Enriching Digital Libraries with Crowdsensed Data
Twitter Monitor and the SoBigData Ecosystem

Stefano Cresci[1], Salvatore Minutoli[1], Leonardo Nizzoli[1,2], Serena Tardelli[1,2](✉), and Maurizio Tesconi[1]

[1] Institute of Informatics and Telematics, IIT-CNR, Pisa, Italy
{stefano.cresci,salvatore.minutoli,leonardo.nizzoli,serena.tardelli, maurizio.tesconi}@iit.cnr.it
[2] Department of Information Engineering, University of Pisa, Pisa, Italy

Abstract. SoBigData is a Research Infrastructure (RI) aiming to provide an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining. A key milestone of the project focuses on data, methods and results sharing, in order to ensure the reproducibility, review and re-use of scientific works. For this reason, the Digital Library paradigm is implemented within the RI, providing users with virtual environments where datasets, methods and results can be collected, maintained, managed and preserved, granting full documentation, access and the possibility to re-use.

In this paper, we describe the results of our effort for integrating the Twitter Monitor, a tool for gathering messages from the Twitter Online Social Network, into the SoBigData RI. The Twitter Monitor provides a simple user interface, enabling researchers and stakeholders, without programming skills, to seamlessly (i) select relevant messages out of the huge Twitter stream by means of language, keyword, user tracking and geographical filters, (ii) store data on user personal Workspace, (iii) and publish them in the SoBigData Resource Catalogue, which implements all the aforementioned Digital Library features.

Thanks to the seamless integration in the SoBigData RI, the Twitter Monitor allows researchers and stakeholders, belonging to different areas and having different backgrounds, to exploit the crowdsensing paradigm for enriching the SoBigData Digital Library. In this way, crowdsensing acquires the key features of openness, accessibility, interoperability and interdisciplinarity that characterize the Digital Libraries framework.

Keywords: Digital libraries · Resource sharing · Online social networks · Crowdsensing

1 Introduction

In the last two decades, *eScience* designed a new research paradigm aiming to produce innovation in collaborative, computationally- or data-intensive research across all disciplines [13,17]. In this context, data sharing plays a key role: the

P. Manghi et al. (Eds.): IRCDL 2019, CCIS 988, pp. 144–158, 2019.
https://doi.org/10.1007/978-3-030-11226-4_12

availability of research datasets, published in open, accessible, fully documented online repositories, ensures the reproducibility of experiments, facilitates the review process and enables re-use for further research [7]. In this open and multi-disciplinary scenario, collaboration and sharing between different disciplines, institutions and scientists become vital.

The *eInfrastructure* paradigm emerged as the most promising approach for enabling those best practices [16]. It is defined as a framework enabling (i) secure, (ii) cost-effective and (iii) on-demand resource sharing across different organizations [8,14]. Within eInfrastructures, it is possible to provide scientists with *Virtual Research Environments* (VRE), defined as *"web-based, community-oriented, comprehensive, flexible, and secure working environments conceived to serve the needs of modern science"* [9]. eInfrastructures also enable the implementation of *Digital Libraries*, virtual environments where research datasets, methods and results can be collected, mantained, managed and preserved, granting full documentation, access and the possibility to perform further analysis [10].

SoBigData[1], a Research and Innovation Action funded by the European Commission under the Horizon 2020 program, creates such eScience ecosystem through the deployment of a Research Infrastructure (RI), which provides an integrated environment for ethic-sensitive scientific discoveries and advanced social data mining and big data applications [15]. As an open research infrastructure, SoBigData promotes repeatable and open science in multiple research fields, including mathematics, ICT, human, social and economic sciences. Interoperability by design enables easy comparison, re-use and integration of state-of-the-art big social data, methods, and services [18]. Hence, SoBigData implements the features of a Digital Library, including accessible and fully documented datasets, open source methods and services, which may impact industrial and other stakeholders (e.g. agencies, non-profit organisations, funders, policy makers).

The SoBigData eInfrastructure grants seamless access to tools, applications, datasets, services, algorithms, catalogues through VREs. The *Twitter Monitor* is a tool, included in the so-called SoBigDataLab VRE, that allows end-users and stakeholders to build, document and share new datasets. It collects data from Twitter in a focused way, by specifying gathering criteria to retrieve only relevant information. This tool is based on the disruptive crowdsensing paradigm, in which people publishing contents on social media platforms act as *social sensors* [2]. Such data can be leveraged to gather information on users activities, preferences, and tastes [1], and to extract public opinion about different topics concerning economy, politics, security, society, and finance [3–5,11,12]. Once retrieved, data can be enriched, documented and published in the SoBigData catalogue, which ensures all the features mentioned as best practices for a Digital Library framework.

Contribution. The contribution of this work is to present our experience in integrating the Twitter Monitor within the SoBigData RI. The purpose of the tool is to enrich and expand the volume and variety of Digital Libraries content, particularly with crowdsensed data. In detail, we provide an example of how a

[1] http://www.sobigdata.eu.

social scientist, laking programming skills, can use the tool to easily perform a data collection task. We then show how the customized datasets of Twitter data can be easily shared in the Digital Library framework implemented in the SoBigData RI. Moreover, We compare the workflow that would have been necessary without using the Twitter Monitor and the SoBigData RI, and we provide a cost-benefit analysis.

