SoBigData RI: European Integrated Infrastructure for Social Mining and Big Data Analytics

Roberto Trasarti[1], Valerio Grossi[1], Michela Natilli[1] and Beatrice Rapisarda[1]

[1]CNR-ISTI (Institute of Information Science and Technologies), via G. Moruzzi 1, 56124 Pisa, Italy

Abstract

SoBigData RI has the ambition to support the rising demand for cross-disciplinary research and innovation on the multiple aspects of social complexity from combined data and model-driven perspectives and the increasing importance of ethics and data scientists' responsibility as pillars of trustworthy use of Big Data and analytical technology. Digital traces of human activities offer a considerable opportunity to scrutinize the ground truth of individual and collective behaviour at an unprecedented detail and on a global scale. This increasing wealth of data is a chance to understand social complexity, provided we can rely on social mining, i.e., adequate means for accessing big social data and models for extracting knowledge from them. SoBigData RI, with its tools and services, empowers researchers and innovators through a platform for the design and execution of large-scale social mining experiments, open to users with diverse backgrounds, accessible on the cloud (aligned with EOSC), and also exploiting supercomputing facilities. Pushing the FAIR (Findable, Accessible, Interoperable) and FACT (Fair, Accountable, Confidential, and Transparent) principles will render social mining experiments more efficiently designed, adjusted, and repeatable by domain experts that are not data scientists. SoBigData RI moves forward from the simple awareness of ethical and legal challenges in social mining to the development of concrete tools that operationalize ethics with value-sensitive design, incorporating values and norms for privacy protection, fairness, transparency, and pluralism. SoBigData RI is the result of two H2020 grants (g.a. n.654024 and 871042), and it is part of the ESFRI 2021 Roadmap.

Keywords

Data Science, Artificial Intelligence, Social Mining, Big Data Analytics

1. Introduction

SoBigData RI (www.sobigdata.eu) is the result of a long-term vision (Fig. 1) and started in 2015 with an initial project called SoBigData funded by H2020 (G.A. n. 654024) and is continuing with a subsequent H2020 project called SoBigData++ (G.A. n. 871042). It also entered the ESFRI RoadMap 2021 and therefore supported in becoming a legal entity in the form of an ERIC as result of a support project called SoBigData PPP (ESFRI Preparatory Phase Project).

At the national level, partners in each country within SoBigData RI participate in national calls to expand their connections internally. For example, the Italian node (which is the coordinator) is participating in the PNRR (*Piano Nazionale di Ripresa e Resilienza*) since SoBigData RI is already in the high priority list in the PNIR *(Piano Nazionale Infrastrutture di Ricerca 2021-2027*).

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy
roberto.trasarti@isti.cnr.it (R. Trasarti); valerio.grossi@isti.cnr.it (V. Grossi); michela.natilli@isti.cnr.it (M. Natilli); beatrice.rapisarda@isti.cnr.it (B. Rapisarda)
0000-0001-5316-6475 (R. Trasarti); 0000-0002-8735-5394 (V. Grossi); 0000-0002-0323-7498 (M. Natilli)

