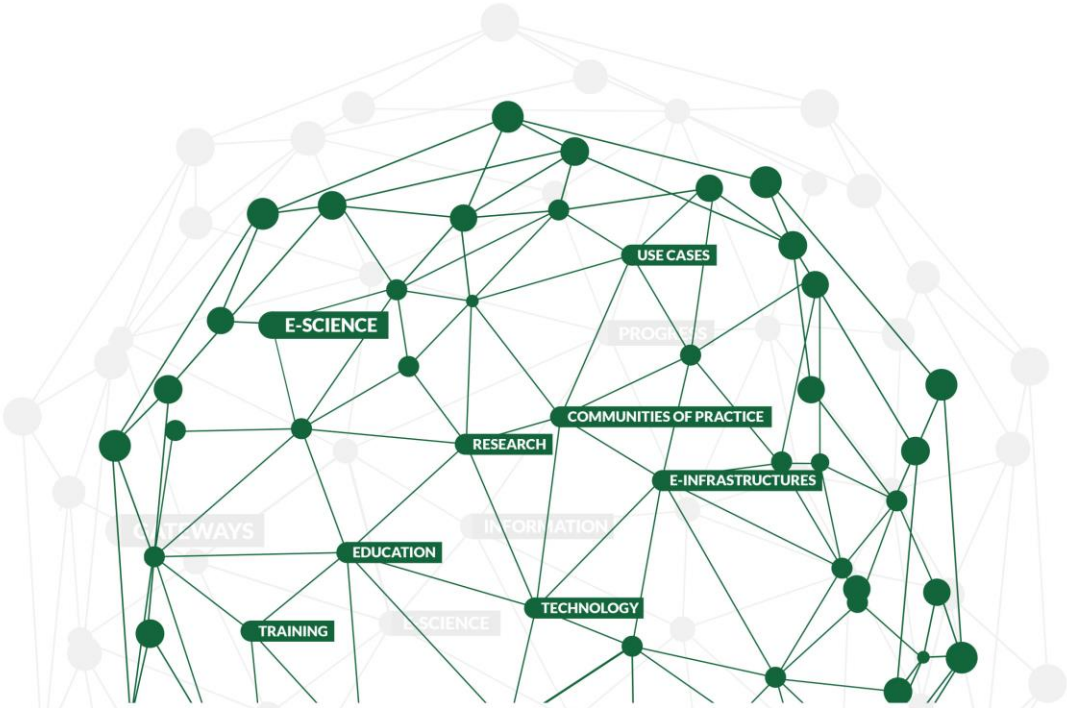
Funded by the
European Union's
H2020 Programme

DATA MANAGEMENT PLAN
October 2015



Sci-GaIA

Energising Scientific Endeavour through Science Gateways
and e-Infrastructures in Africa

PROJECT DOCUMENTATION SHEET	
Project Acronym	: Sci-GaIA
Project Full Title	: Energising Scientific Endeavour through Science Gateways and e-Infrastructures in Africa
Grant Agreement	: GA #654237
Call Identifier	: H2020-INFRASUPP-2014-2
Topic	: INFRASUPP-7-2014
Funding Scheme	: Coordination and Support Action (CSA)
Project Duration	: 24 months (May 2015 - April 2017)
Project Officer	: Leonardo Flores Añover, Unit C.1, DG CONNECT : European Commission
Coordinator	: Dr. Simon J. E. Taylor, Brunel University London (UK) - BRUNEL
Consortium partners	: Brunel University London (UK) - BRUNEL : Sigma Orionis (France) - SIGMA : The UbuntuNet Alliance for Research and Education (Malawi) - UBUNTUNET : University of Catania (Italy) - UNICT : The West and Central African Research and Education Network (Ghana) - WACREN : The Royal Institute of Technology (Sweden) - KTH : The Dar es Salam Institute of Technology (Tanzania) - DIT : Karolinska Institutet (Sweden) - KI : CSIR/Meraka Institute (South Africa) - CSIR
Website	: www.sci-gaia.eu

