



D2.1 DATA LAKE INTEGRATION PLAN

Deliverable No.:	D2.1
Deliverable Title:	Data lake integration plan
Project Acronym:	Fandango
Project Full Title:	FAke News discovery and propagation from big Data and Artificial iNtelliGence Operations
Grant Agreement No.:	780355
Work Package No.:	2
Work Package Name:	Data Access, Interoperability and user requirements
Responsible Author(s):	Sindice (Lead), ENG, LIVETECH, VRT, CERTH, CIVIO, UPM, ANSA
Date:	31.07.2018
Status:	V1.1
Deliverable type:	REPORT
Distribution:	PUBLIC

REVISION HISTORY

VERSION	DATE	MODIFIED BY	COMMENTS
V0.1	07.05.2018	Jeferson Zanim (Siren/Sindice)	First draft.
V0.2	15.05.2018	Jeferson Zanim (Siren/Sindice)	Content structure adjusts and first integration addition.
V0.3	22.05.2018	Theodoros Semertzidis (CERTH)	CERTH contributions.
V0.4	07.06.2018	Jeferson Zanim (Siren/Sindice)	Merged UPM contributions to the document.
V0.5	14.06.2018	Jeferson Zanim (Siren/Sindice)	Merged LvT contributions to the document.
V1.0	30.07.2018	Monica Franceschini, Massimo Magaldi (ENG)	Internally reviewed version
V1.1	03.08.2018	Jeferson Zanim (Siren/Sindice)	Final Version

TABLE OF CONTENTS

1.	Introduction.....	7
2.	Architecture Overview	7
3.	Data Integrations.....	9
3.1.	Data Ingestion	9
3.1.1.	Data Ingestion I – Partner’s Data (FTP).....	10
3.1.2.	Data Ingestion II – Rest APIs.....	11
3.1.3.	Data Ingestion III – Open Data.....	12
3.1.4.	Data Ingestion IV – Websites.....	13
3.1.5.	Data Loading V – RSS Sites	14
3.1.6.	Data Ingestion VI – Social Networks.....	14
3.2.	Data Processing	16
3.2.1.	Siren (Sindice) Data Processing Integrations	16
3.2.1.1.	Siren Integration I – Siren Investigate	16
3.2.2.	UPM Integrations	18
3.2.2.1.	UPM Integration I – Spark	18
3.2.2.2.	UPM Integration II – Hive	19
3.2.2.3.	UPM Integration III – Elasticsearch	21
3.2.2.4.	UPM Integration IV – Neo4J	21
3.2.2.5.	UPM Integration V – Siren <i>Investigate</i>	23
3.2.3.	CERTH Integrations.....	25
3.2.3.1.	CERTH Integration I – HDFS	25
3.2.3.2.	CERTH Integration II – HBase.....	26
3.2.3.3.	CERTH Integration III, VI and V – Apache Zeppelin.....	27
3.2.3.4.	CERTH Integration VI – Elasticsearch.....	27
3.2.3.5.	CERTH Integration VII – Spark	28
3.2.3.6.	CERTH Integration VIII – MLib	28
3.2.3.7.	CERTH Integration IX – Elasticsearch.....	29
3.2.4.	LvT Integrations	31
3.2.4.1.	LvT Integration I – Kafka	31
3.2.4.2.	LvT Integration II – Fandango Webapp	32
3.2.4.3.	LvT Integration III, IV, V – Apache Zeppelin.....	33
3.2.4.4.	LvT Integration VI – Elasticsearch.....	34
3.2.4.5.	LvT Integration VII – Spark.....	35
3.2.4.6.	LvT Integration VIII – Spark.....	36
3.2.4.7.	LvT Integration IX – Spark.....	37
3.2.4.8.	LvT Integration X – HBase.....	37
4.	Conclusion	38
5.	ANNEX – Data Silos for European Content, Climate Change and Migration	39
	Podcasts and Vodcasts	43

LIST OF FIGURES

Figure 1 - Architecture Overview	7
Figure 2 - Data Loading	10
Figure 3 - Siren Integrations	16
Figure 4 - UPM Integrations	18
Figure 5 - CERTH Integrations	25
Figure 6 - LvT Integrations.....	31

LIST OF TABLES

Table 1 – Architecture Components Overview	9
--	---

ABBREVIATIONS

ABBREVIATION	DESCRIPTION
H2020	Horizon 2020
EC	European Commission
WP	Work Package
EU	European Union

EXECUTIVE SUMMARY

This document is a deliverable of the FANDANGO project funded by the European Union's Horizon 2020 (H2020) research and innovation programme under grant agreement No 780355. It is a public report that describes the data lake integration plan for the software development within FANDANGO.

The main goal of this deliverable is to define the data sets that will be collected and processed in the FANDANGO's data-lake, ultimately describing how data is handled and curate in different steps of the process.

Data lakes require data integration solutions that can work with structured and unstructured data, likely with schema-less data storage, and with streams of data that should be processed in near real-time. In other words, data lake requires a completely different approach to data integration and newer data integration technology as compared to traditional data warehouse.

Therefore, this document describes the different data ingestions and integrations currently designed, based on the proposed architecture. For each of those, ownership of source and target repository, type of data, access control, persistence period and purpose are asserted.

As the data lake evolves so will its documentation, becoming more descriptive and precise during the lifespan of the project. Integrations and complementary information will be added into specific sections of deliverables *D2.2 - Data Interoperability and data model design*, *D3.1 - Data model and components* and/or *Project Progress Periodic Reports* to define more detailed data structures available in each repository and its conventions.

1. INTRODUCTION

FANDANGO’s goal is to aggregate and verify different typologies of news data, media sources, social media and open data to detect fake news and provide a more efficient and verified form of communication for European citizens.

To achieve such goal, several different approaches must be used in conjunction to collect a large volume of data. The collection of these datasets is essential to ensure that the Machine Learning algorithms can process the inputs into meaningful information and provide high quality interactions with the user that allows real-time analysis for investigation and validation purposes. Solutions like Spark, which will be used for fast processing of machine learning and graph analysis, needs to work in conjunction with Elasticsearch, that is focused on semantic and statistical computation. Such strategy requires different technologies to be used and interconnected into a single, meaningful, solution. The parts responsible for such interconnections, between data and different parts of the software solution, are the integrations, which will be described in further detail on this document

2. ARCHITECTURE OVERVIEW

To define the data lake integration, it is crucial to analyse the overall architecture of the solution and how data will be collected and processed across different environments. Therefore, the initial architecture overview in Figure 1 serves as base to describe the different parts of the solution.

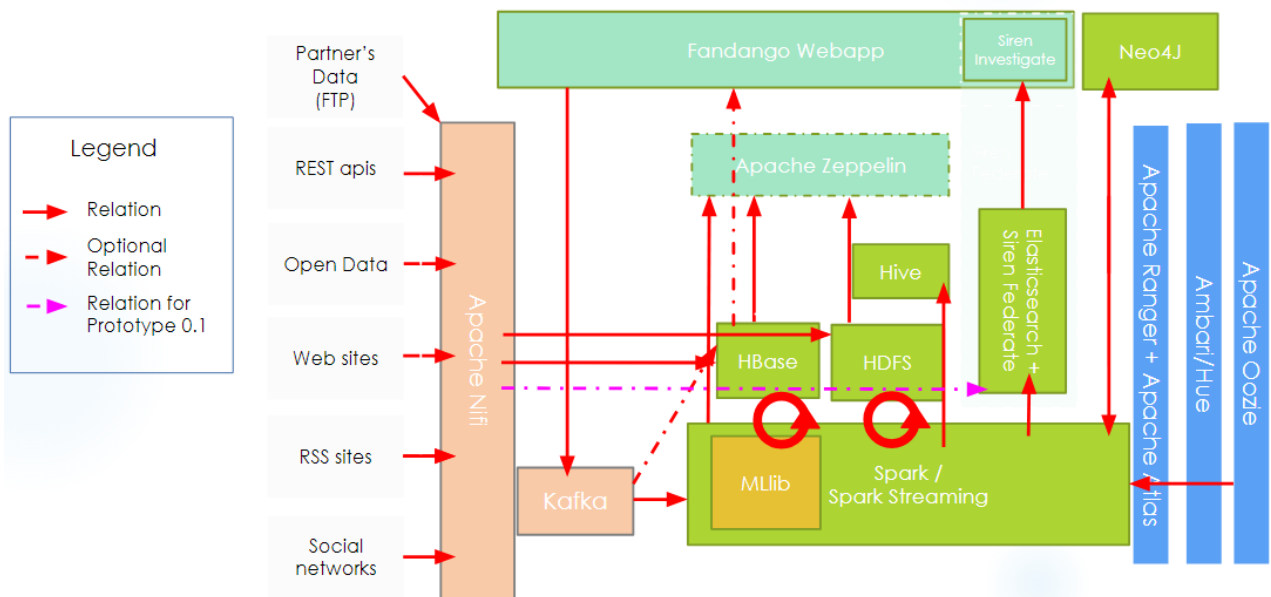


Figure 1 - Architecture Overview

FANDANGO’s features to support journalist in fake-news detection and verification, as well as scoring the news with different trustworthiness scores, requires the development of several different big data processing and analyzing techniques. To optimize the solution and better comply to software quality standards, such as: Functional Suitability, Reliability, Operability, Performance Efficiency, Security, Compatibility, Maintainability and Transferability, FANDANGO relies on well-established products that were brought together to form the proposed architecture. The components of the architecture, which needs to be integrated are described on Table 1 – Architecture Components Overview.

