2.4 INTEROPERABLE META DATA AND PROCESSING COMPONENTS FOR OPEN SOURCE INFORMATION MINING (2018.02)

2.4.1 IDENTIFICATION OF THE ACTION

Service in charge	Text and Data Mining Unit Directorate I: Competencies DG Joint Research Centre Ispra, Italy
Associated Services	

2.4.2 EXECUTIVE SUMMARY

The Internet and its services form a public medium with global reach, easy access and fast information propagation. The use of public information sources is crucial for the mission of many public authorities. A good example is how open source information can support investigations in law enforcement, immigration and customs authorities. However, the amount of information requires the use of automated tools. These tools need to adapt to the rapid evolution of Internet services and must be made of modular components.

Several national public authorities have invested in composite software tool chain for open source information processing. These tool chains are a mix of commercial, bespoke and open source components. Due to a lack of common meta data standards, however, software components (including the underlying resources used by them) of one Member State (MS) authority are often not interoperable with components and services of other MS authorities leading to a lack of reuse. Thus, many authorities are forced to expensively buy or develop software components from scratch which already exist in other MS. Even though, the domain of open source information processing is an ideal area for cooperation since several standards already exist and the processed information is public.

The main objective of this action is to develop and select a set of standards in the domain of processing open source information in order to facilitate interoperability and reuse of software and services between MS authorities.

The Joint Research Centre (JRC) has developed considerable experience in advanced open source information text mining and analysis for open source information processing. Open source information is acquired from the Internet and meta data is computed for different domains, such as media monitoring or law enforcement. Through organising workshops for MS law enforcement authorities and an OSINT community of practice in the last eight years the JRC has gained significant knowledge of the current state of affairs in the field. Additionally, it has developed practical software systems which are shared with the law enforcement community in Europe.

The JRC has learned that MS authorities representing different sectors have started to invest in software tools and services to digest open source information. Such tool sets are always a mix of tools which are specific to the mission of the authority and tools or services which are basically generic. The
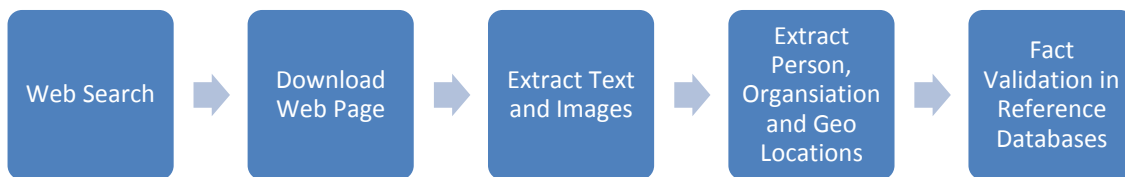
generic components have an enormous potential for reuse. However, due to a lack of coordination and technical standards a common approach to reuse is missing with the following consequences:

- Duplication of same functionality
- Different input and output data standards
- Components from other MS authorities cannot be reused
- Knowledge bases and resources created by one authority cannot be reused by authorities in other MS without substantial effort
- Overly long time to adopt or introduce new software functionality

The dynamic evolution of Internet services providing open source information calls for a pragmatic, agile approach in software development and purchasing. Classical multi-year IT projects are often not a good fit because they do not adapt fast enough to the changing requirements and work environments in this field.

Therefore, the creation of a pool of readily available software processing components (think "Lego") for rapid development of composite applications for acquiring, processing and analysing open source information is of paramount importance.

To give an illustrative example, a composite application with processing components could look like follows:

| Web Search | → | Download Web Page | → | Extract Text and Images | → | Extract Person, Organsiation and Geo Locations | → | Fact Validation in Reference Databases |

In this example, most processing components have generic functionality with potential of reuse. The last two components might be organisation specific with access to case specific person information and internal reference databases.