DELIVERABLE DOCUMENTATION SHEET	
Number	: Deliverable D5.1
Title	: Data Management Plan
Related WP	: WP5 (Management)
Related Task	: Task 5.3 (Data Management)
Lead Beneficiary	: BRUNEL
Author(s)	: Simon Taylor (BRUNEL) – simon.taylor@brunel.ac.uk : Alexandra Cornea (SIGMA) – alexandra.cornea@sigma-orionis.com
Contributor(s)	: Roberto Barbera (UNICT)
Reviewer(s)	: Gheorghe Ghinea (BRUNEL)

Nature	: R (Report)
Dissemination level	: PU (Public)
Due Date	: October 30, 2015 (M6)
Submission date	: October 30, 2015 (M6)
Status	: Final

QUALITY CONTROL ASSESSMENT SHEET			
Issue	Date	Comment	Author
V0.1	01/10/2015	First draft	Alexandra Cornea (SIGMA)
V0.2	22/10/2015	Second draft	Simon Taylor (BRUNEL) WP Leader
V0.4	29/10/2015	Quality check	Camille Torrenti (SIGMA) WP Leader
V0.7	29/10/2015	Quality check	Simon J. E. Taylor (BRUNEL) Coordinator
V1.0	30/10/2015	Submission to the EC	Simon J. E. Taylor (BRUNEL) Coordinator

DISCLAIMER

The opinion stated in this report reflects the opinion of the authors and not the opinion of the European Commission.

The commercial use of any information contained in this document may require a license from the owner of that information.

All Sci-GaIA consortium members are also committed to publish accurate and up to date information and take the greatest care to do so. However, the Sci-GaIA consortium members cannot accept liability for any inaccuracies or omissions nor do they accept liability for any direct, indirect, special, consequential or other losses or damages of any kind arising out of the use of this information.

ACKNOWLEDGEMENT

This document is a deliverable of the Sci-GaIA project, which has received funding from the European Union's Horizon 2020 Programme for Research, Technological Development and Demonstration under Grant Agreement (GA) Nb #654237.

Executive summary

As part of the limited pilot action on open access to research data, Sci-GaIA has implemented a limited pilot action on open access to research data based on the “Guidelines on Data Management in Horizon 2020”. This document specifies the Data Management Plan (DMP) for the project and has created a detailed outline of our policy for data management.

TABLE OF CONTENTS

1 INTRODUCTION

As part of the limited pilot action on open access to research data, Sci-GaIA has implemented a limited pilot action on open access to research data based on the "Guidelines on Data Management in Horizon 2020". As part of the overall project Management work package (WP5), this has been captured in task T5.1 Data Management and specifies the Data Management Plan (DMP) by creating a detailed outline of the project policy for data management. As specified in the Guidelines, this will consider the following:

- Determine if the project will produce new data or combine existing data
- Identify the data sources used and produced during project and the related file formats
- Describe how you will implement a Quality Assurance procedure (QA) for data collection
- Explain your strategy for preventing data loss: files organization and indexing, data backups and storage
- Depending on the dissemination level of each dataset, explain how you will ensure (1) data confidentiality, (2) restricted access, or (3) data high visibility
- Explain how data management tasks and responsibilities are distributed among partners and how they cover the entire data life cycle of the project

This document therefore outlines the first version of the project DMP. The Sci-GaIA DMP primarily lists the different datasets that will be produced by the project, the main exploitation perspectives for each of those datasets, and the major management principles the project will implement to handle those datasets. The purpose of the DMP is to provide an analysis of the main elements of the data management policy that will be used by the consortium with regard to all the datasets that will be generated by the project.

The DMP is not a fixed document. It will evolve during the lifespan of the project. This first version of the DMP includes an overview of the datasets to be produced by the project, and the specific conditions that are attached to them. The next version of the DMP, to be published at M18, will detail and describe the practical data management procedures implemented by the Sci-GaIA project. The data management plan will cover all the data life cycle (figure 1).