SOFTWARE	DESCRIPTION
Nifi	Data flow ingestion tool, open source, distributed and scalable, to model real-time pre-processing workflow from several different sources.
Kafka	Publish-subscribe distributed messaging system, that grants high throughput and back pressure management.
Spark	Fast, in-memory, distributed and general engine for large-scale data processing with machine learning (Mllib), graph processing (GraphX), SQL (Spark SQL) and streaming (Spark Streaming) features.
HDFS	The Hadoop distributed file system, open source, reliable, scalable, chosen as storage.
Elasticsearch + Siren Federate	Distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents. Siren Federate plugin is added to Elasticsearch to allow data set semi-joins and seamless integration with different data sources.
Hive	Query engine (SQL-like language) on HDFS (and HBase) with JDBC/ODBC interfaces.
Oozie	Workflow scheduler.
Ambari	it acts as both a workflow engine and a scheduler. In this case, its main role is to manage the scheduling of Spark jobs and the creation of Hive tables.
Hue	Hadoop HUE to perform dashboards, queries and browse the services.
Siren	Investigative Intelligence UI with connectivity to Elasticsearch, whose aim is to allow reporting, investigative analysis and alerting to users based on the indexed contents.
Rest APIs, RSS, Web Sites, Open Data, Social network	Data sources of the Fandango project. Specific crawlers will connect to these sources of data to get the information needed to verify the news.
FTP	The File Transfer Protocol (FTP) is a standard network protocol used for the transfer of computer files between a client and server on a computer network. In our Architecture it is where Users can place files that will be than ingested in the data lake.
HBase	The Hadoop NoSQL database, to perform random read and writes based on rowkey identifiers.
Zeppelin	The notebook dedicated to data scientists, to run in REPL mode scripts and algorithms on data stored in Hadoop.

Atlas	Apache Atlas provides scalable governance for Enterprise Hadoop that is driven by metadata, adding features for data lineage, governance controls to address compliance requirements and agile data modelling.
Ranger	Framework to enable, monitor and manage data security across the Hadoop platform according to fine-grained policies and a centralized security and auditing.
Web App	Access point to Fandango Infrastructure. The journalist will use the Fandango Web application to insert news and verify the trustworthiness of certain publications.

Table 1 – Architecture Components Overview

3. DATA INTEGRATIONS

To allow the implementation of the data processing and analysis techniques needed to support the FANDANGO’s features, interconnections between the different parts of the solutions and enable the functional requirements designed for FANDANGO, multiple Data Integration processes are required, identified by red arrows in Figure 1. This section is going to describe in further detail each of these integration processes, breaking down into two main steps: data ingestion and data processing. The first one describes the acquisition of external data by the solution and the second its different processing stages within FANDANGO.

3.1. DATA INGESTION

The initial steps in the process is acquiring data from multiple sources. Some of desired data inputs have been mapped and it will be shaped in more details along the requirement evolution and the first software delivery iterations. These can be seen in Figure 2 Figure 2 - Data , and will be described in further detail in the following sections.

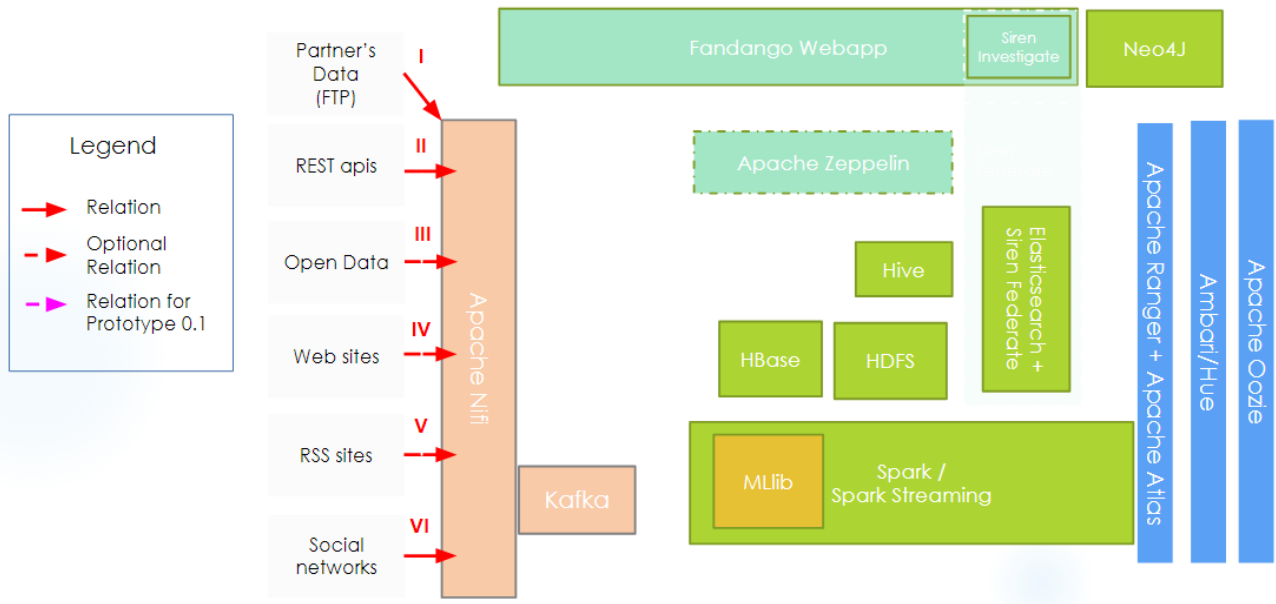


Figure 2 - Data Ingestion

All data ingestion processes will be implemented by CERTH, by following the data model design that is to be defined in WP2.

3.1.1. DATA INGESTION I – PARTNER’S DATA (FTP)

FANDANGO’s user partners own collections of valuable data that may be used in various situations from training machine learning models to support FANDANGO’s fake-news detections feature. The datasets that will be made available are of different types and on a variety of formats.

For each dataset a custom data shipping script will be provided that will ingest the data in the FANDANGO cluster and made available to the processing units.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	The datasets will come from the data collections of ANSA, VRT, CIVIO
Ownership	Each dataset will be owned by the partner that is sharing the data and will be used in FANDANGO only internally. Partner’s data will not be exposed to the public.
Type	Most of the data are in text formats such as plain text, pdf files and word files. Another batch of data will be images and videos in known formats.
Access Control	Internal network only.
Persistence	The data will be kept for the duration of the FANDANGO project. After the finalisation of the project, a new agreement will be conducted.

DATA TRANSFORMATION

The initial ingestion of the data will not follow any transformation procedure. This will permit the partners working on the processing modules to experiment with different configurations with the original data. After the establishment of a solid processing workflow, the defined data transformations and successive updates will be specified into specific annexes to deliverables *D2.2 - Data Interoperability*

and *data model design, D3.1 - Data model and components* and/or *Project Progress Periodic Reports*.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	HDFS
Ownership	The partner that shared the data
Type	HDFS files
Access Control	Internal network
Persistence	Kept until the end of the project

3.1.2. DATA INGESTION II – REST APIS

A list of sources that give access through REST APIs will also be integrated in the FANDANGO data shippers. The REST API data shipper will be able to load data from different sources by changing only a small part of the script with the specificities of each service.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	RSS sources will be defined and integrated during the project
Ownership	Third party
Type	JSON
Access Control	Public access
Persistence	Managed according to the terms of use of each REST API provider

DATA TRANSFORMATION

The data will follow the data model that will be defined. Until then no transformation will be applied.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Elasticsearch
Ownership	Third party
Type	JSON
Access Control	internal network
Persistence	Managed according to the terms of use of the data owner.

3.1.3. DATA INGESTION III – OPEN DATA

A list of open data sources is under development in the FANDANGO project. These sources are mainly text data coming from public organizations either in national or European level. Some of these open data portals, share their data through programmable interfaces, however there is a clear majority that only provided downloadable links to pdf, csv or xls files. Sources such as the Eurobarometer, the Eurostat, the European External Action Service and other organizations are in this category.

