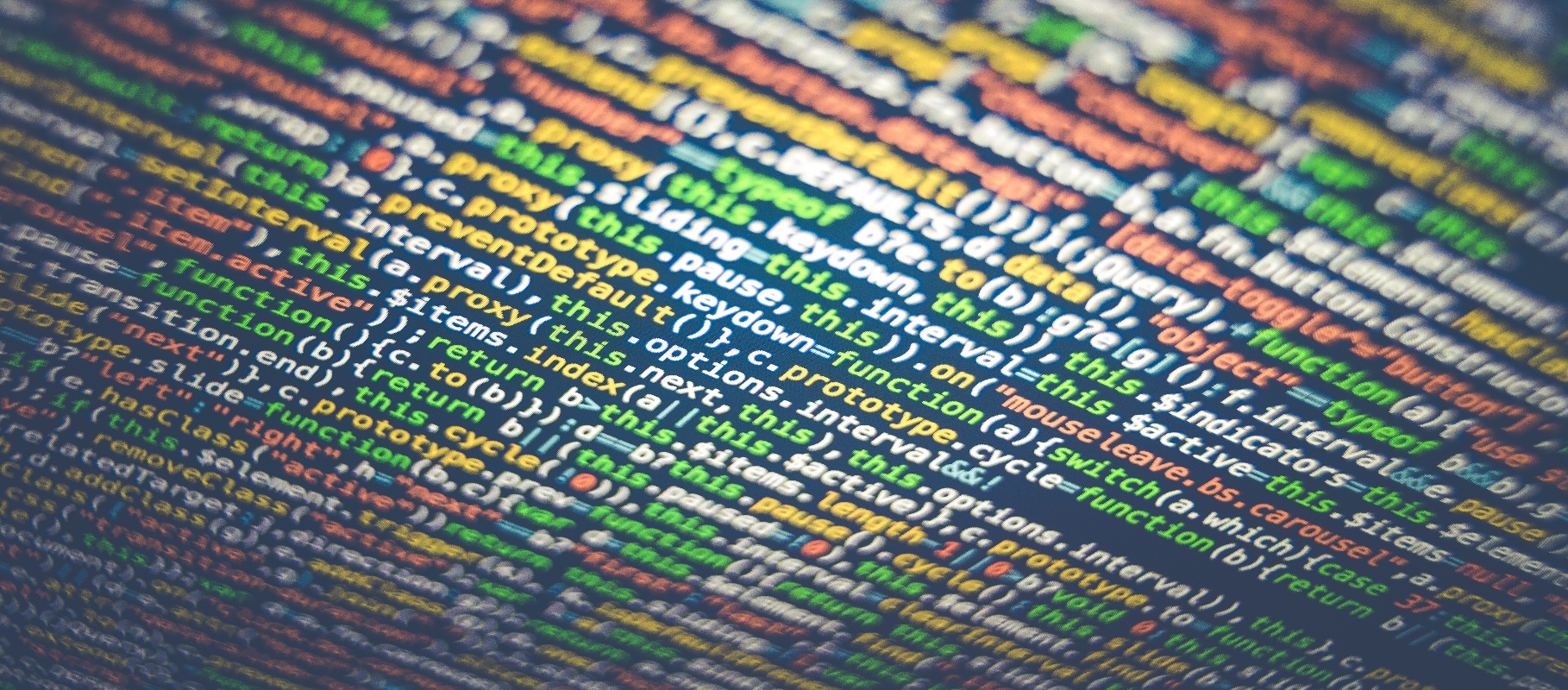


Registry of Available Processing Components

Report for ISA2 Action 2018.02 Processing Open Source Data with Exchangeable Components



Version 0.4  
2019

This is a project related publication by the ISA2 Action 2018.02 – Processing Open Source Data with Exchangeable Components. The Joint Research Centre (JRC) manages the Action 2018.02. The JRC is the European Commission’s science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The document’s content does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication.

