AnCaraS : logiciel d'analyse de l'édition web

David **REYMOND** (*), Brahim **REBAI** (**), Annie Estelle **BORDIER** (*)
mailto:david.reymond@iut.u-bordeaux3.fr, brahim-khalil.rebai@etu.u-bordeaux1.fr, annie-estelle.bordier@u-bordeaux3.fr

(*) MICA-GRESIC, MSHA, Université de Bordeaux 3, 33407 PESSAC, FRANCE,
(**) IMS, 351 Cours de la Libération, F-33405 TALENCE CEDEX, FRANCE

Mots clefs :
Webométrie, Collecteur web, Visualisation de graphes, Analyse des sites web, Indicateurs, Collecte d'informations
Keywords:
webometry, webdatamining tool, graph visualization, web site analysis, web indicators
Palabras clave:
cybermetría, Herramienta de minería web, análisis de sedes web, indicadores.

Résumé

L'outil de webométrie AnCaraS (Analyse et Caractérisation des sites web) s'inscrit comme livrable d'un projet de recherche interdisciplinaire visant à produire des indicateurs de performance de la communication via le média Web[1]. La version prototype de cet outil permet la collecte d'une zone DNS d'un site ou d'un groupe de sites web puis la visualisation d'indicateurs informétriques ainsi que la représentation cartographique des données. Il est donc un logiciel d'analyse de sites web et de leurs réseaux relationnels et se situe par ailleurs comme un outil de régénération complémentaire de méthodes d'évaluation des sites web. En mettant en œuvre les avancées du domaine de la webométrie, les caractérisations produites par AnCaraS sont aussi bien à destination des webmestres et responsables éditoriaux afin de faciliter les actions de maintenance, de suivi, d'alimentation, de syndications et d'affiliation des sites, qu'à d'autres utilisateurs à responsabilité de veille stratégique ou de pilotage d'un ensemble de sites ou d'organisations de grande taille.
Dans cet article nous présentons les spécifications de la première version d'AnCaraS ainsi que son utilisation à travers une étude de cas visant à évaluer l'incidence de la publication professionnelle externe sur la visibilité des sites web universitaires en prenant pour terrain d'étude une des universités d'Aquitaine.

[1] Axe 3 du projet de recherche RAUDIN, *Recherches Aquitaines sur les Usages pour le Développement des dIspositifs Numériques* financé par : Feder N°31462, Conseil Régional Aquitaine et l'université Bordeaux 3. http://raudin.u-bordeaux3.fr/

Introduction :

Le développement de l'outil AnCaraS (ANalyse et Caractérisation des Sites web), s'inscrit dans l'axe 3 du projet RAUDIN (Recherches Aquitaines sur les Usages pour le Développement des dIspositifs Numériques), axe qui vise à produire des indicateurs de performance permettant le déploiement des services numériques de qualité. Un focus particulier est donné à l'édition web, marquée par l'insertion d'hyperliens entre les contenus qui supprime toute possibilité de considération classique et surtout linéaire des traditionnelles analyses de contenus. Les différentes fonctionnalités d'AnCaraS permettent la collecte d'une zone DNS[2], d'un site ou d'un groupe de sites web puis la visualisation d'indicateurs informétriques et la représentation cartographique des données. Il est ainsi un outil d'analyse des sites, de leur réseau relationnel avec pour objectif la régénération des méthodes d'évaluation des sites web. En mettant en œuvre les avancées du domaine de la webométrie les caractérisations produites par AnCaraS sont aussi bien à destination des webmestres et responsables éditoriaux afin de faciliter les actions de maintenance, de suivi, d'alimentation, de syndications et d'affiliation des sites qu'ont d'autres utilisateurs à responsabilité de veille stratégique ou de pilotage d'un ensemble de sites ou d'organisations de grande taille. Nous présentons ici la spécification ainsi que le développement d'une première version d'AnCaraS.

D'abord, nous abordons un rapide état de l'art présentant les différents outils existant issus du logiciel libre, en présentant les différents aspects de l'analyse des sites :

- Collecte de données (sources directes et indirectes via les index) ;
- Extraction, traitement (agrégation à différentes échelles) et génération d'informations (calculs informétriques) ;
- Visualisation et manipulation des graphes.

Puis, nous présentons l'architecture générale d'AnCaraS en décrivant ses différentes fonctionnalités. Enfin, nous mettons en application une analyse de notre terrain d'étude, les sites web des établissements universitaires d'Aquitaine, et situons donc AnCaraS comme outil s'insérant dans l'évaluation de l'édition web. Le logiciel ouvre la voie à des utilisations en monitoring de sites ou de groupes de sites, de veille stratégique concurrentielle en regard de l'étude des réseaux d'affiliations, et de perception globale de données hypertexte.

1 État de l'art

Il existe sur le marché de nombreux outils de collecte de données sur le web et de cartographie. Dans un objectif de développement d'un outil adapté pour la recherche scientifique nous avons comparé les caractéristiques des principaux logiciels libres ou issus du monde de la recherche tel *SocSciBot* de Thelwall [19].
A part ce dernier, allié au logiciel de traitement de graphes *Pajek*, nous n'avons pas trouvé à la date de démarrage du projet, de logiciel collecteur adjoint à un outil de visualisation de graphes que nous puissions adapter à nos besoins de collecter, traiter/agréger, visualiser/manipuler. Nous distinguerons donc dans cet état de l'art les collecteurs (ou *spiders*) web, et les logiciels de visualisation de graphes.

1.1 Les collecteurs

Poétiquement appelés les « chalutiers du web », le lecteur intéressé par une description approfondie des modes de collectes (algorithmes de parcours et de choix informatiques) nécessaires à leur implémentation pourra en référer à Mathieu [14] et à Castillo [8].
La Figure 1, d'après Arroyo [2] présente les différents collecteurs retenus selon les trois caractéristiques essentielles des collecteurs web : fraicheur des données, qualité intrinsèque ou qualité représentationnelle. Les collecteurs de la recherche scientifique se situent dans ce dernier quadrant, en opposition aux collecteurs

[2] Domain Name Server

veilleurs, en quête des dernières nouveautés qui se situeraient (schématiquement), et aux collecteurs de mise en miroir. La Figure 1 permet ainsi de situer les caractéristiques des différents collecteurs et nous avons positionné les différents collecteurs retenus (cf. première ligne du Tableau 1). Ces caractéristiques vont à l'évidence influencer son mode de programmation (privilégier la vitesse sur la qualité de représentation ou de parcours, influence les choix algorithmiques de parcours des graphes). Les critères des choix sur la possibilité d'adapter le collecteur à des traitements spécifiques ne seront pas abordés ici. Nous retenons comme critères son caractère de « libre », le format des données produites pour leur exploitation, la possibilité de scripter le logiciel pour des traitement d'envergure (planification hors heures de bureau, collecte depuis de multiples postes), sa modularité qui ouvre la possibilité simplifiée d'insertion de modules de traitements spécifique, la possibilité d'adaptation à de grandes tailles de site (par exemple les sites des universités), et enfin l'éventuelle présence de calculs informétriques simples (dénombrement de fichier, typages).

