Project Acronym	***SoBigData***
Project Title	***SoBigData Research Infrastructure*** ***Social Mining & Big Data Ecosystem***
Project Number	***654024***
Deliverable Title	***SoBigData e- Infrastructure release plan 1***
Deliverable No.	***D10.2***
Delivery Date	***01 March 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	SoBigData e- Infrastructure release plan 1
Contractual Delivery Date	01 March 2016
Actual Delivery Date	10 March 2016
Author(s)	Leonardo Candela (CNR), Paolo Manghi (CNR), Pasquale Pagano (CNR)
Editor(s)	Leonardo Candela (CNR)
Reviewer(s)	Valerio Grossi (CNR), Paolo Manchi (CNR)
Contributor(s)	Thorsten May (FRH)
Work Package No.	WP10
Work Package Title	JRA3_SoBigData e-Infrastructure
Work Package Leader	CNR
Work Package Participants	USFD, UNIPI, FRH, UT, IMT LUCCA, LUH, KCL, SNS, AALTO, ETHZ
Dissemination	Public
Nature	Report
Version / Revision	1.0
Draft / Final	Final
Total No. Pages (including cover)	26
Keywords	Release planning; D4Science; SoBigData e-Infrastructure;

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
LtR	Learning to Rank
NLP	Natural Language Processing
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
SNA	Social Network Analysis
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
WP	Work Package

TABLE OF CONTENT

DELIVERABLE SUMMARY

This deliverable describes the development plan characterizing the release and development of the SoBigData e-Infrastructure. This is the first of three versions of the plan, each describing the actions associated with a specific version of the infrastructure to be made available at M12 (August 2016), M24 (August 2017) and M36 (August 2018). In particular, the deliverable focuses on the plan leading to the first release of the SoBigData e-Infrastructure at M12.

The deliverable consists of four sections and one appendix:

- **Section 1** briefly introduces the role of this deliverable for the development and delivery of the SoBigData e-Infrastructure.
- **Section 2** describes the main systems that will be used to build the SoBigData e-Infrastructure.
- **Section 3** documents the actual plan guiding the development of the SoBidtata e-Infrastructure.
- **Section 4** concludes the report.
- **Appendix A** describes an initial analysis of the decisions to be taken and steps to be implemented in order to integrate one of the social mining platform and systems, i.e. the Visual Analytics Platform.

EXECUTIVE SUMMARY

SoBigData WP10 is called to support the development of the SoBigData e-Infrastructure in close collaboration with other work packages that are respectively called (a) to operate the infrastructure to provide virtual access to the integrated resources (WP7), (b) integrate existing and newly collected datasets in the infrastructure (WP8), and (c) integrate existing tools and methods for mining social data in the infrastructure (WP9). In particular, WP10 puts in place actions comprising: (i) studies and definition of best practices/policies for the harmonization of federated resources available at the local infrastructure sites; (ii) support for adaptation of existing resources to the identified best practices; and (iii) realization of VREs supporting scientists in benefitting from the integration of the federated resources and infrastructures.

This deliverable describes the development plan characterising the release and development of the SoBigData e-Infrastructure. This is the first of three versions of the plan, each describing the actions associated with a specific version of the infrastructure to be made available at M12 (August 2016), M24 (August 2017) and M36 (August 2018). In particular, the deliverable focuses on the plan leading to the first release of the SoBigData e-Infrastructure at M12.

1 INTRODUCTION

SoBigData is a research infrastructure (RI) for ethic-sensitive scientific discoveries and advanced applications of social data mining to the various dimensions of social life, as recorded by "big data". It is planned to serve the wide cross-disciplinary community of data scientists involved in social mining, i.e., researchers studying all aspects of societal complexity from a data- and model-driven perspective, including data and text miners, visual analytics researchers, socio-economic scientists, network scientists, political scientists, humanities researchers, and more.

In order to serve its "designated community"[1], the project will set up an e-Infrastructure providing "virtual access" to the resources (datasets and services) of interest, namely datasets and methods for social mining. In particular, SoBigData will implement an e-Infrastructure that is "open" and "aggregative" by design, i.e. it is conceived to aggregate into a unifying resource space resources coming from many and heterogeneous providers. In order to do this, it will rely on the D4Science e-Infrastructure [1]. D4Science offers a common ground for hosting the domain specific resources and dynamically building and operating *Virtual Research Environments* (VREs) offering specific and web-based working environments to target communities [2]. In order to be able to serve the needs of the social mining community, it is of paramount importance to invest effort in adapting and extending the resources currently owned by the SoBigData community thus to make them benefitting from the e-Infrastructure capacity. Adaptation and extension of existing resources goes in the direction to integrate them into a unifying space. Depending from the "level of integration" that is achieved for each resource it will be possible to support and guarantee a diverse level of management ranging from a simple discovery to their repurposing to better serve the needs arising in a specific VRE.

This deliverable briefly describes the SoBigData e-Infrastructure that is planned to be developed and operated and presents the incremental plan leading to it.

The reminder of the deliverable is organised as follows.

