R. Rogers
Version of 30 September 2001

IssueAtlas.net: Instructions of Use for the Issue Crawler

Introduction: Issue Network Sampling on the Web

The Issue Crawler software locates 'Issue Networks' on the Web. An Issue Network is a set of inter-linked organizations dealing with the same issue. An Issue Network is located through 'co-link analysis' of issue-oriented web pages, one method used in network analysis, applied here to the Web.

To find an Issue Network, 'starting points' (URLs) are entered into the Issue Crawler software, and each starting point is crawled for its outgoing links. All the outgoing links from the initial URLs are collected, and the common outgoing links, or 'co-links', are kept. These co-links of the starting points are often called the 'population'. The population is then crawled, and the resulting common outgoing links (the 'co-links' of the second iteration of the method) constitute 'the sample'. URLs not dealing with the issue at hand (e.g., software download pages, portal homepages, etc.) are deleted, and the result, potentially, is an 'Issue Network'.

Locating any 'network', and/or an 'Issue Network', is contingent upon the degree of inter-linking between the URLs initially chosen, and found as co-links. If the starting points are unrelated, unlinked, and/or non-linking URLs, the prospect of finding a network lessens considerably.

Consideration of Starting Points

There are different strategies for selecting starting points for the eventual location of an Issue Network on the Web. Careful consideration of starting points often means the difference between finding a specific Issue Network, a multiple Issue Network, or no Issue Network at all. Starting points may be taken from any or more of the following methods, heuristics or devices: sets of known or trusted sources, sets of associatively reasoned sources (hunch URLs), search engine returns or 'related links' from other devices, sources referred to in discussion lists and discussion list archives, sources cited in news stories, web page link lists, etc.

{left side text box}

A Cartographer's Log

It is suggested that the Cartographer create a log file of considerations about starting points, crawler settings, and Issue Network findings, including network properties (type and features) and substantive content. The Cartographer should use a Browser and a Text Editor, or text-processing software, and make and save notes about the entire process of Issue Network location. One of the first considerations to ensure the location of a specific Issue Network is to employ a robust method of choosing starting points, and to enter specific, issue-related link list pages as the starting points for the Issue Crawler. An example of a robust method is to take all the URLs from a particular timespan of a discussion list archive or a particular 'thread', and map the potential Issue Network that the discussion is providing.

{/left side text box}

{right side text box}

Avenues of Analysis: Issue Network Types, Features and Substance
Issue Networks differ according to features, and certain combinations of features may constitute particular Issue Network types. In IssueAtlas.net, the archive of the Issue Crawler results, many of the Issue Networks are comprised of non-governmental organizations (NGOs) and inter-governmental organizations (IGOs). When crawled over time, using the Scheduler, the parties to the Issue Networks may change significantly or less so. One may begin to understand Issue Networks according to these two features: actor composition and stability of actor composition over time. In addition, when reading the actor pages constituting the network, one may note shifts in the terms of discussion. An Issue Network analysis that combines actor composition dynamics and substantive change of debate is one avenue of research suggested by the design of the Issue Crawler.

{/right side text box}

The Harvester

The Issue Crawler has a built-in URL Harvester. The Harvester takes a 'text and URL dump' and strips the dump of text, leaving only unique URLs: http's and/or www's. This allows the user of the Issue Crawler to cut and paste text containing URLs and words into the Harvester, and extract the URLs from the dump. The extracted or 'harvested' URLs are candidate starting points. One may add or delete URLs from this list, before launching the Issue Crawler.

User Privileges and Settings for the Issue Crawler and Issue Network Visualiser

The Issue Crawler has four User Privilege Levels: Map-Maker, Cartographer, Cartographer-Publisher and System Administrator.

The Map-maker has the means to run crawls and visualise findings in an Issue Network map generator, according to the default settings of the Issue Crawler and the Issue Network Visualiser. The default settings on the Issue Crawler are two iterations of co-link method, a reach of two layers deep for each page crawl and co-link analysis method set to 'by pages'. The last default setting means that the co-link analysis is performed on the most specific or 'deep' pages, and the Issue Network is thus comprised of pages as opposed to sites (or hosts). The default setting on the Issue Network Visualiser is a colour-coding showing the governmental, commercial, non-governmental and educational (or scientific) actors in blue, yellow, green and red, respectively, as well as their colour-coded inter-linkings, in two overlapping circles. The software considers country subdomains, e.g., .org.br becomes 'non-governmental' green. Pure country URLS, as .nl, are rendered grey, and may be custom-coloured by the Cartographer.

On the map the inner circle shows the colour-coded Issue Network actors and their inter-linkings, and the outer circle shows the actors that have made and received links with the inner circle of the Issue Network, but have not been linked sufficiently to be party to the Issue Network.