A list of all the open dataset is provided in 5ANNEX – Data Silos for European Content, Climate Change and Migration.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Data publicly available from national and European organisations
Ownership	Open Data publishers.
Type	json, csv, xls, pdf formats are available
Access Control	Internal network
Persistence	Managed according to the terms of use of the data owner

DATA TRANSFORMATION

The data will be fed to the FANDANGO data lake as is. After the loading, the modules that perform the pre-processing will transform them to plain text.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
----------------	-------------

Target System	HBASE
Ownership	Open Data publishers.
Type	Data table
Access Control	Internal network
Persistence	Managed according to the terms of use of the data owner.

3.1.4. DATA INGESTION IV – WEBSITES

A focused web crawler is being implemented in WP3 to load websites that are relevant to news and fake news debunking. The crawler will hold a list of predefined news sources such as newspaper sites, blogs, factcheckers and other related sources. The crawling of these sites will follow a delta approach that will gather only the updates after the original / initial crawling process.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	News sites and other related web pages that will be manually selected
Ownership	Publishers and/or authors
Type	JSON files that contain the text and the URIs of multimedia in each news post
Access Control	Internal network
Persistence	Managed according to the terms of use of the data owner.

DATA TRANSFORMATION

The data will be hold in JSON format and no further transformation is needed.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Elasticsearch
Ownership	Publishers and/or authors
Type	JSON
Access Control	Internal network

Persistence	Managed according to the terms of use of the data owner.
-------------	--

3.1.5. DATA INGESTION V – RSS SITES

The data shippers will provide the means to gather data from RSS feeds. A list of monitored RSS feeds will be created with the help of our users' partners.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	RSS feeds of news agencies and news sites
Ownership	Third party
Type	XML
Access Control	Internal network
Persistence	Managed according to the terms of use of the data owner.

DATA TRANSFORMATION

The data will be transformed into JSON format for a uniform approach on the handling of text data.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	HBASE
Ownership	Third party
Type	Data table
Access Control	Internal network
Persistence	Managed according to the terms of use of the data owners.

3.1.6. DATA INGESTION VI – SOCIAL NETWORKS

A special source for fake news is the social networks. Social networks are the main channels of news propagation and as such FANDANGO must keep a constant eye on what is share there. The first and the most interesting in terms of fake news and news propagation is Twitter. FANDANGO's data shippers will create data gathering with different parameters such as using keywords, hashtags or users' accounts and geolocation queries.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Twitter and other social networks
Ownership	Third party
Type	JSON
Access Control	Internal network
Persistence	Managed according to the terms of use of the data owner.

DATA TRANSFORMATION

No transformations will be applied in the original data.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Elasticsearch
Ownership	Third party
Type	JSON
Access Control	Internal network
Persistence	Will be kept as long as it is permitted by the terms of use of each service

3.2. DATA PROCESSING

Once external data has been made available within FANDANGO’s platform, there are multiple processing stages required to enhance it, analyse it and ultimately make information available to the user, which will then provide more data to the system to continue its learning cycle.

To facilitate the control of the deliverables, project planning and provide better visibility of the require implementations, the different data processing integration have been broken into sub-groups that will be implemented by different partners in FANDANGO project. Each group and its implementations is described in the following sections.

3.2.1. SIREN (SINDICE) DATA PROCESSING INTEGRATIONS

The integrations highlighted in red in Figure 3 are going to implemented by the partner Siren.

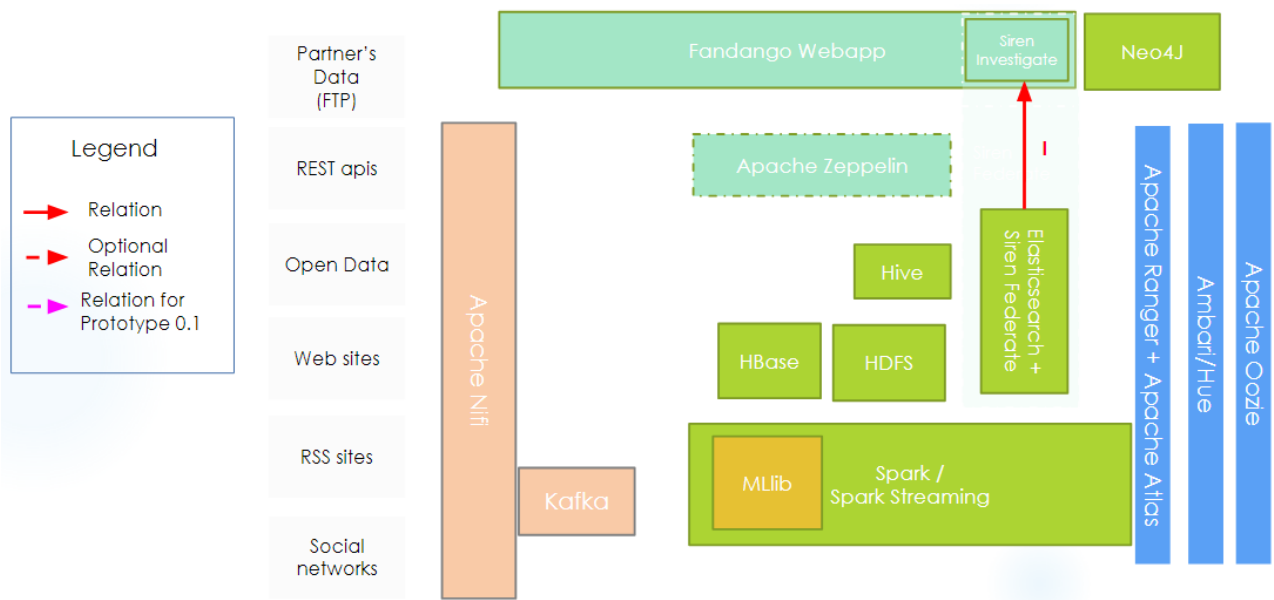


Figure 3 - Siren Integrations

3.2.1.1. SIREN INTEGRATION I – SIREN INVESTIGATE

This integration is responsible for accessing consolidated datasets, made available in Elasticsearch and bringing it to Siren Investigate platform, where users can do investigative analysis through Dashboards and Knowledge Graphs.

The data retrieved is dependent on user request and the assigned credentials, and it is only treated for presentation on the target software. While multi-dataset filters are allowed, data content and granularity is kept unchanged between the systems.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Elasticsearch and Siren Federate plugin
Ownership	FANDANGO
Type	JSON Documents
Access Control	User authentication
Persistence	Data will be preserved indeterminately for analysis purposes. Sanitizing policies might be created after production

DATA TRANSFORMATION

Data is collected and transported without changes to its content. That allows the original data retrieval design to be preserved and ensures that information being presented to the user isn't altered by aggregation or transformation processes.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Siren Investigate
Ownership	FANDANGO
Type	JSON Documents consolidated into Dashboards and Knowledge Graphs
Access Control	User authentication
Persistence	Real-time screen visualization only or CSV export by users

3.2.2. UPM INTEGRATIONS

The integrations highlighted in red in Figure 4 - UPM Integrations are going to be implemented by the partner UPM.

