GUIDELINES TO SETUP AND CONFIGURE AN APPLIANCE FOR THE DEPLOYMENT OF STANDARD COMPLIANT OPEN ACCESS REPOSITORIES
August 2015
## PROJECT DOCUMENTATION SHEET

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

<table>
<thead>
<tr>
<th>Number</th>
<th>Deliverable D3.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Guidelines to setup and configure an appliance for the deployment of standard compliant Open Access Repositories</td>
</tr>
<tr>
<td>Related WP</td>
<td>WP3 (Strengthen and expand Science Gateway and e-Infrastructure related services)</td>
</tr>
<tr>
<td>Related Task</td>
<td>Task 3.1 (Support the creation of federated and interoperable Open Access Document and Data Repositories in Africa, compliant with EU and other international guidelines)</td>
</tr>
<tr>
<td>Lead Beneficiary</td>
<td>UNICT</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Roberto Barbera (UNICT) – <a href="mailto:roberto.barbera@ct.infn.it">roberto.barbera@ct.infn.it</a>, Rita Ricceri (UNICT) – <a href="mailto:rita.ricceri@ct.infn.it">rita.ricceri@ct.infn.it</a>, Mario Torrisi (UNICT) – <a href="mailto:mario.torrisi@ct.infn.it">mario.torrisi@ct.infn.it</a></td>
</tr>
<tr>
<td>Contributor(s)</td>
<td>Bruce Becker (CSIR) – <a href="mailto:bbecker@csir.co.za">bbecker@csir.co.za</a></td>
</tr>
<tr>
<td>Reviewer(s)</td>
<td></td>
</tr>
<tr>
<td>Nature</td>
<td>R (Report)</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>PU (Public)</td>
</tr>
<tr>
<td>Due Date</td>
<td>August 31, 2015 (M4)</td>
</tr>
<tr>
<td>Submission date</td>
<td>August 31, 2015 (M4)</td>
</tr>
<tr>
<td>Status</td>
<td>Final</td>
</tr>
</tbody>
</table>
## QUALITY CONTROL ASSESSMENT SHEET

<table>
<thead>
<tr>
<th>Issue</th>
<th>Date</th>
<th>Comment</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>V0.1</td>
<td>16/08/2015</td>
<td>First public version for internal review</td>
<td>Roberto Barbera (UNICT) Technical Coordinator &amp; WP Leader</td>
</tr>
<tr>
<td>V0.2</td>
<td>21/08/2015</td>
<td>Second version taking into account comments and improvements made by Bruce Becker</td>
<td>Roberto Barbera, Rita Ricceri, Mario Torrisi (UNICT)</td>
</tr>
<tr>
<td>FINAL</td>
<td>31/08/2015</td>
<td>Approved</td>
<td>Simon J E Taylor (Brunel)</td>
</tr>
</tbody>
</table>
DISCLAIMER

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

All intellectual property rights are owned by the Sci-GaIA consortium members and are protected by the applicable laws. Except where otherwise specified, all document contents are: “©Sci-GaIA Project - All rights reserved”. Reproduction is not authorised without prior written agreement.

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

Work-Package 3 (WP3) “Strengthen and expand Science Gateway and e-Infrastructure related services”, as described in the Sci-GaIA Description of Work (DoW), aims at:

1. Expanding and extending activities carried out in past projects in order to consolidate the African e-Infrastructure services and to include the very challenging goal of supporting the creation of an African Open (and Linked) Data Infrastructure, interoperable with and federated to (through the adoption of international standards and guidelines) those emerging in EU and in other regions of the world.

2. Combining Open Access repositories with Science Gateways in order to deal with very important topics such as the discoverability, reproducibility and extensibility of science products.

WP3 builds on the successes achieved by Task 6.1 “Promote, foster & support the creation of e-infrastructure related services” of the past eI4Africa project. In the context of Sci-GaIA, WP3 is expected to expand and extend past activities in order to consolidate the results already achieved and to include the challenging goal of supporting the creation of an African Open (Linked) Data Infrastructure, interoperable with and federated to those emerging in EU and in other regions of the world, through the adoption of international standards and guidelines.

Combining Open Access repositories with Science Gateways interfaced with the underlying e-Infrastructure computing and storage services, WP3 also deals with very important topics related to Open Science, such as the discoverability, reproducibility and re-usability of science products.

