Generating Business Intelligence Part 5 – OSINT

Main Title:	Generating Business Intelligence Part 5 - OSINT		
Date:		**Release Maturity:**	Draft/Live/Final
Author:	Allen Woods		
Owner:	Allen Woods		
Client:			
Version:	0.50		
CADMID	Demonstration		
Line of Development	Information		
Organisation	The Old Geeks Shed.		
Release Classification	unclassified		
Document ID/Number	BI/5		

Note: This document is only valid on the day it was printed

Revision History

Date of next revision:

Revision Date	Previous Revision Date	Summary of Changes	Changes Marked

Approvals

This document requires the following approvals. A signed copy should be placed in the project files.

Name	Signature	Title	Date of Issue	Version

Distribution

This document has been distributed to:

Name	Title	Date of Issue	Version

References

Internal

Ser	Description	Location
1	Site Review Client Side Template	Left click here
2	Why Google Analytics Has Been Banned	Left click here
3	Template case study	Left click here
4	Architectural Schematics	Left click here
5	A Client Side Impact Site review Template	Left click here
6	A Solarwinds Review	Left click here
7	Choosing a developer guidance	Left click here
8		
9		
10		

External

Ser	Description	Location
1	OSINT Wikipedia Article	Left click here
2	Companies House Data Products page	Left click here
3	The GLIEF Downloads	Left click here
4	UK Office for National Statistics	Left click here
5	Standard industrial classification of economic activities (SIC)	Left click here
6	A list of OSINT Tools	Left click here
7	Modernizr.js	Left click here
8	Zero Trust – Wikipedia Article	Left click here
9	Capability Maturity US DoD View	Left click here
10	The EU Eprivacy Directive	Left click here
11	VAT Number Validation	Left click here
12	HMRC VAT API	Left click here
13	UK NCSC Guidance	Left click here
14	US NIST	Left click here
15	OWASP	Left click here
16	HTTP Header Fields	Left click here
17	Device Fingerprinting	Left click here
18	Google Developers SEO Optimisation Guidance	Left click here
19	Search Engine Optimisation Guidance Wikipedia	Left click here
20	ROBOT's Exclusion Standard	Left click here
21	Modernizr.js	Left click here
22	Page Scraping	Left click here

Location Details

Original Path on Development
C:\Users\allen\Documents\ea co\Generating Business Intelligence\OSINT
TLMP Folder Location
Portal URL
http://localhost/geophysical/companyregistersearch.asp

Contents

Table of Figures

Statement of Copyright

The reader can use this document as he or she sees fit. Comment is invited. Contact details are set out below

Fact of the matter is, that the author is retired, not a millionaire and could not protect any copyright in any event….. Help yourselves. The aim of this document is to prompt some thought in respect of just can be done, on a properly founded architectural basis, to join data up and what subsequently falls out of it in respect of the construction and operation of an end to end properly integrated information management architecture. The tools and components in this series have all been built by the author, they are not for sale, the aim is to demonstrate commercial risk. The way operating principles and subsequent implementation are the authors own work.

Allen Woods

Tel: +44 (0)7780 568449

Email: woodsa200@gmail.com

Skype ID: apw808

About the Author

In order to decide for yourselves the experience of the author in respect of whether or not to take this document seriously, this brief biography is provided

- Allen Woods, now retired.
- Ex British Army (1971 – 1995) Taught Arctic Warfare, several years on operations, Funded Himself through College to Study IT
- Chartered Member of the British Computer Society for 20 years
- Member of the Chartered Status Interview Panel for BCS
- In 2010, Finalist of UK "Developer Of The Year" Competition for the MoD Health and Safety Information System (HSIS)
- Primarily Employed in UK Defence Supply Chain and Logistics IT since 1995 until 2019
- Credits: MoD Health and Safety Information System, Various Internal to Defence P&G Portals, CATMIS, IQB Oversight to Defence Voyager Programme IM Transformation
- In respect of contract examination, as part of a major due diligence exercise the author was part of a team, the aim of which was to examine the licence terms, as a matter of contract validation, of 20 major systems, each with an annual maintenance fee in the high 6 low seven figure expenditure range.

It is a wide and varied range of IT work experience covering some 30 years in total. In many ways the author was lucky in that he started working in IT when the PC's were beginning to proliferate and along the way was given the chance to work on a wide variety of tasks that simply would not be possible now. A common theme through them all was legal compliance in one form or another.

His by line is "How the hell did that happen" which is appropriate not least because along the way he has made more than his fair share of mistakes and one of the sub plots of the pack that this document forms part, is that it seems to the author that much is having to be relearned. Above all, this document and the pack is an attempt to help. Particularly non-technical people understand some of the technical complexity they are using almost without thought it seems.

None of this is simple. Even writing this document was a complex task.

Notes On Linked Documents

Internal links on page 2 and elsewhere are unlikely to work if this document is held in isolation. All such documents are available from the author using the contact details on page 4. Just ask for them if needed.

Executive Summary Read Me First

This series of documents is in five parts. The overarching aims are:

- To provide the means to illustrate through emulation, the business advantage to those who operate online marketplaces that an architecturally sound and integrated information architecture brings to them.
- To serve as a warning that entering an online marketplace is not a casual decision to take, particularly for small to medium sized enterprises as the way organisations that own such market places can gather data on them to a considerable degree of detail

The previous four episodes, or chapters, building from initial operating concept through to the development of a forward and reverse supply chain business intelligence capability (in such a way that those who own online markets have a “through life” means to identify and exploit patterns of behaviour of both suppliers and customers), with is document hopefully, shining a light on the reach of the previous documents into the world of “Open Source Intelligence” (OSINT). The aim of this document being to draw the reader’s attention to the art of the possible OSINT brings to extend the business intelligence capability referred to way beyond data held internally.

This document has a key underlying message, in that the focus on “me” in respect of data protection and the more nuanced privacy protection is, for many organisations, is a distraction, albeit a well-intentioned one and that the nature of the commercial risk entering such marketplaces is a significant matter to the point of being existential and while often overlooked, should command more attention than it does.

There are several sources of OSINT introduced into the market place information management infrastructure on which this series is based, firstly there is a copy of the downloadable YK Company Register, followed closely by geophysical data like a UK Post Code list (with latitude and longitude values) and things like definitions od socio economic breakdowns, by geographical area, from the UK national Census archive. However, recently the author published a series of briefing papers on the recent ban of Google Analytics by four EU Data Protection Authorities (Austria, Denmark, France and Italy) which prompted the development of a client-side impact review template and a series of case studies. In the template is a review toolkit (see Annex A of this document) and the author took a decision to see if he could automate, to some degree, the delivery of some of end products of some toolkit components. The result of that work forms the core of this document. And with the result, the author can now carry out a summary compliance review of any web site of any company on the UK Company Register in just under 15 seconds.

However, subsequently, the author took a longer look around various web document repositories he was aware of looking for other “data” to exploit to see what the art of the possible might be. He was staggered at what was available. The result? To be in a position to write his own equivalent tools, but to incorporate other data sets, mainly lists of “interesting IP addresses”, but also making use of some standard legitimate site information files, the net result of which will be to extend the information “network” on any company whose site is reviewed, to include key business relationships, the use of components that may fall foul of data protection law and more besides.

The sheer volume of data, easily available (all that is needed is a reason to look) is simply staggering in the author’s view. What that means is that nowadays, it is considerably easier, than even a few years ago, to build a profile of any business, that is extremely sophisticated, in a very short time-period compared to, say, 10 years ago. As to whether or not companies are building the kind of company profiles the author now knows to be feasible, pages 29 and 30 provide links into the “developer support platforms” nearly all of the major vendors now provide to people like web site builders which explain how to manipulate things like visitor data to the advantage of those who employ them. The trick of course, is for those who employ them to be aware of what the art of the possible is AND of the responsibilities and liabilities the use of such data and appropriate techniques brings into play. And it is that failure to understand

responsibilities, liabilities, and risks that is the greatest danger to many companies, particularly small to medium organisations as they, inevitably, unwittingly sign up to tools and techniques developers recommend, or worse, just use, without having the nature of the risk being explained to them. The net result, the small to mediums act as data gatherers for the very people who make their money from business intelligence gathering.

In conclusion, the authors view is that while the current crop of data protection law is well intentioned, it is not fit for purpose. The sheer scale of data exploitation going on renders it difficult, if not impossible to police effectively. If there was a recommendation, given that much of what is happening in respect of data gathering, it would be that developers of web sites, should be licenced to practice and be obliged to demonstrate they have professional liability insurance at a level to cover all of their portfolio given that it is site operators who are held accountable for data breaches and much more.

The current level of computer related crime is not in the $Tn world-wide, it is the developers who write the software that is not fit for purpose that should be held more accountable for their failure. The concept of “fail fast and fix things” followed closely by “code before documentation”, both being adopted with near casual abandon by the development community, being little more than a licence for mayhem online. Developers it seems, are that rare thing, a bunch of people who are not held properly accountable for their work.

Heads up.

Purpose of this Document

This document forms part five of a series the purpose of which is to describe, in overview, the nature of the effort required to build a business intelligence platform capable of being used to exploit privately and publicly held data for business analysis purposes capable of being used for evidence based decision support with the evidence being both validated and verifiable. This document focusses on the use of "Open Source" business intelligence of a kind that is in the public domain and free to download which is then integrated with internally held data, but on an architecturally sound basis.

This document builds on Part 4, "The Internal Market and Forward and Reverse Supply" and introduces the concept of integration of internal market data with freely available Open Source Intelligence of many kinds

Scope

A list of the full document set can be found at Annex A.

Open Source Intelligence takes many forms, in the author's view it is surprising just what is publicly available if those of a curious mind with a problem to solve care to look. In respect of this document in particular, the main sources of OSINT data used in the screen shots and accompanying descriptions can be found in the external references 1 – 5 on page two. This document also builds on series of exercises (see the internal reference list) setting out an approach to reviewing the client-side impact of code delivery from CDN platforms into end user devices the purposes of such code drops being related to the "finger printing" of device browsers with a view to providing profiles of site visitors. This document is also founded on a series of web site review exercises the aim of which was to illustrate the nature of the data gathering effort going on across the world today. Writing each web site review took some time to investigate and produce, a sub plot of the integration exercises described in this document being to improve the speed of execution of the kind of research each review required.

In the respect of capability, the tools and screens displayed here can retrieve and display a comprehensive web site profile for any of the entries on the UK Company Register that has an active web presence that has been identified as "of interest" by the author. It can do that in fifteen seconds and now includes a detailed assessment of client-side impact, when executed, a site review also incorporates financial summaries and an assessment of the nature of business relationships associated with online advertising.

Annex A contains a list of applications that readers can use to execute their own similar reviews if there is a will. In due course, the tools in the Annex indicated in the body of this text, will be replaced by the owners own code, not that there is anything wrong with anyone, somly because the author knows how to.

Caveats

What is described in this document is an ongoing exercise, its end game being to be able to provide evidence to assist in understanding the nature of market penetration on the basis of the grouping of active companies by "Standard industrial classification of economic activities (SIC)", geographical location, socio economic grouping and some of the formal reporting requirements of the UK Companies Act 2006. All of which may change in respect of function and impact over time,

A Warning

In the author's view, "on line markets" are not, in any meaningful way, markets in the traditional sense of the word rather they are operating environments owned, in every sense of the word, by those who provide the means for others to make use of them to sell their goods and services on-line. While the provision of a marketplace can generate significant revenue from things like sales commission, it is the nature of the business intelligence having complete control of forward and reverse supply brings the real business advantage in a massively connected world.

Entering such marketplaces without more than a passing understanding of the nature of information gathering and associated risk is little more than foolhardy.

The previous documents in this series have been about the nature of the business advantage such markets generate. This document focusses on the additional of business intelligence that can be gathered, that augments what may be held internally, if those with the skill and means look out into the wider web to take advantage of what is known as "Open Source Intelligence". Bluntly, there is so much free data that is available nowadays that its use, when combined with internal data, widens an organisations world view and its ability to make informed decisions considerably. Those who connect Open Source intelligence with internal data on an architecturally sound basis will inevitably, have a considerable commercial advantage over their competition. The diversity of scope of OSINT, is, in the author's view, staggering.

TL:DR this document if the reader wishes. Your risk. Heads up!

Background – The Web Leaks Like a Sieve!