**Contact information**

Name: Unit I.03

Email: bertrand.de-longueville@ec.europa.eu

**ISA2 Home Page**

<https://ec.europa.eu/isa2/home_en>

**Action 2018.02 Home Page**

<https://ec.europa.eu/isa2/actions/processing-open-source-data-exchangeable-components_en>

**Action 2018.02 EC Joinup**

<https://joinup.ec.europa.eu/collection/isa-action-201802-processing-open-source-data-exchangeable-components>

Contents

[Overview 1](#_Toc19180260)

[1 Approach 1](#_Toc19180261)

[2 Stakeholder Use Cases 2](#_Toc19180262)

[2.1 Use Case Site Monitoring 2](#_Toc19180263)

[2.2 Use Case Web Scraping 2](#_Toc19180264)

[2.3 Use Case Entity-Centric Search 2](#_Toc19180265)

[2.4 Use Case Internet Search 2](#_Toc19180266)

[2.5 Use Case Blockchain Analysis 3](#_Toc19180267)

[2.6 Use Case Data Visualisation 3](#_Toc19180268)

[3 Processing Component Registry 4](#_Toc19180269)

[3.1 Software Tools 4](#_Toc19180270)

[3.2 Data Providers 5](#_Toc19180271)

[4 Summary 6](#_Toc19180272)

[5 Conclusions 6](#_Toc19180273)

Overview

The ISA2 Action *Interoperable Meta Data and Processing Components for Open Source Information Mining* sets out to define in close collaboration with its stakeholders a set of metadata standards and guidelines on how to adapt existing tools into composite applications. Furthermore, a set of freely available software components fitting these best practice standards should be developed and demonstrated within a demonstrator application.

1. Approach

The purpose of this ISA2 action is to create interoperable and reusable software systems. Therefore, it is important to get an overview of the typical use cases of the project’s stakeholders and their work environments. Only if the developed standards and software tools are compatible with the typical use cases and work environments of the stakeholders, the stakeholders and other interested authorities will adopt the action’s outputs while mitigating project risks.

For this reason, we organised a kick-off meeting with the stakeholders of the action to discuss the most important use cases as well as create a registry of available processing components and tools.

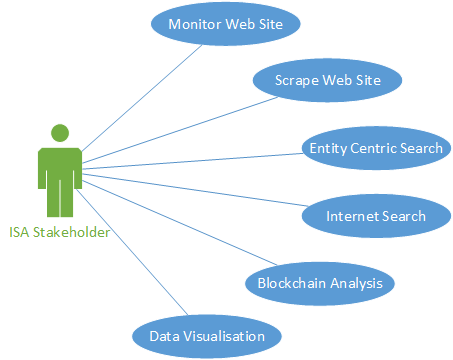
The stakeholder use cases are then taken and fused into a generic use case scenario, which is solved with the action’s demonstrator application.

The following stakeholder organisation took part in this process:

* Dutch Tax Authorities
* Dutch Customs
* University of Applied Sciences Leiden
* Danish Customs and Tax Administration
* Swedish Customs and Tax Administration
* Authority for Consumer and Markets, Netherlands
* Joint Research Centre, European Commission

1. Stakeholder Use Cases

The listed use cases are generic and are supporting more sophisticated use cases specific to the mission of the individual authority. For the purpose of this report a list of the basic generic use cases seems to be sufficient since composite use case scenarios use these cases as building blocks.



* 1. Use Case Site Monitoring

This use case describes the monitoring of websites and forums regarding specific keywords or changes in content.

* 1. Use Case Web Scraping

Scraping of known web sites for monitoring and information extraction. Scraping means to download the source content of a web site up to a specific depth and then processing it to extract data. Recently, a number of tools have emerged providing tools to turn scraped data into specific formats.

* 1. Use Case Entity-Centric Search

Specific search about a person or organisation with taking into account prior knowledge, such as the person’s name, whereabouts, profession, etc. This use case represents a relatively broad field. To assess a person’s background an analyst could use for example social media content. For companies there exist a large number of available online data services, such as public business register information and other public filings. This use case can be supported by a large variety of tools.

* 1. Use Case Internet Search

Internet Search - This use case describes the classical internet search using public search engines, such as Google, Bing, Yandex, DuckDuckGo etc. There are a number of tools, which help to aggregate the found information and further process it.

* 1. Use Case Blockchain Analysis

Analyse Blockchain transactions by using tool who extract transactional data from the Blockchain. The most prominent use of a Blockchain is the implementation of a crypto currency (such as Bitcoin) on top of a Blockchain. This use case is emerging field and of interest to authorities investigating financial transactions and potential financial fraud such as tax evasion and money laundering.

* 1. Use Case Data Visualisation

Visualise data from internet search or data processing. This use case represents also a very broad field. A crucial step in processing open source data and deriving information from it is the way the results are visualised. Only then, the information can be assessed and patterns, relationships and cross-references between numbers or entities can be found.