This deliverable presents the template of a standards-compliant Open Access repository, delivered by the project as a fully-functional, reproducible, proof of principle – a contextualised template of good practice which communities can refer to or even re-use directly. We also report on the current state of activities. Guidelines to install and configure a virtual appliance containing the template repositories are also mentioned.
TABLE OF CONTENT

Table of content ........................................................................................................................................... 6
  1 – Introduction ............................................................................................................................................... 7
    1.1 – Introductory Concepts and Driving Considerations .................................................................................. 7
  2 – Requirements, Choices and Motivations ..................................................................................................... 13
  3 – The Reference Model: the INFN Open Access Repository ........................................................................ 16
    3.1 – Approach and Generalities .................................................................................................................... 16
    3.2 – Types of Resources ............................................................................................................................... 18
    3.3 – Certification and Compliance .............................................................................................................. 20
    3.4 – Knowledge Workflow ........................................................................................................................... 22
    3.5 – Science Gateways and Open Access Repositories ................................................................................ 27
  4 – The Sci-GaIA Open Access Repository .................................................................................................... 30
    4.1 – Current implementation .......................................................................................................................... 30
    4.2 – Installation and Configuration Instructions .......................................................................................... 31
  5 – Summary and conclusions ......................................................................................................................... 33
1 – INTRODUCTION

1.1 – INTRODUCTORY CONCEPTS AND DRIVING CONSIDERATIONS

In the last 30 years or so, scientific computing has steadily evolved from centralized to a more distributed environments. This has been due to the concurrent availability of cost-effective “Commercial Of The Shelf” (COTS) components and decrease of costs of Local Area Networks. In the first half of 90’s, the emergence of cluster computing for High Throughput Computing (HTC) applications was confirmed and “farms” of computers with many-core processors, interconnected by low-latency networks, become the norm. This eventually extended to the domain of High Performance Computing (HPC) to the extent that about 80% of the Top500 machines built in the last 5 years are based on a cluster architecture.

Furthermore, the steep decrease of costs of high-bandwidth Wide Area Networks has fostered in the recent years the spread and the uptake of the Grid Computing paradigm and the distributed computing ecosystem has become even more complex with the recent emergence of Cloud Computing.

At the onset of the 21st century all these developments have led to the new concept of e-Infrastructure - defined as:

“an environment where research resources (hardware, software and content) can be readily shared and accessed where necessary to promote better and more effective research; such environment integrate hard-, soft- and middle-ware components, networks, data repositories, and all sorts of support enabling virtual research collaborations to flourish globally”.

Indeed, e-Infrastructures have been built over several years both in Europe and the rest of the world, to support diverse multi- and inter-disciplinary Virtual Research Communities (VRCs). There is a shared vision for 2020 that e-Infrastructures will allow scientists across the world to do better (and faster) research, irrespective of where they are and of the paradigm(s) adopted to build them.

E-Infrastructure components can be key platforms to support the Scientific Method, the “knowledge path” followed in many aspects by scientists since the time of Galileo Galilei. With reference to Figure 1, Distributed Computing and Storage Infrastructures (local HPC/HTC resources, Grids, Clouds, long term data preservation services) are ideal both for the creation of new datasets and the analysis of existing ones while Data Infrastructures (including Open Access Document Repositories – OADRs – and Data Repositories – DRs) are essential to evaluate existing data and annotate them with results of the

1 Go to http://top500.org/statistics/overtime/, select Category = Architecture, choose Type = Systems Share, and then click on Submit to generate the graph.


4 There are many equivalent definitions and depictions of the Scientific Method, both on the web and on textbooks. In this document we refer to http://home.badc.rl.ac.uk/lawrence/blog/2009/04/16/scientific_method, from which we have re-used the picture included in Figure 1.
analysis of new data produced by experiments and/or simulations. Last but not least, Semantic Web-based enrichment of data is key to correlate documents and data, allowing scientists to discover new knowledge in an easy way, and engage in a more robust scholarly discourse.

One of the cornerstones of the Scientific Method, which is a key driver through the knowledge path, is science reproducibility. In recent years, the issue of the reproducibility of scientific results has attracted increasing attention worldwide, both inside and outside scholarly communities, to which a recent Special Edition of Nature is testament. As striking examples, Begley and Ellis could not reproduce the results of 47 out of 53 “landmark” publications in cancer research and Casadevall et al. have identified more than 2,000 articles listed in Pubmed as retracted since the first identified article was retracted in 1977.

---

5 www.nature.com/nature/focus/reproducibility/.
8 www.pubmed.org.
The problem goes well beyond the topic of cancer. In March 2012 a committee of the US National Academy of Sciences heard testimony that the number of scientific papers that had to be retracted increased more than tenfold over the last decade while the number of journal articles published rose only 44 percent over the same period. At the current rate, by 2045 there will be as many papers published as retracted.

In light of these findings, researchers and other scholarly observers have recently been proposing and conducting initiatives to help the scientific community to address the issue of reproducibility. Some of the most interesting ones are gathered under the umbrella of the Reproducibility Initiative, jointly started by the lab-services start-up Science Exchange and the open access journal PLoS ONE. Scientists can submit studies to Science Exchange that they would like to see replicated. An independent scientific advisory board selects studies for replication and service providers are then selected at random to conduct the experiments. The results are returned to the original investigators, who can then publish them in a special issue of the open-access journal PLoS ONE and are awarded with a "certificate of reproducibility" for studies that are successfully replicated.