The Cartographer has the means to run crawls and visualise findings in an Issue Network map generator, according to customised settings of the Issue Crawler and the Issue Network

Visualiser. On the Issue Crawler, the Cartographer may change the number of iterations of the co-link method, the depth the crawler reaches per page, and the co-link method of analysis to 'by page' or 'by site'. The Cartographer also may privilege the starting points, meaning that the starting points are kept in the second iteration of the method. (With privileged starting points turned on, the starting points are kept only in the second iteration, and not in any subsequent iteration of method.) On the Issue Network Visualiser, the Cartographer may change the categorisation of the nodes in the network from the default setting of governmental, commercial, non-governmental and educational (or scientific). A change in the node categorisation launches a new colour coding. The Cartographer also may use the Scheduler, and regularly schedule crawls between date ranges of the Cartographer's choosing.

The Cartographer-Publisher has the privileges of the Cartographer, and also may publish results to the Atlas at IssueAtlas.net.

The System Administor has the privileges of the Cartographer-Publisher, and also may add and delete users, edit user privileges, and cancel any crawl. System memory overload occasions crawl cancellation.

{left side box}

'Govcom.org' Issue Network Visualisation Scheme

The circular shape of the Issue Network visualisation scheme is initially inspired by astronomical charts, and later thought of in terms of a roundtable. The roundtable is meant to express a 'neo-pluralist' potential of the Web, once high on the agenda of Internet writers and realisers. More specifically, the roundtable connotes the neo-pluralist potential of web-based Issue Networks and the 'access' of the actors to relevant debate or discussion around the issue. The form and substance of their participation (or their realisation of access) is complicated by the 'known' inter-linkings, or entanglements, between actors, shown on the map. The current visualisation scheme, with the addition of a circle that adjoins the bottom of the roundtable, may be thought of in terms of well-known insider and outsider notions, where links as well as non-links between inside and outside parties come into view. The shape, however, is inspired by the Turkish eye, or the downward-looking Sultan's eye, an expression of initial disapproval of whatever came before him. The Issue Network fills the iris, and the white of the eye is comprised of those parties that have aided to bring the Issue Network into being, but are currently not in the network. The entire visualisation scheme becomes a reflection of parties' vying for Issue Network attention and access, desiring to catch the eye. The question of the realization of the neo-pluralist potential (and its 'approval') remain.

{/left side box}

Considerations of Issue Crawler Settings

Among the settings at the disposal of the Cartographer is to privilege starting points. Privileging starting points allows the Cartographer to retain the starting points in the second iteration of the co-link method, and only the second iteration. One may desire to privilege starting points for three reasons, analytical, normative and technical. In the first two rationales, the Cartographer may be certain of the relevance of the starting points for the generation of the Issue Network. Technically speaking, this setting is recommended for the

location of an Issue Network whereby some of the sites are either down or not accepting crawlers.

The default settings of two iterations of co-link method, and two layers deep for the crawler reach, may be altered by the Cartographer. Should the Cartographer have a population of sites at her disposal, the iterations setting may be put to one. Should the Cartographer have a population of sites already, and the starting point pages are link lists, the iteration may be set to one, and the crawl level to one.

Generally, the Cartographer deepens the reach of the crawler when link pages are not used as starting points, and increases the number of iterations for large quantities of starting points. The Issue Crawler has been specified, technically, to accept up to 300 initial starting points. A minimum of two unique starting points is required.

The Cartographer also may ask the Issue Crawler to return an Issue Network comprised of pages or sites. This setting is included for the purposes of specificity of nodes in the Issue Network. With pages set, the Cartographer seeks the most frequently cited single URL per actor (or host) for inclusion in the network. With sites set, the Cartographer seeks the most frequently cited actor (or host), no matter which of the pages on the host's site is cited. Every page on a host that receives a link is counted as a link for the host. Generally, selecting the 'by page' setting is meant to avoid 'issue drift', or the location of multiple-issue networks. For certain research purposes, the use of such a technique may be desirable, i.e., the location of multiple issue networks, or the location of networks where the reason for their 'networking' is not known or surmised in advance.

Issue Network Authority Thresholds

The Issue Crawler returns the sites or pages receiving two or more links from the parties included in the last iteration of the method. Issue Network node quantities may vary considerably, depending on both the number and type of starting points.

Upon completion of an Issue Crawl, the cartographer as well as the visualiser may move the authority threshold upwards, and, later, downwards. Increasing the authority threshold means that the Cartographer is requiring more inlinks for candidate network nodes. The Cartographer is thereby changing the 'threshold' for nodes to be included in the Issue Network. The Cartographer may visualise and view the Issue Network map after choosing an authority threshold.

In order to read on screen and print (on A4 paper) a visualised Issue Network Map, no more than 35 actors in the Issue Network is recommended.

One need not set the actor quantity to a maximum of 35, however. The Issue Network Map generator and visualiser are sets of PERL scripts, producing scalable vector graphics (SVGs). With SVGs in use, and the SVG plug-in or functionality built into the browser, one may retain any authority level, and view the Issue Network by traversing the page, or by zooming in and out. Larger format print-outs (A3, A2, A1, A0) are required for Issue Network maps with node quantities exceeding the baseline figure of 35. Print experiences may differ.