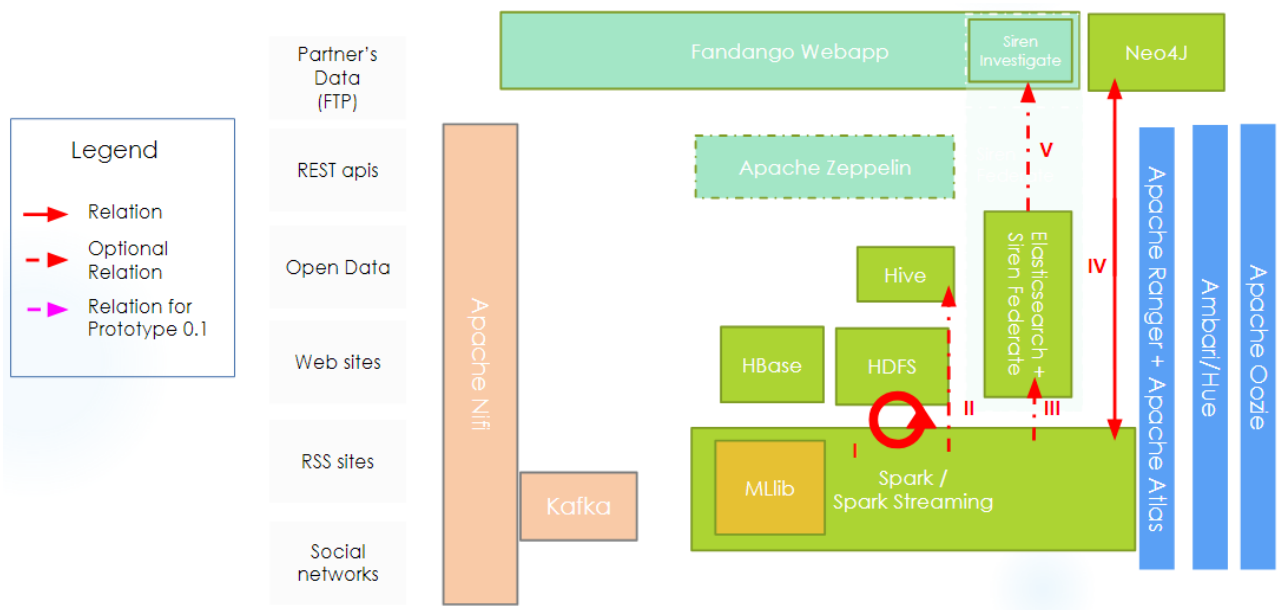


Figure 4 - UPM Integrations

3.2.2.1. UPM INTEGRATION I – SPARK

Spark is a fast and general cluster computing system for Big Data. It provides high-level APIs in different languages including Scala, Java, Python, and R, and an optimized engine that supports general computation graphs for data analysis. It also supports a rich set of higher-level tools including Spark SQL for SQL and DataFrames, MLib for machine learning, GraphX for graph processing, and Spark Streaming for stream processing. (Apache Spark, s.f.)

Moreover, Spark is very flexible, and it allows to preprocess the data in order to store it in a proper format for future data transformations and data analysis.

Indeed, Spark will be employed to preprocess the data provided by the different data sources with the aim of reducing the complexity of the raw data such as images, video content as well as the text files. Since Hortonworks Data Platform (HDP) supports Apache Spark, the integration of such component does not require an external procedure.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Apache NIFI.

Ownership	FANDANGO
Type	Binary and HDFS files for storing images, video content and meta-data.
Access Control	Internal Network
Persistence	The original data will be managed by a third party in order to be removed, processed or stored.

DATA TRANSFORMATION

Data preprocessing can be defined as a data mining technique that involves transforming raw data into an understandable format. The main problem in Real-world data is that it is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Hence, a data preprocessing stage is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

In this project, different data sources will be stored in the Data Lake, and therefore, some data transformations should be applied to Images and other media-content in order to normalize and scale the original data.

Several Data Transformations procedures including centering the data, normalizing will be applied in the different data sources with the aim of standardizing the data for the future Machine and Deep Learning procedures.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Apache ZEPPELIN
Ownership	FANDANGO
Type	HDFS files
Access Control	Internal Network
Persistence	The original data will be managed by a third party in order to be removed, processed or stored.

3.2.2.2. UPM INTEGRATION II – HIVE

Hive is a data warehouse infrastructure built on top of Hadoop. It provides tools to enable easy data ETL, a mechanism to put structures on the data, and the capability for querying and analysis of large data sets stored in Hadoop files. The integration of such component in the HDP is similar to the Spark one since HIVE is a native component of Hortonworks.

Hive defines a simple SQL query language, called HiveQL, that enables users familiar with SQL to query the data. At the same time, this language also allows to work with the MapReduce framework by plugging

custom mappers and reducers to perform more sophisticated analysis that may not be supported by the built-in capabilities of the language.¹

The use of HIVE in the project will be basically to support Spark in the preprocessing methods by providing flexibility and scalability in the required data queries. It may also be employed to perform data analysis tasks over large datasets which are stored in HDFS files using its User Interface as well.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Spark
Ownership	FANDANGO
Type	HDFS files
Access Control	Internal Network
Persistence	The data is managed by a third party to be visualized or processed

DATA TRANSFORMATION

In this scenario, since the transformation step will be carried out by the Spark component, Apache HIVE will not require to perform any data transformation due to It will be employed to make flexible and scalable queries of the data stored in the Spark component as well as to support any other component which requires the usage of a query system for real-time visualization or data analysis.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	FANDANGO WebApp
Ownership	FANDANGO
Type	HDFS files
Access Control	Internal Network
Persistence	The data will be queried in Real-Time screen visualization and managed by a third party.

¹ https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.6.4/bk_data-access/content/ch_using-hive.html

3.2.2.3. UPM INTEGRATION III – ELASTICSEARCH

Elasticsearch is a distributed, RESTful search and analytics engine capable of solving a growing number of use cases. It centrally stores the data and it will be used to process the required queries in order to visualize the results in real-time using the Dashboard of the Siren Investigate.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Spark, HDFS, Apache NIFI
Ownership	FANDANGO
Type	JSON Document
Access Control	Internal Network
Persistence	In-memory processing only

DATA TRANSFORMATION

In this case, the data is collected and transported without suffering from any kind of changes since this component will be used to transport the information between pairs of modules and the original information should remain intact since this module will help users in the real-time visualization procedure.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	FANDANGO Web App
Ownership	FANDANGO
Type	JSON document
Access Control	Internal Network
Persistence	In-memory processing

3.2.2.4. UPM INTEGRATION IV – NEO4J

Neo4J is a Graph Data Base (GDB) mainly oriented to graphs. It means that it uses graphs to represent the data (entities) and the relationship between these. There exist multiple manners of representing these graphs:

- **Undirected Graph:** nodes and links can be exchanged, and its relationship can be interpreted regardless the direction. (i.e. Friend links in Facebook)

- **Directed Graph:** nodes and relationships are not bidirectional (by default). An example of this type is given by twitter following relationships. A user can follow some profiles in this network without these profiles can follow him/her back.
- **Weighted graph:** the relationships between entities are represented by a numerical value (weight). It allows performing some special operations.
- **Labeled graph:** these graphs have labels incorporated that can define the multiple edges and the type of relationship between nodes (i.e. Facebook labeled relationships include friends, job colleague, partner of, friend of
- **Property Graph:** is a weighted graph, with labels where properties can be assigned both to entities (journal, publisher....) as well as relationships (General categories such as name, country, birth place)

In the context of FANDANGO, a particular application of these databases matches with the ambition of Task T4.4 Source credibility scoring, profiling and social graph analytics. This task aims to detect nodes associated to fake content generation and relationships with these entities. For this purpose, it is expected to have a complete definition of the multiple actors involved in the fake news detection paradigm. This will allow to express the complete environment and to exploit sources from multiple entities (News has been published for an author with a bad reputation, so it is likely to be biased or even fake). This paradigm definition is also commonly known as ontology. There are some general-purpose news-related ontologies in the field of news analysis². These approaches will be taken as starting point for the news analysis³.

The integration of NEO4J into the HortonWorks Data Platform (HDP) as a service has been done. The process is summarized as follows:

- in HDP a folder must be created at: `‘/var/lib/ambari-agent/cache/stacks/HDP/2.6/services’` with the name of the service ‘Neo4J’
- go into the folder and clone the <https://github.com/cas-bigdatalab/ambari-neo4j> repository.
- [optional] Change the configuration (General parameters IP, PORTS, SECURITY....) in the configuration file a `‘/master/configuration/neo4j.xml’`
- Start the HDP and go to add a service...

What this repository does, is to create a folder in the `/etc/yum.repos.d/neo4j.repo` and install the most recent version of the software and attached it to the whole stack of services into the HDP platform.

For the interest of the entire research/development community, a DockerHub image has been created integrating the services required by UPM for FANDANGO.⁴

SOURCE

Characteristics of the data origin.

² The IPTC is the global standards body of the news media that provides the technical foundation for the news ecosystem <http://dev.iptc.org/rNews>

³ BBC Ontology: <https://www.bbc.co.uk/ontologies/storyline>

⁴ DockerHub FANDANGO modules https://hub.docker.com/r/tavitto16/fandango_hdp/

CHARACTERISTIC	DESCRIPTION
Main Source	Spark and FANDANGO WebApp
Ownership	FANDANGO
Type	JSON and csv documents
Access Control	Authenticated access
Persistence	The data will be used to make graph analytics and data visualizations by a third party.

DATA TRANSFORMATION

Since NEO4j will be employed in the graph analysis performance, the transformations applied over the origin data will consist of a set of queries and graph operations. This set of operations will be used to find relevant patterns and to analyze the credibility of some sources using graph algorithms but at the end, the output of these operations must be a new graph (if the graph has been modified) as well as a set of metrics or results whether they are required.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	FANDANGO WebApp and Siren investigate.
Ownership	FANDANGO
Type	JSON document
Access Control	Authenticated access
Persistence	The data will be used for graph analytics and data visualizations by a third party.

3.2.2.5. UPM INTEGRATION V – SIREN INVESTIGATE

Siren investigate will be used to visualize the graph database with all the entities and relationships involved in FANDANGO's ontology. In addition, basic modifications and analytics within the graph can also be performed using this module.