Tableau 1 : Comparatif des différents collecteurs (spiders) web selon les critères spécifiques à nos besoins.

Logiciels / **Critères**	SocSciBot	Xenu	*Toke*	Webcheck	Httrack	Wget	cURL	SiteSucker
Libre	Pour la recherche Code non ouvert	✓	✓	✓	✓	✓	✓	✓
Format des données produites	Txt + Pajek	HTML	Pajek	HTML	Arborescence	Arborescence	Arborescence	Arborescence
Scriptable	-	-	-	✓✓	✓	✓	✓	-
Modulaire	-	-	-	✓✓	-	-	-	-
Grandes tailles de graphe	✓	-	-	✓	-	✓✓	✓✓	✓
Calcul informétriques	✓	✓	-	-	-	-	-	-

Le Tableau 1 présente les caractéristiques des différents collecteurs recensés selon ces critères. A part *SocSciBot*, dans une version récente sortie après le démarrage de nos travaux, aucun de ces collecteurs ne présente de possibilité directe de visualisation de données collectées : cartographies des sites et production de données informétriques simples. Souffrant de cette lacune également, *WebCheck* apparaît comme la suite logicielle la plus adéquate à nos besoins mais ne sera pas retenue du fait du langage utilisé (le langage Python).
Le principe retenu sera de développer un outil de collecte qui possèdera les caractéristiques précédentes et qui permettra de nous appuyer sur un outil de visualisation des graphes en garantissant de l'interopérabilité des données produites lors de la collecte. Ce choix nous garanti également de développer les compétences nécessaires à la collecte de données, et aux traitements adjacents de nettoyage ou agrégation qui ne peuvent être évités de par la diversité des productions (format, encodages, choix de composition des pages) et d'ajuster éventuellement au fur et à mesure des évolutions.

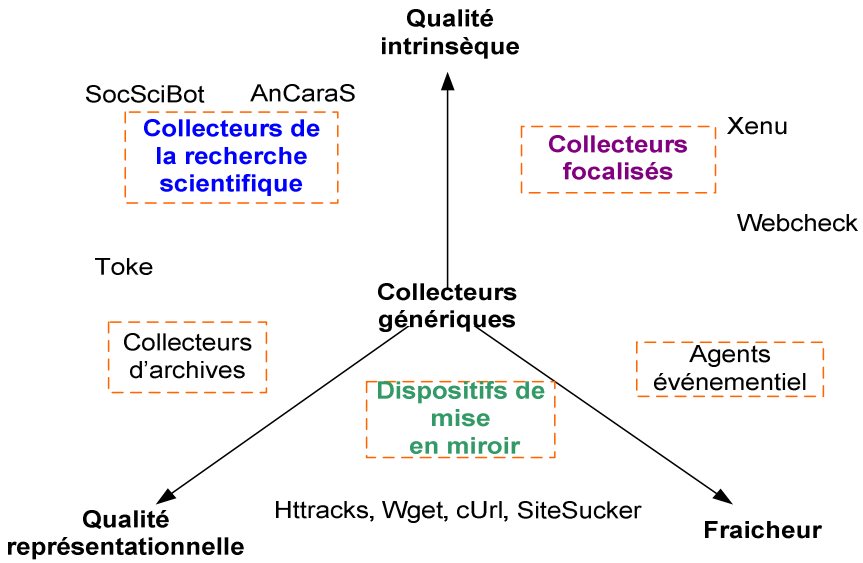

Figure 1: Les collecteurs web selon leurs domaine d'application et les qualités attendues

1.2 La visualisation de graphes

Afin de rendre possible la visualisation des graphes collectés via le Web, il est nécessaire d'inclure une gamme de prétraitements informétriques (filtrage de contenus, agrégation à des échelles permettant de consolider l'appréhension des différents réseaux hypertextes. Nous utilisons le modèle ADM [19,7] de Thelwall pour appuyer les différents niveaux d'aggrégation. Ainsi, au plan conceptuel, s'insère ici un module de traitement de données qui exclut tout collecteur du web n'offrant pas la possibilité de manipuler et traiter les graphes produits. Le format de sortie du collecteur doit en outre s'adapter aux formats standards de graphes afin de pouvoir être importé par un logiciel adéquat.

A l'aide du site "Graph Visualization Software Reference", nous avons pu présélectionner [16] quelques logiciels/bibliothèques de traitement de graphes. Le Tableau 2 récapitule les critères validés appliqués aux suites logicielles ou bibliothèques présélectionnées mettant en exergue les qualités recherchées pour notre objectif : le caractère open-source du produit, les types de formats de données entrées/sorties, le caractère scriptable, modulaire et adaptés aux grandes taille. Conformément aux desseins de production de connaissance [3] à partir de la topologie hypertexte, ou de simple capture d'information [9], il convenait aussi offrir à l'utilisateur une grande interactivité, la navigation éventuellement 3D et de pouvoir gérer des graphes de graphes selon les échelles de navigation utilisées.

Tableau 2: Comparatif de quelques logiciels de traitement de graphes

Critères / Logiciels	Libre	Format des données entrées	Format des données sorties	Scriptable	Modulaire	Adapté à de grandes tailles	Plus court chemin	Graphe de graphes	Navigation	Interactivité	3D	Manipulation
Tulip	Oui	Tlp,dot,GML	Tlp, GML	-	✓	✓✓✓	✓	✓	✓✓✓	✓✓✓	✓	✓✓✓
Pajek	Oui	Pajek, UCINET	EPS, PS, BMP, SVG, HTML, X3D, VRML	-	-	✓	✓	✓	✓	✓✓	✓	✓
Graphiz	Oui	Dot	dot, PS, SVG, images, PDF, GXL	✓✓	-	-	-	-	✓	-	-	-
Gephi	Oui	GDF (GUESS), GraphML, XGMML, Pajek		✓✓	✓✓	✓	✓	✓	✓✓	✓✓	✓	✓✓

1.3 Bilan

Les logiciels *Httrack*, *Wget*, *cURL* et *SiteSucker* ne sont guère utilisables car ils se contentent de recopier les fichiers du site Web sur le disque dur, sans créer le graphe des URLs et sans présenter ni de rapports ni de statistiques. Les logiciels *Xenu* et *Webcheck* créent bien des rapports statistiques sur le site, mais ces rapports sont au format HTML (difficiles à exploiter dans un autre logiciel), et ne créent pas le graphe des URLs. Le logiciel *Toke* n'existe ni en français, ni en anglais, ce qui le rend impossible à utiliser. Le logiciel *SocSciBot* est adapté à notre projet, on ne peut cependant pas lui rajouter de nouvelles fonctionnalités (nouveaux rapports, statistiques, …). **Ce logiciel servira de référence pour nos tests de performance et de fiabilité du collecteur et du traitement post collecte.** Webcheck, au plus près de nos attentes est développé en Python, nous avons opté pour le développement d'un spider en langage Java, interprété et utilisable sur tout système d'exploitation, le développement modulaire, spécifié au format DEVS, nous permet également d'ouvrir la possibilité d'utiliser le spider en mode Web Service, bien qu'actuellement seule la version sur poste client est disponible. Côté visualisation des graphes, nous avons choisi le logiciel Tulip, seul capable de traiter des graphes de très grande taille, modulable et dont les puissantes librairies C++ seraient bridées aux fonctionnalités souhaitées.