Roadmap. This paper is organized as follows: Sect. 2 describes in details the Twitter Monitor sensing tool and its integration in the SoBigData eInfrastructure. Section 3 describes a use-case workflow of an end-user leveraging the Twitter Monitor for a multidisciplinary project developed in his/her VRE. Section 4 focuses on how the Twitter Monitor may contribute in a Digital Library framework. Finally, Sect. 5 draws conclusions and highlights promising directions for future research and experimentation.

2 Twitter Monitor: Features and Integration in the SoBigData RI

Among the aims of SoBigData, there is the capability to provide a set of readily available datasets and methods to scientific communities. Typically, users of the SoBigData eInfrastructure can discover and leverage any of the datasets, released within the SoBigData RI itself, or upload, document and share their owns. The SoBigData RI takes care of hosting, maintaining and granting full, seamless and open access to all the included datasets and methods. In this way, distinguishing features of the eScience and Digital Libraries paradigms, such as collaboration, interdisciplinarity, resource sharing, interoperability, open and seamless access and cost-effectiveness are fulfilled. Recently, another alluring possibility emerged, where end-users and stakeholders are directly given the possibility to build and share new personalized datasets – all within the eInfrastructure – without the need for technical expertise. This novel solution greatly empowers scientists and end-users of the eInfrastructure, thus ultimately further contributing to collaboration, interdisciplinarity and resource sharing, that are fundamental points of the eScience and Digital Libraries paradigms.

The Twitter Monitor, integrated within the SoBigData eInfrastructure, – that we describe in this section – represents one of the first examples of this novel approach. It represents a new data-entry point to the whole eInfrastructure, thus acting as a catalyst for collaboration and sharing between platform users. Given the aims of the SoBigData project, the Twitter Monitor allows easy data collection from social media sources, and in particular, from Twitter[2]. Social media platforms are the most effective, sophisticated and powerful way to gather preferences, tastes, and activities of groups of users in the context of Web 2.0. In turn, this large amount of information may generate in-depth knowledge about topics of interest. As such, because of their massive number of users, their real-time features, and their ease-of-use, social media platforms, such as Twitter, have become a major source of information [1].

[2] https://twitter.com/.

2.1 Twitter Monitor and the Crowdsensing Landscape

In the paradigm of crowdsensing, the crowd of social network users becomes a distributed network of social sensors [1]. Specifically, depending on their awareness and their involvement in the system, users are confronted with either an opportunistic or a participatory sensing approach.

- Participatory crowdsensing: users willingly choose to give their contribution of sensory information to form a body of knowledge. They consciously opt to meet an application request, and they are aware of the sensing action (e.g. by photographing locations or discussing events or by intentionally sending such information to the sensing system). Systems exploiting participatory crowdsensing require intentional participation and must therefore provide incentives to the users to perform such actions.
- Opportunistic crowdsensing: users spontaneously collect and share data as they go for their daily life. In this scenario, relevant data is sensed, intercepted, and collected without user intervention and, in some cases, even without the users explicit knowledge. Opportunistic crowdsensing platforms do not require a specific user base, since they rely on already publicly-available data.

Social networking platforms, such as Twitter, are one of the main source of information for many crowdsensing systems. The Twitter Monitor tool gathers data from Twitter, leveraging the opportunistic crowdsensing approach.

2.2 Features and Usage of the Twitter Monitor

The Twitter Monitor is an interactive tool designed to access the Twitter stream by exploiting the public Twitter Streaming APIs[3], which opens a persistent connection with a stream of tweets. In this way, the tool collects new tweets containing the keywords, plus real-time replies and retweets, until the end of the connection. The tool is able to manage concurrent monitors: it is possible to launch parallel listening sessions (i.e., more than one Twitter crawler at a time) with personalized parameters to collect different sets of data. The Twitter Monitor also offers a set of functionalities, aimed to minimize the loss of data due to network or local machine problems. It is also capable of alerting, detecting and recovering from errors, such as streaming connection failures. In particular, it can automatically handle rate limits imposed by the Twitter APIs[4]. Specifically, the Twitter Monitor accepts three different types of searching parameters, thus allowing high flexibility to the end-users of the eInfrastructure:

- Keywords (so-called *Track* mode): collects tweets containing specific keywords. It is possible to specify simple words, hashtags, by adding the '#' character in front of the word, mentions, by adding the '@' character in front of the word, etc. It is also possible to retrieve data published during the previous week from the crawler start date, thanks to the implementation of the Twitter Search API in the tool. A maximum number of 400 keywords per crawler can be specified.

[3] https://developer.twitter.com/en/docs.html.

[4] https://developer.twitter.com/en/docs/basics/rate-limiting.html.

- Users (*Follow* mode): collects tweets published by or mentioning a specific set of Twitter users. This mode focuses on the users instead of the content of tweets. A maximum number of 5.000 users per crawler can be specified.
- Rectangles (*Location* mode): collects geolocated tweets published within a specific geographical area, defined by the lon/lat coordinates of the corners of a bounding box. The bounding box is represented as a geographical rectangle. A maximum number of 25 bounding boxes per crawler can be specified.