Section 2 describes the main systems that will be used to build the SoBigData e-Infrastructure, namely D4Science, the set of social mining datasets resulting from an inventory, and the set of social mining platforms and systems initially identified within the SoBigData Community (i.e. Elianto, GATECloud, MAtlas, Quick Rank, SNA Toolkit, TagMe, Twitter Streaming Monitoring Framework, Visual Analytics Platform).

Section 3 describes the actual plan guiding the development of the SoBidtata e-Infrastructure. In particular, it sketches the incremental and evolving nature of the release plan and the need to rely on an agile methodology. It also identifies the components that characterise the first release of the e-Infrastructure.

Section 4 concludes the report.

[1] This term is borrowed from the archival community here it is used to refer to an identified group of potential consumers that should be able to understand the particular set of information and benefit from them. The group may consist of multiple communities and may change over time.

In Appendix A it is reported an initial analysis of the decisions to be taken and steps to be implemented in order to integrate one of the social mining platform and systems, i.e. the Visual Analytics Platform.

2 THE DEVELOPMENT AND OPERATIONAL CONTEXT

The SoBigData e-Infrastructure will not be build from scratch rather its development will benefit from an existing e-Infrastructure – D4Science – that will play the role of "aggregative infrastructure" collecting resources and services from the infrastructures and resource providers participating in SoBigdata. Figure 1 gives a conceptual view of the SoBigData e-Infrastructure.

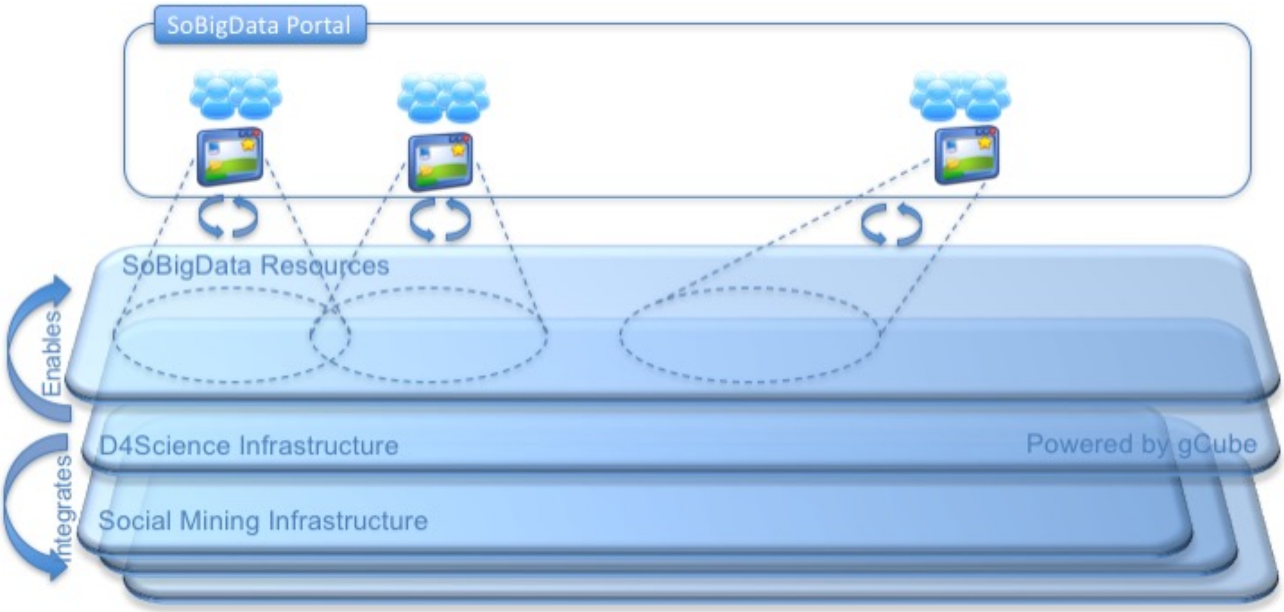


Figure 1. SoBigData Infrastructure

2.1 THE D4SCIENCE HYBRID DATA INFRASTRUCTURE

D4Science[2] [1] is an operational Hybrid Data Infrastructure that (a) supports the creation and management of Virtual Research Environments by dynamically acquiring the resources (data, tools, services, computing) from a resource space, (b) is designed to integrate resources from other e-Infrastructures and (commercial) vendors/providers by using a "system of systems" strategy, (c) is offering unified access to the integrated resources by abstracting from the underlying e-Infrastructures, (d) is designed to maximize resources exploitation and minimize operational costs, and (e) relies on the gCube open source technology that offers a feature rich set of services for data management and collaborative work. Over time D4Science supported a set of projects and initiatives promoting the development of virtual research environments in various application domains [1] including iMarine (fisheries management and marine living resources conservation), EUBrazilOpenBio (biodiversity), and ENVRI (environmental science).

2.2 SOBIGDATA RESOURCES TO BE INTEGRATED: A PRELIMINARY LIST

SoBigData e-Infrastructure is a RI whose primary objective is to support social mining practitioners by providing them with seamless access to datasets and methods of interest. Thus it is of paramount importance to integrate these domain specific resources in a unifying resource space. The se of resources to be integrated is expected to evolve during the project lifetime, e.g. new

[2] https://www.d4science.org/

datasets might become available, new methods might be developed. This section identifies a preliminary list of resources so far identified.

