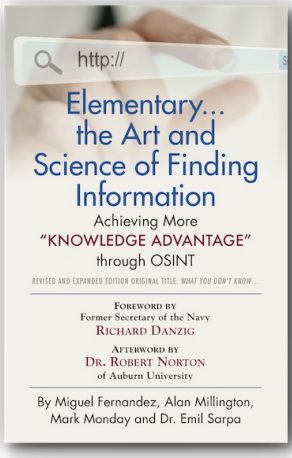
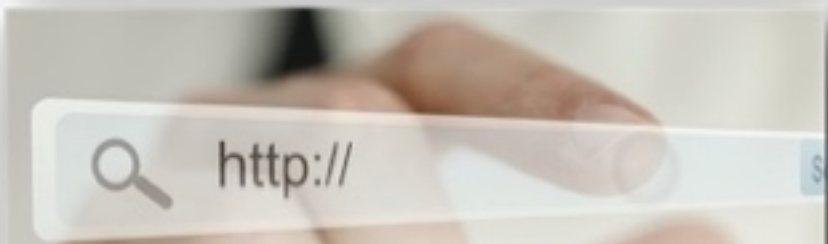


Facts and fake news live side-by-side in the Internet. Elementary... the Art and Science of Finding Information is a guide to achieving information superiority in today's world.

Elementary... the Art and Science of Finding Information: Achieving More "Knowledge Advantage" through OSINT – Revised and Expanded Edition

Original Title: What You Don't Know...

By Miguel Fernandez, Alan Millington, Mark Monday and Dr. Emil Sarpa



Elementary... the Art and Science of Finding Information

Achieving More "KNOWLEDGE ADVANTAGE" through OSINT

REVISED AND EXPANDED EDITION ORIGINAL TITLE: *WHAT YOU DON'T KNOW...*

FOREWORD BY
Former Secretary of the Navy
RICHARD DANZIG

AFTERWORD BY
DR. ROBERT NORTON
of Auburn University

By Miguel Fernandez, Alan Millington, Mark Monday and Dr. Emil Sarpa

Paperback ISBN: 978-1-64718-066-9
Hardcover ISBN: 978-1-64718-067-6

Library of Congress Cataloging in Publication Data
Elementary... the Art and Science of Finding Information by Miguel Fernandez, Alan Millington, Mark Monday and Dr. Emil Sarpa
Library of Congress Control Number: 2019917913

Published by Boolocker.com, Inc., Bradenton, Florida.

BookLocker.com, Inc.
2019

Revised Edition

Table of Contents

1. Introduction and Overview

Knowledge is power. That is elementary!

That's what this book is about: Gaining the knowledge that gives you the power.

We supposedly have lived in, or perhaps through, "The Information Age," not that anyone would know it. There are understandable fears we have "progressed" to "The Post-Information Age," where the noise that humans attach to data and information has drowned out its utility.

Yes, we have tons of data and information available. But data is unprocessed facts, figures without context. Information is the actionable facts or figures that are presented in a meaningful context; they are the things that bring insight, that transform mere knowledge into power.

Despite all the ballyhoo, in too many cases the readily available data and information of the Internet and libraries have yet to be sculpted into insight. Insight is valuable; it leads to good decisions and guides you away from many bad ones. Insight gives you the edge you need, the knowledge that is power, the information supremacy that makes your success more likely. But information supremacy, for now, remains tantalizingly out of reach for many.

There is often more haystack than there are needle-points of insight. One of the first problems for OSR investigators may be finding out which – of the many haystacks in the field – to look through in order to gain insight.

OSR is a practical way to gain insight. There is a need for OSR in today's world. Too few people know how to actually use OSR to their benefit. But when properly used, it can be remunerative.

Sherlock Holmes allegedly said it best more than a century ago in *A Study in Scarlet*: "Yes, I have a turn both for observation and for deduction. The theories which I have expressed there, and which appear to you to be so chimerical, are really extremely practical — so practical that I depend upon them for my bread and cheese."

In today's world OSR is at once a profession, a talent, and a necessary life-skill. Hunch, deduction, or planning – all may work in finding essential information. But planning and an effective SOP offer a better chance to find you the facts you don't know, but need.

,
What you don't know can hurt you.

What you don't know may make a tremendous difference.

What you do know can help you. What you do know, or can learn, will make you money and provide greater satisfaction in life.

Facts and truth matter because they drive decisions and beliefs. Successful people learn, know, use, and value facts. But too often people rely on information sorcery and magical thinking rather than reliable information sources. There is a vast amount of data and information, but a paucity of people who know the rough – let alone finer – details about gathering, vetting, and using the available data or information to gain insight.

Marginalia can be more important in an information search than many people understand.

There is both positive and negative evidence to reap. While most people are aware of the former; fewer can see that the lack of

something can also be indicative. Sherlock Holmes understood the concept of negative evidence and explained it in *The Adventure of the Three Gables*. "I don't think we shall find him in the directory," Holmes says of a businessman whose card provided no contact information. "Honest business men don't conceal their place of business."

Good information shops, the ones that recognize what is missing and then turn that knowledge into insight, are all too rare. The process of elimination, using negative evidence, can be powerful but many researchers forget it in their search for positive information.

Also rare are the researchers who understand the capabilities of Archimedean thinking – the ability to use one set of facts or queries to substitute in information searches for other facts which remain hidden or undiscovered. Commonly employed by better researchers today, it was used during WW II when Allied forces wanted to know how effective air attacks on the French railway system were, but lost too many spies looking at the tracks. Instead they used the prices of transported goods as a guide – if those prices went up, the railways were effectively interdicted, when prices went down freight trains were getting through. In much the same way researchers can keep an eye on employment by a competitor, and whether a new shift is being added in a factory, by counting the cars in the parking lot even when they cannot access the actual hiring figures.

Those who can access and assess reliable knowledge – those who have the tradecraft to operate in the Information Age and Post-Information Age – will thrive and prosper. They will be able to, at the least, detect when someone if trying to fool, manipulate, or direct them. They can make informed choices rather than be socially engineered by the information and the hidden human factors that frame a medium and a message.

Those who fail to understand and properly use the tools, techniques, and the massive amounts of available data will drown in the information tidal wave or die of the thirst for knowledge in what seems like an information desert.

OSR is the gateway to other cyber activities and digital activities. It promises to become a major employment field of the next three decades.

Yes, OSR work can be tedious and even wearisome at stretches but it is exciting to those who weather their way to the end.

OSR uses the best lessons, tools, and the methods that have proved useful in many major research fields.

OSR's steps often combine largely-secret methods and break-through technology with simple, well-known techniques. OSR incorporates and combines the skills, techniques, and procedures learned from many different, but technique-related, research areas. It improves and optimizes knowledge acquisition.

OSR serves more than one master. It is based on the techniques of many. Primary contributors to OSR are the fields of Computer-assisted Research (CaR) in Journalism, Open Source Intelligence (OSINT) in Government, Information Literacy (IL) in Library Science, Opposition Research (OPPO) in Political Science, and Competitor/Competition Intelligence (CI) in Business.

These fields all seek highly-specific, usable, and timely information about someone or something – whether it is a person, company, group, enterprise, activity, or nation. The sources of that information may be people such as experts, leaders, followers, employees, or suppliers. The collected information often leads to actionable ideas.

But the methods of these various research areas are not well-integrated. Moreover, all the individual fields are seriously deficient in ways. They share common threads in their research techniques but they are not congruent.

Each field has strong points, weak points, and some areas that are total blanks. OSINT, for instance, tries to protect sources and methods so thoroughly that the security issues often hinder it in actual use. OSINT's regulations may hinder meaningful and practical social media exploitation; IL is extremely light on security. Some of the information fields rely almost exclusively on execution scripts and automatically-collected databases of others rather than on humans and their thoughtful search strategies.

However, when you concatenate the techniques and practices of all these areas, choose the best parts of each, recognize the weaknesses, fill in blank spaces, and integrate the best methods and capabilities of them all, you come up with bleeding-edge knowledge about current open source research methods and techniques.

The strengths of one field fill in the weaknesses and faults found in other leading research areas. All of the fields hold major pieces of this Information Age puzzle. At the same time all the fields lack pieces and ideas that other areas have.

The more-complete research area known as OSR is in fact a constantly growing, transformational, and integrated system of systems. The indispensable eyes and ears of many fields, OSR techniques offer the advantage that only information research can provide – the knowledge advantage.

OSR takes in the wide picture. It gathers others' knowledge into one place. It tries to integrate all the puzzle pieces so that you, as a knowledge consumer, can go where you need to go, get the information you need when you need it, analyze it for its

usefulness, and apply it to the things in your life, your concerns, and your business. OSR is driven by data and analytics. It is an immediate, essential, urgent, critical, and vital need in today's world. It can provide answers – and occasionally pose new questions – rapidly, accurately, and in a wide scope.

Today's information needs have become a voracious vortex. People require more information today than yesterday; tomorrow they will need more than they did today.

Finding and using useful publicly-available information has become an essential life skill in the Information Age; it is a skill set that flows into every aspect of life. But is that a skill that people and institutions have or use effectively? A few do; many do not.

Not everyone does competent research.

OSR is a means to an end. It is not the end itself. There are steps.

Important parts of OSR include planning, information gathering, management, analysis/assessment, and production. Executing on the knowledge advantage you gained is essential, but is not part of the OSR schema unless you are researching for your own needs.

Some components of OSR are well-known; other parts remain obscure.

Competency requires the knowledge of, and ability to use, a wide variety of systems that apply to the user's needs. Integrated multi-dimensional research – not just popping a word or two into a search engine – produces results rather than failure.

That is not to say results will appear magically. Sometimes all you can do is trudge forward, slogging through the miasma of fact, misinformation, and off-point data that the Internet offers. There are times when open source research seems too much like scutwork

and drudgery. Researchers seldom find themselves boiling along; inevitably it is a slow grind to get the information needed. The speed of any search is controlled by boundless energy and an unwillingness to waste any of that most precious commodity, time. Making progress relentlessly, even if at a crawl, is a principle guiding OSR.

Many workable and surprisingly effective research options and avenues are widely available. Some will work well for you; some miss your unique needs today. But some of the options you don't need today will be required next week. Cocktail lounges stock a full bar, including some ingredients their bartenders may only use once or twice a year. And while bartenders are adept at mixing the common drinks, they also have a manual that tells them how to mix what they need to craft when faced with unusual customer requests. OSR works much the same way.

OSR is about knowing and having access to the options; operators need familiarity with every research area and all types of techniques in order to mix and match them to the ever-changing requirements. Only with that how-to knowledge can OSR users choose wisely from the research smorgasbord. At the same time, just as at a smorgasbord, seldom will anyone need everything available on the resource table. Researchers learn to pick and choose at the information table, just as at a restaurant.

The "Fields" of OSR

Different contributing fields offer a variety of "flavors" to OSR.

Information literacy (IL)) is strong on storage and retrieval of information, particularly from databases and anything found in fiber-based materials. Knowledge of IL helps a researcher in filing and retrieving the information collected elsewhere. IL is also particularly good when it comes to using online databases and non-digital materials. It is not particularly strong in dealing with some other research techniques or security.

Opposition research (OPPO) has proven its strengths through many election cycles. Its techniques are wide-ranging, but they focus largely on individuals. That laser-focus, and the fact that the most effective OPPO techniques tend to be close-held, mean some of this field's special techniques remain largely unknown and often unused by other information professionals.