Moreover, Siren investigate will be communicate with Neo4j in case the latter is required to perform more advance graph analytics.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
----------------	-------------

Main Source	NEO4J, FANDANGO WebApp
Ownership	FANDANGO
Type	JSON, ontology file (owl, rdf).
Access Control	Internal Network
Persistence	Real-time visualizations

DATA TRANSFORMATION

In this process, the data will be transformed whether the graph analysis employs some algorithms that will modify the current graph (i.e. add new entities or relationships or remove some of them). In this case, the transformation will consist of updating the current graph and store such version in the platform to be visualized later on. However, the format of the data must be the same that the original. This transformation will only affect to the information provided by the graph but not in the structure of it.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	FANDANGO WebApp
Ownership	FANDANGO
Type	JSON and ontologies configuration files
Access Control	Internal Network
Persistence	Real Time Visualizations

3.2.3. CERTH INTEGRATIONS

The integrations highlighted in red in Figure 5 are going to be implemented by the partner CERTH.

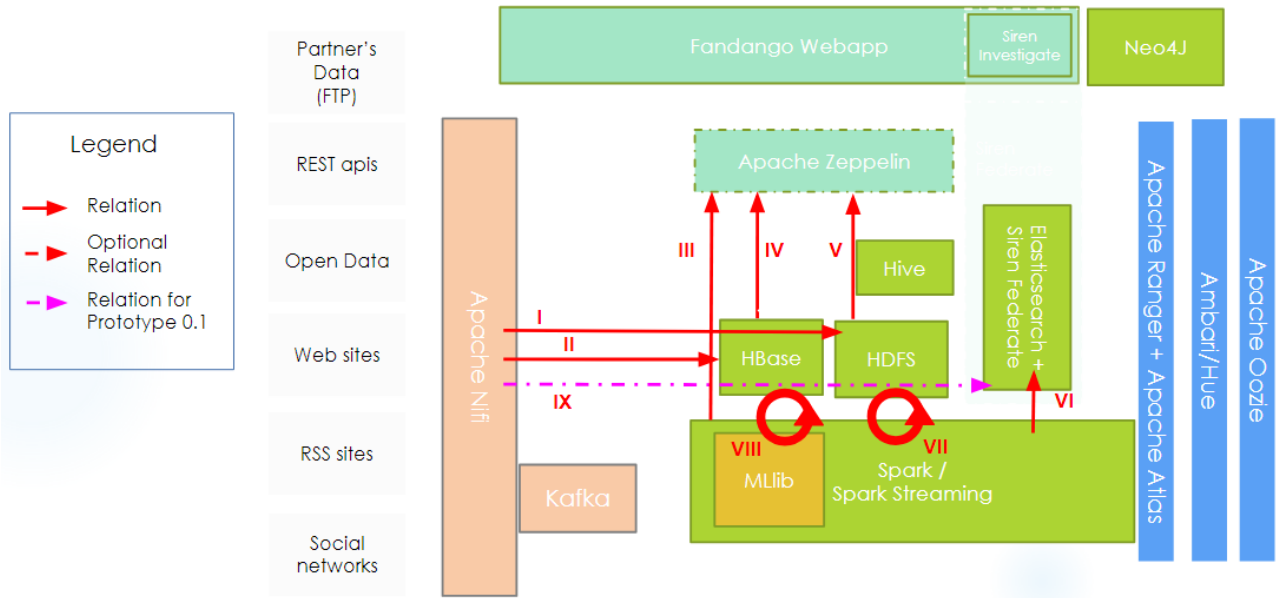


Figure 5 - CERTH Integrations

3.2.3.1. CERTH INTEGRATION I – HDFS

that will be immutable. It is convenient to work with and will hold any type of data of any format. HDFS will be the main storage module FANDANGO will be using to push data

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Data from data shippers through the Apache NIFI
Ownership	FANDANGO
Type	HDFS holding any type of data
Access Control	internal network
Persistence	depends on each source as described in the previous sections.

DATA TRANSFORMATION

As it is described in the data loading section.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	All processing systems e.g. MLlib, Spark, etc.
Ownership	FANDANGO
Type	Mostly JSON files but depends on the original data
Access Control	Internal network
Persistence	As it is defined in the data loading section

3.2.3.2. CERTH INTEGRATION II – HBASE

HBASE will be used for data that are structured such as information coming from RSS feeds or open data portals from European and national organizations.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	RSS and open data
Ownership	FANDANGO unless otherwise defined by the data provider
Type	Data table
Access Control	Internal network
Persistence	As it is defined in the data loading section

DATA TRANSFORMATION

No transformations required.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	All processing modules e.g MLlib and spark
Ownership	FANDANGO
Type	Data table
Access Control	Internal network
Persistence	As it is defined in the data loading section

3.2.3.3. CERTH INTEGRATION III, VI AND V – APACHE ZEPPELIN

Apache zeppelin is a notebook environment. It will be used to work on the data that are stored in the FANDANGO cluster and experiment directly on the data without the need to transfer data to and from the cluster for the research and prototyping purposes. As it is used as a sandbox area, different datasets and formats will be loaded into it. Data in this area will only be persisted while required for the prototyping purposes and will be controlled through user authentication for accessing the notebooks.

3.2.3.4. CERTH INTEGRATION VI – ELASTICSEARCH

Elasticsearch will collect all the processed data and the outcomes of the processing modules for the identification of trustworthiness markers.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Processing modules implemented in WP4
Ownership	FANDANGO
Type	JSON
Access Control	Internal network
Persistence	Deleted after processing

DATA TRANSFORMATION

No transformations required.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Elasticsearch
Ownership	FANDANGO
Type	JSON
Access Control	Internal network
Persistence	Kept for the duration of the project

3.2.3.5. CERTH INTEGRATION VII – SPARK

Spark will be used for processing the available data in the developed modules of WP4.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	HDFS
Ownership	FANDANGO
Type	Any format
Access Control	Internal network
Persistence	Kept in HDFS for the duration of the project

DATA TRANSFORMATION

No transformations are required.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Spark
Ownership	FANDANGO
Type	Any format
Access Control	Internal network
Persistence	In-memory processing only

3.2.3.6. CERTH INTEGRATION VIII – MLlib

The MLlib is used for Machine Learning modules and works in collaboration with spark. As such all the details that apply for spark apply to MLlib as well.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	HDFS or HBASE

Ownership	FANDANGO
Type	Any format
Access Control	Internal network
Persistence	Kept in the storage as it is defined in the data loading section

DATA TRANSFORMATION

No transformations are required.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	MLlib
Ownership	FANDANGO
Type	Any format
Access Control	Internal networks
Persistence	In-memory processing only

3.2.3.7. CERTH INTEGRATION IX – ELASTICSEARCH

This integration will be used initially for the pilot 0.1 and will be evaluated if it will remain in the future versions of the FANDANGO platform.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Apache Nifi
Ownership	FANDANGO
Type	JSON
Access Control	Internal network
Persistence	In-memory only

DATA TRANSFORMATION

No transformations apply here.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Elasticsearch
Ownership	FANDANGO
Type	JSON
Access Control	Internal network
Persistence	For the duration of the project or otherwise defined by the data sources' terms of use,

3.2.4. LVT INTEGRATIONS

The integrations highlighted in red in Figure 6 - LVT Integrations are going to be implemented by the partner LVT.

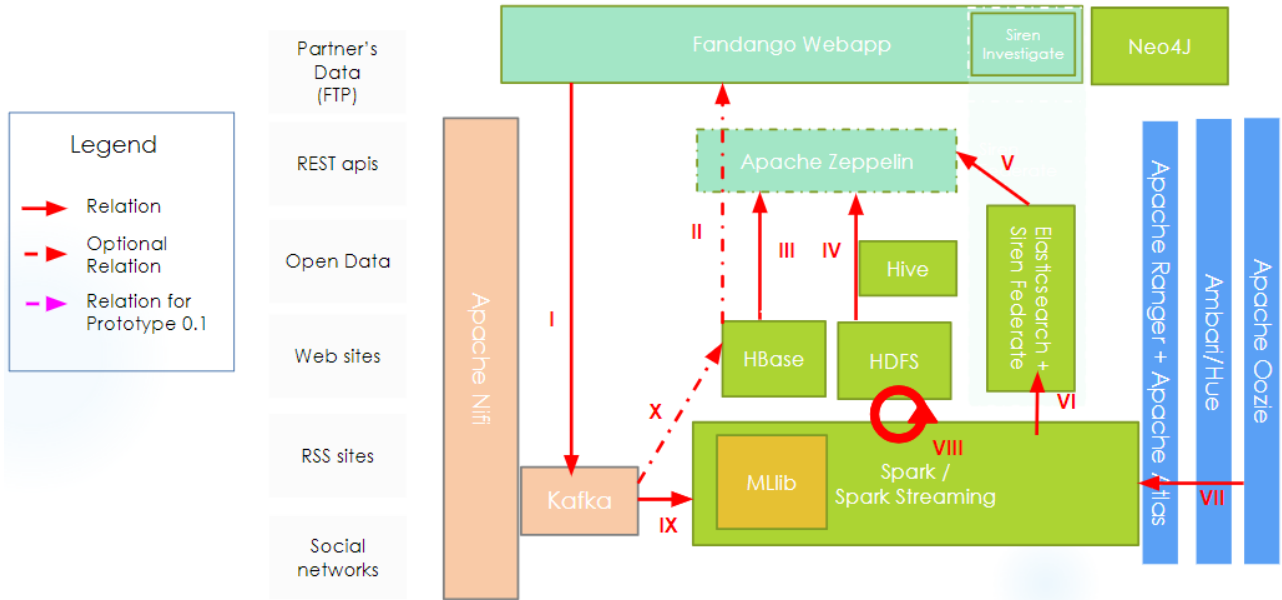


Figure 6 - LVT Integrations

3.2.4.1. LVT INTEGRATION I – KAFKA

Kafka will be used for queueing up the news before their analysis of fake and saving

