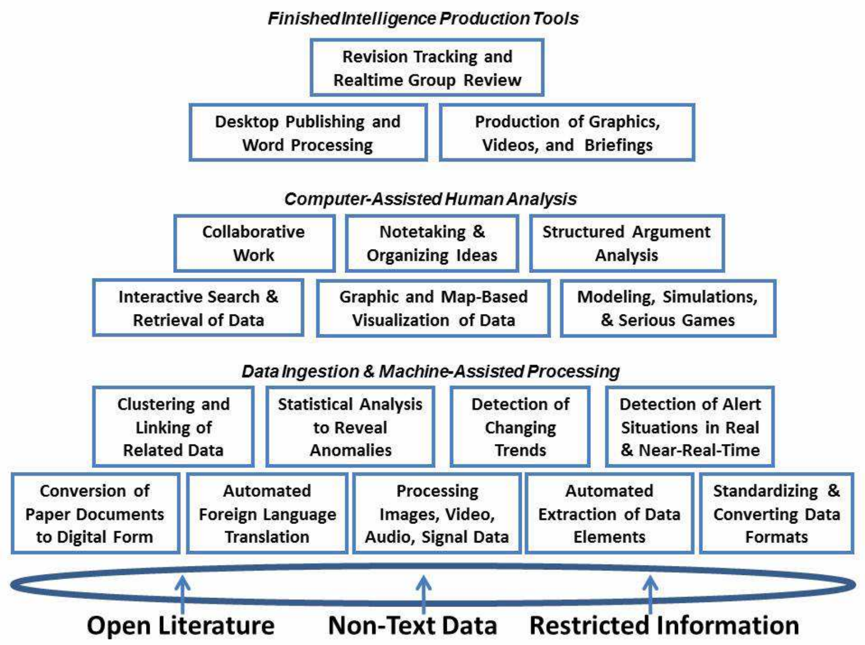
Foreword

By Robert David Steele

In 1986, I was selected from the CIA's clandestine service to help lead a pilot project to bring the CIA into the 21st Century. From that moment almost 30 years ago, I have been obsessed with open sources of information in all languages, mediums and computer-aided tools for analysis—everything the CIA does not utilize today. I took my cue in the mid-1980s from author Howard Rheingold, who explored how computers could be used to amplify human thought and communication, and the CIA Directorate of Intelligence team of Diane Webb and Dennis McCormick.[1]

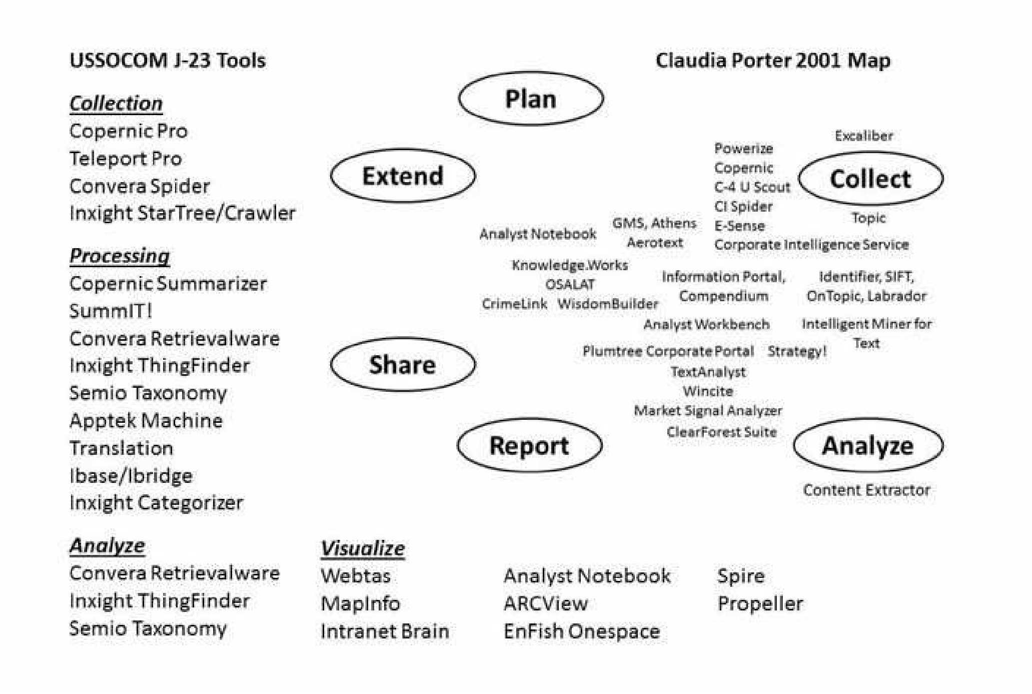

Finished Intelligence Production Tools
Revision Tracking and Realtime Group Review
Desktop Publishing and Word Processing
Production of Graphics, Videos, and Briefings
Computer-Assisted Human Analysis
Collaborative Work
Notetaking & Organizing Ideas
Structured Argument Analysis
Interactive Search & Retrieval of Data
Graphic and Map-Based Visualization of Data
Modeling, Simulations, & Serious Games
Data Ingestion & Machine-Assisted Processing
Clustering and Linking of Related Data
Statistical Analysis to Reveal Anomalies
Detection of Changing Trends
Detection of Alert Situations in Real & Near-Real-Time
Conversion of Paper Documents to Digital Form
Automated Foreign Language Translation
Processing Images, Video, Audio, Signal Data
Automated Extraction of Data Elements
Standardizing & Converting Data Formats
Open Literature
Non-Text Data
Restricted Information

This diagram identifies the 18 functionalities for an intelligence "toolkit." Diagram prepared by Robert D. Steele, 2014.