Competitor intelligence, or competition intelligence, (CI) are widely used in the business community to provide the facts and background needed by many corporate decision-makers. Competitor intelligence focuses solely on competitors, their activities or plans. Competition intelligence adopts a wider view. It takes into account any aspect that affects the field and all business opportunities, anything that provides insight to business leaders. Another way to look at this field is to assume competitive intelligence is focused on specific targets and there is another term, market intelligence, which focuses on an industry or market segment. CI has a long history of use and success but, again, many of the techniques and resources needed for this work remain close-held.

Computer-assisted Research (CaR) is an extension of journalistic techniques. It has great strength in gathering information from social media and excels in webscraping, analysis, and interpretation. CaR, like many other research fields, tends to be light on security and safety.

Open Source Intelligence (OSINT), is used by intelligence, military and police organizations. It features extensive penetration of the Web, use of specialized databases, and is heavier on security than virtually any other research type.

Many people, in the past, associated OSINT with "spying." It had a theatrical aura, one of mystery and even hushed-up notoriety that made many people wary of it and its practitioners.

Worth noting, this term is increasingly being used to refer to almost any type of open source research, not just the investigations in the intelligence field. The term "OSINT" is becoming more popular in the commercial world, especially in the cyber field. There are now multiple commercial cyber security certifications focused on OSINT and its applications. There is also a strong and growing "OSINT" community of hobbyists, security professionals, and world-event followers. The adoption of the term by the cyber security community has also created many open source tools and techniques to help enable new forms of data collection and analysis.

When used in the newer meaning, OSINT is synonymous with OSR.

When OSINT is used in its more restrictive sense – the research conducted by intelligence, military and police organizations – the emphasis on security is positive for the safety of the researcher. But the rigidity of OSINT's numerous – and often-draconian security rules – can interfere with the research process and severely limit results.

Other fields are just starting, in ways that are large and small, to use the ever-expanding methods of research that OSR offers.

What OSR Does

OSR combines the best of the five contributing fields to balance the problems of each with solutions that the others provide. OSR offers an information advantage to those who master the integrated techniques and the tools it uses.

Business, schools, universities, libraries, political organizations, the media, and security organizations world-wide are investing in the equipment and resources needed for information exploitation, under whatever name. Information shops combine and balance the

best technology, processes, and people in their quest for information.

All fields, all students, all people, badly need to acquire the talents required to do the research.

Many fields are putting a foot in the water; some have already advanced to swimming. Some use outsourced services to fill their information needs; others employ their own dedicated personnel and equipment. All are looking for OSR-capable people.

The need is already there for people knowledgeable in areas such as:

- Open Web information
- Information gathering
- Analysis and interpretation
- People, equipment, and programs
- Location-based research
- Social media monitoring and analysis
- Deep Web and Dark Web capabilities
- Database use and Big Data capabilities
- Use of webscrapers and offline browsers
- Multi-media acquisition and video/visuals analytics
- Visualization tools and instruments
- Cyber-security and red-teaming.

The explosion in social networks, groups, forums, multi-media resources, and user-generated sites – and the plethora of material these often provide – also drives the need for capable researchers.

"Capable" is the key word. Some people think of themselves as information gurus because they know how to put a word or two in a website's search box. That's it. And that's magical thinking! They believe they have mastered information searches. In reality, such people are in the pre-kindergarten of the OSR discipline.

Others try OSR but quickly become disgusted with information searching. They find that putting a few words into a search engine just doesn't do it. They quit trying when the reality of their search efforts doesn't match the information promise they anticipated. The reality of the search simply doesn't match the aspiration level when you don't know the tradecraft to get you there. There is more, much more, to obtaining and using quality data.

Very few people have been taught the best ways to explore their information environment. Only a handful know how to get what they need by using integrated information tactics. Fewer still know how important it is to learn and map the structure of key sites, explore subdomains, and find or copy the important folders.

Professionals in the information business learn these, and other, *minutiae.*

They often find the fine details on the Internet. They learn by exploring the filters and quirks of the key sites they use consistently. The pros know how their key sites operate, and how to use them to their best advantage.

Such knowledge and ability is not developed in an hour or two, or even in a month or two. It doesn't come from reading a dozen paragraphs in this, or any, book.

Many of the tools and resources, often the best ones, are anything but intuitive. Search-box-type simplicity may be found on search engines, but it is rare with other tools and techniques. The ways to use advanced tools and techniques – even those available on search engines – have to be researched, learned, and practiced. A fan in the stands understand what a baseball bat is used for and the basics. It takes a professional touch to swing a home run out of the many bits and pieces of knowledge that go into a professional understanding of the bat's use. OSR professionals need to develop that same level of

understanding about their tools and techniques, and they need to practice their swing before coming up to the plate.

Learning how to achieve in the Information Age is seldom simple, not when you consider that the Internet and computers are overlaid by an extensive matrix of other, older, information resources which remain both usable and useful. Nor is it made any easier by the determination of many who know how to harvest information to keep their knowledge of that capability to themselves. There is often a demonstrated reluctance by professionals to share even the simplest techniques that are needed to exploit the wide world of information. Competition is unwelcome to some.

OSR Resources

Web search engines, rightly or wrongly, are believed by many to be the 21st Century successor to Samuel Johnson's library. While the search engine is one information source, it is not the only – nor is it always the best – resource. Some useful resources are found on the Internet; many are not. Research is always a work in progress; new techniques and fresh resources appear all the time. Older resources still prove valuable, as well.

Important information resources include products of academia, from courseware to research papers. Other resources include commercial and public information services that provide news and special reports. Some of these resources are expensive but they are often available cost-free to library patrons. Groups and individuals produce everything from leaflets and graffiti to letters and posters; getting on mailing or emailing lists helps in collecting information for medium- and long-term projects. Email information can sometimes be useful. Social media postings and even overheard street-corner discussions can qualify as open source information resources.

The resources for information gathering are almost inexhaustible for those who want to mine the knowledge trove, who understand how to use the instruments of OSR.

OSR Phases

"Common knowledge" holds that there are three distinct phases of OSR research:

- Acquire or "find"
- Process
- Distribute.

That is the short view: There are always collection, processing and presentation steps. OSR, in reality, is more complicated; what seem like three simple steps are actually composed of many parts:

- Collect
- Secure and preserve information
- Process
- Verify and validate
- Analyze
- Interpret
- Review and polish
- Publish.

The SOP you create while reading these pages takes all the phases into account.

Although all stages of the process are important, the researcher or research desk must first collect the information before it can be turned into the finished product, turned into insight. Until information has been collected no other phase exists.

Seldom do you know with any certainty what will, or won't be, valuable. For that reason expert researchers save a lot of material, even things that don't appear useful at first. They know that things which initially seem to be peripheral may become important later.

The rule is collect widely early, winnowing the information during the analysis and report writing phases.

Asked about what to do with a particular piece of information in *The Adventure of the Six Napoleons*, Holmes answered in the way every good researcher would: "To remember it – to docket it. We may come on something later which will bear upon it." The Holmes of literature understood that solutions are based on a mélange of ideas and facts, not one only, and that all must be put in a safe place for possible later use.

Researchers need to designate a repository such as **Hunchly ($$$$, Web capture):** https://hunch.ly/ where they can squirrel away all the bits and pieces of information they collect on the way to developing an insightful report.

As you build your personal or team SOP, remember that most research plans include:

- Preliminary search strategy
- Research plan
- Security plan
- Research tools selection
- Assignment of research tasks
- Research phase
    - Documentation and research preservation
    - Revision of research plan
    - Continued research
- Validation of information and sources
- Analysis
- Vetting and drafting of the product
    - Finalization of the product
    - Distribution of the product.

Physical Assets

The physical assets needed for a modern OSR project can range from a laptop computer connected to the Internet and a library card

to specialized servers and use of the pricey, but commercially available, software that some intelligence agencies use.

Arguably the most important assets are the human senses. OSR has been carried out through the millennia with nothing more than the senses. (In Occupied France during WW II, Allied agents were taught to take note of the smell of coffee brewing as they walked down the street; that normally indicated they were outside a building housing high-level members of the Nazi Occupation.)

Little things mean a lot in OSR. Windshield wipers are a desirable feature on cars, but most Moscow cars didn't have them at the height of the Cold War. Soviet surveillance vehicles were well-equipped with them, however, and CIA agents routinely checked the traffic around them for cars with windshield wipers to see if they were being followed.

See, smell, hear – the senses are the OSR sensors.

Predictive, Current, Reactionary?

Research can be predictive, current, or reactionary. OSR may try to see what is ahead (predictive), know the present situation (current), or understand what has passed (reactionary).

The tools and techniques of greatest use to you will depend on what type of research you are doing. You must look over the collection of tools and techniques, master those that are most useful to you, and include them in your personal SOP. No two Christmas trees look precisely alike though each may be beautiful; no two search plans should be the same, either.

The OSR craft constantly changes – sometimes borrowing things from others, at other times developing useful new techniques. You need to do the same. People in the infoshop practice a craft that is working its way toward, but has yet to become, being an industry.

Successful researchers develop, over time, their own matrix of skills, assets, resources, and techniques. These are similar to those used by investigative journalists, reference librarians, post-graduate scholars, business leaders, intelligence operators, and law enforcement officers. Learning how "these people" do their job – through books, manuals, information posted on the Internet, and personal contacts – are among the ways to constantly improve your qualifications and achieve better research results. But books on how to be a private detective or reporter, while they may provide useful ancillary tips, do not make anyone a successful information collector.

Moreover, no one starts the first chapter of any book as a novice and emerges an expert in the field by the final chapter. Information professionals grow into the field; they are not born into it.

Mastery of the techniques and tradecraft comes automatically to few, if any. Most people find help and guidance more useful than learning by mistakes, when blasphemies abound. This book tries to guide you around some of the common missteps. But in the final analysis success using the tactics of OSR does not come from this book or any other resource. Success derives from practice, from the experience gained by selectively applying resources and techniques to specific situations.

If there is one piece of advice that should be kept in mind, it is this: Use your experience and your gut feelings when responding to the challenges. Avoid cookie-cutter approaches. Allow intuition to guide you. Give it free reign. Think outside the box. Train the brain to defy the orthodox. Conservative, by-the-book thinking (and searching) is often far from the ideal in OSR. Control the process; don't be controlled by it. No research system works flawlessly all the time. Use your intuition to find those pathways to success that neither this nor any other book or class can provide.

Good results are far more important than slavish adherence to any checklist. The spirit of initiative should never be squelched. User drive, brainstorming, time-outs for thinking, and even some unconventionality are often as valuable as any thought-through search script. The success of all research rests primarily with the individuals doing the work.

Your SOP book will be unique to you, designed specifically for your needs and purposes. Good personal notes on what worked for you, what didn't, and what might work better, become your best manual. These will give you the edge.

You will get out of this effort what you put into it! Nothing less; nothing more.

***

Project:
If you haven't read the Preface – and many of us skip that part of a book – go back and review it before moving forward. It contains important information.

Project:
In the overview section of your SOP define the current audiences you are now serving, including yourself and family, and the types of information you are likely to need to provide. Also select an audience you are not currently serving, but would like to work with, and the types of information you would likely need to provide that audience.

