Large Scale Data Processing & OSINT Overview

What's this about and why now?

This document provides an overview of the LSD & OSINT project to ensure our stakeholders and others have better sight of the work we're doing and plan to do. The document includes points of contact for each of the work packages and activities, so if you have any questions please get in touch.

It's one of a number of things we intend to do to improve how we communicate:

- Project team members will be posting updates on their work. It'll be pithy and informal, but you can probably expect to see some sort of update in this group when they're kicking off or wrapping up a new piece of work.
- We'll share these monthly summaries with external stakeholders via email (in the absence of good collaborative tools with MOD).

We're putting this document out now to provide some context to the subsequent informal and monthly updates. Spoiler: if you've read the IPP, some of this will look familiar...

WP1 - S&T Support to future OSINT Capabilities

This work package aims to deliver science, technology and innovation in support of MOD's future OSINT enterprise. Success for this work package will constitute S&T developed or coordinated by Dstl being exploited by those in MOD seeking to advance or benefit from OSINT collection, processing and analysis. In addition to the Fundamental OSINT Research Activity and the OSINT Experiments and Trials, there are several enabling tasks to ensure Dstl has the right policy and licenses to practice in place to conduct its research.

Fundamental OSINT Research

Web Fundamental Science - Research in this activity will focus on three, programme-funded PhD studentships with two, top-tier UK Universities over the period 2015-2019. The PhD PhD topics will include: innovation of a conceptual 'geology map of the Web' that will allow Defence to understand conceptual sub-populations of Web content; development of an appropriate and evidenced statistical sampling strategy to apply to social media Web resources; and estimation of the capability and intent of strategic industries using open source information, the geology map of the Web and the sampling strategies.

Web Risk Modelling - When Defence staff use the Web, they reveal much information about Defence's interest to an increasing number of unknown profiling systems that track queries and the Web pages that are visited. This activity will develop a software implementation of a Web Risk Model (based on academic literature) and quantitatively assess its ability and system stability in balancing adaptively Defence Web querying with that of the external population's querying pattern.

Deep and Dark Web Mining Capabilities – This research activity will seek to understand the Deep and Dark Web from a Defence perspective and provide initial and explicit evidence of its utility for Defence missions. The research shall produce a deliverable, technical report of the evidence gathered throughout the year with conclusions and recommendations to inform how Defence should/could exploit Deep and Dark Web activity.

MACE – In collaboration with the US, Dstl are investigating the benefits of adopting a Multi-Agency Collaborative Environment (MACE) model in the UK whereby OSINT analysis is outsourced to a consortium of approved companies.

OSINT Experimentation & Trials

This activity will combine the essential elements of technology, intelligence problems, underpinning architecture and multi-disciplinary teams to develop and test technology and methods in support of OSINT collection and analysis. OSINT Experiments will aim to understand and demonstrate the value of OSINT processing capabilities applied to specific operational problems through the development and integration of S&T. OSINT Trials will aim to expose tools and techniques to analysts within their working environment for testing and feedback to inform near-term and future operational capability. In the short-term, this activity will work on:

OSINT Imagery Collection – A user-driven experiment to assess the utility and problems of collecting, serving and processing of open source imagery and video media in support of UK and overseas CT and crisis response.

Target Audience Analysis –The work will assess the potential for OSINT to contribute towards a digital Target Audience Analysis (TAA).

Scientometrics & Expert Mapping – The processing, visualisation and analysis of open source academic information in support of DIAS in collaboration with Dstl Knowledge Management Exploitation (KME).

Global Event Alerting – The collection and processing of news media to identify events of potential interest to MOD in a timely fashion. This sounds like a bit of a strange one, so ask if you've got questions.

Further work will develop our user-level links with MOD OSINT users to identify other areas of development and experimentation.

WP2 – Large Scale Data Processing

MOD needs to understand the benefits which emerging "Big Data" technologies can provide, and how they are best applied to Defence problem-sets, in order to support decision makers. This work package will address a broad set of requirements across MOD to help users understand the benefits; how these technologies could be integrated as part of a wider ISR architecture; and the associated DLOD implications.

Big Data in Defence: Opportunities & Problems

Under this activity Dstl will work with MOD stakeholders to identify potential experiment topics and datasets where Big Data tools and techniques could be used to provide useful insights and recommendations to improve on current processes. The topics will be passed over to the experimentation team, where they will be reviewed and prioritised for the data science investigations. This activity builds on the Survey of Big Data Opportunities in MOD Management Information, carried out in FY14/15 in support of Chief Technology Office (CTO).

Data Science Investigations

This activity will deliver the stakeholder-prioritised investigations developed under the Opportunities & Problems study. Given the wide interest in the topic amongst stakeholders, there is a need to expand the areas of application for investigations compared to previous years (which predominantly focussed on Big Data support for Intelligence analysis) in order to meet needs from across MOD.

The main output from investigations will be the evidence to demonstrate the potential benefits to MOD of adopting these capabilities for a variety of business needs. The project will undertake three internal and one external EMR investigation each year.

Constellation

Constellation is a next-generation information gathering, sharing, analysis, collaboration and visualisation platform which is being developed by the US Defence Threat Reduction Agency (DTRA) in support of the Counter Weapons of Mass Destruction (C-WMD) mission. Working with stakeholders in the CB and potentially RN communities, both internal and external to Dstl, a series of exercises and vignettes will be designed and undertaken in order to assess the Constellation working environment (up to Official Sensitive) as it evolves. Outcomes of case studies will be documented and may be utilised by MOD stakeholders as evidence when considering whether Constellation in the UK should be escalated to higher security domains.

Deep Learning

This activity will undertake research into Deep Learning (DL) to understand and demonstrate the benefits of these approaches to Defence. It will take forward agnostic approaches developed within industry and academia, in addition to funding external research to further the science, and apply these against large scale MOD databases. The activity will be split into three areas:

Applied DL Research – A low risk task using proven open source techniques for military and security requirements. This will be delivered by Dstl staff to develop, test and apply DL Models to classified datasets. Initial application areas will include Face Recognition (in collaboration with OGDs) and Object detection and classification from multimedia data (e.g. firearms)

Novel DL Research – Using DL techniques in areas not currently studied in the open literature or commercial bodies due to restrictions in dataset types, classifications and/or financial viability. This will be performed through a series of challenges using the data science website Kaggle (http://www.kaggle.com/competitions) to access hundreds of highly qualified researchers around the world. Initially this activity will focus on Object detection and classification from overhead data (e.g. buildings, vehicles).

Fundamental DL Research – Advancing DL techniques through better mathematical understandings, integration of novel hardware and advanced learning methods. This will be delivered through PhD scholarships potentially through the Alan Turning Institute and/or the University Defence Research Collaboration.