Although the initiative of Science Exchange is commendable, it is however limited to the health domain, authors have to pay to have their results reproduced, and the choice of studies to be reproduced is entirely decided by the advisory board.

Furthermore, some very important considerations are in order.

1. As pointed out by C. Drummond, reproducibility and replicability are different concepts and "replicability is not reproducibility".
2. The “re-‘s” of the Scientific Method go beyond re-PLICability and re-productibility and indeed include both re-peatability and re-usability (see Figure 2).
3. In the last 2-3 decades science has become more and more computationally intensive and computer simulations are actually “reconciling” the inductive and deductive approaches of the Scientific Method. Then:
   a. “An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment, [the complete data] and the complete set of instructions which generated the figures”.
   b. “Scientific communication relies on evidence that cannot be entirely included in publications, but the rise of computational science has added a new layer of inaccessibility. Although it is now accepted that data should be made available on request, the current regulations regarding the availability of software are inconsistent.

10 [http://validation.scienceexchange.com](http://validation.scienceexchange.com).
11 [https://www.scienceexchange.com](https://www.scienceexchange.com).
We argue that, with some exceptions, anything less than the release of source programs is intolerable for results that depend on computation. The vagaries of hardware, software and natural language will always ensure that exact reproducibility remains uncertain, but withholding code increases the chances that efforts to reproduce results will fail.\textsuperscript{15}

c. “The publication and open exchange of knowledge and material form the backbone of scientific progress and reproducibility and are obligatory for publicly funded research. Despite increasing reliance on computing in every domain of scientific endeavor, the computer source code critical to understanding and evaluating computer programs is commonly withheld, effectively rendering these programs “black boxes” in the research workflow.”\textsuperscript{16}

---


According to a recently published seminal book\textsuperscript{17}, Open Science “refers to a scientific culture that is characterized by its openness. Scientists share results almost immediately and with a very wide audience”.

Five schools of thought on Open Science have been identified so far\textsuperscript{18}, characterised by their central assumptions, the involved stakeholder groups, their aims, and the tools and methods used to achieve and promote these aims (see Figure 4). The infrastructure school is concerned with the technical infrastructure that enables emerging research practices on the Internet, for the most part software tools and applications, as well as computing networks. The infrastructure school regards Open Science as a technological challenge and focuses on the technological requirements that facilitate particular research practices, such as Grid and, more recently, Cloud Computing.


The Sci-GaIA very much supports the Open Science “paradigm” and has a strong focus on the application of the guidelines of the infrastructure school. The whole Consortium is committed to setup a platform for re-producible and re-usable science across Europe and Africa and one the deemed key components is a template for an Open Access Repository (OAR) that could be easily deployable wherever it is required in the region targeted by the project.

This deliverable introduces the requirements setup upfront, the choices made and their motivations (see Section 2) and presents the Sci-GaIA Open Access Repository (see Section 4) which has the double function of being the project’s document repository (in order to comply with the Open Data Pilot launched by the EC) and the template to be cloned in many places in Africa to make science “made in Africa” more visible, re-producible and re-usable.

As anticipated in the Description of Work, the Sci-GaIA Open Access Repository is based on the INFN Open Access Repository whose implementation details and functionalities are discussed in Section 3.
2 – Requirements, Choices and Motivations

Open Access repositories are powered by Digital Asset Management Systems (DAMS), which are intertwined structures incorporating both software and hardware that take care of management tasks and decisions surrounding the ingestion, annotation, cataloguing, storage, retrieval and distribution of digital assets. Types of digital assets include, but are not exclusive to, photography, logos, illustrations, animations, audio-visual media, presentations, spreadsheets, Word and/or PDF documents, and a multitude of other digital formats and their respective metadata.

There is a plethora of DAMS’s available and some of the most common used in the Open Access domain are listed in Table 1. Others, more business- and/or social-oriented, are listed in ref. 20.