Project:
In the digital library section you recently made, search the Internet and then download these book files:

- NATO Open Source Intelligence Handbook.pdf
- NATO Open Source Intelligence NATO Reader.pdf
- Intelligence Exploitation of the Internet.pdf
- Untangling the Web.pdf
- Desktop OSINT Handbook.pdf
- Exploring Social Media Web Sites.pdf

- 2oolkit on the Go.pdf
- Army OSINT Manual/US Army ATP 2.22-9.pdf

Save these documents to the digital library section of the SOP you have created on your computer. Review them and make note in your SOP of any helpful material. This project will come up again.

14. Tradecraft of Knowledge

Modern information collectors, no matter who or where they are, rely on tradecraft, both on and off the Internet. Tradecraft helps meet the many challenges facing both individual researchers and information shops. Information voyages require an understanding of tradecraft to navigate the shoals and sandbars of the Internet.

Save, Save, Save…

Backup and save what you are doing – probably on every half hour or hour. Autosave and automatic backup on computers are wonderful and should be used, but never rely on them. Save what you are doing to least two locations – one site on the computer and one off the computer, such as a thumb drive.

Losing hours – even days – of work because of some computer glitch and having no way of retrieving the work will make anyone cry. A "save" to a couple of locations takes seconds, seconds that can save hours or even weeks of work.

Command Line

Knowledge of the “command line” is essential for advanced OSR work. Yes, the average Windows user gets along quite well without knowing how to use the command line. But average users seldom do efficient and effective research.

Ant failure to familiarize yourself with the command line can lead to information impotence.

Command line operations are now an advanced method that may even be required in some research activities such as Webscraping. Before the age of Windows the command line was the only way to

use personal computers. The command like is the stick shift of OSR, overtaken by new techniques but still critical in some cases.

Windows usually requires a mouse or touch pad to move around the screen, select sites or programs, and initiate the many operations. The command line is an older way of using a computer, typing in commands at screen prompts. This method is a holdover from the MS-DOS (Microsoft Disk Operating System).

Windows still has a command prompt feature, allowing users who understand the process to simulate the MS-DOS process. Some even refer the Windows command line version as the "DOS prompt" although, on Windows, users are not actually using the DOS system. A similar method in Apple computers is termed "terminal commands."

In Windows the command line prompt can be reached from the Apps Screen, the Main Screen, or the Start Menu depending on the program version. To use the system valid commands, as well as optional parameters, must be typed into the command prompt.

Many useful commands can be initiated from the prompt. The method may be used for running programs, to start scripts or batch files, begin administrative actions, or solve a variety of computer issues. Some OSR programs, apps, and resources only operate from the command line. Users who don't know, or learn how to use, the command line are behind from the start in OSR operations.

As with everything on a computer, the commands must be typed exactly. Misspellings or improper syntax vitiate everything. The computer cannot recognize or correct mistakes of an operator.

Good information about use of the command line can be found on the Internet, and in books.

Hidden Elements and Metadata

There are those who maintain that OSR is all about "documents." If you include "documentation" within that meaning it becomes a persuasive argument. Open Source Research is about searching for information, but more importantly information that documents and proves the information that is collected.

Often the documentation you need is contained – sometimes apparent, often hidden – in the documents that an OSR operator uncovers,

Documents, and the sites they are on, contain hidden – and widely unknown – elements that can often be exploited by knowledgeable researchers. Those hidden elements may relate to accounts, material that was removed, the computer, or server. Or there may be highly-personal information. While such "secret" elements can often be removed, even people who know how to do so seldom follow through. Hidden data may include a wide variety of information such as:

- Metadata
- Various document versions
- Hidden text, objects, slides, rows, columns or worksheets
- Comments
- Tracked changes
- Document reviewers' identities
- Routing information
- Presentation notes
- Server or computer information
- Custom XML Data.

Researchers who know the details of how a program or piece of technology works, and the hidden elements that such programs bring to light, have the ability and knowledge to exploit those secretive elements.

FOCA, found at https://www.elevenpaths.com/es/labstools/foca-2/, stands for Fingerprinting Organizations with Collected Archives. It helps a researcher unlock many types of documents and find the hidden elements that aid in analysis of Web pages, word processing and PDF files. FOCA also makes metadata visible. It can be considered the black light of the OSR field, making thing stand out when they are otherwise seemingly invisible.

Too many online users ignore metadata. Sure, it's widely collected by Internet providers and professionals but it can't be all that important, can it?

Officially, metadata classifies material: It is this, but not this and was produced by this person, not that person. Metadata is often called "information about information." That's the techie version; another version that puts a completely different light on metadata is "information about somebody and their computer activity."

Metadata may be created during collection – such as the date that something was downloaded – or it may be hidden within whatever was collected.

Metadata can be powerful. It narrows and identifies; it is often a peek over the transom, showing how something was produced or by whom. Metadata can simplify analysis and interpretation. While it is not always important, it can be. To squeeze the most information out of any document or site it is crucial to look for and carefully study all the available metadata.

Metadata is often attached to online and digital materials including pictures, word processing files, and web pages. Metadata types will differ, depending on the file type. Metadata on a word processing file may tell who wrote the item and when, the file size, who approved it, and a good deal more background. The background information – metadata – on a picture may tell when and where the

picture was taken, what kind of camera was used, or who owns the publication rights. On web pages metatags may provide key words and a description of the material. And the metadata found on any file type often provide dates and times.

Metadata is usually hidden inside a file; it often has to be ferreted out. The metadata search procedure varies from file type to file type, and depends on the particular program you are using on your computer. Researchers can use the Internet to learn the specific procedures to search for common types of metadata files.

While the research community generally accepts that metadata is fair game for discovery and use, some still; argue against that contention. Although governments and their minions say that metadata doesn't matter and collecting it is not intrusive, others say it can be too revealing when the pieces are put together.

A basic view of the metadata of a document or picture may be as simple as right-clicking on the object and then a click on the "properties." While some basic metadata is often available this way there are tools that may provide additional information, including the identity of the creator and other facts that can be effectively used by researchers.

Common metadata information includes formats, dates, processes or equipment used, and update information.

Webpages often contain many pieces of metadata. The metadata in web pages, including meta tags such as <description> and <keywords>, are worth examining. In fact, a truly-complete search requires a researcher or research desk to examine the source code of some web pages to see if there are useful tidbits of knowledge there that are hidden when looking at the page as it shows on the screen.

For webpages, the so-called Robots.txt file serves as a major meta-tag. This meta-tag on a site tells the search engine spider what not to look at, index, follow, or archive about the page. "Go on, there's nothing to see here," is another way of thinking about that tag. Researchers can't know whether there really is nothing of value on the page until they look. But there is always a question why the system operator doesn't want other people to look at the page and its contents.

No law mandates that a search engine's spider must follow the orders of the Robot.txt file. Many spiders do; some do not. However, that doesn't mean the information in the forbidden pages is unavailable to knowledgeable searchers. There are ways around a Robots.txt file, but you have to know the prohibition exists before you can start looking for what system operators forbid you to see. The way to find and defeat a robots.txt file is covered in Chapter 10.

There are other hidden things on a web page and some tags are meant to direct, or keep out, search engines.

Tags of all kinds are usually invisible to the Web user but they help guide search engines – and researchers – to the subjects they seek. They also assist in identifying a page's content. Located at the start of an html document – in the <head> section – meta tags describe what the website operator says are the important parts of the page contents. Expert researchers use meta tags to hone their searches, homing in on needed material.

For the researcher, a handful of Web tags provide the most value. To view the meta tags, right click on the page and, from a drop-down menu select "view page source." The meta tags will often be found on the first parts of the page; they will be denoted by the opener, <tag name> and the closer, </tag name>.

Search engines often use the **Title Tag** to fill in the search result that users get. It is what the site owner titled the page – in theory what the page is about. Page owners writing about rust who wanted to make information about that subject less accessible the page developer could use a Title Tag such as "kitchen knives" and that might throw off the search. When the Title Tag is unclear, if the words of the page title do not accurately describe the contents, or if the tags lack context, search engines have no way of knowing.

The **Description Tag** is often used by search engines to fill in the descriptive part of the search result – in a few words it expands the title and may serve as the snippet beneath the title when the search result is displayed.

The **Robots Tag** serves the same role as a Robots.txt message – saying to the search engine "you're not allowed to go there." While a robots.txt notice puts an entire site or sections of the site off limits, the Robots meta tag forbids the indexing of a particular page or following any links on that page. The **No Follow Tag** is used to dissuade search engines from following links on the page – links that may be internal or situated elsewhere on the Web. Since following links is a very important part of the OSR process it is important to know when there is a link on the page that the system operator wants to keep secret.

Alt text is a meta tag used in image optimization. Images have become increasingly important over the years but a search engine text file has no way to display those audio-visuals. This meta tag shows that there are images on the page and describes them. It is a textual supplement to a picture or other video content.

Metadata is a complex subject, one that too few professional researchers pay sufficient attention to. While it seems to many to be a niche area, the study of this subject separates true professionals from amateurs.

An **Understanding Metadata Primer** pdf can be found at https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata. **A Complete Guide to Meta Tags** is available at https://searchenginewatch.com/2016/10/05/complete-guide-to-meta-tags-in-seo/.

Tools that will read a variety of metadata types include:

- **ExifTool:** https://www.sno.phy.queensu.ca/~phil/exiftool/
- **GetMetadata:** https://www.get-metadata.com/.

Another type of hidden data is subdomains. These are parts of the domain, sectioned off from the main domain name. The subdomain generally has its name listed before the actual domain name and is separated by a period as in subdomain.domain.com.

Subdomains are one of the ways site managers organize their pages.

They are often used to create testing versions of a website, private pages, online e-commerce stores, and sites for mobile use (the latter may be designated by an "m." before the domain name).

Discovery of subdomains can be difficult, but this is an integral part of an effective and complete information search. The search for, and exploitation of, subdomains is usually a "techie" project as many subdomain search engines are written in python computer language. Users planning subdomain searches must have more than a passing familiarity with use of the command line. The search can be made through sites and services such as:

- **Amass:** https://github.com/OWASP/Amass
- **Knock:** https://github.com/guelfoweb/knock
- **SubBrute:** https://github.com/TheRook/subbrute
- **Sublist3r:** https://github.com/aboul3la/Sublist3r
- **SurfaceBrowser:** https://securitytrails.com/.

Source Code

There is a lot to a web page that few users ever see. Information may be there, hidden in the source code that never displays on the visible page.

To call any research complete, the source code of key pages should be searched – a taxing and often-enervating job. Not only does the researcher or research desk have to know html – how a web page is constructed – but most of today's web pages are made up of "spaghetti code," long lines of text and tags that have to be reviewed and their purpose understood.

Looking at source code is comparable to panning for gold: Virtually all of what you see is as valuable as sand and gravel but occasionally there is a useful nugget of information. Source code examination may reveal information not visible on the page presented to the public. While the displayed page may be barren of information, the source code may include email addresses, phone numbers, or some other content that will otherwise go unseen.

Source code is easier to view in some browsers than in others. In Firefox the Alt key brings up the menu bar. On the bar select "Tools," then "Web Developer" and then click "Page Source" on the menu. For other browsers use the Internet and look up the name of the browser with a query such as "source code on (browser name)." To view the source code when you have no other way do so, go to the **GenerateIt** site at https://www.generateit.net/seo-tools/source-viewer/ and enter the URL of the page you want to examine. Copy the source code text that is displayed and paste it in a word processing page; then go through it line by line.