Readers will be aware (hopefully) that the state of security in the field of information technology is of major concern across the world, one of the most significant incursions over the last two years known as "solarwinds" has been estimated to have cost the US Federal Government some $100bn dollars (and counting) to put right. By any standard, the "Solarwinds" incursion was a remarkable technical achievement given its comprehensive nature. In the author's view "solarwinds" represents a systematic and systemic failure on the part of all concerned and out of it came the concept of "zero-trust" which in turn brought into play the idea that developing the kind of security and information management awareness of the kind needed does not happen overnight, nor is there any single product that can "do it all", as a consequence any "zero-trust" exercise will inevitably be based on the idea of capability maturity, which will, of necessity, involve the co-ordination of multiple lines of development both cultural and technical. **In the author's view, one of the primary concepts to understand is the idea of the "organisation boundary" over which nothing should pass over, in or out, without the express approval of the organisation itself,**

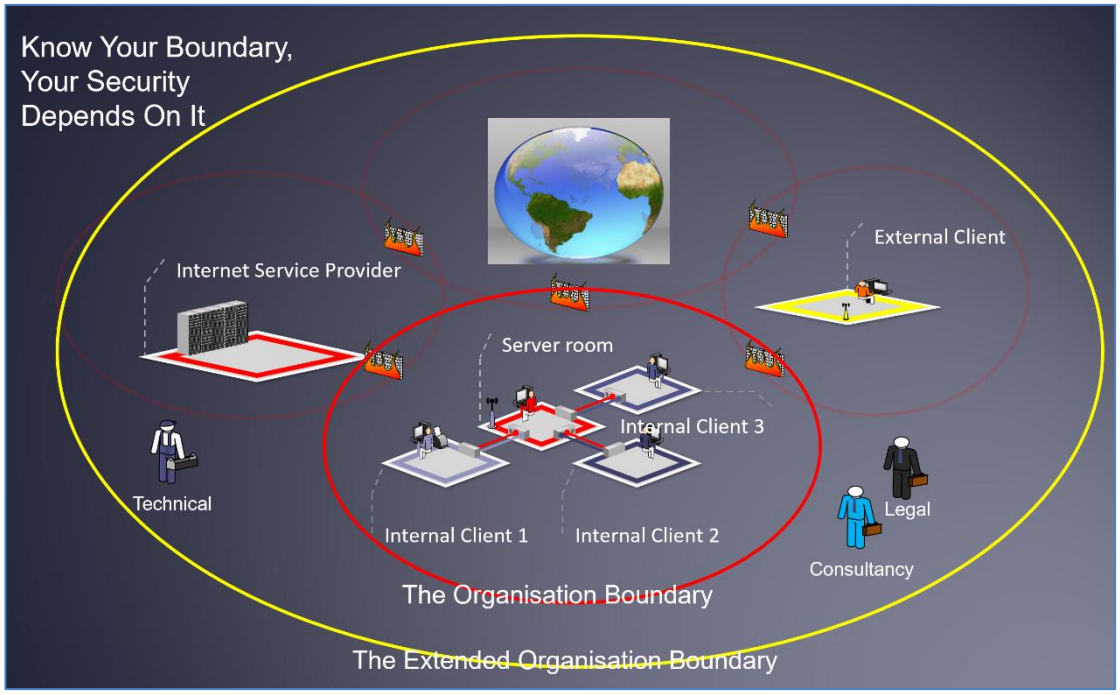

Figure 1 - The Organisation Boundary

The concept of the organisation boundary as set out above is also consistent with the idea of an individual's "private sphere" as set out in the EU E Privacy Directive on its page 3 Para 24. Arguably, the true nature of any privacy risk to the individual lies in the impact, client side, of

any exchange of data between the individual and the organisation which is made more complicated by its nature as a maturing series of electronic conversations.

Figure 2 is an attempt to illustrate the complexity of uncontrolled data exchange across the organisation boundary with the "cloud" indicated on the right-hand side of the schematic. Increasingly, in a massively interconnected world that simply did not exist even a few years ago, the boundaries are becoming fuzzy and nebulous with an increasing risk to the viability of the organisation as a living thing given that inevitably, the nature of what is (as opposed to what it is thought to be) the organisation boundary expands. A key concept to understand being that in effect, a web site is a window for the world to investigate the organisation if it finds it interesting enough. If there are also doors to go through, then the organisation is only as secure as its weakest or least protected doorway.

The remainder of this document is about demonstrating the nature of the business risk associated with the doorway known as a "web site". And, when data and information that can be drawn from even the most secure and well-constructed web site is linked to other "stuff" that is freely available just what can be done with it all.

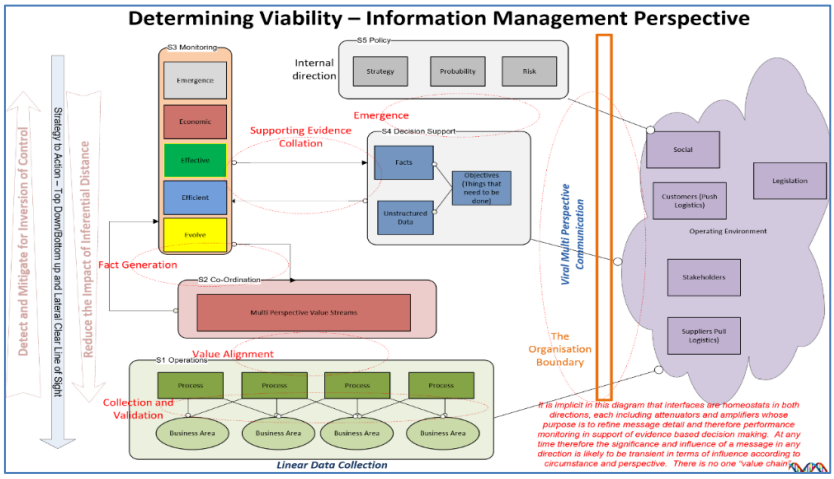

Figure 2 - Data to Information Transition and Attenuation

It is a recurring theme of this series of documents that it is a growing understanding of the relationships between things and how to exploit them is where the power of data to information transition lies in respect of the provision of evidence based, contextually sound, validated, and verifiable decision support lies. Once grasped as a concept, the commercial advantage it brings is enormous and one of if not the reason the GAFAM's are the kinds of companies they are.

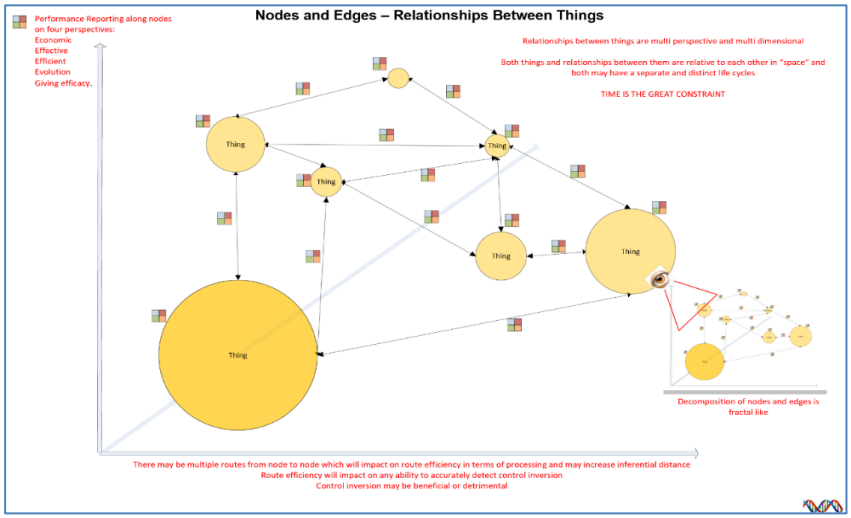

Figure 3 - Understanding the relationships between things is where the power of information lies

Those who unwittingly fall for the twin seductions of “free” and “simple” are in fact giving away hard business intelligence about themselves and their stakeholders (or all kinds) that the likes of the majors, referred to often as the “GAFAMS” can exploit. The kind of analytical capability that is described in this document, is the same kind of thing they have at their disposal (but given their data processing resources way more powerful than that described here) with the majority of their data gathering done for them, for free, by those who fall for the seductions referred to.

The Nature Of The Beast

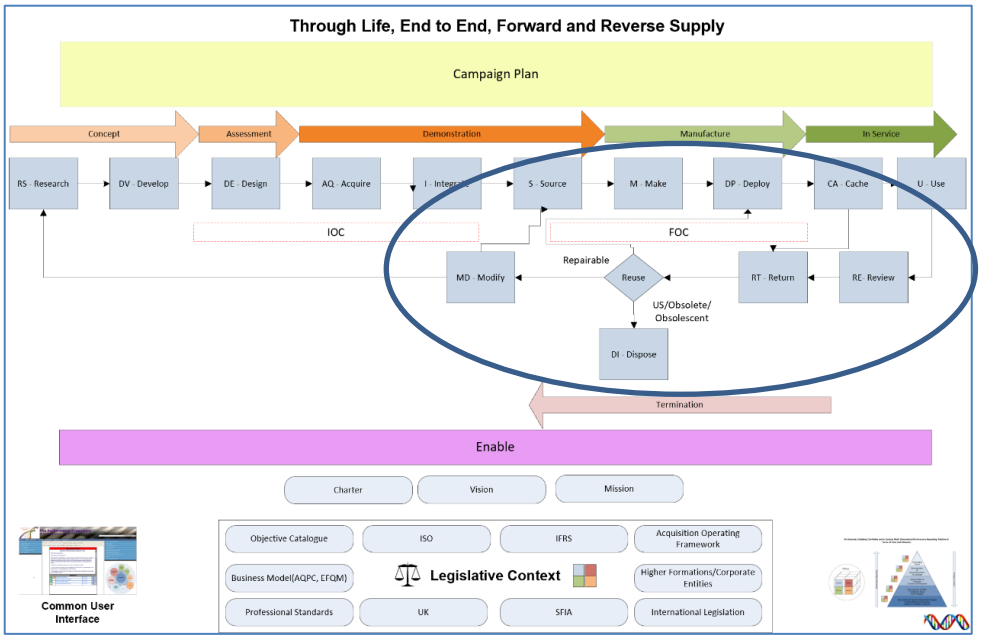

Figure 4 - Reverse Supply - They know their suppliers too.

The previous documents in this series are about the need to understand operating principle, organisation form function and purpose and the nature of the management of internal forward supply. All of those subject areas, are about the maintenance of organisation effectiveness (meeting customer needs) and internal viability concern the organisation as a single entity. While an individual organisation may enter an online marketplace with the aim to increase sales and profits, the value of the business intelligence to the market owner should never be underestimated. The market owner, beside understanding the nature of its customer base, also has detailed knowledge of its supplier base too.

Do that and through life, coherent data management and information generation is more than feasible. Bear in mind the following pages are live, working software.

The Key Compliance Risk and Liability – Device Fingerprinting

In the authors view, perhaps they key business risk on the internet at the time of writing, is the concept of browser or device fingerprinting. Its essence is being able to profile a site visitor, usually covertly, by taking a snap shot of the working state of an end user device and or the end user browser. The scale of fingerprinting effort is enormous. To give an idea of the scale, just one well known and well used software component known as “Google Analytics” (GA) is reported to have some 320,000,000 installations on web sites across the world. Given that it, like many other third party components is usually installed on multiple site pages, that suggests that the nature of the scope of data gathering, just for GA is in the millions of site visits, per hour across millions of pages. A review of the nature of GA and how it works can be found and read here.

Given that the vast majority of account holders are likely to be small to medium sized enterprises, then it is the small to mediums that are supplying Google with hard, validated and verified evidence of business activity that can then be sold on, not as data, but rather as a series of targeted services under the banner "Real Time" or "Header" bidding. In short, an auction that inevitably those with the deepest pockets will win.

The Core Business Advantage? Joining Internal Data and OSINT Up.

