Social Mining & Big Data Ecosystem

SoBigData

RESEARCH INFRASTRUCTURE

Project Acronym	***SoBigData***
Project Title	***SoBigData Research Infrastructure*** ***Social Mining & Big Data Ecosystem***
Project Number	***654024***
Deliverable Title	***Best practices and guidelines towards interoperability***
Deliverable No.	***D10.1***
Delivery Date	***29 February 2016***
Authors	***Leonardo Candela (CNR), Paolo Manghi (CNR), Pasquale Pagano (CNR)***

SoBigData receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654024

DOCUMENT INFORMATION

PROJECT	
Project Acronym	SoBigData
Project Title	SoBigData Research Infrastructure Social Mining & Big Data Ecosystem
Project Start	1st September 2015
Project Duration	48 months
Funding	H2020-INFRAIA-2014-2015
Grant Agreement No.	654024
DOCUMENT	
Deliverable No.	D10.1
Deliverable Title	Best practices and guidelines towards interoperability
Contractual Delivery Date	29 February 2016
Actual Delivery Date	23 March 2016
Author(s)	Leonardo Candela (CNR), Paolo Manghi (CNR), Pasquale Pagano (CNR)
Editor(s)	Leonardo Candela (CNR)
Reviewer(s)	Valerio Grossi (CNR)
Contributor(s)	N/A
Work Package No.	WP10
Work Package Title	JRA3_SoBigData e-Infrastructure
Work Package Leader	CNR
Work Package Participants	USFD, UNIPI, FRH, UT, IMT, LUH, KCL, SNS, AALTO, ETHZ
Dissemination	Public
Nature	Other
Version / Revision	1.0
Draft / Final	Final
Total No. Pages (including cover)	15
Keywords	Interoperability; Best practice; Standard;

DISCLAIMER

SoBigData (654024) is a Research and Innovation Action (RIA) funded by the European Commission under the Horizon 2020 research and innovation programme.

SoBigData proposes to create the Social Mining & Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by "big data". Building on several established national infrastructures, SoBigData will open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research.

This document contains information on SoBigData core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as SoBigData Board members. Any reference to content in this document should clearly indicate the authors, source, organisation and publication date.

The document has been produced with the funding of the European Commission. The content of this publication is the sole responsibility of the SoBigData Consortium and its experts, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 27 member states of the European Union. It is based on the European Communities and the member states' cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

GLOSSARY

ABBREVIATION	DEFINITION
Research Infrastructure	Facilities, resources and services that are used by a research community to conduct research and foster innovation in their fields. Include: major scientific equipment (or sets of instruments), knowledge-based resources such as collections, archives and scientific data, e-infrastructures, such as data and computing systems and communication networks and any other tools that are essential to achieve excellence in research and innovation. They may be 'single-sited', 'virtual' and 'distributed'.
RI	Research Infrastructure
VA	Virtual Access
Virtual Access	**Open and free access** through communication networks to resources needed for research, without selecting the researchers to whom access is provided.
Virtual Research Environment	Innovative, web-based, community-oriented, comprehensive, flexible, and secure working environments conceived to serve the needs of modern science.
VRE	Virtual Research Environment

TABLE OF CONTENT

DELIVERABLE SUMMARY

Interoperability is one of the key problems to be resolved when building a system as a “collection” of independently developed constituents (systems on their own) that should cooperate and rely on each other to contribute to realise the system tasks. The SoBigData e-Infrastructure is a “system” strongly characterised by such an aggregative nature and thus its development is largely exposed to the interoperability issue. This deliverable describes a growing set of approaches, practices and guidelines aiming at overcoming interoperability issues in the SoBigData e-Infrastructure. It has an “ongoing” nature, i.e. the actual and always up to date set of practices and guidelines characterising the interoperability approaches in SoBigData e-Infrastructure are captured by a set of dedicated Wiki pages.

EXECUTIVE SUMMARY

Interoperability is a crucial problem to be resolved when building a new system out of a “collection” of independently developed constituents (systems on their own) that should cooperate and rely on each other to accomplish the system tasks. This is the case of the SoBigData e-Infrastructure whose most important constituents / resources are systems that have been developed independently of it and are going to be integrated into a unifying resource space.