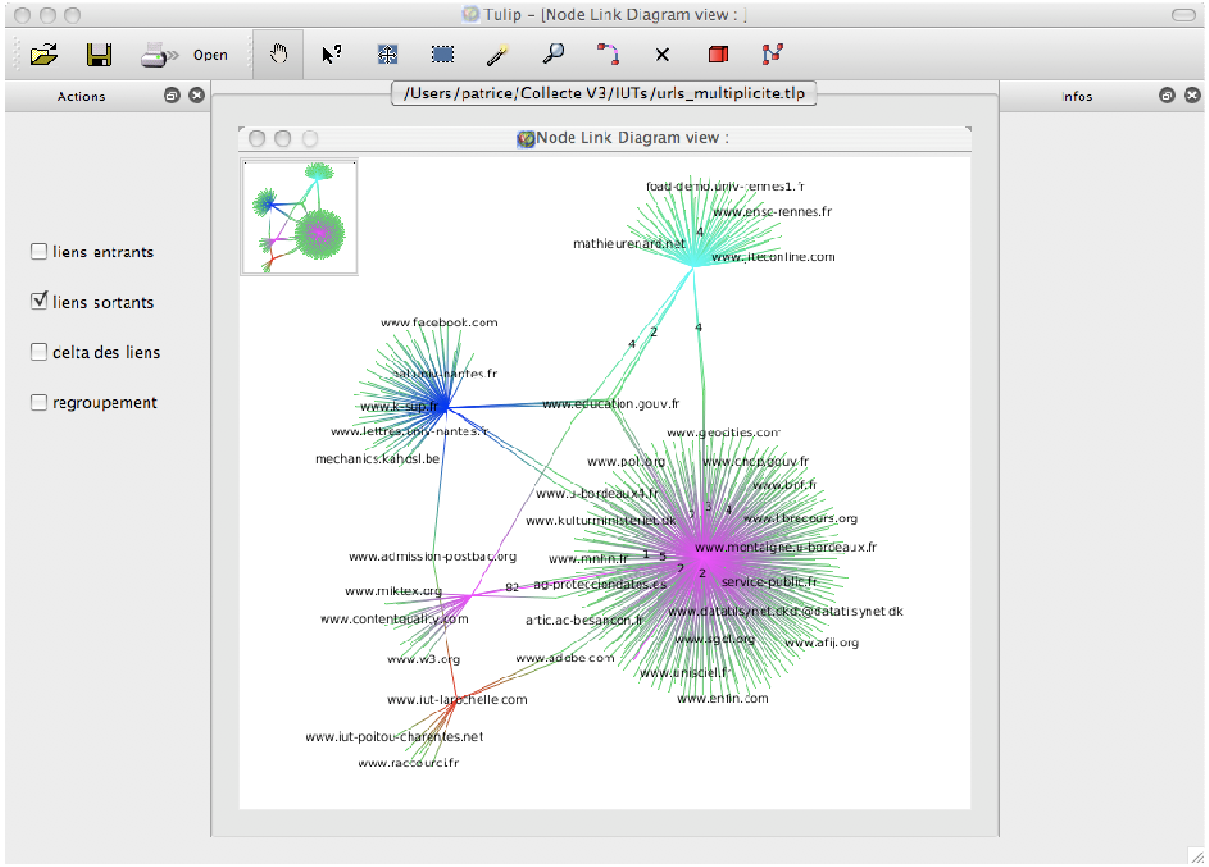

Illustration 1: Exemple de visualisation d'un réseau relationnel existant entre 4 structures universitaires. L'outil AnCaraS permet le choix des relations à présenter (colonne de gauche), le changement d'échelle (absent dans cette portion d'interface), la navigation et la manipulation du graphe grâce à Tulip.

2 AnCaraS : un outil d'analyse de l'édition Web

AnCaraS est composé de différents modules en charge des différentes fonctionnalités décrites précédemment. Il rassemble en un seul outil ces fonctions de collecte, de traitement et est un outil d'analyse de données basé sur la webométrie. Les fonctions de visualisation et de manipulation de ces données sous forme de graphes, issus de Tulip rajoutent une dimension interactive utile pour traiter de la complexité des graphes engendrés par ces collectes de données.

La Figure 2 présente l'architecture d'AnCaraS. Le logiciel dispose d'un collecteur multi instances prenant en entrée un ou plusieurs URLs et des règles de collecte (profondeur interne du site et de la zone DNS pour le suivi récursif d'url, la possibilité de bannir des URL ou des sites) qui contraindront le périmètre de la collecte.

Le module extrait récursivement les URLs des pages visitées qu'il stocke dans la pile des URLs à visiter, sous réserve d'approbation du système de gestion du périmètre. En option les contenus textuels seront extraits et stockés. Enfin, il est possible de compléter la collecte en recherchant les liens entrants externes (*backlinks*) en utilisant les moteurs de recherche (Yahoo, Google...) via les API des index.

À la fin de ce processus, le module de traitement des données prend le relais et produit les caractérisations de différents niveaux agrégés selon un modèle inspiré du modèle ADM *(Advanced Document Model)* de Mike Thelwall [19], revisité par Björneborn [7]. Les données de caractérisations informétriques (dénombrement et typage des fichiers) et les graphes des réseaux relationnels issus du traitement sont exportés au format XML.

Le dernier module permet la visualisation des caractérisations précédente en parallèle des graphes de réseaux générés et des statistiques relationnelles (liens entrants et sortants des différentes unités web) à l'aide du logiciel Tulip. L'outil Tulip développé par le LaBRI [3], permet ainsi d'accéder à une gamme de représentation des graphes ainsi que leur manipulation interactive favorisant l'accès à la topologie relationnelle des sites étudiés trouve de nombreuses applications en traitement des graphes en général. *AnCaraS* lui confère la capacité de produire des cartographies Web selon des modèles d'analyse établis en webométrie, pour sa dimension de statistiques et dénombrement des contenus pour évaluer l'impact des productions [1] ou pour sa dimension réseaux relationnels [20] pour évaluer le maillage inter-unités web.