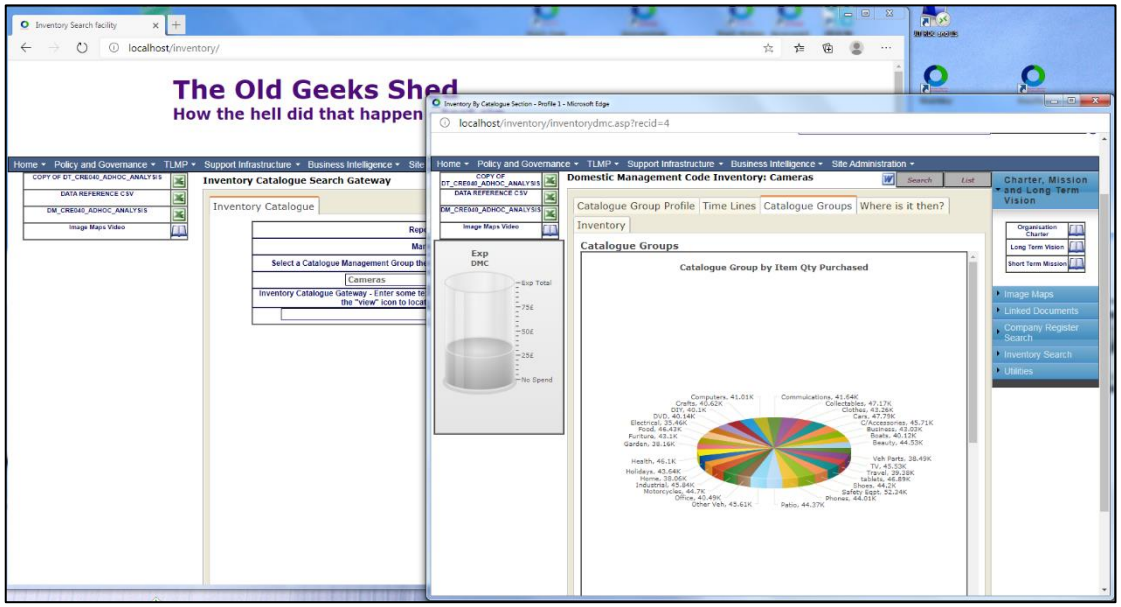

Figure 5 - Forward and Reverse Supply

Part four of this series sets out the nature of an online market place, that covers forward (customer facing) and reverse (supply side) activities on an end to end basis. Figure 5 is a sample screen shot from a model, built by the author, the purpose of which is to demonstrate the business advantage of being a position to own such a marketplace. The advantage combining OSINT and internal data of the comprehensiveness an owned market place gives, is that the owner and they alone, are able to identify business opportunities that would otherwise remain hidden and exploit them. While those who provide the data, both customers and suppliers, will only see the parts of internal data that refers to them. The author's view is that combination of data sets is a significant risk to both suppliers and purchasers and is arguably the reason why companies known collectively as the "GAFAM's" are the successful businesses they are.

It may be of interest to the reader to know that the chart illustrated in figure 5 is "clickable", each pie segment being a gateway into the wider information environment providing navigation support for top down/bottom up as well as linear navigation left to right of arc. Navigation providing the means to "drill down" in multiples of views from high level overview to the performance of individual inventory items, suppliers, purchasers and more. Itself a powerful business intelligence capability. The following pages illustrate how that internal analysis capability can be extended to include OSINT data of various kinds to develop a more complete picture of market activity.

External Review Tools

External OSINT – Introduction

This next section describes, in overview, the interface built to demonstrate OSINT data and its utilities, between a series of open source data sets and how they might be connected up to free to use software to profile an organisation. Readers should note, that given that the toolkit provides here can also be associated with an organisation management infrastructure that also includes things like a customer management system, an inventory management platform, both covering the concept of "forward and reverse" supply management, then what is described here is the means to enhance any supplier review capability that goes beyond items supplied by them into the market place and extends that to cover the viability of supplier businesses.

While the internal market place described elsewhere in this series, is a model, make believe if the reader wishes it, nevertheless, the nature of the integration of selected OSINT data into such a market place represents a significant enhancement to the maintenance of market dominance of a kind that the major IT companies have achieved.

Annex A contains a toolkit of much wider scope than what is listed here. During any company review exercise, if necessary and prompted by the tools listed in this document, it is recommended a more extensive review is undertaken using the tools listed in Annex A.

The Search Engine Front End

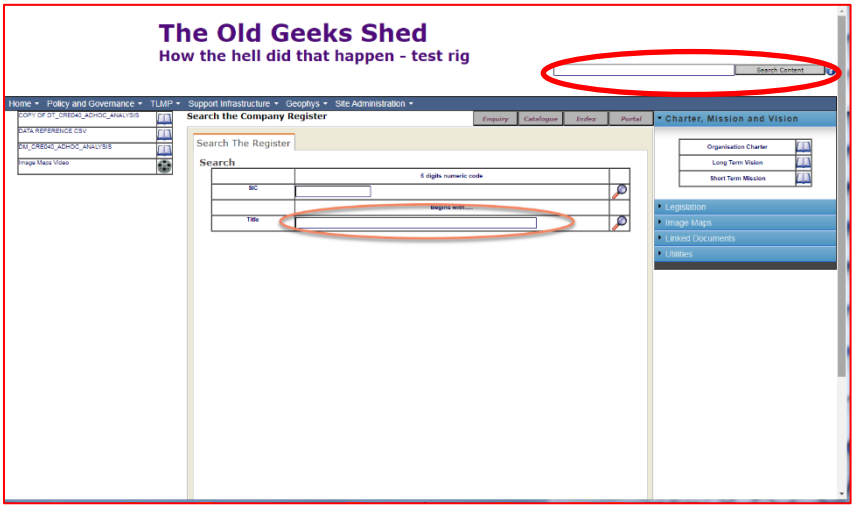

Figure 6 - Search Engine Gateway

The "front end page" is illustrated in figure 5. For simplicity of operation, the design approach is to keep entry points into the site analysis capability as simple as possible for the user giving an easy to use gateway based on few straightforward ways to get to a specific company of group of companies. The screen shot presented here has two search criteria, they are:

1. By SIC code, which provides the means to locate all companies in the same market or business sector.
2. By company name or part thereof.

However, on the live version, given that the copy of the company register held on the platform is in an indexed relational database, to which has been added additional profile criteria like latitude and longitude of approximate location and a list of all UK towns and cities associated with the UK post code structure, then the number of search entry points can be increased to include criteria like Post Code, Town or City, VAT number and more besides.

Note too that in the image above and on every page in the supporting site, there is a search capability that can also tap into a legislation librarian, internal documentation, performance monitoring support and more besides.

Search Results By Search Phrase

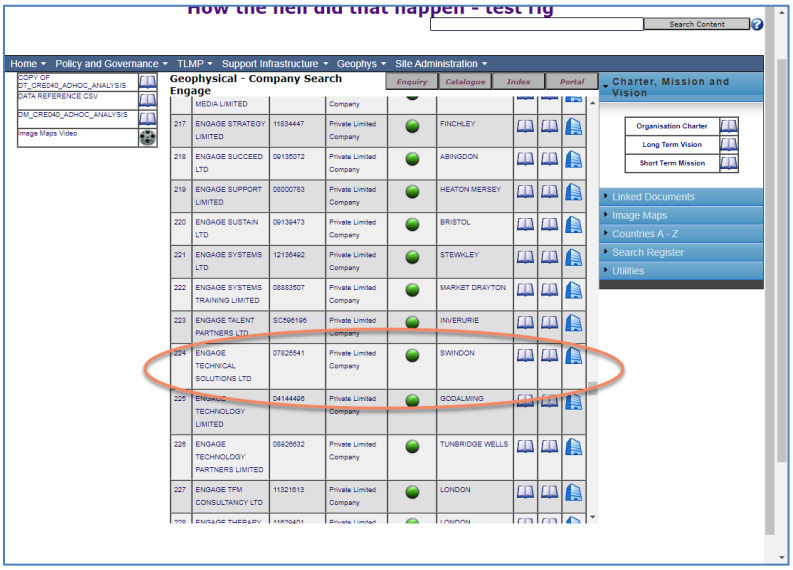

Figure 7 - Search result all companies with search phrase in title

The results of entering a search phrase look like the list presented in figure 6. The company selected (entirely at random) for the remainder of the exercise described here. There are three immediate search options which are:

1. To draw down the company certificate as granted by Companies House
2. To draw down the company summary reports as filed at Companies House
3. To connect into the on-site company profile.

Company Register Record

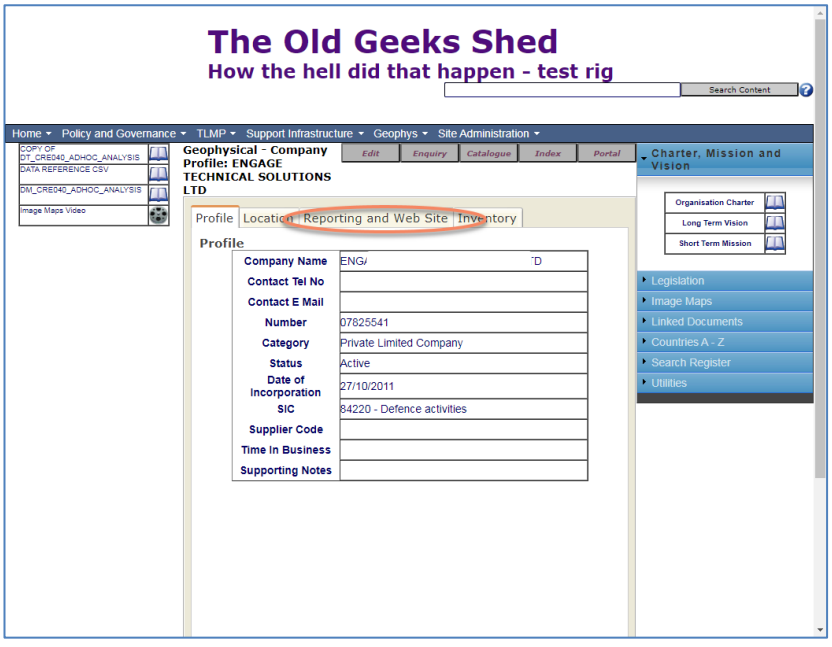

Figure 8 - Company Profile

Once the company chosen for review has been selected, then the companies house register entry for it is displayed. There are five tabs which provide access to:

1. A summary corporate profile

2. A geographical profile which can be connected to an internal and external mapping capability (the external one is not Google Maps).
3. A profiling tool kit which is a subset of a client-side impact site review template
4. A web site code review history.
5. An inventory list if the company under review has been identified as a supplier of goods and services into the site market place

The readers attention is drawn to the client side impact site review template as it provides more detail on the rationale behind the development of this part of the information management portal that this series of documents describes.

The Profiling Toolkit

The company profiling toolkit consists of a series of “point and click” components built into a company record. The interface is pictured below

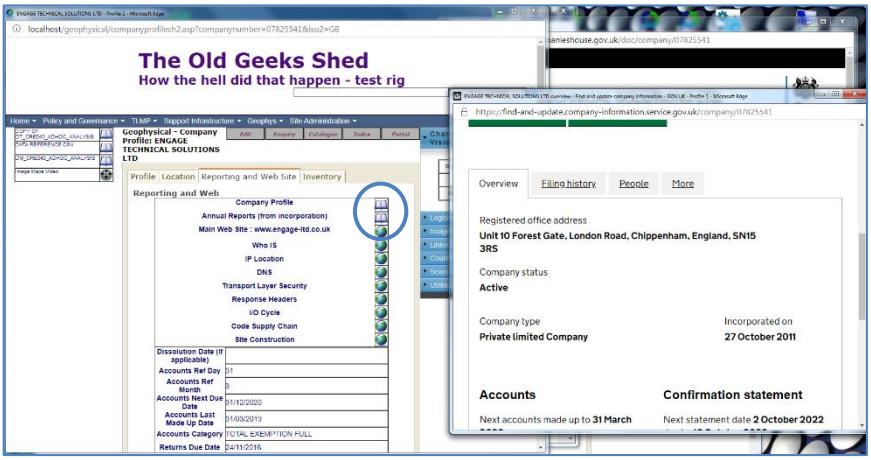

Figure 9 - Sample pProfiling toolkit interface

The “point and click” provides the means to view standard companies house reports including a company profile and give access to mandatory annual financial reports. The remainder of the options in the toolkit provide the means to pass search criteria from a company profile (typically the company domain or the domain IP address) into free to use on line review tools of various kinds.