In order to enact the development of the SoBigData e-Infrastructure a number of solutions, practices and approaches have to be conceived to enable existing systems to operate in the context of the aggregating infrastructure. This “solutions space” is characterised by at least two orthogonal aspects: (a) the typology of resource to be integrated and (b) the level of integration/interoperability to be achieved. In fact, resources range from datasets to stand alone systems and methods while levels of integration range from the discovery of the resource to the actual exploitation. By implementing a given solution for a selected resource, the resource will be conceptually added to the SoBigData e-Infrastructure and made interoperable with the rest of resources to the extent that is characterising the selected solution. Along the infrastructure lifetime it is likely that in the context of the same resource are implemented one or more interoperability solutions with to goal to make it more and more interoperable with the rest, to potentially enlarge the set of scenario where the given resource can be re-used.

This document is a sort of placeholder of the real deliverable that is implemented through an ever updated set of Wiki pages. These Wiki pages will contain the details of each of the proposed solution by clearly describing the resources the specific solution is devised for as well as the level of interoperability that is achieved by implementing it.

1 INTRODUCTION

Interoperability is among the most critical issues to be faced when building a new system as a "collection" of independently developed constituents (systems on their own) that should cooperate and rely on each other to contribute to implement the system tasks.

Although it is a core problem in many systems and application scenario, there is no single definition of this problem which is accepted by the overall ICT community. Even the abstract or generic ones fail to capture the entire problem space as demonstrated by the following definitions. The IEEE Glossary defines interoperability as "*the ability of two or more systems or components to exchange information and to use the information that has been exchanged*" [4]. This definition highlights that to achieve interoperability between two entities (provider, consumer) two conditions must be satisfied: (i) the two entities must be able to exchange information and (ii) the consumer entity must be able to effectively use the exchanged information, i.e. the consumer must be able to perform the tasks it is willing to do by relying on the exchanged information. Wegner defines interoperability as "*the ability of two or more software components to cooperate despite differences in language, interface, and execution platform. It is a scalable form of reusability, being concerned with the reuse of server resources by clients whose accessing mechanisms may be plug-incompatible with sockets of the server*" [11]. He also identifies in *interface standardization* and *interface bridging* two of the major mechanisms for interoperation. Heiler defines interoperability as "*the ability to exchange services and data with one another. It is based on agreements between requesters and providers on, for example, message passing protocols, procedure names, error codes, and argument types*" [7]. He also defines *semantic interoperability* as ensuring "*that these exchanges make sense -- that the requester and the provider have a common understanding of the 'meanings' of the requested services and data. Semantic interoperability is based on agreements on, for example, algorithms for computing requested values, the expected side effects of a requested procedure, or the source or accuracy of requested data elements*". Park and Ram define syntactic interoperability as "*the knowledge-level interoperability that provides cooperating businesses with the ability to bridge semantic conflicts arising from differences in implicit meanings, perspectives, and assumptions, thus creating a semantically compatible information environment based on the agreed concepts between different business entities*" [9]. They also define semantic interoperability as "*the application-level interoperability that allows multiple software components to cooperate even though their implementation languages, interfaces, and execution platforms are different*" [10]. They also state that some of the standards emerging at the time they wrote their paper – e.g. XML, SOAP (Simple Object Access Protocol), UDDI (Universal, Description, Discovery, and Integration), and WSDL (Web Service Description Language) – can resolve many application-level interoperability problems. Assuncion and van Sinderen discussed *Pragmatic Interoperability*, i.e., the interoperability dealing with the mutual understanding in the use of data between the collaborating systems [1]. By analyzing the definition of this term as presented in 44 different papers, they conclude that there are two main interpretations: (i) system level, i.e., sharing the same understanding of the intended and actual use of exchanged system message in a given context; and (ii) business level, i.e., going beyond service use by considering also the compatibility business intentions, business rules, organizational policies, and the establishment and maintenance of trust and reputation mechanisms between collaborating business parties. Their survey results also show that much research emphasis has been given to the system level interpretation. Moreover, they compare this interoperability with syntactic and semantic ones and

affirm that pragmatic interoperability can only be achieved if collaborating systems are also syntactically and semantically interoperable.