Some pages contain externally linked CSS files. Such pages will not actually have the file, but the file will be linked and the page will have a path to get to the external file. You may have to use the Internet to learn how to follow CSS links.

Reviewing a page's source code is usually boring, but like panning for gold it may eventually pay off. Resources of help in reviewing Source Code include:

- **NerdyData (Technology source code):** https://nerdydata.com/search
- **The Silver Searcher (Source code search engine):** https://github.com/ggreer/the_silver_searcher.

Freedom of Information and Sunshine Laws

Governments write, collect, and archive scads of information and documents. The general rule in the US is that these records belong to the public, not to the officials, functionaries, or the people who drafted the records. In some cases just going to a public agency and asking nicely – remember, there is almost always a person involved in getting the information – will get you the information or the records. But you cannot count on a smile, "please," and nice words.

The US federal Freedom of Information Act (FOIA) applies to "records" of most federal agencies under the aegis of the President. The courts and Congress are exempt. At the state level there are similar public records laws, sometimes termed "sunshine laws" or "open records acts." The details of these state and municipal laws vary widely and involve different procedures, but they allow access to official records that can be invaluable.

Journalists in the United States rely heavily on the FOIA and similar public records laws. Because of the complexity of these laws the FOIA open record techniques are used primarily by reporters and lawyers. But if the information is available under the FOIA or public records acts anyone, not only reporters, can demand it. Knowing the proper verbiage and the techniques to demand documents – and how to outwit the bureaucrats trying to outwit the records requesters – becomes important.

Entire books are written on the subject; there are a number of online resources that explain how to frame an FOIA demand and

what to expect. (Expect little if you don't follow up.) Anyone in the information business, not just reporters, needs to have at least a passing acquaintance with the FOIA and state open records procedures.

In general, federal FOIA requests must be in writing and follow the regulations of the particular agency. The law says the request must "reasonably describe" the record or records, an argument that is frequently used by some reluctant agencies to deny access. Having exact titles, locations, and dates is helpful.

Because sunshine laws are written in such a way that they exempt information from disclosure when government employees might not be able to locate the documents, knowledgeable researchers often go into great detail about the records needed and even suggest physical locations and/or systems where the documents might be located.

The law applies only to documents that already exist. An agency cannot be required to compile or create documents in order to fulfill a request.

Hope for, but never expect, quick results. Normally a federal agency is given 20 business days (or 10 calendar days for an expedited request) to either provide the information or explain why it is withholding the data. However, there is an "unusual circumstances" exception that some agencies seem inclined to use on a more or less permanent basis. Expect all requests to take longer than the statutory time limit. An acknowledgment by the agency that they have received your request does not qualify as fulfillment under the law, though some will try to use this excuse for not meeting the specified time limit.

To speed up the process some information professionals request expedited processing, but not everyone or every request is entitled to that. Researchers who are going to call for federal expedited

processing must be familiar with the conditions under which it can be claimed; a detailed statement about those conditions will be required.

Federal agencies are allowed to charge "reasonable" fees for direct costs such as searching and copying the record, although reductions and waivers can be requested by researchers. It is always wise to indicate the amount you're willing to pay as part of your request letter, then ask the agency to contact you if the search and copying fees exceed that amount.

Federal law doesn't require a press pass to get the reduced fees that are available to reporters. The federal definition of news media is wide enough that many information seekers can use it to cut their costs. In some cases researchers who gather information of interest to some part of the public and use their skills to turn the material into a work that is distributed to an audience may be considered part of the "news media." That claim can reduce the costs. There is also a “public interest” fee waiver available.

When submitting any FOIA request it is essential to know and understand the rules, often arcane, of the particular agency you are querying. Different agencies require different submission methods, but when using whatever method is required be certain to clearly label your communication as an FOIA request. Make certain there is no question about what it is.

It is quite possible, even probable, that the agency will claim some exemption, in which case you will have to do further investigation to see if the exemption claim is valid. Some are; many are not.

Obsessive secrecy is common among bureaucracies. Claims of exemptions run the gamut. The agency may claim the information is classified. The agency may say the information or document is exempt because of internal personnel rules or practices, or is exempted by some statute. It is not unheard of for an agency to

claim there is a trade or commercial secret involved or that some attorney or deliberation privilege is involved. Claims of exemption because of personal privacy are not uncommon. Some agency personnel are experts at twisting the language of legal exemption beyond all reason or recognition. Always check, and if appropriate challenge, any exemption claim. Exemption claims are often tortured readings, or even misstatements, of the law

Getting a document may only be the first step. Sometimes the document is so heavily redacted – parts are blacked out – that it is impossible to make sense of, or even read coherently.

Expect, and prepare, to have to submit an administrative appeal regarding what you were, or were not, provided. Save every scrap of information related to the FOIA request, including envelopes. Make certain the appeal is filed within agency timelines and falls within the reasons allowed to contest the initial decision. State and local open records process may be even more convoluted than the federal one.

There are detailed resources and guides available on the Internet to help researchers understand the rules and how to submit a properly formatted request and appeal. Some of them are at **Reporters Committee for Freedom of the Press,** under the Legal Resources menu, at https://www.rcfp.org/.

The FOIA Request Generator aids users in creating their FOIA demands at http://www.refp.org/foi_letter/generate.php.

Dataleak Sites

While OSR tries to catch publicly available information that is insufficiently protected or mistakenly made public, there is part of the information universe devoted to "outing" anything that others would like to keep secret. These Dataleak Sites are often clandestine for at least part of their activities. Some open source advocates may, in fact, believe that all "information should be free"

and feel that nothing should be secret. Other people insist some Dataleak sites are associated with groups or organizations – possibly even foreign intelligence agencies – whose intentions are to make someone look bad or to force someone to reveal sources and methods.

Open Source investigators – usually those dealing with issues at the national and international level – may want to explore such sites to see whether and how they may fit into their search plan and if they should be routinely searched as part of the SOP. Workers for, or employees at, some government agencies will be legally prohibited from even accessing such sites, however, and researchers should determine if they fall under the exclusion rules before going to any Dataleak sites.

These sites include:

- **Al Jazeera Investigations (Arabic region):** https://www.aljazeera.com/investigations/
- **BalkanLeaks:** https://balkanleaks.eu/
- **Cryptome:** https://cryptome.org/
- **GlobalLeaks (Whistleblower software):** https://www.globaleaks.org/
- **Wikileaks:** https://wikileaks.org/.

Backlinks

Backlinks is the term used to describe the links that other sites make to the page that a user is viewing. Backlinks may be described as a vote of confidence. They can be one indicator of how much trust to place in a site's information – the quality of the backlinks sites is always worth looking into when evaluating authoritativeness.

"Birds of a feather flock together" is a useful phrase to consider when looking at backlinks. If the backlinks are from good sites they are very, very good but when the backlinks go to questionable sites you may want to mark your potential resource as "horrid."

Search engines often use the number of backlinks as a factor in ranking a site but seldom evaluate the quality of the linking site.

Small SEO Tools has a free backlinks checker tool that is available at https://smallseotools.com/backlink-checker/ and there are many for-pay sites, as well.

Who Is and IP Lookups…

Tracking a website owner or operator, whether an individual or an entity, is often a starting point in verification. Identifying the people or organizations who run a site may prove to be important. WhoIs lookup sites were more instructive a few years ago but still remain useful.

Using WhoIs sites effectively today is more challenging. Now many site-sellers provide users – for a price, for a price – with a level of privacy that approaches anonymity. Such site sellers effectively refuse to divulge the ownership information without a court order, thwarting any effort to determine who owns or operates a site. Still, many site owners won't pay their site-sellers the price for the privacy so WhoIs lookups remain useful regarding those sites.

WhoIs sites take the URL you enter and search domain name registries and registrar tables. The information they return can be used for a variety of purposes.

A variation is reverse DNI sites. These will get you the domain name and some information on a site if you have the IP address – sets of numbers from 0 to 255 that are separated by periods. You may be able to use this information to determine the name of the Internet service provider assigned to a particular IP address.

Enter the address of the website into a site like IP Lookup which determines the IP address and shows information such as the host, location, and other WhoIs data regarding the address, or addresses,

entered. A number of sites claim to provide a plethora of data, including meta information, hosting data, and site age. BuiltWith is somewhat different in that it reveals what was used to build a site. Occasionally that can be useful information.

- **BuiltWith:** https://builtwith.com/
- **DomainTools ($$$$):** https://www.domaintools.com/
- **ICANN:** https://whois.icann.org/en
- **IP2Location:** https://www.ip2location.com/
- **IPAdress (Tracing):** https://www.ip-adress.com/ip-address/lookup
- **IP Fingerprints (Location):** ipfingerprints.com
- **IPLookup:** http://ip-lookup.net/domain.php
- **Lookupserver:** http://lookupserver.com/dns
- **Riherds (Reverse DNI lookup):** http://remote.12dt.com/
- **Robtex:** https://www.robtex.com/
- **SiteDossier:** http://www.sitedossier.com/
- **URLSCAN:** https://whois.icann.org/en
- **Whoisology**: https://whoisology.com/
- **WhoisRequest**: http://whoisrequest.com/
- **XL-WhoIs:** https://le-tools.com/XL-Whois.html.

Sam Spade is a downloadable utility freeware package that does IP lookups and many other chores for OSR researchers. Access it at http://www.majorgeeks.com/files/details/sam_spade.html.

Cookie Management

Cookies are small text programs that are downloaded onto the devices that access the Internet. They remember, and may later transmit, something about you – sometimes on behalf of the site you visit and sometimes on behalf of a company or agency that wants to know details about you and your interests. Cookies may be designed to tell system operators about the device, or about the browsing habits of the device user. They are also used as a recognition method so the website's operator can identify users. They may allow you to return to the site repeatedly without logging

in or otherwise serve your needs and interests. They may also track your activities. In short, cookies may be good for you; they may also be dangerous to your research work.

There are different cookie "flavors," although any particular cookie may fall into more than one category.

First-party cookies come from a server or a domain managed by the website publisher. Third-party cookies come from a server or domain other than the website's publisher; this third party processes the collected information and often sells or uses it for its own purposes. Third party cookies can be used to develop an extensive profile of you, your searches, and your techniques. When cookies are installed from a server or domain managed by the website publisher, but the information collected is managed by a third-party, they are not classed as first-party cookies.

Cookies are also classed by the length of time they remain active on the user's computer.

Session cookies sweep up and store data when the user accesses a website. They are usually used to store the information needed to provide whatever service is requested by the user and are active only for that occasion.

Persistent cookies are stored on the device of the user. They can be accessed and managed over whatever period of time the cookie developer decides. The time can be short or the cookie could remain active indefinitely.

There are a number of different roles for cookies. Technical cookies allow the web user to use the site effectively. These cookies focus on the technological needs of the site.

Personalization cookies allow users to access the service, personalizing it by matching the display to the requirements of the user's device.

Analytics cookies allow website operators to track and analyze the behavior of the users.

Advertising cookies control the advertising space that a publisher has included.

Behavioral advertising cookies store and utilize information about the user's activity by monitoring browsing habits and developing a profile based on those habits.