1. Processing Component Registry
   1. Software Tools

|  |  |  |
| --- | --- | --- |
| Name | Description or URL | Use Cases |
| Bazzell OSINT Scripts | Buscador is a preconfigured Linux virtual machine that contains a collection of OSINT scripts and tools.  <https://inteltechniques.com/buscador/> | Site monitoring |
| Chainalysis Reactor | Reactor is a software suite to help investigator understand and mine cryptocurrency transactions.  <https://www.chainalysis.com/> | Blockchain Analysis |
| Coosto | Sentiment Analysis and social media data mining  <https://www.coosto.com/en> | Web scraping, Analytics |
| EMM OSINT Suite | EMM OSINT Suite is an open source intelligence base line tool which provides scraping, entity extraction and analysis functions  <https://wiki.emm4u.eu/confluence/display/osint/Home> | Internet Search, Web scraping, Entity-centric search, |
| EMM Media Monitor / NewsDesk | EMM Media Monitor is a suite of applications that provides near-real time media monitoring functions. The NewsDesk toolkit allows creating newsletters based on monitored sources.  <https://emm.newsbrief.eu/>  [https://newsdesk.emm4u.eu](https://newsdesk.emm4u.eu/) | (News monitoring), Analytics |
| Europol Tools (Sirius platform) | The SIRIUS project is a Europol tool platform. It provides the EU law enforcement community with tools and knowledge to facilitate internet-based investigations.  <https://www.europol.europa.eu/activities-services/sirius-project> |  |
| FMiner | Fminer is a visual web scraping and data extraction tool  [http://www.fminer.com](http://www.fminer.com/) | Web scraping |
| iRN | Internet Research and Investigation Network (iRN) is a system of Dutch authorities to facilitate online investigations. It provides secure access to the internet and offers added services, such as a possibility to share downloaded content. | Internet Search, Entity-centric search |
| KNIME Analytics Platform | KNIME is a data science tool platform. It allows to create data scienc workflows based on processing nodes.  [http://www.knime.com](http://www.knime.com/) | Analytics, Reporting and Integration  platform |
| Maltego | Maltego is an interactive data-mining tool to find relationships between pieces of information from various locations in the Internet. It is used to mine data from social media sites.  <https://www.paterva.com/web7/> | Entity-centric Search |
| Neo4J | Neo4J is a database which allows to represent interconnected data. The database uses graph structures to allow queries agains nodes and edges used to represent the data.  [http://www.neo4j.com](http://www.neo4j.com/) | Graph Database Toolkit |
| Reverse Image Search | TinEye is a search engine which allows to search for an uploaded image,  [https://www.tineye.com](https://www.tineye.com/) | Entity-centric Search, Internet Search |
| Wallet Explorer | Wallet Explorer is an online Blockchain analysis tool.  <https://www.walletexplorer.com/> | Blockchain Analysis |
| Web Ripper | Web Ripper is a visual scraping and data extraction tool.  <http://visualwebripper.com/> | Web scraping |
| Website Watcher | This is a tool to monitor changes on web sites based on simply changes in content or changes of specific keywords.  <https://aignes.com/de/> | Site monitoring |
| X1 Social Discovery | X1 provides a case-centric workflow to search and collect data from social media sites and normal websites.  <https://www.x1.com/products/x1_social_discovery/> | Entity-centric Search |
| XENON | Custom Internet search tool developed by Dutch Internet Service Centre | Internet Search |

* 1. Data Providers

|  |  |
| --- | --- |
| Name | Description or URL |
| Domain Registration Service | Registrar with information who owns a domain |
| EMM NewsBrief | Provides RSS data feeds of monitored categories  [https://emm.newsbrief.eu](https://emm.newsbrief.eu/) |
| LexisNexis | Access to vast number of databases, such as legal, commercial data  <https://www.lexisnexis.com/> |

1. Summary

This report summarises the results of the first stakeholder meeting for the action. The purpose of thise meeting and this report was to take stock of the currently used tool kits. The stakeholders use a large number of different tools from very simple to complex platforms to conduct open source investigations. For every investigation, analysts pick different tools to collect data, process it and visualise the information. A big challenge is that the data cannot flow seamlessly from one processing step to the next. A core set of tools and standards to support a more integrated work flow could provide huge benefits for analysts and increase their efficiency.

1. Conclusions

The landscape of available tools appears to be rather scattered and fragmented. There are a number of small tools each serving a single purpose. These tools cannot easily be composed in an integrated fashion. One exception is the KNIME data science tool platform. It is composed of processing nodes that can be assembled into data science workflows. Furthermore, KNIME is open source and extendable. However, there are processing nodes missing for general purpose Open Source Intelligence (OSINT) applications since KNIME has been primarily designed to analyse in-house data.

