PROJECT Nº HOME/2012/ISEC/AG/4000004374

D2.1 System Requirements Report

Start Date of Project :   01/01/2014   Duration :   24 months

PROJECT FUNDED BY THE EC DIRECTORATE GENERAL FOR HOME AFFAIRS	
Due date of deliverable	M6
Actual submission date	June 2014 (M6)
Lead partner	MCC
Participant(s)	HIB
Work package	WP2
Classification	PU
Version	1.2
Total number of pages	25

DISCLAIMER

The work associated with this report has been carried out in accordance with the highest technical standards and SAMi2 partners have endeavored to achieve the degree of accuracy and reliability appropriate to the work in question. However since the partners have no control over the use to which the information contained within the report is to be put by any other party, any other such party shall be deemed to have satisfied itself as to the suitability and reliability of the information in relation to any particular use, purpose or application.

Under no circumstances will any of the partners, their servants, employees or agents accept any liability whatsoever arising out of any error or inaccuracy contained in this report (or any further consolidation, summary, publication or dissemination of the information contained within this report) and/or the connected work and disclaim all liability for any loss, damage, expenses, claims or infringement of third party rights.

List of Authors

Partner	Authors
HIB	Elena Muelas Cano, Paloma Jimeno Sánchez-Patón, Raúl Santos de la Cámara

Table of contents

List of Tables

© SaMi2 Consortium

Glossary

Acronym	Meaning
DoW	Document of Work
GUI	Graphical User Interface
NLP	Natural Language Processing
OSN	Online Social Network
SOA	Service-Oriented Architecture
WP	Work Package

References

1. **University, Middlesex.** *D7.1 Ethical and Legal Requirements.*

2. **HI-Iberia.** *D2.3 State of the art.*

3. **Consortium, SAMI2.** *SAMI2 Document of Work.*

1 Executive Summary

The present document aims to identify the requirements that best respond to project needs and that ensure the fulfillment of commitments from all partners involved in the project.

The objective of SAMI2 project is to fight against illegal content and crime, gathering data from OSNs and processing them with the adequate methodology to ensure the maintenance of the custody chain for their future forensic use. In order to meet these goals, 2 types of requirements will be established: technical and based on end-users. Ethical and legal requirements will be described in deliverable D7.1 (1) by Middlesex. All these requirements will result in the different features and services of the system that will be fulfilled using the current State of the Art technologies (see deliverable D2.3 (2)).

Thus, this deliverable will settle the basis that will scope and give form to SAMI2 and that will serve as input of the WP3 (design) and the WP6 (trials and testing), where a validation plan will be created based on this requirements document.

2 Objectives

The success of the requirements gathering depends on two things: stakeholders who can explain their needs and experts who can determine their feasibility. This requires:

- A close relationship with stakeholders and partners

- Effective requirements management

The first objective is achieved maintaining a special relationship with the stakeholders and employing the requirement techniques that best suit with their needs.

The second objective tries to avoid improvisation. The product cannot be built without understanding its nature and the role that each module will play in it. Thus, an organized and well-structured Work Package is defined in the Description of Work (DoW) (3). This Work Package, number 2, is in charge of user requirements and it is based on use case definitions and the current State of the Art.

In addition, the collection and management of requirements have taken into account the best practices of requirements engineering such as collecting them using a consistent format and ensuring that all parameters of the collection process (e.g., unique identification of requirements, ownership, and tracking of the evolution of the requirements) are made using a traceable process.

Finally, in order to ensure the effectiveness and feasibility of the requirements, police and other validation organisms will certify them in one of all the meetings that will be celebrated along the project life cycle.

Summing up, the main objective of this document is an adequate formulation and validation of requirements. This is very important because will add value to the product and reduce the deployment and maintenance costs.

2.1 Scope and methodology

Requirements in this document will be divided in 2 types depending of their domain of interest: technical and end-user oriented. This deliverable will not contain detailed definitions at technical level but high level functional requirements. According to that, for each of these requirements, a table containing the requirement definition and traceability parameters such as name, description, rationale and partner involved will be provided.

The first set of requirements has an important role due to it will indicate the parameters and technical constraints that will have to take into account to fulfill end-users needs. These results will feed directly into the core platform work of WP3 for the design of the SAMi2 solution.