2.2.1 DATASETS

A preliminary list of datasets worth to be integrated in the SoBigData e-Infrastructure has been identified and described by D8.1 Data Management report [7]. The list of datasets is expected to grow during the project lifetime. The picture emerging from the census concluded in November 2015 identified a total of 63 datasets covering five of the six thematic clusters:

- Human Mobility Analytics: 15 datasets;
- Social Data: 6 datasets;
- Social Network Analysis: 15 datasets;
- Text and Social Media Mining: 21 datasets;
- Web Analytics: 6 datasets.

No dataset for Visual Analytics has been identified yet.

Regarding accessibility, the survey identified that there are 22 datasets suitable for virtual access (10 public and 12 with restricted access) while there are 38 datasets suitable for transnational access and 27 are actually private.

2.2.2 ELIANTO

Elianto[3] is a platform hosting an open-source, user friendly and re-active Web interface to support the crowdsourced creation of gold standard datasets for entity linking and salient entities recognition. It supports human labelling of semi-structured documents through a guided two-step process. Collections of unstructured or structured documents can be uploaded on the platform and the guided annotation task easily monitored.

2.2.3 GATECLOUD

GATECloud[4] is a unique, cloud-based infrastructure for large-scale, data-intensive NLP and text mining research. Important infrastructural issues are dealt with by the platform, completely transparently for the researcher: NLP algorithm distribution, load balancing, efficient data upload and storage, deployment on the virtual machines, security and fault tolerance. Another unique feature of this specialized NLP Platform-as-a-Service is its support for researchers who want to develop and run their own text mining/NLP pipelines on big data.

It offers a growing number of NLP and text mining services for multiple European languages. Currently only around 30% of the algorithms and tools from the GATE infrastructure have been made available through GATECloud, and the number will continue to grow during the project lifetime. GATECloud offers a web interface for data upload, social media data collection, programming-less execution of the GATECloud text analytics services, and results download.

[3] http://elianto.isti.cnr.it/

[4] https://gatecloud.net/

2.2.4 MATLAS

M-Atlas[5] is a mobility querying and data mining system centered onto the concept of trajectory. Besides the mechanisms for storing and querying trajectory data, M-Atlas has mechanisms for mining trajectory patterns and models that, in turn, can be stored and queried. M-Atlas is equipped with a querying and mining language making the analytical process possible and providing the mechanisms to master the complexity of transforming raw GPS tracks into mobility knowledge.

Other important facets of M-Atlas include (i) the privacy-preserving data publishing and mining techniques, designed to transform trajectory datasets into anonymous forms in such a way that strong privacy-protection guarantees can coexist with high data utility (ii) the analysis of different forms of mobility data, such as mobile phone call records, characterised by complementary weaknesses and strengths with respect to GPS trajectories.

2.2.5 QUICK RANK

Quick Rank[6] [4] is an efficient Learning to Rank (LtR) toolkit providing several C++ implementations of specific algorithms. In particular, the toolkit offers an implementation of the following algorithms:

- *GBRT*: J. H. Friedman. Greedy function approximation: a gradient boosting machine. Annals of Statistics, pages 1189–1232, 2001.
- *LamdaMART*: Q. Wu, C. Burges, K. Svore, and J. Gao. Adapting boosting for information retrieval measures. Information Retrieval, 2010.
- *Oblivious GBRT / LamdaMART*: Inspired to I. Segalovich. Machine learning in search quality at yandex. Invited Talk, SIGIR, 2010.
- *CoordinateAscent*: Metzler, D., Croft, W.B.: Linear feature-based models for information retrieval. Information Retrieval 10(3), pages 257–274, 2007.

2.2.6 SNA TOOLKIT

The SNA Toolkit is a collection of datasets and methods offered by the various project partners, with particular attention to scalable methods for Community Discovery, Link Analysis, Evolutionary Analysis and Multidimensional Network modeling. The methods in the SNA toolkit are currently used by many students and scientists through the various interfaces provided by the individual partners. For instance, the DEMON[7] software developed by CNR has become a reference in the area of community discovery in complex networks, it is widely used.

2.2.7 TAGME

TagMe[8] [6] is a “topic annotator” that is able to identify meaningful sequences of words in a short text and link them to a pertinent Wikipedia page. TagMe results are one of the best topic-

[5] http://www.m-atlas.eu/

[6] http://quickrank.isti.cnr.it/

[7] http://kdd.isti.cnr.it/~giulio/demon/

[8] http://acube.di.unipi.it/tagme/

annotators to date [5] due to its high efficiency and effectiveness in annotating also very short texts, such as tweets and news.

It is an open-source software (released under Apache License 2.0) and is freely accessible over the Web via a RESTful API[9]. The API allows users to annotate text in English or Italian and, in addition, to execute some useful basic functionalities for text mining: such as, Wikipedia graph exploration, computing the relatedness between two Wikipedia Pages, and returning the spots which point to a Wikipedia page and contain a specified term. These functionalities and TagMe itself are the building blocks upon which more sophisticated tools can be designed for solving more sophisticate text analysis and information retrieval applications.