<table>
<thead>
<tr>
<th>DAMS</th>
<th>Home page</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKAN</td>
<td><a href="http://ckan.org/">http://ckan.org/</a></td>
<td>Free</td>
</tr>
<tr>
<td>CONTENTdm</td>
<td><a href="http://www.oclc.org/contentdm.en.html">http://www.oclc.org/contentdm.en.html</a></td>
<td>Commercial</td>
</tr>
<tr>
<td>Digibib</td>
<td><a href="https://digibib.com">https://digibib.com</a></td>
<td></td>
</tr>
<tr>
<td>Digital Commons</td>
<td><a href="http://digitalcommons.bepress.com/">http://digitalcommons.bepress.com/</a></td>
<td>Commercial (hosted service)</td>
</tr>
<tr>
<td>DigiTool</td>
<td><a href="http://www.exlibrisgroup.com/category/DigiToolOverview">http://www.exlibrisgroup.com/category/DigiToolOverview</a></td>
<td>Commercial</td>
</tr>
<tr>
<td>DiVA-Portal</td>
<td><a href="http://www.diva-portal.org">http://www.diva-portal.org</a></td>
<td>Free (hosted service)</td>
</tr>
<tr>
<td>dLibra</td>
<td><a href="http://dingo.pscn.pl/dlibra/">http://dingo.pscn.pl/dlibra/</a></td>
<td>Commercial</td>
</tr>
<tr>
<td>Drupal</td>
<td><a href="https://www.drupal.org/">https://www.drupal.org/</a></td>
<td>Free</td>
</tr>
<tr>
<td>DSpace</td>
<td><a href="http://www.dspace.org/">http://www.dspace.org/</a></td>
<td>Free</td>
</tr>
<tr>
<td>Earmas</td>
<td><a href="http://www.earmas.net/">http://www.earmas.net/</a></td>
<td>Free</td>
</tr>
<tr>
<td>EPrints</td>
<td><a href="http://www.eprints.org/software/">http://www.eprints.org/software/</a></td>
<td>Free</td>
</tr>
<tr>
<td>EQUELLA Repository</td>
<td><a href="http://www.equella.com/">http://www.equella.com/</a></td>
<td>Commercial</td>
</tr>
<tr>
<td>ETD-db</td>
<td><a href="http://scholar.lib.vt.edu/ETD-db/index.shtml">http://scholar.lib.vt.edu/ETD-db/index.shtml</a></td>
<td>Free</td>
</tr>
<tr>
<td>Fedora</td>
<td><a href="http://www.fedora-commons.org/">http://www.fedora-commons.org/</a></td>
<td>Free</td>
</tr>
<tr>
<td>Fez</td>
<td><a href="http://apsr.anu.edu.au/currentprojects/fez06.htm">http://apsr.anu.edu.au/currentprojects/fez06.htm</a></td>
<td>Free</td>
</tr>
<tr>
<td>Greenstone</td>
<td><a href="http://www.greenstone.org/">http://www.greenstone.org/</a></td>
<td>Free</td>
</tr>
<tr>
<td>HAL</td>
<td><a href="https://hal.archives-ouvertes.fr/">https://hal.archives-ouvertes.fr/</a></td>
<td>Free (hosted service)</td>
</tr>
<tr>
<td>Invenio</td>
<td><a href="http://invenio-software.org/">http://invenio-software.org/</a></td>
<td>Free</td>
</tr>
<tr>
<td>Islandora/Fedora</td>
<td><a href="http://islandora.ca/">http://islandora.ca/</a></td>
<td>Free</td>
</tr>
<tr>
<td>intraLibrary</td>
<td><a href="http://www.intrallect.com/solutions/managing_content/">http://www.intrallect.com/solutions/managing_content/</a></td>
<td>Free</td>
</tr>
<tr>
<td>MyCoRe</td>
<td><a href="http://www.mycore.de/">http://www.mycore.de/</a></td>
<td>Free</td>
</tr>
<tr>
<td>Open Repository</td>
<td><a href="http://www.openrepository.com/">http://www.openrepository.com/</a></td>
<td>Commercial (hosted service)</td>
</tr>
<tr>
<td>OPUS</td>
<td><a href="http://www.kobv.de/entwicklung/software/">http://www.kobv.de/entwicklung/software/</a></td>
<td>Free</td>
</tr>
</tbody>
</table>

TABLE 1 – LIST OF THE MOST USED OPEN ACCESS DIGITAL ASSET MANAGEMENT SYSTEMS

<table>
<thead>
<tr>
<th>System</th>
<th>URL</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>PURE</td>
<td><a href="https://www.st-andrews.ac.uk/staff/research/pure/">https://www.st-andrews.ac.uk/staff/research/pure/</a></td>
<td>Free (hosted service)</td>
</tr>
<tr>
<td>SciELO</td>
<td><a href="http://scielo.org/php/index.php">http://scielo.org/php/index.php</a></td>
<td>Free (hosted service)</td>
</tr>
<tr>
<td>VITAL</td>
<td><a href="https://www.iii.com/products/vital">https://www.iii.com/products/vital</a></td>
<td>Commercial</td>
</tr>
<tr>
<td>WEKO</td>
<td><a href="http://weko.wou.edu.my">http://weko.wou.edu.my</a></td>
<td>Free</td>
</tr>
<tr>
<td>XooNips</td>
<td><a href="http://xoops.org/modules/repository/">http://xoops.org/modules/repository/</a></td>
<td>Free</td>
</tr>
</tbody>
</table>

Our requirements for a DAMS to be used in and promoted by the Sci-GaIA project were the following:

- Open source;
- Distributed under a free license;
- Deployable on a local infrastructure (i.e., not a hosted service);
- Standard compliant;
- Well supported;
- Scalable, up to $O(10^6) - O(10^7)$ resources.