The Scheduler and the Scheduler Settings

The Cartographer may schedule the Issue Crawler to chart the Issue Network at regular intervals. The purpose of the Scheduler is to chart changes in the actor composition, and eventually substance, of the Issue Network over time. Generally, one also may take notice of the rise and fall of a network, though it is expected that given an acceptable quantity of viable actors, the Issue Crawler may continue to capture core network movements.

There are two types of scheduled Crawls available. One may schedule the Issue Crawler to use either the original starting point URLs or the URLs from the last iteration of the method. In the former case, the Crawler is 'refreshing' the network on the Cartographer's original terms, and in the latter case the Crawler is refreshing the network more on its own terms.

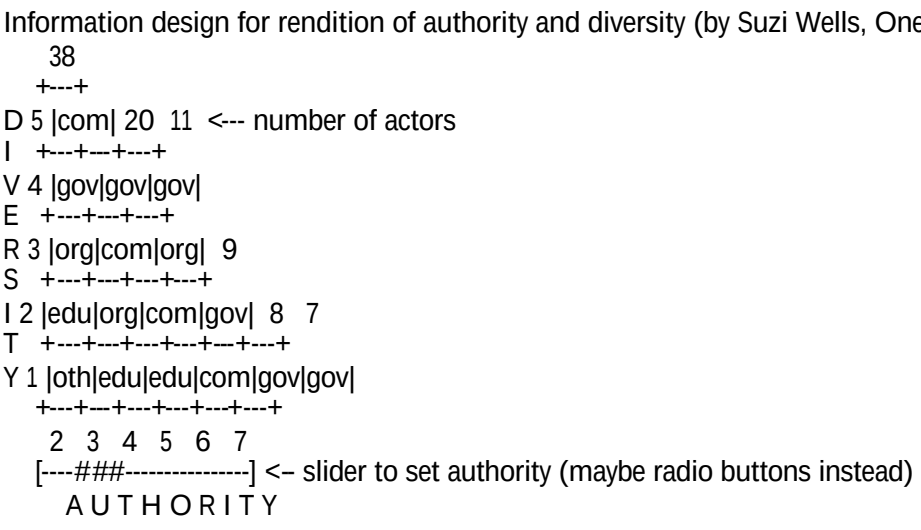
Network Manager

To Map-makers and Cartographers, the heart of the Issue Crawler is the Network Manager. The Network Manager shows the Map-maker's and Cartographer's queued and completed crawls, and allows for the further tuning of the network, including setting the authority threshold, and eventually viewing the Issue Network map.

The Network Manager includes a facility for downloading and uploading an XML file. The Issue Crawler produces its results in the form of an XML file, which saves the information the PERL scripts require to visualise the Issue Network, including the inter-linkings between nodes inside and just outside the network, the categorisation of nodes and the colour coding and visualisation scheme. The downloading feature is included to allow the Issue Crawler results to be read 'in the raw', to be analysed with additional means, or to be rendered with different means and scripts. Similarly, the uploading feature is available for rendering XML data, gathered by other means, and formatted in a way that the visualiser will understand.

Currently, there is some information in the XML file that is not utilised by the analytical means built into the Network Tuner (where only authority is raised or lowered) or by the visualiser. Date stamps (latest page modification dates) have been made of each of the pages per Issue Crawl. The purpose of keeping the date stamp of the page returned in each crawl has been to create a 'live issue atlas on the web', the original name of this project. To 'enliven' the atlas, a network freshness metric is written that measures the 'freshness' or, as it's been called in the project, the 'heat' of the issue network. One then has the crawler automatically refresh the network at intervals that account for its 'heat'. The hotter an issue network, the more frequently it would refresh itself.

In the XML file, there is also a 'diversity' field, awaiting the design of a setting on the Network Tuner. Issue Network diversity is a measure of the quantity of node categories resident in the Issue Crawl results. An issue network constituted only by .org's has a diversity of one, by .com's and .org's a diversity of two, and so forth. The normative idea here is that one may wish to raise the authority threshold to a higher and higher level, whilst still retaining a particular level of diversity, or, to the project, a particular level of 'pluralism'.



The Archive - Issue Atlas

The Archive is the centerpiece of IssueAtlas.net, and is considered an Atlas of Issue Network maps, dealing largely with global or even globalization issues. Unlike the Issue Crawler, the site is publicly accessible at http://www.issueatlas.net.

The Archive is a database of saved Issue Crawl results by map-makers and cartographers alike. One queries either 'published' maps, or all saved crawls, including published maps. Queries may be made for issue crawl results in a date range. Queries also may be made for one or more URLs, one or more organization names or one or more key words that have been assigned to crawl results or published maps by the Map-maker or by the Cartographer. Choosing 'show all networks' will produce a list of all the results in the database.

One of the main purposes of the Archive for comparative research is to enquire into the presence of one or more actors across the Issue Networks.