2.4.3 OBJECTIVES

The main objective of the action is to develop interoperable standards in the domain of processing open source information in order to facilitate and reduce the costs of public service organisations carrying out developments in this context. The specific objectives of the action are to:

- Define a set of meta data standards to facilitate the interchange of software components in the field of open source information processing,
- Define guidelines on how to adapt existing tools relevant for open source information processing and assemble them into composite applications (e.g. processing tool chains),
- Create a set of freely available and interchangeable software components for open source information processing compliant with the aforementioned standards,
- Develop a composite demonstrator application.

With regard to ISA² objectives the envisaged activity, using a holistic approach, will contribute to the development of interoperable cross-border and cross-sector solutions and guidelines for assembling open source information processing application that will indirectly support various Union policies.

2.4.4 SCOPE

The goal of the action is to build upon existing standards wherever possible. Existing standards may have to be adopted in a way that simplifies their use. Often this approach is already foreseen by existing standards in creating so-called profiles which leave out certain parts.
Only where there is a clear gap in existing standards new development is needed in a way to fit into the landscape of existing standards. In other words, the action does not intend to develop "yet another set of standards".
The needed activities to choose existing standards, adapting them and to fill potential gaps are part of the action's plan.

2.4.5 ACTION PRIORITY

2.4.5.1 Contribution to the interoperability landscape

Question	Answer
How does the proposal contribute to improving interoperability among public administrations and with their citizens and businesses across borders or policy sectors in Europe? *In particular, how does it contribute to the implementation of:* • *the new European Interoperability Framework (EIF),* • *the Interoperability Action Plan and/or* • *the Connecting European Facility (CEF) Telecom guidelines* • *any other EU policy/initiative having interoperability requirements?*	Recently, an ever-growing exploitation of open source information by many MS authorities across different sectors and borders can be observed, who are often forced to develop costly solutions for processing open source information from scratch. The activity will boost development of interoperable metadata data standards to facilitate the interchange of software components in the field of open source information processing, including underlying resources, and provide a pool of freely available (ideally open source) software components compliant with the aforementioned standards. In particular, the activity will involve contribution from wide range of MS organisations in the process of analysis and design of the data standards and evaluation of the resulting software components through testing a pilot

	application. The resulting data standards, software components and related reports and guidelines will be available through a publicly open web portal that will facilitate information exchange. Furthermore, the envisaged developments are highly related to other EU policy areas requiring interoperability. For instance, various communications of the EC emphasized the need to improve the cross-sectoral interoperability of EU information systems[33] (law enforcement, immigration, customs, etc.), in whose context exploitation of open source information plays an ever-growing role as well since it needs to be jointly processed and analysed with other type of data. Therefore, alignment of the open source information processing chains with respect to interoperability with the aforementioned systems is of paramount importance and will be taken into account.
Does the proposal fulfil an interoperability need for which no other alternative action/solution is available?	There are certain existing meta data standards. Often more than one for a specific purpose. However, it is needed to select a set interoperable meta data standards and pools of compliant software components for processing open source information that are agreed on the EU level

2.4.5.2 Cross-sector

Question	Answer
*Will the proposal, **once completed** be useful,*	Provided that the activity will turn out

---

[33] (1) EU Agenda on Security : COM (2016) 205 "Stronger and Smarter Information Systems for Border and Security", (2) Communication from the Commission to the European Parliament and the Council: Overview of information management in the area of freedom, security and justice.URL: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0385:FIN:EN:PDF, (3) Regulation of the European Parliament and of the Council establishing a European Travel Information and Authorisation System (ETIAS) and amending Regulations (EU) No 515/2014, (4) DIRECTIVE (EU) 2016/681 of the European Parliament and of the Council of 27 April 2016
on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime. URL: http://eur-lex.europa.eu/eli/dir/2016/681/oj

Question	Answer
from the interoperability point of view and utilised in two (2) or more EU policy sectors? Detail your answer for each of the concerned sectors.	successfully it will pave the way towards a set of common meta data and component access standards for processing open source information. Cross-sectoral interoperability aspect is embraced through diversifying the range of end-user organisation participating in the activity. Finally, once developed, the common standards will also indirectly contribute to the improvement of the interoperability of EU information systems across various domains (law enforcement, customs, border control, migration, etc.), in whose context exploitation and sharing of open source information, whether processed or in raw format, is deemed to play an increasing role.