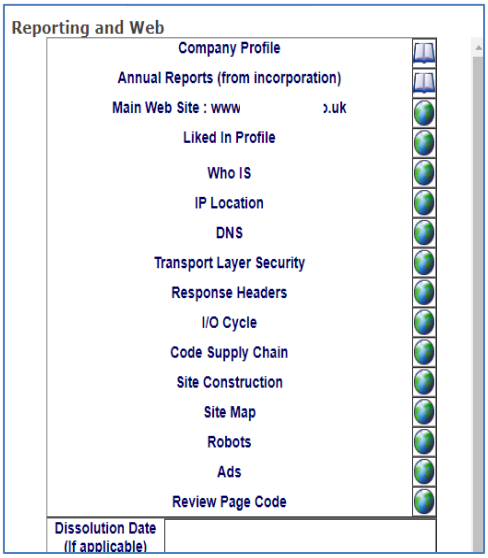

Figure 10 - Sample review toolkit

In addition to providing the means to connect to and use external tools, there are options to view key site configuration files often held in a web site root directory. The remainder of the

images in this document will describe the nature of the information delivery for each button in the toolkit. It should be noted that the options listed for part of a wider toolkit available in the client-side impact site review template. It should also be born in mind that the long term aim of the "Review Page Code" capability listed will be to remove reliance on third party software. In "top down" button order the purpose of each is:

Company Profile and Financial Reports

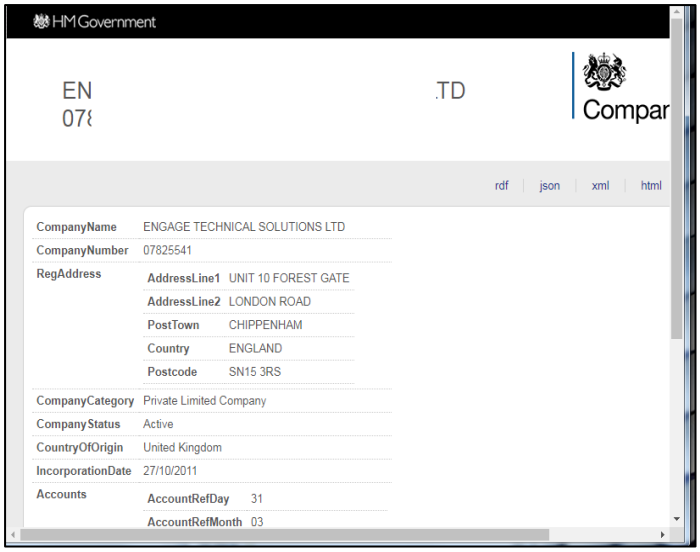

Figure 11 - Company Profile and Financial submissions

As mentioned elsewhere, the first two options request, from the Companies house web site, key company documents namely a company profile and access to other key details like mandatory company financial reports

Company Home Page

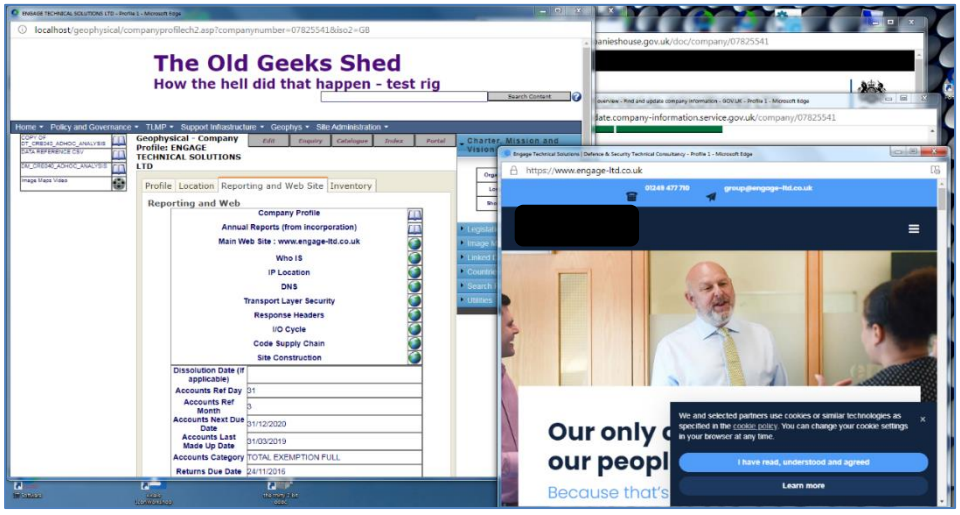

Figure 12 - Company Home Page

Button 3 will launch the web site of the company under review using the operator device default browser. The display of the home page giving access to all pages under the site and if of interest to use browser "Developer Tools" to review things like the site CDN manifest (a list of components built into the site supplied by third parties) and any other review facilities the browser in use supports. The use of the browser also provides the means to use "view Source" and cut and paste page source into the built in code analyser.

Social Media Profile

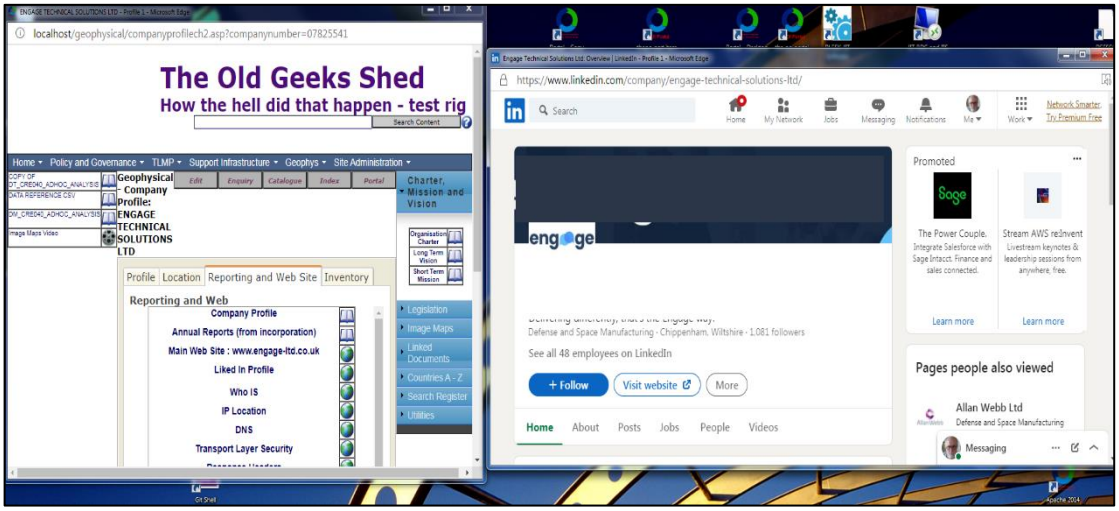

Figure 13 - Social Media Profile

Button 5, will (assuming the operator has an account with the social media platform of interest), launch any company public facing profile in a platform. Quite deliberately, such pages cannot be “scraped”, the aim is to provide the means to view social media content. It should be noted that page scraping of social media platforms is an increasingly common practice that company’s (particularly small to medium sized enterprises) should be aware of as it is a rich source of company information related to business activity and more. In the author’s view a particular risk for the legal profession is a potential compromise of client confidentiality.

Who Is Record

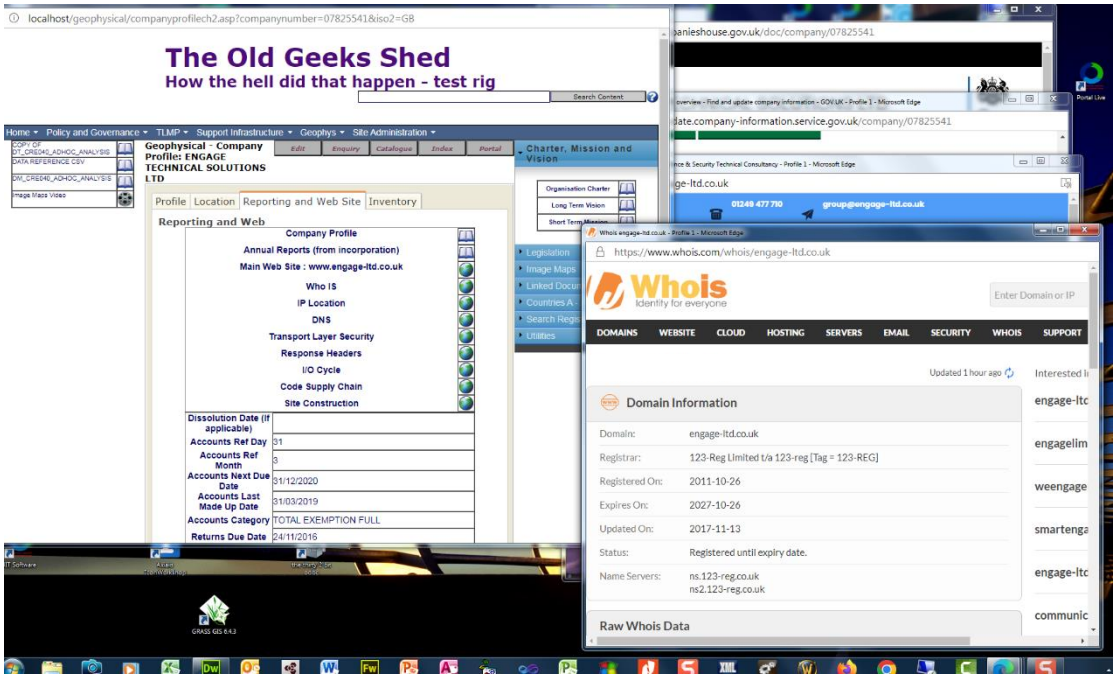

Figure 14 - Who Is Record

Button 6 gives access to the site “Who Is” record. “Who Is” provides the means to identify which individual or organisation a domain name is registered to and when. It should be noted that a key piece of information drawn from a “Who is” record is to confirm the location of a domain owner and host as it is the case that more often than may be realised a web site may have a

national Top Level Domain (TLD), .uk for example, but may not be physically hosted in the country a TLD indicates or be a national of the TLD concerned.

IP and Geophysical

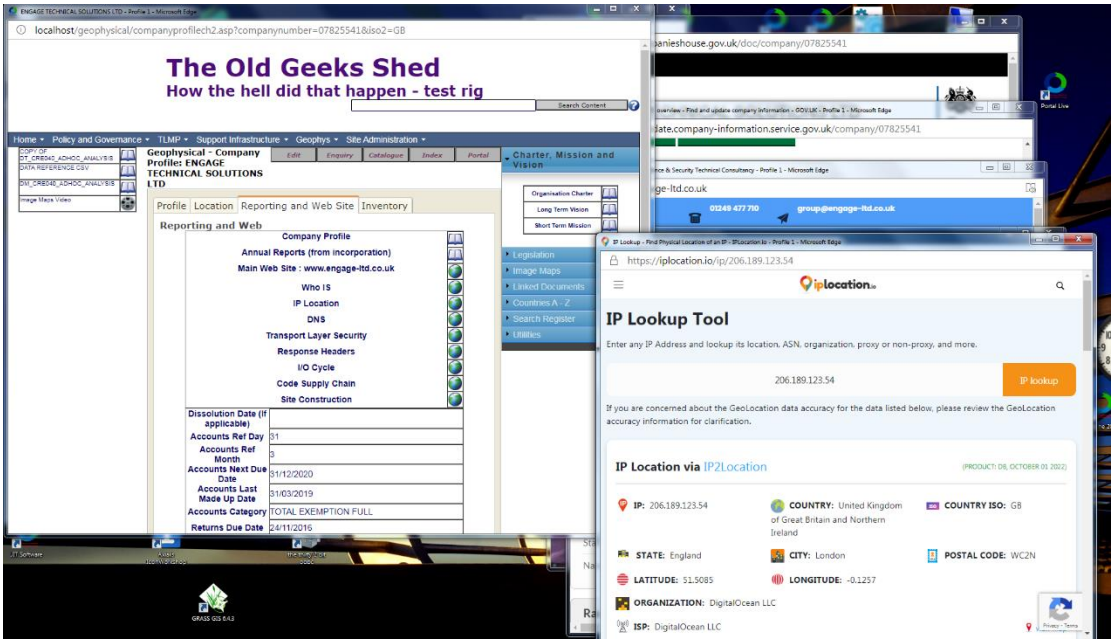

Figure 15 - IP and Geophys

Button 7 will query and retrieve a site IP address and the physical location of the host server. The IP address location being key, for regulatory purposes, for determining which jurisdiction the host machine comes under. ***It should be noted that if a site is operated by a company in one country and the "Whois" record has the domain registered in a second and the IP address of the host is in a third, then that raises the real prospect of site operations being governed by three forms of legislation in three separate countries.***