The collection of news from different sources (Web site, Open Data, Webapp, etc..) will produce hundreds/thousands of data, for this reason we will need Kafka, otherwise we can't process all the news together. In addition, we don't want to save all the news we retrieve into databases, but only the fake news our Machine Learning System recognizes.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Apache Nifi, Fandango App
Ownership	FANDANGO
Type	JSON
Access Control	Authenticated access
Persistence	Real-time screen visualization only

DATA TRANSFORMATION

Kafka doesn't apply any kind of transformation, therefore all news must have an agreed json format before putting them in Kafka.

For example:

```
{ "source": aaaa,
  "date": dddd
  "title": xxxx,
  "body": yyyy,
  "urls_image": ["xxx","xxx",...,"xxx"]
  etc..
}
```

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	HDFS, HBase
Ownership	Fandango
Type	JSON
Access Control	Internal network
Persistence	In-memory processing; deleted after processing

3.2.4.2. LVT INTEGRATION II – FANDANGO WEBAPP

The Webapp provides a set of interfaces to manage in a single touchpoint:

- data from the different FANDANGO layers
- distributed data resources and transfers them efficiently and securely.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Fandango platform's
Ownership	FANDANGO

Type	JSON
Access Control	Public access and Authenticated access
Persistence	In-memory processing

DATA TRANSFORMATION

It is a Webapp control, so it doesn't need any kind of transformations.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Fandango platform's
Ownership	FANDANGO
Type	JSON
Access Control	Public access, authenticated access
Persistence	Real-time screen visualization and HIVE/HBase to save the configurations

3.2.4.3. LVT INTEGRATION III, IV, V – APACHE ZEPPELIN

Apache zeppelin is a python notebook environment. It will be used to work on the data that are stored in the FANDANGO cluster and experiment directly on the data without the need to transfer data to and from the cluster for the research and prototyping purposes. As it is used as a sandbox area, different datasets and formats will be loaded into it. Data in this area will only be persisted while required for the prototyping purposes and will be controlled through user authentication for accessing the notebooks.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	HBase, HDFS, Elasticsearch
Ownership	FANDANGO
Type	JSON
Access Control	Internal network
Persistence	In-memory processing

DATA TRANSFORMATION

HBase already contains data in json format, therefore it doesn't require any transformations.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Apache Zeppelin
Ownership	FANDANGO
Type	Serialized object, text, JSON and others
Access Control	Internal network
Persistence	In-memory processing

3.2.4.4. LVT INTEGRATION VI – ELASTICSEARCH

It centrally stores the data and it will be used to analyze news using natural language processing tools.

Elasticsearch is necessary because it implements a search engine that will be used to search similar news based on semantic context.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Spark
Ownership	FANDANGO
Type	JSON
Access Control	Internal network
Persistence	In-memory processing

DATA TRANSFORMATION

Elasticsearch will apply NLP pipelines depending on the language of the news. The main transformations that will be applied are: Tokenization, Lemmatization, Syntactic Parser, Ngram etc.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
----------------	-------------

Target System	Elasticsearch
Ownership	FANDANGO
Type	JSON
Access Control	Internal network
Persistence	Kept indeterminately

3.2.4.5. LVT INTEGRATION VII – SPARK

Apache Oozie is a workflow scheduler that is used to manage Apache Hadoop jobs. Oozie combines multiple jobs sequentially into one logical unit of work as a directed acyclic graph (DAG) of actions. Oozie can weave a Spark job into your workflow. The workflow waits until the Spark job completes before continuing to the next action.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Oozie
Ownership	Fandango
Type	JSON
Access Control	Authenticated access and internal network
Persistence	In-memory processing

DATA TRANSFORMATION

Apache Oozie will not require to perform any data transformation

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Spark
Ownership	FANDANGO
Type	JSON
Access Control	Internal network

Persistence	In-memory processing
-------------	----------------------

3.2.4.6. LVT INTEGRATION VIII – SPARK

Spark is useful to analyze millions of data in short time. Therefore, in this project, Spark will be employed to process the news provided by the different data source. This is done using the MLlib, a library for Machine Learning of Spark. The interaction with the saved news is useful to catch feedbacks of the users about the truth of a news. These feedbacks will be processed by the machine learning models.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	HDFS
Ownership	FANDANGO
Type	JSON
Access Control	Internal network
Persistence	Kept indeterminately

DATA TRANSFORMATION

In order to re-training the Machine and Deep Learning algorithms in MLlib, we will apply different data transformations to make the formats compatible with the libraries you use.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Spark
Ownership	Fandango
Type	JSON
Access Control	Internal network
Persistence	In-memory processing

3.2.4.7. LVT INTEGRATION IX – SPARK

Spark is useful to analyze millions of data in very short time, therefore in this project, Spark will be employed to process the news provided by the different data sources using the MLlib, a library for Machine Learning in Spark.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
Main Source	Kafka
Ownership	Fandango
Type	JSON
Access Control	Internal network
Persistence	In-memory processing, Kafka persistence, deleted after processing.

DATA TRANSFORMATION

In order to use the Machine and Deep Learning algorithms in MLlib, we will apply different data transformations to make the formats compatible with the libraries you use.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	HDFS
Ownership	FANDANGO
Type	JSON, Text file
Access Control	Internal network
Persistence	Data saved until they are removed

3.2.4.8. LVT INTEGRATION X – HBASE

HBASE will be used to save data that are structured, such as news analyzed or that come from Kafka.

SOURCE

Characteristics of the data origin.

CHARACTERISTIC	DESCRIPTION
----------------	-------------

Main Source	Kafka
Ownership	FANDANGO
Type	JSON
Access Control	Internal network
Persistence	Saved until they are removed

DATA TRANSFORMATION

Apache HBase will not require to perform any data transformation.

TARGET

Characteristics of the data target.

CHARACTERISTIC	DESCRIPTION
Target System	Apache Zeppelin
Ownership	Fandango
Type	In-memory structure
Access Control	Internal network
Persistence	In-memory processing

4. CONCLUSION


While the purpose of this document is providing an initial overview of required data integrations and how data is going to be curated in order to direct the initial development and facilitate the overall coordination of activities between the partners, this structure should evolve along the development of project and will be updated in later stages. Further definitions are already planned on deliverables *D2.2 - Data Interoperability and data model design* and *D3.1 - Data model and components*.

Nonetheless, the architectural structure, data ingestion and data integration definitions will direct the development of the first versions of the solution and play a crucial role in refining requirements and validating the platform.

5. ANNEX – DATA SILOS FOR EUROPEAN CONTENT, CLIMATE CHANGE AND MIGRATION

This annex lists the initial data silos used collect articles for the FANDANGO project.