The main limitations of the Twitter Monitor are related to the APIs needed for data acquisition. The Streaming API opens a persistent connection to the real-time tweet stream, therefore tweets published before the opening of the connection cannot be retrieved. To overcome this restriction, we exploit the Search API, which returns the tweets produced no more than one week prior to the crawler start date, but it gives access only to a subset of all the tweets published on the platform. In addition, the search parameters have some limitations in flexibility and extensibility. For example, it's not possible to use regular expressions to retrieve tweets, to search for all inflections of a keyword.

2.3 Twitter Monitor Integration into the SoBigData eInfrastructure

The Twitter Monitor application has been seamlessly integrated in the SoBigData RI as a tool included in the `SoBigDataLab` VRE. It was implemented as a collection of modules. It leverages many functionalities[5] made available by the SoBigData RI, which is build on top of the D4Science eInfrastructure [8], supported by the gCube software system [18]. The key benefits, allowed by the presence of the Twitter Monitor within the SoBigData eInfrastructure, are achieved via a combination of functions of the Twitter Monitor itself, and by means of a tight integration with the functionalities offered by the eInfrastructure. The main functionalities of the SoBigData RI, used in this application, are (i) the management of user interaction, (ii) the management of a user Workspace, (iii) the management of processing modules hosted on a set of nodes, and (iv) the management of database systems. The Twitter Monitor application is composed of three main modules: `TwMonScheduler`, `TwMonCron` and `TwMonCrawler`. In particular, the `TwMonScheduler` is visible to the users in the list of available Algorithms in the SoBigData environment. The `TwMonCron` is launched periodically by the SoBigData RI, by means of appropriate configurations. The `TwMonScheduler` is launched, when needed, by the `TwMonCron` by means of SoBigData RI APIs. Figure 1 shows the structure of the application and the interaction of the modules with the SoBigData RI.

The `TwMonScheduler` is a Statistical Manager Algorithm, and it extends the `StandardLocalExternalAlgorithm` class by overriding methods that allows it to interact with the SoBigData RI. It displays a user interface for collecting the

[5] The documentation for all the platform libraries, functions and methods, mentioned in this subsection, can be found at https://gcube.wiki.gcube-system.org/gcube/GCube_Documentation.

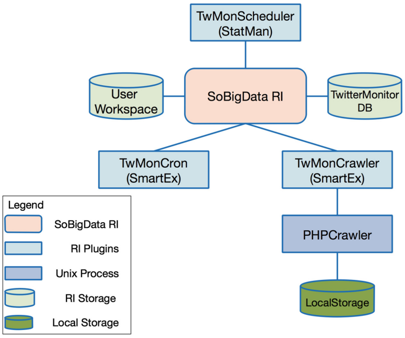


Fig. 1. Twitter Monitor integration in the SoBigData RI environment.

input parameters that will be used to filter the Twitter messages. The user interface is built by the SoBigData RI, based on settings specified by the module. In particular, the SoBigData RI calls a method `setInput`, that the module is required to implement. This method must define the list of needed parameters, along with each own type. Based on this list, the SoBigData RI displays, for each parameter, an appropriate widget to let the user enter the corresponding value. This module stores the parameters in a database called `TwitterMonitorDB`, managed by the SoBigData RI, containing a record for each crawler launched. The reference to this database is obtained by means of a service discover procedure: the `ICFactory` class is used to obtain a `ServiceEndpoint`, given the database identifier (in our case `TwitterMonitorDB`, unique among the SoBigData RI) and the needed credentials. The SoBigData RI finds the node on which the database is currently deployed, and it returns all the information that the `TwMonScheduler` actually needs to connect to the database. The database record also contains a unique identifier for the crawler, the state of the crawling process and a link to the final results. The `TwMonScheduler` periodically queries the database to check if the crawling process terminated, and, when the output result is available, it updates the user interface to notify the user with a message and with a link to download the output file. The output file is stored in the user Workspace available to each registered user so that they can also access it later. Since the processing time could be very long, depending on the period of time selected by the user, he/she is allowed to disconnect from the platform. The crawling process still continues to run until the specified end time.

The `TwMonCron` is a process (actually, a `SmartExecutor` plugin) started periodically by the SoBigData RI. Each time it is launched, it queries the TwitterMonitorDB database to check the state of the crawling processes (called `TwMonCrawler` and described later). It manages to stop each process that reached its end time, and to start the new crawlers inserted by `TwMonScheduler`. It can also detect if some process has terminated before its end time, possibly due to an error. In this case, it launches it again. In order to check, start and stop the `TwMonCrawler`, the `SmartExecutorProxy` API is used: this API allows to manage the state of plugins independently of the actual node on which they are running. More than one TwMonCrawler can run on the SoBigData RI, and each one generally runs in a different node. The SmartExecutor API manages to run a new TwMonCrawler on the most convenient node, and find a particular instance of TwMonCrawler, among all the available nodes, given its unique ID. In particular, its methods getStateEvolution and getPluginState return information on whether the plugin is running or not. The SmartExecutorProxy.launch method allows starting the TwMonCrawler, with given parameters. The SmartExecutor-Proxy.stop method is used to stop a TwMonCrawler, when the end time has been reached.