In spite of this heterogeneity and uncertainty in characterizing the problems, various approaches and solutions aiming at removing or mitigating it have been developed in different application domains. Such solutions fall in two main categories: (a) *agreement-bases approaches* and (b) *mediator-based approaches*.

Agreement-based approaches consist in agreeing on a set of principles and practices that achieve a limited amount of homogeneity among heterogeneous entities and practices. It is one of the most effective approaches to reach interoperability. Standards belong to this category and the value of standards is clearly demonstrable. The major drawbacks of these solutions reside in the fact that standards and agreements are challenging to agree between different organisations/across diverse application domains. They often end up being complex combinations of features reflecting the interests of many disparate parties. Moreover, by nature they infringe autonomy of the entities adopting them.

Mediator-based approaches have been proposed to resolve scenarios where there is the need to guarantee an high level of autonomy among the partaking entities. These approaches consist in isolating the interoperability machinery and implementing it in components specifically conceived to link the entities partaking to the interoperability scenario. These solutions have been initially conceived in the Information Systems domain [12] and are nowadays used in many cases and realised in many different ways. The most important part of such kind of approaches is represented by the 'mediation function', i.e. the interoperability machinery that is implemented to transform / reconcile the different data models and interaction modes. With respect to agreement-based approaches, mediators are strong in supporting the criteria of autonomy of the entities interconnected.

It is worth to notice that these two classes of approaches are not mutually exclusive, rather solutions belonging to them can be combined each other in real and complex interoperability scenarios.

This deliverable captures the set of approaches agreed when developing the SoBigData e-Infrastructure [2]. Such an e-Infrastructure is strongly characterised by an ever growing set of social mining datasets and social mining platforms and systems identified by the SoBigData Community. This set of resources is gradually going to (a) be integrated into a unifying resource space and (b) "interoperate" in such a context. The actual level of interoperability will depend on the solution put in place for each resource and on the peculiarities of the specific resource.

In the reality this document is only a sort of placeholder of the real deliverable that is implemented through an ever updated set of Wiki pages. These Wiki pages will contain the details of each of the proposed solution by clearly describing the resources the specific solution is devised for as well as the level of interoperability that is achieved by implementing it.

The rest of the document gives a very brief overview of the approaches identified so far and links to the Wiki pages where detailed information are documented un continuously updated.

2 INTEROPERABILITY PRACTICES AND GUIDELINES

The entire set of guidelines and practices designed to enact interoperability in the SoBigData e-Infrastructure are captured by the following wiki page

https://wiki.d4science.org/index.php?title=SoBigData_Interoperability_Guidelines

This section briefly introduces the approaches identified per class of resource to be integrated in the SoBigData e-Infrastructure.

2.1 DATASET INTEGRATION

Datasets are collections of data that are considered as a unit for management purposes. An initial list of datasets to be considered / integrated in the SoBigData e-Infrastructure is described in D8.1 [5].

The first level of interoperability to be achieved consists in enabling users (primarily SoBigData practitioners) to discover, in a seamless way, information (aka metadata) on the available datasets. This will be achieved by explicitly *registering/publishing dataset* information through the dataset catalogue service hosted by the SoBigData Infrastructure. Once registered, datasets can be discovered by search (Google-like and faceted search by tags) and/or browse. For each dataset the catalogue provides the user with a rich set of metadata including:

- descriptive information like title, author/provider, keywords, and description aiming at providing details on the dataset,
- coverage oriented information to characterise the extent (e.g. spatial extent, temporal extent) of the dataset,
- classification oriented information like tags,
- access oriented information like web protocols for accessing the actual data in any of the formats it is made available,
- usage oriented information like licence.

To collect such metadata, SoBigData

- defines a dataset ***application profile*** [6] by building on existing formats like DataCite[1] and DCAT[2],
- requires data providers to register their datasets into the catalogue.

In case datasets are already published in other repositories, e.g. Geospatial data repositories compliant with CWS standard, the integration might be semi-automatic by harvesting the metadata and repurposing them to comply with the SoBigData application profile.