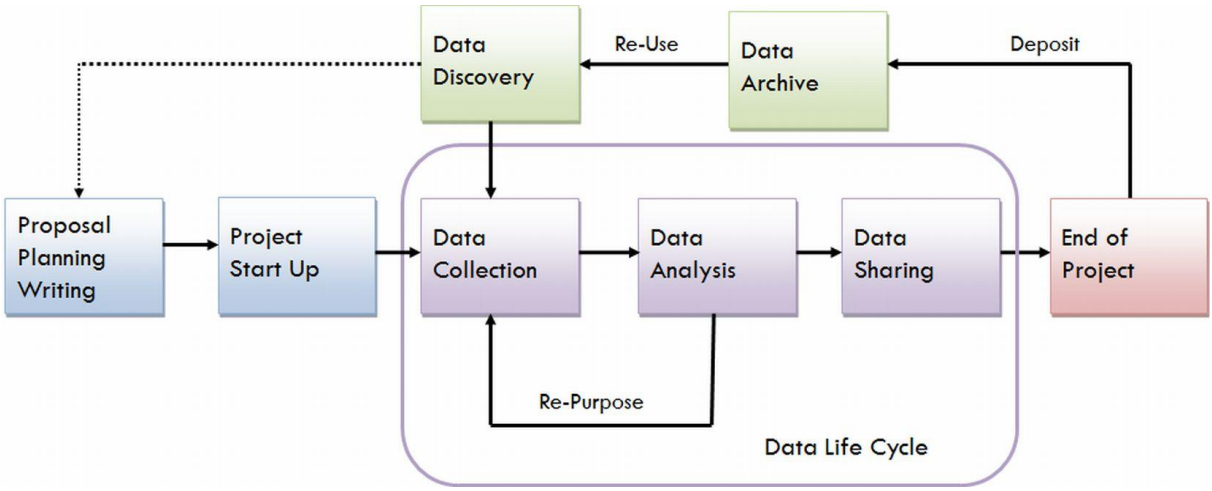


Figure 1: Steps in the data life cycle. Source: From University of Virginia Library, Research Data Services

2 DATASET LIST

All Sci-GaIA partners have identified the datasets that will be produced during the different phases of the project. The list is provided below, while the nature and details for each dataset are given in the subsequent sections.

This list is indicative and allows estimating the data that Sci-GaIA will produce – it may be adapted (addition/removal of datasets) in the next versions of the DMP to take into consideration the project developments.

#	Dataset (DS) name	Responsible partner	Related WP(s)
1	DS1_Newsletter-Subscribers_SIGMA_V01_DATE	SIGMA	WP4
2	DS2_e-Infrastructure-Survey_WACREN_V01_DATE	WACREN	WP1
3	DS3_User-Forum-Members_CSIR_V01_DATE	CSIR	WP2
4	DS4_Open-Access-Repositories&Services_UNICT_V01_DATE	UNICT	WP3
5	DS5_Event-Membership_SIGMA_V01_DATE	SIGMA	WP4
6	DS6_Educational-Materials_BRUNEL_V01_DATE	BRUNEL	WP1
7	DS7_Project-Deliverables_V01_DATE	BRUNEL	WP5

Table 1: Dataset list

3 GENERAL PRINCIPLES

3.1 PARTICIPATION IN THE PILOT ON OPEN RESEARCH DATA

The Sci-GaIA project participates in the Pilot on Open Research Data launched by the European Commission along with the Horizon 2020 programme. The consortium strongly believes in the concepts of open science, and in the benefits that the European innovation ecosystem and economy can draw from allowing reusing data at a larger scale. Therefore, all data produced by the project can potentially be published with open access – though this objective will obviously need to be balanced with the other principles described below.

3.2 IPR MANAGEMENT AND SECURITY

Project partners obviously have Intellectual Property Rights (IPR) on their technologies and data, on which their economic sustainability relies. As a legitimate result, the Sci-GaIA consortium will have to protect these data and consult the concerned partner(s) before publishing data.

Another effect of IPR management is that – with the data collected through Sci-GaIA being of high value – all measures should be taken to prevent them to leak or being hacked. This is another key aspect of Sci-GaIA data management. Hence, all data repositories used by the project will include a secure protection of sensitive data.

An holistic security approach will be undertaken to protect the 3 mains pillars of information security: confidentiality, integrity, and availability. The security approach will consist of a methodical assessment of security risks followed by an impact analysis. This analysis will be performed on the personal information and data processed by the proposed system, their flows and any risk associated to their processing.

Security measures will include the implementation of PAKE protocols – such as the SRP protocol – and protection against bots such as CAPTCHA technologies. Moreover, the WP/Task leaders identified in Table 1 will implement monitored and controlled procedures related to data collection, integrity and protection. Additionally, the protection and privacy of personal information will include protective measures against infiltration as well as physical protection of core parts of the systems and access control measures.

