justement pourquoi on l'étudie : pour construire les faits qui nous donneront prise sur lui. Comme le dit Bruno Latour, les « données » devraient plutôt s'appeler des « obtenues », car c'est au prix d'un lourd travail que nous pouvons prétendre en disposer. Ce travail est lourd parce qu'il demande beaucoup de rigueur, parce que le propre du travail scientifique c'est de produire des données qui *réfèrent* au terrain. Il faut que les données désignent en quelque sorte le terrain, il faut qu'elles soient reliées à lui, sans quoi elles ne valent rien.

La rigueur scientifique pourra être évaluée par la robustesse de la chaîne d'opérations qui aura conduit le chercheur du terrain aux données. Cette chaîne de référence doit pouvoir se parcourir dans l'autre sens, des données au terrain, par celui qui s'intéresse aux résultats. Dans notre cas, il s'agit de comprendre en quoi les cartographies réalisées rendent compte du web comme terrain. Si on ne peut pas le comprendre, c'est que la chaîne n'est pas assez robuste, le travail scientifique pas assez rigoureux, les données potentiellement biaisées, pas assez fiables pour dégager des conclusions significatives. Le rôle de l'ingénieur dans ce travail est de s'assurer de la cohérence de la chaîne de référence, c'est-à-dire de réaliser un dispositif qui rende compte des spécificités du terrain comme de la méthode. Technique et méthodologie sont ici indissociables. D'un point de vue technique, ceci signifie que les données ne sont pas indépendantes de la façon dont on les produit et dont on les manifeste. Concrètement, une base de données n'est pas dissociable de son dispositif de lecture. Sans un logiciel d'accès aux données, ces dernières ne peuvent pas être considérées comme le fruit d'une méthodologie.

Pour RITIMO une autre question s'ajoute à celle-ci. C'est que les résultats de notre travail

concernent l'association, pour sa réflexion et pour son action. Les données ont une valeur en tant que telles, et la base est en quelque sorte leur lieu de stockage. Pas de données sans base de données, ou en tout cas sans « structure de données » au sens large (ça peut être un format de fichier par exemple). L'enjeu se trouve alors là où les contraintes techniques rencontrent la vie des données. En l'occurrence c'est la diffusion et la gestion des données qui compte. Comme nous l'avons vu le fait que les données soient publiques ou privées, centralisées ou distribuées, compte beaucoup pour les acteurs associatifs. **Comment maîtriser la circulation des données ?**

Concevoir un dispositif qui réponde aux deux questionnements relève de la gageure. Le simple fait de stocker et publier des données complexes est en soi une question difficile. Mais en plus de cela, la conception coince où il faut à la fois donner à voir quelque chose et rendre les données transportables. Plus précisément, il est facile de faire deux choses distinctes : stocker les données dans une base où on pioche pour produire des visualisations, et produire des visualisations et des indicateurs sous forme de documents qu'on peut transporter. Mais la base de données est difficilement transportable. Et les documents ne sont ni complets ni éditables. Donc d'un côté on maîtrise la construction des données mais on ne peut pas les transporter, et de l'autre on maîtrise la circulation des documents mais on ne peut pas en manipuler le contenu. Si les données étaient en petit nombre, nous pourrions remplacer la base par un document XML ; mais les données sont en trop grande quantité pour que ce soit faisable. Une base de données est également transportable au sens technique du terme. Mais le dispositif de lecture de cette base, c'est-à-dire le système de décodage qui permet d'accéder concrètement aux données, ne l'est pas. Dans le cas de données simples il serait possible de faire un petit programme qu'on déplace avec la base, mais les performances pour nos données sont si faibles que ce n'est pas jouable. En fait il est utile de travailler en profondeur les modèles mathématiques de stockage des données, sans quoi les performances rendent le système peu utilisable. C'est ce qui rend la création d'un dispositif transportable si difficile. C'est cette problématique que nous appelons la *transportabilité des données complexes* : **Comment stocker les données dans un dispositif transportable qui permette à un usager de consulter et d'éditer le contenu, dans de bonnes conditions, et malgré leur complexité ?**