[1] https://schema.datacite.org/
[2] https://www.w3.org/TR/vocab-dcat/

2.2 METHOD INTEGRATION

A method is an implementation of a social mining algorithm / procedure. An initial list of methods to be considered / integrated in the SoBigData e-Infrastructure is described in D10.2 [2].

The integration of this typology of social mining asset relies on:

- the SoBigData Infrastructure information system / registry for discovery purposes;
- the gCube-based data analytics engine [3] equipping the SoBigData Infrastructure for enactment / operation purposes.

A very lightweight integration is achieved by explicitly *registering/publishing method* information through the catalogue service hosted by the SoBigData Infrastructure. Each method is expected to be equipped with a sort of web-based "landing page", i.e. a web-page where users are provided with any information (documentation, download, examples) enabling the user to make use the specific method. The link to this landing page represents a key information to be maintained in the catalogue in addition to the rest of information needed for discovery purposes.

For an effective integration method owners are requested to:

- make their method compliant with the guidelines of the hosting platform according to the methodology described in a dedicated Wiki page[3]. This activity might require some modification / adaptation of the method implementation, e.g. for input parameters specification. The cost of this adaptation depends on the complexity of the method;
- publish the algorithm through the platform. In case the method is implemented with a R script, the platform is provided with a facility supporting this publishing phase[4];

Once integrated, the method becomes a SoBigData social mining asset that:

- will benefit from a distributed and scalable computing platform;
- can be exploited in the context of many virtual research environments and it is suitable for being repurposed / applied to datasets;
- will be automatically made available via a web-based GUI as well as with web-based protocols (SOAP and Rest);
- is monitored and assessed by SoBigData tools, e.g. detailed statistics on usage are transparently collected.

2.3 APPLICATION INTEGRATION

An application is a stand-alone system offering one or more social mining methods. In some cases it offers also some social mining datasets. An initial list of applications to be considered / integrated in the SoBigData e-Infrastructure is described in D10.2 [2].

[3] https://wiki.gcube-system.org/How-to_Implement_Algorithms_for_the_Statistical_Manager
[4] https://wiki.gcube-system.org/Statistical_Algorithms_Importer

The integration of this typology of social mining asset relies on:

- the SoBigData Infrastructure information system / registry for discovery purposes;
- the gCube-based data analytics engine [3] equipping the SoBigData Infrastructure for enactment / operation purposes.

A very lightweight integration is achieved by explicitly *registering/publishing application* information through the catalogue service hosted by the SoBigData Infrastructure. Each application is expected to be equipped with a sort of web-based "landing page", i.e. a web-page where users are provided with any information (documentation, download, examples) enabling the user to make use the specific application. The link to this landing page represents a key information to be maintained in the catalogue in addition to the rest of information needed for discovery purposes.

For an effective integration application owners are requested to:

- reconsider their application architecture thus to extrapolate the methods and the datasets;
- make their methods compliant with the guidelines of the hosting platform according to the methodology described in a dedicated Wiki page[5]. This activity might require some modification / adaptation of the method implementation, e.g. for input parameters specification. The cost of this adaptation depends on the complexity of the method;
- publish the algorithm through the platform. In case the method is implemented with a R script, the platform is provided with a facility supporting this publishing phase[6];
- transform the datasets in publishable assets and publish them as described in Sec. 2.1;

Once integrated, the application actually become a number of SoBigData social mining assets that:

- will benefit from a distributed and scalable computing platform;
- can be exploited in the context of many virtual research environments and it is suitable for being repurposed / applied to datasets;
- will be automatically made available via a web-based GUI as well as with web-based protocols (SOAP and Rest);
- are monitored and assessed by SoBigData tools, e.g. detailed statistics on usage are transparently collected.

2.4 SERVICE INTEGRATION

A service is a web-based facility offering one or more social mining methods. An initial list of services to be considered / integrated in the SoBigData e-Infrastructure is described in D10.2 [2].

The integration of this typology of social mining asset relies on:

- the SoBigData Infrastructure information system / registry for discovery purposes;
- the gCube-based hosting platform equipping the SoBigData Infrastructure for operational purposes.