2.2.8 TWITTER STREAMING MONITORING FRAMEWORK

Twitter Streaming Monitoring Framework is a tool for gathering data from Twitter in a focused way. Using mechanisms offered by Social Media for collecting data not only useful information are retrieved. To reduce the consequent noise in the dataset the tool offers the possibility to specify gathering criteria and collect only data relevant for the experiment purposes.

It helps data scientist to build datasets for their experiments, in real time and in a focused way. The gathering criteria can be expressed by the users in three ways: as one or more keywords, stating a user, or drawing a geographical area to monitor. In this way only tweets matching the gathering criteria are used for building the dataset. Moreover, the tool offers different facilities to manage different gathering processes at the same time.

2.2.9 VISUAL ANALYTICS PLATFORM

The Visual Analytics Platform is a local Infrastructure which offers services for the visualization and especially the visual analysis of research data. It is developed and maintained by the Fraunhofer Group. The services offered by the platform are primarily interactive visualization techniques. As a starting point, the repository contains implementation of generic techniques, yet the platform should be capable to attract both users and contributors alike.

The Visual Analytics Platform will draw upon ideas, which already have been implemented in the information and visual analytics server (or IVA-Server for short). From an analytics perspective, the IVA-Server is a test bed to configure, store and execute analytical services as workflows.

Being a visualization server, the IVA-Server offers dedicated service functions to transform visualization content to a client format and to handle interaction events from a visualization client. Workflows are composed of atomic software artefacts, which include data import and transformation routines, analytical techniques and – of course – visualization techniques. Visualization techniques can be applied to any (intermediate) step of the workflow. An executed workflow represents an analytical service, and multiple services can be managed and executed at the same time.

[9] http://tagme.di.unipi.it/

3 THE SOBIGDATA E-INFRASTRUCTURE RELEASE PLAN

Like many projects and initiatives, release deadlines in SoBigData are fixed / imposed by contractual obligations, i.e. it is planned to make available the first version of the e-Infrastructure at M12 (August 2016), while the second and third version are planned to be released at M24 (August 2017) and M36 (August 2018) respectively.

However, because of the highly aggregative nature of the SoBigdata e-Infrastructure where key resources (social mining data and methods) are coming from seven national infrastructures and the enabling infrastructure is actually built by properly exploiting the D4Science services (cf. Sec. 2) it is possible to envisage an incremental and evolving strategy following some agile principles. One of the major advantages of this strategy is to release a working e-Infrastructure into the users' hands as quickly as possible thus to make "course corrections" as soon as possible.

The deadlines are satisfied while the features (actually the datasets and methods) to be made available by an infrastructure release actually depend from a series of factors including:

- the "level of integration" expected, i.e. SoBigData has developed a series of guidelines and best practices to make existing resources interoperable in the context of the e-Infrastructure [3]. Depending on the amount of guidelines and best practices implemented the resource will result integrated and interoperable with the rest to some extent. In the course of the project, the level of integration will be properly tuned depending on the needs and expectations raised by the community;
- development teams' velocity differences, i.e. the integration of existing resources is expected to be driven by resource owners and service providers. These actors are in the best position to reconsider their products, to identify the major limitations with respect to the guidelines for interoperability, and to plan concrete steps and timelines leading to enhanced versions of their resources;
- a negotiation between users' expectations/desiderata and effort needed to satisfy them, i.e. the availability of a resource within the e-Infrastructure range from the simple discovery to the actual reuse. Diverse users might have different expectations with respect to the actual reuse of a resource and reconsidering the implementation of a given resource thus to serve the expected reuse scenario has a cost for the resource provider;
- dependencies among resources in a user scenario, i.e. when combining a series of existing resources in a certain exploitation scenario the workflows assume that the resources are integrated to a well expected level/degree. Thus the willingness to implement a certain exploitation scenario poses specific requirements on diverse resources and on how these resources have to be integrated.

These factors and the willingness to provide users with a working e-Infrastructure as quickly as possible call for a planning activity that is almost continue and where it is possible to correct the course as the project go.

Because of this SoBigData will largely rely on an "issue tracking system", i.e. a service for managing and maintaining lists of "issues" or "tickets". Each ticket focuses on a specific activity/problem and describes the current state of the art and plan including a due date.

In the following it is first described how the SoBigData e-Infrastructure is planned to be incrementally developed and then how the issue tracking system will actually be used to collaboratively develop and maintain an up to date development and delivery plan.

3.1 SOBIGDATA E-INFRASTRUCTURE VERSION 1 AND FUTURE VERSIONS

The first version of the SoBigData e-Infrastructure will be comprised of the following basic components:

- The ***SoBigData Portal***, i.e. a Liferay[10] based web-portal customized to interface with the D4Science infrastructure and equipped with gCube portlets. This portal will act as the “one stop shop” for the entire SoBigData e-Infrastructure. Through it users will have access to the resources and Virtual Research Environments created to serve the needs of the SoBigData community and scenarios;
- The ***SoBigData Virtual Organisation***, i.e. an organizational structure and a set of basic services created and operated in the context of D4Science to serve the needs of SoBigData. This Virtual Organisation realizes the actual operational context for realizing and operating the SoBigData e-Infrastructure and its resources in autonomy with respect to the other communities and initiatives supported by D4Science;
- The ***SoBigData Resource Catalogue***, i.e. a core services where all the resources contributing to form the SoBigData e-Infrastructure are expected to be registered thus to make it possible for clients to discover them and be informed on their characteristics for, e.g. properly using them. This catalogue is expected to serve both (a) human users willing to know the offering of the e-Infrastructure in terms of datasets and services / methods and (b) other services willing to dynamically discover resources to consume / interact with to deliver their services;
- Services supporting the ***creation and operation of VREs***, i.e. a rich array of gCube-based services enacting the creation and operation of Virtual Research Environments by relying on the available resources (those appearing in the Catalogue);
- Services supporting the ***collection of resources metrics***, i.e. an array of services automatically collecting per-resource usage metrics.