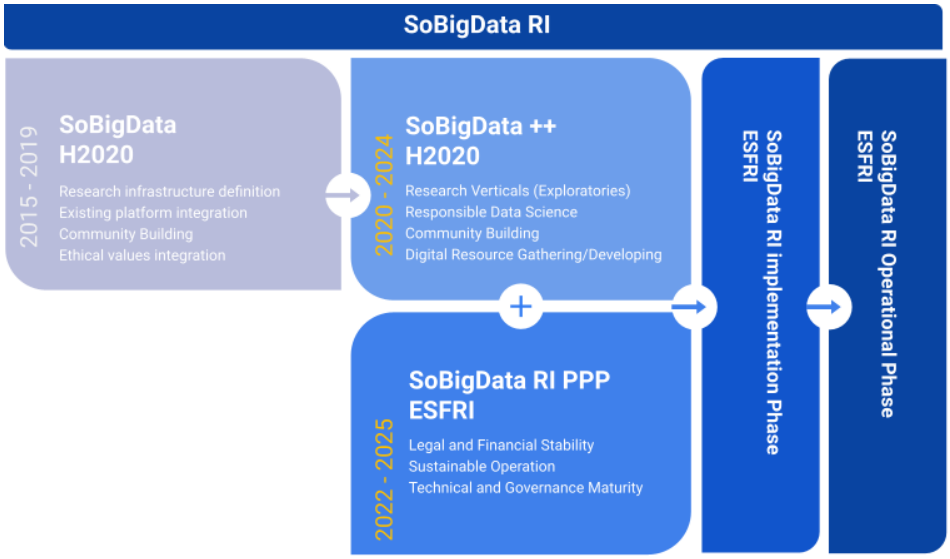

Figure 1: Overview of the SoBigData RI Vision and related projects

In particular, the currently active project is SoBigData++ with the objective of building effective national research systems within a federated European platform by planning, designing, and integrating services in a research infrastructure called SoBigData RI. The aim is to allow users to access methods and algorithms provided by RI as services in the cloud of e-infrastructure to be executed using the SoBigData RI computational resources. This aspect will follow the concepts of meta-modelling as well as FAIR and FACT principles. Optimal transnational cooperation and competition: this aspect will be granted in 13 European countries, offering world-leading research expertise from multiple disciplines, as well as Big Data computing platforms, big social data resources, and cutting-edge computational methods [1]. SoBigData RI plays a key role in impacting the ERA in several aspects. Firstly, training the next generation of responsible social data scientists engaged in the challenging research questions and ambassadors of critical data literacy to facilitate data citizenship and data democracy. Secondly, providing an accelerator of data-driven innovation that streamlines the collaboration with industries and startups to develop pilot projects and proofs-of-concept. Finally, democratizing the benefits of AI, data science, and big data through a network of excellencies within an ethical-legal framework that harmonizes individual rights and collective interests. SoBigData RI is also instrumental in many training activities related to a novel Ph.D. program centered on Data Science in Pisa - https://datasciencephd.eu/ (the Italian node will work synergically with the Ph.D. in Data Science); a second level master on Big Data Analytics and Social Mining; a national program on Ph.D. in Artificial Intelligence. The SoBigData RI promotes the platform's services to other EU projects (in the previous years, we already served many H2020 projects (i.e., HUMAN AI Net, HUMMING BIrd, AI4EU, Tailor, XAI). During the ESFRI PPP, SoBigData RI will become an EOSC service provider. These aspects will impact the creation of synergies between SoBigData RI and other RIs in the European landscape.

In the following sections, we will describe the different pillars of the SoBigData RI, starting

from the Exploratories, which are thematic virtual laboratories where the researchers study the complexity of social phenomena and collect data to produce new methods and algorithms which are published in the RI Catalogue to be shared with the community or executed in the cloud computational resources offered by the platform.

2. Exploratories

SoBigData RI has made a relevant design choice for creating its e-infrastructure and community by introducing the concept 'exploratory', i.e., a vertical social mining research environment focused on specific societal challenges. The aim was to exemplify the cutting-edge, multidisciplinary research supported by the RI and drive the integration into the e-infrastructure of the resources available within the national laboratories [2, 3]. Our exploratories are environments where concrete, substantive multidisciplinary social mining research has been carried out. Moreover, exploratories serve to attract many users, both via transnational access and virtual access, as well as students and innovators attending the project's training and innovation initiatives. Exploratories are the vehicle for fostering cooperation and synergies across different lines of activity within the research infrastructure, promoting networking, access, and joint research. SoBigData RI currently includes 7 exploratories:

Demography, Economy and Finance 2.0 is devoted to the study of traditional, complex socio-economic and financial systems as well as emerging ones (e.g., block-chain and cryptocurrency markets). The exploratory aims to extend the existing approaches grounded on statistical physics for the analysis of real-world (economic and financial) networks by considering innovative models to analyze the dynamics of and on networks and infer their structural details from partial information extending the rich formalism of entropy-based null models towards the development of non-linear Exponential Random Graphs and employ renormalization methods.

Migration Studies tries to estimate flows and stocks from available, real-time data, building models that track indicators extracted from unconventional and official data sources. This exploratory evaluates migrants' integration in new communities through social network and retail data analysis since migration can generate cultural changes in the local and incoming population.

Social Impact of AI and explainable ML studies how the AI can replicate/simulate/infer people's behaviours in the field of computational social choice. This interdisciplinary research area deals with the aggregation of agents' preferences to reach a consensus decision that achieves some social objectives by finding "simpler" ways to explain this to humans.