The TwMonCrawler, a SmartExecutor plugin, is the process that actually collects the information required by the user. It is launched by the SoBigData RI, on behalf of a TwMonCron request: it retrieves the crawling information from the TwitterMonitorDB database and runs a PHP script, as a separate process, providing it with all the needed parameters. The PHP process connects to the Twitter services to receive the selected messages, and it stores them in a local file. The TwMonCrawler continuously monitors this PHP process and, if it stops before the defined end time, it will promptly run it again. When the end time is reached, the TwMonCrawler copies the local output file into the user Workspace. This is done by using the HomeLibrary, that allows accessing the Workspace of the user who launched the crawling process. The JCRWorkspaceFolder API is then used to create folders (createFolder) and the WorkspaceUtil API is used to create files in the user Workspace (createExternalFile). The FolderItem.-getPublicLink creates an HTTP link useful to quickly download the output file, without traversing the Workspace folders. It also creates a link (functionality provided by the SoBigData RI) to this output file, and stores it in the database. The crawling process is finally marked as finished in the database.

3 Twitter Monitor Use-Case in the SoBigData Research Infrastructure

In this section, we describe a simple use-case of the Twitter Monitor within the SoBigData RI. The purporse of this example is to show how the seamless integration of the Twitter Monitor into the SoBigData RI enables researchers, lacking programming skills, to benefit of the crowdsensing approach for gathering Online Social Media data, and to share them contributing to enrich a Digital Library.

We imagine a social scientist, with very basic Computer Science skills, trying to discover which are the most mentioned locations on Twitter related to the *World Tourism Day* (the 27th of September) trending topic. Firstly, we show a general layout of the SoBigData RI components involved in the process, namely the SoBigDataLab and the ResourceCatalogue VREs, summarizing the main available features. Secondly, we describe what the user should do to perform this task *without* the Twitter Monitor and SoBigData RI. Then, we highlight how our tool, interacting with the SoBigData RI, *greatly simplifies* the process. In both cases, we suppose that the user has already created a Twitter App and obtained the tokens necessary to access the Twitter APIs. Finally, we carry out an explicit cost-benefit analysis.

3.1 SoBigDataLab and ResourceCatalogue VREs Features and Usage

Figure 2 depicts a general layout of the features and interactions of two among the many VREs provided by the SoBigData RI, precisely those involved in our use-case example. The SoBigDataLab VRE provides users with a collection of methods and tools, included in the shared Method Engine library, and a personal Workspace, hosting his/her own datasets and results, and accessible only to him/her. Datasets and results can enter the Workspace by means of (i) an upload from the user filesystem, (ii) an import from the shared ResourceCatalogue Digital Library, (iii) an output from a method included in the Method Engine Library. The user can also send data, hosted in his/her Workspace, as an input to a method, and he/she can publish them, together with proper documentation, to the ResourceCatalogue Digital Library. All the aforementioned operations can be performed by means of a web-based, user-friendly interface, provided by the SoBigData RI.

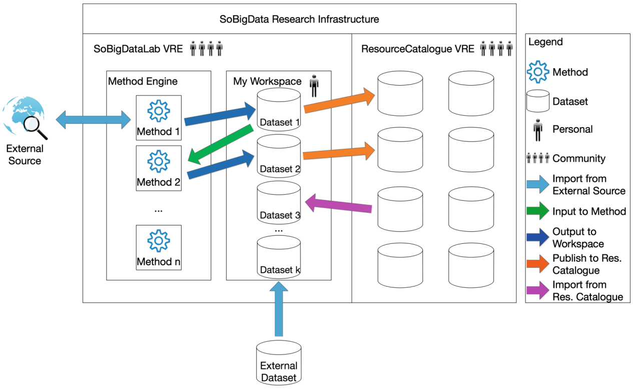


Fig. 2. Layout of the SoBigDataLab and ResourceCatalogue VREs. Each user can access open source methods in the SoBigDataLab VRE. Method engines take datasets as input and perform different tasks. They can acquire information from external sources. Outputs and results are saved into the user personal Workspace, where the user can also upload external datasets. Research data can be shared and published to the ResourceCatalogue VRE, directly from the Workspace.

3.2 Enriching a Digital Library with Crowdsensing Data Without the Twitter Monitor and the SoBigData RI

Twitter, as well as all other popular social media, provides companies, developers and users with programmatic access to its data through Twitter APIs. To use the APIs, collect and store large datasets, one must be familiar with pivotal technologies, such as basic programming languages, Web knowledge, REST and API concepts, JSON[6], user authentication standards such as OAuth[7], and database structures. In particular, the user must be able to (i) authenticate via the OAuth protocol, (ii) establish and maintain a connection to the streaming APIs, specifying the proper parameters to obtain the desired messages, (iii) handle various Twitter errors[8] related to rate limits, connection failures, etc., (iv) consume all messages as soon as they are provided by Twitter APIs, and (v) store the dataset in a proper repository. Therefore, users without proper programming skills are prevented to access this type of data. Moreover, the user should also be able to extract mentioned locations by applying Named-Entity Recognition (N.E.R.) techniques. Finally, the scientist must find a suitable platform for sharing data (e.g. Zenodo[9] or Figshare[10]) and results, in order to enable reproducibility, validation and re-use.