[1] They identified 18 functionalities that are essential for any all-source analytic toolkit.

Later, as the senior civilian responsible for creating the Marine Corps Intelligence Activity (MCIA) from scratch, I helped draft and implement the winning statement of work for the Joint National Intelligence Development Staff (JNIDS) competition, only to see it torpedoed by the ostensibly “joint” Admiral in favor of an anti-submarine project.

In 1992 I began the first completely open international conference on open sources and methods, and over the years was deeply impressed by the quality of the presentations we attracted. One in particular stands out, that of Claudia Porter, speaking in 2001, on “Tools of the Trade: A Long Way to Go.”



USSOCOM J-23 Tools

Collection
Copernic Pro
Teleport Pro
Convera Spider
Inxight StarTree/Crawler

Processing
Copernic Summarizer
SummIT!
Convera Retrievalware
Inxight ThingFinder
Semio Taxonomy
Apptek Machine Translation
Ibase/Ibridge
Inxight Categorizer

Analyze
Convera Retrievalware
Inxight ThingFinder
Semio Taxonomy

Visualize
Webtas
MapInfo
Intranet Brain
Analyst Notebook
ARCView
EnFish Onespace
Spire
Propeller

Claudia Porter 2001 Map

Claudia Porter’s 2001 graphic provides a useful context for the evaluation of analytic tools. Note that in the last decade many of the companies have gone out of business or been acquired. Software has progressed, but not nearly enough.

From my vantage point, little has changed in the intelligence and security field over the past quarter century. In fact, the commercial sense-making has made surprisingly modest progress. The stall-out struggles to tap the vast quantities of multi-lingual, multi-disciplinary, multi-domain information in 183 languages. Furthermore, 33 of those languages are vital to creating understanding and wisdom. Many tools are behind the Big Data and OSINT curves.

CyberOSINT: Next Generation Information Access is an attempt to put some of the most sophisticated tools in a comparative context. All of the tools addressed by this first edition are useful after a fashion.

They are also not good enough.

In some instances, they will be judged to be unaffordable, not interoperable, or unable to scale across the security barriers in government, much less across the eight information networks that comprise the totality of what can be known.[2]

Many of the challenges stem from source diversity and integrity issues. A holistic analytic process is needed; it will fuel additional innovation toward next-generation information access.

Venture funding is available. More pressure must be exerted on the vendors, large and small, established and newly launched. Only a handful of companies discussed in this report are on the fast track financially. More important is that the functionality is missing as well, outside of isolated domains with relatively tame data sets. Mary Meeker, a well-known analyst, has said that less than one percent of "Big Data" is being processed - the same percentage holds for NSA's mass surveillance. Google is on record as being unable to process more than four percent (I personally think it is closer to two percent) of the full web - shallow (yes), deep (no), and dark (never).

We have a very long way to go.

In my view, the systems described in this volume have something to offer.

However, most of the capabilities in this volume will be displaced within five years by open source alternatives able to work with a distributed open source network that resides on end-user devices and displaces both the cloud and enterprise servers.

I could be wrong. But I could not possibly be as wrong as we have all been this past quarter century.

We have failed to meet the demand for a serious all-source analytic toolkit able to ingest all forms of data in all languages, and make sense of that data in historical, current, and future-oriented contexts.