There are many comparisons of DAMS’s available on the web and we went through several\(^{21,22}\) of them but an important element for the choice was to have the most direct possible know-how of the DAMS to be adopted, in order to have something solid but, at the same time, deployable in a very short amount of time (the present deliverable was planned at month 4 of the project and the deliverable about the Data Management Plan was planned at month 6).

In this respect, a key element was the fact that a group of technologists at INFN Catania, led by the representative of UNICT in Sci-GaIA, is developing a pilot for an INFN Open Access Repository since almost two years.

For this reason, we have choosen Invenio\(^{23}\) and the main motivations for the choice were the following:

1. It is fully compliant with all most important library standards, such as, for example: DCMI\(^{24}\), Marc21\(^{25}\) and OAI-PMH\(^{26}\);
2. It is co-developed by an international collaboration comprising institutes such as CERN\(^{27}\), DESY\(^{28}\), EPFL\(^{29}\), FNAL\(^{30}\), SLAC\(^{31}\) and used as institutional DAMS by about 30 scientific institutions worldwide\(^{32}\);

\(^{21}\) http://www.rsp.ac.uk/start/software-survey/results-2010/.
\(^{24}\) http://dublincore.org/.
\(^{25}\) http://www.loc.gov/marc/bibliographic/.
\(^{26}\) https://www.openarchives.org/pmh/.
\(^{27}\) http://www.cern.ch/.
\(^{28}\) http://www.desy.de/.
\(^{29}\) http://www.epfl.ch/.
\(^{30}\) http://www.fnal.gov/.
3. INSPIRE\textsuperscript{33}, SCOAP\textsuperscript{3}(\textsuperscript{34}) and ZENODO\textsuperscript{35} (the OpenAIRE\textsuperscript{36} flagship archive) repositories are based on Invenio;

4. The CERN Document Server\textsuperscript{37} operates since 2002 and manages more than 1.3 million records in high-energy physics, covering articles, books, journals, photos, videos, and more;

5. UNESCO\textsuperscript{38} and UEMOA\textsuperscript{39} are leading an initiative\textsuperscript{40} to create a virtual library based on Invenio in 8 African countries (Benin, Burkina Faso, Côte d'Ivoire, Guinea Bissau, Mali, Niger, Senegal and Togo).

The INFN Open Access Repository, the reference model adopted by Sci-GaIA, enhances Invenio with several add-ons developed by INFN and, for the sake of completeness, it is presented in the next section.

\textsuperscript{31} http://www.slac.stanford.edu/.
\textsuperscript{32} http://invenio-software.org/wiki/General/Demo.
\textsuperscript{33} https://inspirehep.net/.
\textsuperscript{34} http://scoap3.org/.
\textsuperscript{35} https://zenodo.org/.
\textsuperscript{36} http://www.openaire.eu/.
\textsuperscript{37} http://cds.cern.ch/.
\textsuperscript{38} http://www.unesco.org/.
\textsuperscript{39} http://www.uemoa.int/.
\textsuperscript{40} http://www.unesco.org/new/en/dakar/about-this-office/single-view/news/a_virtual_library_for_8_uemoa_countries_under_discussion_in_dakar/#.Vc7c1PRRQ5z.
3 – THE REFERENCE MODEL: THE INFN OPEN ACCESS REPOSITORY

3.1 – APPROACH AND GENERALITIES

The INFN Open Access Repository\(^\text{41}\) (INFN-OAR) is the key enabler of a vision where a DAMS can allow researchers to deposit their science products and, at the same time, be part of an Open Science platform ensuring science reproducibility and re-usability (see Figure 5).

The INFN-OAR allow single researchers to upload their products but also connects to external archives to harvest products having at least one INFN author. External archives can contain papers, such as arXiv\(^\text{42}\), data, such as EUDAT\(^\text{43}\), and software, such as GitHub\(^\text{44}\).

Furthermore, INFN-OAR allows each resource to be citable and discoverable, through unique identifiers, and reproducible/re-usable, thanks to the connection to a Science Gateway and to Grid, Cloud and local High Performance Computing (HPC) infrastructures.

The home page of the INFN Open Access Repository is shown in Figure 6 below.

\(^{41}\)http://www.openaccessrepository.it/.

\(^{42}\)http://www.arxiv.org/.

\(^{43}\)http://www.eudat.eu/.

\(^{44}\)http://www.github.com/.
FIGURE 6 – HOME PAGE OF THE INFN OPEN ACCESS REPOSITORY

It is worth noting that INFN-OAR allows federated login and it has been registered as a Service Provider of several Identity Federations including the GrIDP\(^{45}\) “catch-all” one and the eduGAIN\(^{46}\) inter-federation. Moreover, the service includes REST APIs to automatically harvest documents and data from arXiv, SCOAP\(^{3}\) and EUDAT, also as a “cron job”.