Ce travail est de longue haleine mais le besoin qui en est à l'origine se retrouve dans de nombreux projets tels que les travaux de WebAtlas sur l'Arctique ou les Nanotechnologies, ou le e-Diaspora Atlas en collaboration avec le programme TIC Migrations à la Maison des Sciences de l'Homme. Les objectifs sont, dans l'ordre :

1) Décrire correctement l'ensemble de pages web stockées (structure de données de plus bas niveau). Il s'agit d'indexer les pages, représentées par leurs liens et leurs contenus, dans un arbre performant.
2) Articuler la définition d'un site web par dessus l'ensemble de pages indexées : couche de qualification, de description. A ce niveau les données du corpus sont stockées. Dans cette approche il est possible de déterminer précisément ce qu'est un site (contrairement à notre approche actuelle où site=domaine).
3) Construire une API[4] de lecture et d'édition de ces données, et l'implémenter dans plusieurs langages comme Java, Perl... A ce stade les données forment un système autonome (module) qu'on peut mobiliser dans des projets.
4) Fabriquer un logiciel qui intègre ce module pour transporter les données. Ce logiciel permet notamment de générer les indicateurs que l'on souhaite, et de naviguer dans les données. Le logiciel intègre en lui-même les données, il est donc transportable.

Ce type de système est exigeant à réaliser car il demande d'implémenter des structures de

4 Interface de Programmation Applicative

données particulières, ce qu'on confie habituellement à un logiciel de bases de données tout fait (comme MySQL). Il requiert donc des compétences de haut niveau en mathématiques et en informatique. En substance il s'agit de rien de moins que de réinterpréter les couches profondes d'un moteur de recherche.

Par contre les avantages d'un tel dispositif valent l'effort demandé. Car par sa nature de « moteur » il est à même de fournir les performances nécessaires à un usage « grand public ». Et par sa modularité, il peut néanmoins communiquer efficacement avec d'autres systèmes. On peut l'utiliser comme un fichier pour transporter les données ou comme un logiciel autonome pour consulter les contenus.

La grande différence entre cette solution et la Bibliothèque Numérique, c'est que la B.N. est nécessairement hébergée à un seul endroit. La solution que nous proposons permet au contraire de diffuser les données, en même temps que les moyens de les maîtriser. Puisque la transportabilité des données est un enjeu important pour le milieu associatif, nous pensons qu'il faut repenser la B.N. autour de cette contrainte. Ce qui signifie qu'il faut dégager de nouvelles articulations entre le pool collectif de données et des poches de données maîtrisées localement par des acteurs différents. Il faut permettre la collectivisation sans pour autant forcer les acteurs à partager leurs ressources. La dimension collective relève alors d'un choix « politique », qui n'est plus pris en otage par la technique et sa prescription communautaire.

. 6.4. En guise de conclusion

Ce projet est un commencement. Il a produit le corpus web le mieux maîtrisé jamais rendu public à notre connaissance, et il préfigure des travaux que des documentalistes seront amenés à faire lorsque leurs pratiques professionnelles évolueront pour se tourner vers une plus grande prise en compte du web. Notre projet a joué le rôle de test en grandeur nature, pour mettre en œuvre les théories du web comme réponses possibles à des enjeux non-scientifiques, dans notre cas ceux de la société civile. La confrontation entre les acteurs effectifs d'un domaine et le travail scientifique censé en rendre compte met au jour un certain nombre de points clés dans l'approche de la cartographie du web.

D'abord le simple fait que ce projet existe, et soit rendu public, a un impact. Comme ce type d'approche est innovant à plusieurs titres (approche peu connue en Europe, nouveaux outils, collaboration « sciences-société »), la dimension communication et pédagogie a une importance particulière. Ce projet a permis de sensibiliser plusieurs acteurs de la société civile à certaines problématiques propres au web et, nous l'espérons, aura ouvert quelques perspectives quant à ce qu'on peut faire avec ces nouvelles technologies. Le débat est ouvert et nul doute que la société civile s'emparera des approches qu'elle jugera pertinentes. En particulier les cartographies du web, qui permettent de comprendre de quelles façons on peut brasser de grandes quantités de données numériques, font partie d'une culture émergente. Ces objets sont aujourd'hui difficiles d'accès mais peu à peu ils se diffuseront et sembleront d'un abord plus simple.