3.3 Enriching a Digital Library with Crowdsensing Data with the Twitter Monitor and the SoBigData RI

The SoBigData RI provides users with various VREs, giving access to many datasets, tools, methods and services. In this way, the task of the social scientist is greatly simplified. The key VREs for our use-case are the SoBigDataLab VRE and the ResourceCatalogue VRE (cfr. Subsect. 3.1). The SoBigDataLab VRE provides personal user Workspace and a set of tools, in this case the Twitter Monitor and the N.E.R. tools. The ResourceCatalogue VRE gives access to a rich set of datasets, and it enables the user to easily make his/her data and results available to the scientific community.

Figure 3 shows a schema of the workflow to accomplish the task in the SoBigData RI. Firstly, the user must access the SoBigDataLab VRE and choose the Twitter Monitor tool in the Method Engine panel[11]. Here, by means of a web-based, user-friendly interface (cfr. Fig. 4), the user can set general parameters of the crawler:

- the name of the crawler, to label and easily retrieve the collected dataset. In this case, the user names it “WTD_Crawler”;
- the language of the tweets to be retrieved (optional). In this case, “en”;

[6] http://www.json.org/.
[7] https://oauth.net/.
[8] https://developer.twitter.com/en/docs/basics/response-codes.html.
[9] https://zenodo.org/.
[10] https://figshare.com/.
[11] https://sobigdata.d4science.org/group/sobigdatalab/method-engine.

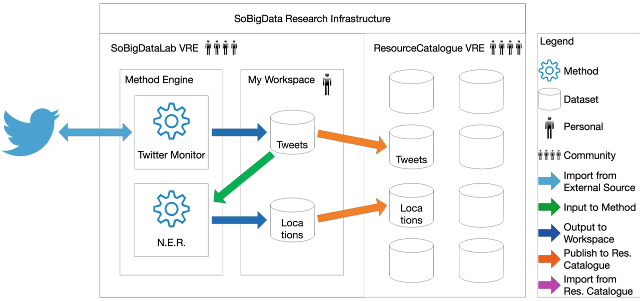
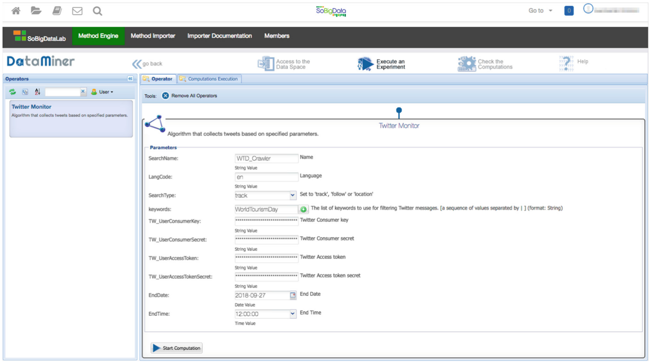


Fig. 3. Workflow of the use-case: the user accesses the SoBigDataLab VRE and launches a Twitter Monitor crawler. The data retrieved is saved into his/her personal Workspace and can be used to initialize a N.E.R. task to extract locations from tweets. The results are again saved into the Workspace. The user may want to publish data and results into the ResourceCatalogue VRE, to make it accessible to the community.



Twitter Monitor
Algorithm that collects tweets based on specified parameters.

Twitter Monitor
Algorithm that collects tweets based on specified parameters.

Parameters

SearchName: WTD_Crawler Name
String Value

LangCode: en Language
String Value

SearchType: track Set to 'track', 'follow' or 'location'

keywords: WorldTourismDay The list of keywords to use for filtering Twitter messages. [a sequence of values separated by |] (format: String)

TW_UserConsumerKey: ****************************** Twitter Consumer key
String Value

TW_UserConsumerSecret: ****************************** Twitter Consumer secret
String Value

TW_UserAccessToken: ****************************** Twitter Access token
String Value

TW_UserAccessTokenSecret: ****************************** Twitter Access token secret
String Value

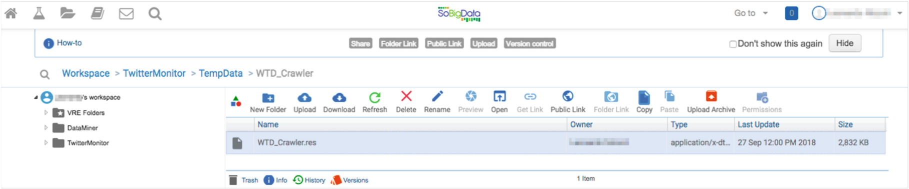
EndDate: 2018-09-27 End Date
Date Value

EndTime: 12:00:00 End Time
Time Value

Fig. 4. The user must set general parameters to launch a Twitter Monitor crawler (e.g., tweet language and keywords).