DNS Layer Security Profile

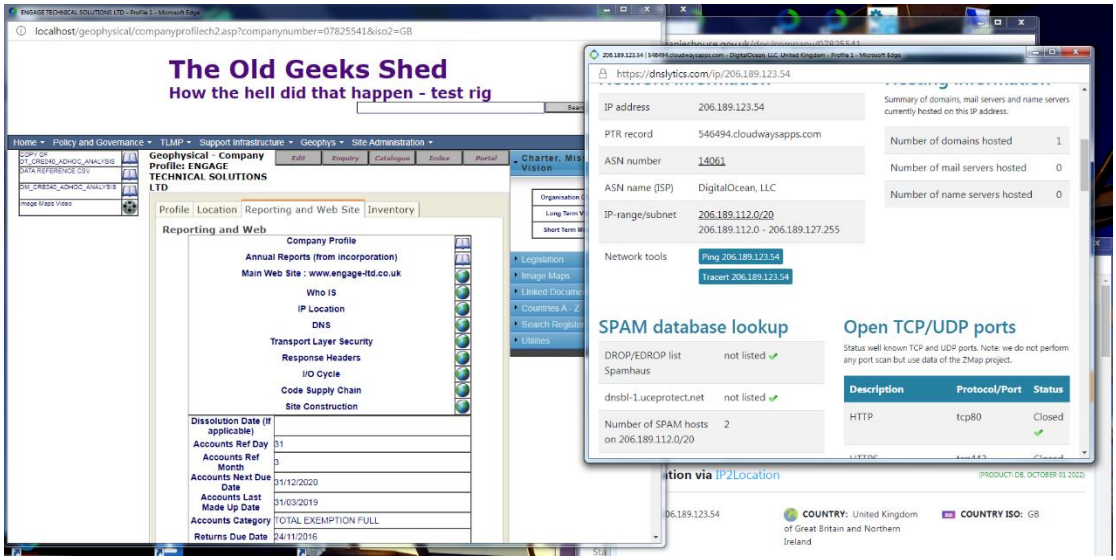

Figure 16 - TLS Profile

The text below is extracted from the Wikipedia article on the [Domain Name System]

"*The **Domain Name System** (**DNS**) is the hierarchical and decentralized naming system used to identify computers reachable through the Internet or other Internet Protocol (IP) networks. The resource records contained in the DNS associate domain names with other forms of information. These are most commonly used to map human-friendly domain names to the numerical IP addresses computers need to locate services and devices using the underlying network protocols, but have been extended over time to perform many other functions as well. The Domain Name System has been an essential component of the functionality of the Internet since 1985.*"

Typically, at domain level, DNS security is neglected. It should not be as DNS hacking is on the rise and represents perhaps the biggest on line security risk organisations of all kinds face.

Transport Layer Security Profile

Button 9 retrieves the Transport Layer Security Certificate details of the site under review. If DNS security failures represent a significant risk to the integrity of a web site, out of date or improperly configured TLS certification runs a close second. There is a mistake beleife that a TLS certificate guarantees site security, it does not, it ensures secure communication and messaging from a host machine to a visitors machine and as a consequence is one of many aspects of site security that need to be considered.

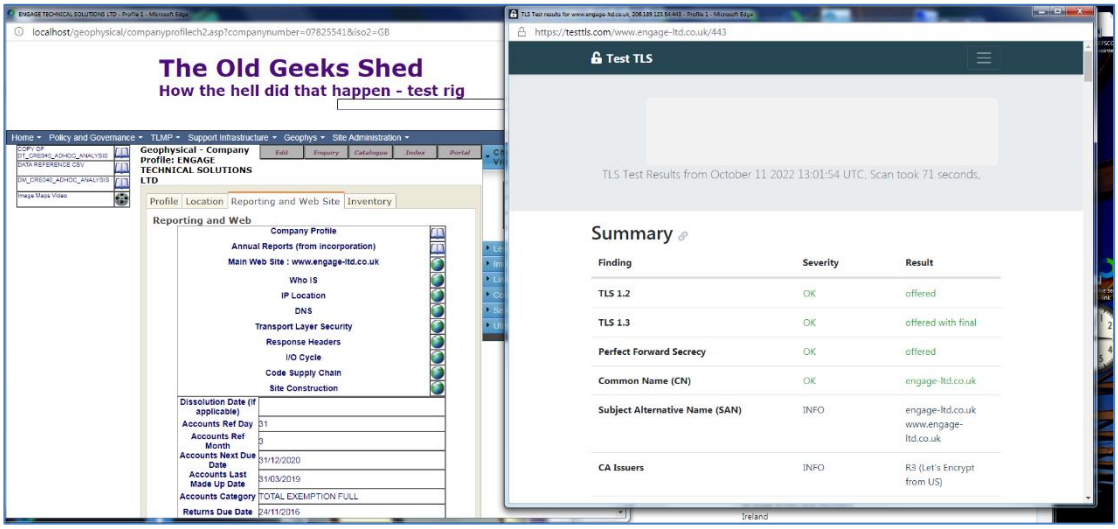

Figure 17 - TLS Profile

From the associated Wikipedia article on Hypertext Transfer Protocol Secure (HTTPS):

"***Hypertext Transfer Protocol Secure** (**HTTPS**) is an extension of the Hypertext Transfer Protocol (HTTP). It is used for secure communication over a computer network, and is widely used on the Internet.[1][2] In HTTPS, the communication protocol is encrypted using Transport Layer Security (TLS) or, formerly, Secure Sockets Layer (SSL). The protocol is therefore also referred to as **HTTP over TLS**,[3] or **HTTP over SSL**.*

The principal motivations for HTTPS are authentication of the accessed website, and protection of the privacy and integrity of the exchanged data while in transit. It protects against man-in-the-middle attacks, and the bidirectional encryption of communications between a client and server protects the communications against eavesdropping and tampering.[4][5] The authentication aspect of HTTPS requires a trusted third party to sign server-side digital certificates. This was historically an expensive operation, which meant fully authenticated HTTPS connections were usually found only on secured payment transaction services and other secured corporate information systems on the World Wide Web. In 2016, a campaign by the Electronic Frontier Foundation with the support of web browser developers led to the protocol becoming more prevalent.[6] HTTPS is now used more often by web users than the

original non-secure HTTP, primarily to protect page authenticity on all types of websites; secure accounts; and to keep user communications, identity, and web browsing private."

Security Response Headers

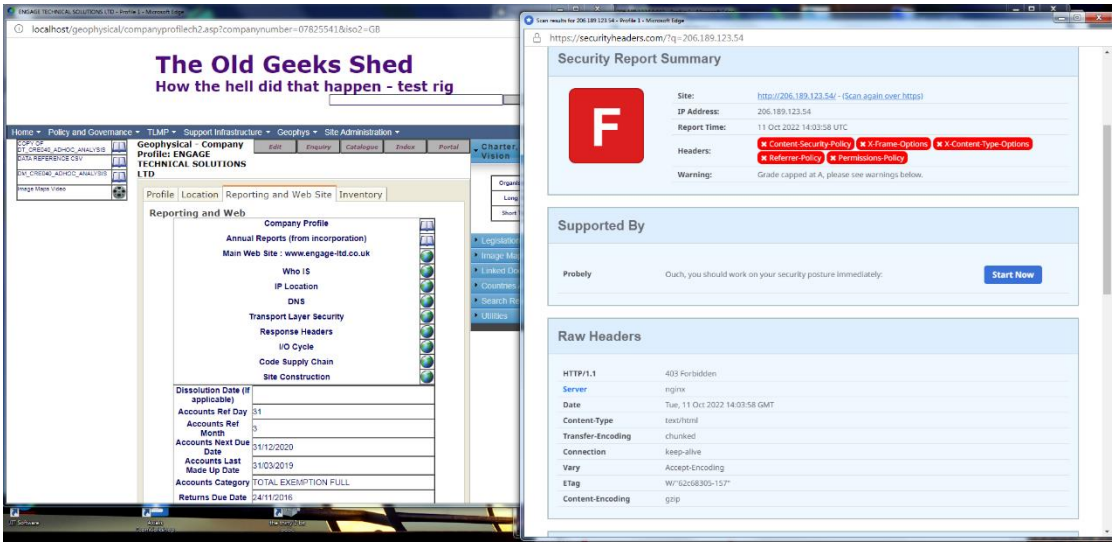

Figure 18 - Security Response Headers

The third security related element of the kind of review supported by the OSINT tools et out in this document focusses on "security Response Headers", on of the more informative locations to find information on this subject area can be found [here]

Like much to do with the security of web site content, response headers are often overlooked. If, on inspection, a site being reviewed failed the kind of response header review the selected tool provided, then a detailed analysis including port scans, penetration testing and more would be recommended. A combination of DNS and response header failures would result in a site under review being declared insecure.

Site Construction

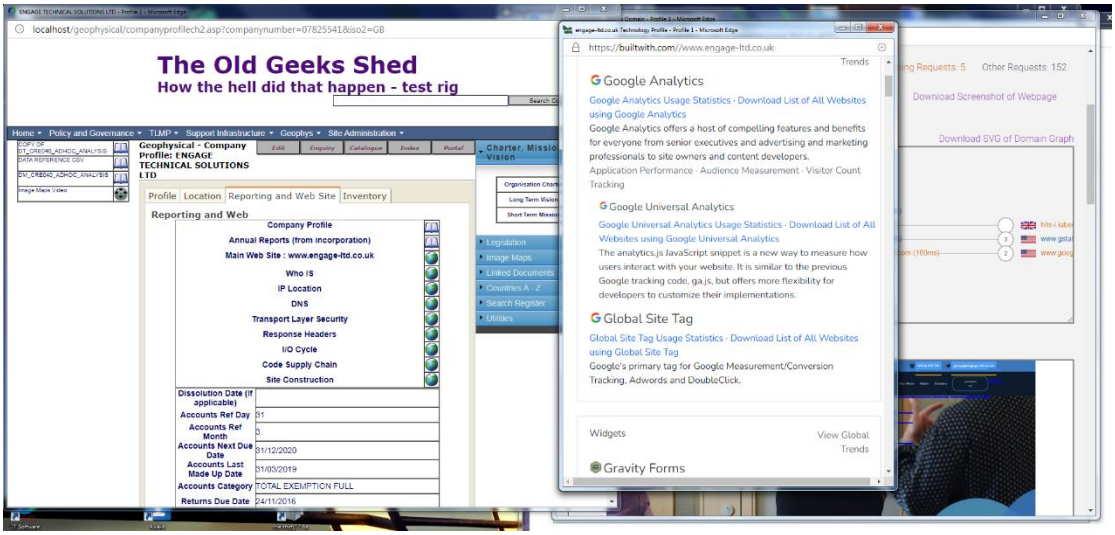

Figure 19 - Site Construction

The next of the external page review views, illustrated in figure 20, focusses on the discovery, in overview, of the way a web site is constructed. As indicated in the previous section, web

sites are composites made up of multiples of pages and components sourced from many parts of the world. In the author's experience, it is a rare thing nowadays to find a web site that is completely self-contained (though in recent months the composite web site seems to be being replaced by more and more self-contained),

The point of understanding in overview of how a web site is built is that it exposes the component provider where possible. With practice, given the way web pages are rendered, then it becomes possible to identify page load sequence, the order in which components are rendered in an end user or site visitor browser.

Understanding the page load and rendering sequence is key to understanding the nature of multiple conversations between site visitors and other organisations that site operators will inevitably may not be aware. In point of fact, when a visitor calls, as requests to load code are made during the rendering process, where the requests are to domains outside of the control of a site operator a separate conversation is struck up between the visitor and a third party that the site operator has no sight of at all. For each conversation, the communication is direct and the host can supply more than what is hoped for and an opportunity presents itself to fingerprint the end user device and browser. In GDPR and other similar regulation terms, that raises issues of the nature of the controller processor relationships that are in operation. It is the case that many component makers are aware of that complication and make provision to mitigate their risk through terms and conditions that elsewhere are described as "sadistic". Their aim? To make site operators solely responsible and liable for any compliance risk. It should be born in mind too that the nature of the legal risk in particular, certainly in the EU, is more wide ranging than just the GDPR.

It is necessary therefore, to have some understanding, in overview, of how a web site is put together, from that comes a need to understand where page components come from in the world. The EU E Privacy Directive (Page 3, para 24) and similar regulation/legislation demand it as it requires site operators to respect the end user "sphere of privacy"

Code Supply Chain

For the average small to medium sized enterprise, of limited technological expertise, that is or has commissioned a web site, the one section of this whole series that the author hopes they read or are briefed on, is this which is about what is known as a "code supply chain".

It is the case nowadays, that a web site is a composite of many components of various kinds, written by any number of people and organisations who can be located anywhere in the world and that combination raises, or should raise, many concerns in respect of who has written code and what the code concerned does as opposed to what computer code is actually claimed to do.