The second set of requirements is also very important. It will form the core of SAMI2 and will ensure the success of our product, adding value against other possible competitors. End-users will be involved during all project development and adequate methods will be carried out to ensure that their needs are gathered in a correct form. The end user requirements will be crucial to determine the design of the data collection and processing features in WP3.

It is also important to mention that, a set of legal and ethical requirements that SAMI2 will have to meet to avoid their future in validation, will be described in WP7.

As a conclusion, many of those listed features have to be parsed in exact way by close-up analysis. At this stage also the feedback from other partners will be highly important, because the information received from them clarifies the real intentions and development needs in this work package. Finally, WP3 will make use of these requirements using them as guidelines for the product design.

3 Requirements

3.1 Technical Requirements

This section contains the requirements that need to be accomplished to deliver a product, service or result with the specified features and functions. It will be based on technical limits.

3.1.1 Scope

Crawlers will be able to offer analysis and correlation of heterogeneous and freely accessible data from OSNs in order to search and extract appropriate information. They go across interested web pages and download, parse and process them. System will cover all OSNs where crawlers are available.

RQ_WP21-01	RQ_NAME **Crawlers specifications**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMI2 will support analyzing the following OSNs:** - **Twitter (must)** - **Facebook (could)** - **Google+ (could)**		RQ_RATIONALE The challenge is to cover the higher number of OSNs as possible.
RQ_PRIORITY Medium		RQ_ASSESSMENT N/A

Table 1 Crawlers specifications

Compatibility and interoperability will ensure the integration of the system into already running systems. Compatibility is related to integration; in order to exchange information and to access to systems using common interfaces, systems are needed to be compatible.

RQ_WP21-02	RQ_NAME **Compatibility with already running systems**	RQ_PARTNER HIB
RQ_DESCRIPTION **N/A**		RQ_RATIONALE The challenge is to ensure the compatibility with already running systems easing the system integration

RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 2 Compatibility with already running systems

RQ_WP21-03	RQ_NAME **Storage formats**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMI2 must be developed to aid software with processing of XML and JSON data.**		RQ_RATIONALE Ensuring the compatibility and interoperability between systems with the aim of sharing information in a secure, reliable and easy way.
RQ_PRIORITY Medium		RQ_ASSESSMENT N/A

Table 3 Storage formats

RQ_WP21-04	RQ_NAME **Scalable architecture**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMi2 architecture should be easy to grow if needed, improving its performance after adding hardware, proportionally to the capacity added or better.**		RQ_RATIONALE The aim is to accept increased volume of data without impacting in the system performance
RQ_PRIORITY Medium		RQ_ASSESSMENT N/A

Table 4 Scalable architecture

Semantic technologies require a great deal of computing power which scales linearly with the data corpora. For this reason, mitigation strategies will be developed ensuring information processing as fast as possible.

RQ_WP21-05	RQ_NAME **Information processing near to real-time**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMI2 analysis must work in near real-time. Analysis of crawling OSNs should be ready for querying in less than 1 minute**		RQ_RATIONALE The challenge is to ensure an information processing as quick as possible to take appropriate reaction activities

RQ_PRIORITY	RQ_ASSESSMENT
Medium	N/A

Table 5 Information processing near to real-time

Well-documented, open, standardized and extensible interfaces will be developed to connect the extracted information with the systems used by the security and safety institutions.

RQ_WP21-06	RQ_NAME **Well-documented standardized interfaces**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMi2 must develop well-documented, open, standardized and extensible interfaces in modular environments**		RQ_RATIONALE **The correct functioning of SAMI2 when is deployed in a modular environment**
RQ_PRIORITY Medium		RQ_ASSESSMENT N/A

Table 6 Well-documented standardized interfaces

RQ_WP21-07	RQ_NAME **Graphical User Interface (GUI)**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMi2 should incorporate an intuitive Graphical User Interface (GUI) when works as standalone application**		RQ_RATIONALE The challenge is to provide an easy and intuitive platform that the user can interact with.
RQ_PRIORITY Medium		RQ_ASSESSMENT N/A

Table 7 Graphical User Interface

RQ_WP21-08	RQ_NAME **Open Source software**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMi2 should make use of Open Source software in the NLP tools, search engines and programs to access the DBs**		RQ_RATIONALE The challenge is to ensure the efficiency and robustness of the SAMI2 product reducing costs and bureaucracy as much as possible

RQ_PRIORITY	RQ_ASSESSMENT
Medium	N/A

Table 8 Open Source software

Software able to run on two or more platforms and to provide the similar logical functionality as well as similar look and feel is needed.