Subsequent versions will be continuously released thanks to the integration of ***community specific resources***, i.e. the datasets and facilities described in Sec. 2.2 as well as new ones that will be identified during the project. An example of the complexity of this integration task is reported in Appendix A where the case of the Visual Analytics Platform is sketched. When dealing with the exercise of actually integrating a facility that is implemented like a “system on its own” to operate in the context of e-Infrastructure and eventually outsource part of its business logic to other components there are a plenty of details and choices to consider and reconsider. The guidelines and best practices the project has developed [3] have to be transformed in concrete steps taking into account the peculiarities of the specific resource to be integrated. That’s why the project is embracing an agile-driven planning and development strategy supported by a ticketing system and a largely distributed development team.

[10] http://www.liferay.com/

3.2 DYNAMIC PLANNING

The entire set of development and deployment tasks leading to the various releases of the SoBigData e-Infrastructure will be captured and monitored via a Redmine instance dedicated to this project and available at

https://support.d4science.org/projects/sobigdata-eu

By using this environment the teams called to develop and deliver the subsequent versions of the SoBigData e-Infrastructure will be able to capture both

- software development tasks (actually complex workflows consisting of multiple and interrelated tasks) needed to gradually adapt each target resource to the operational environment by implementing the project guidelines and best practices [3];
- software deployment tasks needed to plan the actual action of bringing a certain version of the resource into effective action in the context of the SoBigData e-Infrastructure.

By using this knowledge base every project member is informed of the state of the art and the planned activities with their detailed time plan and can intervene to identify potential issues and propose corrective actions (e.g. reconsider a deadline, reconsider an implementation decision) to be agreed with the proper team.

4 CONCLUSION

This deliverable described the development plan characterising the building and release of the SoBigData e-Infrastructure. This e-Infrastructure is primarily conceived to serve the needs of social mining practitioners by providing them with “virtual access” to the resources of interest, namely datasets and methods for social mining. In particular, SoBigData will implement an e-Infrastructure that is “open” and “aggregative” by design, i.e. it is conceived to aggregate into a unifying resource space resources coming from many and heterogeneous providers.

The development and release plan is driven by an incremental and evolving strategy following some agile principles. One of the major advantages of this strategy is to release a working e-Infrastructure into the users’ hands as quickly as possible thus to make “course corrections” as soon as possible.

This is the first of three versions of the plan, each describing the actions associated with a specific version of the infrastructure to be made available at M12 (August 2016), M24 (August 2017) and M36 (August 2018). In particular, the deliverable focused on the plan leading to the first release of the SoBigData e-Infrastructure at M12. It described what are the services to be made available and how the infrastructure will be developed to make it possible to integrate further resources and release enhanced versions.

APPENDIX A. VISUAL ANALYTICS PLATFORM: A PRELIMINARY PLAN

A.1 Introduction and Rationale

The Visual Analytics Platform is a local Infrastructure which offers services for the visualization and especially the visual analysis of research data. It is developed and maintained by the Fraunhofer Group in the SoBigData project. With this local infrastructure we plan to offer services which integrate to a larger ecosystem of data providing services and other analytical services. The services offered by the platform are primarily interactive visualization techniques. As a starting point, the repository contains generic techniques, but the platform should to attract both users and contributors alike. While any infrastructure imposes some standards to its its services, contributors may add or modify techniques to adapt to specific use cases and designs. Visualization services are designed as modular units, which are usually part of larger service compositions.

A vital platform must attract a balance between users and contributors. In business terms, it reflects a two-sided-market. Hence, the development efforts should prioritize the visible value for the different user groups. The different user groups and their immediate implications to the design of the platform, define the outset of the following discussion.

The rationale of this chapter is to specify the architecture of a local service platform and to localize the specific leverage points to interconnect between the local and the global infrastructure. Because we are at the beginning of SoBigData, some parts of the the architecture, specifically the interplay between different layers of the technology stack, are not fixed yet. Wherever possible we will identify services, which are likely provided elsewhere in the "global" SoBigData e-Infrastructure – this especially applies to so-called "cross cutting" services that have to be used by all other techniques (like virtualization, user-management, etc.). Instead of *defining* such services, we will outline them as *requirements* from the perspective of a local infrastructure.