In web site development in particular, there are two seductions, that of "free" and "simple" that are fell for quite frequently by the unwitting, the risk being further hidden by the conceptual mistake that "open source" means free of responsibility and liability. "Open Source" is free of neither. Furthermore, the use of any third party code, called or requested from a domain outside of the organisations home represents a considerable loss of control for the following reasons:

1. Code for components can be amended by the developer on their decision and theirs alone. While patch notes may be issued, for may component providers that may not be the case. In any event, there is an administrative overhead associated with keeping track of code modifications.
2. Code for components can be stored anywhere in the world. Code stored outside of the UK, will not be subject to UK legislation and indeed, depending on the way supporting terms and conditions are written, may mean that the site operator may run the risk of being subject to legislative sanction from other jurisdictions
3. The use of third-party code is entirely dependent on the code security steps taken by code authors which may not be sufficiently robust enough to meet UK (or EU legal requirements and standards).

4. Those writing components are not above including functionality that is there for their benefit. A recent study on the effect of several EU countries and the banning of Google Analytics contains several samples of such things.

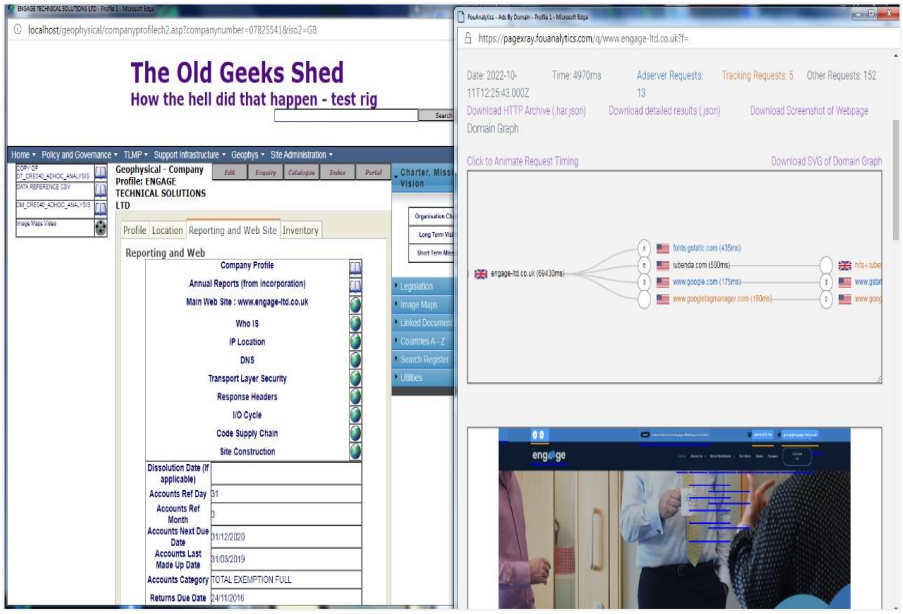

Figure 21 - Code Supply Chain Illustrated

Figure 20 - Code Supply Chain - raw

It should also be noted that the volume of code expressed as “lines of code” (LOC) downloaded by site visitors can be significant. As just one example, there is a commonly used javascript library called “JQuery”, it should be noted that the library is well respected and used by all manner of people, including the author, it consists of a core library (some 12,000 LOC). There are other core supporting components related to the provision of user interface support, which if frequently used, that hold, on average, 2,000 r more lines each. Then there is a library of many thousands of “add ins” each having their own LOC count. Just the use of JQuery can bring about the loading of many thousands of LOC from multiple physical hosts and locations worldwide.

The number of components and supporting code can be quite large in terms of the associated LOC. Figure 21 giving an indication of the potential complexity of a code supply chain has been taken from an on-line news outlet, for just one page of a 2000 page web site. Figure 21 is not particularly large or complex, there are many, many web sites, usually operated by significant companies. Figure 20 is a screen shot of an analysis capability the review toolkit this document is about gives ready access to. Note that in the site the screen shot focusses on, it is from a UK company, but it calls code from several US based organisations at least one of which takes a visitor browser fingerprint. For both, the international reach of the components indicates both a compliance and a data gathering complexity that should not be ignored, but all to often is.

What should also be of concern is the ingenuity applied to the data gathering effort. Clever does not quite cut it to describe it, much of it, through various exercises in subterfuge, invisible to site operators and visitors.

And with that, moving from the use of tools, to a more detailed look at the way a web site might work and how to exploit that.

Server Side Scans (So Far)

Delving Deeper

As described earlier, the internet leaks like a sieve. Many, many organisations and individuals understand that and exploit it for advantage. One of the key operating concepts of exploiting to advantage, an operational web presence is referred to as "Search Engine Optimisation". Companies like Google and Microsoft, in their technical documentation provide guidance on the means they provide to optimise a web presence in order to attract visitors to a site. A process known as web crawling is one of the primary means of cataloguing web content using a piece of software designed for the purpose known more frequently nowadays as a "crawler" or "bot".

Increasingly, the concept of web page scraping is applied to catalogue site content, in detail for any number of reasons, but, again and to reiterate, usually for commercial gain. There are means to restrict crawling and scraping attempts but they hinge on the idea that those operating web sites are aware that it is happening and what it can be used for. This part of this document explains how a limited page scraping capability has been built into the site review toolkit presented here. The scraping described in the content that follows will not be used for commercial gain. However, given that the law on scraping is ambiguous, then it falls to site operators to take actions to block scraping efforts, bearing in mind that search engine optimisation is a form of scraping given that sites and pages are crawled and their content catalogued for search engine use.

For the benefit of site operators, the list below, taken from the Wikipedia article referred to above, the following guidance is offered on restricting crawling and scraping activity:

"The administrator of a website can use various measures to stop or slow a bot. Some techniques include:

- *Blocking an IP address either manually or based on criteria such as geolocation and DNSRBL. This will also block all browsing from that address.*
- *Disabling any web service API that the website's system might expose.*
- *Bots sometimes declare who they are (using user agent strings) and can be blocked on that basis using robots.txt; 'googlebot' is an example. Other bots make no distinction between themselves and a human using a browser.*
- *Bots can be blocked by monitoring excess traffic*
- *Bots can sometimes be blocked with tools to verify that it is a real person accessing the site, like a CAPTCHA. Bots are sometimes coded to explicitly break specific CAPTCHA patterns or may employ third-party services that utilize human labor to read and respond in real-time to CAPTCHA challenges.*
- *Commercial anti-bot services: Companies offer anti-bot and anti-scraping services for websites. A few web application firewalls have limited bot detection capabilities as well. However, many such solutions are not very effective.[28]*
- *Locating bots with a honeypot or other method to identify the IP addresses of automated crawlers.*
- *Obfuscation using CSS sprites to display such data as telephone numbers or email addresses, at the cost of accessibility to screen reader users.*
- *Because bots rely on consistency in the front-end code of a target website, adding small variations to the HTML/CSS surrounding important data and navigation elements would require more human involvement in the initial set up of a bot and if done effectively may render the target website too difficult to scrape due to the diminished ability to automate the scraping process.*
- *Websites can declare if crawling is allowed or not in the robots.txt file and allow partial access, limit the crawl rate, specify the optimal time to crawl and more.*
- *Load database data straight into the HTML DOM via AJAX, and use DOM methods to display it. No visible data in the source document means that it can't be scraped."*

However, regardless of attempts to block, once a page is rendered by an end users browser, then it is the work of moments to review the code either visually or automatically as inevitably web site source code is delivered into the end user device in the majority of case.

In addition to page content, by convention, there are a number of kinds or types of files, usually stored for access by crawlers and bots from anywhere that are designed to catalogue site content. In short, put anything of any kind up on the web and because of the way the web works, it should be assumed that such content is lost (hence the authors copyright message). The remainder of this section will attempt to described how, easily accessible source code and other artefacts can be used for business intelligence gathering purposes.

Code View

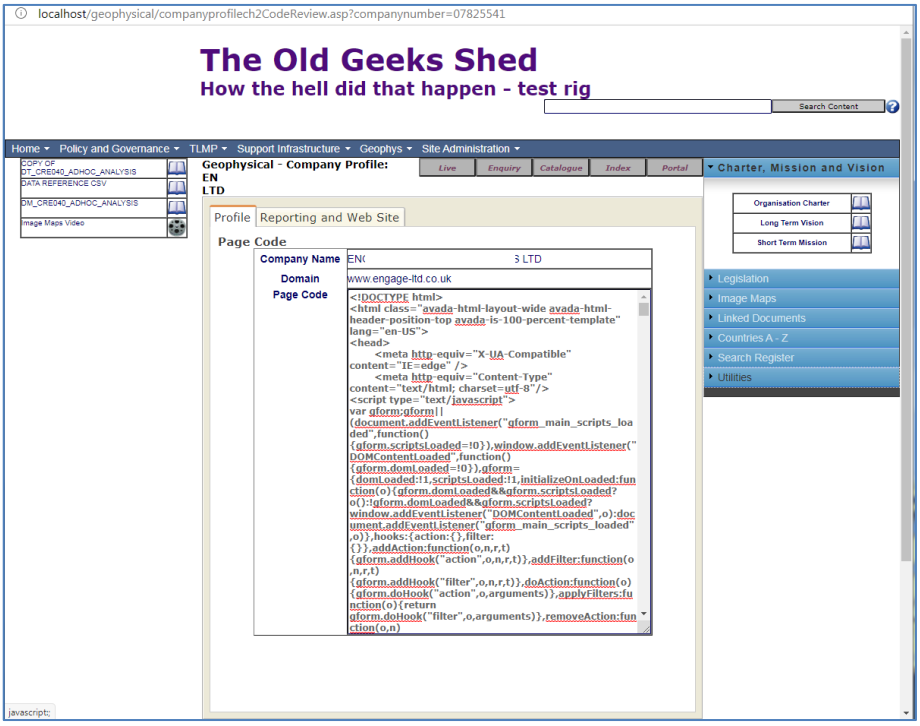

Figure 22 - Code Review

Having set out some of the issues related to code review and web pages, the issue then is why to do it and how. The why, quite simply, is that arguably the greatest risk to privacy protection and the greatest business risk is what happens at the client end of any electronic conversation between a site visitor and the operating organisation. It should be understood that code scraping or parsing is a feature of many of the tools used in the supporting client side review template on which this document is based.

The “how” has a number of issues, the first major one being how to get at code delivered client side, the second being what to look for in respect of generating business intelligence. In respect of the how, once a domain of interest has been identified, a web page can be read in its entirety, electronically and with increasing sophistication, code delivered client side by third parties can be read and examined too. The illustration above shows a sample page read (in this case simply cut and pasted into a text box to be stored for more sophisticated review based on parsing each line of code.

The second part of “how, what to look for, takes a little research but in the authors growing experience, the scope of evidence that can be used for the purposes of examining code is growing almost by the day. In no particular order, the author has copies of lists of domains, beacon files, web crawlers, adtech domains and more that can be incorporated in a simple text search facility in the body of any code in a site. Supporting code, to detect host domains and IP addresses for list entries have been written as has a simple frequency scoring mechanism.

What the use of such lists for parsing demonstrates, unequivocally, is the scale and scope of data gathering in any web site. In short, once scanned, code can reveal a considerable number of business relationships, many of course known to a site operator, but with increasing frequency, many more that site operators may not realise the implications of. For an example of the nature of the relationships that may not be fully understood, the readers attention is drawn to a Data Protection Impact Assessment (DPIA) written by author focussing on the decision to close the live version of the site providing the illustrations in this document.

Data collected by a code scan, when linkded to other data collected on a site review as well as market place data of the kind described in part four of this series increases the scope of business profiling to cover so much more than what is held by a market place owner.

The next two parts of this document describe a further enhancement, that is the means to scan read and interpret the content of files other than web pages themselves.