[5] https://wiki.gcube-system.org/How-to_Implement_Algorithms_for_the_Statistical_Manager
[6] https://wiki.gcube-system.org/Statistical_Algorithms_Importer

In particular, if the service is a Java web service compliant with JAX-RS[7], JAX-WS[8] service owners are requested to:

- produce some simple configuration files according to the guidelines given in the SmartGears Wiki page[9];
- deploy their service in a SmartGear hosting node;

Once integrated, according to this pattern the service becomes a SoBigData social mining asset that:

- will benefit from a distributed and scalable hosting infrastructure, e.g. the SoBogData Infrastructure manager can create one or more instances of such a service;
- can be exploited in the context of many virtual research environments and it is suitable for being repurposed / applied to datasets;
- is monitored and assessed by SoBigData tools, e.g. detailed statistics on usage are transparently collected.

If the service is a non-Java based web service the level of integration is more superficial and supports only discovery. Service owner is actually requested to register the service instance in the SoBigData information system / registry by rich metadata including a web-based access point to use to interact with the service instance.

[7] https://jax-rs-spec.java.net
[8] https://jax-ws.java.net
[9] https://wiki.gcube-system.org/gcube/SmartGears

3 CONCLUSION

This document is a placeholder of the actual deliverable that is implemented via a set of dedicated Wiki pages all available at

https://wiki.d4science.org/index.php?title=SoBigData_Interoperability_Guidelines

These pages capture the array of interoperability solutions and guidelines that have been agreed to develop the SoBigData e-Infrastructure. In particular, these pages describe the set of practices and approaches that have to be put in place by the social mining resource (datasets, methods and platforms) owners to make the selected resources “integrated” in the unifying space of the SoBigData e-Infrastructure resource space.

REFERENCES

[1] Asuncion, C., van Sinderen, M. (2010) Pragmatic interoperability: A systematic review of published definitions. In: P. Bernus, G. Doumeingts, M. Fox (eds.) Enterprise Architecture, Integration and Interoperability, IFIP Advances in Information and Communication Technology, vol. 326, pp. 164–175. Springer Boston

[2] Candela, L., Manghi, P., Pagano, P. (2016) SoBigData e-Infrastructure Release plan 1. SoBigData Project Deliverable D10.2, March 2016

[3] Coro, G.; Candela, L.; Pagano, P.; Italiano, A.; Liccardo, L. Parallelizing the execution of native data mining algorithms for computational biology. Concurrency and Computation: Practice and Experience, Wiley, 2014, doi: 10.1002/cpe.3435

[4] Geraci, A. (1991) IEEE Standard Computer Dictionary: Com- pilation of IEEE Standard Computer Glossaries. IEEE Press, Piscataway, NJ, USA (1991)

[5] Grossi, V., Romano, V., Trasarti, R. (2015) Data Management report. SoBigData Project Deliverable D8.1, December 2015

[6] Heery, R. and Patel, M. (2000) Application Profiles: Mixing and Matching Metadata Schemas. Ariadne, Issue 25. http://www.ariadne.ac.uk/issue25/app-profiles/

[7] Heiler, S. (1995) Semantic interoperability. ACM Comput. Surv. 27(2), 271–273. DOI 10.1145/210376.210392

[8] Paepcke, A., Chang, C.C.K., Winograd, T., Garçia-Molina, H. (1998) Interoperability for Digital Libraries World-wide. Communications of the ACM 41(4), 33–42. DOI 10.1145/273035.273044

[9] Park, J., Ram, S. (2004) Information Systems Interoperability: What Lies Beneath? ACM Trans. Inf. Syst. 22(4), 595– 632. DOI 10.1145/1028099.1028103

[10] Ram, S., Park, J., Lee, D. (1999) Digital libraries for the next millennium: Challenges and research directions. Information Systems Frontiers 1(1), 75–94. DOI 10.1023/A:1010021029890

[11] Wegner, P. (1996) Interoperability. ACM Comput. Surv. 28(1), 285–287. DOI 10.1145/234313.234424

[12] Wiederhold, G. and Genesereth, M. (1997) The conceptual basis for mediation services. *IEEE Expert*, 12(5), 38-47. DOI 10.1109/64.621227