3.3 PERSONAL DATA PROTECTION

For some of the activities to be carried out by the project, it may be necessary to collect basic personal data (e.g. full name, contact details, background), even though the project will avoid collecting such data unless deemed necessary.

Such data will be protected in compliance with the EU's *Data Protection Directive 95/46/EC*[1] aiming at protecting personal data. National legislations applicable to the project will also be strictly followed, such as the <*Italian Personal Data Protection Code*[2]>.

[1] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0046&from=en

All data collected by the project will be done after giving data subjects full details on the analysis to be conducted, and after obtaining signed informed consent forms.

[2] http://www.privacy.it/privacycode-en.html

4 DATA MANAGEMENT PLAN

4.1 DATASET 1: NEWSLETTER SUBSCRIBERS

DS1_Newsletter-Subscribers_ SIGMA_V01_DATE	
Data Identification	
Dataset description	Mailing list containing email addresses and names of all subscribers to the Sci-GaIA's newsletter
Source (How have the data been collected? From which tool/survey does the data come from?)	This dataset is automatically generated in <Mailchimp/Mailjet/…> by visitors signing up to the newsletter form available on the project website. Additional subscribers can be manually added to the mailing list by the partner in charge of the project communication after receiving informed consent from the data subjects
Partners activities and responsibilities	
Partner owner of the data; copyright holder (if applicable).	SIGMA
Partner in charge of the data collection	SIGMA
Partner in charge of the data analysis	SIGMA
Partner in charge of the data storage	SIGMA
Related WP(s) and task(s)	WP4, T4.1
Standards	
Info about metadata (production and storage dates, places) and documentation?	N/A
Standards, format, estimated volume of data	This dataset can be imported from, and exported to a CSV, TXT or Excel file. Currently, at the time of this deliverable, the list is containing contact information of around 7000 people, and is smaller than 1 Mb
Data exploitation and sharing	
Data exploitation (purpose/use of the data analysis)	The mailing list will be used for disseminating the project newsletter to a targeted audience. An analysis of newsletter subscribers may be performed in order to assess and improve the overall visibility of the project

Data access policy / Dissemination level : confidential (only for members of the Consortium and the Commission Services) or Public	This dataset does not contain confidential information. However, the information is sensitive because it implies managing personal data. Therefore, access to the dataset is restricted to the project dissemination and communication leader
Data sharing, re-use, distribution, publication (How?)	None
Embargo periods (if any)	None
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	The mailing list contains personal data (names and email addresses of newsletter subscribers). People interested in the project voluntarily register, through the project website, to receive the project newsletter. They can unsubscribe at any time.
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The mailing list will be regularly backed up in Excel file format all along the project. Back-ups are safely stored in SIGMA's server.

4.2 DATASET 2: E-INFRASTRUCTURE SURVEY

DS2_e-Infrastructure-Survey_WACREN_V01_DATE	
Data Identification	
Dataset description	Dataset containing details of people who have participated in the Sci-GaIA e-Infrastructure Survey
Source (How have the data been collected? From which tool/survey does the data come from?)	This dataset is captured using Limesurvey as people take part in the survey. The link is http://surveys.sci-gaia.eu/index.php/531683
Partners activities and responsibilities	
Partner owner of the data; copyright holder (if applicable).	WACREN
Partner in charge of the data collection	WACREN
Partner in charge of the data analysis	WACREN

Partner in charge of the data storage	WACREN
Related WP(s) and task(s)	WP1, T1.3
Standards	
Info about metadata (production and storage dates, places) and documentation?	N/A
Standards, format, estimated volume of data	This dataset can be imported from, and exported to a CSV, TXT or Excel file. Currently, at the time of this deliverable, the list contains 0 people and their responses, and is smaller than 1 Mb
Data exploitation and sharing	
Data exploitation (purpose/use of the data analysis)	The data will be analysed to give indications of the impact of e-Infrastructures in Africa. This will appear in D1.3.
Data access policy / Dissemination level : confidential (only for members of the Consortium and the Commission Services) or Public	This dataset does not contain confidential information. However, the information is sensitive because it implies managing personal data. Therefore, access to the dataset is initially restricted to the task leader. However, if the participant has indicated that they are happy to have their personal details shared then this will be made available to the project team and within D1.3 (i.e. participants wish to have their e-Infrastructure project efforts shared with the international community).
Data sharing, re-use, distribution, publication (How?)	See above.
Embargo periods (if any)	None
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	The survey specifically asks if the participants are happy to share their details. If so, they indicate this in the survey document and add their details.
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The list will be stored within the Limesurvey tool or exported to Excel and stored in the WACREN beneficiary's computer. These will be held at WACREN. The list will be deleted six months after the end of the project. Participants who are happy to share their details will have their data stored within D3.1 (see project deliverables dataset).