2.4.5.3 Cross-border

Question	Answer
*Will the proposal, **once completed,** be useful from the interoperability point of view and used by public administrations of three (3) or more EU Members States? Detail your answer for each of the concerned Member State.*	1) Administration to Administration: The new set of standards will allow re-using existing processing components. For example processing of text for tax administration purposes can be shared across European MS authorities. 2) Administration to Business: A set of standards facilitates the procurement of bespoke or off-the-shelf components from commercial providers. Again with the ability to be shared between authorities.

2.4.5.4 Urgency

Question	Answer
Is your action urgent? Is its implementation foreseen in an EU policy as priority, or in EU legislation?	The action as such is not urgent. However, since many MS authorities are currently investing or plan to invest into solutions for open source processing the action proposal has higher impact if implemented now than at a later stage. While there is no specific EU policy related to the exploitation of open source information, it has been widely acknowledged across different domains and in various countries that open source information often constitutes relevant complementary information in the decision making processes of whatever kind, or it constitutes the only available information on certain topic/entity etc. Many organisations, which started or plan to exploit open sources as a crucial source of information, stumble across both budgetary and technical issues (including, i.e., interoperability problems) that prevent them from taking the full advantage of open source information. Thus, creation of common standards and a pool of freely available components for assembling open source information processing pipelines would alleviate the situation enormously. Furthermore, it would allow to immediately benefiting those organisations, which are at an early stage of considering exploitation of open source information.
How does the ISA^2 scope and financial capacity better fit for the implementation of the proposal as opposed to other identified and currently available sources?	This project matches the ISA^2 goals. No other sources envisage funding of such an activity.

2.4.5.5 Reusability of action's outputs

Name of reusable solution to be produced (for new proposals) or produced (for existing actions)	Registry of available processing components
Description	A description of available processing components which are commercial, open source or available from MS authorities for sharing
Reference	
Target release date / Status	Q2 / 2018
Critical part of target user base	Public services in Member States and EU organisations

Name of reusable solution to be produced (for new proposals) or produced (for existing actions)	Existing Meta Data Standards to be adopted for open source information processing applications
Description	Report of existing meta data standards and missing standards for definition
Reference	
Target release date / Status	Q2 / 2018
Critical part of target user base	Public services in Member States and EU organisations

Name of reusable solution to be produced (for new proposals) or produced (for existing actions)	Meta Data and Interoperable Components for Open Source Information Processing
Description	Report with project result: Adopted and newly designed standards, component access guide lines
Reference	
Target release date / Status	Q3 / 2019
Critical part of target user base	Public services in Member States and EU organisations

Name of reusable solution to be produced (for new proposals) or produced (for existing actions)	Core Components for Open Source Information Processing
Description	Pool of Runnable software components, with documentation and source code repository
Reference	

Target release date / Status	Q4/2019
Critical part of target user base	Public services in Member States and EU organisations

2.4.5.6 Level of reuse of existing solutions

Question	Answer
Does the proposal intend to make use of any ISA², ISA or other relevant interoperability solution(s)? Which ones?	The proposal aims to make use of the work done in ISA's Core Data and PM² solutions.

2.4.5.7 Interlinked

Question	Answer
Does the proposal directly contribute to at least one of the Union's high political priorities such as the DSM? If yes, which ones? What is the level of contribution?	By the end user community it addresses it contributes directly to the Security Union as part of the Justice and Fundamental Rights priority.

2.4.6 PROBLEM STATEMENT

The problem of	Fragmented, non-coordinated, and redundant efforts across sectors and borders in the context development of IT tools for processing open source information by public organisations
affects	Technical interoperability and reuse of existing solutions.
the impact of which is	Increased Time needed by an organisation to develop and introduce new software functionality for processing open source information
a successful solution would be	Definition of meta data standards to facilitate the interchange of software components in the field of open source information processing and establishing guidelines to adapt existing tools

	and resources and assemble them into composite applications. Creation of a pool of generic and freely available interchangeable software components compliant with the standards.