Ainsi, les différents modules du logiciel couvrent l'instrumentation et leur interopérabilité nécessaire à l'étude des sites web par projection des éditions dans leur mode textuel en intégrant la dimension hypertextuelle. Les données de caractérisations produites permettent ainsi d'englober la volumétrie et la complexité des productions et d'appréhender une visualisation inédite de ces productions d'information et de communication en cohérence avec les indicateurs clés [9] utiles à leur amélioration : taille, impact visibilité et rayonnement essentiellement sont facilement appréhendés pour décrire les diverses unités du web. Les différentes mesures produites permettent sur un groupe de sites web :

- De détecter les liens morts,
- D'analyser la topologie relationnelle interne (mesure des affiliations, syndication et partenariats),
- D'analyser la topologie relationnelle externe à un groupe de sites (luminosité du site et produire un typage des sites « pointés », cités, citant et co-associés),
- D'analyser l'impact relationnel du groupe de site par extraction d'une partie des sites « citant » le groupe.

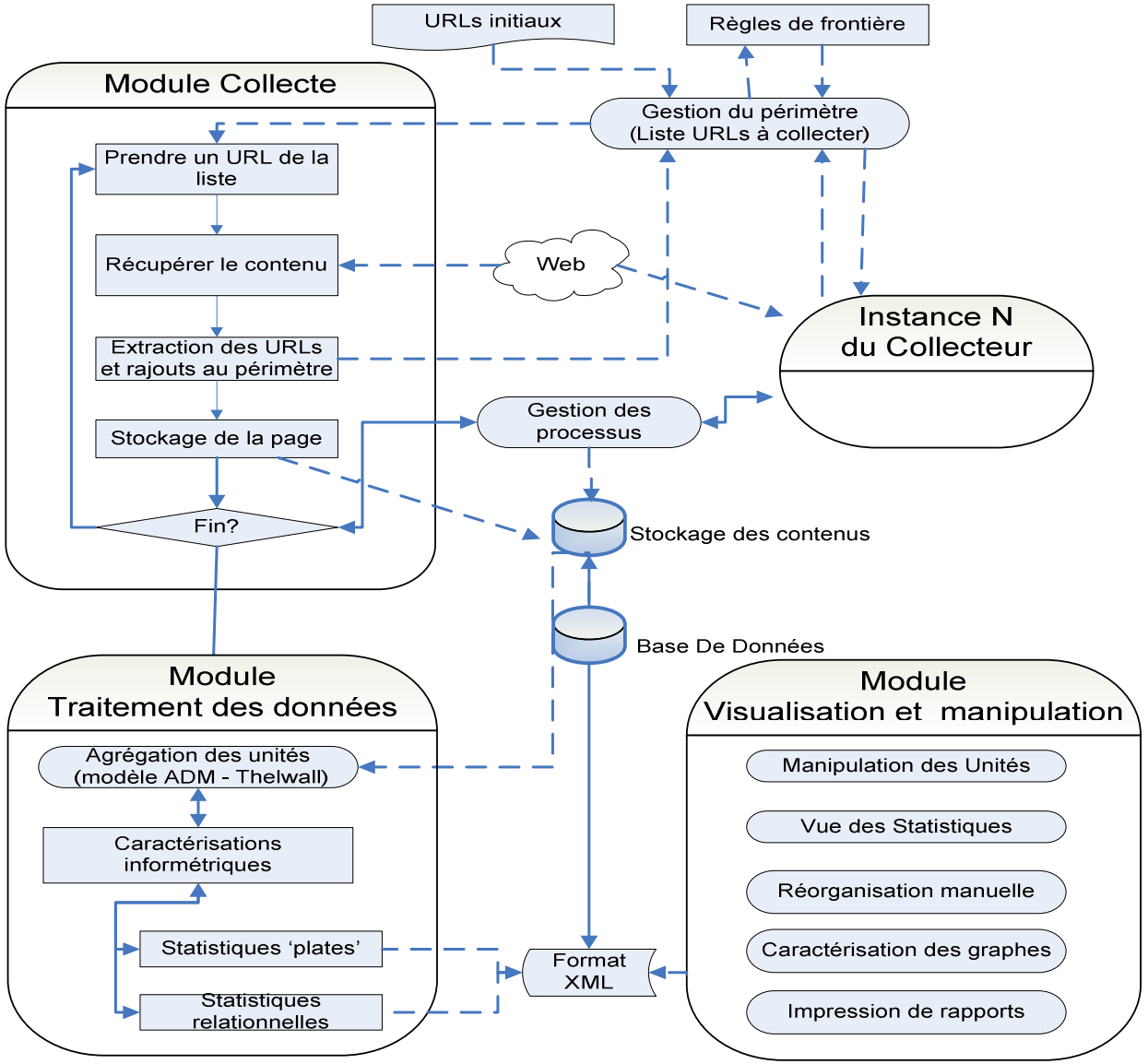

Figure 2 : Architecture générale du logiciel AnCaraS

3 Étude de cas : évaluation du rayonnement des sites web universitaires d'Aquitaine

En continuité d'une étude des pratiques éditoriales du personnel des établissements membres de l'université d'Aquitaine, nous mettons ici en application une analyse d'une partie du terrain du projet RAUDIN. En écho à d'autres études [4,[10,[11,15] qui soulignent que les pages personnelles hébergées par les établissements favorisent, tout au moins en Israël et en Grande-Bretagne, la visibilité[3] de la zone éditoriale de leur établissement, nous visons dans notre cadre à estimer le phénomène d'incidence positive ou négative éventuelle de productions professionnelles mais hébergées à l'extérieur de la zone éditoriale, sur la visibilité de la zone éditoriale (définie sur notre terrain comme sa zone DNS).

L'étude de cas porte sur les données collectées de plusieurs sites web mentionnés par les membres d'une université d'Aquitaine au cours d'une enquête (cf. infra) qui ont déclaré publier dans la zone éditoriale de leur établissement à titre professionnel et ont également annoncé qu'ils produisaient des contenus à titre professionnel mais hébergés à l'extérieur de cette zone. Ainsi certaines productions des membres font partie de la zone éditoriale web alors que d'autres, définis ici comme étant « externes » à cette zone constituent des espaces de publications pouvant potentiellement participer à sa visibilité. Il s'agit donc, à travers essentiellement une analyse structurale du réseau créé par les liens et les relations [12], d'évaluer les répercussions précises de ces productions web sur la zone éditoriale des universités : dans quelle mesure les sites Web « externes » participent-ils ou non à la visibilité de la zone éditoriale des sites Web universitaires ?