Societal Debates and Misinformation uses the state-of-the-art algorithms for supervised prevalence estimation that can track collective sentiment even when expressed on fine-grained ordinal scales. We analyze discussions using data from online social networks to investigate topics relevant to society and phenomena such as opinion polarization and echo chambers.

Sports Data Science analyses the way scientists, fans, and practitioners conceive sports performance is changing rapidly due to the proliferation of new sensing technologies that provide reliable data streams extracted from every game. Combining this flow of (big) data with robust data science and AI tools, we now have the chance to unveil the great complexity

underlying sports performances and work towards many challenging goals, from automatic tactical analysis to data-driven performance ranking, from game outcome prediction to injury forecasting.

Sustainable Cities for Citizens develops innovative methodologies to help Italian municipalities and other decision-makers in producing policies and strategies for urban sustainability and improve the environmental performance concerning digitization, energy, water, and waste management and pollution. The methodologies are also directed to citizens to be used as a tool to increase their sustainability awareness. This exploratory also focuses on analyzing human mobility using mobile phone traces, vehicular GPS, and social media data are all Big Data sources and proxies of individual and collective behaviours.

Network Medicine aims to create a new approach to biomedical science, combining principles and approaches from systems biology and network science to both understand the causes of human diseases and find and develop new personalized treatments. This exploratory also studies and develops new compressed indexes and algorithms for storing, indexing, and mining knowledge graphs and key-value stores that are the backbone of modern Computational Biology platforms.

Disaster response and recovery focusing on the research of methods and tools to analyze, monitor, and improve post-disaster reconstruction processes in socio-economic areas, spatial planning, and environmental health in cooperation with national and international institutions. ICT-enhanced solutions to the disaster management cycle's response and recovery phases can improve the efficacy of "search & rescue" activities, emergency relief, reconstruction, and rehabilitation. This exploratory was explicitly created to tackle the problems with a strong connection with the local institution in the area of L'Aquila; for this reason, it has a dedicated access point[1]).

Exploratories had and still have a significant impact on the use of the platform, both in terms of the number of users and experiments carried out; they can be seen as *think tank* producing research, dataset, and algorithm which are refined and offered by the RI.

3. Ethical, Legal, Socio-Economic and Cultural Framework

SoBigData RI adheres to the EU vision on Responsible Research and Innovation and operationalizes values driving the ongoing reform of the EU Data Protection and Fundamental Rights legislation [4]. For this reason, SoBigData RI is operationalizing the ELSEC (Ethical, Social, Legal, Economic, and Cultural Aspects) values within AI and machine learning. FAIR (Findable, Accessible, Interoperable and re-usable) principles [5] are not enough in our work, but we also need FACT (Fairness, Accuracy, Confidentiality, and Transparency) [6] principles. In other words, the goal is to develop a sustainable and ethical framework. This aim relates to several aspects, starting from the ethical and legal ones and including gender balance aspects. From this perspective, SoBigData contributes and will contribute in moving the RI forward from the simple awareness of ethical and legal challenges in social mining to the development of concrete

[1]This exploratory relates to the "Territori Aperti" project - https://territoriaperti.univaq.it/

tools that operationalize ethics with value-sensitive design, incorporating values and norms for privacy protection, fairness, transparency, and pluralism.

This can be demanding in an environment where regulations tend to change. The fact that the potential of technology is not always captured in time by the legislative system can result in gaps that need to be anticipated and, if possible, prevented. We are facing a continuous evolution of the ethical-legal framework. After GDPR, many legislative initiatives are developing new ethical-legal dimensions impacting data processing for statistics and scientific purposes and innovation.

There are numerous challenges to be faced. Since many private companies and industries have difficulties including and exploiting ELSEC principles in their project, we can act as facilitators showing how to include them in practice (as already done in SoBigData). Another fundamental challenge is the potential conflict of interest that can arise. In some cases, there may be a tension between the requirement to operate ethically and the need to develop and market as many services as possible, to as many users and as many different companies as possible.