Nous nous plaçons donc dans une perspective de démocratisation des cartographies, et nous avons dû modifier nos outils pour qu'ils s'accordent à cet enjeu. Et pour commencer parlons d'un petit bouleversement à l'échelle d'un petit bout de science. Sachez, si vous ne le saviez pas, que la théorie des graphes a longtemps prospéré sans que quiconque n'éprouve le besoin de se représenter la chose autrement que sous la forme d'une matrice d'adjacence. La représentation visuelle, et donc la spatialisation de graphes, est un exercice relativement récent. Mais même dans cette perspective les algorithmes de spatialisation sont longtemps restés des instruments très analytiques. Autrement dit ce travail s'apparentait à de la modélisation pure, et dure. Dans cette logique là, faire une carte avec un graphe s'apparente à un détournement. Mais précisément pour faire une carte plus lisible on est obligé de modifier jusqu'aux algorithmes de spatialisation. Par exemple il est nécessaire de faire

en sorte que les nœuds ne se recouvrent pas, pour que l'œil les distingue bien ; il est aussi nécessaire que les textes ne se chevauchent pas. Ces deux interventions, et d'autres plus subtiles, se jouent dès l'algorithmique de la spatialisation. Ce faisant on sort de la modélisation scientifique pour entrer en quelque sorte dans la sphère de la communication visuelle. Mais c'est aussi le rôle de la science de se rendre intelligible auprès de ceux qu'elle concerne.

Les cartographies suscitent toujours des réactions, elles ne laissent pas indifférent. Mais ce sont des objets ambivalents, et le signe en est qu'elles peuvent être parlantes sans être directement compréhensibles. Elles représentent la plupart du temps des données complexes, et si elles les rendent accessibles, elles ne simplifient pas non plus cette complexité dans laquelle on peut se perdre. C'est pourquoi elles ne sont pas « lisibles », mais paradoxalement (ou pas) c'est ce qui fait leur valeur. Un histogramme ou un camembert sont esthétiquement des images à lire, elles n'informent pas outre ce qu'elles manifestent intentionnellement. Nos cartographies sont des compte-rendus où le regard peut, et doit même, se perdre et naviguer. Nous proposons dans ce projet un usage ouvert, créatif pourrait-on dire, des cartographies. Mais nous cherchons bien sûr des réponses à la question de leur sens, sachant que ce type d'approche permet d'avancer plusieurs éléments. En particulier nous avons pu observer l'intérêt des cartes pour expliquer la nature complexe du web et des réseaux sociaux en général, ainsi que pour permettre aux acteurs d'échanger autour d'un support qui « révèle » ce qui autrement n'apparaît pas.

En deçà des cartographies, le travail même de constitution du corpus, qui nous a pris beaucoup de temps, est riche d'enseignements. En premier lieu il nous a forcé à revoir et ajuster l'ensemble d'outils nécessaires. Si on ne peut pas encore dire qu'ils sont suffisamment évidents pour que n'importe qui s'en empare, de grands pas ont été faits vers une démocratisation de l'exploration du web. Ce bénéfice profite d'ailleurs en premier lieu aux chercheurs que WebAtlas forme pour analyser le web dans leur propre domaine, et d'ores et déjà certains acteurs de la société civile s'en emparent (par exemple la Fondation pour le Progrès de l'Homme).