Dans un premier temps, nous évaluerons la cohérence interne de la zone éditoriale de Bordeaux 1 en effectuant une collecte exhaustive de celle-ci. Nous espérons recueillir des données concernant le nombre d'hyperliens entre les différents sous-sites et le site institutionnel de façon à observer plus précisément les différents rattachements entre ceux-ci. Ceci nous permet de dresser un état descriptif de la zone éditoriale.
Puis, en effectuant une collecte qui réunit les sites externes cités par les personnels et le site Web de Bordeaux 1 nous pourrons visualiser l'interconnexion existante entre ces différents sites et mesurer le degré de participation de ces sites externes au rayonnement de la zone éditoriale de l'université. Dans le contexte organisationnel universitaire, une telle étude permet de définir et d'évaluer l'édition Web de l'université, de repérer ses faiblesses mais aussi ses points forts afin de modifier et renforcer par la suite sa stratégie en accord avec les objectifs de communication de la structure.

Nous montrons ainsi qu'AnCaras, qui permet de visualiser le réseau relationnel d'une ou plusieurs structures à un instant T, s'insère dans la veille et le pilotage stratégiques autour de l'édition Web et s'adresse donc à toute personne ressource en charge de la communication Web, de son pilotage et/ou de son évaluation. Les hyperliens entre les différents sites peuvent être d'ordre structurels (le site institutionnel vers une composante tel une UFR ou un laboratoire), fonctionnels (le site d'un UFR vers le service d'inscription en ligne de l'université, vers les différents services de documentation, etc.). Peu de place est réservée (statistiquement) aux hyperliens de type personnel que l'on peut trouver dans les productions privatives. La présence ou l'absence d'hyperliens témoignent partiellement de l'organisation interne, sur ce paradigme, la visualisation inter-site permet ainsi d'établir, mesurer de cette cohérence et dresser tout au moins un état descriptif.

La zone éditoriale sur laquelle nous portons notre étude est l'université de Bordeaux 1 qui se compose de 73 sites (72 sous-sites et le site institutionnel). Notre première collecte, qui porte essentiellement sur cette zone, nous permet de visualiser et vérifier la connectivité existante entre les sous sites et le site institutionnel[4]. Nous repérons ainsi par exemple les sites internes isolés qui finalement aurait pu participer à ce maillage inter-site. Ainsi, en offrant la possibilité à ses acteurs de publier dans sa zone éditoriale, l'université accroît sa cohésion ainsi que sa visibilité sur le Web. L'*Illustration 2* montre ce réseau. AnCaras permettant la navigation à l'intérieur du graphe généré et il est possible de qualifier les différentes relations observées.

---

3 La visibilité d'une unité Web se définit [1] par le nombre d'hyperliens que cette unité reçoit.

4 AnCaraS recense 2759 liens entre les sous-sites et le site institutionnel de Bordeaux 1

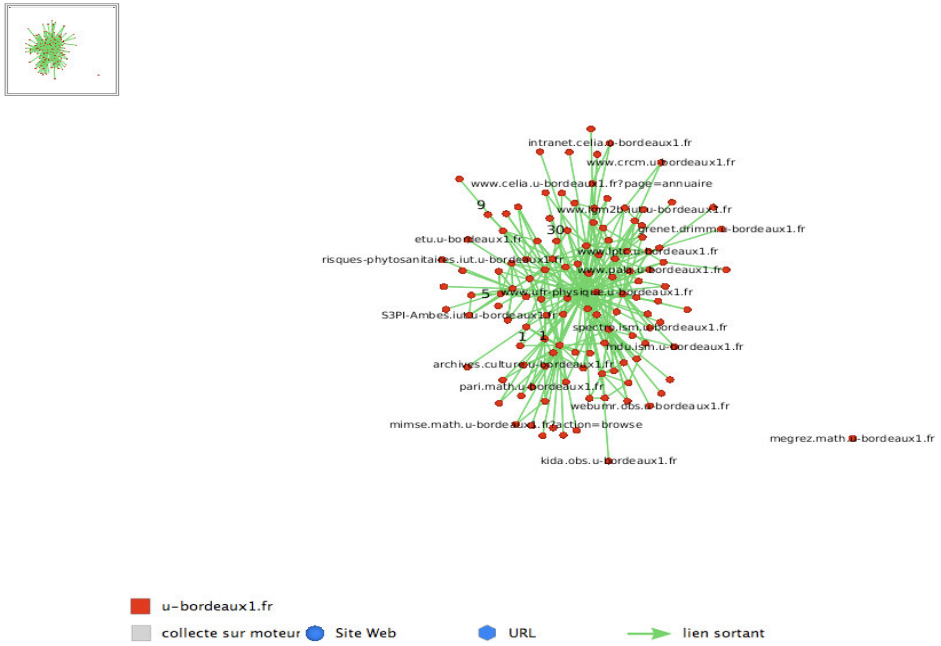

Illustration 2 : Graphe représentant le maillage inter-sites de la zone éditoriale de Bordeaux 1

L'enquête de terrain que nous avions mené en janvier-février 2009 auprès des acteurs[5] de l'université de Bordeaux 1 entre autres[6] afin de connaître leur comportement face à la production de contenu Web, a pu mettre en évidence deux points concernant les pratiques éditoriales de ceux-ci. Concernant Bordeaux 1, le taux de réponse à l'enquête, de 5,51%, reste relativement faible mais il s'agit aussi de l'université où les acteurs ont mentionné le plus d'adresses URLs hors zone éditoriale. Parmi les répondants, 50,7% de ceux qui publient à titre professionnel sur la zone DNS de l'université publient, au même titre, sur des espaces Web externes. D'un autre côté, 54,5% de ceux qui ont déclaré ne pas publier sur la zone éditoriale de l'établissement à titre professionnel, publient pourtant sur des sites Web externes dont ils nous ont indiqué les URLs. Nous tentons donc de vérifier s'il existe des liens entre ces sites déclarés et la zone DNS de Bordeaux 1 afin d'apprécier leur participation à la visibilité de l'université. Nous effectuons dans un premier temps une collecte simple[7] des 16 URLs 'externes'. Celle-ci nous permet de voir que deux URLs sur seize (soit 12,5%) ont une connexion directe au site institutionnel *(Illustrations 3 et 4)*.

5 Il s'agit des différents personnels, hors étudiants.

6 L'étude a été menée auprès de l'ensemble des personnels des universités d'Aquitaine. Nous ne retenons ici que la partie d'un établissement.

7 Collecte sur sites Web hors zone DNS avec des profondeurs internes et externes de 0, nous collectons ainsi les pages d'accueil, pages principales d'un site Web, afin de vérifier s'il existe un premier lien entre les différents sites, dès leur point d'entrée, liens que nous pourrions qualifier de structurel ou de rattachement organisationnel.