4.3 DATASET 3: USER FORUM MEMBERS

DS3_User-Forum-Members_CSIR_V01_DATE	
Data Identification	
Dataset description	Forum list containing email addresses and names of all subscribers to the Sci-GaIA User Forum. Dataset also contains all Forum posts.
Source (How have the data been collected? From which tool/survey does the data come from?)	This dataset is automatically generated by visitors signing up to the User Forum at discourse.sci-gaia.eu.
Partners activities and responsibilities	
Partner owner of the data; copyright holder (if applicable).	CSIR
Partner in charge of the data collection	CSIR
Partner in charge of the data analysis	CSIR
Partner in charge of the data storage	CSIR
Related WP(s) and task(s)	WP2, T2.1
Standards	
Info about metadata (production and storage dates, places) and documentation?	N/A
Standards, format, estimated volume of data	This dataset can be imported from, and exported to a CSV, TXT or Excel file. Currently, at the time of this deliverable, the list is containing contact information of around 20 people, and is smaller than 1 Mb
Data exploitation and sharing	
Data exploitation (purpose/use of the data analysis)	The dataset is the Forum discussions that support the project's activities. This is "self-exploiting" in the sense of continued discussion. The Forum's themes and content will be analysed without reference to specific users in D2.1 Outcomes of the Web-based Forum.

Data access policy / Dissemination level : confidential (only for members of the Consortium and the Commission Services) or Public	This dataset does have some personal data. Users have control over the visibility of this and the degree to which this is shared with other Forum users. Access to personal data is otherwise restricted to the task leader. Posts in the Forum are visible to all users.
Data sharing, re-use, distribution, publication (How?)	As noted above.
Embargo periods (if any)	None
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	This dataset does have some personal data. Users have control over the visibility of this and the degree to which this is shared with other Forum users. People interested in the Forum voluntarily register and can deregister at any time.
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The forum is held on the discourse server in their own file format. The location of the server is being investigated.

4.4 DATASET 4: OPEN ACCESS REPOSITORIES & SERVICES

DS4_Open-Access-Repositories&Services_V01_DATE	
Data Identification	
Dataset description	This is the list of open access data repositories and services supported by Sci-GaIA's infrastructure services. Where appropriate it will list the data management policy for a particular service.
Source (How have the data been collected? From which tool/survey does the data come from?)	This is a simple list that is added to when new data repositories and services are added to our infrastructure services.
Partners activities and responsibilities	
Partner owner of the data; copyright holder (if applicable).	UNICT
Partner in charge of the data collection	UNICT
Partner in charge of the data analysis	UNICT

Partner in charge of the data storage	UNICT
Related WP(s) and task(s)	WP3, T3.2
Standards	
Info about metadata (production and storage dates, places) and documentation?	Under development
Standards, format, estimated volume of data	Under development
Data exploitation and sharing	
Data exploitation (purpose/use of the data analysis)	Various and will reflect the services. Each will be captured along with the service description.
Data access policy / Dissemination level : confidential (only for members of the Consortium and the Commission Services) or Public	Through IdP, i.e. only those will appropriate security credentials can access the service. This will be detailed against each service.
Data sharing, re-use, distribution, publication (How?)	This will vary from service to service and will be captured with a service description.
Embargo periods (if any)	None.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	The policy for each service will be captured.
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	As above.