Once logged in, users can add their ID’s and INFN-OAR supports several credential providers, such as Google Scholar\(^{47}\), GitHub, ORCID\(^{48}\), ResearcherID\(^{49}\), ResearchGate\(^{50}\), and SCOPUS\(^{51}\) (see Figure 7). In the particular case of ORCID, which is becoming a “de facto” standard for author disambiguation, when available ORCID’s can be shown as links next to the authors’ names in the resources stored on INFN-OAR. (see Figure 8).

\(^{45}\) [http://gridp.garr.it/](http://gridp.garr.it/).
\(^{47}\) [https://scholar.google.com/](https://scholar.google.com/).
\(^{48}\) [http://www.orcid.org/](http://www.orcid.org/).
\(^{50}\) [https://www.researchgate.net/](https://www.researchgate.net/).
3.2 — Types of Resources

INFN-OAR can store several types of digital assets, from papers (see Figure 9), to datasets (internal or linked to external sources, see Figure 10), to software (Figure 10), including entire virtual machines containing all the code to reproduce and extend the analysis performed in a given paper with a given dataset.
D3.1 – Guidelines to setup and configure an appliance for the deployment of standard compliant open access repositories

August 2015

Author: UNICT

FIGURE 9 – EXAMPLE OF A DOCUMENT RESOURCE

FIGURE 10 – EXAMPLE OF A DATA RESOURCE
3.3 — Certification and Compliance

As already mentioned in Section 2, Invenio is fully standard-based and a lot of effort has been devoted to make INFN-OAR officially certified and compliant with the most relevant and widely known Open Access initiatives.

As shown in Figures 12 and 13, INFN-OAR is an Open Archives Initiative\(^52\) (OAI) conforming repository and it is an official OpenDOAR\(^53\) data provider.

\(^{52}\) [https://www.openarchives.org/](https://www.openarchives.org/)

It is also worth underlying that INFN-OAR is one of the officially certified OpenAIRE archives, compliant with version 3.0 of its guidelines\(^\text{54}\) (see Figures 14 and 15).

\(^{54}\) https://guidelines.openaire.eu/.
When a resource is uploaded to/harvested in INFN-OAR, the project under which it was produced can be specified. INFN-OAR includes the databases of FP4, FP5, FP6, FP7 and H2020 project as well as those of the National Operating Program of the Italian Ministry of Education, University and Research and of the INFN internal projects.

3.4 – KNOWLEDGE WORKFLOW

In the previous sub-sections we have shown how, thanks to the adoption of important bibliographic standards, INFN-OAR is both an Open Access Initiative conforming and an official OpenDOAR data provider, able to automatically harvest resources from different sources, including the Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP³), using RESTful API’s. It is also one of the official OpenAIRE archives, compliant with version 3.0 of its guidelines.

INFN-OAR allows SAML-based federated authentication and it is one of the Service Providers of the eduGAIN inter-federation. Furthermore, it is also connected to DataCite⁵⁵ for the issuance and registration of Digital Object Identifiers (DOIs).

But what makes INFN-OAR really different from other repositories is its capability to connect to Science Gateways and exploit Distributed Computing and Storage Infrastructures worldwide, including EGI⁵⁶ and EUDAT ones, to easily reproduce and re-use/extend scientific analyses.

Indeed, INFN-OAR is a key enabler of the “share, validate, preserve, reproduce” workflow depicted in Figure 16 and allows to “walk through the knowledge path in a circular way”.

55 http://www.datacite.org/
56 http://www.egi.eu/
Starting from the upper-left panel of Figure 16 and going clockwise, either a researcher or a citizen scientist can:

1. Search and discover a research product, e.g. a scientific publication or an analysis object;
2. Be re-directed to an Open Access Repository where that product is stored as one of its resources;
3. Be-redirected to Science Gateway where that analysis can be reproduced or even extended;
4. Write a new paper about the new extended analysis;
5. Upload the new paper on the Open Access Repository, assign a DOI to it, and “connect” the new paper to the old one, to the needed dataset(s) and to virtual appliance containing the software to read and analyse it(Them).