RQ_WP21-09	RQ_NAME	RQ_PARTNER
	Cross-platform software	HIB
RQ_DESCRIPTION		RQ_RATIONALE
The software must be developed so that portability across platforms is maximized.		The challenge is to carry out a deep understanding of the required functionalities and capabilities of the different platforms to provide this requirement. The usage of virtualization at the middleware layer (e.g., Java Virtual Machine) or the OS layer (e.g., hypervisors) will be explored.
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 9 Cross-platform software

Text to be processed will be found in different text formats. Search-engines needed to search and organize these texts must be independent of the file format, ensuring the extensibility of the product.

RQ_WP21-10	RQ_NAME	RQ_PARTNER
	Analysis tools independent of the text format	HIB
RQ_DESCRIPTION		RQ_RATIONALE
SAMi2 analysis tools must be able to process text data independently of the format of the file.		The challenge is to cover the higher amount of text formats as possible. An intermediate SAMi2 specific format will be used to store disparate sources under s similar paradigm.
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 10 Analysis tools independent of text format

RQ_WP21-11	RQ_NAME	RQ_PARTNER
	Decentralized Security	HIB
RQ_DESCRIPTION		RQ_RATIONALE

The SAMi2 system should be secure to unintended users. If a security breach occurs, isolation of the rest of the components of the system should be carried out.	The information gathered by the SAMi2 system has the potential of being quite sensitive. Information security measures will be included in the design. Adaptive capacity of the secure methods to detect, early, security breaches
RQ_PRIORITY	RQ_ASSESSMENT
Medium	N/A

Table 11 Decentralized Security

RQ_WP21-12	RQ_NAME **Security and Safety Ontologies**	RQ_PARTNER HIB
RQ_DESCRIPTION		RQ_RATIONALE
SAMI2 must include ontologies related to social networks, safety and security or equivalents.		The challenge is to cover as many ontologies as possible to use with the SAMi2 semantic backend.
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 12 Security and Safety Ontologies

RQ_WP21-13	RQ_NAME **Cryptographic structures**	RQ_PARTNER HIB
RQ_DESCRIPTION		RQ_RATIONALE
SAMi2 must contain cryptographic structures able to provide tamper resistance and unique identification		The challenge is to achieve a strong encryption for critical data (e.g., information that is meant for forensic purposes).
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 13 Cryptographic structures

RQ_WP21-14	RQ_NAME **Logging tools**	RQ_PARTNER HIB
RQ_DESCRIPTION		RQ_RATIONALE

SAMi2 must have authentication tools. User name and password must be required to access to the activities related to action procedures		Secure access is required as the system will contain information for the safety providers' eyes only.
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 14 Logging tools

RQ_WP21-15	RQ_NAME **Backup snapshots**	RQ_PARTNER HIB
RQ_DESCRIPTION **Backup snapshots of all illegal detected activities must be carried out every 12 hours including traceable proof to the original OSN contents.**		RQ_RATIONALE It will allow having evidences of possible crimes although the information has been deleted by authors.
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 15 Backup snapshots

RQ_WP21-16	RQ_NAME **Parallel Processing**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMi2 must incorporate techniques for Big Data analysis. This includes extensive parallel processing and analytics algorithms like Map-Reduce**		RQ_RATIONALE The challenge is to provide timely end-to-end processing
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 16 Parallel processing

RQ_WP21-17	RQ_NAME **Availability**	RQ_PARTNER HIB
RQ_DESCRIPTION **The service must be available 7 days X 24 hours and have backup with contingency solutions**		RQ_RATIONALE The challenge is to provide a seamless service that is not interrupted by any single point of failure.

RQ_PRIORITY	RQ_ASSESSMENT
Low	N/A

Table 17 Availability

3.1.2 Analysis

This section will contain the requirements that will specify how to analyze the data.

Natural Language Processing Tools will allow the syntactic and lexical analysis of the texts extracted trough the "crawler". It is needed that the software used to parse and process these texts be available for English and Spanish. For this task, the support of corpus of reference to train Spanish and English tagger is required.