A.2 Current Starting Point – IVA-Server

The Visual Analytics Platform will draw upon ideas, which already have been implemented in the information and visual analytics server (or IVA-Server for short). From an analytics perspective, the IVA-Server is a testbed to configure, store and execute analytical services as workflows. Consequently, its architecture combines a number of **core services**:

- Secure User management
- Data source management
- Workflow configuration
- Workflow execution & Concurrency

Being a visualization server, the IVA-Server offers dedicated service functions to transform visualization content to a client format and to handle interaction events from a visualization client. Workflows are composed of atomic software artefacts, which include data import and transformation routines, analytical techniques and – of course – visualization techniques. Visualization techniques can be applied to any (intermediate) step of the workflow. An executed workflow represents an analytical service, and multiple services can be managed and executed at the same time.

Currently the atomic functions are *not* services itself; hence, the IVAS-workflows do not represent full-fledged service compositions. Instead the functions actually are service implementations only, which are registered to the execution engine. They are added to the service functions of any executed workflow, which are in turn exposed by the execution service. In summary, the workflow and execution services act as the gatekeepers to control the configuration and execution from the client side. The available **generic service functions** include,

- A visualization of the output of any given step (given that the data schemas are compatible)
- Managing user interaction
- The modification of parameter settings of individual steps
- Rearrangement of the workflow. Routines may be added and connected to the workflow.

These service functions essentially allow for high-level (visual) programming and visual feedback in flexible analytical workflows. However, with a typical application setting, the workflow itself is not visible and remains fixed during an entire session. Thus, interaction is managed by a boundary service which maps REST calls from the client to the corresponding steps in the running workflow. In the IVA-Server architecture, the role of this service is to decouple the workflow management and execution from the communication protocol and client application itself.

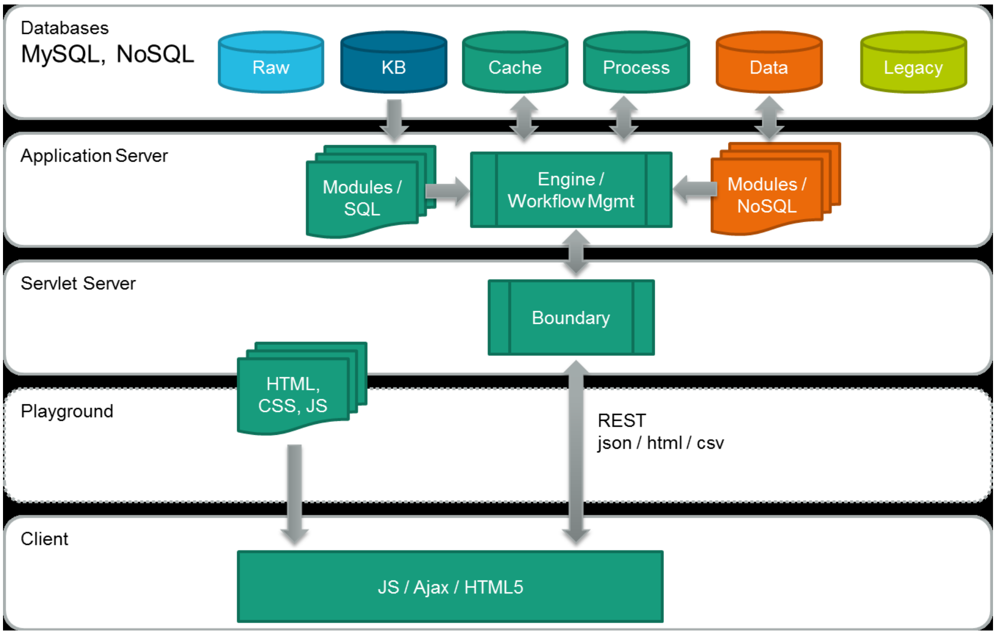
The core services and the generic service function are part of the infrastructure. The workflows itself represent the services that generate the operative value. The main drivers for setting up an infrastructure are the developments towards marketability and the ability for new researchers to step into an existing solution, instead of developing everything from scratch to slowly work towards the state-of-the art. Another driver is the growth of the infrastructure by the contribution of its users in terms of **operative functions (= steps)** used to build the workflows from. These steps include:

- Connectors to data sources
- Data transformation
- Analytical techniques
- Visualizations

Again, these routines are not services, because their use is not independent from their implementation. All routines are build in terms of some constraints imposed by the workflow configuration and execution (apart from the Visualizations, they are all implementations against an interface defined the current IVA-Server SDK).

The IVA-Server has been developed with a slightly different focus, than offering an infrastructure of services. The core services offer an efficient integration and recombination of the operative functions for a workflow composition. Every workflow defines a service, with the service functions defined by the boundary interface. From this perspective, the IVA-Server primarily is a programming and execution environment for analytical services.

However, as a single software artefact, a workflow is far less reusable than its individual components. In SoBigData we will focus on two features, that complement the existing architecture and strengthen its value as an infrastructure. The first feature is an isolation of smaller software artefacts for reuse as (micro-) services. In particular, this refers to individual visualizations, but also may apply to other operative functions. The second feature is the improvement of accessibility of the service collection. IVA-Server already features a repository of the steps available and visible during workflow composition.



A.3 Technology Stack

The technology stacks of the IVA-Server and also the Visual Analytics Platform are loosely defined by the data processing pipeline. In this section we have a focus on the differences between the IVA-Server and the Visual Analytics Platform stacks, because these form the leverage points for the following releases.