ROBOTS.TXT

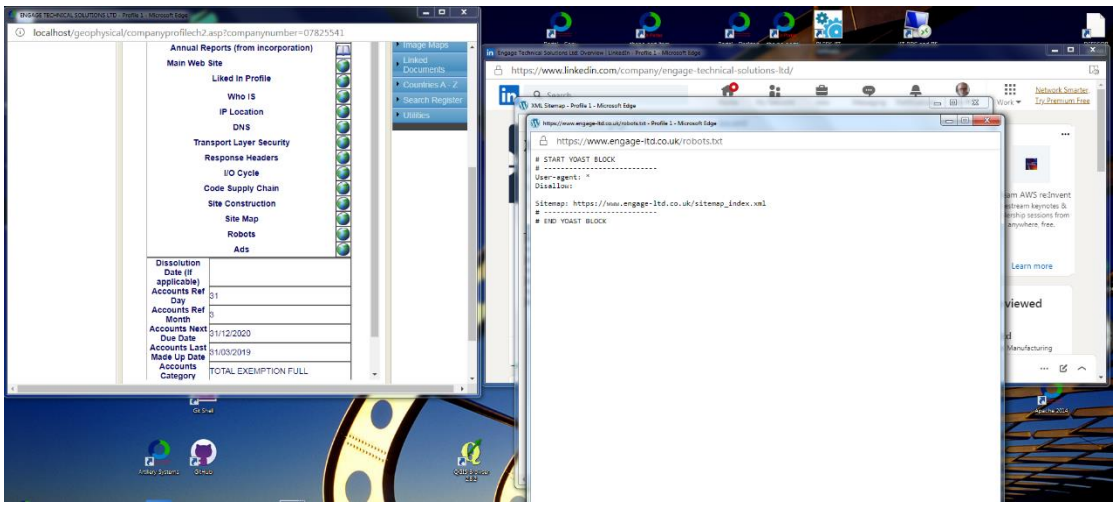

Figure 23 - Robots.TXT

The way the internet works and is governed is often expressed in the form of a “standard” or a set of rules that those using the web, for any purpose are expected to follow. In principle, the standards are followed, some many of them have to be followed in order to maintain connectivity and more besides. Some however can be written and not followed. As a standard that has been written, the Robots Exclusion Standard is one of those standards that can and often is ignored. Its purpose, explained in the link in this sentence, is to provider guidance to web crawlers and bots about files and file locations in a site that an operator would rather not be catalogued or have some other cataloguing restrictions placed on crawling. Typically, a ROBOTS.TXT file can be found in a site root directory. Figure 23 provides an illustration of a simple robots.txt file. It can be read, if present, by anyone simply by adding the file name to a domain URL:

<domain name>/robots.txt

Once requested, its contents will be displayed in the default text editor of the querying device if such an editor is installed. A very good guide to how to use robots.txt can be found in the Google Developers SEO guide here. When the author operated his own site, his version consisted of a simple “no follow” directive which legitimate bots and crawlers obeyed. However, “what the geek giveth, the geek can taketh away” and not all bots and crawlers obey the files contents. One of the things a complex robots.txt file inevitably gives away is the locations of site content that operators would rather not be catalogued, which inevitably does is draw the attention of those of a curious enough to target material flagged in a robots.txt file.

Readers should bear in mind that a ROBOTS.TXT file is but one of possibly many files that may be found in a root drive by convention. A second is:

SITEMAP.XML

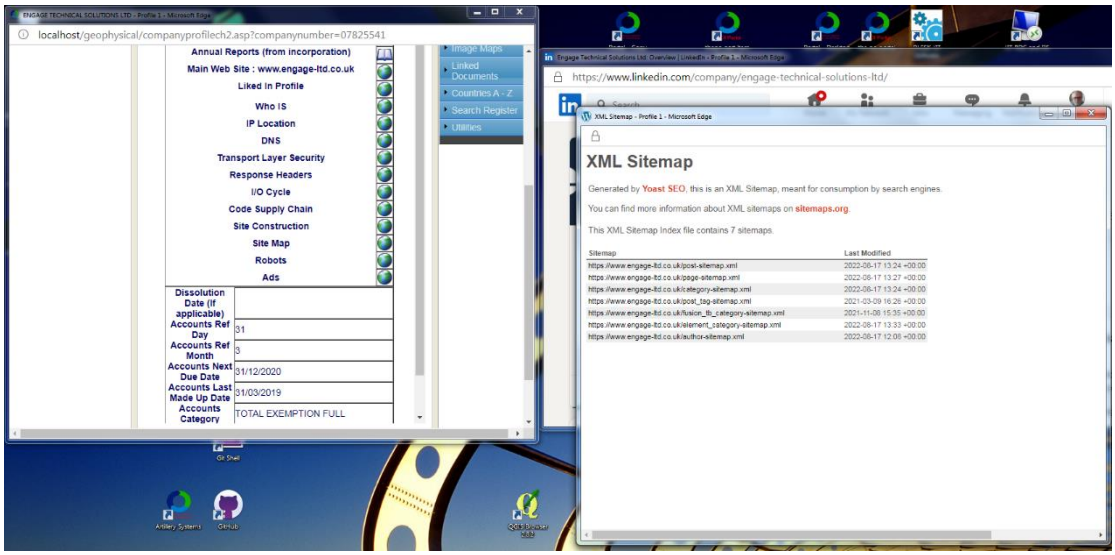

Figure 24 - Site Map Schema

Like ROBOTS.txt sitemaps in XML form are based on a standard way of describing the structure of a web site, An explanation of the form the standard follows can be read here. The purpose is to support efforts related to rising to prominence in search engine listings known as "search engine optimisation" (SEO) the basic idea being that given the size of search engine catalogues nowadays, thee is a need to design web sites to make it easier, technically speaking, for a search engine bot or crawler to catalogue both a page and page content. However, as with other forms of server side content a site map lays out a site structure and that often includes sensitive material excluded in something like a ROBOTS.TXT.

If, as is normal, a sitemap is in a web server root folder, like ROBOTS.TXT, accessing and reading it can be achieved by requesting the map itself in a similar way to that for the ROBOTS.TXT file

<domain name>/sitemap.xml

As with everything to do with operating a web site, how sitemaps.xml is deployed and configured is a matter of serious consideration, not least because it is a source of business intelligence

Ads.txt

Of the three root folder files that are freely accessible, potentially the most fruitful for those seeking to gather business intelligence on a company or organisation, in the author's opinion especially when there is access to something like the UK Company Register, is a file called "Ads.txt" and its close cousin "app-ads.txt" often deployed on mobile devices as part of an applications infrastructure.

As described in the "ads.txt" article, linked to above, the purpose of both is to provide the means, for search engine optimisation purposes (amongst other things) to specific various forms of business relationships between a web site operator and its associated resellers and other stake holders. As the Wikipedia articles above explain, the Ads.txt file (a sample is illustrated in figure 25) has a legitimate business purpose, to the benefit of site operators, it can be read by anyone who knows how to. Like the other files mentioned so far, requesting a copy of ads.txt follows

the convention set out for sitemap.xml and robots.txt, simply add the file name to the primary domains and if an ads.txt file is in the root drive, it will be retrieved for review.

<domain name>/ads.txt

Figure 25 - Ads.txt

What the file gives, if it is present, is a detailed profile of the way a site operator does business. In short, extremely valuable business intelligence from “the horses mouth”. Linking the content to something lie company records and building in relationships between companies give a considerable advantage to those who know to do so and can do so.

Then There Is The Rest (And of Course Page Scraping).

The three files described in overview in the preceding sections are not the only files that are sometimes placed in a site root that are useful for business intelligence gathering purposes. Inevitable, there are others. All that is needed is a list to test against. There are thousands of them, that there is, is indication of poor site design, the presence of each can be tested for, most being entirely innocent, many with a purpose directly related to visitor device profiling like “modernizr.js”. Web, or page scraping of a more comprehensive nature than what is described here, being a capability that some computer languages are particularly well suited to.

Bluntly, the web leaks like a sieve and the presented risk to small to mediums in particular, is existential

This document focusses on the way internal sales data can be associated, as a series of relationships between suppliers, with supplier online presences. The market owners ability to expanded, in terms of the number of relationships the detail of the way suppliers do business away from the market place concerned. A key question is, is it really possible, both technically and architecturally (the answer is yes) and are those who operate online markets, of any kind doing so?

The fact of the matter is, that the majors, known as the GAFAM's are doing so as are many others. In the developer support pages of GAFAM web sites, they make the means to identify such patterns of behaviour to those who choose to exploit them. This from Microsoft, or this from Google, and from Oracle, from Meta and this from AWS are all indications that each provides advice and guidance to its site operator customer in respect of how to take advantage of their commercial platforms just in respect of improving their clients chances of being in a better chance to be chosen when end users search for things they may wish to learn about, purchase or whatever. What is often not understood, is the business advantage the use of the kinds of SEO capabilities combined with visitor tracking bring to them as the provider of such services. Entering the market places of any of the majors, without studying the way their market places work is to introduce into a business a considerable commercial risk which does not usually make itself apparent. That is not to suggest anything illegal is happening, it is to suggest there is an element of "caveat emptor" in the grand scheme of things and any such move into a major on line market place needs to be considered very, very carefully.

Just an Observation - Controlled Borders…

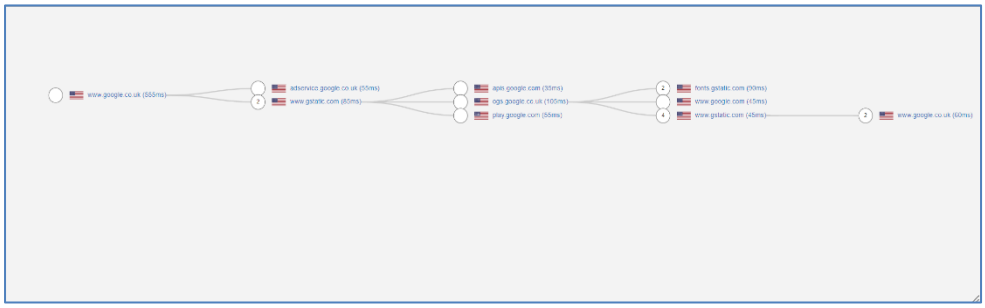

Figure 26 - Just an observation. Controlled Borders. Google

Given that the scale of data gathering is so diverse, it is worth trying to understand how vulnerable those who collect significant amounts of data they send to others might be, and how tightly the others control their own organisation boundaries.

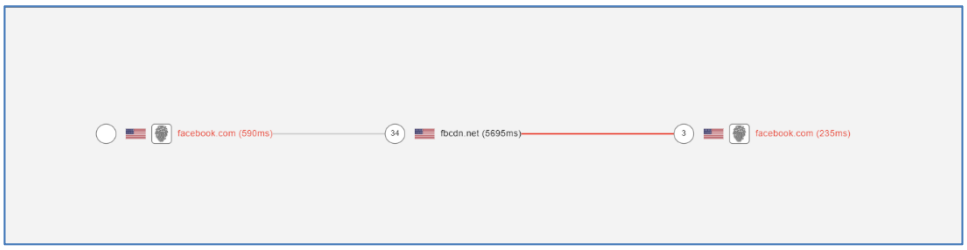

Figure 27 - Just an observation, Controlled Borders, Facebook

The images in this section are taken from three web site code supply chains that together are companies that know how to exploit data and have the processing and analytical capabilities to do so.

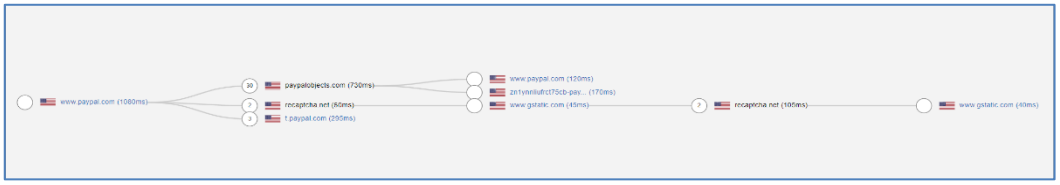

Figure 28 - Just an observation. Controlled Borders. Paypal

Each has a tightly controlled border, no third-party components, no links to ad tech sites, and each giving away no data as a result. But the site operators of each are data gather's fed by the millions of companies they supply services to and the billions of end user service consumers they appeal to worldwide.