EU Open Data Portal

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	RESOURCES FORMAT	NOTE/COPYRIGHT
EU Open Data Portal 	http://data.europa.eu/euodp/en/home	<p>The European Union Open Data Portal (EU ODP) provides access to an expanding range of data from the European Union (EU) institutions and other EU bodies.</p> <p>Only EU institutions, agencies and bodies can provide data for the EU Open Data Portal – a single point of access for EU data.</p> <p>These data can be used and reused for commercial or non-commercial purposes.</p>	<ul style="list-style-type: none"> • Social questions • Science • Environment • Employment and working conditions • Economics • Finance • Trade • Production, technology and research • Industry • European Union 	<ul style="list-style-type: none"> • European • Euro zone (when relevant) • National • Other <p>GEOGRAPHICAL COVERAGE:</p> <p>France (3014)</p> <p>Italy (2964)</p> <p>Austria (2948)</p>	EN, FR, GE	Data	<ul style="list-style-type: none"> • ZIP (8800) • HTML (7496) • text/tab-separated-values (7326) • PDF (876) • XML (818) 	<p><i>Data can be reused free of charge and without any copyright restrictions.</i></p> <p>(REUSE OF EU DATA HAS TO BE SHARED WITH THE PORTAL)</p>

		<p>Total datasets available: 12,209</p> <p>PUBLISHERS</p> <hr/> <p>European Parliament (54 datasets)</p> <p>Council of the European Union (3 datasets)</p> <p>European Commission (11072 datasets)</p> <p>European Central Bank (31 datasets)</p> <p>European External Action Service (0 datasets)</p> <p>European Economic and Social Committee (2 datasets)</p> <p>Committee of the Regions (2 datasets)</p> <p>European Investment Bank (2 datasets)</p> <p>European Ombudsman (1 datasets)</p> <p>European Data Protection Supervisor (6 datasets)</p> <p>EU body or agency (1035 datasets)</p>	<ul style="list-style-type: none"> • Agriculture, forestry and fisheries • Energy • Transport • Business and competition • International relations • Geography • Education and communications • Law • International organisations • Politics • Agri- foodstuffs 	<p>Germany (2929)</p> <p>Spain (2908)</p> <p>Belgium (2874)</p> <p>Netherlands (2867)</p> <p>Denmark (2849)</p> <p>Ireland (2842)</p> <p>Portugal (2839)</p> <p>Sweden (2823)</p> <p>Luxembourg (2814)</p> <p>Greece (2813)</p> <p>United Kingdom (2804)</p> <p>Finland (2802)</p> <p>Slovenia (2730)</p> <p>Hungary (2727)</p> <p>Poland (2713)</p> <p>Czech Republic (2685)</p> <p>Latvia (2683)</p> <p>Estonia (2679)</p> <p>Slovakia (2678)</p> <p>Lithuania (2678)</p> <p>Bulgaria (2632)</p> <p>Romania (2623)</p> <p>Cyprus (2562)</p> <p>Malta (2551)</p> <p>Croatia (2299)</p> <p>Norway (1116)</p> <p>Switzerland (1038)</p>		<ul style="list-style-type: none"> • Excel (405) • CSV (379) • application/msaccess (157) • RDF (155) • OCTET_STREAM (146) • TXT (75) • DOC (58) • webservice/sparql (49) • text/n3 (49) • application/x-dbase (45) • XLSX (36) • JSON (18) • JPEG (16) • PNG (13) 	<p>Already listed in the Fandango Agreement</p>
--	--	--	---	--	--	--	---

				<p>Liechtenstein (988)</p> <p>Serbia (966)</p> <p>Former Yugoslav Republic of Macedonia (963)</p> <p>Bosnia and Herzegovina (953)</p> <p>Montenegro (950)</p> <p>Albania (922)</p> <p>Iceland (902)</p> <p>Russia (896)</p> <p>Turkey (885)</p> <p>Monaco (794)</p> <p>Ukraine (790)</p> <p>San Marino (777)</p> <p>Belarus (772)</p> <p>Andorra (767)</p> <p>Moldova (749)</p> <p>Vatican City (740)</p> <p>Kosovo (737)</p> <p>Åland Islands (726)</p> <p>Gibraltar (711)</p> <p>Guernsey (708)</p>			<ul style="list-style-type: none"> • PPT (11) • application/msword (9) • WEBSETERVICE/SPARQL (7) • TIFF (7) • RSS (5) • N3 (3) • DOCX (3) • KML (2) • GIF (2) • interactive web pages (1) • file (1) • application/x-compress (1) • application/javascript (1) • OWL (1) • MXD (1) • E00 (1) • Access (1)
--	--	--	--	---	--	--	--

EUROPEAN DATA PORTAL

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	RESOURCES FORMAT	NOTE/COPYRIGHT
European Data Portal	https://www.europeandataportal.eu	<p>The European Data Portal harvests the metadata of Public Sector Information available on public data portals across European countries.</p> <p>Information regarding the provision of data and the benefits of re-using data is also included.</p> <p>At the very last count, it provides access to 817,747 datasets. Among the others:</p> <ul style="list-style-type: none"> - Italy – 38,259 - Spain – 28,693 - Netherlands – 22,672 - Belgium – 7,315 - Greece – 6,709 - <p>(THOSE DATASET SHOULD BE CHOSEN CAREFULLY, TO AVOID UNNECESSARY INGESTIONS)</p>	<p>Datasets by categories:</p> <ul style="list-style-type: none"> - Agriculture, Fisheries, Forestry & Food - Energy - Regions & Cities - Economy and Finance - Health - Population & Society - Government & Public Sector - International Issues - Transport - Environment - Science & Technology 	All EU Countries	All EU members' languages	CSV, XLA, XML, RDF, JSON	<p>Provides Datasets URI and SPARQL Queries</p> <p>(and 78 catalogues on EU Countries Datasets)</p>	<p>Already listed in the Fandango Agreement</p> <p>LICENCE NEEDED</p> <p>OPEN LICENSE ASSISTANT</p> <p>https://www.europeandataportal.eu/en/content/how-license</p>

EUROPEAN PRESS ROOM (Press releases)

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	RESOURCES FORMAT	NOTE/COPYRIGHT
European Union Newsroom – Press Releases	https://europa.eu/newsroom/	<p>Press releases by EU Institutions of the last 30 days.</p> <p>Press releases databases:</p> <ul style="list-style-type: none"> Committee of the Regions Council of the EU and European Council Court of Justice of the European Union European Central Bank European Commission European Court of Auditors European Data Protection Supervisor European Economic and Social Committee European Investment Bank European Ombudsman European Parliament Eurostat – Statistical Office 	<ul style="list-style-type: none"> • Asylum and migration • Business • Business, taxation and competition • Consumer affairs and public health • Culture, education and youth • Economy and the euro • Employment and social rights • Energy, environment and climate • Enlargement, external relations and trade • EU regional and urban development • Food, farming and fisheries 	All EU Countries	EN FR DE	Audio Video Text	<p>E-mail and RSS: (https://europa.eu/newsroom/rss-feeds_en#press-releases)</p> <p>PODCASTS AND VODCASTS</p> <ul style="list-style-type: none"> • European Parliament • European Commission • Daily press briefing 	<p>Already listed in the Fandango Agreement</p> <p>(FREE TO USE?)</p> <p>We should ask for clarification:</p> <p>https://europa.eu/european-union/contact/write-to-us_en)</p>

			<ul style="list-style-type: none"> • Institutional affairs • International aid, development and cooperation • Justice and citizens' rights • Research and innovation • Security and defence • Statistics • Transport and travel 				<ul style="list-style-type: none"> • Podcast • Daily press briefing • Podcast • Press conferences • Podcast 	
--	--	--	--	--	--	--	--	--

EUROPEAN MEDIA MONITOR (NewsExplorer)

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	RESOURCES FORMAT	NOTE/COPYRIGHT
European Media Monitor News Explorer	http://emm.newsexplorer.eu/	The NewsExplorer uses JRC developed technology to automatically generate daily news summaries, allowing users to see the major news stories (news clusters) in various languages for any specific day and to compare how the same events have been reported in the media written in different languages; The list of most mentioned names and find further automatically derived information (e.g. variant name spellings, titles and phrases, list of the most recent	<ul style="list-style-type: none"> • Clustered news • Countries • People • Other names • Alerts 	Austria Belgium Germany Spain	English, Spanish, Greek, Netherlands, Greek and other 15 languages	Text only	Available on the web pages and RSS format (http://emm.newsexplorer.eu/rss?type=clusters&language=it)	WE SHOULD ASK THEM DIRECTLY Already listed in the Fandango Agreement

		<p>articles and list of related persons and organizations).</p> <p>NewsExplorer carries out the following tasks:</p> <ul style="list-style-type: none"> • cluster all news articles of the day, separately for each language, into groups of related articles; • for each cluster, identify names of people, places, organizations; • apply approximate name matching techniques to all names found in the same cluster, in order to identify which name variants may belong to the same person; • link the monolingual clusters with the related clusters in the other languages; • identify the most typical article of each cluster and use its title for the cluster; • store the extracted information in a database, learning more about each person, etc. every day; • occasionally, the Wikipedia online encyclopedia is automatically searched for images and for further multilingual name variants. 	<ul style="list-style-type: none"> • Timeline 	<p>France</p> <p>U.K.</p> <p>Italy</p> <p>Netherlands</p> <p>United States</p>				
--	--	---	---	--	--	--	--	--