The problem of	Incompatibility of meta data standards and interfaces in the context of IT tools for processing open source information
affects	Capacity to reuse of existing open source information and relevant resources created by other institutions
the impact of which is	Knowledge bases and resources created by one authority cannot be reused by authorities in other Member States without substantial effort
a successful solution would be	Definition of meta data standards to facilitate information access and exchange in the field of open source information processing

2.4.7 IMPACT OF THE ACTION

2.4.7.1 Main impact list

The beneficiaries of the project are MS authorities and EU/international organisations which use open source information for their daily tasks. Especially users in law enforcement, customs, finances, public health are primary beneficiaries.

Impact	Why will this impact occur?	By when?	Beneficiaries
(+) Savings in money	The availability of freely available core components for assembling open source information processing pipeline will reduce the expenditures	Q1/2020 onwards	Member States and EU/internation al organisations
(+) Savings in time	The availability of freely available core components and guidelines for assembling open source	Q1/2020 onwards	Member States and EU/internation al

	information processing pipeline will speed up the development process and potentially eliminate some procurements on the end-user side.		organisations
(+) Improved cross-border and cross-sector exploitation and reuse of existing open source information	Through introduction of common meta data standards and corresponding APIs access and sharing of information will be easier	Q1/2020 onwards	Member States and EU/internation al organisations
(+) Improved interoperability at EU level	Due to consideration in the action meta data formats of the EU-level It systems conversion and integration of open source information in the related workflows will be easier	Q1/2020 onwards	Member States and EU/internation al organisations

2.4.7.2 User-centricity

The action will be strictly end-user driven, in particular, one will aim at involving possibly high diversity of end users with respect to different sectors and countries being involved. Information of the specific workflows related to processing open source information by end user will be collected at a very early stage of the project in order to best embrace end user needs in the scope of the planned activities and developments. A network of end-users will be established in order to safeguard end-user interests and sustainability of the to-be-developed deliverables and steer potential future developments.

2.4.8 EXPECTED MAJOR OUTPUTS

Please see major outputs already listed in 1.1.5.5.

Output Name	Pilot Application
Description	Pilot composite application using core components for open source information processing
Reference	
Target Release Date / Status	Q4/2019

2.4.9 ORGANISATIONAL APPROACH

2.4.9.1 Expected stakeholders and their representatives

The JRC has created a community of practice for Open Source Information exploitation with yearly meetings. The community is comprised of MS authorities, EU institutions and international organisations. The community as such will be brought in as a stakeholder. The following organisations have explicitly expressed support for the action:

Stakeholders	Representatives	Involvement in the action
Authority for Consumers & Markets, The Netherlands	Remco Siderius	Provision of expertise in analysis, design and evaluation of the deliverables
Financial Investigation Service of the Tax Authority, The Netherlands	G.H. De Grutter	Provision of expertise in analysis, design and evaluation of the deliverables
Tax Authority, Denmark	Michael Krogh Jacobsen	Provision of expertise in analysis, design and evaluation of the deliverables
HS Leiden, Forensics Institute	Jos Griffioen	Provision of expertise in analysis, design and evaluation of the deliverables
Swedish Tax Agency	Joanna Kozakiewicz	Provision of expertise in analysis, design and evaluation of the deliverables
Dutch Customs Administration	Liesbeth Kremer	Provision of expertise in analysis, design and evaluation of the deliverables
Joint Research Centre	Gerhard Wagner	Provision of expertise in open source information processing tools and standards development. Action management

Furthermore, the following organisations have voiced interest and will most likely join:
Swedish Tax Administration, Police Slovenia, Europol

2.4.9.2 Identified user groups

- Member State public service organisations (law enforcement, customs, tax, public health, etc.)