In this challenging scenario several actions are put in place: 1) the design of a pipeline for data science experiments and services that adheres to the legal-ethical principles but is also effective; 2) an Ethical and legal board (with experts in IT law, IT ethics and data protection) which accommodate within-infrastructure ethical issues as well as will aim at answering new questions on the scope, interpretation, and application of the GDPR along with the expanding role of AI, machine learning and data mining; 3) white papers on ethical issues from actual cases and practical solutions for social science; 4) a variety of privacy-enhancing and discrimination discovery algorithms; 5) guidelines for legal, ethical, methodological and infrastructural issues arising from working with social data, to help scientists to focus on their research; 6) clear policies and agreements for the sharing of resources according to the Privacy and Ethic EU regulations; 7) tools that enable technologies and ethical values transfer to industry and PA, which are not always aware of the responsible data science pipeline, and show how to exploit its full potential in a sustainable way for a long-term added value.

4. SoBigData Research Infrastructure Services

In this section, we describe the services offered by the RI, which are designed for different stakeholders and attract and foster a solid and active user community[2]. There are two kinds of services according to the access type: (i) Virtual Access (VA) gives the user the possibility to navigate and discover datasets, methods, services, and other resources (e.g., papers, experiments, etc.) thanks to the online tools provided by the RI; (ii) Transnational Access (TNA) is an integral part to the RI services to grow the social mining community and widen the reach of the RI.

4.1. Virtual Access

The leading virtual access service is the RI catalogue which contains all the metadata of the resources, and it is accessible through the web interface. To access existing resources of the e-infrastructure, the user must log in to perform a free registration or use any academic/social credentials supported by the EOSC. The access to the SoBigData RI also grants access to the Catalogue, SoBigData Lab, and the e-learning area. The SoBigData VA services have the main

entry point at the site: www.sobigdata.eu. The gateway of the SoBigData RI publishes all the services related to e-infrastructure organized into six main areas: 1) the **catalogue** enables the user to search an item given a set of keywords; 2) **SoBigData Lab** where users can execute methods on the e-infrastructure; 3) the description ant the link to of all the **applications** available; 4) the **e-learning area** that enables the user to access all the training material related to the SoBigData community; 5) the portal to access to our **HPC** facilities, and 6) on the left the **work space**, an online environment to support secure and controlled data storage and sharing. The catalogue is the primary tool for discovering and searching for an item inside the SoBigData RI. All the elements inside the SoBigData RI are discoverable through this service. The user can insert a set of keywords, and the list of the results will be visualized. The search result lists items included in the catalogue and their classification (e.g., Method, Training Material, Dataset). The complete description is provided on the dedicated page, accessible by clicking on the item. These features can be added to the search filter, which will be recalculated in real-time. The search result can be sorted alphabetically concerning the insertion date or popularity. The SoBigData Lab integrates different methods that can be invoked under the same environment through SoBigData e-infrastructure. A method is the implementation of an algorithm/procedure or is an algorithm that requires an engine to be executed. Different kinds of integration are available based on the programming language in which a method is implemented. Once a method is integrated into the platform, the final user has a homogeneous web form for inserting parameters and invoking it independently from the programming language employed. In particular, JupyterHub is easily accessible by clicking on the link on the top of SoBigData Lab VRE. After starting the server by selecting one of the default profiles available, the user can start to use Jupyter Notebook as the local version. Several services are offered in the form of SoBigData libraries, consisting of a set of thematic methodologies available on the cloud and usable both in a JupyterHub environment or in the SoBigData engine and applications.

4.2. Transnational Access

TA visit is a powerful tool to drive to share knowledge and expertise throughout Data Science. By benefitting from a SoBigData TA visit, the researcher is also guided and supported to become a responsible and ethical Data Scientist; this, in turn, will promote one of our primary objectives of creating and maintaining a trusted Research Infrastructure with a highly respected status in the field. TNA, by definition, shares and disseminates expertise from many of the partners of our community. The goal is to provide researchers and professionals with access to big data computing platforms, big social data resources, and cutting-edge computational methods. A TNA visitor will interact with the local experts, discuss research questions, run experiments on non-public big social datasets and algorithms, and present results at workshops/seminars.

The RI expertise offer covers multiple and wide-ranging skills and knowledge. The exploratories described in Section 2 provide areas that pull together the strands of the RI across lateral and vertical thematic environments for supporting TNA visits. The RI can now offer a visit to one of 17 European SoBigData nodes providing wide-ranging and varied expertise. The visit is a short-term scientific mission (STSM) between 2 weeks and 2 months, covering the travel and living expenses costs[2].