One fundament for the technology stack are the data(base) management technology. The IVA-Server offers a number of connectors to different data sources and internally uses standardized data structures across its implementations to maximize internal compatibility. By virtue of this connectors, the IVA-Server is able to access external services, that are persisting their results to accessible storages. Of course, the concept of connectors is not constrained to proper data sources, and other kinds of services may be used as a source. However, by this design, the connection between the IVA-Server and its data sources is asymmetric: The workflows hosted by the IVA-Server are able to “choose” its source connectors, and consequently any upstream data-processing pipelines, and exchange sources even online. However, other services may not choose the IVA-Server as a target. In terms of utility, this might not even be a problem. In terms of usability, this asymmetry makes the design and use of multiple infrastructures more complicated.

Thus, **eliminating such a service asymmetry is one of the design goals** for the Visual Analytics Platform. Services offered by the visual analytics platform should appear no different than other services in the infrastructure.

The visualization is rendered on the client. Their technology is based on the HTML5-standard, which basically is an extension to HTML to include multimedia documents. The rendered images of the visualization are embedded in the HTML DOM as a scalable vector graphic (SVG). For interactive or animated visualization the modification of the HTML DOM is done with Javascript scripts, which are either triggered by user events, or events issued from the endpoint of the REST-interface. On top of the script modification of the HTML DOM we settle libraries like D3 or react, which allow for a more convenient approach build interactive visualization. Other than the definition of the REST-endpoint, the IVAS does not impose constraints on the implementation of visualization techniques in the client.

A given visualization cannot be located at a single spot in the IVAS-architecture. It is naturally part of the client usually being responsible for the rendering and handling simple interaction. It is part of the webserver hosting the html and scripting and it is part of the boundary service, containing the server endpoint of the REST-interface. Thus, **isolating a visualization for reuse is another design goal** for the visual analytics platform. "Isolating" does not necessarily mean to create a single piece of software. For visualization it actually means to create a self-contained package, which is ready for deployment as a service on the platform.

A.4 Release Plan

Integration towards the SoBigData e-Infrastructure

The release plan basically defines the stepwise elimination of "open requirements" about integration and use. From the current implementation level it heads into the two directions of the technology stack. The first direction – integration, towards the infrastructure – defines how the local Visual Analytics Platform will be integrated/connected to the entire SoBigData infrastructure. Corresponding open questions are, for example

	Issue	Status
Service Description	How are services specified described and registered across the infrastructure?	Planned/Pending: Service specification and registering will be defined in the following WP10 Deliverables.
Service Execution	Where are services to be executed?	Solved: Services are usually hosted by the contributors of the platform. Consequently visualization services will be hosted on IGD-Servers.
Service Execution	What is actually registered as a service? (Executable package or running software)	Solved: Services will be registered as a resource. Both forms are allowed. With the services hosted on our premises, only the running service will be registered as a service. Note that also platforms and data are counted as resources.
Service Execution	Are we able to migrate service instances to different hosts (i.e. on demand)?	Partially Solved: The migration of service instances is not a mandatory feature for the SoBigData eInfrastructure. As a default policy, the services are hosted by the respective project partners. In general, this is a question, which is basically dependent on the hosting architecture and access policies.
Workflow Execution	How is the interplay between different services be organized?	Open: The global infrastructure does not prescribe composition and execution environments for entire workflows and the connection of services. This even might require a global standardization of services, which is impossible to impose to a global, open ecosystem. Hence, manual composition of service is a likely alternative (including the implementation of mediators, to connect incompatible services, where necessary).

		Pending: The local infrastructure (IVA-Server) does provide a composition and execution environment on standardized software modules. How external services can be integrated into this infrastructure, will be tested for the following releases.
Workflow Execution	Are workflows operated by central execution services or do they emerge from the connected behaviour of otherwise independent service instances?	Technically, both are possible. However, services that communicate directly to other services (e.g. by calling them) do not remain independent (= highly reusable) software. We favor a central execution here (either manual or automatic).
Workflow Definition	How are complex workflows defined?	Open: For entirely manual composition and execution (see above), no definition is needed at all.
Workflow Definition	Are workflows always linear pipelines?	Open: Meaningful visualizations often relate raw data and processed data, or data from multiple sources in general. Only digesting the results from the end of a single pipeline often excludes the context for interpretation. Hence, visualization – and perhaps other services as well – need to access intermediate results from different steps of the pipeline. A number of strategies can be implemented to solve this problem.
Workflow Definition	Service Nesting	Open: The IVA-Server architecture currently features the composition of services from analytical methods that are not services. External Services from other sources (i.e. partners) might be used from withing the IVA-Server/Visual Analytics platform by different means.

Services and their platforms can be nested. Hence, most of these questions affect the internal behaviour of a local platform, but they also define the external interface between the platform and an entire ecosystem. For the IVA-Server some of these questions have already been answered. We assume, that **this very document may act as a condensation point** for the discussion about requirements. We furthermore assume, that the **SoBigData e-Infrastructure may allows for flexible solutions to these answers.** Nevertheless, we propose the synchronize releases between the global and local infrastructure to match mutual requirements and specific solutions.