RQ_WP21-18	RQ_NAME **NLP software available for English and Spanish**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMI2 must cover the main languages involved in the project (English and Spanish) allowing the parsing and processing of the web texts gathered by the "crawler"**		RQ_RATIONALE The challenge is to cover the higher number of languages as possible increasing the applicability of the project.
RQ_PRIORITY Medium		RQ_ASSESSMENT N/A

Table 18 NLP sw available for English and Spanish

NLP tools able to be trained with new data is required

RQ_WP21-19	RQ_NAME **NLP tools with adaptive training based on changing inputs**	RQ_PARTNER HIB
RQ_DESCRIPTION **The corpora of data used to train NLP methods may evolve over time so the training should be able to adapt to these changes.**		RQ_RATIONALE The challenge is to ensure adaptive training based on changing input data
RQ_PRIORITY Medium		RQ_ASSESSMENT N/A

Table 19 NLP tools with adaptive training

Distributed File System will be used to storage data gathered from OSNs. In order to ensure the speed and security of the access and processing of data, a data partition and clustering algorithm will be used for the file system management. In this algorithm, data will be replicated in different machines with the aim of reducing the impact of a machine power outage, data loss or switch failure.

RQ_WP21-20	RQ_NAME	RQ_PARTNER
	Data integrity	HIB
RQ_DESCRIPTION		RQ_RATIONALE
SAMi2 must ensure data integrity in databases.		The aim is to protect databases from outside attacks with tools like firewalls, antivirus software and periodic scans for malicious code
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 20 Data integrity

3.1.3 Deployment

This section will contain the requirements that need to be accomplished in the real-life deployments of SAMI2.

RQ_WP21-21	RQ_NAME	RQ_PARTNER
	Integration with legacy systems	HIB
RQ_DESCRIPTION		RQ_RATIONALE
SAMi2 should enable integration with legacy information management systems already in use including databases and data analysis tools.		Simplify and accelerate the introduction of Big Data technology
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 21 Integration with legacy systems

SAMI2 must be used as standalone application, through a Graphical User Interface (GUI) of its own.

RQ_WP21-22	RQ_NAME	RQ_PARTNER
	Standalone application	HIB
RQ_DESCRIPTION		RQ_RATIONALE

SAMI2 must have the possibility of being used as standalone module, e.g. not needing to be integrated in another module to work	The challenge is to create a module as independent as possible
RQ_PRIORITY	RQ_ASSESSMENT
Medium	N/A

Table 22 Standalone application

SAMI2 must be able to be used as modular application component being integrated with other tools

RQ_WP21-23	RQ_NAME **Modular application**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMI2 must be able to be used as a module connected to other tools**		RQ_RATIONALE The challenge is to design SAMI2 as a module easy to integrate with other tools
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 23 Modular application

RQ_WP21-24	RQ_NAME **Service-oriented architecture (SOA)**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMI2 must use a SOA software architecture design pattern based on distinct pieces of software providing application functionality**		RQ_RATIONALE A service that is well-defined, self-contained and does not depend on the context or state of other services
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 24 SOA service

3.1.4 Standards

This section will contain the requirements related to the use of standards in SAMI2.

RQ_WP21-25	RQ_NAME **RDF**	RQ_PARTNER HIB
RQ_DESCRIPTION		RQ_RATIONALE

SAMI2 semantics must be expressed in standard languages RDF and OWL.	The conceptual description or modeling of information implemented in web resources (semantic web) will be expressed using well-known standards for ease of reuse and interoperability.
RQ_PRIORITY	RQ_ASSESSMENT
High	N/A

Table 25 RDF

RQ_WP21-26	RQ_NAME **SPARQL**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMI2 must be able to accept SPARQL queries to retrieve semantic data.**		RQ_RATIONALE Semantic data cannot be queried using SQL or similar approaches. SPARQL is a W3C sanctioned standard for semantic information retrieval.
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 26 SPARQL

RQ_WP21-27	RQ_NAME **REST**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMI2 must provide external functionality by means of HTTP interfaces using the REST paradigm.**		RQ_RATIONALE REST is a de-facto standard for web services across the Internet.
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 27 REST

RQ_WP21-28	RQ_NAME **JSON**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMI2 must provide data on its external interfaces using JSON.**		RQ_RATIONALE JSON is an open de facto standard format that uses human-readable text to transmit data objects consisting of attribute-value pairs.
RQ_PRIORITY		RQ_ASSESSMENT

| Medium | N/A |

Table 28 JSON

RQ_WP21-29	RQ_NAME **XML**	RQ_PARTNER HIB
RQ_DESCRIPTION **SAMI2 could provide data on its external interfaces using XML.**		RQ_RATIONALE XML is a more formal alternative to JSON but it is also much more verbose. Thus JSON will be the preferred format although XML outputs might be used.
RQ_PRIORITY Medium		RQ_ASSESSMENT N/A

Table 29 XML

3.2 End-user Requirements

3.2.1 Keep the custody chain

The forensic module must ensure that the custody chain is respected at every moment. For this reason, special procedures in the gathering, storage and delivery of the evidences.