Cookies on your system are neither good nor bad; it is how others use them that determine that. Because they can be used to obtain, store, and use information about an investigator's or a researcher's online activity it is usually best to limit the number and types of cookies on your equipment. However, particularly with technical cookies, doing so might make a site totally or partially unusable. Cookie issues could best be resolved at the time they arise. A major factor in any decision about cookies is whether the researcher or research desk believes the system operator who has access to your cookie information will use them in, or against, your interests. Unfortunately you will seldom know if or when they are being installed on your system.

It is generally a wise practice for researchers to restrict or monitor cookie placement when on a site – and then clean cookies from the system after getting off, but before going to another site. That is also a pain. Few do it, but it should be considered as part of security measures when accessing sites that might be hostile.

It is definitely a good idea to prevent third-party tracking cookies from being installed on your computer or to eliminate them if they do slip through.

Some services will disrupt tracking cookies:

- **Ghostery:** https://www.ghostery.com/
- **NoScript:** https://noscript.net/
- **Privacy Badger:** https://privacy-badger.en.softonic.com/
- **UBlockOrigin:** https://ublock.org/.

The **Privacy Badger** of the Electronic Freedom Foundation at https://privacy-badger.en.softonic.com/ can help solve many cookie problems.

Kicked Off the Page…

"You have used up your free access" – the message makes it clear you cannot see the article or access the information you wanted. And the snippet you glimpsed briefly looked like it might be the exact information you needed!

There are several ways to approach the problem and get the article out of the cookie jar. The article may been denied to you because you have used up your permitted accesses for the month or whatever period of time the site set.

That time period was based on a "cookie" that the site set, sight unseen, in your browser system when you accessed the site before. So…. Use a different browser! Put the URL you previously clicked on into a new browser and see if that brings up what you want.

A second way of handling this situation is to stay on the browser that has the cookie, look on the Internet for instructions such as "how do I remove the cookies from my browser." Follow the instructions and remove the particular cookie set by the site, then try accessing the article again. Most information professionals believe it is best to remove only the one cookie that applies to the site, not all the cookies on your machine. Some other cookies are useful to you and you would need to restock your helpful cookie jar.

If those don't work, try using the Incognito window in Firefox or similar so-called anonymization services found in other browsers.

Some pages, particularly on commercial sites, will appear on the screen briefly, only to be covered within seconds by a form that tells you to subscribe, whitelist the site, turn off your ad blocker, or do something else you don't want to do. But you really, really, want to see what the site has just hidden.

You can dutifully follow orders, of course. For those who don't necessarily want to obey orders from site managers as their first alternative, try saving the page to your desktop. Go back to where you first encountered the URL of the page and click on it again. Quickly, *before* the message that says you cannot look at that page unless you pay or do X or Y, save that page. Quickly, in this case, is usually a matter of split seconds so you must be prepared to do a fast "save." After speedily saving the page, minimize the covered-over online page, and open the copy you have saved to your computer. Often the page you wanted to read will be saved and shown rather than the form that covered the material you originally wanted to look at.

A variation on this is quickly – again the technique requires speed and attention – do a "select all" on the drop down menu, then a "copy." The text should now be in your computer. You only have to figure out a way to display it. The easiest: open an empty text or word processor file and paste the text in it. Inelegant, perhaps, but it works.

In many cases you probably know the title or headline. Copy it into your browser, with quotation marks around it, and hit the "enter" key. There is a chance someone reposted it elsewhere on the Internet and a search engine may have picked up that repost.

If you can see who the author is, search online for that person. If you can locate an email address, write the person and tell him or

her how much you would like to read the article but cannot access it. It often helps to explain why you want to read the piece that they obviously researched so well. Many writers will respond with a copy of the article or requested information.

There are several ways to collect material off pages the system operator doesn't want you to see. You just have to decide what works best, and quickest, in the particular situation.

Failing Links and Dead Pages

A somewhat similar situation is a failed link, called Web Rot by some. When a Web page appears to have disappeared into the ether but the site remains, don't give up. Go to the site's homepage and use any on-site search engine to look for key words. It is all-too-common that page addresses will change but the actual text or page remains somewhere on the site. Or try a search inputting the title, headline, or author into the site's search engine.

If there is no on-site search engine at the location you are searching use the "site:" command to, limit the search to that site and enter a key word or title: site:urlofsite "key word(s) or title."

To access pages from the past, a primary means is **The Wayback Machine**, below. Its search engine can resurrect useful sites and pages that have been taken down, changed, or disappeared. This site has many uses but one is to use it when looking at locations that may have minutes of meetings and similar periodic postings.

To a much more limited extent the cache page results on a search engine may show what a website looked like when a spider indexed it previously.

Also check the wide variety of archives online; it may already be there:

- **Archive-It:** http://www.archive-it.org
- **Archive.is:** http://www.archive.is

- **ArchiveTeam (Archiving projects):** https://www.archiveteam.org/index.php?title=Main_Page
- **Arquivo.pt(Portuguese archive):** http://www.arquivo.pt
- **Icelandic Web Archive:** http://www.vefsafn.is
- **Library of Congress Web Archives:** https://www.loc.gov/programs/web-archiving/archived-web-sites/
- **Perma.cc (Limited to Harvard affiliates):** https://library.harvard.edu/services-tools/permacc
- **UK Web Archive:** https://www.webarchive.org.uk/
- **Czech Republic Web Archive:** http://www.webarchiv.cz
- **TheWayBackMachine:** http://www.archive.org
- **Webrecorder:** http://www.webrecorder.io.

Another option that provides several ways to find, view, and download archive sites is use of the **Firefox Archive Add-on** at https://addons.mozilla.org/en-US/firefox/addon/view-page-archive.

This gives you options to find, view, and download archive pages from sites such as:

- Archive.is
- Baidu Snapshot
- Bing Cache
- Exalead Cache
- Gigablast Cache
- Google Cache
- Megalodon
- Memento Time Travel
- Naver Cache
- Qihoo 360 Search Snapshot
- Sogou Snapshot
- Wayback Machine
- WebCite
- Yandex Cache.

To stay ahead of failing links, develop a habit of saving any of the information that you may need later when you first find it. Download all potentially-useful material to your own project library immediately, whenever you come across it. Snare anything and everything that you judge useful. Do it at the outset of your investigation. It may disappear today or tomorrow; in fact, expect it to disappear before you need to return to it.

Programs and addons that make downloading easier and faster include:

- **Down Them All:** https://www.downthemall.net/
- **Resurrect Pages**: https://addons.mozilla.org/en-US/firefox/addon/resurrect-pages.

Promises Unkept

Some unscrupulous site managers – eager to drive traffic to their location – will try to trick Web search engines to steer you there even though they have nothing for you. A few do it to try to get you to click on links that will download malware to your computer or that will pay them money.

Some of the more common tricks of the trade include adding false keywords to meta-tags – for instance the names of popular stars on pages that have nothing to do with entertainment.

Then there is the "hidden text" scam, where words of the same color as the background are placed in the page text. Search engine spiders don't detect color; they just pick up the text words. Similar is the "tiny text" scam, words written so small that they may appear to the viewer as a straight line, if they can be seen at all.

Of course, the difference between your search and the page contents may not be due to a greedy system operator. The page may have been updated and changed but the search engine spider may not have returned yet to document that change.

Restrictions on Identifying Information Sources

For those researchers able to go beyond simply observing, for those who can ask questions of others and get answers, there is a potential difficulty in identifying information sources. Whether the contact is in person, by phone, through email, or over a social media account, both you and the person you are talking to need to understand the rules and establish them at the start of any conversation.

Unfortunately, while "off the record" or "on background" are terms of art with specific meanings there is enough confusion that – unless both sides explain to each other what they understand by them – misunderstandings are almost inevitable.

When no rules are set down – and any must be agreed to in advance, not as an afterthought – anything said is "on the record." Everything can usually be used as long as the quotes are accurate and the statement is consistent with what the interviewee said.

"Off the record" or "on background" may be rules laid down at the start of an interview by either side. Normally the person asking the questions wants everything to be "on the record," but either party may ask – or demand – that something is "off the record" or "on background" so long as that condition is made clear before the statement at issue is made.

Whenever there are restrictions, both sides must clearly understand what that means and how much is restricted. If both parties agree that part of any discussion is "off the record" or "on background" both parties have to be clear about what part is which.

"On the record" means that whatever is said or done may be directly attributed to someone by name, title, organization, or in any other personally identifiable way. Unless something else is agreed to – in advance – everything is considered "on the record."

"Off the record" means that what a person says can't be attributed to that person – not under any circumstance, not in any way, period, end of discussion.

An "off the record" discussion is often a comment "for your ears only." It is designed to help questioners to the extent that they now know what information they need to be looking for – but should be seeking in some other place, with some other person. The rule with "off the record" backgrounding is that the information must be obtained from another source and attributed only to that source if it to be used.

"Off the record" discussions have potential drawbacks. There are several reasons people may speak "off the record." Some sources may want to provide context. Some are whistleblowers. Others may want to gain the upper hand in internal politics or succeed in infighting over policy or personnel issues. At times "off the record" is used to "float trial balloons"and gauge a reaction before an official announcements is made. When the only person who knows something – and knows no other person has that information – tells an interviewer something "off the record," the technique can be used cynically to prevent information from being used at all

It is possible, but extremely rare, for a public speaker to demand that an audience adhere to an "off the record" rule. That seldom works out well.

"Speaking on background" allows use of the information on whatever terms are agreed to by the source. Usually this prohibits attribution by name, organization or any other identifier. "Industry sources" or "company officials" is often the way speaking on background is denoted. In some cases "speaking on background" information will be unsourced, as in a statement such as "according to anonymous sources."

There is another term called “deep background.” That means different things to different people. Some consider it equivalent to “off the record” while others see it as another way of saying “speaking on background.” Since that phrase is dark-brown muddy at best, the meaning should be clarified before any conversation continues.

Public speakers may use a version of “speaking on background” known as the Chatham House Rule. When the Chatham House Rule is invoked by any speaker, the event (including academic debates, educational lectures, news conferences, political rallies, and public government meetings) is altered in character. The source becomes a confidential one.

The Chatham House Rule provides anonymity to speakers. It is intended to aid open discussion. There is no set penalty for breaking the Chatham House rule, but generally rule-breakers will be ostracized or refused admittance to any subsequent talks of the sponsoring organization. The rule is: “When a meeting, or part thereof, is held under the Chatham House Rule, participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed.”

Created in 1927, the rule was refined in 1992 and 2002. As constituted, it is designed to prevent tying any individual, either directly or indirectly, to anything said. It does not, however, prohibit distribution of information about what was said.

There is a similar case in which an object, such as a small sculpture, may be placed in front of the speaker and as long as comments are made “behind the…” sculpture they are unattributable to a person, or by a time or place.

A somewhat distant "relative" of these is the “embargo.” Usually used in the news business, embargoes are an agreement between

the reporter or media and a news source to withhold publication of the information until some set time and/or date. Embargoes give publications the time to explore the information and evaluate its importance, but they often provide the source the ability to release the same information to several publications simultaneously so that no news provider is favored, nor do any feel left out.

In OSR the researcher's word is binding, whether on embargos or in other aspects. In *The Valley of Fear* Sherlock Holmes set an example of how important it is for researchers to remain true to their word. Chided for not investigating a person whose identity he had agreed not to look into, the detective explained his inaction simply: "Because I always keep faith."

Marketing Resources

Marketing tools often include such things as biographies, press releases, contact information, and identification about personnel, job openings, mission statements, webinars, financial information, white papers, speeches, newsletters, executive presentations, reports, and announcements. Never overlook the possibilities of searching out and using marketing materials whenever you are researching people or organizations. You can use these tools as jump-off points to search other sites, particularly those on social media.