Upon completion of the virtuous cycle, new knowledge has been added to the existing one and both the new and the existing one are citable, searchable and discoverable. Step 1, 2, and 3 of the above workflow are described in the following.
Step 1 of the workflow depicted above is shown in Figures 17 and 18. A user is using the CHAIN-REDS Semantic Search Engine\(^{57}\) to look for the analysis of some experiment (in the particular case, the ALEPH\(^{58}\) experiment).

\[\text{FIGURE 17 – USING A SEARCH ENGINE TO LOOK FOR A RESEARCH PRODUCT}\]

\[\text{FIGURE 18 – INFORMATIVE PANEL RELATIVE TO ONE OF RESEARCH PRODUCTS FOUND}\]

\(^{57}\)\url{http://www.chain-project.eu/semantic-search},
\(^{58}\)\url{http://aleph.web.cern.ch/aleph/aleph/Public.html}.\]
In the informative panel relative to one of research products found (see Figure 18), both the location of the product in the INFN-OAR and that of the dataset used for that particular analysis (located in one of EUDAT storage elements) are reported.

Step 2 of the workflow depicted above is shown in Figure 19. Clicking on the link to the research product, the user is re-directed to INFN-OAR where he/she can explore the “package” containing all the elements associated to the research product of interest. In the particular case shown in Figure 19, these are the paper describing the analysis, the virtual machine containing all the software needed to reproduce the analysis and the analysis object itself.

Inspecting the detailed description of the analysis object (see Figure 20), contained in INFN-OAR like any other types of resources, the user can find the link “RUN PAGE”, which re-directs him/her to a Science Gateway where that particular analysis can be reproduced. It is worth noting that, if the user is authorised to access the Science Gateway, a Single Sign On is established between INFN-OAR and the Science Gateway. This enables the third step in the workflow described above.
Following the link “RUN PAGE”, the user lands on the pages of the Science Gateway (in the particular case shown in the figure, the CHAIN-REDS Science Gateway) where the dataset corresponding to the particular analysis of interest can be searched (see Figure 21) and inspected in detail (see Figure 22), with a view of its metadata that are automatically retrieved from INFN-OAR using the Invenio RESTful API.

![Science Gateway](http://science-gateway.chain-project.eu/).
In the detailed information page shown in Figure 22, the user is also presented with all the analyses available for the particular dataset of interest and he/she can choose which one he/she wants to reproduce.

When the user clicks on the “Analyse” button, using a special JSAGA\(^6\) adaptor for OCCI\(^6\)-compliant cloud middleware, the Science Gateway starts a dedicated virtual machine already configured with all the experiment software and runs the analysis.

Both the CHAIN-REDs Cloud Testbed\(^6\) and the EGI Federated Cloud\(^6\) can be used as e-Infrastructures where the ALEPH virtual machines can be executed.

### 3.5 — SCIENCE GATEWAYS AND OPEN ACCESS REPOSITORIES

In the previous sub-section we have described the knowledge workflow and all its steps. The “connection” between INFN-OAR and the CHAIN-REDs Science Gateway is bi-directional so the user may start the (re-)analysis of some dataset(s) from the latter and not from the former.

This is shown in Figures 23, 24 and 25: a user can visit the CHAIN-REDs Science Gateway, select the ALEPH Analysis as application (see Figures 23 and 24), list all available datasets (inspecting their metadata), choose one of the available analyses and execute it (see Figure 25).

---


\(^6\) [http://www.chain-project.eu/cloud-testbed/](http://www.chain-project.eu/cloud-testbed/).

\(^6\) [https://www.egi.eu/infrastructure/cloud/](https://www.egi.eu/infrastructure/cloud/).
ALEPH Analysis

ALEPH was a particle physics experiment installed on the Large Electron-Positron collider (LEP) at the CERN laboratory in Geneva/ Switzerland. It was designed to explore the physics predicted by the Standard Model and to search for physics beyond it. ALEPH first measured events in LEP in July 1989. LEP operated at around 91 GeV – the predicted optimum energy for the formation of the Z particle. From 1995 to 2000 the accelerator operated at energies up to 200 GeV, above the threshold for producing pairs of W particles. The data, taken consisted of millions of events recorded by the ALEPH detector, allowed precision tests of the electro-weak Standard Model (SM) to be undertaken. The group here concentrated on our analysis efforts mainly in Heavy Flavour (beauty and charm) physics, in searches for the the Higgs boson, the particles postulated to generate particle mass, and for physics beyond the SM, e.g. Supersymmetry, and in W physics.

This application performs the search for the production and non-standard decay of a scalar Higgs boson into four tau leptons through the intermediation of the neutral pseudo-scalars Higgs particle.

The analysis was conducted by the ALEPH collaboration with the data collected at centre-of-mass energies from 100 to 209 GeV.

Results are published in JHEP 1005 (2010) 045 DOI: 10.1007/JHEP05(2010)049

FIGURE 24 – RUN PAGE OF THE ALEPH ANALYSIS APPLICATION

FIGURE 25 – LIST OF THE AVAILABLE DATASETS AND CORRESPONDING ANALYSES
Reproducibility of science products is very important but it is not the only important feature of the Open Science paradigm. As discussed in Section 1, several other “re-‘s” matter and the possibility to re-use/extend existing analyses is key.