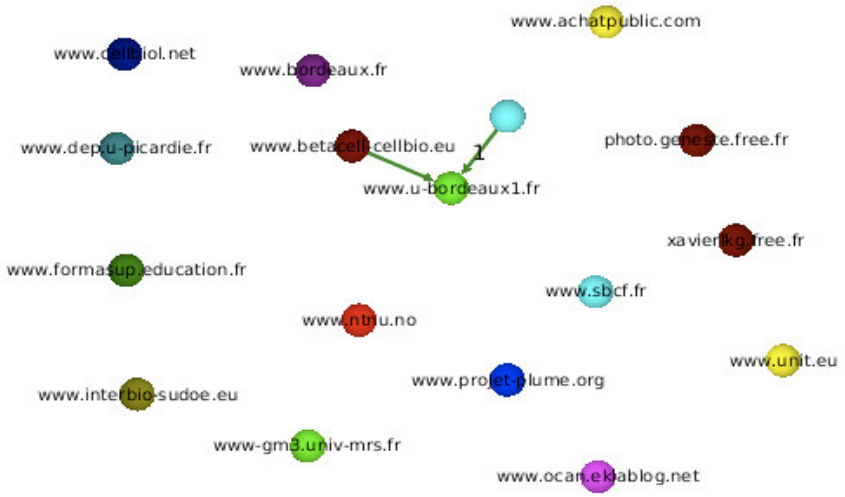

Illustration 3: Graphe représentant les liens entrants du site institutionnel de Bordeaux 1 et les sites "externes"

Afin de préciser notre analyse, nous effectuons une deuxième collecte réunissant cette fois-ci le site institutionnel, tous les sites de sa zone DNS ainsi que les sites externes déclarés. Cette collecte nous permet dans un premier temps de constater que les sites externes apparaissent de façon isolée dans le réseau relationnel site institutionnel-sous-sites Bordeaux 1-sites « externes »: L'*illustration 5*, nous montre clairement un groupe de nœuds (de couleur différente) isolé qui correspond à la zone des sites externes. Dans un deuxième temps, l'*illustration 6,* montre que la zone de sites externes établie toutefois des liens vers la zone éditoriale de Bordeaux 1. Cependant, cette visibilité « sert » essentiellement au site Web institutionnel. Effectivement, comme nous le montre l'*illustration 6*, 80% des liens sortants de la zone des sites externes sont dirigés vers le site Web institutionnel qui représente par sa taille seulement 25% de la zone éditoriale entière mais a une visibilité équivalente à celle des sous sites réunis. Après manipulation du graphe et création d'un nœud représentant tous les sous sites, nous pouvons avoir une meilleure visibilité de la relation existante entre la zone éditoriale de Bordeaux 1 et ces sites externes *(Illustrations 7).* Finalement, cette collecte nous permet d'observer que 3 sites (soit 18,75% des sites externes) citent la zone éditoriale de Bordeaux 1. Tous sont liés aux sous sites de l'université, ces derniers tenant alors une place d'« agrégateur » dans le réseau relationnel et le rayonnement de Bordeaux 1[8].

---

[8] Grâce à ses sous-sites, Bordeaux 1 récupère 6 liens entrants dont 2 sont des liens d'un site externe ne citant pas le site institutionnel.

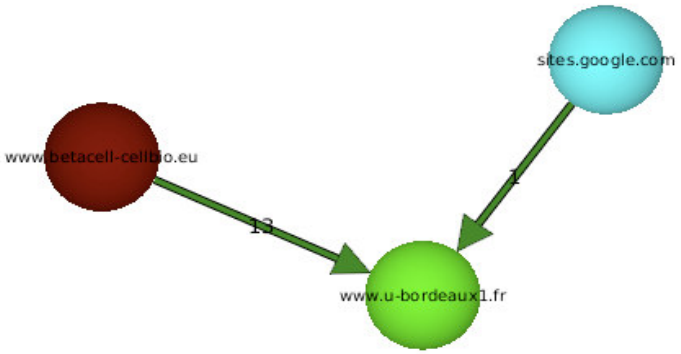

Illustration 4: Graphe valué et dirigé représentant la connexion directe de ces 2 sites au site Web institutionnel de Bordeaux 1

Nous relevons donc plusieurs cas de figure concernant les sites externes : certains sites externes fonctionnent comme des « satellites » de la structure en établissant des liens directs entrants vers la zone éditoriale universitaire. Treize sites sur seize (soit 81,25% des sites externes) apparaissent malgré tout comme étant isolés, et ne participent donc pas au rayonnement de celle-ci sur le Web. Nous poursuivons l'analyse en traitant de la visibilité de ces sites par la collecte des hyperliens provenant depuis le web en général sur ces sites.

Grâce à sa fonction de collecte Yahoo! grâce à l'API ouverte par l'éditeur, *AnCaraS* interroge l'index pour obtenir récursivement les liens entrants (*backlinks*) référencés sur un site, et complète la collecte précédente en déterminant l'inscription plus précise (même si elle est incomplète de par la construction des index, cf. les travaux de Bar-Ilan [4], Boutin et Perrin [6]) des sites « externes » au sein du web. Ces collectes des liens entrants nous permettent d'estimer les liens que l'on peut considérer comme « perdus » par la zone éditoriale universitaire à ne pas être citée par l'ensemble des 13 sites externes qui n'établissent aucun lien avec elle. La collecte Yahoo! nous produit en septembre 2010 un total de 755273 hyperliens pointant depuis le web vers les sites « externe ». Ce nombre estimatif montre cependant la notoriété importante de ces sites hébergeant les productions des membres de l'organisation.