Facial Recognition

Facial recognition uses faces as data. The concept remains unproven – in fact it is known to be failure-prone under many circumstances and instances. And it remains unregulated; users can employ it for any reason or purpose.

At present facial recognition should be considered a lead to be explored, a "maybe" that is worth following up.

Facial recognition combines cameras and AI algorithms to scan the faces of people in crowds, photos, or videos and then match the

digital codes of their facial patterns to those in a database of people whose identities are already known. Facial recognition is one of the biometric identification systems that can remotely identify people by name without their consent, or even knowledge. Paired with the now-ubiquitous surveillance cameras, facial recognition systems allow identification and tracking of any individual in a camera's field. Recently systems have even been programmed to reportedly detect the emotions displayed by the subjects.

Powerful facial recognition software is on the cusp of becoming available to the general public at reasonable prices. There are no restrictions on how it may be used in the circumstances. Everyone, including OSR researchers, are denied any control over whether or how their face is linked to their identity in databases, and how pictures can be shared. There are no opt-out rights. Facial recognition can be, and is, carried out without anyone's permission. Anyone can be photographed anywhere, and if a picture appears on social media the users of facial recognition software can often identify that person. Once identified, open sources can be used to fill in other details.

Widely available to governments, the systems are now deployed worldwide.

National governments have been using various systems for years, trying to spot terrorists, suspects, and others deemed security threats.

The facial recognition capability has now drifted down to local governments and security contractors, making it possible to track and locate people suspected of criminal activity, to identify protestors, or to locate missing people.

Commercial providers also have carved out a role in allowing people to enter a protected area, or even to sign on to a computer system. It is being used to speed up loading of planes and there is

no doubt additional uses will be carved out in the near term. Non-governmental uses are being developed even now, but there is – as yet – no widespread use of facial recognition by commercial or civilian users, or researchers.

Searching for the term "facial recognition" on the Internet, or at **Facefirst ($$$$):** https://wwwfacefirst.com, may provide additional information.

Systems currently in use are improving in sensitivity and accuracy, although the glitches that remain in correctly identifying some racial and gender groups, as well as close relatives, have proven difficult to resolve. Angle-related problems will probably be resolved sooner rather than later but they remain significant difficulties to accuracy.

Translations and Language Use

There are many languages in the world. People who write or speak about your subject often do so in languages other than yours.

Widely-used Web languages include:

- English
- Chinese
- Spanish
- Japanese
- Portuguese
- German
- Arabic
- French
- Russian.

Never limit yourself to your own lingo. Search in the language of the group, organization, or individual you are following. But you can be assured that good searches are generally harder to conduct in a foreign language than in the researcher's native language.

In social media, particularly, but even on news sites, you need to be aware of any of the alternative or non-standard terms, sub-cultural language, or slang that may be commonly used in the area. Localized language or the private *patois* that is common to a profession or group may be misleading, if not un-understandable.

Often you will need to type words in the writing system of the language. The Roman alphabet does not always suffice. For those occasions the **Google Input** tools which can be found at https://www.google.com/inputtools/ may provide the needed keyboard.

To make your research easier, and the result more accurate, when searching in a foreign language develop a list – cheat sheet if you will – of key foreign words, phrases, and symbols. Those are aids that will probably provide good results. Start your list with the words and commands you are likely to see on foreign language search engines or on sites written in that language. Alternatively, or in conjunction with any words list, use image searches to see pages with pictures that appear closest to your needs.

Every language is complex. Just as there are different words for the same thing in English, most languages have multiple words or shadings of meaning for the same concept. Sherlock Holmes, in *The Adventure of the Three Garridebs,* solved a problem by noticing the subtle spelling and language differences that were found in an advertisement and commenting, "Yes, it was bad English but good American."

Different spellings of the same word will have a significant effect on search engine results. Sometimes words which are spelled differently sound similar. And sometimes a word spelled the same way may have two different meanings and even two different pronunciations. Since many languages – and this applies particularly to concepts – don't translate cleanly, it may be difficult to get to the true meaning of a passage.

When you find what appears to be useful information you have to tease meanings, and they may be subtle differences of meaning, from the collected information. The options are:

- Human translations
- Mechanical translations.

Language barriers create the same problems as an unbreakable code – you don't know what the other person is saying, even when that knowledge may be critical. For that reason humans – linguists, interpreters, and translators – can be key support resources for researchers.

There is a difference among the three. Interpreters change spoken language from one tongue to another; translators work with written materials. Linguists work with both spoken language and written materials.

Linguists – or good translators and interpreters – may be helpful in dealing with cultural nuances, slang, and dialects. Keep in mind that when searching for information about a subject in a foreign country or a foreign language you need to search for the word(s) people actually use in their speech – which may not be a simple translation of an English language word.

The gold standard of OSR is, and will continue to be accuracy in everything. Translations require accuracy. Human translations are considered superior to machine ones. Yet even if a researcher has a qualified humans available they can be easily overwhelmed by the amount of work. While some would argue for a full and complete translation of everything, human limitations often make that impractical.

To mitigate the problem researchers have developed four formats or levels of translation. The open source researcher has to choose among them when using human resources. Going from the easiest and fastest, they are:

- Gisting
- Summary
- Extract
- Full translation.

Gisting is designed to show the general meaning of a piece. It is a look-see. Gisting translations are often used to determine whether there is enough information, of high-enough quality, to forward the item on for a more accurate and complete translation.

Summary translation requires the reading of the entire text, and the summarization of the main thrusts. It falls short of an exact translation. No one would suggest it is accurate enough to put quote marks around.

The Extract is a precise translation of part(s) of a document, text, or resource. Extracts often result when Gisting or a Summary reveals the importance of parts of a document that also contains there is information that may be of limited value.

Full Translations are what the words imply – a complete rendering of the document or verbiage. Full translations are time- and labor-intensive, but they may be necessary for complete understanding. Technical reports and position papers are among the document types that usually require a full translation.

While it is always best if the researcher speaks the language like a native – and thinks in that language like a native – or uses a human translator, there are mechanical workarounds when the best isn't available. Linguists always are at a premium, but they are especially scarce when you need one *tout d'suite.* Because of that, there has been a serious move within and outside the open source community to use machine translations on Web materials.

Many consider mechanical translations to be second string solutions but they can be immensely helpful. There are machine language

translations of scores of languages, from Afrikaans and Hawaiian to Yiddish and Zulu.

Mechanical translators may have problems sensing – and properly rendering – things like shades of meaning, idiomatic or slang speech, uncommon words, metaphorical allusions, or informal expressions. But machine language translations may be the only thing available. Make the most of whatever you have and try to reduce the chances for errors.

Use machine translations for topic identification and as an aid in determining whether some passage should be reviewed and revised by a human translator. Generally it is possible to get decent – that is not to say absolutely accurate – translations from a good mechanical translator. Machine translations can usually be used as indicators that something is probably worth a human review. Never use machine translations alone for information that is critical to your work; they should never be considered reliable. Machine translations can provide the "gist" of text, but "gisting" is at best an approximation.

Translation programs can also be used to locate Web information about your subject that are written in foreign languages, pages you might not normally see. First type the English language word you want to search for into the translation system and ask it translate the word into to the language you want. When the word is translated, copy and paste it from the translation page into the regular search page. That will usually bring up pages written in the foreign language. Then use translation programs to translate those pages into English. Check the probability of accuracy by translating from the English language version back to the foreign language and compare the texts. They usually won't be exact duplicates, of course, but if the final version doesn't look close, the translation may be questionable. Remember, too, that some wording – in any language – simply does not translate well.

If you are doing extensive work in a foreign language it is important to choose the correct type of translation engine. When you are going to use machine translators extensively, understand that there are differences in how translation programs work. Dig down. Details matter. Make certain you use the best type of translation engine for the language and your topic.

- **Dictionary-based:** Translates words without correlation to context
- **Example-based:** Translates by analogy
- **Interlingual:** Rules-based translation
- **Statistical:** Uses bilingual corpora.

Babelfish, Systran, and Google Translate are among the most popular translation programs. Some information professionals use two or more translation programs and compare the results. Others use Google Translate for the body text, but when words don't come back correctly those words can be isolated and dumped into Systran. The options are many, and all the options are yours. Popular translation engines include:

- **Babelfish:** http://babelfish.com/
- **Bing/Microsoft Translate:** https://www.bing.com/translator
- **Foreignword.com**: http://foreignword.com/
- **Google Input Tools (Typing in a foreign language):** https://www.google.com/inputtools/try/
- **Google Translate:** https://translate.google.com/
- **Google Translate Add-on:** https://addons.mozilla.org/ en-US/firefox/addon/to-google-translate
- **Online Translate:** http://www.online-translator.com/
- **PROMPT- Online:** http://translation2.paralink.com/
- **Systran:** http://www.systransoft.com/
- **Word2Word:** http://www.word2word.com/.

Generally you either paste the text you want translated into a window of the translation program, click a button and read the result – or you click somewhere and the entire page will be translated.

This year machine translations are better than last year, and that was better than the year previous. Overall, most online translations have above 90 percent accuracy but they vary from language to language and site to site. While significant room for improvement remains, searches made in the language of the area rather than English, as well as use of the correct script, generally improves the result.

Although machine translations should always be considered inexact and in need of human verification, one workaround to be better-assured of mechanical results is to triangulate. Do translations on three separate machine language translation engines, matching up the wording and meaning to find the common denominators among them while discarding the outliers.

Those who use mechanical means for translations of important documents, or any technical papers, should always make certain the output is reviewed by at least one qualified human translator. If important documents or technical papers were originally translated by a human they should be reviewed by a second qualified person. When the outcome of the research depends on what was actually said or written, accuracy is vital. In translations, as with many other things in OSR, "two is one and one is none."

Deception, bias, and incompetence are of particular concern in any translations made during research. Even people doing a translation can intentionally or unintentionally add, delete, modify, or otherwise filter the information. Local hires may not have been fully vetted to determine their ability level or things in their background that may make them prone to deception or bias. Be aware of the potential for deception, bias, and/or incompetence in all translations.

If you suspect deception, bias, or just incompetence, test the accuracy of any translation by giving the same material to one or two additional linguists and compare the results. At a minimum you will know how far to trust the translations you are getting.

Emojis

Emojis – those little expressive symbols that have become popular in messages – are a means of communication. They should never be ignored or forgotten. Modern information professionals need to be aware of emojis whenever they are used, and of their appearance – which may vary depending on the platform they appear on. There are over 2,500 official emojis, which provides a great deal of opportunity to use them to convey a wide variety of thoughts and messages. Besides the official batch there are many additional unofficial ones.

Simple emojis can be used to communicate complex ideas. For example, they have reportedly been used in the past for selling drugs – one type of leaf symbol meaning "marijuana" and a dollar sign symbol meaning "for sale." Crystal meth might be noted as a gemstone; injectable drugs, including heroin, could be shown by a syringe.

Emojis can be used in many ways. As a code, one symbol stands for a word or even a concept completely different from the picture. Emojis can also be used as a cipher, using the symbols to replace an alphabet or other writing system.