En second lieu, un corpus de grande qualité est précieux. Il est précieux pour la science, car il permet à des experts (de la solidarité internationale) d'évaluer ce qui manque à nos méthodes, et en l'occurrence nous avons vu que les crawlers ne sont pas satisfaisants sous leur forme actuelle. Mais il est aussi précieux pour les gens qu'il concerne. Plus un corpus est de bonne qualité plus il est pérenne. Même si les acteurs peuvent apparaître ou disparaître, la structure générale observée demeure, ainsi que la répartition des thèmes. Par ailleurs il est bien plus facile de modifier un corpus existant que d'en refaire un de zéro. Un corpus de grande qualité est un investissement. Enfin précisons qu'un tel corpus sera une trace précieuse de l'état de notre domaine à un moment donné, sur laquelle on pourra revenir dans le futur pour faire des comparaisons.

Le web recèle des ressources insoupçonnées. L'une des qualités de notre travail en profondeur est de répertorier des acteurs qui autrement sont sous-représentés. Des acteurs qui par exemple ont un « petit » site web, et qui sont négligés par les moteurs de recherche. De manière générale les moteurs de recherche comme Google ne sont pas aussi fiables qu'on le croit. Beaucoup de ressources leur échappent et leurs critères de classification sont fortement contraints par les performances qu'ils doivent obtenir. Songez qu'en quelques dixièmes de secondes ils sont censés retrouver ce qu'il y a mieux dans le web tout entier ! L'intérêt d'un corpus est aussi de pouvoir construire un système plus fiable, plus fin et mieux maîtrisé, en complément des autres moteurs comme Google. Certes nous n'en sommes pas encore là, parce que c'est difficile, donc exigeant, donc cher. Donc long ; mais petit à petit nous nous en rapprochons. Nous avons présenté un tel prototype lors de la journée de restitution, proposé par notre partenaire de recherche, l'association Magellan. Comme nous l'avons déjà largement évoqué, la centralité du dispositif pose un certain nombre de problèmes à la collectivité des acteurs. Mais malgré cela, les personnes présentes ont réagi positivement et c'est peut-être là le début d'une prise de conscience : il est possible à terme d'avoir prise sur son territoire informationnel, fût-il aussi complexe que le web. Et c'est là un enjeu

que nous pensons important pour la société civile.

7. Suites à donner à ce projet

La Bibliothèque Numérique n'existe pas mais nous n'avons pas abandonné pour autant !

Avant toute chose, il est important d'asseoir un socle technologique plus solide avant d'envisager d'aller plus loin. En premier lieu il est nécessaire de disposer d'une technologie de crawl efficace et maîtrisable. Ce sont sans doute les technologies de Magellan qui à cet égard sont les mieux placées. En second lieu il faut régler la question de la portabilité des données numériques complexes.

Dans un cas comme dans l'autre, WebAtlas est impliquée dans des projets similaires qui requièrent les mêmes besoins. Ces deux préalables sont donc en chantier.

Ensuite se pose la question de la collectivité des acteurs concernés par ce travail. Leur réaction est positive et selon l'implication de chacun, de petits projets émergent localement. A ce niveau nous pensons qu'un temps de réflexion est nécessaire.

En effet la principale difficulté concerne la capacité des acteurs à s'emparer d'une technologie web « complète ». A cet égard l'approche qui consiste à proposer un service « clefs en mains », même si elle n'a pas complètement abouti, se révèle ne pas être la meilleure solution. Des décisions d'ordre politique doivent être prises et des ajustements sont nécessaires avec des projets existants.

Malgré tout, les acteurs que nous avons rencontrés ont bien compris l'intérêt de disposer d'un moteur de recherche « maison », basé sur un corpus bien maîtrisé qui existe aujourd'hui. En des temps plus favorables aux associations, des projets concrets auraient émergé dès la journée de restitution. En particulier l'acteur le plus logique pour supporter un tel projet, Globenet, s'est révélé très sensible à la démarche. Malheureusement cette association connaît d'importantes difficultés et ne peut donc le prendre en charge pour l'instant.

Cette aventure aura largement participé à réorienter l'approche de WebAtlas. Les principaux enseignements bénéficieront directement à ses projets, et des technologies plus modulaires, plus en accord avec la logique des acteurs, seront bientôt disponibles.

Il sera alors temps de reformuler un projet avec les acteurs qui se révéleront les plus impliqués, sur la base des petites collaborations qui ont lieu dans la lancée de ce projet.