- EU Communities of Practice (e.g., ENLETS – European Network for Law Enforcement Technologies and Services)
- EU Institutions (COMM DGs, EU Agencies, etc.)
- International Institutions (e.g., International Criminal Court, IAEA)

2.4.9.3 Communication and dissemination plan

Both online and offline communication channels will be used. A web-based information sharing platform will be established to:

a) report on the action progress,
b) disseminate information on the deliverables,
c) gather end-user feedback,
d) facilitate information exchange between the different stakeholders involved.

At the end of the project a workshop to present the outcomes will be organised. This workshop will be used to set up a post-action dissemination plan to inform other relevant communities will be elaborated with the participating end-users. Furthermore, on-site trainings and workshops to MS authorities will be organised. Finally, ISA² Member States-network will be exploited for action result dissemination.

2.4.9.4 Key Performance indicators

Description of the KPI	Target to achieve	Expected delivery (months after kick-off)
Level of end-user interest with respect to the definition of the project	Participation of at least 15 experts from different authorities and countries to define the scope of the project	M +1
Level of end-user satisfaction with the respect to the report on "Overview Existing Meta Data Standards"	An average level of 4 in a scale from 1 (not satisfied) to 5 (very satisfied) in an end-user survey	M +6
Level of end-user satisfaction with the respect to the report on "New Meta Data Standards and Component Access Defined "	An average level of 4 in a scale from 1 (not satisfied) to 5 (very satisfied) in an end-user survey	M +12
Level of readiness of the Core Components developed	At least 5 MS authorities have adopted the deliverables for operational work	M +18
Level of interest of end-user community in the events dedicated to the dissemination of the outcomes of the activity	At least 20 experts from 5 different sectors/countries participating in the result dissemination events	M +24

2.4.9.5 Governance approach

The project management board will be composed of the Head of the Text and Data Mining Unit of the Competencies Directorate of DG JRC, project manager Text and Data Mining Unit of the Competencies Directorate of DG JRC, and one representative from each participating Member States or other EU/international organisation. The board will meet 3 times during the execution of the project (at the beginning, at the end of 2018 and at the closing). Additional meetings could be organised if deemed necessary. An electronic web-based platform for monitoring the progress of the project and facilitation of the communication of the project management board will be put in place.

Since the continuous participation of the end-user community is crucial for accomplishing the goals of the project a pool of reserve end-user organisations will be maintained in case of unexpected resignation of the partners that agreed to participate in the project. This pool will be centred on the Member State expert OSINT community created circa 10 years ago by DG JRC.

2.4.10 TECHNICAL APPROACH AND CURRENT STATUS

The project is structured into separate phases. After project initialisation with the main goal of setting up a stakeholder group, the main project phase is an iterative design and implementation phase to minimize risks and optimizes results with immediate stakeholder feedback. Finally, in the Closing/Evaluation phase the project results are presented and limited on-site trainings are provided to interested MS authorities. The various phases are detailed below.

1. **Initialisation Phase**: Take stock of state of play, form stakeholder group
   a. Take stock of currently used open sources information processing workflows to define state of play (data standards and software used) and best practices
   b. Create a stakeholder group of interested MS authorities, research and education institutions and EU partners willing to participate, give feedback and test the results in practice
2. **Definition Phase**: Define scope of project with stakeholders
   a. Create a registry of used existing components which can potentially be shared or adapted for interoperability
   a. Create a list of already used meta data standards
   b. Create a list of missing standards and components with the greatest potential for reuse
3. **Execution Phase**: Design and Implementation
   a. Define data formats and component access (see Annex 1 for detailed description)
      i. Analyse existing meta data standards to be adopted or amended, find missing ones (gap analysis)[34].
      ii. Define missing meta data formats for component interoperability
      iii. Define guidelines on how to adopt existing or newly developed software components to be interchangeable[35].