[2]http://www.sobigdata.eu/content/call-2022-23-sobigdata-transnational-access

Figure 2: SoBigData RI in numbers

5. Technological infrastructure

The technological model of SoBigData RI has the form of a system of systems considering the principles of **autonomy** of constituents (independence and evolution), **openness** (join and leave; dynamic reconfiguration), and **distribution** (interdependence and interoperability). The cited principles represent the building blocks of the policies governing the rules of participation of the national nodes. From the technical point of view, the RI is a hyper-converged infrastructure, including both the storage area network and the underlying storage abstractions are implemented virtually in software rather than physically in hardware [7, 8]. All physical resources of the infrastructure are manageable through a single platform, reducing inefficiencies and reducing the total cost of ownership. Moreover, the technological model adopted considers some requirements for integrating with EOSC to provide its services and integrating with services providers such as OpenAIRE, Zenodo, EuroHPC, etc.

6. Conclusion

The paper presented the main impacts and services that SoBigData RI has in the research community, mainly related to social mining and AI. Our ambition is to update our repertoire of social mining continuously and AI methods (from knowledge extraction, multilingual text classification, and enhanced privacy technology to federated machine learning) with cutting-edge techniques spurring from partners' research activities or wrapping useful new openly-available methods. In this perspective, Figure 2 completes the SoBigData RI description providing some numbers related to our activities from 2015 to now. At the moment (March 2022), we have more than 9000 registered users, we fully support more than 80 TNA visits, and we trained more than 750 in our courses. We organized events disseminating and organizing workshops on thematic areas, e.g., Special Action on Gender, and with the presence at European Parliament.

In the next future, we think that SoBigData RI will contribute to the creation and maintenance of local ecosystems that go in the direction of dynamic and open data spaces aligned with the ambition of the European strategy for data.

7. Acknowledgments

This work is supported by the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n.871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu).

References

[1] F. Giannotti, R. Trasarti, K. Bontcheva, V. Grossi, Sobigdata: Social mining & big data ecosystem, in: P. Champin, F. Gandon, M. Lalmas, P. G. Ipeirotis (Eds.), Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018, ACM, 2018, pp. 437–438. URL: https://doi.org/10.1145/3184558.3186205. doi:10.1145/3184558.3186205.

[2] V. Grossi, F. Giannotti, D. Pedreschi, P. Manghi, P. Pagano, M. Assante, Data science: a game changer for science and innovation, Int. J. Data Sci. Anal. 11 (2021) 263–278. URL: https://doi.org/10.1007/s41060-020-00240-2. doi:10.1007/s41060-020-00240-2.

[3] V. Grossi, B. Rapisarda, F. Giannotti, D. Pedreschi, Data science at sobigdata: the european research infrastructure for social mining and big data analytics, International Journal of Data Science and Analytics 6 (2018) 205–216.

[4] N. Forgó, S. Hänold, J. van den Hoven, T. Krügel, I. Lishchuk, R. Mahieu, A. Monreale, D. Pedreschi, F. Pratesi, D. van Putten, An ethico-legal framework for social data science, International Journal of Data Science and Analytics 11 (2021) 377–390.

[5] A. Gvishiani, M. Dobrovolsky, A. Rybkina, Big data and fair data for data science, in: Resilience in the Digital Age, Springer, 2021, pp. 105–117.

[6] J. van de Hoven, G. Comandè, S. Ruggieri, J. Domingo-Ferrer, F. Musiani, F. Giannotti, F. Pratesi, M. Stauch, Towards a digital ecosystem of trust: Ethical, legal and societal implications, Opinio Juris In Comparatione (2021) 131–156.

[7] M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, V. Marioli, P. Pagano, G. Panichi, C. Perciante, F. Sinibaldi, The gcube system: Delivering virtual research environments as-a-service, Future Gener. Comput. Syst. 95 (2019) 445–453. URL: https://doi.org/10.1016/j.future.2018.10.035. doi:10.1016/j.future.2018.10.035.

[8] M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, P. Pagano, G. Panichi, F. Sinibaldi, Enacting open science by d4science, Future Gener. Comput. Syst. 101 (2019) 555–563. URL: https://doi.org/10.1016/j.future.2019.05.063. doi:10.1016/j.future.2019.05.063.