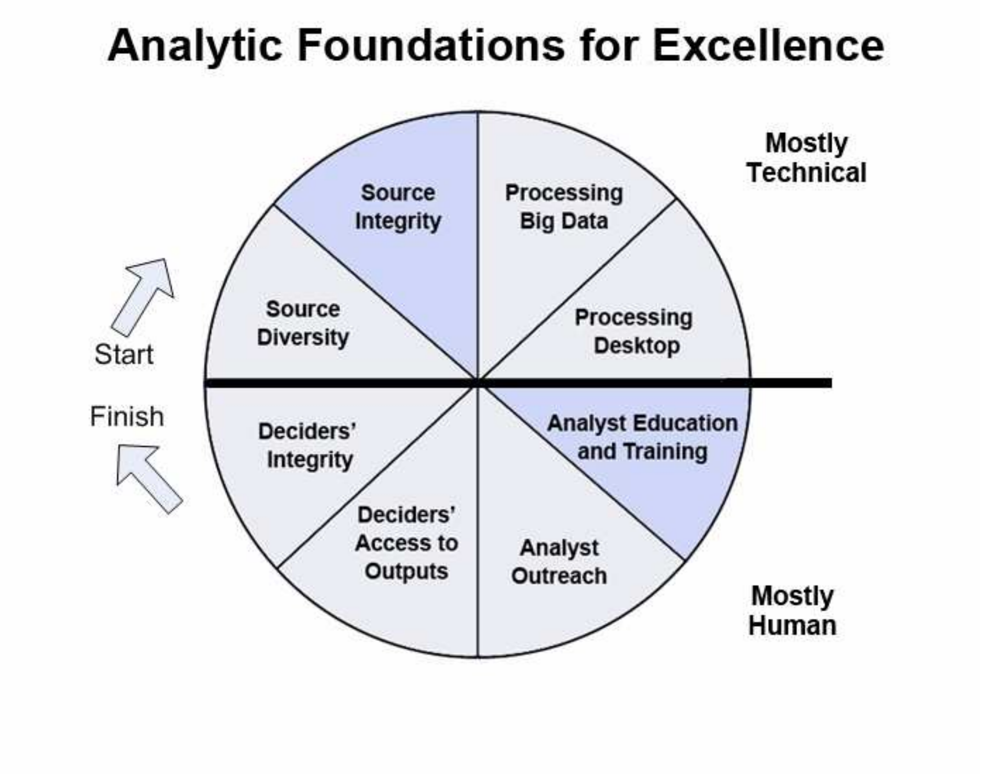
Here are seven points the reader should consider regarding the state of CyberOSINT.

While remembering that my motto is "the truth at any cost lowers all other costs," here is the Naked Truth, in my opinion:

[2] Academic, civil society including labor and religion, commerce especially small business, government especially local, law enforcement, media including bloggers, military, and non-governmental/non-profit. The ability to share information and collaborate in sense-making across all boundaries is not possible today.

1 **Automation from smart software is becoming increasingly important.** No fewer than five of the CATALYST building blocks deal with data ingestion. In 1988 my priority for data collection focused on the need to be able to scan crumbled captured documents in the field and get near real time translations.

Today it still takes up to ten days to get a captured Dari document ingested, translated, and returned. We need more smart software at every point of the information pipeline.



2 **Smart software does not replace smart people. It complements their capabilities.** The most serious mistake made by law enforcement, intelligence and security officials has been to assume that secret sources and methods are the primary source of value, and that young inexperienced analysts are "good enough." Not true.

3 **NGIA systems are integrated solutions that blend software and hardware to address very specific needs. Our intelligence, law enforcement, and security professionals need more than brute force keyword search.** We have, as Claudia Porter documented in 2001 and we document now, a long way to go. In my view the next big leap is going to be achieved at three points: Data ingestion and conversion; per-

vasive geospatial tagging of all data; and integrated human conversations on top of securely shared visuals.

4 **Much work remains to be done. Accuracy for classification, even for the best methods, scores about 85 percent. Audio and video content is largely inaccessible. Slang in many languages is not handled well. False information is not detectable by machine processes.** And this is just in relation to the digital information arena.

We have lost ground in relation to identifying top experts (ProQuest killed the RANK Command that DIALOG offered for processing citation analytics), and most analog information remains "invisible")

5 **Outputs that are more eye candy than food for the mind - and confusing outputs divorced from their source mix - are a disservice to all in the community.** This problem is compounded by senior officers ignorant of the craft of analyst, who confuse colorful meaningless charts with "intelligence"

6 **Each system is only as good as the combination of its source mix, its internal algorithmic integrity, and the skill of the analyst using the system.** A major flaw shared by all buyers and users of these systems is the assumption that the data they have is the data they need. I personally find computational mathematics suspect because the threshold settings and internal configuration of algorithms are not revealed to colleagues or to licensees.

I do not trust any ranking based on assumptions hidden in code. Those systems that require a great deal of training tend to operate at twenty percent of capacity, and hence do not render value. Interoperability of data, interoperability of processes, and interoperability of humans matters. We must begin to address those three challenges in a responsible manner.

7 **This monograph is not the last word on next-generation information access.** Among data points that may be explored in future editions are revenue and return on investment information and an evaluation matrix that scores vendors on their source mix, internal integrity, training required, utility of output, and percentage of open source code.

This monograph is a starting point for those who might wish to demand a "full spectrum" solution, one that is 100% open source, and thus affordable, interoperable, and scalable.

I do suggest that any law enforcement, security, or intelligence professional learn the context of these next-generation systems and become familiar with specific vendors' products and services.

One cannot create the future unless each professional has a good grasp of what is available today, as well as the rich possibilities not yet realized but clearly within our grasp in the very near future.

Robert David Steele

Former Marine major, Central Intelligence Agency, and open source intelligence officer and Chief Executive Officer, Earth Intelligence Network (501c3)
Oakton, Virginia

January 22, 2015