--	--	--	--	--	--	--	--	--

EUROSTAT

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	RESOURCES FORMAT	NOTE/COPYRIGHT
Eurostat	http://ec.europa.eu/eurostat	<p>Eurostat is the statistical office of the European Union.</p> <p>Database of European Statistics, by theme and A to Z</p>	<p>Statistic by Theme:</p> <ul style="list-style-type: none"> - General - Regional - Economy and Finance - Population and Social conditions - Industry, Trade and Services - Agriculture and Fisheries - International Trade - Transport - Environment and Energy 	All EU Countries	English, German, French (mainly)	Text only	<p>http://ec.europa.eu/eurostat/data/database</p> <p><u>SDMX</u> <u>Web</u> <u>Services</u></p> <p><u>Json</u> and <u>Unicode</u> <u>Web</u> <u>Services</u></p> <p>BULK DOWNLOAD</p> <p>http://ec.europa.eu/eurostat/data/bulkdownload</p>	<p>Already listed in the Fandango Agreement</p> <p>All Eurostat databases and electronic publications are available free of charge via the website.</p> <p>EUROSTAT requires notification of use and the INDICATION of provenance (EURSTAT)</p>

			- Science, Technology and Digital Society					
--	--	--	---	--	--	--	--	--

EUROBAROMETER

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	RESOURCES FORMAT	NOTE/COPYRIGHT
Eurobarometer (EU Commission) on Public Opinion	http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/General/index	<p>Eurobarometer was established in 1974. Each survey consists of approximately 1000 face-to-face interviews per country. Reports are published twice yearly. Reproduction is authorized, except for commercial purposes, provided the source is acknowledged.</p> <p>Special Eurobarometer reports are based on in-depth thematic studies carried out for various services of the European Commission or other EU Institutions and integrated in the</p>	<p>Public Opinion and... contains: Eurobarometer A-Z, Eurobarometer Timeline, Eurobarometer 40 years, Eurobarometer Almanac</p> <p>Eurobarometer Interactive is the search engine with FAQ</p>	All the EU members since 1974	All the EU languages since 1974	Text only	<p>No RSS</p> <p>The Data is generally provided in PDF format, but there is the possibility to download XLS format from the Open Data Portal</p>	<p>Already listed in the Fandango Agreement</p> <p>Reuse is authorized, provided the source is acknowledged. The Commission's reuse policy is implemented</p>

		Standard Eurobarometer's polling waves.	<p>Links (to Twitter)</p> <p>Archives points to the old website, renovated in 2016</p>				<p>(https://data.europa.eu/)</p> <p>ARCHIVES</p> <p>http://ec.europa.eu/comfrontoffice/publicopinion/archives/en.htm</p>	<p>by the Decision of 12 December 2011 - reuse of Commission documents [PDF, 728 KB]</p>
--	--	---	--	--	--	--	--	--

OTHER POSSIBLE DATA SILOS

(Pending FURTHER verification)

EUROPEAN E-RESOURCES CENTRE

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	RESOURCES FORMAT	NOTE/COPYRIGHT
European Library and E-	http://ec.europa.e	Online Search of the resources on EU policies, law and more in the	Open Access Resources.	All European Countries	English	Text only	Direct Search on the website	Not listed in the Fandango Agreement Doc

Resources	u/libraries/	European Commission Library's electronic collections	Books, eBooks, Journal Articles and more				(possible other access)	REQUIRES DIRECT CONTACT
-----------	------------------------------	--	--	--	--	--	-------------------------	-------------------------

EUROPEAN EXTERNAL ACTION (EU Foreign Policy)

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	RESOURCES FORMAT	NOTE/COPYRIGHT
EUROPEAN EXTERNAL ACTION SERVICE (EEAS)	http://ris.is.europa.eu/	The EEAS is the European Union's diplomatic service. It helps the EU's foreign affairs chief – the High Representative for Foreign Affairs and Security Policy – carry out the Union's Common Foreign and Security Policy.	The website contains many documents, publications and infographic on EU Foreign Policy, Security and Defence	Almost all EU Countries' Research Projects	English, French, Italian, Spanish	Text, pictures, graphics	RSS or Direct download	Not listed in the Fandango Agreement Doc REQUIRES DIRECT CONTACT

EUROPEAN SCIENTIFIC DATA (Zenodo)

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	NOTE/COPYRIGHT	RESOURCES FORMAT
ZENODO	https://zenodo.org/	Free and Open Digital Archive built by CERN and OpenAIRE to facilitate scientific data exchange among researchers	Zenodo offers access to 1831 Scientific 'Communities'	European Countries	English	Text	Not listed in the Fandango Agreement Doc REQUIRES DIRECT CONTACT	Output via OAI-PMH Direct Search

EUROPEAN UNIVERSITY INSTITUTE (EUI)

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	NOTE/COPYRIGHT	RESOURCES FORMAT
European University Institute (Firenze)	https://www.eui.eu/	The European University Institute (EUI) is a unique international Centre for doctorate and post-doctorate studies and research, situated in the Tuscan hills overlooking Florence.	The website has a search feature on the Historical Archives of the Eu Institutions and on a unique collection	Almost all European Countries	English, French, German, Italian	Text	Not listed in the Fandango Agreement Doc	Direct Search

		<p>Since its establishment 40 years ago by the six founding members of the then European Communities, the EUI has earned a reputation as a leading international academic institution with a European focus.</p> <p>The European Commission supports the EUI through the European Union budget</p> <p>The EUI library boasts around half a million volumes in the Institute's specialist areas, attracting external researchers with an interest in Europe. The campus also hosts the Historical Archives of the European Union (HAEU), which provides an unparalleled insight into the workings of the European Union.</p>	of 150 private archives of pro-European association and personalities				REQUIRES DIRECT CONTACT	
--	--	--	---	--	--	--	-------------------------	--

RISIS (European Scientific Research)

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	NOTE/COPYRIGHT	RESOURCES FORMAT
------	-----	--	--------------	-------------	----------	-------	----------------	------------------

<p>RESEARCH INFRASTRUC TURE FOR RESEARCH AND INNOVATIO N POLICY STUDIES (RISIS)</p>	<p>http://ris.is.eu/data/</p>	<p>The overall objective of the project is to build a distributed infrastructure on data relevant for research and innovation dynamics and policies</p> <p>RISIS aims at opening to European researchers a large number of (linked) datasets covering 7 themes:</p> <ol style="list-style-type: none"> 1. Research funding: Datasets that contain information about research projects funded by the EC (EUPRO, CORDIS), by trans-border funding programs between EC member states (JOREP), and others funders (OPEN-AIR, Open Funder database). 2. Datasets on dominant sciences and technologies (nanotechnology dataset). 3. Datasets covering firm innovation dynamics 4. Public sector research in Europe with several data on European Higher Education Institutions (RISIS-ETER) and on European public research organizations 	<p>The datasets cover five critical dimensions: ERA dynamics (3 datasets), firm innovation dynamics (3 datasets), public sector research (3 datasets), research careers (3 datasets) and a repository on research and innovation policy evaluations.</p> <p>Several of these datasets are accessible on line</p>	<p>Almost all EU Countries ' Research Projects</p>	<p>English</p>	<p>Text, pictures, graphics</p>	<p>Not listed in the Fandango Agreement Doc</p> <p>REQUIRES DIRECT CONTACT</p>	<p>Access only through accreditation</p>
---	--	--	--	--	----------------	---------------------------------	--	--

		<p>(under development) and on their academic performance (Leiden ranking).</p> <p>5. Research careers with access to the European mobility survey (MORE) and the German panel on doctoral students and their careers (early career facility) and, at a later stage, with access to a platform and/or dataset integrating multiple national sources (under development);</p> <p>6. A specific repository, SIPER, on policy evaluations, articulated with the OECD- World Bank Innovation policy platform) and giving access to the accumulated knowledge on policy instruments and policy mixes.</p> <p>7. Several datasets that provide linked data, such as data from statistical offices, geographical classifications, patents (USPTO), open science (Open-Air), and others. For more information see the SMS Data Store.</p>						
--	--	--	--	--	--	--	--	--

OpenAIRE

NAME	URL	DESCRIPTION, PROVIDER AND AVAILABILITY	MAIN SECTORS	GRANULARITY	LANGUAGE	MEDIA	NOTE/COPYRIGHT	RESOURCES FORMAT
The OpenAIR E 2020 Project	https://www.openaire.eu	50 partners, from all EU countries, and beyond, will collaborate to work on this large-scale initiative that aims to promote open scholarship and substantially improve the discoverability and reusability of research publications and data. The initiative brings together professionals from research libraries, open scholarship organizations, national e-Infrastructure and data experts, IT and legal researchers, showcasing the truly collaborative nature of this pan-European endeavor.	The website has a search feature on 24 million publications and almost 700 thousand datasets on the 2020 Projects	Almost all European Countries	English, French, German, Italian	Text	Not listed in the Fandango Agreement Doc REQUIRES DIRECT CONTACT	Direct Search