4.5 DATASET 5: EVENT MEMBERSHIP

DS5_Event-Membership_ SIGMA_V01_DATE	
Data Identification	
Dataset description	A list of participants at the Sci-GaIA workshops and training events.
Source (How have the data been collected? From which tool/survey does the data come from?)	The dataset is generated from attendees joining the Sci-GaIA events.
Partners activities and responsibilities	
Partner owner of the data; copyright holder (if applicable).	SIGMA
Partner in charge of the data collection	SIGMA
Partner in charge of the data analysis	SIGMA
Partner in charge of the data storage	SIGMA
Related WP(s) and task(s)	WP4, T4.2 & T4.3
Standards	
Info about metadata (production and storage dates, places) and documentation?	N/A
Standards, format, estimated volume of data	This dataset can be imported from, and exported to a CSV, TXT or Excel file. Currently, at the time of this deliverable, the list is containing contact information of 0 people as the events are yet to take place.
Data exploitation and sharing	
Data exploitation (purpose/use of the data analysis)	An analysis of event attendees may be performed in order to assess and improve the overall visibility of the project

Data access policy / Dissemination level : confidential (only for members of the Consortium and the Commission Services) or Public	This dataset does not contain confidential information. However, the information is sensitive because it implies managing personal data. Therefore, access to the dataset is restricted to the project dissemination and communication leader.
Data sharing, re-use, distribution, publication (How?)	None
Embargo periods (if any)	None
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	The list contains personal data (names and email addresses of newsletter subscribers). People interested in the events voluntarily register, through the project website.
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The mailing list will be regularly backed up in Excel file format all along the project. Back-ups are safely stored in SIGMA's server.

4.6 DATASET 6: EDUCATIONAL MATERIALS

DS6_Educational-Materials_ BRUNEL_V01_DATE	
Data Identification	
Dataset description	Educational materials created for the training workshops.
Source (How have the data been collected? From which tool/survey does the data come from?)	This has been developed by UNICT and Brunel to support the training events and subsequent educational modules.
Partners activities and responsibilities	
Partner owner of the data; copyright holder (if applicable).	UNICT, BRUNEL
Partner in charge of the data collection	BRUNEL, UNICT
Partner in charge of the data analysis	NA

Partner in charge of the data storage	UNICT
Related WP(s) and task(s)	WP1, T1.1
Standards	
Info about metadata (production and storage dates, places) and documentation?	N/A
Standards, format, estimated volume of data	There are many types of data involved in this ranging from word documents to videos. The estimated size as deployed in OPENEDX will be determined at the time of the associated deliverable.
Data exploitation and sharing	
Data exploitation (purpose/use of the data analysis)	This will be published under an open commons licence for anyone to exploit.
Data access policy / Dissemination level : confidential (only for members of the Consortium and the Commission Services) or Public	Open access according to the open commons licence.
Data sharing, re-use, distribution, publication (How?)	This will be available to all.
Embargo periods (if any)	None
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data.
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The data will be held and backed up at UNICT servers.

4.7 DATASET 7: PROJECT DELIVERABLES

DS7_Project-Deliverables_ BRUNEL_V01_DATE	
Data Identification	
Dataset description	The deliverables of the project.
Source (How have the data been collected? From which tool/survey does the data come from?)	Generated by WP leaders.
Partners activities and responsibilities	
Partner owner of the data; copyright holder (if applicable).	BRUNEL
Partner in charge of the data collection	BRUNEL (and WP leaders)
Partner in charge of the data analysis	BRUNEL (and WP leaders)
Partner in charge of the data storage	SIGMA/EC
Related WP(s) and task(s)	WP5 and all WPs
Standards	
Info about metadata (production and storage dates, places) and documentation?	N/A
Standards, format, estimated volume of data	This will be determined by the end of the document. It will be a combination of WORD/PDF documents and supporting information.
Data exploitation and sharing	
Data exploitation (purpose/use of the data analysis)	The deliverables present the outcomes of the project for public use.

Data access policy / Dissemination level : confidential (only for members of the Consortium and the Commission Services) or Public	Open access for all deliverables apart from financial information. This is restricted to the consortium and Commission Services.
Data sharing, re-use, distribution, publication (How?)	Open expect noted above.
Embargo periods (if any)	None
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Any personal data will be handled according to the datasets appear in any deliverable as noted above.
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	SIGMA and EC – indefinitely.

5 CONCLUSION

This document contains the data management policy for Sci-GaIA. The policy will be periodically revised at Project Management Board meetings.