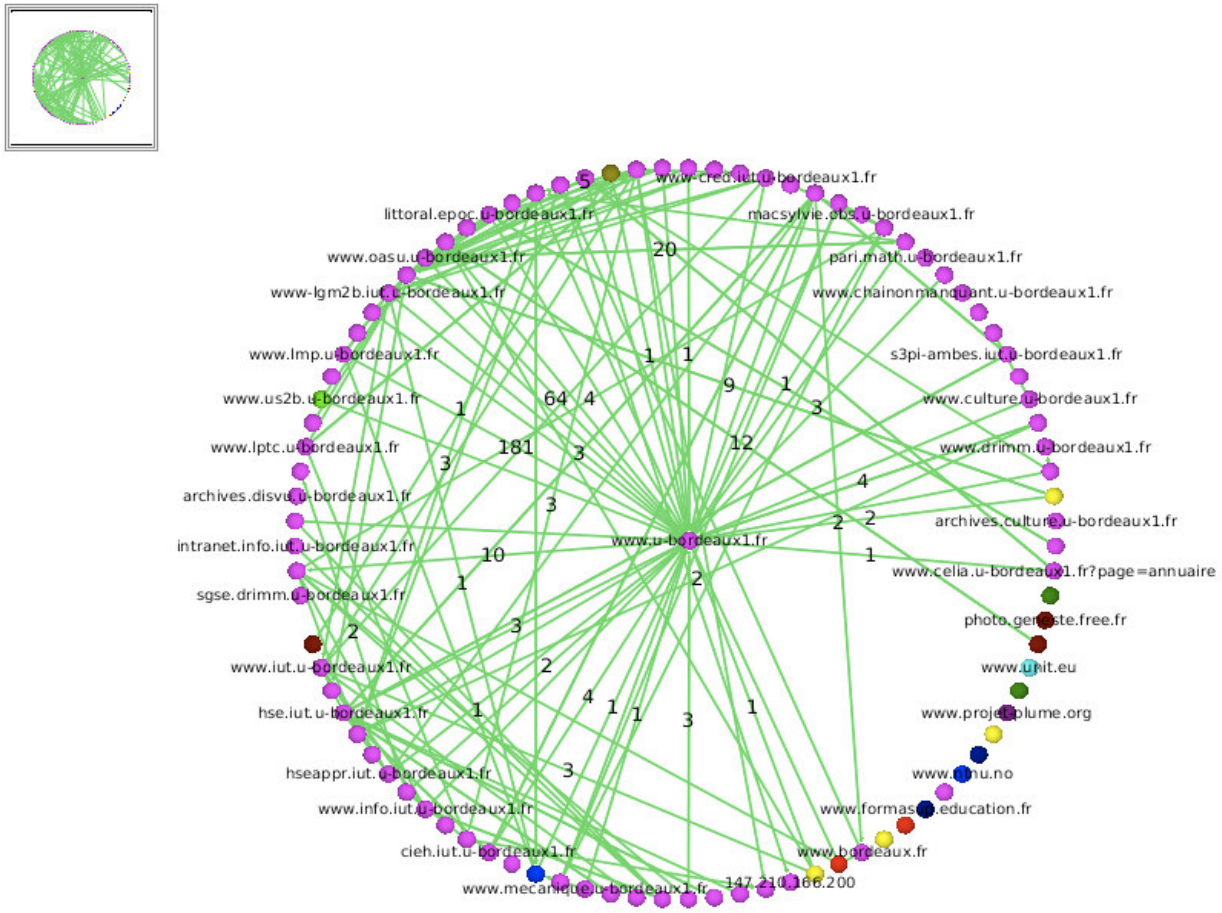

Illustration 5 : Graphe de liens entre le site institutionnel de Bordeaux 1 (au centre), sa zone éditoriale (de la même couleur) et les sites externes (en règle générale, isolés).

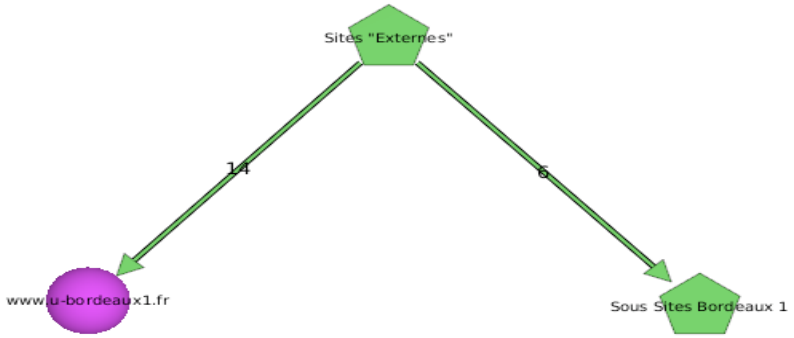

Illustration 6 : Les liens des sites externes vers le site institutionnel de Bordeaux 1 et ses sous sites

En guise de conclusion, l'étude de cas menée nous permet de conclure que les sites « externes » à la zone éditoriale de Bordeaux 1, collaborent à sa visibilité de par les hyperliens générés vers la zone de l'université dans une proportion relativement faible (25%) avec un degré de citation relativement important puisqu'un total de vingt hyperliens entrants (cette multiplicité est en général ignorée). Cependant et au regard des données de départ recueillies (taux de réponse faible des répondants à l'enquête, nombre faible de liens), cette incidence peut être qualifiée comme étant moindre. Effectivement, dans le cas d'un moteur de recherche comme Google, qui fonctionne selon un algorithme relationnel donnant de l'importance aux sources les plus citées (ie. le PageRank), nous pouvons conclure que ces sites externes à la zone éditoriale de l'université de Bordeaux 1 même s'ils n'ont qu'un impact faible puisque représentant qu'une infime partie du Web ne participe que très faiblement à la visibilité de l'université.

Pour conclure au plan stratégique de mise en œuvre d'une politique de communication Web, en regard des indicateurs informétriques utilisés pour établir un score aux différents sites il est d'évidence qu'il va de l'intérêt de l'institution à favoriser l'hébergement pour des raisons évidentes de poids (la taille des productions "externes" interviendrait alors pour le compte de la zone éditoriale) contribuant ainsi à son impact. Nous venons de montrer (avec toute la prudence possible due au faible nombre de productions analysées) que l'activité de production de contenu web par des membres d'une organisation universitaire jouit d'une notoriété Web très importante. Pour autant ces productions ne peuvent être hébergées forcément toutes au sein de la zone éditoriale institutionnelle (des cas d'expertises dans d'autres ministères ou institutions, d'activités particulières dans d'autres domaines ayant un besoin établit de communiquer sur le web) mais la visibilité de la zone éditoriale y gagnerait à inciter les experts à mentionner leur établissement de rattachement sous forme d'hyperlien. La zone éditoriale se verrait ainsi renforcée récursivement par la notoriété des sites pour lesquels ils œuvrent tout en conservant une légitime autonomie d'activités "externes".

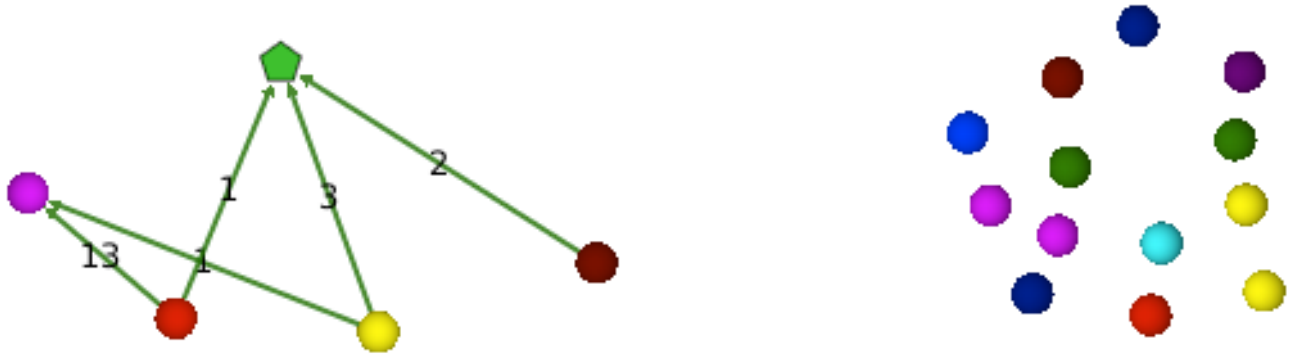

Illustration 7 : Graphe valué et dirigé représentant la relation existante entre les sites "externes" et la zone éditoriale de Bordeaux 1 : 2 sites externes dirigent des liens vers le site institutionnel (à gauche), 3 sites externes vers les sous-sites, les autres sites de la collecte étant isolés.