- the type of searching parameters ("track", "follow" or "location" mode, cfr. Subsect. 2.2). In this case, the user uses the "track" mode;
- the list of parameters for filtering the messages that the user wants to collect. In this case, the user chooses the keyword "WorldTourismDay", which is the official hashtag of the event;
- the user tokens to authenticate via the Twitter OAuth protocol;
- the crawler end date and time.

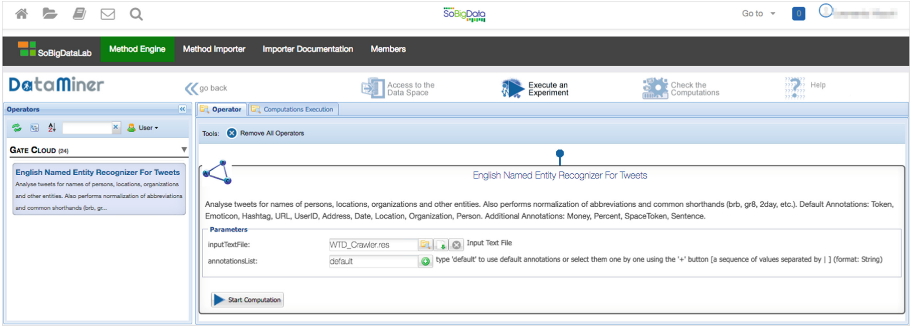
At the end of the crawling process, the user has collected 1,000 English tweets containing the desired keyword.



Name	Owner	Type	Last Update	Size
WTD_Crawler.res	[redacted]	application/x-dt...	27 Sep 12:00 PM 2018	2,832 KB

Fig. 5. In the Workspace, the user can find all his/her datasets and results, perform some basic actions (rename, delete, open, copy files, etc.), download files in local, publish his/her datasets and results into the ResourceCatalogue, and retrieve datasets to re-use from the SoBigData ResourceCatalogue or other external sources.

The dataset is automatically saved in the user personal Workspace[12], as shown in Fig. 5. The platform interface allows the user to download it, or to use it as an input for another method available in the VRE.



English Named Entity Recognizer For Tweets

Analyse tweets for names of persons, locations, organizations and other entities. Also performs normalization of abbreviations and common shorthands (brb, gr...

English Named Entity Recognizer For Tweets

Analyse tweets for names of persons, locations, organizations and other entities. Also performs normalization of abbreviations and common shorthands (brb, gr8, 2day, etc.). Default Annotations: Token, Emoticon, Hashtag, URL, UserID, Address, Date, Location, Organization, Person. Additional Annotations: Money, Percent, SpaceToken, Sentence.

Parameters

inputTextFile: WTD_Crawler.res Input Text File

annotationsList: default type 'default' to use default annotations or select them one by one using the '+' button [a sequence of values separated by |] (format: String)

Fig. 6. Example of a dataset re-use for a Name Entity Recognition task.

In this case, the user wants to extract locations mentioned inside the text of collected tweets. To do so, he/she can leverage the N.E.R. Method Engine available in the VRE [19]. This tool allows the user to input data from his/her Workspace, in this case the dataset just created, and perform the task with a simple click of a button, as shown in Fig. 6. The results are again saved in the user Workspace. The user has therefore completed the task without the need of writing a single line of code and, most importantly, by investing only a small amount of work time.

Datasets and results can be used for further analysis, such a simple visualisation (Fig. 7), and they can be uploaded in the ResourceCatalogue VRE, where other scientists can use them to validate the work or to perform further research (Fig. 8).

[12] https://sobigdata.d4science.org/group/sobigdata-gateway/workspace.

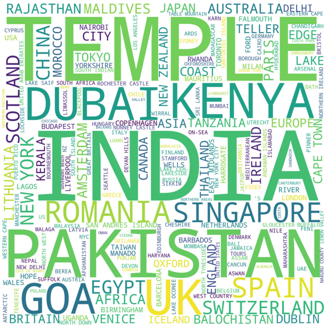
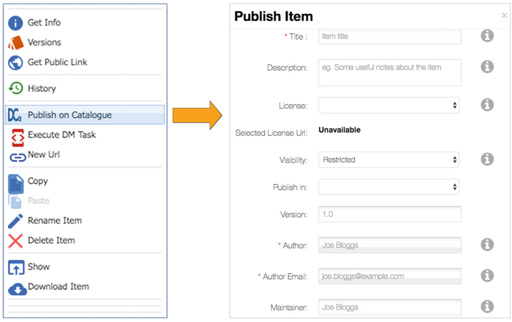


Fig. 7. World Tourism Day locations wordcloud.



Publish Item

* Title: Item title

Description: eg. Some useful notes about the item

License:

Selected License Url: Unavailable

Visibility: Restricted

Publish in:

Version: 1.0

* Author: Joe Bloggs

* Author Email: joe.bloggs@example.com

Maintainer: Joe Bloggs

Fig. 8. The user can publish datasets and results into the catalogue directly from his/her Workspace.