---

[34] There are already several meta data standards, most notably the European Commission's ISA Core Vocabularies (see https://ec.europa.eu/isa2/solutions/core-vocabularies_en) which can be adopted as underlying basis.

  b. Implement core components and pilot application
     i. Develop core set of software components (proof of concept) for open source information processing to be shared with MS authorities and EU institutions[36]
     ii. Develop a best practice pilot application[37] which demonstrates how the core software components can be assembled into a composite application[38]
  c. Gather feedback from stakeholders to minimize risks and maximise applicability of results
4. Closing/Final Evaluation Phase:
   a. Present project results to interested MS authorities and EU institutions at a workshop
   b. Disseminate project results and carry an evaluation thereof through on-site trainings/workshops to/for MS authorities

It is important to emphasize that the Definition and Execution phase will also encompass studying the latest developments in the context of EU-level security-related IT systems[39] and existing EU customs-related information exchange platforms (e.g., CIS, FIDE)[40] in order to safeguard interoperability therewith whenever applicable. In addition, recently introduced and future envisaged data exchange formats at EU level for sharing security-related information, e.g., Passenger Name Record (PNR)[41] or European Travel Information and Authorisation System (ETIAS)[42] records, would also be considered if deemed relevant.

2.4.11 COSTS AND MILESTONES

2.4.11.1 Breakdown of anticipated costs and related milestones

Phase: Initiation Planning Execution Closing/Final evaluation	Description of milestones reached or to be reached	Anticipated Allocations (KEUR)	Budget line ISA/ others (specify)	Start date (QX/YYYY)	End date (QX/YYYY)
Initiation and	Announcement, Forming Stakeholder Group, Kick-	10	ISA	Q1 / 2018	Q2 / 2018

---

[35] As far as possible reuse of existing approaches, such as REST-like interfaces for loosely coupled internet applications.

36 The JRC has already developed certain components which can be shared.

37 The exact subject of the pilot application is identified by the involved stakeholders.

38 The pilot application might not be shared with the public if it deals with requirements of law enforcement or other sensible topics.

[39] Communication from the Commission to the European Parliament and Council. COM (2016) 205. "Stronger and Smarter Information Systems for Borders and Security"
http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52016DC0205

[40] Communication from the Commission to the European Parliament and the Council:
Overview of information management in the area of freedom, security and justice.
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0385:FIN:EN:PDF

[41] DIRECTIVE (EU) 2016/681 of the European Parliament and of the Council of 27 April 2016 on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime. http://eur-lex.europa.eu/eli/dir/2016/681/oj

[42] Regulation of the European Parliament and of the Council establishing a European Travel Information and Authorisation System (ETIAS) and amending Regulations (EU) No 515/2014

planning	off				
Planning	Project Scope Defined	10	ISA	Q1 / 2018	Q2 / 2018
Execution	Management and Supervision	20	DG JRC	Q1 / 2018	Q4 / 2019
Execution	Review Existing Meta Data Standards	25	ISA	Q2 / 2018	Q2 /2018
Execution	New Meta Data Standards and Component Access Defined	100	ISA	Q3 / 2018	Q4 / 2018
Execution	Core Components designed and developed	85	ISA	Q1 / 2019	Q2 / 2019
Execution	Pilot application developed and tested	40	ISA	Q3 / 2019	Q3 / 2019
Final Evaluation	Result Presentation	15	ISA	Q4 / 2019	Q4 / 2019
Closing/Final Evaluation	On-Site Trainings and Result Dissemination	15	ISA	Q4 / 2019	Q4 / 2019
	Sums	**ISA: 300** JRC: 20 Total: 320			

2.4.11.2 Breakdown of ISA^2 funding per budget year

Budget Year	Phase	Anticipated allocations (in KEUR)	Executed budget (in KEUR)
2018	Initiation and planning	20	
2018	Execution	125	
2019	Execution	125	
2019	Closing/Final Evaluation	30	

2.4.12 ANNEX AND REFERENCES

Description	Reference link	Attached document
Letter of Support from Stakeholders		• Letter of Support, NL Tax • Letter of Support, DK Tax • Letter of Support, NL

		Consumer Market Authority
		• Letter of Support, NL HS Leiden Forensics Institute
		• Letter of Support Swedish Tax Agency
		• Letter of Support Dutch Customs