4 Conclusion

Actuellement en version alpha, le logiciel AnCaraS se constitue comme une base instrumentée de production de connaissance à partir du Web. En intégrant un module paramétrable de collecte récursive des sites, depuis la profondeur à la définition du périmètre il offre à l'usager la possibilité de reproduire les contenus textuels ou la cartographie ou topographie hypertexte d'un site ou d'un groupe de site. En suivant un module de traitement calcule les agrégations utiles à des analyses synthétiques à différentes échelles du web (Zone éditoriale, DNS, Site, Répertoire, ou URL) en produisant notamment des calculs statistiques sur les divers types de contenus publiés propres au Web (HTML, Images, Vidéos, etc.). Enfin, le module de cartographie permet la représentation d'indicateurs type mais aussi la navigation au sein des graphes ou des graphes de graphes produits ce qui ouvre la voie à la matérialisation des connections inter-sites (ou autre unité sus-citée) soulignant les réseaux d'affiliations, les interconnections fortes ou isolements particulier.

Au plan de l'étude cas nous avons pu utiliser Ancaras pour d'une part collecter, traiter puis visualiser des graphes relationnels inter-sites propices à notre questionnement pour estimer dans quelle mesure les sites Web « externes » participent peu à la visibilité de la zone éditoriale des sites Web universitaires. Ceci montre l'intérêt stratégique au plan de la communication Web de favoriser l'hébergement de productions de membres ou tout au moins de solliciter d'établir des hyperliens de revendication d'appartenance à l'institution. Nous avons montré (avec toute la prudence possible du au faible nombre de productions analysées) que l'activité de production de contenu web par des membres d'une organisation universitaire jouit d'une notoriété Web très importante et pâtit de l'absence de renvois d'hyperliens vers les institutions de rattachement des membres « web-publiant ».

5 Bibliographie

[1] AGUILLO, I., "Web, webometrics and the ranking of universities". In *Proceedings of the 3rd European Network of Indicators Designers Conference on STI Indicators for Policymaking and strategic decision*, CNAM, Paris, mars 2010. A paraître.

[2] ARROYO, N. « Métodos y herramientas para la extracción de datos en Cibermetría : el software académico y comercial », Master, Departamento de Biblioteconomía y Documentación - Universidad de Salamanca, 2005.

[3] AUBERT, D., CHIRICOTA, Y., DELEST, M., DOMENGER , J.P., MARY P., MELANÇON, G. (2007). Visualisation de graphes avec Tulip : exploration interactive de grandes masses de données en appui à la fouille de données et à l'extraction de connaissances, *EGC'07: Extraction et Gestion de Connaissances*, Namur, Belgique.

[4] BAR-ILAN, J. (2004). A microscopic link analysis of academic institutions within a country - the case of Israel. *Scientometrics, 59*(3), 391-403

[5] BAR-ILAN, J. (2008). Informetrics at the beginning of the 21st century-A review, Journal of Informetrics, 2(1), january 2008, pages 1-52.

[6] BOUTIN ET PERRIN (2005), Construction du réseau d'interaction entre sites web : test de robustesse de la méthode à partir de plusieurs sources d'information, *Actes du Colloque île Rousse*, Juin 2005

[7] BJÖNEBORN L. (2004), *Small-world link structures across an academic Web space: A library and information science approach*, Thèse de doctorat, University of Copenhagen*, Departement of Information and Library Science*. 2004: Copenhagen, Denmark.

[8] CASTILLO, C. (2004), « Effective Web Crawling », Thèse de doctorat, University of Chile.

[9] CHAKRABARTI S. (2002), *Mining the Web: Discovering Knowledge from Hypertext Data*, Morgan Kaufmann Publishers.

[10] CHU, H. (2005). Taxonomy of inlinked Web entities: What does it imply for Webometric research? *Library & Information Science Research 27*(1), 8-27.

[11] HARRIES, G., WILKINSON, D., PRICE, E., FAIRCLOUGH, R., & THELWALL, M. (2004). Hyperlinks as a data source for science mapping. *Journal of Information Science, 30*(5), 436-447.

[12] LEMIEUX, V., OUIMET, M. (2004), *Analyse structurale des réseaux sociaux*, Les Presses de l'Université de Laval, Quebec.

[13] LIU, B. (2007), *Web data mining*. In *Springer Berlin Heidelberg New York*.

[14] MATHIEU F. (2004), *Graphes du web, Mesures d'importance à la PageRank*, http://papyrus.lirmm.fr/Document.htm&numrec=031984291916600, Thèse de doctorat informatique, Université de Montpellier II.

[15] MAMOUD-CHARNI S. ET REYMOND D. (2009), "Analyse descriptive de l'espace web académique tunisien. Constats et perspectives", 2e séminaire *Veille Stratégique scientifique et technologique (VSST)*, Nancy.

[16] PINAUD B. et KUNTZ P. (2004). Un guide sur la toile pour sélectionner un logiciel de tracé de graphes. In *Actes conférence Veille Stratégique scientifique et technologique (VSST)*, volume 1, pages 339-347.

[17] REBAÏ, B.K., ZACHAREVICZ G., REYMOND, D., CORBÉ, P. (2010) *AnCaraS: a New Webometrics Web-Spider; G-DEVS-Based Validation of Concepts*. Dans *Spring Simulation Multiconference 2010*. Orlando, FL, USA: The Society for Modeling and Simulation International.

[18] TAGUE-SUTCLIFFE, J. (1992), *Introduction to Informetrics.* Information Processing and Management, 1992. **28**(1).

[19] THELWALL, M. (2002), "Methodologies for crawler based Web surveys", *Internet Research: Electronic Networking and Applications* Vol. 12, N° 2.

[20] THELWALL, M. (2009), *Introduction to Webometrics. Quantitative Web Research for the Social Sciences*, Morgan & Claypool Publishers, collection *Synthesis Lectures on Information Concepts Retreival and Services*, 113 p.