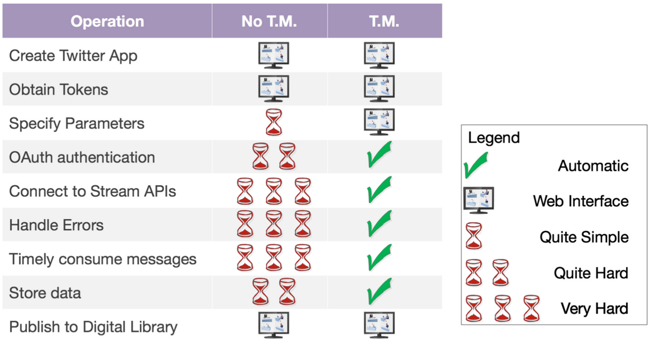
3.4 Cost-Benefit Analysis of the Usage of the Twitter Monitor Tool Integrated in the SoBigData RI

Figure 9 shows a qualitative cost-benefit analysis of the usage of the Twitter Monitor tool for a user aiming to enrich the Digital Library with crowdsensing data.

The common starting point of the two approaches consists in creating a Twitter App and obtaining the tokens, necessary to access the Twitter APIs. This can be easily done leveraging an ad-hoc Web interface[13]; hence, it is not a difficult or time-consuming activity for a user with no technical expertise. Then, the user needs to set the parameters necessary to authenticate to the APIs and to filter the relevant messages. This task does not require much effort, nevertheless it can be a first obstacle for a user that is totally unaware of the basics of programming. As shown in Subsect. 3.3, Twitter Monitor provides a dedicated web-based, user-friendly interface to accomplish this task (cfr. Fig. 4). Much more effort is required for the authentication, via the OAuth protocol, and the connection to the Twitter Stream APIs, which implies to handle all the possible errors and to timely consume and store the retrieved data. Implementing this workflow is far beyond the possibilities of an unskilled user. Instead, all the above mentioned operations are automatically performed in the background by the Twitter Monitor, requiring no effort to the user.

Finally, it is possible to leverage the features provided by the SoBigData RI, in which the Twitter Monitor tool is integrated, to publish the obtained dataset to the SoBigData ResourceCatalogue, which implements all the aforementioned features of a Digital Library. Also in this case, this can be done by means of the platform interface. Otherwise, the user would have to upload data on an external service (e.g.: Zenodo), which can be a time consuming operation in case of large datasets.

[13] https://apps.twitter.com/.



Operation	No T.M.	T.M.
Create Twitter App	🖥	🖥
Obtain Tokens	🖥	🖥
Specify Parameters	⧗	🖥
OAuth authentication	⧗ ⧗	✓
Connect to Stream APIs	⧗ ⧗ ⧗	✓
Handle Errors	⧗ ⧗ ⧗	✓
Timely consume messages	⧗ ⧗ ⧗	✓
Store data	⧗ ⧗	✓
Publish to Digital Library	🖥	🖥

Legend

✓	Automatic
🖥	Web Interface
⧗	Quite Simple
⧗ ⧗	Quite Hard
⧗ ⧗ ⧗	Very Hard

Fig. 9. Cost-benefit analysis of the usage of the Twitter Monitor tool for a user, lacking programming skills, aiming to enrich the Digital Library with crowdsensing data.

4 The Twitter Monitor as a Source for Digital Libraries

The release of research data to other potential users has been extensively discussed in literature [6,7]. The complexity that arises in making research data available is mainly threefold:

1. the difficulty in getting people to collaborate and share data, due to their concerns about the potential misuse of their work, intellectual property rights and credit attribution [6]. Thanks to the SoBigData RI, people can specify the type of license for each resource that they publish in the catalogue (Fig. 8). Furthermore, research data remains within the SoBigData community, as people need to have an account to access openly available datasets;
2. the need of data owners to personally benefit from data sharing and to be incentived in terms of ease of sharing [7]. Data sharing is a way of increasing collaboration and citation rate. To encourage users to share, the SoBigData RI includes the ResourceCatalogue VRE, which act as a Digital Library and makes the publishing task very simple;
3. the uncertainty of how combined sets of data can operate together once shared [7]. The SoBigData RI provides open source methods and services, enabling easy comparison, re-use and integration of shared data into new research, which will, in turn, be shared. In this way, data owners can actually see the contribution of their work to other research.

Because of these reasons, SoBigData RI facilitates data sharing and integration in a variety of scenarios.

The dataset catalogue service, hosted by the ResourceCatalogue VRE in the SoBigData RI, enables users to discover, in a seamless way, information and metadata on the available datasets and datasets itself. The VRE is, to all effects, a Digital Library, in which datasets are accessible to other researchers or stakeholders. In this way, research publications, dataset descriptions and the actual datasets can be linked. This is important to validate the processes applied to data collection, treatment and analysis. Moreover, data and results can be reused, allowing new applications and further research.

The data collected with the Twitter Monitor can be published and made available to the community. This task can be performed in a very straightforward way, by means of the SoBigData RI. Datasets can be published together with proper documentation and metadata, and the infrastructure itself takes care of maintenance and accessibility. In this way, the Twitter Monitor contributes to enrich the SoBigData Digital Library with Twitter datasets, enabling users to apply the crowdsensing paradigm to their research activities.