The **CodeEmoji** site at https://codemoji.org/#/encrypt does a good job of explaining cipher encryption with emojis. When used in these ways the messages require decoding – they may be simple cipher, or complex enough to require the use of codebreakers. That is beyond this book. Nonetheless, when sent in the open, and even if they require decoding, emojis are fair game for open source work and should be treated as any page or a message needing translation.

Researchers must be aware, whenever and wherever encountering an emoji, and particularly many in a series, that information is being conveyed even when the message may not be apparent to the uninitiated. A useful guide to the **official emojis** is found at https://emojipedia.org.

Audio-to-Text Transcription

Transcription tools make it possible to turn audio materials into a written transcript. The development of audio-to-text conversion tools saves time as well as frustration. Things often go better – not to mention faster – when you can search the transcript.

- **Otter (Meeting/interview transcription):** https://otter.ai/login
- **Temi: (Meeting/interview transcription, $$$$):** https://www.temi.com/.

More than two dozen transcription tools are outlined at the **Tow-Knight Center for Entrepreneurial Journalism**, which can be found at https://medium.com/journalism-innovation/the-best-new-ways-to-transcribe-c4c342abf172.

Truncating URLs…

Learn to truncate a website's URL, going back one forward slash (/) at a time until you reach the home page address. This will allow you to see whether you can look into any of the website's file folders. When you can, you never know what you will find. Sometimes, when trying to do a complete job, it is worth clicking every link on the home page – or on every page and then truncating each web address. This technique previously provided more information about a site and the material on it than it does now because site owners were not as careful then in their set-ups. Hpowever, it remains a worthwhile exercise whether you uncover any useful information or not. When you can say you truncated all of the URLs visible on a site it shows you have done due diligence, not a job that was just "good enough."

Databases

Databases are usually created and maintained by businesses, governments, experts, or at least by people who are knowledgeable in a particular field. They are high-value targets for any researcher. Databases are often found on scholarly sites, ones offering better-quality resources.

Databases organize information. They are almost invariably devoted to a specific subject. Search engines cannot and do not download or index database information – which is often massive. Even if search engines had the bandwidth to download and save the mountains of information in databases, they are often protected by firewalls or sign-ons that keep search engines out. Most search engines will find and record the existence of databases – they just cannot move through the front door since they can neither sign in nor enter a term in the search field.

You can often find databases related to your subject by adding "db" or "database" behind the general subject you are looking them up in your search. Or, when you get on a site that seems likely to have a database, type "database" or "db" into the site's search engine. This will often reveal the existence of any database. If one shows up in a site search, click on the link just as with any other search and proceed from there to use, or try to use, the database.

Some databases that may be useful to you, or might suggest the types of databases that are available to you, include:

- **American Religion Data Archive:** http://www.thearda.com/index.asp
- **Ancestry:** https://www.ancestry.com/
- **Britannica Online ($$$$):** https://www.britannica.com/
- **CountryWatch:** http://www.countrywatch.com/
- **DOD and Military Electronic Journals:** http://www.au.af.mil/au/aul/periodicals/dodelecj.htm
- **Dudley Knox Library Databases:** http://libguides.nps.edu/az.php

- **EBSCO ($$$$):** https://www.ebsco.com/
- **FirstGov:** https://www.usa.gov/
- **Georgetown University Library Dissertations:** https://guides.library.georgetown.edu/dissertations
- **Google Scholar:** https://scholar.google.com/
- **Government Databases (Louis J. Blume Library):** http://lib.stmarytx.edu/c.php?g=288016
- **Government Printing Office:** https://www.gpo.gov/fdsys/search/advanced/advsearchpage.action
- **Libraries on the Web:** http://www.lib-web.org/
- **Monash University Databases and Resources**: http://guides.lib.monash.edu/subject-databases
- **National Inventory of Dams:** https://catalog.data.gov/dataset/national-inventory-of-dams
- **ProQuest:** https://search.proquest.com/
- **Reference Desk:** http://www.earthstation9.com/index.html
- **Search Systems:** http://publicrecords.searchsystems.net/
- **The Academic Web Link Database Project:** http://cybermetrics.wlv.ac.uk/database/
- **The Educator's Reference Desk (Lesson plans):** https://eduref.org/
- **The Invisible Web:** http://www.invisible-web.net/
- **UNBISNET:** http://unbisnet.un.org/
- **University of Michigan Library (Maps and atlases):** https://www.lib.umich.edu/clark-library/collections/maps-atlases
- **US National Archives and Records Administration:** https://www.archives.gov/
- **World Basic Information Library:** http://military.wikia.com/wiki/World_Basic_Information_Library.

Plagiarism Searches

Plagiarism – appropriating someone else's work and publishing it as your own – has led to the downfall of many people. It is often associated with copyright violations. Plagiarism checks have also become common among instructors in higher education.

Checking published documents – particularly the theses of people holding higher educational degrees or the writings of politicians – is common among operators running a political opposition research program but is an oft-forgotten technique in many other OSR reviews. Computers have made this type of search far easier than it was once, while the ubiquity of materials on the Internet has made it easier to find and copy others' writings. Some sites will check to see if entire websites have been misappropriated. Resources available to the researcher include:

- **Copyscape ($$$$):** https://www.copyscape.com/
- **HelioBlast:** https://helioblast.heliotext.com/
- **Plagiarisma ($$$$):** http://plagiarisma.net/
- **PlagiarismSearch ($$$$):** https://plagiarismsearch.com/
- **Unicheck ($$$$):** https://unicheck.com/.

Dumpster Diving

Valuable Open Source information is literally thrown away every day, waiting to be collected by the thoughtful researcher. Dubbed "dumpster diving," or "trash picking," the wastebasket becomes a friend to researchers and a foe of anyone they are collecting on.

Few people outside the investigative community give much thought to what they are throwing away. Even organizations that try to recycle often do it for economic advantages – paper, metal and plastic may bring back money – or for social reasons. Too few do it for security.

Because it can be financially remunerative, many people who are looking for a cash turnaround rather than information are dumpster divers. Occasionally their efforts turn up in media reports when

something that probably should not have been thrown away is found and becomes a news story. Protection against dumpster divers of any sort is a major security measure – one that is often overlooked.

How useful dumpster diving is can be readily seen by the fact that a highly-placed US intelligence official was convicted and sentenced to life in prison for working with Moscow operatives. He had thoughtlessly thrown away important evidence of his betrayal, not thinking it would end up on a prosecutor's desk. Expecting anything to be buried forever in a trash heap can be a major mistake.

In the United States the Supreme Court has said that, as a general rule, anything left in curbside trash cans is considered "abandoned" and is free for the taking. Municipal ordinances, often designed to assure recycling is economically viable, may make the trash the property of the local government. For that reason dumpster diving may be illegal in some locations, but where it is lawful it is one way to acquire critical information

Trash cans, dumpsters, recycling bins, and even members of the nightly cleaning crew can help researchers piece together inside information. Recycling containers are often sought out, not only because they don't contain old banana peels and coffee-impregnated paper cups, but because that is where people trash their notes, their drafts, and the information miscellany that can tell as much about an operation as an inside snitch. People may shred things, but too many rely on shredders that cut pages into thin strips that can be readily pasted back together instead of cutting paper into the rice-grain size slivers that make good ticker-tape parade snow. Cross-cut shredders that actually do the job are a good security investment in many situations. Some of the best shredders will even take care of CDs and credit cards. Hard drives and thumb drives require special erasure and disposal techniques.

While a single letter, paper, document, or even a slip of paper with a phone number may seem insignificant at first glance, when it is combined with other knowledge that one piece of trash may fill out the mosaic of crucial clues. Moreover, many things that can contain Personally Identifiable Information (PII), including Social Security numbers, make their way to the trash.

Dumpster diving – also called trash picking – is safest when people wear long sleeve shirts, jeans, heavy leather gloves, and leather boots because people do throw away sharp and dangerous items.

If you can do it to others, others can do it to you. Remember that. From a self-protection standpoint, keep in mind the fact that any trash containing sensitive information may be intercepted at many points – while the material is at your location, while in transit, or even at the dump or recycling facility. Open recycling bins and trash cans are never safe storage for sensitive material on the way to the dump or recycle site.

Threats posed by loose paperwork are almost too many and diverse to mention. Important items include:

- Attendance records
- Correspondence
- Coursework or information about training, marketing or sales
- Credit card information, including offers
- Customer or order information
- Delivery and deliverable information
- Development plans for future projects
- Digital media such as floppy or CD disks
- Emails
- Employment data
- Financial records of any type
- Insurance information
- Internal notes and memoranda
- Maintenance records

- Medical files or information
- Payroll Information
- Personal notes
- Price lists or invoices
- Reports
- Rosters or phone tree information
- Schedules
- Shipping or other labels
- Supplier data
- Travel information.

Internet of Things

The Internet of Things (IoT) promises – or perhaps the word is threatens – to open vast vistas of personal information. The IoT contains information loads that are far beyond what any government was capable of gathering only a few years ago.

This information grocery store is widely available to anyone who wants to shop in it and learns how to exploit it.

The IoT links devices and resources, giving them a communications pathway through the Internet. IoT devices collect, transmit, receive, and exchange data. Usually the data transmission and any IOT activity is unknown, unseen and unmoderated by humans. Are the IoT exchanges openly available information? That remains to be legally determined in many cases! But until, and unless, courts weigh in decisively expect that the area will be exploited.

Lines are continually being blurred; what only governments were once to be able to do, civilians can now do better. Linking the IoT and the ever-developing techniques of today and tomorrow promises to provide everyone the intelligence – read spying – capabilities which were available only to first-world governments in the past.

Today the IoT links things as diverse as buildings, security systems, defibrillators, televisions, automobiles, insulin pumps, digital assistants, and refrigerators. The number of items connected to the IoT, and their vulnerabilities to exploitation, has exploded.

The major problem with smart TVs, smart refrigerators, or other parts of the IoT is that they are often smarter than their owners.

A smart TV may well be spying on you – revealing what you watch and how long. Unknown to you the TV may well be linked to your computer and have stored, somewhere in its memory, every document or keystroke including user names, credit card information, and passwords. Your car may be leaking information about your speed, location, or upkeep record. Digital assistants are more than willing to help you buy anything from food to underwear. They can help you because they have collected databases of information about you, your daily contacts, your wants, and your needs.

The IoT sends all kinds of information about you, your habits, your environment, and your likes and dislikes, to servers that are located who knows where for who knows what purposes. That is the outgo; there is also the potential for incoming. Some parts of the IoT are interactive – that "other side" may be able to reach in and "correct" problems by turning off your heart monitor, your car's braking system, or changing the temperature inside your refrigerator.

You need to read – read for understanding – every privacy notice that pops up on any electrical device you buy. But knowing what a company says it will do with your information is not the same as how it, or hackers, can tinker with the underlying technology to affect you. Dealing with "smart anything" requires an understanding of what information you are sending, where, to whom, for what purpose, and who can share in that information. Security suggests you should know what others can do to, or through, the smart technology you are using. This usually turns out

to be a black hole, an information blank. Even the creators of the technology may not know or understand how their particular piece of technology can or would be misused. Technology developers think in terms of use, not misuse.

But the fact is, with the IoT, areas that were previously open only to government interception have become available to many others, for all sorts of reasons, and all sorts of exploitation.

Civilian commercial information sites increasingly move into areas that once were the exclusive domains of government. Today's technologies enable anyone with even modest computer skills to do things that were once reserved for advanced nation states. Some of the newest techniques may be deemed illegal by some governments, but the capabilities are there and they can be exploited by anyone who is knowledgeable.