For this reason, the CHAIN-REDS Science Gateway offers to experienced users also the possibility to start a personal, persistent, virtual machine containing all the software and the dataset(s) related to a given analysis. This is shown in Figure 26.

![Science Gateway Page to Manage and Access the VM to Extend Analyses](image)

It is very important to note here the new, stable analyses (and their results), generated running the VM, may be registered by the user in the INFN-OAR (with a DOI) to further extend the analysis catalogue shared across the whole Virtual Research Community he/she belongs to.
4 – THE SCI-GAIA OPEN ACCESS REPOSITORY

4.1 – CURRENT IMPLEMENTATION

The current version of the Sci-GaIA Open Access Repository (Sci-GaIA-OAR) is shown in Figure 27. It is available in two languages, English and French, and, being based on the INFN-OAR, it inherits all its functionalities. It is already conforming with OAI specifications (see Figure 28) and the registration as a data provider of both OpenDOAR and OpenAIRE is in progress, as it is the registration as a Service Provider of the GrIDP “catch-all” federation and of the eduGAIN inter-federation.

[Figure 27 – Home page of the Sci-GaIA open access repository]

---

64 http://oar.sci-gaia.eu/.
At the time of writing (end of August 2015), it already contains all issued Sci-GaIA deliverables as well as the newsletter, the fact sheet of the project, some templates, and some components of the visual identity.

From the Open Science point of view:

- The Knowledge Base of e-Infrastructures has already been integrated in the project website at [http://www.sci-gaia.eu/e-infrastructures/knowledge-base](http://www.sci-gaia.eu/e-infrastructures/knowledge-base);
- The Sci-GaIA-OAR is already “registered” in the Knowledge Base as well as in the Semantic Search Engine\(^{65}\) that allows visitors to search for science products stored in the more than 4,000 repositories included so far in the Knowledge Base;
- The Sci-GaIA-OAR is also “connected” to the Africa Grid Science Gateway\(^{66}\) (the Africa Grid and the CHAIN-REDS Science Gateway are both based on the Catania Science Gateway Framework\(^{67}\)) and some applications to demonstrate science reproducibility and re-usability are already in progress to be integrated in that Science Gateway.

At the same time, the deployment of a service to issue Handle.Net-based Persistent IDentifiers (PIDs), to be assigned to the resources of Open Access repositories in the region, is in progress and, when completed, the Sci-GaIA-OAR will be connected to it.

### 4.2 – INSTALLATION AND CONFIGURATION INSTRUCTIONS

As already mentioned in Section 1, the Sci-GaIA Open Access Repository has the double function of being the project’s document repository (in order to comply with the Open Data Pilot launched by the EC) and the template to be cloned in many places in Africa to make science “made in Africa” more visible, re-producible and re-usable.

---


\(^{67}\) [http://www.catania-science-gateways.it/](http://www.catania-science-gateways.it/).
Concerning the second scope, the virtual appliance containing a clone of Sci-GaIA-OAR is available, as one of its own resources, at [http://oar.sci-gaia.eu/record/19](http://oar.sci-gaia.eu/record/19) and this demonstrates the achievement of milestone M3.1.

5 SUMMARY AND CONCLUSIONS

One of the objectives of the Sci-GaIA project is to strengthen and expand e-Infrastructure and Science Gateway related services. In particular, to expand and extend activities carried out in past projects in order to consolidate the African e-Infrastructure services and to include the very challenging goal of supporting the creation of an African Open (and Linked) Data Infrastructure, interoperable with and federated to (through the adoption of international standards and guidelines) those being consolidated in the EU and in other regions of the world.

Sci-GaIA also aims at exploiting Open Access repositories together with Science Gateways in order to deal with very important Open Science related topics such as the discoverability, reproducibility and reusability of science products. All this will make African science and scientists more visible and will allow the extension of the principles of the European Research Area well beyond its southern border.

Within this very fertile, exciting and fascinating landscape, paving the way for a second scientific revolution, the breakthrough concept of Open Science Commons is emerging as the ensemble of management principles that may actually implement Open Science guidelines. According to the Open Science Commons principles, “research data, scientific instruments, digital services (including those for data-intensive science), software, written knowledge (e.g., scientific publications, educational and training resources), expertise from people”, etc. are openly and widely shared across virtual research communities encompassing a variety of different disciplines.

Due to the huge importance of the above concepts and the key role they can play for the development of Africa and for making science “made in Africa” more visible worldwide, most of the work carried out by Sci-GaIA in its first months has been concentrated on setting up an Open Science Commons Platform for Africa comprising all the tools and services needed to enable Open Science workflows.

In the present deliverable, we have started presenting some key concepts and driving considerations, we have listed our requirements, choices and corresponding motivations, and we have finally shown the current implementation of the Sci-GaIA Open Access Repository, which is one of the main components of the Open Science Commons Platform for Africa mentioned above.