In the era of big data as the new oil of the digital world, the tight integration of a data-acquisition tool – such as the Twitter Monitor – within an eInfrastructure, contributes to bridge the gap between data and non-technical research communities. In turn, this effort further strengthens the collaboration, sharing and interdisciplinarity within the flourishing eScience ecosystems.

5 Conclusions

In this paper, we described how the Twitter Monitor tool can enrich the Digital Library hosted in the SoBigData RI, with data collected from the Twitter Streaming APIs. We showed the features of the tool, and we described how we seamlessly integrated it within the SoBigData RI. By means of a simple use-case and a cost-benefit analysis, we provided a practical example of workflow, enabling a non-specialist user to retrieve social media content, store it in his Workspace, re-use it for further analysis and share data and results on the SoBigData Resource Catalogue, which implements the functionalities of a Digital Library. The Twitter Monitor contributes to the Digital Library framework by enabling research that exploits the crowdsensing paradigm.

Aknowledgements. This research is supported in part by the EU H2020 Program under the schemes `INFRAIA-1-2014-2015: Research Infrastructures` grant agreement #654024 *SoBigData: Social Mining & Big Data Ecosystem.*

References

1. Avvenuti, M., Bellomo, S., Cresci, S., La Polla, M.N., Tesconi, M.: Hybrid crowdsensing: a novel paradigm to combine the strengths of opportunistic and participatory crowdsensing. In: Proceedings of WWW 2017 Companion, pp. 1413–1421. ACM (2017)
2. Avvenuti, M., Cimino, M.G., Cresci, S., Marchetti, A., Tesconi, M.: A framework for detecting unfolding emergencies using humans as sensors. SpringerPlus **5**(1), 43 (2016)

3. Avvenuti, M., Cresci, S., Del Vigna, F., Fagni, T., Tesconi, M.: CrisMap: a big data crisis mapping system based on damage detection and geoparsing. Inf. Syst. Front. 1–19 (2018)
4. Avvenuti, M., Cresci, S., Marchetti, A., Meletti, C., Tesconi, M.: Predictability or early warning: using social media in modern emergency response. IEEE Internet Comput. **20**(6), 4–6 (2016)
5. Avvenuti, M., Cresci, S., Nizzoli, L., Tesconi, M.: GSP (Geo-Semantic-Parsing): geoparsing and geotagging with machine learning on top of linked data. In: Gangemi, A., et al. (eds.) ESWC 2018. LNCS, vol. 10843, pp. 17–32. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_2
6. Bezuidenhout, L., Chakauya, E.: Hidden concerns of sharing research data by low/middle-income country scientists. Glob. Bioeth. **29**(1), 39–54 (2018)
7. Borgman, C.L.: The conundrum of sharing research data. J. Am. Soc. Inf. Sci. Technol. **63**(6), 1059–1078 (2012)
8. Candela, L., Castelli, D., Pagano, P.: D4Science: an e-infrastructure for supporting virtual research environments. In: Proceedings of IRCDL 2009, pp. 166–169 (2009)
9. Candela, L., Castelli, D., Pagano, P.: Virtual research environments: an overview and a research agenda. Data Sci. J. **12**, GRDI75–GRDI81 (2013)
10. Candela, L., et al.: Setting the foundations of digital libraries. D-Lib Mag. **13**(3/4), 1082–9873 (2007)
11. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. IEEE Trans. Dependable Secure Comput. **15**(4), 561–576 (2018)
12. Cresci, S., Lillo, F., Regoli, D., Tardelli, S., Tesconi, M.: $FAKE: evidence of spam and bot activity in stock microblogs on Twitter. In: Proceedings of ICWSM 2018, pp. 580–583. AAAI (2018)
13. Deelman, E., Gannon, D., Shields, M., Taylor, I.: Workflows and e-Science: an overview of workflow system features and capabilities. Future Gener. Comput. Syst. **25**(5), 528–540 (2009)
14. Foster, I., Kesselman, C., Tuecke, S.: The anatomy of the grid: enabling scalable virtual organizations. Int. J. High Perform. Comput. Appl. **15**(3), 200–222 (2001)
15. Giannotti, F., Trasarti, R., Bontcheva, K., Grossi, V.: SoBigData: social mining & big data ecosystem. In: Proceedings of WWW 2018 Companion, pp. 437–438. ACM (2018)
16. Hey, T., Trefethen, A.E.: Cyberinfrastructure for e-Science. Science **308**(5723), 817–821 (2005)
17. Newman, H.B., Ellisman, M.H., Orcutt, J.A.: Data-intensive e-science frontier research. Commun. ACM **46**(11), 68–77 (2003)
18. Simeoni, F., Candela, L., Lievens, D., Pagano, P., Simi, M.: Functional adaptivity for digital library services in e-infrastructures: the gCube approach. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 51–62. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04346-8_7
19. Tablan, V., Roberts, I., Cunningham, H., Bontcheva, K.: GATECloud.net: a platform for large-scale, open-source text processing on the cloud. Phil. Trans. R. Soc. A **371**(1983), 20120071 (2013)