As more items in our daily environment are controlled by, or use, the Internet for communication, the amount of data they can provide to the information interloper grows. Legacy devices, produced at a time when manufacturers were more interested in getting a working product than in providing security for users, are of special concern. But even modern IoT products face the talents and techniques of a hacktivist army – some of whom would and could turn devices into a lethal weapon. The threat of information theft from the IoT, and the parallel possibility of using the capacities for many nefarious purposes, is significant. Of serious concern: In time IoT exploitation will be used for crimes, including murder.

The IoT has serious implications – both good and bad – for open source researchers. The collection of "smart devices," cars, planes, and buildings connected to, and controlled through, the Internet, grows apace. The exponential growth of linked devices is expected to balloon the size of the IoT to tens or hundreds of billions of connected items.

Security measures for many IoT devices is largely an afterthought. The unsecured information on IoT devices is readily available to researchers, hackers, cyber criminals, and operatives of unfriendly foreign powers. Users seldom know of the security threats that may be posed by any compromised IoT devices, whose systems may be interrogated and used for any number of purposes.

Patching IoT devices is often difficult or even impossible. Security verification – making certain that the orders being sent to or from the item through the IoT – is nil. Knowledgeable intruders can often do what they will with connected devices. Devices usually lack any useful forensics capability. IoT items may include:

- Entertainment devices such as toys, DVRs, TVs, and music systems
- Home automation items, such as controllers of EVAC, electricity, and lock systems
- Hubs and routers controlling a fleet of other IoT devices
- Medical devices such as drug dispensers and heart monitors
- Office equipment including printers and computer peripherals
- Security alarms and cameras
- Smart appliances such as stoves and refrigerators
- Transportation vehicles and systems, including cars, trains and aircraft
- Wearables such as watches, fitness trackers, and smart clothing.

Hackers can also use IoT devices for such mundane things as DDoS attacks and in Botnets, or accessing and controlling devices using well-worn default passwords and usernames. They can be used to burrow into private networks.

Users of IoT devices seldom know what data the item collects. How data is stored, whether it is encrypted, if there is third party access, and how long the information is retained are other questions

that can seldom be answered by IoT users. Opt-out measures, even when available, are usually difficult and often will destroy the usability of the item.

In most cases the IoT should not be considered part of the open source information world. However, because of continuing lack of oversight by manufacturers, coders, and users of the IoT, this area continues to be wide-open for exploitation. Vulnerabilities promise to grow geometrically.

A major search engine that can be used to find IoT resources is **Shodan,** which can be explored at https://www.shodan.io/. An acronym for Sentient Hyper Optimized Data Access Network, Shodan collects data from computer ports making it useful for penetration testing. There is a limited-use free version as well as a for-pay product.

Other IoT sites of interest include:

- **Censys:** https://censys.io/
- **Thingful:** https://www.thingful.net/
- **Zoomeye:** https://www.zoomeye.org/.

Mobile Phones

Mobile phones have become a popular way of accessing the Internet and information. They can be used in the stead of computers for research, but screen size, keyboard size, and a dearth of special research programs – not a lack of computing power – make them last choices for most OSR professionals.

Security is another significant problem when using a cell for OSR. Phones are precise identifiers. A phone is tied to a unique number that can and does identify you in much the same way a Social Security number or home address will. Few people share cell phones or phone numbers, making a number almost as individual as DNA. You, or others, can be traced by the phone number and calls made on it. Phone numbers are included on all sorts of

documents and postings. Once s researcher ties a phone number to a person that is gold.

Mobile security is imperative in theory, but scant in reality. Android phones are open source and are unvetted. The IPhone is safer, provided users are careful about the security settings. But safer is not the same thing as safe.

The upshot is that mobile phones, in particular, leak like colanders. This benefits researchers who are on the lookout for others who are using them. But savvy researchers generally stay away from them as a tool. Those who do use them make certain the phone's history is cleared and they take steps to make certain the phone will not show location. For some people using old-style burner phones that may be as simple as removing the battery when not in use, or for many newer phones as difficult at leaving it behind when trying to prevent anyone from tracing your activities and routes.

Phone messaging apps may – or may not – be highly vulnerable to intrusion by intelligence agencies, but they usually are of significant concern.

Private messages are always off limits for OS researchers; none-the-less one-way channels and two-way groups may be accessible for some researchers. Important phone apps include:

- **Telegram:** https://telegram.org/
- **WhatsApp:** https://www.whatsapp.com/.

The ubiquity of phones and cell phones and their reach is, overall, a positive for OSR. Everybody, it seems, has a phone, even grade-schoolers and some of the poorest people in the world. People who don't have running water in third-world countries will often still have a ringing phone.

People tend to break out their cell phones to record unusual, abnormal, or violent events of all types. People post commentary

and visuals – still pictures and videos – to social media and photo-sharing sites, or sites such as YouTube. Cell phone users provide more material for researchers than would be available if computers were the only source of postings.

But the yin and yang of mobile phones poses problems for the researcher. Phone numbers are useful identifiers. Postings from phones can provide much good information, particularly in developing disasters. But some parts of the Internet are not computer-friendly. Researchers who want to check sites like Tinder, Snapchat, or YikYak cannot search those sites from their computer. Some portions of the electronic world can only be reached by cell phones. Researchers cannot live with them; they cannot live without them.

For open source researchers the answer to many of the problems is to set up and use one of the many cell phone emulators available for use on the Internet. Most are designed with the game-playing segment of the public in mind, but with some thought they can be used in place of an actual cell for OSR work. Emulators are seldom easy to set up and in many cases are complicated to use. But when you need one they can be a valuable piece of technology to have available. They include:

- **Android Studio:** https://developer.android.com/studio/
- **BigNox:** https://www https://bignox.com/
- **BlueStacks:** https://www.bluestacks.com/
- **Droid4x:** https://droid4x.en.uptodown.com/windows
- **Genymotion:** https://www.genymotion.com/#!/
- **KoPlayer:** http://www.koplayer.com/.
- **PrimeOS (Operating system):** https://primeos.in/.

Another use of the phone for open source researchers is checking phone numbers to confirm the name of the listed owner. This is not always an easy task and it is complicated when people on some sites will list their phone numbers in both numerals and text, as in

123 four five six 7890 or a similar version. They may do so in order to avoid commercial webscrapers who grab numbers and sell them to phone solicitors; they may also do it for reasons that are hardly altruistic or defensive.

Despite the problems of using phones and their numbers, there are online services that can provide information. For those who have a phone number that they want to link with a name, the **CallerIDTest** at https://www.calleridtest.com/ may be helpful.

Others may benefit by using the phone number lookup search boxes that can be found at **ThatsThem,** https://thatsthem.com/, the **TruePeopleSearch** site, https://www.truepeoplesearch.com/#, or the **411.com ($$$$)** site at https://www.411.com/. Many of the background check sites also offer some form of reverse-phone lookup.

***

Project:
In your SOP document write in the techniques section whether you see any possibility of using dumpster diving.

Project:
Whether or not you anticipate using dumpster diving, describe protective measures you will take against the tactic in the security section of your SOP.

Project:
In your SOP describe how or whether you will use a cell phone or emulator and any details about what defensive measures you will employ with phones.

Project:
In your SOP write whether you will explore and use the IoT, and if so indicate how you will learn more about it and access it.

Project:
If you have a need to search in foreign countries, in your SOP outline the techniques you will use to find information if that differs from the method you will use in your home country.

Appendix G. Basic Search Plan Format

Name(s) of Researcher(s):

_____
_____

Client:

_____
_____

Priority – Is time more important than detail level?

_____
_____

Product format:

_____
_____

WENTK – Who else needs to know:

_____
_____

LTIOV – Last Time Information of Value:

_____

Date of Start:

_____

Tentative search timeline, including drafting and review:

_____
_____
_____
_____
_____

Scope and Focus:

Original Question:

Question to be researched:

Who, What, When, Where, Why, and How elements:

Context of the question. The user's actual need:

What does requestor already know; and doesn't know

Search technique:

- Hunter-Focused
- Gatherer-Broad
- Mixed

What must be found out:

_____
_____
_____
_____
_____

What resources, tools and methods will be used:

_____
_____
_____
_____

Assets I already have (knowledge/capabilities):

_____
_____
_____
_____

Assets needed (refinement/tools/help):

_____
_____
_____
_____

Information I can get in time (deadline):

_____
_____
_____
_____

Likely location(s) (Physical area of search):

_____
_____
_____

Languages to be used:

_____
_____
_____

Product format desired: _____

Identify what is known and who knows it:

- How much do I know about the subject?
- Has this question been addressed previously?
  - If so, by whom?
  - If so, where is that information?
  - If so, can I get it there?

_____
_____
_____
_____
_____
_____
_____
_____
_____

Resource types to be used in collection (Circle all to be used):

- Alert services
- Audio-visuals
- Blogs
- Books and magazines
- Chatrooms
- Corporate sites/materials
- Databases

- Directories
- Email
- Forums
- Games
- Government sites
- Gray material and ephemera
- Libraries
- Listserves/discussion lists
- Meta-search engines
- News media
- Non-Internet resources
- Observation
- Photos or satellite views of important sites
- RSS feeds
- Search engines
- Social media
- Subject matter experts
- The Dark Web
- The Hidden Web
- The Open Web
- Think tanks or subject matter experts
- Usenet groups
- Wikis
- Other resources:_____
  _____
  _____
  _____

Browsers and Search Engines to be used

_____
_____
_____
_____
_____

Location(s) To Be Saved To:

_____
_____
_____
_____
_____

Security Plan Outline:

- Own Critical Information: _____
- Possible Threats:_____
- Vulnerabilities: _____
- Risks:_____
- Countermeasures:_____

Keywords: Found in question researched and from searched items:

Example Word	**Similar**	**Broader**	**Narrower**	**Related Words**
Smoking	Smoke, smoker	Tobacco	Cigarettes, Cigars	Nicotine, Lung cancer

Names, handles, personal identifiers:

_____
_____
_____
_____
_____

Family Members:

_____
_____
_____
_____
_____
_____

Groups, Organizations, Associates:

Government and Court Searches:

Venues to be Searched (Name/Date):

Sites where information found:

Sites Devoid of Useful Information:

Actual Browsers Used/Date-Time:

Actual Search Engines Used:

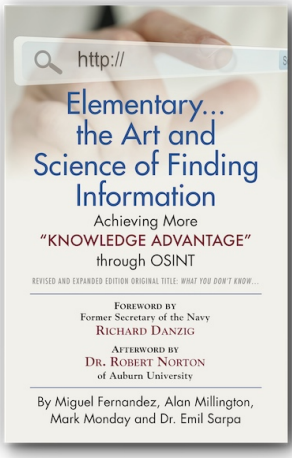
Vetting and Validation of Sources Methods

Researcher Comments:

Coming soon!

Join Us to Explore More of
The Wide World of Information
at:

opensourceresearch.net



Facts and fake news live side-by-side in the Internet. Elementary... the Art and Science of Finding Information is a guide to achieving information superiority in today's world.

Elementary... the Art and Science of Finding Information: Achieving More "Knowledge Advantage" through OSINT – Revised and Expanded Edition

Original Title: What You Don't Know...

By Miguel Fernandez, Alan Millington, Mark Monday and Dr. Emil Sarpa