Release Plan, Visual Analytics Platform

Feature	Details / Rationale	Planning Date
Platform Site (v0.1)	Essentially a website for the convenient access to the core services of the platform. In the first release, the primary goals are the • publishing of the portal • developer section	TBD

	• registration service site See the attached image for a mock-up.	
Software architecture for visualization services (v0.1)	Visualization services are different, because of their eponymous feature. Their main output is visualization code or images. The architecture includes the technology stack (libraries, toolkits) used for its implementation. *(Target: Developers)* *Note: Development for the SoBigData Exploratories will be used to test the architecture drafts.*	TBD
Visualization Testing Site (Platform Site v0.2)	Draft version for a core service of the platform. Users may test the function of visualization services operating on different source data and/or formats. Behind the scenes, the service attaches to a running service instance and offers the means to change input or parameters to test the feasibility of the service. Furthermore, this service allows integrators to become familiar with the REST API and to estimate the costs for integration. *(Target: Users)*	TBD
Visualization Service Description (preliminary version v0.1)	For registration and finding services on the platform, these will be defined in terms of a description standard. This standard will at least include a categorization of the visualization type, input data, interaction, and output, if applicable. It needs to be defined in parallel to the global infrastructure service description standard and it will adopt this description as a more general subset in the following releases. *(Target: Developers & Users)*	TBD
Software Development Kit & Packaging Instructions (v0.1)	Templates for the development are the most convenient strategy to enforce implementation and architecture standards. The SDK include boilerplate templates to create a simple working visualization from scratch. It also includes instructions to package the executable software for registration and deployment on the platform. *(Target: Developers)*	TBD
Registration Service (v0.1) [Layered*)]	A core service of the Visual Analytics Platform, it will ensure new visualization techniques to be included to the platform's internal service repository. Note that the global infrastructure maintains its own registration service. Service Descriptions on the local	TBD

	platform represent a superset of the description of the global platform. *) Hence, to avoid duplicate registration effort, the services can be layered in the following releases. *(Target: Developers & Users)*	
Deployment Service	A core service (or service function) of the Visual Analytics Platform. Services registered on the Visual Analytics Platform may be hosted on IGD’s servers. Given any package of a executable during registration, the service may be deployed on demand. Note that IGD will maintain a pragmatic uptime reliability in the first release. *(Target: Developers & Users)*	TBD
Visualization Gallery (Platform Site v0.2)	The visualization gallery is the visible manifestation of the repository of registered services (see mock-up). It requires the visualization services to be registered with an example dataset or at least a screenshot. The gallery function will be used to show featured services, the entire list of services (if applicable) or the result of a service search. *(Target: Developers & Users)*	TBD
Registration Service (v0.2)	This version reflects the first integration between the service description on the Visual Analytics Platform and the SoBigData e-Infrastructure. This version aims at ensuring the compatibility between the different descriptions, with the SoBigData e-Infrastructure defining the more generic representation. *(Target: Developers)*	TBD
Software architecture for other services (v2)	The visual analytics platform will contain other services than “pure” visualizations. These may include reusable data transformation, pre-processing techniques or other generic calculations. Either this means a revision of the first architecture, or the specialization of this architecture as needed. *(Target: Developers)*	TBD
Search Service [Layered*)]	A core service (or service function) of the platform. It allows a user to search and browse the services matching a given user query. It is depending on the service description and the registration service. Note that this service, like the registration is also covered by the SoBigData e-Infrastructure. As a nested platform, this service may be layered. However, a	TBD

	more detailed description language focusing on visualization may warrant for more detailed search functions. *(Target: Developers & Users)*	

REFERENCES

[1] Candela, L., Castelli, D., Manzi, A., Pagano, P. (2014) Realising Virtual Research Environments by Hybrid Data Infrastructures: the D4Science Experience. International Symposium on Grids and Clouds (ISGC) 2014, Proceedings of Science PoS(ISGC2014)022

[2] Candela, L., Castelli, D., Pagano, P. (2013) Virtual Research Environments: An Overview and a Research Agenda. Data Science Journal, Vol. 12, pp. GRDI75-GRDI81, DOI 10.2481/dsj.GRDI-013

[3] Candela, L., Manghi, P., Pagano, P. (2016) Best practices and guidelines towards interoperability. SoBigData Project Deliverable D10.1, March 2016

[4] Capannini, G., Dato, D., Lucchese, C., Mori, M., Nardini, F. M., Orlando, S., Perego, R., Tonellotto, N. (2015) QuickRank: a C++ Suite of Learning to Rank Algorithms. Proceedings of the 6th Italian Information Retrieval Workshop (IIR 2015). Cagliari (Italy)

[5] Cornolti, M., Ferragina, P., Ciaramita, M. (2013) A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web* (WWW '13). ACM, New York, NY, USA, 249-260. DOI 10.1145/2488388.2488411

[6] Ferragina, P., Scaiella, U. (2010) TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management* (CIKM '10). ACM, New York, NY, USA, 1625-1628. DOI 10.1145/1871437.1871689

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

[8] Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela, L., Castelli, D., Pagano, P. (2014) The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures. Program: electronic library and information systems, Vol. 48 Iss: 4, pp.322 – 354, DOI 10.1108/PROG-08-2013-0045