RQ_WP21-30	RQ_NAME **Respect of the custody chain**	RQ_PARTNER MCC
RQ_DESCRIPTION **SAMi2 must use standard evidence processing to keep the custody chain through the correct storage and information handling procedures in order to validate the information gathered**		RQ_RATIONALE The challenge is to avoid that the gathered evidences become invalid due to an incorrect procedure when keeping the custody chain.
RQ_PRIORITY High		RQ_ASSESSMENT N/A

Table 30 Custody Chain

RQ_WP21-31	RQ_NAME **Confidentiality**	RQ_PARTNER MCC
RQ_DESCRIPTION		RQ_RATIONALE

All persons involved in the project should be classified	The challenge is to ensure that the project information is only accessed by authorized personnel
RQ_PRIORITY	RQ_ASSESSMENT
High	N/A

Table 31 Confidentiality

RQ_WP21-32	RQ_NAME **Cross-line searching**	RQ_PARTNER MCC
RQ_DESCRIPTION **Searching must be carried out in concordance with other OSNs**		RQ_RATIONALE The challenge is to cross the data obtained in several OSNs
RQ_PRIORITY		RQ_ASSESSMENT
Medium		N/A

Table 32 Cross-line searching

RQ_WP21-33	RQ_NAME **Dynamic searching**	RQ_PARTNER MCC
RQ_DESCRIPTION **Searching must be able to adapt to changing hashtags around a same topic**		RQ_RATIONALE The challenge is to be able to follow messages related to a topic, even when hashtags are slightly modified or totally changed
RQ_PRIORITY		RQ_ASSESSMENT
High		N/A

Table 33 Dynamic searching

RQ_WP21-34	RQ_NAME **Easy Graphical User Interface**	RQ_PARTNER MCC
RQ_DESCRIPTION **Graphical User Interface should be intuitive and easy to use**		RQ_RATIONALE The challenge is to avoid that the User Interface entails a barrier for the use of the system
RQ_PRIORITY		RQ_ASSESSMENT

Medium	N/A

Table 34 Easy GUI

RQ_WP21-35	RQ_NAME **Natural Image Processing**	RQ_PARTNER MCC
RQ_DESCRIPTION **SAMI2 should be able to analyze images in the same mode that text is processed**		RQ_RATIONALE The challenge is to be able to extract information about images like similarities, where they were taken…
RQ_PRIORITY Medium		RQ_ASSESSMENT N/A

Table 35 Natural Image Processing

RQ_WP21-36	RQ_NAME **Intern servers**	RQ_PARTNER MCC
RQ_DESCRIPTION **Servers should be located intern to police**		RQ_RATIONALE The challenge is to be able to settle server in the police perimeter to protect confidential information
RQ_PRIORITY High		RQ_ASSESSMENT N/A

Table 36 Intern servers

RQ_WP21-37	RQ_NAME **Arabian analysis**	RQ_PARTNER MCC
RQ_DESCRIPTION **SAMI2 must be able to detect illegal content or activities in messages written in Arabian**		RQ_RATIONALE The challenge is to be able to detect terrorism actions or illegal activities by part of jihadist groups
RQ_PRIORITY Medium		RQ_ASSESSMENT N/A

Table 37 Arabian analysis

4 Conclusions

This deliverable has contained the most important requirements of SAMI2 platform, both technical and from end-users. Its objective has been to obtain an abstract high-level statement of the specifications that SAMI2 must provide, in a structured way, to meet with its main services or functionalities.

In this document, both functional and non-functional requirements have been specified. The first ones determine the objectives and services that the system should fulfill, i.e. what the system should do. Without them, the system quality and performance will decrease. The second ones look for judging the system operation, they do not define actions to carry out. Without them, the system could not work.

This document serves as first step for the use case creation, for the later design of the system and finally, to validate it.