It should be noted, that each company does not sell data, instead they capture it, for analysis and then sell on the analysis results in some other form, typically under the banner “ad tech” by auction. Then there is this…

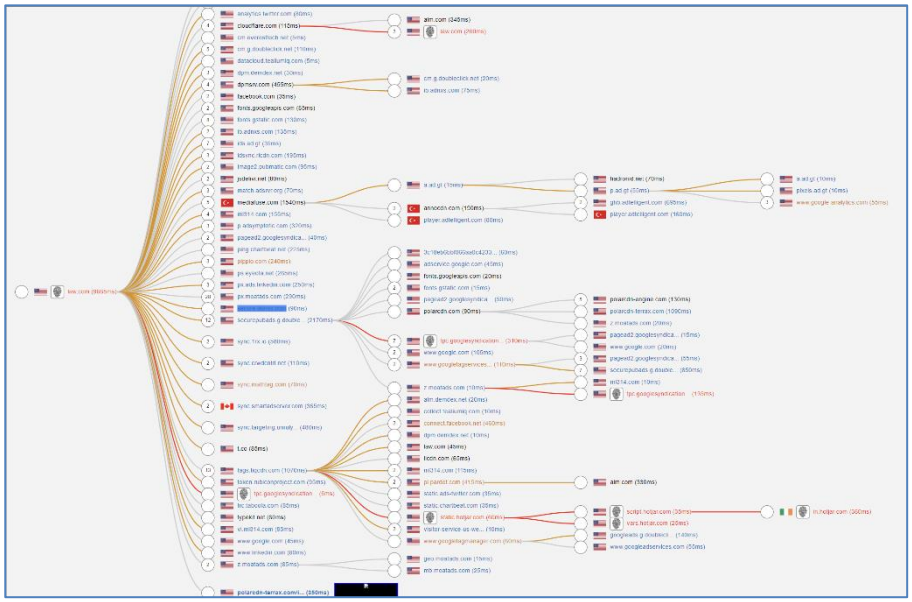

Figure 29 - Just an observation. Uncontrolled borders. A legal fraternity web site.

Readers are invited to draw their own conclusions. Using code components without checking client side impact is not wise, nor legal given the nature of consent. However, what the use of such components also does, is to give those of a curious mind, to exploit the use of such code and the files mentioned earlier, for their own reasons. Usually commercial gain. In the case of the site concerned, in the author’s view, there may be a risk to the integrity of client confidentiality. Because it too has a comprehensive “ads.txt”.

If You Can Do This, Who Else Can?

Short answer, anyone with the skills and motive can do the same.

Conclusions

This is part five of a series describing the nature of a model designed and implemented to demonstrate that an online market place, of any kind gives the owners of the market place a considerable advantage when it comes to being in a position to understand patterns of behaviour of both suppliers and customers on a “though life” and “forward and reverse supply” basis.

One of the underlying themes of this document is to also capitalise on the availability of OSINT. In the content, it is described how the UK company register, with the basic company record extended structurally, to incorporate just a domain and IP address can be used by those with the will to enhance the data gathering capabilities of those who already have a considerable business intelligence analysis capability anyway to build a more complete profile of the suppliers to their marketplaces in whatever form they take.

In the author’s view, the combination of data already held on through life transaction processing and the ability to profile things like external business relationships give hose who own the markets to achieve a dominance that has never been seen before and explains why those who own the markets are multi billion dollar companies in which even through customer recommendation schemes, the benefit to them is that purchasers of goods and services do their product quality assurance for them.

In that kind of operating environment, while the current crop of data protection law is well intentioned, quite simply. it is not fit for purpose. The sheer scale of data exploitation going on renders it difficult, if not impossible to police effectively. If there was a recommendation, given that much of what is happening in respect of data gathering, I enabled purely by those who write the software that makes it all possible, that they should be licenced to practice and be obliged to demonstrate they have professional liability insurance at a level to cover all of their portfolio given that it is site operators who are held accountable for data breaches and much more.

The current level of computer related crime is not in the $Tn world-wide. The concept of "fail fast and fix things" followed closely by "code before documentation", both being adopted with near casual abandon by the development community, both being little more than a licence for mayhem online. Why? If developes use code they have not checked and validated, from an external domain or CDN, then they are surrendering control of data and content to 3rd parties. As simple as that. Developers it seems, are that rare thing, a bunch of people who are not held properly accountable for their work. That should be unacceptable, in any other form of engineering, it would be.

Heads up 2.

Recommendations

Bear in mind that a web site is a window into your organisation that the world can look in to if there is an interest. Consider that as interest grows, doorways are opened into the organisation and an increasingly sophisticated conversation takes place between the visitor and the web site (which is not the same as talking to an employee, even with a chatbot). If data is the new gold, then handing data over to anyone, without understanding how it is being done or where it is going and for what purpose, is handing over gold that has been validated or verified by the horse with the mouth.

"Open Source" does not mean free of risk, responsibility or liability, if code is called form afar, the effect is that those doing so are handing over considerable control to other people, over which, there is no real authority or constraint.

Above all, the employment of a web site developer, either as an individual or as a company, should be a matter of serious consideration. Ultimately, it is they who are responsible for web site (and other applications) behaviour, it is they who with a casual shrug of the shoulders, walk away without concern or shame, in the event that things go belly up (as they do with increasing regularity).

Recently, the author was made aware of a UK company, that commissioned a web site, from an American Company, who in turn outsourced coding to a team in India (an increasingly common way of doing business it seems). The developers placed the company on one of the major cloud hosting platforms in the US, at a stroke generating a significant number of technical and legal compliance issues that the company is still blissfully unaware of. The site has cost the company commissioning it, a six-figure sum so far, just to build, with a monthly rental charge of SIRO £500 pcm, at the moment. Because of the complexities of the way the site works, that company now has a dependency on the US developers that will be near impossible to shake off and in the event that the UK decides to try, will bring the growing realisation that they will have to write off the monies spent already and then spend a similar amount of money and time to rebuild. Why? Because amongst other things, the developer are using a framework over which they too have no control nor ownership produced by a GAFAM.

Choosing a developer, is a major strategic decision, choose wisely, guidance on developer selection is available here. Heads up 3

What Next?

The tools, and more, illustrated in the images in this document, used to test the nature of the viability of embedding them more closely with "in market" data, are listed at Annex A. It is

recommended that readers experiment and learn how to use them. As for the author, he will be building his own site review toolkit with the aim of replacing the applications illustrated with his own because he can. This is the equivalent of a new bit of a trainset in the shed.

The next section, part 6, will focus on closer integration of the desktop by means of building components integral to office automation software, focusing on Microsoft Office.

Document Assurance

Document Ownership and Amendments

This document is owned by the Old Geek and its application and monitoring will be carried out by the Old Geeks terrier.

Document Review Timetable

This document will be reviewed on or about 31/05/2024 or in the event that circumstances may change and a review may be necessary before then.

Quality Management and Compliance

This document is training document. The tools and web portal illustrated in this document will be used to illustrate the capabilities described in this series of documents on request.

Document Sign Off

This document has been reviewed by the The Old Geek and approved for inclusion in the Shed Training Pack catalogue.

Printed/Typed Name	Signature	Date

This document has been reviewed by the SWMBO and approved for training use.

Printed/Typed Name	Signature	Date

This document has been reviewed by The Author and approved for distribution.

Printed/Typed Name	Signature	Date

Annex A. The Document Set

Part 1. Sets out the reason to produce the series itself focusing on the complexity of building a business intelligence platform and the advantages to be had by starting on the basis of "organisation form, function and purpose" as opposed to the silo inducing "People, Process and Technology". It raises the idea that data collection is linear and procedural, while reporting tends to be organic and viral. Setting the tone for the rest of the series.

Part 2. Describes and illustrates some operating principles that the author applied over his time as an information management practitioner (some 30 years) setting out a series of concepts on which to base development on the basis of "POSIWID" (the purpose of a system is what it does)

Part 3. Expands on part two introducing concepts associated with the organisation as a system introducing the idea of "capability maturing". It introduces the idea of mapping an organisation form and purpose and then using the resulting map as a basis for the construction of a web-based software infrastructure (a risk register, image mapper, legislation librarian and more) as a precursor to the integration of operational data.

Part 4. Uses a "market place" analogy as the basis for describing the nature of forward and reverse supply chain operations, which when linked into the map based infrastructure described in part two, gives the means to structure reporting capability, through the means of controlled and architecturally linked alignment and attenuation exercises to construct the first stage of a coherent decision support capability that can be verified and investigated "top down, bottom up and left to right of arc) as alignment and attenuation drives validation at the level of data collection

Part 5. This extended the reach of the now formed business intelligence capability to take into account the wider internet, illustrating how to link internal data to external OSINT data sources like the UK Company register, elements of the UK Census and key geolocation data types such that business sectors and geographical areas could be examined on the basis of socio-economic issues and more.

Part 6. This document extends the business intelligence capability further to include the means to become more and more aware of the behaviour patterns of customers, stakeholders, and other service users if those who have access to vast quantities of data have the will and the means to do so.

Part 7. Addresses a method and means to introduce the concept of "desk top to server" integration using a number of software tools (add-ins to the MS Office suite, stand-alone programmable components and more) and training aids to provide "end to end" contextually coherent information delivery.

Part 8. The wrap up. This document contains details of "lessons learned", recommendations, names "thought leaders", provides a reading list and so on so that those of a mind can research in greater detail if they so wish. In the author's view, it is in the readers interest to do some further research, not least because the software and other material presented in this set of documents is not for sale. Seeking alternatives, if readers chose to do that, will take some time.

Annex B – Extended Tool Kit

Over the course of the past three years, the author has used a number of tools to support technical web site review to prove elements of the nature of compliance. In addition, the author has produced many of his own. The toolkit in this Annex was initially produced for the associated DPIA exercise. The tools are easy to use, in the main, basically, all that is required is to cut and paste a URL and into a text box, click a button and await the results. Interpretation of the results may require practice.

The images in the body of the review are taken after using tools in this list. Where appropriate, in an image caption, the tool used for an image is indicated. This means that review results can be replicated, live if the reader takes a notion

There is nothing contentious in the toolkit all of them are part of a standard software test environment, nor are they the only ones of their kind.

Ser	Tool	Purpose	Link Address.
1	Site Response Header check	Use to prove the viability of site http response headers	Left click here
2	IP address geolocation	Used to identify the physical location of a server or other fixed device	Left click here
3	Fou Analytics Web site page Lineage	Used to identify, in a tree/node schematic the nature of links between a home domain and any CDNs.	Left click here
4	Whois is a catalogue of web site registrants.	It will tell visitors who owns a registered domain. It should be noted that because of discussions related to privacy matters, names of individuals are not visible on any returned entry. It should be noted that in the event a search fails, users may be directed to other registrant catalogues.	Left click here
5	WeblogExpert	A server side log file analysing toolkit. Much less risky than using a site visit analysis than using a third part tool via a remote call to a CDN. Cheap too	Left click here

6	Traceroute	A command line utility used to track connections or “bounces” between devices on an IP network it should be noted that the link provided here is for MS Windows platforms. Different operating systems will have similar capabilities.	Left click here
7	Ping “Time To Live” Table	To test if an IP address is live, it is possible to “ping” it, or send a signal saying “are you there”. Ping results give a “time to live result which can be used to determine the IP host operating system	Left click here
7	Code review	A review of source code	Contact the author for a copy
8	Case Study	behaviour on a single page of a single web site.	Contact the author for a copy
9	Web supply chain investigation	A tool that can be used to investigate the nature of domain linking from site header information	Left click here
10	Domain IO Scan	On page load, this tool provides a means to check the nature of connectivity in some detail	Left click here
11	Site security certificate checker	SSL certs are an indicator that the site is using encrypted comms between visitor and host device. They run out or expire. Nor do they provide complete server side protection (they are not meant to.	Left click here
12	DNS Viz	A tool to detect and determine DNS	Left click here

		security status from TLD down.	
13	“Built With”	A tool that gives a sophisticated description of the components and tools a web site has been built with.	Left click here
14	The “Way Back Machine”	One of several web archive platforms that can be used to review a domain and its site history	Left click here
15	Visual Traceroute	A visual trace route tracking capability	Left click here
16	GSuite tool kit	A single site with all of the tools in earlier serials	Left click here
17	Web Server Detector	Detects the nature of server side software that is used to manage browser sessions and the allocation of server side resources	Left click here
18	Cookiepedia	A catalogue of well known and used cookies, their format and purpose	Left click here
19	Code “Prettifier”	Typically, javascript code dropped into a client machine is compressed, or obfuscated to make it difficult to read. Under the jargon “prettify” the tool listed here reformats code to make it more readable	Left click here
20	DNS Lytics	A sophisticated DNS analysing tool	Left click here
21	Abuse IP DB	And IP address blacklist catalogue	Left click here
22	Netstat	Windows TCP detector, run as command line software	Left click here
23	Process Hacker	Desktop system reviewer	Left click here
24	A Malware Database	Contains list, regularly updated, of detected malware.	Left click here

25	Virus Assessment	A tool providing a collation of site audits for malware of various kinds and whether or not a site has been infected	Left click here

There are of course other tools and architectural elements that should also form part of a review, in a Windows World, if operated, a client side active directory should be considered(lose control of a Windows Active Directory (AD) and any semblance of control of a system or platform is lost) and learning about the nature of the .Net or .Com controls on the client machine may be considered too. However, both (and other components) are beyond the scope of this template. Seek advice on their implications if readers think it is appropriate to do so.